ARIMA AND INTEGRATED ARFIMA MODELS FOR FORECASTING AIR POLLUTION INDEX IN SHAH ALAM, SELANGOR

Lim Ying Siew, Lim Ying Chin and Pauline Mah Jin Wee

International Education Centre (INTEC), Universiti Teknologi MARA Section 17 Campus, 40200 Shah Alam, Selangor.

Keywords: Air Pollution Index (API), Integrated Autoregressive Moving Average (ARIMA), Fractionally Integrated Autoregressive Moving Average (ARFIMA)

Abstract

Air pollution is one of the major issues that has been affecting human health, agricultural crops, forest species and ecosystems. Since 1980, Malaysia has had a series of haze episodes and the worst ever was reported in 1997. As a result, the government has established the Malaysia Air Quality Guidelines, the Air Pollution Index (API) and Haze Action Plan, to improve the air quality. The API was introduced as an index system for classifying and reporting the ambient air quality in Malaysia. The API for a given period is calculated based on the sub-index value (sub-API) for all the five air pollutants, namely sulphur dioxide (SO_2), nitrogen dioxide (NO_2), ozone (O_3), carbon monoxide (CO) and particulate matter below 10 micron size (PM_{10}). The forecast of air pollution can be used for air pollution assessment and management. It can serve as information and warning to the public in cases of high air pollution levels and for policy management of many different chemical compounds. Hence, the objective of this project is to fit and illustrate the use of time series models in forecasting the API in Shah Alam, Selangor. The data used in this study consists of 70 monthly observations of API (from March 1998 to December 2003) published in the Annual Reports of the Department of Environment, Selangor. The time series models that were being considered were the Integrated Autoregressive Moving Average (ARIMA) and the Integrated Long Memory Model (ARFIMA) models. The lowest MAE, RMSE and MAPE values were used as the model selection criteria. Between these two models considered, the integrated ARFIMA model appears to be the better model as it has the lowest MAPE value. However, the actual value of May 2003 falls outside the 95% forecast interval, probably due to emissions from mobile sources (i.e., motor vehicles), industrial emissions, burning of solid wastes and forest fires.

Introduction

Air pollution is one of the major issues that has been affecting human health, agricultural crops, forest species and ecosystems. Air quality monitoring is part of the initial strategy in the pollution prevention program in Malaysia. Since 1980, six major haze episodes were officially reported in Malaysia that is in April 1983, August 1990, June 1991, October 1991, August to October 1994 and July to October 1997. The 1997 haze episode was the worst ever experience in the country [3]. As a result, the government has established the Malaysian Air Quality Guidelines, the Air Pollution Index (API) and the Haze Action Plan in an attempt to improve the air quality.

There are possible health effects of exposure to air pollution. Recent studies have examined possible health effects of the 1997 forest fires. For example, respiratory disease outpatient who visited the Kuala Lumpur General Hospital increased from 250 to 800 per day and the data assembled indicated an increase in cases of asthma, acute respiratory infection, and conjunctivitis [1]. A study conducted by Nasir *et al*. [8] suggested that in the 1997 haze episode the total health effects were estimated to include 285,227 asthma attacks, 118,804 cases of bronchitis in children, 3889 cases of chronic bronchitis in adults, 2003 respiratory hospital admission, 26,864 emergency room visits and 5,000,760 restricted activity days. In addition, among the five pollutants, ozone was demonstrated to cause stress to the skin. It possesses a strong oxidizing potential and is therefore very reactive to the affected part [7]. Blockage of sunlight may also promote the spread of harmful bacteria and viruses that would otherwise be killed by ultraviolet B. Components of smoke haze, including polycyclic aromatic hydrocarbons known as carcinogens are also potentially dangerous and their effects may not be apparent for

years. The consequences may be more severe to children, for whom the particulates inhaled are high relative to their body size [4].

