Forecasting Air Pollution Index in Klang by Markov Chain Model

Nurul Nnadiah Zakaria, Rajalingam Sokkalingam, Hanita Daud, Mahmod Othman

Abstract: ***The main purpose of analyze future air quality is to maintain the environment in good and healthy condition. Current techniques applied to forecast the air pollution index were ARIMA, SARIMA, Artificial Neural Network, Fuzzy Time Series, Machine Learning, etc. Thus, each technique has its own advantages and disadvantages in the variables, model selection and model accuracy determination. This study aims to forecast air pollution index by developing a Markov Chain model in Klang district, Selangor state which is one of the most polluted area in Malaysia. The Markov Chain model development is a stochastic process sequence that depends on the previous successive event in time. In this model development, state transition matrix and probability are the main concept in determine the future behavior of Air Pollution Index which depends on the present state of the process. The result shows that the developed model is a simple and good performance model that will forecast and evaluate the distribution of the pollution level in long term.***

Index Term: Markov Chain, Air Pollution Index (API), Stationary Distribution, Mean Return Time, Long Term Forecasting.

I. INTRODUCTION

Air pollution is one of the major issues that has been affecting human health, agricultural crops, species of forest and ecosystems in Malaysia. It may have serious effects on the future industrial development and human activities. Since 1980, Malaysia has had a series of haze events; the worst ever was reported in 1997. Many areas, particularly the urban and industrial, were severely affected by heavy pollutants, causing major air pollution throughout the region (DOE, 2019c). In 2015, Malaysia was facing with the haze again over a month, due the Southeast Asian Haze Crisis. This crisis was caused by Indonesian palm oil producers burning off forest growth in remote areas, to clear land for planting. It was considered as the worst haze crisis after 1997 due to the prolonged haze duration, which lasted for more than two months and negatively impacted Malaysia and Singapore (2019c).

In Malaysia, the air quality data is controlled and monitored by Department of Environment (DOE), Ministry of Natural Resources and Energy. DOE will monitor the API reading continuously if there are any changes in the ambient air quality status through the Continuous Air Quality Monitoring Station (CAQM). These CAQM stations are located at strategic places such as industrial, urban, sub-urban and rural areas. The parameters that will be measured are Sulphur dioxide (SO_2), nitrogen dioxide (NO_2), carbon monoxide (CO), ozone (O_3), and particulate matter with the size of less than 10 micron (PM_{10})(2019a) . Starting in 2015, DOE established the New Ambient Air Quality Standard in order to replace the Recommended Malaysian Air Quality Guidelines that has been used since 1989 (2015). From mid-year 2018 onwards, DOE has improved the calculation for API by adding one more parameter namely, particulate matter with the size of less than 2.5 micron ($PM_{2.5}$)(2019b). The data were saved as hourly averages in different measurement unit and the index will be based on which pollutant recorded the highest concentration at that time.

On the other hand, there are many methods that have been used to analyze and forecast the future API either hybrid or non-hybrid techniques. Both methods were complex and difficult to derive and analyze for API data. Therefore, this study attempts to apply Markov chain for forecasting air pollution index. Markov chain is a simplest way of introducing statistical dependence into a model for a stochastic process. It is a random process where all information about the future is contained in the present state.

II. LITERATURE REVIEW

There are several non-hybrid and hybrid techniques are used in monitoring and forecasting air pollutant. Non-hybrid techniques are Seasonal Autoregressive Integrated Moving Average (SARIMA), Fuzzy Time Series (FTS), Artificial Neural Networks (ANN) etc. Rahman and Lee (2016) claim that the classical method SARIMA is more accurate compared to FTS, but both methods have ability to investigate and forecast the API trends. Rahman *et al.* (Rahman et al., 2017) noticed that ANN was used to predict the state of atmospheric air in an industrial city for capability of the operative environmental decision. Hybrid techniques are Empirical Mode Decomposition-SVR-Hybrid (EMD-SVR-Hybrid), Empirical Mode Decomposition – Intrinsic Mode Functions-Hybrid (EMD-IMFs-Hybrid), ANN-Support Vector Machine (SVM) etc. Zhu *et al.* (Zhu et al., 2017) study's found that the EMD-SVR-Hybrid and EMD-IMFs-Hybrid developed model were used

Revised Manuscript Received on September 22, 2019.

