American Journal of Environmental Sciences 5 (5): 625-632, 2009
ISSN 1553-345X

Forecasting and Time Series Analysis of Air Pollutants in Several Area of Malaysia

[1]Mohd Zamri Ibrahim, [1]Roziah Zailan, [1]Marzuki Ismail and Muhd Safiih Lola
[1]Department of Engineering Science, Faculty of Science and Technology,
University Malaysia Terengganu, 21030, Kuala Terengganu, Malaysia
[2]Department of Mathematic, Faculty of Science and Technology,
University Malaysia Terengganu, 21030, Kuala Terengganu, Malaysia

Abstract: Problem statement: In keeping abreast with Malaysia’s rapid economic development and to meet the nation's aspiration for an improved quality of life, clean-air legislation limiting industrial and automobile emissions was adopted in 1978. **Approach:** Yet, to this day, air pollution from both sources still poses a problem for the nation. In order to predict the status of future air quality in Malaysia, a Box-Jenkins ARIMA approach was applied to modeling the time series of monthly maximum 1 h carbon monoxide and nitrogen dioxide concentrations in the east coast states of Peninsular Malaysia, i.e., Terengganu, Pahang and Kelantan, respectively, as well as to a comparison with the representative west coast state represent of Hulu Kelang. **Results:** In all the states, both carbon monoxide (CO) and Nitrogen dioxide (NO_2) concentrations have shown a fairly consistent upward trend since 1996. Nevertheless, the values forecast to 2016 for all states excluding NOx for Hulu Kelang did not exceed the permissible values given by either NAAQS or DOE Malaysia which are 35 and 30 ppm, respectively, at a 1 h average for CO and 0.053 and 0.17 ppm, respectively, for NOx. **Conclusion/Recommendations:** The forecasting values of each of the concentration parameters are still within a well-conserved condition as they do not exceed the limits of either NAAQS or DOE Malaysia excluding the values for nitrogen dioxide for Hulu Kelang.

Key words: ARIMA forecasting, time series, carbon monoxide, nitrogen dioxide, east coast peninsular Malaysia

INTRODUCTION

The time series forecasting approach is of useful for predicting future air quality status from various aspects of development in each country. The forecasting method analyzes the sequence of historical data in a period of time to establish the forecasting model. The ARIMA method has been extensively studied and used in previous research proven to be effective in the forecasting field. Forecasting methods applying the ARIMA time series method for pollution field have been expounded upon in many previous publications.

Air pollution data are obtained from the Air Quality Division of Alam Sekitar Malaysia Sdn. Bhd. (ASMA) which was awarded a concession by the government of Malaysia to set up a systematic and comprehensive monitoring network for air quality for the nation and to establish the National Environmental Data Centre since 1995. Currently, there are 52 continuous monitoring stations for ambient air and 20 manual air quality monitoring operations, managed and maintained by ASMA throughout Malaysia. The monitoring system employs the state-of the art instrumentation to continuously monitor the major pollutant gases in the air as well as to provide precise and accurate monitoring data[1].

Two critical pollutants, carbon monoxide and nitrogen dioxide, are considered because each of the data sets covers at least 10 years with no missing data in between and shows a fairly apparent either trend or seasonality, or both. Scientific research has proven that these two gases have many negative health effects, including some deadly diseases. Carbon monoxide is a significantly toxic gas that can lead to significant toxicity of the central nervous system and heart. Nitrogen dioxide is also toxic to humans since it can form nitric acid with water in the eyes, lungs, mucus

Corresponding Author: Mohd Zamri Ibrahim, Department of Engineering Science, Faculty of Science and Technology, University Malaysia Terengganu, 21030, Kuala Terengganu, Malaysia
Tel: +096683328 ext 3328 Fax: +6096694660

membranes and skin. Exposure to high concentrations of NO_2 can cause lung irritation and potentially lung damage. In this study, pollutant data of selected monitoring stations of Pahang, Terengganu, Kelantan and Hulu Kelang from the years 1996-2006 were analyzed to establish the forecasting model of these parameters as well as to observe the upcoming trend of these pollutants. Subsequently, the root causes of the pollution problem in this study area will be deliberated.

