Universiti Teknologi MARA
Cawangan Kedah
Kampus Sungai Petani

Voice of Academia

Academic Series of Universiti Teknologi MARA Kedah

VoA 2021
Volume 17 Issue 1

COMMITTEE PAGE

Samsiah Bidin
Universiti Teknologi MARA

Shafina Md Salleh
Universiti Teknologi MARA

Sharina Saad
Universiti Teknologi MARA

e-ISSN: 2682-7840

© Voice of Academia is jointly published by the Universiti Teknologi MARA Caawangan Kedah, Malaysia and Penerbit UiTM (UiTM Press), Universiti Teknologi MARA Malaysia, Shah Alam, Selangor.

TABLE of CONTENTS

Voice of Academia
e-ISSN: 2682-7840

JOURNAL VOICE OF ACADEMIA, VOL.17, ISSUES. 1. 2021, P.P 20 - 29

Voice of Academia
Academic Series of Universiti Teknologi MARA Kedah
Available online at http://voa.uitm.edu.my

A STUDY ON AIR POLLUTION INDEX IN SABAH AND SARAWAK USING PRINCIPAL COMPONENT ANALYSIS AND ARTIFICIAL NEURAL NETWORK

Norwaziah Mahmud[1], Nur Elissa Syazrina Zulkifli[2], Nur Syuhada Muhammat Pazil[3]

[1,2] *Faculty of Computer & Mathematical Sciences,Universiti Teknologi MARA Perlis Branch*

[3] *Faculty of Computer & Mathematical Sciences, Universiti Teknologi MARA*

ARTICLE INFO

Article history:

Received July 2020

Received in revised form Aug 2020

Accepted Oct 2020

Published Jan 2021

Keywords:

Air Pollution Index (API), Principle Component Analysis (PCA), Artificial Neural Network (ANN), Varimax rotation

Corresponding Author:

norwaziah@uitm.edu.my

ABSTRACT

This study focuses on the identification of Sabah and Sarawak air quality trends based on the data derived from the Department of Environment (DOE). Five Malaysia's monitoring stations in Sabah and Sarawak were selected based on five air pollutants for four years (2015-2018). This study aims to classify the indicators of variable predictors using the Principal Component Analysis (PCA) method and to compare the best model to predict Air Pollution Index (API) in Sabah and Sarawak using the Artificial Neural Network (ANN) model. After running the varimax rotation, only two pollutants (PM_{10} and NO_2) are the most significant pollutants out of the five pollutants. These two pollutants were used as input layers in Model B and the five pollutants were used as input layers in Model A. These two models were used to compare the best model in the ANN method. The output of ANN models was evaluated through the coefficient of determination (R^2) and Root Mean Square Error (RMSE). To identify the best model, the highest value of R^2 and the smallest value of RMSE were declared. The findings indicate that the ANN technique has been successfully implemented as a decision-making tool as well as in solving problems for proper management of the atmosphere.

1. INTRODUCTION

Recently, air pollution is one of the major environmental issues throughout the world. Pollution can take many forms, for examples, air pollution, water pollution, ground pollution, and noise pollution. Air pollution can be defined as polluting the atmosphere by a different contaminant that is endangering all life (Azid et al., 2014). It is also a type of environmental pollution that affects the air and is usually caused by smoke or other harmful gases, mainly oxides of carbon, sulfur, and nitrogen. Air pollution does not only endanger humans' health, similarly, it is inappropriate to live in the future where humans are still contributing to the increasing of the air pollution.

Based on a previous research, Ku Yusof et al. (2019) found that there were three Malaysia's major air pollution sources which were caused by industries, motor vehicles, and open burning. For the past few years, a country which was blamed to cause the burning of the forest was Indonesia. Smoke from the burning spread to Malaysia, and caused the worst polluted areas in Sabah and Sarawak. Sabah and Sarawak faced the highest record of air pollution reading. The highest Air Pollution Index (API) reading was in Sri Aman, Sarawak, which was about 395. This was the first area in Malaysia that recorded a hazardous level since the transboundary haze caused by burning agricultural practices started choking the country. There were three areas in Sarawak which continued to hover at very unhealthy levels. While in Sabah, there were two areas which recorded unhealthy API levels. The areas were Kuching (236), Samarahan (202), Sri Aman (225), Sibu (190), and Sarikei (183) (Chung, 2019). However, the forest burning in Indonesia could not be blamed exclusively as the local industry in Malaysia also led to air pollution. The smoke emitted by factories was not adequately filtered which had caused the smoke to become dangerous and had affected humans' health.

