MATEC Web of Conferences 47, 05001 (2016)
DOI: 10.1051/matecconf/20164705001

Short Term Prediction of PM_{10} Concentrations Using Seasonal Time Series Analysis

Hazrul Abdul Hamid[1,a], Ahmad Shukri Yahaya[2], Nor Azam Ramli[2], Ahmad Zia Ul-Saufie[3] and Mohd Norazam Yasin[1]

[1]*Faculty of Civil and Environmental Engineering, Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Johor, Malaysia*
[2]*School of Civil Engineering, Universiti Sains Malaysia, 14300 Nibong Tebal, Penang, Malaysia*
[3]*Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia*

Abstract. Air pollution modelling is one of an important tool that usually used to make short term and long term prediction. Since air pollution gives a big impact especially to human health, prediction of air pollutants concentration is needed to help the local authorities to give an early warning to people who are in risk of acute and chronic health effects from air pollution. Finding the best time series model would allow prediction to be made accurately. This research was carried out to find the best time series model to predict the PM_{10} concentrations in Nilai, Negeri Sembilan, Malaysia. By considering two seasons which is wet season (north east monsoon) and dry season (south west monsoon), seasonal autoregressive integrated moving average model were used to find the most suitable model to predict the PM_{10} concentrations in Nilai, Negeri Sembilan by using three error measures. Based on AIC statistics, results show that ARIMA (1, 1, 1) x $(1, 0, 0)_{12}$ is the most suitable model to predict PM_{10} concentrations in Nilai, Negeri Sembilan.

1 Introduction

There are many sources of air pollution such as stationary sources, mobile sources, and open burning sources [1]. Air pollution emissions degrade air quality whether in urban or rural settings. An issue of great concern has been the detrimental effect of low air quality onto human health, chronically or acutely. Understanding the behaviour of air pollution statistically would allow predictions to be made accurately. In the early days of abundant resources and minimal development pressures, little attention was paid to growing environmental concern in Malaysia [1]. The Department of Environment is one of the bodies in Malaysia that is responsible in monitoring the status of air quality throughout the country to perceive any significant change which may cause harm to human health and environment.

In Malaysia, there are 52 monitoring locations throughout the country that belong to the Department of Environment [2]. The parameters monitored include Total Suspended Particulates, Particulate Matter (PM_{10}), Sulphur Dioxide (SO_2) and several airborne heavy metals. This 52 monitoring station was categorized in to 4 categories which is industrial, urban, sub-urban and background station. This research will focus on PM_{10} concentrations since previous researches have

[a] Corresponding author : hazrul@uthm.edu.my

shown that particles larger than 10 micrometer in aerodynamic diameter did not penetrate the body's defences in nose, mouth, and upper airways so it is unlikely to cause respiratory effects [3]. PM_{10} is also a useful indicator of several sources of outdoor air pollution, such as fossil-fuel combustion and also contaminants resulting from motor vehicle emissions [4]. Increase in ambient PM_{10} concentration can lead to significant impacts on the respiratory health especially for children, the elderly and susceptible individuals, which are normally associated with reduced lung function, asthma, pneumonia, bronchitis and emphysema [5]. At extremely high levels and long term exposure, it may even cause death. Malaysia Ambient Air Quality Guidelines state that the 24-hour mean for PM_{10} concentration should not exceed 150μg/m^3 and 12-month average should not exceed 50 μg/m^3.

Selecting appropriate time series models for the data is an important step to make prediction. These time series models may become the basis for estimating the parameters to meet the evolving information needs of environmental quality management. The developed models also can be used by related bodies to provide an early warning to the respective population. Since Malaysia has two seasons which is wet season (north east monsoon) and dry season (south west monsoon) [6], this research will focused on seasonal autoregressive integrated moving average models (Seasonal ARIMA). This research was carried out to find and proposed the most suitable time series model to predict the PM_{10} concentrations in Nilai, Negeri Sembilan, Malaysia using seasonal time series model.