MATERIALS AND METHODS

Box-Jenkins ARIMA modeling: Monthly data covering the periods of 1997-2006 were acquired from the Air Quality Division of Alam Sekitar Malaysia Sdn. Bhd. (ASMA). The Box-Jenkins ARIMA model was used to model the time series behavior to generate the forecasting trend. ARIMA stands for Autoregressive Integrated Moving Average, with each term representing steps taken in the model construction until only random noise remains. The methodology consisting of a four-step iterative procedure was used in this study. The first step is tentative identification, where the historical data are used to tentatively identify an appropriate Box-Jenkins model. It is followed by estimation of the parameters of the tentatively identified model. After that, the diagnostic checking step must be executed to check the adequacy of the identified model in order to choose the best model. A better model should be identified if the model is inadequate. Finally, the best model is used to establish the time series forecasting value[6].

Identification steps: The MINITAB® statistical software package was used in this study. The first consideration of the data that are used is to ensure their stationarity condition. If the n values fluctuate with constant variation around a constant mean μ, it shows that the time series is stationary. The stationary time series value z_b, z_{b+1}, …, z_n can be determined through the behavior of the autocorrelation function (acf). If the acf of the time series values either cuts off fairly quickly or dies down fairly quickly, the time series value should be considered stationary. However, if it dies down extremely slowly, it should be considered non-stationary.

If the data are not stationary, a differencing process should be performed until an obvious pattern such as a trend or seasonality in the data fades away. This means, taking the divergence between consecutive observations, or between observations a year apart. The first differences of a non-stationary time series value y_1, y_2, …y_n are described as $z_t = y_t - y_{t-1}$ where t = 2,…, n. If the differences of a time series are still not stationary, the second differences should be implemented. The second differences of time series value y_1, y_2, …y_n are $z_t = y_t - 2y_{t-1} + y_{t-2}$ for t = 3, 4, ..., n[2].

Parameter estimation steps: Then, the plot of the acf and partial autocorrelation function (pacf) of the stationary data was examined to identify what autoregressive or moving average terms are suggested. The acf at lag k, denoted by ρ_k, is defined as:

$$\rho_k = \gamma_k/\gamma_0 \quad (1)$$

Where:
γ_k = The covariance at lag k
γ_0 = The variance

Since both covariance and variance are measured in the same units, ρ_k is a unitless and lies between -1 and +1. In the time series data, the main reason for correlation between z_t and z_{t-k} originates from the correlations that they have with intervening lags, z_{t-1}, z_{t-2}, …, z_{t-k+1}. The pacf measures the correlation between observations that are k time periods apart after controlling for correlations at intermediate lags. In other words, the pacf is the correlation between z_t and z_{t-k} after removing the effect of intermediate z's[3].

An acf with large spikes at initial lags that decays to zero or a pacf with a large spike at the first and possibly at the second lag indicates an autoregressive process. An acf with a large spike at the first and possibly at the second lag and a pacf with large spikes at initial lags that decay to zero indicate a moving average process. If both the acf and pacf exhibiting large spikes that gradually die out, this indicates both autoregressive and moving averages processes.

Autoregressive models (AR), Moving Average model (MA) and Autoregressive Integrated Moving Average models (ARIMA): An autoregressive model of order p, AR (p) has the form of:

$$z_t = \rho_1 z_{t-1} + \rho_2 z_{t-2} + \ldots + \rho_p z_{t-p} + \varepsilon_t \quad (2)$$

Each AR term corresponds to the use of a lagged value of the residual in the forecasting equation for the unconditional residual. The term 'autoregressive' refers to the fact that this model expresses the current time series values z_t as a function of past time series values $z_{t-1}, z_{t-2}, \ldots, z_{t-p}$. The $\rho_1, \rho_2, \ldots, \rho_3$ are unknown parameters relating z_t to z_{t-1}, z_{t-2}, ..., $z_{t-p.}$

A moving average forecasting model uses lagged values of the forecast error to improve the current forecast. A first-order moving average term uses the most recent forecast error, a second-order term uses the forecast error from the two most recent periods and so on. An MA(q) and has the form of:

$$zt = \varepsilon_t-\theta_1\varepsilon_{t-1}+-\theta_2\varepsilon_{t-2}- ...-\theta_q\varepsilon_{t-q} \quad (3)$$