Ong (2019) explained that because of the bad air quality, the number of patients receiving treatment for conjunctivitis, asthma, and skin rashes in Labuan has increased by 40%. There is a different impact on each pollutant towards humans' health. For example, the particulate matter under 10 microns (PM_{10}) may lead to lung cancer while ozone (O_3) can decrease the function of the lung and cause coughing. Besides, the existence of carbon monoxide (CO), may affect fetal growth, while nitrogen monoxide (NO) can cause respiratory distress with symptoms such as cough, nasal congestion, and sore throat. Additionally, sulfur dioxide ($S0_2$) may also affect people with asthma as the space in the respiratory tract becomes limited (Rani et al., 2018).

Five requirements were used to measure APIs for air pollutants such as O_3, CO, nitrogen dioxide (NO_2), SO_2, and PM_{10} (Azid, Juahir, Latif, Zain, & Osman, 2013). Abd Rahman, Lee, Suhartono, and Latif (2015) explained that the API in Malaysia was developed based on the API introduced by the United States Environmental Protection Agency (USEPA). There are many studies which evaluate the air quality index, however, not many focus on Sabah and Sarawak. Therefore, this study focuses to evaluate the air quality index in Sabah and Sarawak by using PCA and ANN.

This study aims to classify the indicators of the ~~variables~~ predictors by using the PCA method. This study also aims to compare the best model to predict the API in Sabah and Sarawak. The data used in this study were secondary data obtained from the DOE of Malaysia that contained the air pollution index in Malaysia. The years involved in this research were from 2015 until 2018 which were on daily basis for certain areas in Sabah and Sarawak. The data were categorized into several pollutants which were ozone (O_3), carbon monoxide (CO), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), and particulate matter less than 10 microns (PM_{10}). Five continuous air monitoring stations were selected. The stations were Kuching (CA0004), Kota Kinabalu (CA0030), Sibu (CA0026), Labuan (CA042), and Tawau (CA0039).

2. LITERATURE REVIEW

2.1 Principal Component Analysis (PCA)

The PCA is a powerful tool for data analysis because data patterns can be difficult to find in high-dimensional data, where the luxury of graphical representation is not available. Mohamad, Ash'aari, and Othman (2015) explained that this technique is used for defining linear combinations of the initial variables which are useful for accounting for variations in those variables. PCA provides the most important and relevant variables indicating the source of the variance, since less important variables are omitted from the entire data set during the study, with very limited loss in the original information. PCs are the eigenvector of a matrix of covariance or a matrix of correlation, and each PC derives a maximum proportion of the total variance. A PC that contains eigenvalue larger or equal to 1 is used and considered significant to obtain new variables (Alonso, Arribas, Manzoor, & Caceres, 2019).

Dominick, Juahir, Latif, Zain, and Aris (2012) explained that after rotation, the loading factor is important as it represents how much the variable contributes to that particular PC and to what degree one variable is identical to another. The values of factor loading which are larger than 0.75 are known as strong, the values between 0.50 to 0.75 are known as moderate, and the values between 0.30 to 0.49 are known as weak factor loading (Azid et al., 2014). Besides, in order to measure the sampling adequacy, the Kaiser-Meyer-Olkin (KMO) test will be carried out. According to Azid et al. (2014), the value of the KMO test must be larger than 0.5.

2.2 Artificial Neural Network (ANN)

Abd Rahman, Lee, Latif, and Suhartono (2013) explained that forward and backward were the two main phases that were included in the network. According to Azid et al. (2014), the training data were spread by the hidden layer during the forward process and the resulting value contrast to the actual values for calculating the error among them. After that, the error that was calculated was propagated back towards the hidden layer. ANN's study was divided into three parts which are data training (60%), testing (20%) and validation (20%). In the training phase, the data point was used to predict and study the shape of the parameters. In the testing phase, it was chosen to analyze the generalization capabilities of the networks that are said to be trained, while for the validating phase, it was liable for carrying out the last inspection on the network trained.

