Journal of Environmental Science and Technology

ISSN 1994-7887

science alert

ANSInet
an open access publisher
http://ansinet.com

OPEN ACCESS Journal of Environmental Science and Technology

ISSN 1994-7887
DOI: 10.3923/jest.2017.96.106

Research Article
Modeling of Daily PM_{10} Concentration Occurrence Using Markov Chain Model in Shah Alam, Malaysia

[1]Norsalwani Mohamad, [1]Sayang Mohd Deni and [2]Ahmad Zia Ul-Saufie Mohamad Japeri

[1]Faculty of Computer and Mathematical Sciences, Universiti Teknologi Mara (UiTM), Shah Alam, Malaysia
[2]Department of Computer and Mathematical Sciences, Universiti Teknologi Mara (UiTM), Pulau Pinang, Malaysia

Abstract

Background and Objective: The analysis of the behavior of daily PM_{10} occurrence is becoming important nowadays and the results obtained may be useful for the prediction and decision making purposes. This study considered the behavior of PM_{10} concentration that related with its dependency nature. Therefore, this study is attempted to determine the sequences of polluted and non-polluted days affected by PM_{10} concentration based on the optimum order of a Markov chain model. **Methodology:** Twelve years of monitoring records which is from 2002-2013 and have been analyzed for this purpose. The PM_{10} concentration data that possess Markov chain properties show that the successive event is dependent on the previous event and is suited for further analysis using this model. **Results:** The optimum order of the Markov chain model for Shah Alam monitoring station shows that the order of two and three are optimum for threshold values less than 120 μg m^{-3} and a simple order is optimum for a threshold value of 150 μg m^{-3}. The results mean that the occurrence of the polluted or non-polluted days affected by PM_{10} is dependent on the 2 or 3 days before the observed day for threshold value less than 120 μg m^{-3}. For a threshold value of 150 μg m^{-3}, the occurrence depends only on a day before the observed day. Besides that, the distribution of polluted events is well fitted based on the optimum order for each threshold value used. **Conclusion:** The information of polluted (non-polluted) occurrences is important in monitoring the PM_{10} concentrations which can be used for predicting related future events and helpful in providing the necessary precautionary measures to public and protect their health.

Key words: PM_{10} concentration, Markov chain model, occurrence, optimum order, sequence of polluted (non-polluted) days, prediction, dependency, behavior

Received: December 23, 2016 Accepted: January 16, 2017 Published: February 15, 2017

Citation: Norsalwani Mohamad, Sayang Mohd Deni and Ahmad Zia Ul-Saufie Mohamad Japeri, 2017. Modeling of daily PM_{10} concentration occurrence using markov chain model in shah alam, Malaysia. J. Environ. Sci. Technol., 10: 96-106.

Corresponding Author: Ahmad Zia Ul-Saufie Mohamad Japeri, Department of Computer and Mathematical Sciences, Universiti Teknologi Mara (UiTM), Pulau Pinang, Malaysia Tel: + (60)4-3823428

Competing Interest: The authors have declared that no competing interest exists.

Data Availability: All relevant data are within the paper and its supporting information files.

INTRODUCTION

The PM_{10} particulate matter with an aerodynamic diameter of less than 10 μm, is identified as a type of air pollution that causes the greatest concern to public health and also the environment. This pollutant is also the main air pollutant exists during the haze events in Malaysia since 1980[1,2]. Besides that, this pollutant can result in short and long term health impacts. The presence of PM_{10} in ambient air may cause severe health effects such as asthma, throat irritation, respiratory problems and even hospital admission[3]. There are five major sources of PM_{10} emissions: Power plants and heat, motor vehicle exhausts, industrial sources and open burning[4]. However, the most predominant sources of PM_{10} emissions in Malaysia are heavy traffic and industrial emissions[5].

Statistical modelling has allowed environmental authorities to carry out daily air pollution forecasts since this model provides a good insights in short term predictions of future air pollution levels. The regression models and Artificial Neural Network (ANN) are commonly used in predicting the PM_{10} concentrations by previous studies[6-8]. Besides that, Central Fitting Distributions (CFD) and Extreme Value Distributions (EVD) can also yield good results to fit the mean concentrations of PM10[9]. There have been many efforts made in monitoring this air pollutant. However, the sequence of polluted and non-polluted days affected by PM_{10} still receives less attention among researchers. A Markov chain model is dependent on its previous state and this model is highly suited to the pattern of observations. Hence, once the patterns are identified, it is possible to predict the possibility of future events based on the information of previous day event.

Markov chain models are intended to be simple models requiring only two parameters and fitting various aspects of occurrence patterns. Simple Markov chain models are widely used in describing the sequences of daily rainfall occurrences all over the world including Malaysia Chin[10], Deni *et al.*[11] and Gabriel and Neumann[12]. The use of this method is also beneficial in describing the sequences of daily PM_{10} occurrences due to this pollutant being controlled by weather conditions and showing similar persistence[13]. Rahimi *et al.*[14] used Markov chain models in order to study the persistence of days affected by PM_{10} in Tehran and found that the first order of two states Markov chain models had a good fitting on the data of five selected stations. Furthermore, this model had been applied to other air pollution data such as those in studies by Lin[13] and Lin and Huang[15].

Even though there are still few studies on PM_{10} concentration that apply this model, the advantage of using a Markov chain model as a model suitable for future prediction considering previous events, have made this model useful to be applied in this study. Subsequently, this model is also frequently used to forecasts the weather at some future time by given the current state as reported by previous studies such as Mangaraj *et al.*[16], Deni *et al.*[11] and Chin[10].