Here:

$\varepsilon_{t-1},\varepsilon_{t-2},...,\varepsilon_{t-p}$ = The past random shocks

$\theta_1,\theta_2,...,\theta_q$ = Unknown parameters relating z_t to ε_{t-1}, ε_{t-2}, ..., ε_{t-p}

The autoregressive and moving average specifications can be combined to form an ARMA (p,q) specification:

$$z_t = \rho_1 z_{t-1}+\rho_2 z_{t-2}+\ldots+\rho_p z_{t-p}+\varepsilon_t-\theta_1\varepsilon_{t-1}-\theta_2\varepsilon_{t-2}-...-\theta_q\varepsilon_{t-q} \quad (4)$$

The point estimate for each parameter in a Box-Jenkins model is associated with its standard error and t-value. Each parameter is tested to determine whether it is zero (null hypothesis, H_o) or different from zero (alternative hypothesis, H_a)[3]. If the t>1.96, we can reject H_o: $\theta_1 = 0$ in favor of H_a: $\theta_1 \neq 0$ by setting α equal to 0.05.

Seasonal ARIMA model (SARIMA): Seasonality is defined as a pattern that repeats itself over fixed interval of time. In general, seasonality can be found by identifying a large autocorrelation coefficient or large partial autocorrelation coefficient at a seasonal lag. Often, autocorrelation at multiples of the seasonal lag will also be significant, such as at lag 24 or even lag 36. The seasonal differencing is the difference between an observation and the corresponding observation from the previous year. It used to obtain the stationary seasonal time series data, $z_t' = z_{t-} z_{t-s.}$ The seasonally differenced series, z_t, is the change between observations separated by s time periods, where s is the number of seasons. For monthly data, s = 12, for quarterly data, s = 4 and so on.

For the seasonal model, we used the Akaike Information Criterion (AIC) for model selection. The AIC is a combination of two conflicting factors: the mean square error and the number of estimated parameters of a model. Generally, the model with smallest value of AIC is chosen as the best model[5].

Forecasting stages: The final stage for the modeling process is forecasting, which gives results as three different options that are forecasted values and upper and lower limits that provide a confidence interval of 95%. Any forecasted values within the confidence limit are satisfactory. Finally, the accuracy of the model is checked with the Mean-Square error (MS) to compare fits of different ARIMA models. A lower MS value corresponds to a better fitting model.

RESULTS

Plots of raw data: The application chosen for this study is the concentration in ppm of pollutants (CO and NO_2) for the east coast and a comparison to Hulu Kelang representative of the west coast area of Peninsular Malaysia. The data used were monthly data from 1997-2006. The raw data for each parameter of every state are included in Fig. 1-4.

The model development process was begun by studying the original acf of the raw data. If the non-stationary condition emerges, the differentiation process will be executed to obtain the stationary time series. The number of lags to display the acf is 30. Then, the acf and pacf of difference were examined to determine the best combination of ARIMA model for each time series.

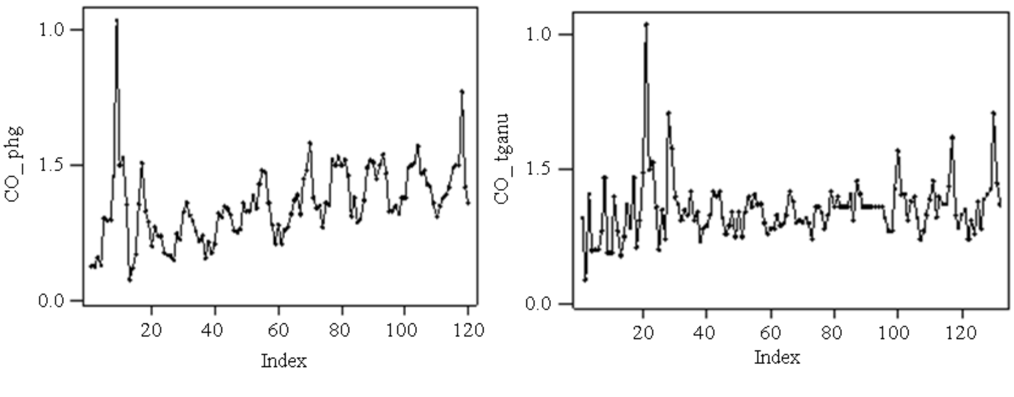

Fig. 1: Raw CO data (Pahang and Terengganu)