Nurul Nnadiah Zakaria, Department of Fundamental and Applied Sciences, Universiti Teknologi PETRONAS, Seri Iskandar 32610, Perak, Malaysia. nurulnnadiah94@gmail.com

Rajalingam Sokkalingam, Department of Fundamental and Applied Sciences, Universiti Teknologi PETRONAS, Seri Iskandar 32610, Perak, Malaysia

Hanita Daud, Department of Fundamental and Applied Sciences, Universiti Teknologi PETRONAS, Seri Iskandar 32610, Perak, Malaysia

Mahmod Othman, Department of Fundamental and Applied Sciences, Universiti Teknologi PETRONAS, Seri Iskandar 32610, Perak, Malaysia

in forecasting the regional air quality indexes and both models were effective to apply in air quality data. Meanwhile, Wang *et al.* (Wang et al., 2015) showed that ANN-SVM is the model that leads to the improvement of model accuracy in forecasting the pollutant concentrations.

Markov chain, named after mathematician of Russia, Andrey Markov (Shannon, 1948), is a system in mathematical that suffer transitions of one state to another state in a chain (Anderson and Goodman, 1957). It is a random process that supplied with Markov property. The shifts of state in the process are called as transitions while the probabilities related to the changes of state could be defined as transition probabilities (Tettey et al., 2017). For transition probabilities and all states absolutely represent a Markov chain. There are several advantages of the Markov chain model in forecasting air pollution index. Firstly, because the pollution level has finite or limited number of states which is DOE classified the air pollution level into five levels (Good, Moderate, Unhealthy, Very Unhealthy, Hazardous) (2019b). Besides, the main components in developing a Markov chain model are state transition matrix and probability where they will summarize all the essential parameters of dynamic change. In addition, this technique does not require deep insight into the mechanisms of dynamic change and relatively easy to derive from the API data.

In Malaysia, Markov chain technique was commonly used in manpower planning. Saad *et al.* was introduced this technique to investigate the track movements of lecturers in universities (Saad et al., 2014). Zhou *et al.* noticed that Markov Chain is the most significant techniques and constraints for the good prediction of the probability of bikes rental and returns in a bike sharing system at each station in China (Zhou et al., 2018). In addition, Markov chain techniques have also been used to describe the probabilistic behaviors of wind-direction data (Masseran, 2015). The developed model was used to analyze the natural geographic direction behavior based on maximum likelihood method and the linear programming formulation. Thus, this study develops Markov chain model to forecast the air pollution index in Malaysia.

III. METHODOLOGY

This section was divided into three parts namely study area, data collection and Markov chain model development.

A. Study Area

The study area is Klang district which include the Klang City, Port Klang and the part of Shah Alam that is located in Selangor state. The study area that covers a total of 626.78 square/km of land with 53.75 km of coastline (JUPEM, 2016).

B. Data Collection

Secondary datasets were used for 5 years from 2013 until 2017 from Department of Environment (DOE), Malaysia. DOE categorized the air quality status into five level (2019b). Therefore, the datasets were classified into five state in this Markov Chain model, as in Table I below.

Table I. Air Pollution Index

State	API	Air Quality Status
1	0-50	Good
2	51-100	Moderate
3	101-200	Unhealthy
4	201-300	Very Unhealthy
5	300 and above	Hazardous

C. Markov Chain Model Development

Markov chain model was developed to forecast the Air Pollution Index future behavior. Based on the Figure 1, the model development was divided into three phases; state transition matrix and probability, confirmation of ergodic Markov chain, and stationary probability distribution.