Since 1996, the National Environmental Research Institute [9], Denmark, has developed a comprehensive and unique integrated air pollution forecasting model system, called the THOR system. The system has been used to forecast the air pollution from accidental releases such as power plants, industrial sites and natural or human made fires. Dominici *et al.* [6] studied on the improved semi-parametric time series models of air pollution and mortality.

Clearly then from the above discussions, the modelling and ability to forecast API in Malaysia can be useful to many organisations like Environmental Agencies, Medical Research Institutes, Hospitals, etc. and the public at large. Therefore, the objective of this research is to model the API time series data in Shah Alam, Selangor so that such a model may be able to provide somewhat an estimate of the future API values.

Methodology

The Data Set

The Air Pollution Index (API) was introduced as an index system for classifying and reporting the ambient air quality in Malaysia. The API for a given period is calculated based on the sub-index value (sub-API) for all the five air pollutants, namely sulphur dioxide (SO_2), nitrogen dioxide (NO_2), ozone (O_3), carbon monoxide (CO) and particulate matter below 10 micron size (PM_{10}) which are included in the Malaysia API system. The API reference value has been based on the Malaysia Ambient Air Quality Guidelines (MAQG) of 1989 as shown in Table1 [2].

Table 1: API and Health Effect

API readings	Alert level	Health Effect Descriptor
0-50	No alert	Good
51-100	No alert	Moderate
101-150	Early alert	Unhealthy
151-200	Early alert	Unhealthy
210-300	On alert	Very unhealthy
301-500	Warming	Hazardous
>500	Emergency	Hazardous

Alam Sekitar Malaysia Berhad (ASMA) is a company that monitors the ground level ambient air continuously 24 hours a day. ASMA is responsible to install, operate, and maintain a network of 50 continuous air quality monitoring stations throughout Malaysia for the Department of Environment, DOE [7].

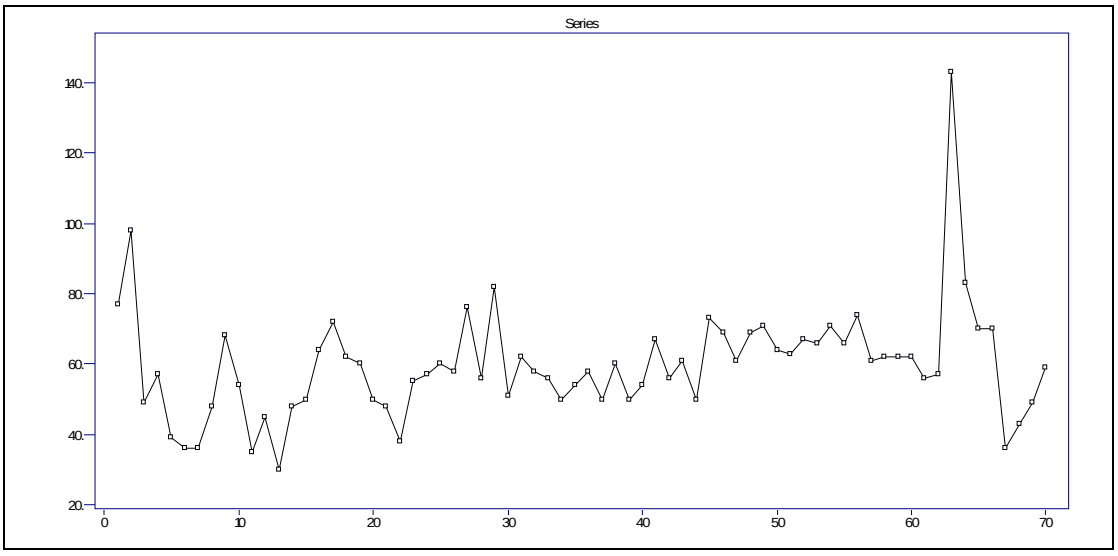

Figure 1: The time series plot of the monthly API data observed at the Shah Alam monitoring station from March 1998 to December 2003

The data used in this study consists of 70 monthly observations of API (from March 1998 to December 2003) published in the Annual Reports of the Department of Environment, Selangor [10]. The time series plot of API is shown in Figure 1.