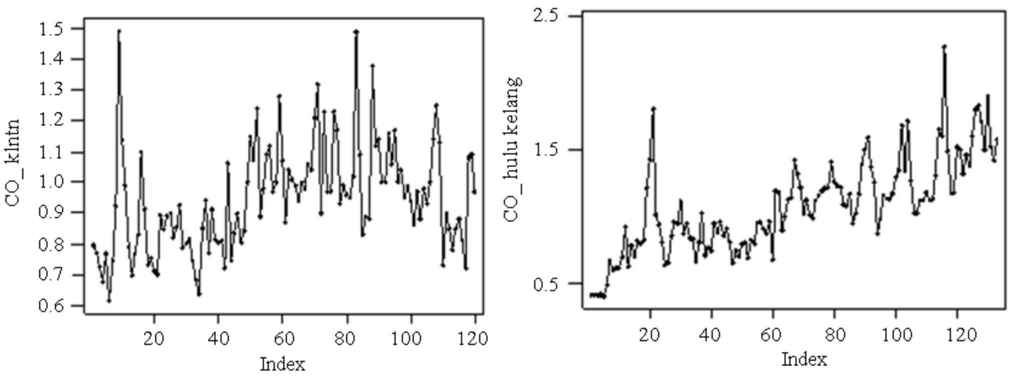

Fig. 2: Raw CO data (Kelantan and Hulu Kelang)

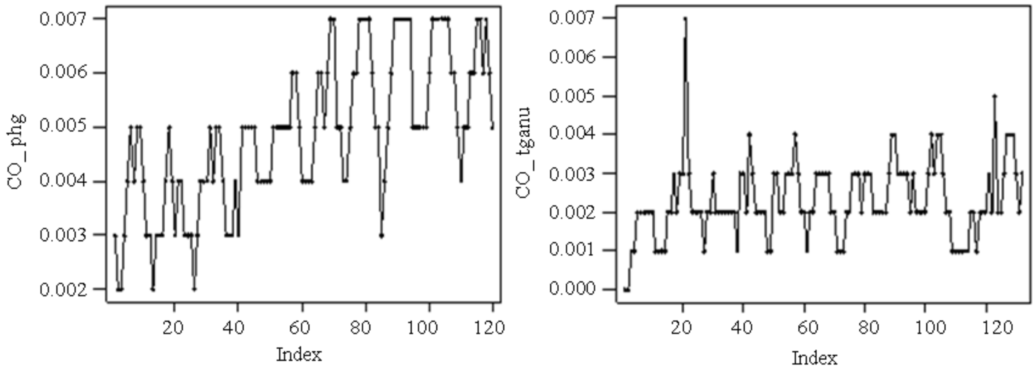

Fig. 3: Raw NOx data (Pahang and Terengganu)

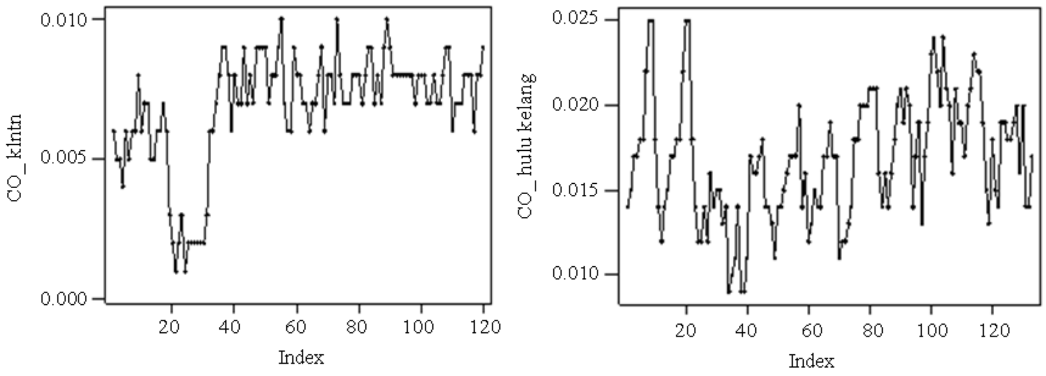

Fig. 4: Raw NOx data (Kelantan and Hulu Kelang)