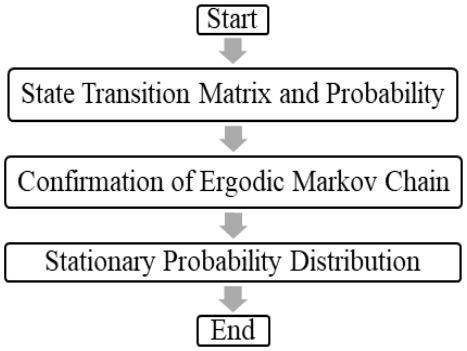

Fig 1. Model Development Framework.

The main concept in developing the model are state transition matrix and probability. Let S be a state space of the events of X, where a random process of $X = \{Xt; t = 0,1,2, \dots, T\}$ and $S = \{1,2, \dots, q\}$. To fulfil the Markov property, Markov process can be defined as;

$$p_{ij} = p(X_{t+1} = j | X_t = i) \quad (1)$$

Let p_{ij} be the transition probability from state i to move to state j where i and j must be in the state space of X event as shown in Equation 1. The Markov process will occur when one of the state shifts to another state. Let P be a transition matrix that describes all the transition probabilities for every state in S. Then P is denoted by:

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1s} \\ p_{21} & p_{22} & \cdots & p_{2s} \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ p_{s1} & p_{s2} & \cdots & p_{ss} \end{bmatrix} \quad (2)$$

Where,

p_{ij} : Probability of an hour being in state j, given that the previous hour was in state i.

The observed frequency for state transition matrix can be written as following;

$$N = \begin{bmatrix} n_{11} & n_{12} & \cdots & n_{1s} \\ n_{21} & n_{22} & \cdots & n_{2s} \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ n_{s1} & n_{s2} & \cdots & n_{ss} \end{bmatrix} \quad (3)$$

Where,

n_{ij} : Observed frequency of a sequence for state i followed by state j.

And is respectively the same for other transition probability of p_{ij}. The transition probability must be between 0 and 1 and the row of P, which is i;

$$\sum_{j=1}^{s} p_{ij} = 1 \quad (4)$$

for each of the column, *j*. This transition describes the probabilities for the next hour of pollution level given by current hour's level. The confirmation of ergodic Markov chain must be made to identify the existence of limiting distribution, by classifying the state of *P*. It can be divided into three sections; Irreducible Markov Chain, Periodicity Markov chain and Recurrent and Transient states as in (Grinstead and Snell, 2006, Pinsky and Karlin, 2011).

Stationary probability distribution will describe the behavior of the air pollution in long term forecasting. The distribution maintains for all future time with steady-state probabilities that are independent from initial conditions (Ibe, 2013). For ergodic Markov chain, the limiting distribution was existing for stationary probability distribution and can be represent as;

$$\pi_j = \lim_{n\to\infty} P(X_n = j | X_0 = i) \quad (5)$$

π_j will provide the time proportion which the process of stochastic stays in certain state. The probability occurrences of state *j* are higher when the value of the π_j is high (Grinstead and Snell, 2006). After determining the limiting distribution, we then identify the average time for specific states to return back to itself, m_j. It can be denoted as;

$$m_j = \frac{1}{\pi_j} \quad (6)$$

IV. RESULTS AND FINDINGS

This section shows the results and findings of the Markov chain model development. The state space for API was decided by DOE as shown in Table 1.

A. State Transition Matrix and Probability

The state transition matrix and probability were obtained for 2013 until 2017 as in Table 2 and Table 3. Based on the result, the highest observed frequency for transition moderate state in current hour to moderate state for the next hour is 23773 and the transition for good to good level is 17633. Therefore, these two highest observed frequencies shown in Table II, the highest probability whereas 0.9756 and 0.9692 in state transition probability as in Table 3 and respectively same for other transition probability of *P* based on Equation 2 and Equation 3. From the result in Table III, the chain of the developed model can be constructed to show the transition between five states of API as shown in Figure 2 by using Mathematica software.