Time Series Modelling Procedure

For the purpose of time series modelling in this study, the first 58 observations (March 1998 to December 2002) were used to fit the ARIMA and integrated ARFIMA models while the subsequent 12 observations (from January 2003 to December 2003) were kept for the post sample forecast accuracy check.

A brief description of the time series models and definitions used in this study are as follows. A stationary ARMA (*p*, *q*) model is defined as a sequence of random variables $\{X_t\}$, given by

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}$$

where $\{Z_t\}$ is a sequence of uncorrelated random variables with zero mean and constant variance, denoted as $\{Z_t\} \sim WN(0,\ \sigma^2)$, [5].

A process $\{X_t\}$ is called an ARIMA (*p*, *d*, *q*) process [5] if *d* is a nonnegative integer such that $(1-B)^d X_t$ is a causal ARMA (*p*, *q*) process. The ARIMA (*p*, *d*, *q*) processes satisfies the difference equation of the form

$$\phi^*(B) \equiv X_t \phi(B)(1-B)^d X_t = \theta(B) Z_t,\ \{Z_t\} \sim WN(0,\ \sigma^2),$$

where $\phi(z)$ and $\theta(z)$ are polynomials of degrees *p* and *q* respectively, and $\phi(z) \neq 0$ for $|z| \leq 1$. The $\phi^*(z)$ has a zero of order *d* at $z = 1$. The process $\{X_t\}$ is stationary if and only if $d = 0$, in which case it reduces to an ARMA (*p*, *q*) process.

A long memory process [5] or a fractionally integrated ARMA, ARFIMA (*p*, *d*, *q*) processes with $0 < |d| < 0.5$ is a stationary process with much more slowly decreasing autocorrelation function $\rho(k)$ at lag *k* as $k \to \infty$ which satisfies the property of $\rho(k) \sim Ck^{2d-1}$. The ARFIMA processes satisfy the difference equation of

$$(1-B)^d \phi(B) X_t = \theta(B) Z_t, \text{ where } \{Z_t\} \sim WN(0, \sigma^2),$$

$$\phi(z) = 1 - \phi_1 z - \ldots - \phi_p z^p \text{ satisfying } \phi(z) \neq 0 \text{ and}$$

$$\theta(z) = 1 + \theta_1 z + \ldots + \theta_q z^q, \text{ satisfying } \theta(z) \neq 0$$

for all *z* such that $|z| \leq 1$, and *B* is the backward shift operator. The operator $(1-B)^d$ is defined by the binomial expansion of

$$(1-B)^d = \sum_{j=0}^{\infty} \pi_j B^j \text{ with } n_0 = 1 \text{ and } \pi_j = \prod_{0<k\leq j} \frac{k-1-d}{k} \text{ for } j = 0, 1, 2, \ldots.$$

For the purpose of this study, we let $\{Y_t\}$ be the time series that represents the API and $\{y_t\}$ be the observed time series. Since there is a gradual decrease in level of $\{y_t\}$, we differenced the series at lag 1 to obtain a new series that is more or less constant in its level and this we denote it by $\{X_t\}$. To this $\{X_t\}$, we fitted an ARMA (*p*, *q*) process. The entire process of model fitting was done using the computer software "ITSM 2000, version 7.0", [5].

The criteria chosen to measure the accuracy of the forecast in this study are the mean absolute error (MAE), the root mean squared error (RMSE) and the mean absolute percentage error (MAPE) are given below.

$$\text{MAE} = \frac{\sum_{i=1}^{n} |x_i - \hat{x}_i|}{n}, \qquad \text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \hat{x}_i)^2}{n}}, \qquad \text{MAPE} = \frac{\sum_{i=1}^{n} \left| \frac{x_i - \hat{x}_i}{x_i} \right|}{n} \times 100\%$$

where x_i and $\hat{x}_i$ are the actual observed values and the predicted values respectively while n is the number of predicted values.