After estimating all the possible models, the best fitted models were selected through a diagnostic checking procedure the t-value test for parameter estimation and the Q* value to determine if the model is satisfactory. Figure 5-8 show the forecasting graph for both pollutants.

Carbon monoxide forecasting model for Pahang, Terengganu and Kelantan: The prediction trend for carbon monoxide concentration is shown in Fig. 5 and 6 for each state from 1997 up to 2016. The selected models are the ARIMA model with first differentiation. However, the Kelantan data excluded the stage of differentiation as the acf of the original data is assumed stationary.

The carbon monoxide for Pahang exhibits the ARIMA (1,1,1). With a t value for AR (1) of 7.4>1.96 and for MA (1) of 44.62>1.96 and a Q* value equal to

13<18.3 at df = 10, we can assume that the best model for CO in Pahang is the mathematical expression:

$$z_t = 7.9\times10^{-4}+0.57z_{t-1}+\varepsilon_{t-}0.99\varepsilon_{t-1} \quad (5)$$

The CO model for Terengganu is the ARIMA (4,1,1). From the parameter estimation stages, the t values of each parameter are valid, as t for AR (1) = |-5.97|>1.96, AR (2) = |-3.22|>1.96, AR (3) = |-3.86|>1.96, AR (4) = |-2.86|>1.96, MA (1) = 474.85>1.96 and the Q* value is 9.3<14.06 at df = 7, thus, we can take this model as the best fitted:

$$z_t = -3.38\times10^{-5}+-0.52z_{t-1}-0.31z_{t-2}-0.37 z_{t-3}-0.25 z_{t-2}+\varepsilon_t -1.0\varepsilon_{t-1} \quad (6)$$

The CO for Kelantan is best described by the ARMA model (1,2) with a t value for AR (1), of 33.62>1.96 and for MA (1) = 5.98>1.96, for MA (2) = 3>1.96 and a Q* value equal to 14.9<16.91 at df = 9, we can assume that the best model for CO Kelantan is:

$$z_t = 0.02+0.98z_{t-1}+\varepsilon_t-0.57\varepsilon_{t-1}-0.28 \varepsilon_{t-2} \quad (7)$$

Whilst, the CO for Hulu Kelang is suit with ARMA (1,1,1) model. The parameter estimation shows that the t value of AR (1), |-2.7|>1.96 and MA (1), 192.26>1.96 and Q* value equal to 19.7<23.2 at df = 10, the model is expressed as

$$z_t = -0.0003-0.233z_{t-1}+\varepsilon_t-0.99\varepsilon_{t-1} \quad (8)$$

Nitrogen Dioxide forecasting model for Pahang, Terengganu and Kelantan: The predicted trends for nitrogen dioxide concentrations are displayed in Fig. 5. The selected models are the SARIMA model with differentiation at the seasonal level, the ARIMA model without differentiation for Terengganu and ARIMA with the first differentiation for Kelantan.

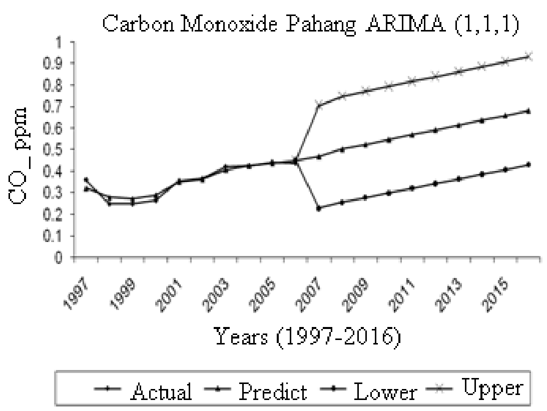

Fig. 5: Prediction model for Carbon Monoxide (Pahang and Terengganu, 1997-2016)