Although, the first order of Markov chain models is simple and the calculation is easier than the higher order, but according to Chin[10], a Markov chain model cannot be assumed to always be one because sometimes it is inadequate to give an appropriate model. Besides that, since the effect of being exposed to PM_{10} is more than 24 h as reported by World Health Organization (WHO)[17], thus, there is a need to use a higher order of Markov chain models in describing the sequence of PM_{10} concentration (consider more than one previous events) in order to obtain a better prediction of PM_{10} occurrences and improving the quality of reports from the data generations. Thus, in this study, simple and higher orders are considered. The aim of this study was to determine the optimum order of Markov chain model in describing the sequence of polluted (non-polluted) days of PM_{10} concentration in Shah Alam, Malaysia.

MATERIALS AND METHODS

The air quality in Malaysia is monitored by the Department of Environment (DoE) through 52 continuous monitoring stations. These stations are strategically located in order to detect any significant changes of air quality in that area. This study considers the PM_{10} concentration data from Shah Alam, an urban area. SekolahKebangsaan Raja Muda, Shah Alam, is where the monitoring station was placed. The coordinates of this monitoring station reads 3.08° North latitude and 101.51° East longitude. The main contributor of PM_{10} concentration in this area is the emissions from motor vehicles, since Shah Alam is the state capital of Selangor, Malaysia and due to the increasing number of vehicles, as well as rapid urban development[18].

Twelve years worth of PM_{10} concentration data provided by DoE from year 2002 until year 2013 were used to achieve the objective of this study. In this study, a polluted day is defined as a day when the PM_{10} concentration exceeds the threshold value, while a non-polluted day is defined as a day when the PM_{10} concentration is less than the threshold value. For example, a day with PM_{10} concentration of more than 50 μg m^{-3} is a polluted day if the threshold value is 50 μg m^{-3} and if the value is less than 50 μg m^{-3}, it is considered a non-polluted day. The threshold values considered in this study are 50 μg m^{-3} (WHO guideline), 52 μg m^{-3} (Background concentration of PM_{10} at this station);

Table 1: Transition probabilities of the occurrence of PM_{10} concentration for Shah Alam station with threshold value of 50 μg m^{-3}

	Preceding days						Observation day (t)		
Order	t-4	t-3	t-2	t-1	Matrix condition		1	0	Total
0					n_1	n_0	2402	1981	4383
1				1	n_{11}	n_{10}	1900	503	2403
				0	n_{01}	n_{00}	502	1478	1980
2			1	1	n_{111}	n_{110}	1609	292	1901
			0	1	n_{011}	n_{010}	291	211	502
			1	0	n_{101}	n_{100}	182	321	503
			0	0	n_{001}	n_{000}	320	1157	1477
3		1	1	1	n_{1111}	n_{1110}	1407	203	1610
		0	1	1	n_{0111}	n_{0110}	202	89	291
		1	0	1	n_{1011}	n_{1010}	124	58	182
		0	0	1	n_{0011}	n_{0010}	167	153	320
		0	0	0	n_{0001}	n_{0000}	227	929	1156
		1	0	0	n_{1001}	n_{1000}	93	228	321
		0	1	0	n_{0101}	n_{0100}	50	161	211
		1	1	0	n_{1101}	n_{1100}	132	160	292
4	0	0	0	0	n_{00001}	n_{00000}	166	762	928
	0	0	0	1	n_{00011}	n_{00010}	111	116	227
	0	0	1	0	n_{00101}	n_{00100}	39	114	153
	0	0	1	1	n_{00111}	n_{00110}	111	56	167
	0	1	0	0	n_{01001}	n_{01000}	43	118	161
	0	1	0	1	n_{01011}	n_{01010}	29	21	50
	0	1	1	0	n_{01101}	n_{01100}	32	57	89
	0	1	1	1	n_{01111}	n_{01110}	154	48	202
	1	0	0	0	n_{10001}	n_{10000}	61	167	228
	1	0	0	1	n_{10011}	n_{10010}	56	37	93
	1	0	1	0	n_{10101}	n_{10100}	11	47	58
	1	0	1	1	n_{10111}	n_{10110}	91	33	124
	1	1	0	0	n_{11001}	n_{11000}	50	110	160
	1	1	0	1	n_{11011}	n_{11010}	95	37	132
	1	1	1	0	n_{11101}	n_{11100}	100	103	203
	1	1	1	1	n_{11111}	n_{11110}	1253	155	1408

100, 120 and 150 μg m^{-3} (New Malaysia ambient air quality standard[19]). The purpose of using various levels of threshold values is to investigate the effect of these values with the optimum order of Markov chain model.

The example of calculation in order to achieve the aim of this paper is illustrated at each section. For the example of calculation, all the values used were based on the 12 years data of PM_{10} concentration at Shah Alam monitoring station with threshold value of 50 μg m^{-3} and the transitions probabilities of the occurrence of PM_{10} concentration is shown in Table 1.