Results

In this section, we present the results of the study.

Let $\{ Y_t \}$ be the API and $X_t = \nabla Y_t$ where $\{ X_t \}$ is an ARMA (p, q) process. The best model fitted based on the AICC criterion that is given as follows:

$$X_t = 1.047X_{t-1} - 0.6656X_{t-2} - 0.1976X_{t-3} + Z_t - 1.771Z_{t-1} + 1.772Z_{t-2} - 0.7764Z_{t-3}$$

where $\{Z_t\} \sim WN(0, 0.027020)$.

The monthly forecast results of the API values using the ARIMA (3, 1, 3) model for the year 2003 are shown in Table 2.

Table 2: Forecasts of the API values from January 2003 to December 2003 using the ARIMA (3, 1, 3) model

Month	Actual	Forecast	95% Confidence Interval
January	62	61.23	(44.36, 84.50)
February	62	63.20	(45.25, 88.27)
March	56	65.61	(44.43, 96.88)
April	57	66.76	(43.81, 101.73)
May	143	65.70	(42.73, 101.04)
June	83	63.20	(41.08, 97.22)
July	70	60.93	(39.60, 93.73)
August	70	60.18	(39.03, 92.79)
September	36	61.15	(39.08, 95.68)
October	43	62.96	(39.12, 95.68)
November	49	64.17	(38.88, 105.92)
December	59	63.81	(38.23, 106.52)

In Figure 2, the graph of the predicted values given by the ARIMA (3, 1, 3) model and the actual values of the API, together with their 95% forecast intervals are shown. We note that the actual API values fall within the 95% confidence interval.

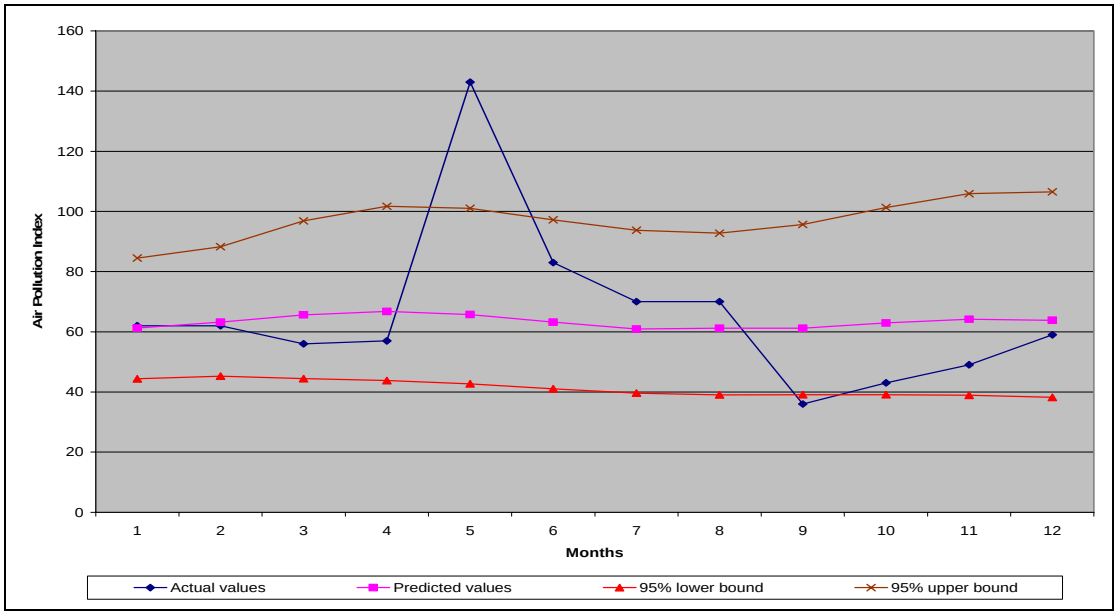

Figure 2: Graph of the API values with 12 predicted values of the ARIMA (3, 1, 3) model and the actual values from January 2003 to December 2003

For ARFIMA modelling, again, we let $\{Y_t\}$ be the API and $X_t = \nabla Y_t$ where $\{X_t\}$ is now an ARFIMA (p, d, q) process given by,

$$(1-B)^{-0.5} X_t = Z_t - 0.001520 Z_{t-1} + 0.3578 Z_{t-2},$$

where $\{Z_t\} \sim WN(0, 0.036980)$.