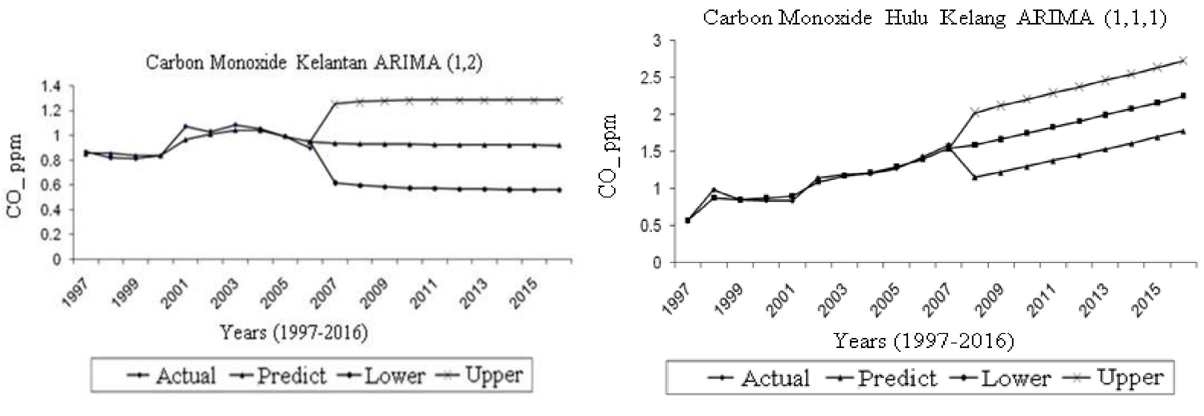

Fig. 6: Prediction model for Carbon Monoxide (Kelantan and Hulu Kelang, 1997-2016)

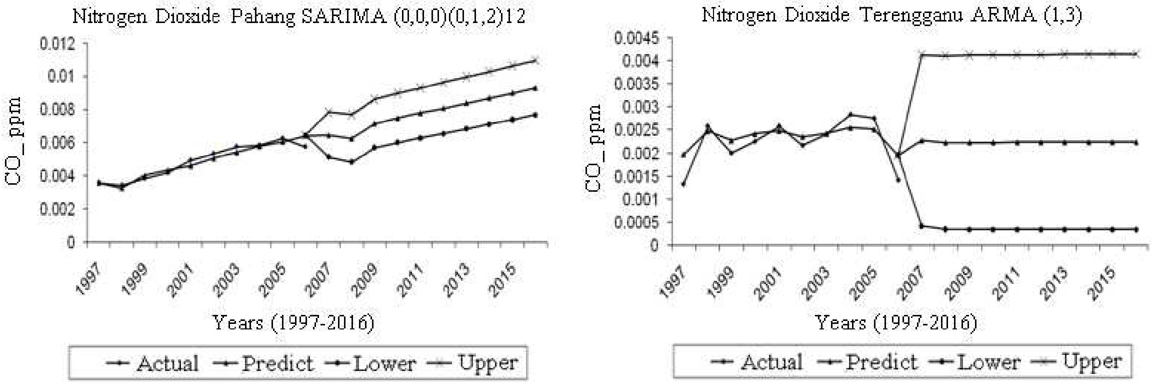

Fig. 7: Prediction model for Nitrogen Dioxide (Pahang and Terengganu, 1997-2016)