Markov chain property: The purpose of checking the Markov chain property is to statistically test whether or not the successive events are independent. Furthermore, according to Moon *et al.*[20], the successive events can form or possess Markov chain models when the events are dependent on each other. As for the statistics, α is defined as in Eq. 1 if the successive events are independent:

$$\alpha = 2\sum_{i,j}^{m} n_{ij} \ln(P_{ij} / P_{.j}) \quad (1)$$

where, the P_{ij} denotes the conditional probability of the jth day event depends on the ith day event and $P_{.j}$ is the probability of the jth day event. Equation 1 is distributed a symptomatically as χ^2 with degree of freedom of $(m-1)^2$. The m is the total number of state (in this case: m = 2) and the marginal probabilities for jth column of the transition probabilities ($P_{.j}$):

$$P_{.j} = \sum_{i}^{m} n_{ij} \Big/ \sum_{i,j}^{m} n_{ij} \quad (2)$$

For example, to obtain the value of the statistics, α for threshold value of 50 μg m^{-3}, the calculation is shown as given below and all the values used in the calculation are obtained from the transition probabilities:

$$\alpha = 2\sum_{i,j}^{m} n_{ij} \ln(P_{ij} / P_{.j})$$

$$= 2\left[\begin{array}{l} \left(n_{00} \ln\left(\frac{\frac{n_{00}}{n_{00}+n_{01}}}{\frac{n_0}{n_0+n_1}}\right)\right) + \left(n_{01} \ln\left(\frac{\frac{n_{01}}{n_{01}+n_{00}}}{\frac{n_1}{n_0+n_1}}\right)\right) + \\ \left(n_{10} \ln\left(\frac{\frac{n_{10}}{n_{10}+n_{11}}}{\frac{n_0}{n_0+n_1}}\right)\right) + \left(n_{11} \ln\left(\frac{\frac{n_{11}}{n_{11}+n_{10}}}{\frac{n_1}{n_0+n_1}}\right)\right) \end{array}\right]$$

$$= 2\left[\begin{array}{l} \left(1478 \ln\left(\frac{\frac{1478}{1478+502}}{\frac{1981}{1981+2402}}\right)\right) + \\ \left(502 \ln\left(\frac{\frac{502}{502+1478}}{\frac{2402}{1981+2402}}\right)\right) + \\ \left(503 \ln\left(\frac{\frac{503}{503+1900}}{\frac{1981}{1981+2402}}\right)\right) + \\ \left(1900 \ln\left(\frac{\frac{1900}{1900+503}}{\frac{2402}{1981+2402}}\right)\right) \end{array}\right]$$

$$\alpha = 663.89$$

Determination the optimum order of Markov chain models for occurrence of PM_{10} concentration: The sequence of polluted (non-polluted) days of daily PM_{10} concentration is denoted as X_1, X_2, X_3,..., X_t,...X_n, for n-arbitrary days. A two-state Markov chain model was considered in this study where one denotes a polluted day 1 and 0 denotes a non-polluted day 0. The sequence of polluted (non-polluted) days is assumed to follow a Markov chain of a first order at time t, when X_1t depends on previous events, X_{t-1}. Thus, the two conditional probabilities of the first order can be given by $P_{10} = (X_t = 0 | X_{t-1} = 1)$ and $P_{11} = P(X_t = 1 | X_{t-1} = 1)$. The transition probability matrix P, which describes the 2-state Markov chain model is as in Eq. 3[16]:

$$P = \begin{pmatrix} P_{00} & P_{01} \\ P_{10} & P_{11} \end{pmatrix} \quad (3)$$

where, $P_{ij} = P(X_1 = j | X_0 = i)$ i. j = 0,1.

Note that $P_{00}+P_{01} = 1$ and $P_{11}+P_{10} = 1$. The definition of the conditional probabilities is as follows:

P_{00}: The probability of a day being non-polluted given that the previous day was also a non-polluted day

P_{01}: The probability of a day being polluted given that the previous day was a non-polluted day

P_{10}: The probability of a day being non-polluted given that the previous day was a polluted day

P_{11}: The probability of a day being polluted given that the previous day was also a polluted day

As for the assumption that the Markov chain is stationary, the transition probabilities of the k^{th} order are as in Eq. 4 and the joint probability distribution for X_1, X_2, X_3,..., X_t,...X_n is as in Eq. 5[10]:

$$P_{i_1,...,i_{k+1}} = P\left(X_t = i_{k+1} \middle| X_{t-1} = i_k,..., X_{t-k} = i_1\right); \quad i = 0, 1 \quad (4)$$

$$P(X_n = i_n,..., X^1 = i_1) = P(i_n, i_{n-1},..., i_1) \quad (5)$$

Akaike's Information Criteria (AIC) and Bayesian Information Criteria (BIC) are two decision criteria which commonly used by the researchers in describing the optimum order of the Markov chain models. The PM_{10} concentration is determined when the minimum loss function is obtained. For instance, Berchtold and Raftery[21], Singh and Kripalani[22], Deni *et al.*[11] and Dastidar *et al.*[23] applied these two loss functions in their studies. Both criteria are based on the likelihood functions for the transition probabilities of the fitted Markov chain models. The maximum likelihood function for the k^{th} order chain can be written as:

$$L_k(X_1,..., X_n) = \prod_{s_1,...,s_k} \hat{P}_{s_1,...,s_k}^{n_{s_1,...,s_k}} \quad (6)$$

where, $\hat{P}_{s_1,...,s_k}^{n_{s_1,...,s_k}}$ is the estimated transition probabilities of each of the sequence going from state s_1 to s_2, from state s_2 to s_3 and from state s_{k-1} to s_k (s_k is the state of the most recent observation). The $n_{s_1,...,s_k}$ denotes the associated transition counts. The maximum likelihood estimator used in Eq. 7 of the transition probabilities is given by:

$$\hat{P}_{s_1,...,s_k} = \frac{n_{s_1,...,s_k}}{\sum_{s_1} n_{s_1,...,s_k}} \quad (7)$$