The monthly forecast results of the API values using the integrated ARFIMA (0, –0.5, 2) model for the year 2003 are shown in Table 3.

Table 3: Forecasts of the API values from January 2003 to December 2003 using the integrated ARFIMA (0, –0.5, 2) model

Month	Actual	Forecast	95% Confidence Interval
January	62	59.18	(33.60, 85.98)
February	62	58.56	(30.31, 87.25)
March	56	58.53	(25.18, 92.55)
April	57	58.13	(22.56, 94.44)
May	143	57.70	(20.60, 95.52)
June	83	57.26	(19.00, 96.22)
July	70	56.84	(17.63, 96.71)
August	70	56.44	(16.41, 97.05)
September	36	56.05	(15.31, 97.29)
October	43	55.68	(14.31, 97.46)
November	49	55.32	(13.38, 97.58)
December	59	54.97	(12.51, 97.65)

Figure 3 shows the predicted values given by the integrated ARFIMA (0, –0.5, 2) model and the actual values of the API, together with their 95% forecast intervals. The actual API values all fall within the 95% forecast intervals.

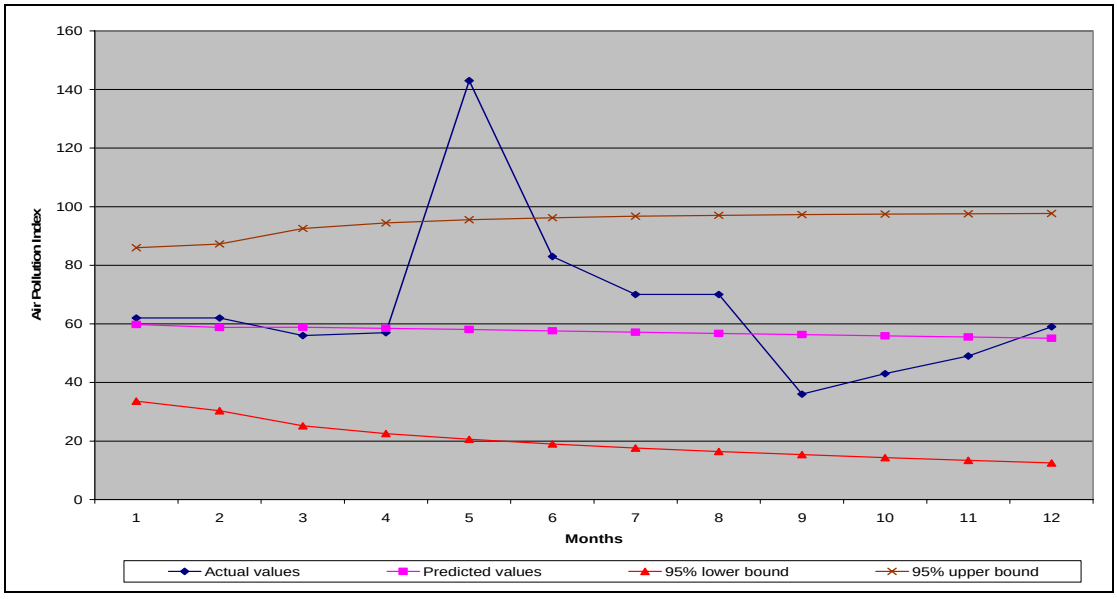

Figure 3: Graph of the API values with 12 predicted values of the integrated ARFIMA (0, –0.5, 2) model and the actual values from January 2003 to December 2003

In Table 4, the MAE, MAPE and RMSE values of the ARMA (3, 1, 3) and the integrated ARFIMA (0, –0.5, 2) models are shown.