Table II. State Transition Matrix

State of API	Good	Moderate	Unhealthy	Very Unhealthy	Hazardous
Good	17633	550	10	0	0
Moderate	551	23773	44	0	0
Unhealthy	10	44	1081	6	0
Very Unhealthy	0	0	6	49	3
Hazardous	0	0	0	3	60

Table III. State Transition Probability

State of API	Good	Moderate	Unhealthy	Very Unhealthy	Hazardous
Good	0.9692	0.0302	0.0005	0	0
Moderate	0.0226	0.9756	0.0018	0	0
Unhealthy	0.0088	0.0386	0.9474	0.0053	0
Very Unhealthy	0	0	0.1034	0.8448	0.0517
Hazardous	0	0	0	0.0476	0.9524

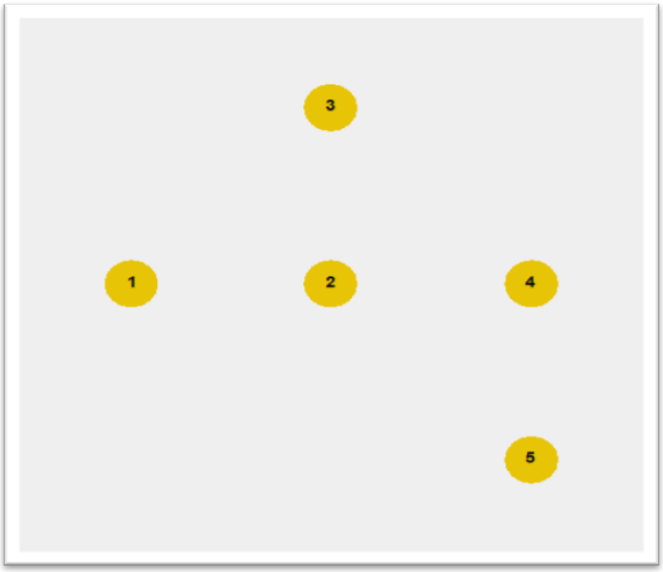

Fig 2. Chain of the Develop Model.

Based on the results in Table III, these probabilities were used to find the predicted probability for the next 5 years (2018-2020) based on Equation 7 until Equation 12 as below;

$$\pi_{Good} = 0.9692\pi_{Good} + 0.0302\pi_{Moderate} + 0.0005\pi_{Un} \quad (7)$$

$$\pi_{Moderate} = 0.0226\pi_{Good} + 0.9756\pi_{Moderate} + 0.0018 \quad (8)$$

$$\pi_{Unhealthy} = 0.0088\pi_{Good} + 0.0386\pi_{Moderate} + 0.947 \quad (9)$$

$$\pi_{Very\ Unhealthy} = 0.1034\pi_{Unhealthy} + 0.8448\pi_{Very\ Unh} \quad (10)$$

$$\pi_{Hazardous} = 0.0476\pi_{Very\ Unhealthy} + 0.9524\pi_{Hazardou} \quad (11)$$

$$\pi_{Good} + \pi_{Moderate} + \pi_{Unhealthy} + \pi_{Very\ Unhealthy} + \pi_{H} \quad (12)$$

B. Confirmation of Ergodic Markov Chain

Based on the transition probability matrix in Table 3 and the chain of the develop model in Figure 2, the states can be can classify in order to confirm the ergodic Markov chain. All the states are accessible because state *j* can access from state *i* while state *j* also accessible from state *j*, for *n* where $P_{ij}^{n} > 0$ and *m* = 0,1, 2, *M*. If both state *i* and *j* is accessible, it can imply that state *i* and *j* are communicating with each other in this chain. Thus, all the states in this chain are communicable among the states and can be represented as $j \Leftrightarrow i \ \forall\ i,j$. In other words, it can be confirmed that all states are symmetric, transitive and reflexive relation. Furthermore, this Markov chain model is irreducible because all states in this chain are communicate. Next, there is no transient that exists during the process and can reveal that the state is recurrent, means that state *i* will definitely return to itself during the process.