The maximum likelihood computed is used to decide the optimum order of two different Markov chain models, say the Markov chain models of the k^{th} and m^{th} orders where, k<m and k = 0, 1,..., m-1. Thus the maximized likelihood ratio statistics is given by:

$$\eta_{k,m} = 2\ln\lambda_{k,m} \quad (8)$$

Where:

$$\lambda_{k,m} = \frac{L_k(X_1,..., X_n)}{L_m(X_1,..., X_n)}$$

Table 2: Loss function, AIC and BIC values of Shah Alam monitoring station

Order		Threshold value (50 μg m^{-3})			
k	m	$\eta_{k,m}$ ($_kH_m$)	AIC	BIC	Degree of freedom
0	1	1327.781	1325.781	1319.396	1
0	2	1519.525	1513.525	1494.368	3
0	3	1621.670	1607.670	1562.972	7
0	4	1666.950	1636.950	1541.168	15
1	2	191.743	187.743	174.972	2
1	3	293.889	281.889	243.576	6
1	4	339.169	311.169	221.772	14
2	3	102.146	94.146	68.604	4
2	4	147.425	123.425	46.799	12
3	4	45.280	29.280*	-21.804*	8
4	4	0.000	0.000	0.000	0

AIC: Akaike's information criteria, BIC: Bayesian information criteria, *Minimum loss function

For example, the maximized likelihood ratio statistics for $_0H_1$ is calculated where k = 0 and m = 1. The parameter estimation (P_{00}, P_{11}, P_{10} and P_{01}) is then substituted into the equation and the value of $_0H_1$ is given by:

$$
\begin{aligned}
{}_0H_1 &= 2\left[\begin{matrix} n_{00}\left(\ln P_{00} - \ln P_0\right) + n_{01}\left(\ln P_{01} - \ln P_1\right) + n_{10}\left(\ln P_{10} - \ln P_0\right) + \\ n_{11}\left(\ln P_{11} - \ln P_1\right)\end{matrix}\right] \\
&= 2\left[\begin{matrix} 1478\left(\ln 0.747 - \ln 0.452\right) + 502\left(\ln 0.254 - \ln 0.548\right) + \\ 503\left(\ln 0.209 - \ln 0.452\right) + 1900\left(\ln 0.791 - \ln 0.548\right)\end{matrix}\right] \\
&= 1327.78
\end{aligned}
$$

As stated earlier, in determining the optimum order of Markov chain model, two loss functions are used, namely AIC and BIC. Tong[24] proposed that the loss function (AIC) is to define risk on the basis of the AIC criteria, while the BIC criteria, introduced by Schwarz[25] is to define risk on the basis of the BIC criteria. The only difference between these two criteria is the form of the penalty function. These criteria attempt to find the value of k that minimizes the loss function. The equation of AIC and BIC are as in Eq. 9 and 10, respectively:

$$\text{AIC}(k) = \eta_{k,m} - 2v \tag{9}$$

$$\text{BIC}(k) = \eta_{k,m} - v\ln(n) \tag{10}$$

where, $v = (s^m - s^k)(s-1)$ is the degree of freedom, s represents the number of states which in this case is 2 (polluted and non-polluted) and n is the number of the sample size. For example, for the value of AIC and BIC when k = 0 and m = 1 is given as follows:

$$
\begin{aligned}
\text{AIC}(0) &= \eta_{0,1} - 2\left[\left(2^1 - 2^0\right)\left(2-1\right)\right] \\
&= 1327.781 - 2(1) \\
&= 1325.781
\end{aligned}
$$

$$
\begin{aligned}
\text{BIC}(0) &= \eta_{0,1} - \ln(4383) \\
&= 1327.781 - 8.386 \\
&= 1319.396
\end{aligned}
$$

All the values obtained from the loss function of AIC and BIC in determining the optimum order of Markov chain model for a threshold value of 50 μg m^{-3} are presented in Table 2. The comparison between the minimum values of the loss function was done in order to choose the optimum order. For example, based on Table 2, the minimum values for both functions are at order three, which means that the optimum order for this threshold is at third order of Markov chain model. Many studies had also used these two loss functions in determining the optimum order of Markov chain model[11,21-23].

Fitting the higher order of Markov chain model: The information obtained from the transition probabilities was also used to calculate the frequency distribution of the order of Markov chain model, which was used to assess the performance of the higher order. The first order of Markov chain model was considered only for one preceding day and, similarly for the second order, the observed day depends on two preceding days and as does the other order. The joint probabilities of the k^{th} order of the Markov chain model is the following[11]:

$$P\left(i_n,...,i_2 \middle| i_1\right) = P_{i_{2i_1}}....P_{i_{k-1}...i_1}\prod_{j=1}^{n-k} P_{i_{k+j},...,i_j} \tag{11}$$

$$n \geq k+1;\ k = 0, 1.....$$

The conditional probabilities of events of n polluted days with the k^{th} order of the Markov chain can be written as Eq. 12[11]. From this equation, [n] represents n times. For example, the conditional probability of two consecutive polluted days can be written as P(011|0). The expected number of polluted days is computed by multiplying the conditional probabilities obtained from Eq. 12 with the total number of non-polluted days. For instance, the number of

non-polluted days of this station for a threshold value of 50 μg m^{-3} is 1981 days (Table 1). The chi-square test with a degree of freedom of d = v-1 was employed in this study to compare the observed and expected distributions of polluted events:

$$P(0,[n]|0) = \begin{cases} \prod_{i=1}^{n} P_{[1]0} P_{0[n]0}, & n \le k-1 \\ \prod_{i=1}^{k} P_{[i]0} P_{k+1}^{n-k} P_{0[k]}, & n \ge k \end{cases} \quad (12)$$