2 Materials and Method

2.1 Study area

Nilai is a town located in Negeri Sembilan, Malaysia and this location classified as an industrial area by Department of Environment Malaysia. Geographically located at latitude $2^0 45'$ N of the equator and longitude$102^0 15'$ E of the prime meridian, Nilai is a rapidly growing town due to its proximity and easy connection to Kuala Lumpur using the existing highway. The data used in this study is hourly PM_{10} concentration in Nilai, Negeri Sembilan taken from April 2008 to March 2009. Mean top bottom method were used to replace the missing values where the data were filled with the average of data available above and below the missing values.

2.2 Procedure of time series analysis

A time series is a sequence of values $\{y_1, y_2, y_3, ..., y_{t-1}, y_t, ...\}$ observed through time. Time series analysis involves the statistical method of the analysis of a sequence data. There are three phases in time series modelling which is identification, estimation and testing [7].

2.3 Seasonal ARIMA (*p*, *d*, *q*) x (*P*, *D*, *Q*)$_s$

Time series having a trend or a seasonal pattern are not stationary in mean. So, ARIMA models cannot really cope with seasonal behaviour. Since this study focus on wet season and dry season in Malaysia, seasonal ARIMA time series model has been used. Dry season period is from April to September (south west monsoon) and dry season is from September to March (north east monsoon) [6]. Seasonal ARIMA (*p*, *d*, *q*) x (*P*, *D*, *Q*)$_s$ are defined by six parameters as follow:

$$\begin{gathered}\underbrace{\left(1-\varphi_1 B-\varphi_2 B^2-...-\varphi_p B^p\right)}_{AR(p)}\underbrace{\left(1-\beta_1 B^s-\beta_2 B^{2s}-...-\beta_P B^{Ps}\right)}_{SAR(P)}\underbrace{\left(1-B\right)^d}_{I(d)}\underbrace{\left(1-B^s\right)^D}_{I_s(D)} y_t = \\ c+\underbrace{\left(1-\psi_1 B-\psi_2 B^2-...-\psi_q B^q\right)}_{MA(q)}\underbrace{\left(1-\theta_1 B^s-\theta_2 B^{2s}-...-\theta_Q B^{Qs}\right)}_{SMA(Q)}\varepsilon_t\end{gathered} \tag{1}$$

where:

AR(p)	= autoregressive part of order p
MA(q)	= moving average part of order q
I(d)	= differencing of order d
SAR(P)	= seasonal autoregressive part of order P
SMA(Q)	= seasonal moving average part of order Q
$I_s(D)$	= seasonal differencing of order D
s	= the period of the seasonal pattern appearing

2.4 Stationarity of time series

Time series methods are typically based on stationarity. For stationary time series, the value of mean and variance are assumed constant with time [8]. There are many types of test that can be used to determine the stationarity of the time series. The most commonly used test for stationarity is augmented Dickey-Fuller test [9]. The Augmented Dickey-Fuller (ADF) test uses the following equation:

$$\mathrm{ADF} = \alpha_0 + p_1 y_{t-1} + \sum_{j=2}^{p-1} \beta_j \nabla y_{t-j} + e_t \tag{2}$$

where:

α_0 = drift components
e_t = error term

Abdel-Aziz and Frey [10] stated that the hypothesis to determine the stationary of the data is as follows:

H_o : the time series data is non-stationary
H_1 : the time series data is stationary

The null hypothesis will be rejected if the ADF value is greater than the critical value.

2.5 Model identification and estimation

This study used Akaike Information Criterion (AIC) to identify the suitable model. AIC is the most commonly used criteria to select the best model in time series analysis [11]. The smallest AIC statistics indicates the best time series model. The Akaike Information Criterion (AIC) is define as follow:

$$\mathrm{AIC} = -2\ln \hat{\sigma}_a^{\,2} + 2M \tag{3}$$

where,

$\hat{\sigma}_a^{\,2}$ = maximum likelihood estimate of $\sigma_a^{\,2}$
M = effective number of observations that is equivalent to the number of residuals that can be calculated from the series

2.6 Validation

The error measures were used to judge the developed model. Three performance indicators which are mean absolute percentage error (MAPE), normalized absolute error (NAE) and root mean square error (RMSE) as shown in Table 1 has been used to evaluate and validate the time series model [12, 13].

Table 1. Error measures.