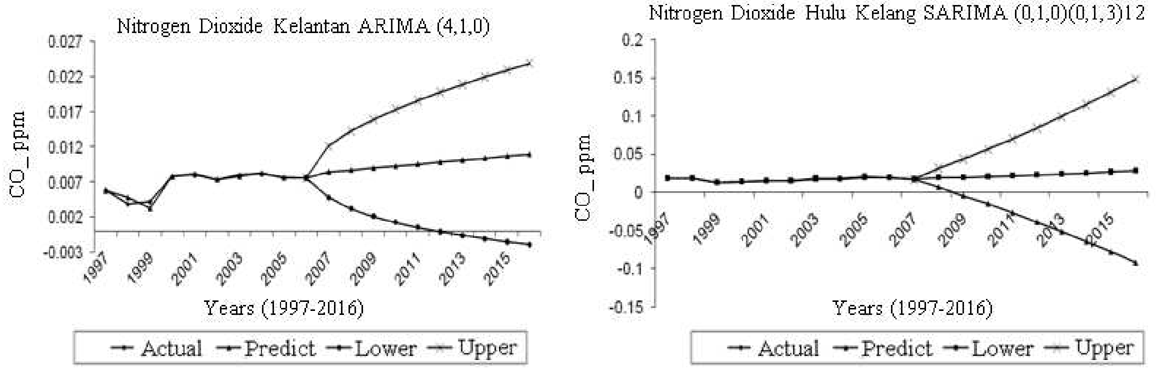

Fig. 8: Prediction model for Carbon Monoxide and Nitrogen Dioxide (Hulu Kelang, 1997-2016)

The most suitable model for NO_2 in Pahang is the SARIMA model because the raw data show a seasonal trend. The selected model is SARIMA (0,0,0)(0,1,2)12. We used the Akaike Information Criterion (AIC) for this model selection. The smallest AIC, which is -071.73708 determined that this is the best model. The given mathematical expression for this model is:

$$z_t = 1.375\varepsilon_{t-1}-0.379\varepsilon_{t-12}+\varepsilon_t \qquad (9)$$

Predicted NO_2 for Terengganu is suited to the ARMA (1,3) model. The parameter estimation gives t values of AR (1) = 13.55>1.96 and MA (1) = 17.86> 1.96, MA (3) = 3.05>1.96 and a Q* value equal to 9.4<15.5 at df = 8, so the model is expressed as:

$$z_t = 2.78\times10^{-4}+0.98_1z_{t-1}+\varepsilon_t-0.56\varepsilon_{t-1}-0.16\varepsilon_{t-2}-0.29\varepsilon_{t-3} \qquad (10)$$

In the other hand, NO_2 for Kelantan is adequate with ARIMA (4,1,0). The parameter estimation stage shows the t values of each parameter are |-11.21| for AR (1), |-6.46| for AR (2), |-4.84| for AR (3) and |-4.4| for AR (4) and Q* value is 18.4<20.09 at df = 8 and give out the equation:

$$z_t = 2.1\times10^{-5}-0.97z_{t-1}-0.74 z_{t-2}-0.56z_{t-3}-0.39 z_{t-4} \qquad (11)$$

Whilst, the suitable model for Hulu Kelang NO_2 is SARIMA (0,1,0)(0,1,3)12. The smallest AIC which is -895.12 determine that this is the best model. The given mathematical expression explained this model is:

$$z_t = z_{t-12}-+ \varepsilon t-1.239\varepsilon_{t-1}+0.51 \varepsilon_{t-2}-0.266 \varepsilon_{t-3} \qquad (12)$$

DISCUSSION

Carbon monoxide forecasting model for Pahang, Terengganu and Kelantan: From the forecasting graph in Fig. 4, it can be seen that the CO concentration for Pahang state increases steadily from the initial value of 0.36-0.68 ppm in 2016. For Terengganu state, the concentration of CO shows a slight increase to 0.42 in 2016 from an actual value of 0.26 in 1997. As for Kelantan, the value of the forecast concentration lies in the range of 0.9 ppm, close to the actual value for the initial year of 0.86 ppm. The rate of CO increase in Pahang and Terengganu states is rapid, unlike Kelantan. So far, the predicted values of CO for the states are still under the regulatory limits 30 ppm[7] or 35 ppm[4] for 1 h average CO concentration.

Nitrogen Dioxides forecasting model for Pahang, Terengganu and Kelantan: The predicted values of NOx for Pahang and Kelantan both increase evenly. The NOx in Pahang rises from 0.0035-0.009 at 2016 whereas, NOx in Kelantan rises from an initial actual value of 0.005 up to 0.011 ppm by the year 2016. The NOx concentration for Terengganu increases from 0.0013 ppm and varies steadily between 0.002 for the forecasted years. Pahang still stands to be the most polluted state on the east coast and it is most developed state among the three. Still, the NOx value is less than the DOE and NAAQS standards, which are 0.17 ppm and 0.053 ppm respectively.