In assessing the best fitted higher order Markov chain model, the expected distribution which was closer to the observed distribution of polluted events was considered. The information from the transition probabilities was used to calculate the frequency. For example, the conditional probabilities of the first order of Markov chain model for polluted events in Shah Alam with a threshold value of 50 μg m^{-3} are given by:

$$P(0[1]|0) = P_{01}P_{11}^{0}P_{10} = (503/2403)(1900/2403)^{0}(502/1980) = 0.05307$$

$$P(0[2]|0) = P_{01}P_{11}^{1}P_{10} = (503/2403)(1900/2403)^{1}(502/1980) = 0.04196$$

$$P(0[3]|0) = P_{01}P_{11}^{2}P_{10} = (503/2403)(1900/2403)^{2}(502/1980) = 0.03318$$

$$P(0[25]|0) = P_{01}P_{11}^{24}P_{10} = (503/2403)(1900/2403)^{24}(502/1980) = 0.00019$$

The calculation of the conditional probabilities was continued until the maximum duration of polluted days for a sequence of polluted days was met. For example, the maximum number of polluted days for the Shah Alam station was 25 days for the 12 years worth of data used. Then, to get the expected frequencies of the first and higher orders, the conditional probabilities obtained from Eq. 12 were multiplied with the number of non-polluted days as mentioned earlier.

RESULTS AND DISCUSSION

Characteristics of PM_{10} concentration: The descriptive statistics of PM_{10} concentration data in Shah Alam monitoring station are shown in Table 3. Based on the Table 3, the maximum PM_{10} concentration value of 587 μg m^{-3} had occurred at Shah Alam monitoring station in August (11/8/2005) which may due to the haze event caused by trans-boundary pollution from Kalimantan and Sumatera in Indonesia[26]. The background concentration of PM_{10} is based on the median value, thus, the value of 52 μg m^{-3} was used for threshold value based on background concentration. The average daily of PM_{10} concentration from year 2002 until 2013 as illustrated in Fig. 1. Figure 1 shows that, higher values recorded were between 161st-231st days due to the occurrence of the dry season (Southwest monsoon) in Malaysia that occurred in the months of June until August[9].

Table 3: Descriptive statistics of PM_{10} concentration data in Shah Alam monitoring station

Descriptive statistics	PM_{10} concentration (μg m^{-3})
Mean	56.29
Median	52.00
Minimum value	14.00
Maximum value	587.00

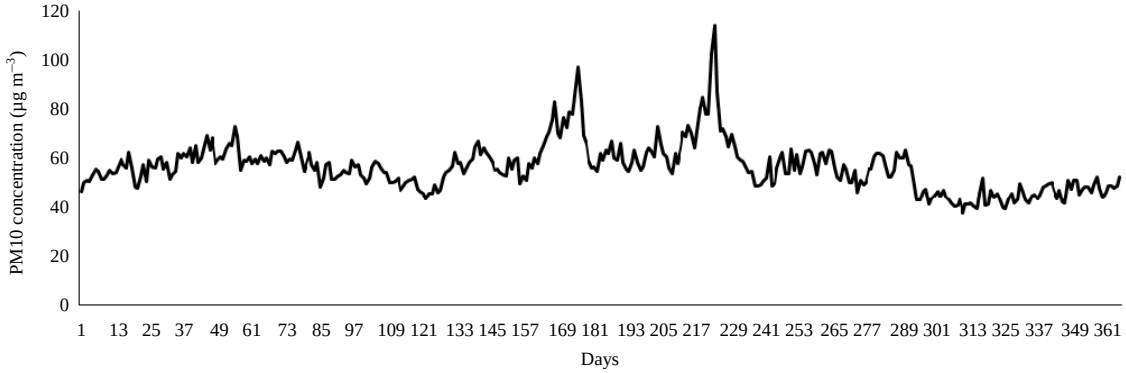

Fig. 1: Average daily of PM_{10} concentration from year 2002 until 2013

Table 4: Conditional probabilities of the sequence of polluted and non-polluted days of PM_{10} concentration and the values of α for all threshold values

	Threshold value (μg m^{-3})				
Conditional probabilities	50	52	100	120	150
P_1	0.548	0.476	0.045	0.014	0.006
P_{11}	0.791	0.767	0.523	0.508	0.538
P_{111}	0.846	0.822	0.596	0.516	0.500
P_{1111}	0.874	0.853	0.613	0.563	0.571
P_{11111}	0.890	0.870	0.605	0.444	0.750
P_0	0.452	0.524	0.955	0.986	0.994
P_{00}	0.746	0.789	0.977	0.993	0.997
P_{000}	0.783	0.825	0.980	0.995	0.998
P_{0000}	0.804	0.846	0.983	0.995	0.998
P_{00000}	0.821	0.856	0.984	0.996	0.998
α	663.890	716.663	218.520	100.050	58.570

The conditional probabilities of the sequence of polluted and non-polluted days of PM_{10} concentration and the values of α for all threshold values which obtained from Eq. 1 is shown in Table 4, respectively. These values were used to check whether the successive events (polluted or non-polluted) are independent of each other or not. The events could form Markov chain models or possess Markov chain properties if the events were dependent on each other. Table 4 shows that, the values of α for all threshold values used show that the successive events are dependent on each other and possess the Markov chain property where the value of α is larger than χ^2 with a value of 3.84 at a 5% level with 1° of freedom. Therefore, future analysis could be done since the events of PM_{10} concentration occurrence are dependent on each other. The values of the conditional probabilities that were used to analyse the persistency of the events in the area of study are shown in Table 4. Based on the Table 4, the conditional probabilities for both events show an increasing values where these values indicate the strong relationship between the observed and the previous event, which also means that the probabilities of getting polluted or non-polluted day influencing by previous events is higher regardless of threshold value used. For example, the probability of getting a polluted day based on the previous day that was also a polluted day (P_{11}), is greater than the unconditional probability (P_1). Indirectly, it also means that the higher persistency of polluted days indicates the occurrence of two or more consecutive polluted days for a given threshold value. Besides, the probability of two consecutive polluted days (P_{11}) is found higher than the probability of polluted day (P_1), which may be due to the behavior of PM_{10} concentration dependency since the effect of PM_{10} according to WHO is more than 24 h.