Error Measure	Formula	Description
Mean Absolute Percentage Error (MAPE)	$MAPE = \dfrac{\sum_{i=1}^{N}\left\lvert\dfrac{(O_i - P_i)}{O_i}\right\rvert}{N} \times 100$	The smallest value of mean absolute percentage error indicates the best model
Normalized Absolute Error (NAE)	$NAE = \sum_{i=1}^{N}\lvert P_i - O_i\rvert \Big/ \sum_{i=1}^{N} O_i$	A small value for the normalized absolute error means that the model is appropriate.
Root Mean Square Error (RMSE)	$RMSE = \sqrt{\dfrac{1}{N}\sum_{i=1}^{N}(P_i - O_i)^2}$	A smaller RMSE value means that the model is more appropriate.

* N is the number of monitoring records, O_i is the observed monitoring records and P_i is the predicted monitoring records.

3 Results and Discussion

Descriptive statistics for hourly PM_{10} concentration in Nilai, Negeri Sembilan from April 2008 to March 2009 shown in Table 2. As an industrial area, the maximum concentration of PM_{10} is 96µg/m^3, less than Malaysia Ambient Air Quality Guideline (MAAQG) and these data is positively skewed which means that most of the readings are below the mean value.

Table 2. Descriptive statistics for PM_{10} concentration for Nilai.

Mean	55.56
Std deviation	14.51
Median	55
Mode	51
Maximum	96
Skewness	0.18
Kurtosis	0.05

Since this study considered the wet season and dry season in Malaysia, each set of data was start at 1st of April and end at 31st of March. Dry season period is from April to September which is during south west monsoon and wet season is from September to March which is during north east monsoon. Figure 1 show the time series plot for PM_{10} concentration. Since the observation from time series plot sometimes may not able to give clear information about the stationarity, the augmented Dickey-Fuller (ADF) test was performed to check the stationarity of this series.

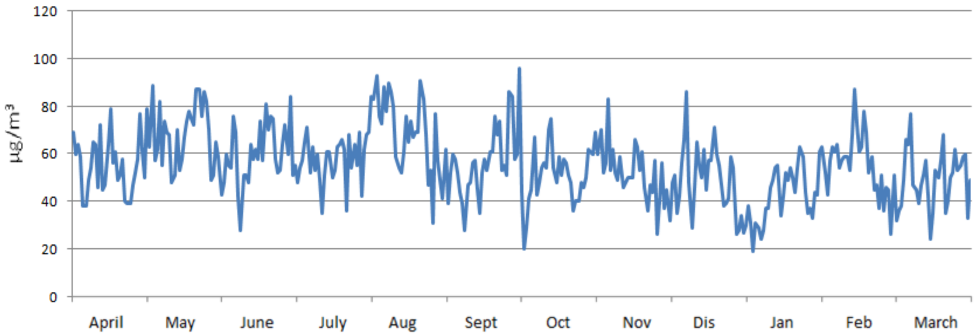

Figure 1. Time series plot of PM10 concentration for Nilai, Negeri Sembilan.

Result of Augmented Dickey Fuller test shows that the monitoring records of PM_{10} concentration from April 2008 to March 2009 are stationary with ADF statistic –4.9718 (p-value is less than 0.001). Seasonal Autoregressive Integrated Moving Average model (Seasonal ARIMA model) has been considered to find the most appropriate model for these set of data by comparing the AIC statistics values. Seasonal ARIMA (1, 0, 1) x $(0, 0, 1)_{12}$ with the smallest AIC statistics value 7.789 is the most appropriate model that give the most accurate forecasting for PM_{10} concentration in Nilai, Negeri Sembilan for composite years from April 2008 to March 2009. Time series equation based on the most appropriate time series model for Nilai is as follows.

$$y_t = 57.15 + 1.27 y_{t-1} - 0.31 y_{t-2} + 0.04 y_{t-3} - 0.009 y_{t-4} + 0.88 e_{t-1} + e_t \quad (4)$$

The performance indicators for the time series model are summarized in Table 3. For the purpose of seeing more clearly, the 15 days observed and forecasted of PM_{10} concentration using the most appropriate time series model were plotted in Figure 2.