According to transition matrix, it was shown that the state is aperiodic because the period is one for all values of *n* in the chain. Apart from that, this model also can be identified as regular Markov chain because of the step from one state to other state is uniform in n-steps. Since all states are aperiodic and recurrent, then all states can be considered is ergodic. For the overall state in the developed Markov chain model for API can be concluded as ergodic Markov chain.

C. Stationary Probability Distribution

Next, stationary probability distribution is to evaluate the long run proportion of the air pollution behavior. It will exist when the Markov chain is ergodic. Then, for certain state, it will not depend on the initial state for long-time proportion of the chain.

Table IV. Stationary Probability Distribution

CAQM Station	π_{Good}	π_{Modera}	$\pi_{Unhealt}$	$\pi_{Very\ Unhe}$	π_{Hazard}
Klang	0.4156	0.5556	0.0260	0.0013	0.0014

Based on the result above, π_j is the steady-state probabilities for stationary probability distribution in this study by using Equation 5. The highest probabilities in Table IV is in the good state which is 0.5556. Good state can be considered as the most prominent compared to other state because of the high proportion in a long-run. From this point, it can be explained that 55.56% of variability for the occurrences based on the observed data. Meanwhile, the lowest proportion, zero probability are in very unhealthy state which is 0.0013 compared to hazardous state 0.0014. It means that Klang district is not in a high risk for haze to occur.

D. Mean Return Time

Table V. Mean Return Time

CAQM Station	m_{Good}	m_{Moder}	m_{Unheal}	$m_{Very\ Unhe}$	m_{Hazard}
Klang	2.4063	1.7997	38.4237	755.8874	695.8963

Table V shown that the mean return time for each API state in study area by using Equation 6. The highest average time for specific state to return again is 32 days (755.8874 hours) for unhealthy state and 29 days (695.8963 hours) for hazardous state to return to itself. Besides, for good and moderate state only took a few hours which 2.4063 hours and 1.7997 hours respectively. Unhealthy state took around one and a half day (38.4237 hours) to return back to itself means that the probability for air pollution in that state is still in the low average time.

V. CONCLUSIONS

The model from the Equation 7 until 12 was applied to air pollution index from 2013-2017 for Klang district. Based on the result, it can be concluded that the develop Markov chain model provides a good approximation for describing the occurrence of the sequence of API; good (0.9692), moderate (0.9756), unhealthy (0.9474), very unhealthy (0.8448) and hazardous (0.9524) respectively. Based on the stationary probability distribution Equation 5, the probability of the pollution level for the next 5 years can be predicted, where the highest towards lowest probabilities in the good, moderate, unhealthy, very unhealthy and hazardous state are 0.4156, 0.5556, 0.0260, 0.0013 and 0.0014 respectively. It is concluded in next five years that the Klang district is not in a high risk for haze to occur. Developing this model is very useful to apply in another study area for forecasting the air pollution index.

ACKNOWLEDGMENTS

This research work is supported by the Graduate Assistantship (GA Scheme) under Universiti Teknologi PETRONAS.