Optimum order of Markov chain model in describing the sequence of PM_{10} concentration at Shah Alam monitoring station: The AIC and BIC values for Shah Alam monitoring Station with different threshold values of PM_{10} concentration is shown in Table 5. Table 5 shows the threshold value less than 120 μg m^{-3}, the higher order or an order of more than one is optimum for both AIC and BIC, which means that the occurrence of polluted (non-polluted) events for this station is dependent on the events of two or three days before the observed day. However, according to Katz[27], the AIC has the tendency to overestimate the optimum order. It was also found that the BIC estimate for rainfall data of the Tel Aviv data is unity. Besides that, Dastidar *et al.*[23] stated the use of the BIC also gives a mathematical formulation with a principle of parsimony in model building. Thus, this study considers the optimum order obtained from the minimum loss function of the BIC. The table also shows an order of three is optimal for threshold values of less than 100 μg m^{-3}, which best describes the sequence. While for a threshold value of 120 μg m^{-3}, the optimum order is two. As for a threshold value of 150 μg m^{-3}, a simple order is the optimum order that best describes the sequence of polluted (non-polluted) days of PM_{10} concentration at this station. Thus, it can be concluded that the higher order is more appropriate in describing the sequence of polluted (non-polluted) days of PM_{10} concentration at this station for threshold value less than 120 μg m^{-3}.

Besides that, the results obtained also show that there are less dependency on previous events when the threshold value is increasing, which indicates that it is not accurately predict occurrences of PM_{10} concentration when the threshold value used is more than 120 μg m^{-3} for this station. Indirectly, this study also suggests the reason why the limit or threshold value of PM_{10} should be revised to suit with Malaysia

Table 5: AIC and BIC values for Shah Alam monitoring station with different threshold values of PM_{10} concentration

Threshold value (μg m^{-3})	Order	Akaike's Information Criteria (AIC)	Bayesian Information Criteria (BIC)
50	0	1636.95	1541.17
	1	311.17	221.77
	2	123.43	46.80
	3	29.28	-21.80*
	4	0.00	0.00
52	0	1711.84	1616.06
	1	280.52	191.12
	2	115.84	39.21
	3	29.77	-21.32
	4	0.00	0.00
100	0	492.25	396.47
	1	57.20	-32.19
	2	24.59	-52.04
	3	-4.00	-55.08*
	4	0.00	0.00
120	0	242.44	197.74
	1	44.35	6.03
	2	1.04	-24.51*
	3	0.00	0.00
	4	-	-
150	0	124.49	105.33
	1	9.35	-3.43*
	2	0.00	0.00
	3	-	-
	4	-	-

*Minimum loss function

now-a-days so that better prediction based on previous events can be made for early precaution to the public and the environment, as suggested by DoE[28].

Appropriate order for polluted events: Since the third order is found to be the optimum order in the sequence of PM_{10} concentrations, therefore, the investigation on the fitting will be carry out further by considering the distribution of polluted events at this monitoring station. The observed and expected frequency distribution is computed as shown in Table 6. The chi-square goodness of fit test is considered as to select the most successful and the best fitted model for each threshold used. All the expected frequencies are more than 5 days and met the assumption of chi-square test. Table 6 shows that, at $\alpha = 0.05$ level of significant, there is enough evidence to conclude that the observed and expected days of polluted events at threshold value of 50 and 52 μg m^{-3} for first and second order of Markov chain model is seems not satisfy in representing the distributions of polluted event at this station since the null hypothesis is rejected. However, the threshold value of 50 and 52 μg m^{-3} for third order produce better fit since the chi-square statistics value is lower than other order and the observed and expected days of polluted events also well describe the distribution. This study also can conclude that higher order (order three) produce better fit than other order since the value of chi-square produced is lower than other order at all threshold value used.

Figure 2 provides the observed and expected frequencies for the threshold value used in this study based on the appropriate order of the first three orders of the Markov chain models, known as order of one (MCM1), order of two (MCM2) and order of three (MCM3) for the distribution of the polluted events at this station. However, the results obtained show that the order more than one (MCM2 and MCM3) are the most appropriate order that best describe the distribution of the polluted events at Shah Alam monitoring station. The observed and expected frequencies of polluted events based on the best fitted order of Markov chain model at threshold values of 50, 52, 100 and 120 μg m^{-3} is shown in Fig. 2, respectively. Obviously, the expected frequencies of polluted events obtained from the order really describe the observed distributions, since among χ^2, χ^2 for order three (MCM3) is the best fitted order for threshold values of 50, 52 and 100 μg m^{-3}, while for a threshold value of 120 μg m^{-3}, the distribution of polluted events does fit at the order of two (MCM2). Besides that, Fig. 2 also shows the number of polluted events decreasing when the threshold value of PM_{10} increases. In conclusion, threshold value used is also important in determining the optimum order and also describing the distribution of polluted events at this monitoring station.