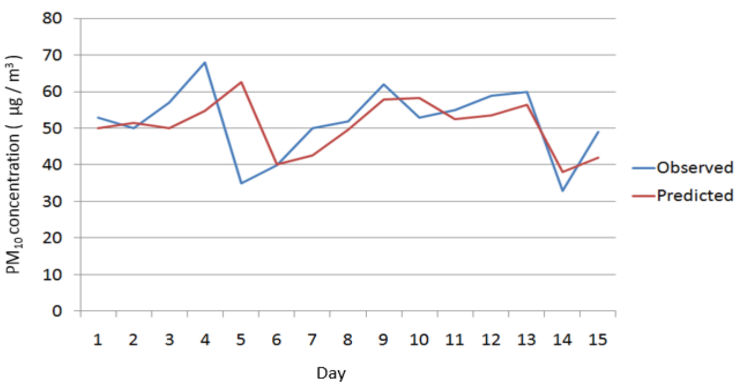

Figure 2. Observed and predicted of PM10 concentration using Seasonal ARIMA (1,0,1) x $(0,0,1)_{12}$.

Table 3. Model performance indicator.

Monitoring Site	Performance indicator (Error measure)		
	RMSE	NAE	MAPE
Nilai	11.9920	0.89053	10.2130

4 Summary

The hourly average actual monitoring record of PM_{10} concentration was used to find the best time series models. By using three error measure, seasonal ARIMA (1, 0, 1) x $(0, 0, 1)_{12}$ model is the most suitable model to forecasting the PM_{10} concentration in Nilai, Negeri Sembilan. The results of this study provide useful information on air quality status in Nilai Negeri Sembilan and can be used for prediction of future reading and also for air quality management.

References

[1] R. Afroz, M.N. Hassan and N. A. Ibrahim, Review of air pollution and health impacts in Malaysia, Environment Research, **92**(2), 71-77, (2003).

[2] Department of Environment Malaysia, Malaysia Environmental Quality Report 2013, Kuala Lumpur: Department of Environment, Ministry of Natural Resources and Environment, Malaysia, (2013).

[3] D.W. Dockery, Health effects of particulate air pollution, Ann Epidemiology, **19**, 257-263, (2009).
[4] N. Kunzli, R. Kaiser, S. Medina, M. Studnicka, O. Chanel, P. Filinger, M. Herry, J.F. Horak, T.V. Puybonnieux, P. Quenel, J. Schneider, R. Seethaler, J.C. Vergnaud and H. Sommer, Public health impact of outdoor and traffic related air pollution: a European assessment, The Lancet, **356**, 795-801, (2000)
[5] B.M.T. Shamsul, Paras pendedahan kepada PM_{10} dan hubungannya dengan simptom masalah pernafasan di kalangan pekerja Majlis Perbandaran Petaling Jaya. Master of Medical Sciences Thesis. Universiti Kebangsaan Malaysia, (2002)
[6] L.H.L. Oliver, S. Ahmad, K. Aiyub, Y.M. Jani and T.K. Hwa, Urban environmental health : Respiratory illness and urban factors in Kuala Lumpur city, Malaysia. Environment Asia, **4**(1), 39 – 46, (2011)
[7] T.M.J.A. Cooray, *Applied Time Series Analysis and Forecasting*, Alpha Science, Oxford, (2008).
[8] W.S.W. William, *Time Series Analysis-Univariate and Multivariate Methods,* Pearson Education, United States, (2006)/
[9] R.A. Carmona, *Statistical Analysis of Financial Data in S-Plus*, Springer, United States, (2004).
[10] A. Abdel-Aziz and H.C. Frey, Development of hourly probabilistic utility NOx emission inventories using time series techniques: Part I-univariate approach, Atmospheric Environment, **37**, 5379-5389, (2003).
[11] H. Bozdogan, Model selection and Akaike information criterion (AIC): The general theory and its analytical extensions, Psychometrika, **52**(3), 345-370, (1987).
[12] J. Tayman and D.A. Swanson, On the validity of MAPE as a measure of population forecast accuracy, Population Research and Policy Review, **18**, 299 - 322, (1999).
[13] H. Junninen, H. Niska, K. Tuppurainen, J. Ruuskanen and M. Kolehmainen, Method for impotation of missing values in air quality data sets, Atmospheric Environment, **38**(9), 2895-2907, (2004).