REFERENCES

[1] ANDERSON, T. W. & GOODMAN, L. A. 1957. Statistical Inference about Markov Chains. The Annals of Mathematical Statistics, 28, 89-110.
[2] DOE. 2015. Air Pollution Index of Malaysia [Online]. Department of Environment. Available: http://apims.doe.gov.my [Accessed 01/03/2019 2019].
[3] DOE. 2019a. Air Quality [Online]. Department of Environment. Available: https://www.doe.gov.my/portalv1/en/info-umum/kuality-udara/114 [Accessed 10/4/2019 2019].
[4] DOE. 2019b. Air Quality Standards [Online]. Department of Environment. Available: https://www.doe.gov.my/portalv1/en/info-umum/english-air-quality-trend/108 [Accessed 31/03/2019 2019].
[5] DOE. 2019c. Chronology of Haze Episodes in Malaysia [Online]. Department of Environment. Available: https://www.doe.gov.my/portalv1/en/info-umum/info-kualiti-udara/kronologi-episod-jerebu-di-malaysia/319123 [Accessed 23/03/2019 2019].
[6] GRINSTEAD, C. & SNELL, L. 2006. Grinstead and Snell's Introduction to Probability, American Mathematical Society.
[7] IBE, O. 2013. Markov Processes for Stochastic Modeling, Elsevier Science.
[8] JUPEM. 2016. Atlas kebangsaan Malaysia. Kuala Lumpur: Jabatan Ukur dan Pemetaan Malaysia.
[9] MASSERAN, N. 2015. Markov Chain model for the stochastic behaviors of wind-direction data. Energy Conversion and Management, 92, 266-274.
[10] PINSKY, M. A. & KARLIN, S. 2011. 4 - The Long Run Behavior of Markov Chains. In: PINSKY, M. A. & KARLIN, S. (eds.) An Introduction to Stochastic Modeling (Fourth Edition). Boston: Academic Press.
[11] RAHMAN, N. H. A. & LEE, M. H. 2016. Evaluation Performance of Time Series Approach for Forecasting Air Pollution Index in Johor, Malaysia. Sains Malaysiana, 45, 1625-1633.
[12] RAHMAN, P. A., PANCHENKO, A. A. & SAFAROV, A. M. 2017. Using neural networks for prediction of air pollution index in industrial city. IOP Conference Series: Earth and Environmental Science, 87, 042016.
[13] SAAD, S. A., ADNAN, F. A., IBRAHIM, H. & RAHIM, R. 2014. Manpower planning using Markov Chain model. AIP Conference Proceedings, 1605, 1123-1127.
[14] SHANNON, C. E. 1948. A mathematical theory of communication. The Bell System Technical Journal, 27, 379-423.
[15] TETTEY, M., ODURO, F. T., ADEDIA, D. & ABAYE, D. A. 2017. Markov chain analysis of the rainfall patterns of five geographical locations in the south eastern coast of Ghana. Earth Perspectives, 4, 6.
[16] WANG, P., LIU, Y., QIN, Z. & ZHANG, G. 2015. A novel hybrid forecasting model for PM10 and SO2 daily concentrations. Science of The Total Environment, 505, 1202-1212.
[17] ZHOU, Y., WANG, L., ZHONG, R. & TAN, Y. 2018. A Markov Chain Based Demand Prediction Model for Stations in Bike Sharing Systems.
[18] ZHU, S., LIAN, X., LIU, H., HU, J., WANG, Y. & CHE, J. 2017. Daily air quality index forecasting with hybrid models: A case in China. Environmental Pollution, 231, 1232-1244.

AUTHORS PROFILE

My name is Nurul Nnadiah Zakaria, currently affiliated with Department of Fundamental and Applied Sciences, Universiti Teknologi PETRONAS, Seri Iskandar 32610, Perak, Malaysia. My area of interest is nurulnnadiah94@gmail.com

My name is Rajalingam Sokkalingam, currently working Department of Fundamental and Applied Sciences, Universiti Teknologi PETRONAS, Seri Iskandar 32610, Perak, Malaysia

I am Hanita Daud, currently working with Department of Fundamental and Applied Sciences, Universiti Teknologi PETRONAS, Seri Iskandar 32610, Perak, Malaysia

I am Mahmod Othman, affiliated with Department of Fundamental and Applied Sciences, Universiti Teknologi PETRONAS, Seri Iskandar 32610, Perak, Malaysia