Table 6: Chi-square test, degree of freedom and p-value of observed and expected length of polluted days for Shah Alam monitoring station

Order	Threshold value (μg m^{-3})		Length of polluted event																			Chi-square	Degree of freedom	p-value
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19			
1	50	Observed	209	90	49	24	24	16	10	13	12	8	2	5	1	6	6	13				173.30	15	2.20E-16*
		Expected	105	83	66	52	41	32	26	20	16	13	10	8	6	9	6	8						
	52	Observed	195	110	59	32	29	14	15	7	6	2	2	2	5	4	3					112.64	14	2.20E-16*
		Expected	113	87	66	51	39	30	23	18	14	10	8	6	5	6	5							
	100	Observed	53	18	9	6	9															4.19	4	0.38
		Expected	45	24	12	6	7																	
	120	Observed	15	8	7																	0.11	2	0.945
		Expected	15	7	7																			
	150	Observed	6	7																		0.08	1	0.782
		Expected	6	6																				
2	50	Observed	209	90	49	24	24	16	10	13	12	8	2	5	1	2	4	6	7	4	6	79.58	18	1.02E-09*
		Expected	210	45	38	32	27	23	20	17	14	12	10	8	7	6	5	8	6	6	9			
	52	Observed	195	110	59	32	29	14	15	7	6	2	2	2	5	3	3	1				112.94	15	2.20E-16*
		Expected	201	50	42	34	28	23	19	16	13	11	9	7	6	5	7	5						
	100	Observed	53	18	9	6	9															0.27	4	0.992
		Expected	53	17	10	6	8																	
	120	Observed	15	8	7																	0.10	2	0.949
		Expected	15	7	7																			
	150	Observed	6	7																		0.11	1	0.743
		Expected	5	7																				
3	50	Observed	209	90	49	24	24	16	10	13	12	8	2	5	1	2	4	6	7	5	5	25.38**	18	0.115
		Expected	210	90	36	22	20	17	15	13	11	10	9	8	7	6	5	8	6	6	9			
	52	Observed	195	110	59	32	29	14	15	7	6	2	2	2	5	3	4					23.29**	14	0.056
		Expected	201	99	50	34	21	16	15	10	9	9	8	7	6	5	5							
	100	Observed	53	18	9	6	9															0.11**	4	0.999
		Expected	53	18	9	6	8																	
	120	Observed	15	8	7																	0.00**	2	1.000
		Expected	15	8	7																			

*p-value <α = 0.05, **Smallest value of χ^2 for each threshold value

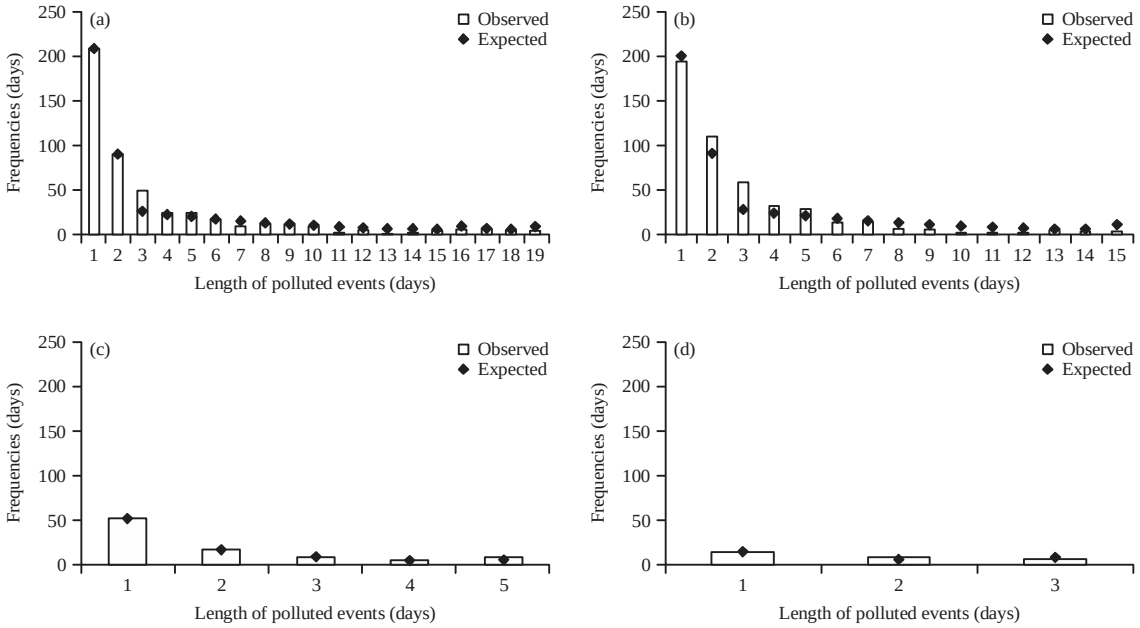

Fig. 2(a-d): Observed and expected frequencies of polluted events based on the best fitted order of Markov chain model at threshold values of (a) 50 μg m^{-3}, (b) 52 μg m^{-3}, (c) 100 μg m^{-3} and (d) 120 μg m^{-3}

CONCLUSION

This study has successfully discussed a complete description of the occurrence of PM_{10} concentration at the Shah Alam monitoring station by using Markov chain model. For threshold value of less than 120 μg m^{-3}, the optimum order for this station is order of 2 and 3. This results indicate that the occurrence of polluted (non-polluted) days depends on the two or three days before the observed day and early prediction can be made by responsible authorities as they can predict the event of two or three days prior. As for a threshold value of 150 μg m^{-3}, an order of one is the optimum order, which indicates that prediction can be made by referring to the day before the observed day. As a conclusion, the higher order Markov chain model is appropriate in making the prediction of PM_{10} concentrations based on the minimum loss function at this monitoring station.

SIGNIFICANCE STATEMENTS

This study discover the future prediction of PM_{10} concentration events by considering the previous day events which beneficial in monitoring the effect of PM_{10} concentrations at the area of study. This study will help the researcher or authorities to provide the necessary information and early prediction of PM_{10} occurrences at particular area. Therefore, the effect of PM_{10} concentrations may be reduce by taking early precaution especially to high risk people such as children and elderly people.