Comparison of Carbon Monoxide and Nitrogen Dioxides forecasting model of East Coast area with Hulu Kelang: From the study, we can see that Pahang shows an incremental trend for both parameters, unlike, Terengganu and Kelantan. The increase in the pollutant levels can be related to the development level of the states. The construction of the East Coast Highway which connects Kuala Lumpur to Kuantan and Pahang and continues to Kuala Terengganu caused a large impact on the escalation of the pollutant concentrations. With the new linkage, many investors from other states will be interested to initiate business here. The industrialized sector will also further develop and increase the amount of transportation for both states simultaneously. Recently, the government also gave more attention to the east coast area by holding important events here. It mainly did this in order to expand the economy and tourism sector here, as the east coast of Peninsular Malaysia is well known as the major area for tourism. However, the actual and forecast pollutant values for all the states are considered harmless as they are under the permissible values of the DOE and NAAQS.

This condition differs from the study at west coast area which shows a higher value of pollutants, as illustrated in Fig. 6. The comparison between both areas shows that carbon monoxide for Hulu Kelang will increase tremendously from 0.56-2.25 ppm by 2016, while the nitrogen dioxide will increase from 0.0179 ppm to 0.028 by 2016 as well as exceeding the DOE and NAAQS standards. These values reflect the norm that the west coast area is highly polluted as compared to the east coast.

Talib *et al*.[7] also highlight that the highest concentration of CO was recorded in the Nilai Industrial Area with a concentration 4.35±0.80 ppm respectively, whiles the highest concentration of NO_2 was recorded in the Sepanggar Industrial Area (0.057±0.027 ppm). These two values are higher than the actual and predicted values of both parameters in the present study. A report from the Malaysia Meteorological Department also points out that, generally, the rainfall from the west coast of Peninsular Malaysia is more acidic than on the east coast of Peninsular Malaysia. This situation supports the finding of a less polluted condition in the East Coast area.

CONCLUSION

In summary, Hulu Kelang appears to be the most polluted state when compared with East Coast cities. Nevertheless, the forecasting values of each of the concentration parameters are still within a well-conserved condition as they do not exceed the limits of either NAAQS or DOE Malaysia excluding the values for nitrogen dioxide for Hulu Kelang. This condition appears to be the reason that the cities on the East Coast of Peninsular Malaysia are still not as developed as those in the West Coast area.

ACKNOWLEDGEMENT

The researchers would like to thank DOE Malaysia for providing pollutants data from 1997-2006 and the Ministry of Higher Education (MOHE) for allocating a research grant to accomplish this study.

REFERENCES

1. ASMA., 2008. Alam Sekitar Malaysia Sdn Bhd. http://www.enviromalaysia.com.my
2. Bowerman, B.L. and R.T. O’Connel, 1993. Forecasting and Time Series. An Applied Approach, Duxbury Press, Belmont, California.
3. Ediger, V.S. and S. Akar, 2007. ARIMA forecasting of primary energy demand by fuel in Turkey. Energy Policy, 35: 1701-1708. http://cat.inist.fr/?aModele=afficheN&cpsidt=18513226
4. Environment Protection Agency, 2008. National Ambient Air Quality Standards (NAAQS). http://www.epa.gov/air/criteria.html

5. Hong Wu, 1997. A time series analysis of United States carrot exports to Canada. http://en.scientificcommons.org/5979914
6. Hyndman, R.J., 2001. Box jenkins modeling. http://www.itl.nist.gov/div898/handbook/pmc/section4/pmc445.htm
7. Talib, M.L., M.O. Rozali, S. Norela, M.N. Ahmad Daud and N.J. Permata, 2002. Air quality in several industrial areas in Malaysia. Proceedings of the Regional Symposium on Environment and Natural Resources, Apr. 10-11, Malaysia, pp: 703-710. http://pkukmweb.ukm.my/~rsenr3/rsenr1/P703-710.pdf