Table 4: The MAE, MAPE and RMSE values of the ARMA (3, 1, 3) and the integrated ARFIMA (0, –0.5, 2) models

Model	MAE	RMSE	MAPE
ARIMA (3, 1, 3)	16.786	25.821	24.70%
ARFIMA (0, –0.5, 2)	15.896	27.299	20.86%

The MAE and MAPE values of the integrated ARFIMA (0, –0.5, 2) model are smaller when compared to those of the ARIMA (3, 1, 3) model. The RMSE value for ARIMA (3, 1, 3) is found to be smaller than the ARFIMA (0, –0.5, 2) model. The actual API value for May 2003 falls outside the 95% forecast intervals when forecasting using both the ARIMA (3, 1, 3) and ARFIMA (0, –0.5, 2) model. The ARIMA (3, 1, 3) model seem to be unable to forecast well as the actual API value for September 2003 is also found to be below the 95% lower bound of the forecast interval.

Conclusion

Based on this data set, the integrated ARFIMA model appears to have a slightly better forecasting performance compared to that of the ARIMA although both models are unable to forecast all values within the 95% forecast interval.

The actual API value of May 2003 which falls outside the 95% upper bound of the forecast interval may be due to the emissions from mobile sources like motor vehicles, industrial emissions, burning of solid wastes and forest fires. As such, factors which could affect the API should be taken into consideration in modelling of API for a better forecast ability as the modelling processes in this project were done based on the data of API only.

References

1. Awang M. B., Jaafar A. B., Abdullah A. M., Ismail M.B., Hassan M.N., Abdullah R., Johan S. and Noor H. (2000), Air Quality In Malaysia: Impacts, Management Issues And Future Challenges, *Respirology*, Vol. 5, pp 183-196.
2. Afroz R., Hassan M.N. and Ibrahim N.A. (2003), Review Of Air Pollution And Health Impacts In Malaysia, Environmental Research, Volume 92, Issue 2, pp 71-77.

3. Nasir M.H., Choo W.Y., Rafia A., Theng L.C., Noor M. M. H (2000), Estimation Of Health Damage Cost For 1997-Haze Episode In Malaysia Using the Ostro model, *Proceeding Malaysian Science And Technology Congress*, Confederation Of Scientific And Technological Association In Malaysia,COSTAM, Kuala Lumpur, in Press.
4. Faridah Mohamad (2002), Impacts Of Exposure To Ambient PM_{10} On Hospital Outpatient Visits For Haze-Related Diseases And School Children Lung Function, *Masters Thesis*. Universiti Putra Malaysia, Malaysia.
5. Beardsley R., Bromberg P.A., Costa D.A., Devlin R., Dockery D. W., Frampton M. W., Lambert W., Samet J. M., Speizer F. E., Utell M.(1997), Smoke Alarm: Haze From Fires Might Promote Bacterial Growth. *Science America*, pp 24-25.
6. National Environment Research Institute, NERI (2003), THOR. http://www.dmu.dk/1_viden/2_miljoe-tilstand/3_luft/4_spredningsmodeller/5_thor/default_en.asp. Accessed on 27 April 2004.
7. Dominici, F., McDermott, A., Hastie, T. J.(2004), Improved Semi-Parametric Time Series Models Of Air Pollution And Mortality. http://www-stat.stanford.edu/~hastie/Papers/dominiciR2.pdf. Accessed on 27 April 2004.
8. Awang M.B., Hassan M.N., Noor Alshurdin M.S., Abdullah A.M. (1997), Air pollution in Malaysia, in *IMR Quaterly Bulletin*, No. 43, pp 26-42.
9. *Annual Report* (1998-2003), Department of Environment, Selangor.
10. Brockwell, P.J. and Davis, R.A. (2002), *Introduction To Time Series And Forecasting*, 2nd Edition, Springer-Verlag, New York.