ACKNOWLEDGMENT

The authors are greatly thankful to Universiti Teknologi Mara (UiTM), under Grant 600-RMI/IRAGS 5/3 (36/2015). Special recognition goes to the Department of Environment (DoE) and Alam Sekitar Malaysia Sdn. Bhd (ASMA) for providing the air quality data for this study.

REFERENCES

1. Awang, M.B., A.B. Jaafar, A.M. Abdullah, M.B. Ismail and M.N. Hassan *et al*., 2000. Air quality in Malaysia: Impacts, management issues and future challenges. Respirology, 5: 183-196.
2. Field, R.D., G.R. van der Werf and S.S. Shen, 2009. Human amplification of drought-induced biomass burning in Indonesia since 1960. Nat. Geosci., 2: 185-188.
3. Fellenberg, G., 2000. The Chemistry of Pollution. John Wiley and Sons Ltd., England.
4. DoS., 2013. Compendium of environment statistics. Department of Statistics, Malaysia, pp: 1-278.
5. Ul-Saufie, A.Z., A.S. Yahaya, N.A. Ramli and H. Abdul Hamid, 2012. Robust regression models for predicting PM_{10} concentration in an industrial area. Int. J. Eng. Technol., 2: 364-370.

6. Chaloulakou, A., G. Grivas and N. Spyrellis, 2003. Neural network and multiple regression models for PM_{10} prediction in Athens: A comparative assessment. J. Air Waste Manage. Assoc., 53: 1183-1190.
7. Ul-Saufie, A.Z., A.S. Yahaya, N.A. Ramli and H.A. Hamid, 2012. Future PM10 concentration prediction using quantile regression models. Proceedings of the 2nd International Conference on Environmental and Agriculture Engineering, Volume 37, May 27-June 1, 2012, Singapore, pp: 15-19.
8. Huebnerova, Z. and J. Michalek, 2014. Analysis of daily average PM_{10} predictions by generalized linear models in Brno, Czech Republic. Atmos. Pollut. Res., 5: 471-476.
9. Yusof, N.F.F.M., N.A.R. Ramli and A.S. Yahaya, 2011. Extreme value distribution for prediction of future PM_{10} exceedences. Int. J. Environ. Prot., 1: 28-36.
10. Chin, E.H., 1977. Modeling daily precipitation occurrence process with Markov Chain. Water Resourc. Res., 13: 949-956.
11. Deni, S.M., A.A. Jemain and K. Ibrahim, 2009. Fitting optimum order of Markov chain models for daily rainfall occurrences in Peninsular Malaysia. Theor. Applied Climatol., 97: 109-121.
12. Gabriel, K.R. and J. Neumann, 1962. A markov chain model for daily rainfall occurrence at Tel Aviv. Q. J. Royal Meteorol. Soc., 88: 90-95.
13. Lin, G.Y., 1981. Simple Markov chain model of smog probability in the South coast air basin of California. Prof. Geogr., 33: 228-236.
14. Rahimi, J., J. Bazfarshan and A. Rahimi, 2014. Study of persistence of days infected pollutant particulate matter (PM_{10}) in city of Tehran using Markov chain model. J. Environ. Sci. Technol., 15: 79-90.
15. Lin, G.Y. and L.C. Huang, 1985. Statistical models for predicting air pollution in taipei. Proc. Nat. Sci. Counc. ROC(A), 9: 47-59.
16. Mangaraj, A.K., L.N. Sahoo and M.K. Sukla, 2013. A markov chain analysis of daily rainfall occurrence at Western Orissa of India. J. Reliab. Stat. Stud., 6: 77-86.
17. WHO., 2000. Particulate Matter. In: Air Quality Guidelines, WHO (Ed.). 2nd Edn., WHO Regional Office for Europe, Denmark, pp: 1-40.
18. Shuhaili, A., A. Fadzil, S.I. Ihsan and W.F. Faris, 2013. Air pollution study of vehicles emission in high volume traffic: Selangor, Malaysia as a case study. WSEAS Trans. Syst., 12: 67-84.
19. DoE., 2013. New Malaysia ambient air quality standard. http://www.doe.gov.my/portalv1/wp-content/uploads/2013/01/Air-Quality-Standard-BI.pdf
20. Moon, S.E., S.B. Ryoo and J.G. Kwon, 1994. A Markov chain model for daily precipitation occurrence in South Korea. Int. J. Climatol., 14: 1009-1016.
21. Berchtold, A. and A.E. Raftery, 2002. The mixture transition distribution model for high-order Markov chains and non-Gaussian time series. Stat. Sci., 17: 328-356.
22. Singh, S.V. and R.H. Kripalani, 1986. Analysis of persistence in daily monsoon rainfall over India. J. Climatol., 6: 625-639.
23. Dastidar, A.G., D. Ghosh, S. Dasgupta and U.K. De, 2010. Higher order Markov chain models for monsoon rainfall over West Bengal, India. Indian J. Radio Space Phys., 39: 39-44.
24. Tong, H., 1975. Determination of the order of a Markov chain by Akaike's information criterion. J. Applied Probab., 12: 488-497.
25. Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat., 6: 461-464.
26. Mahmud, M. and N.H. Abu Hanifiah, 2009. Air pollution during the haze event of 2005: The case of perai, pulau pinang, Malaysia. Malays. J. Soc. Space, 2: 1-15.
27. Katz, R.W., 1981. On some criteria for estimating the order of a markov chain. Technometrics, 23: 243-249.
28. DoE., 2013. Annual report 2013. Department of Environment (DoE), Ministry of Natural Resources and Environment, Kuala Lumpur, Malaysia.