2.3 Application of Artificial Neural Network

In real-world issues, there are some applications of ANN, for example, the application of ANN in the stock market index. Currency predicting is a major financial issue that needs more attention. ANN is one of the best ways to model the market value, which can simply adjust to the changes of the market and does not accommodate usual formulas (Guresen, Kayakutlu, & Daim, 2011). This study appraises the efficacy of models of neural networks that are considered to be complex and efficient in the forecasts stock-market. ANN is also used in rainfall forecasting in Queensland, Australia (Abbot & Marohasy, 2012). Researchers usually use climate indices to predict rainfall in Queensland, but the models have so far restricted them to consider-combinations of linear correlations individually. By using ANN, it has the capacity to review big values of climate indices and another input together in order to search the settlement independently of the relationship being considered. Many rainfall prediction applications use a feed-forward neural network using the generic MLP equipped with the back-propagation algorithm.

2.4 Other Methods used to Evaluate Air Pollution Index

Many methods have been used to study the API. For example, Multiple Linear Regression models (MLR). Based on Ku Yusof et al. (2019), MLR models have been used to estimate the value of PM_{10} during haze and non-haze seasons. Based on the results, it is shown that carbon monoxide clearly has a strong relationship with PM_{10} at 66.32%. Besides, the method that is used to evaluate API is the Fuzzy Time Series (FTS). The fuzzy set has been used as membership, which means that the concentrations of air pollution quality are characterized in terms of good, moderate and poor.

The cumulative logistic model is used to examine and compare the probability of every kind of haze (Zhu, Zhang, & Chen, 2017). Another method that is used in determining the air pollution index is fuzzy logic. According to Nejadkoorki (2011), fuzzy logic is used to recognize air pollution and areas at risk by numerical value to assist the Air Pollution Control District (APCD).

3. ESTIMATION METHOD

The data analysis by using PCA and ANN is explained below. There were several steps that needed to be conducted in order to achieve the objectives.

3.1 Data Collection

The prediction model was developed in this study using 36525 (5 variables × 7305 observations).The number of missing data in the data set was very small (~3%) compared to the overall data set. The data were collected by using the Division of Air Quality, DOE.

3.2 Data Pre-Treatment

If there was a missing data set, the nearest neighbour method would be used to treat the cases through the use of the XLSTAT 2019 add-in software. This method looked at the distance between each point and the closest point to it. So, the nearest neighbour method was used to estimate the missing value based on the gap endpoints.

3.3 Principal Component Analysis (PCA)

By using the PCA, which was considered to be one of the most prevalent and useful statistical methods to uncover the potential structure of a set of variables, the feature of a huge data set could be reduced. PCA had the ability to display the most significant variables that might indicate the source of the pollutants because the less significant variables were removed from the data set with minimal loss of original data. This approach was used by converting them in order to describe the variability of a large set of interrelated variables, and a smaller set of uncorrelated (independent) variables, called principal components (PCs). The PCA must be performed to produce PCs. The PCs were used as input variables in the API prediction model using the ANN approach.

3.4 Varimax Rotation

It was advisable to rotate PCs generated by the PCA using varimax rotation because they were not readily interpreted. The use of varimax rotation was intended to reduce the complexity of the components by increasing large loads and reducing small loads within the component. The varimax rotation method was implemented because this approach simplified the variable structure and made it easier and more accurate to be interpreted. In the varimax rotation process, only PCs with their eigenvalues greater than 1 were used and deemed to be important for obtaining new variables, known as varimax factors (VFs) or factor loads. The number of VFs given by varimax rotations was equal to the number of variables that were consistent with common characteristics and might include non-observable, hypothetical, and latent variables. VFs were values used to calculate the correlation among variables. For this analysis, the selection threshold was set for VFs with absolute values above 0.75. By using XLSTAT 2019 add-in software, PCA analysis was implemented (Azid et al., 2014).

3.5 ANN – API Prediction Model

In this study, the feed-forward ANN was used for prediction purposes and to determine the most important parameters affecting API values. This model was divided into three layers which were known as the input layer, hidden layer, and output layer. With different input variables, two different feed-forward ANN models were developed. Model A was built with five variables as input layers based on the original data, while Model B was created by factor scores of rotated PCs with an input value of eigenvalues greater than 1.

The procedure of trial-and-error between one to ten hidden nodes in the network structure was tested to approximate any level of accuracy and the best model for prediction values was sought. Figure 1 and Figure 2 present the structure of the ANN model. A data set was divided into three classes which are training (60%), testing (20%), and validating (20%). The training phase was used to estimate and understand the data set parameter trends. Meanwhile, the testing phase was used to determine the generalization potential of the supposedly trained network, while validating was responsible for carrying out the final check on the trained network.

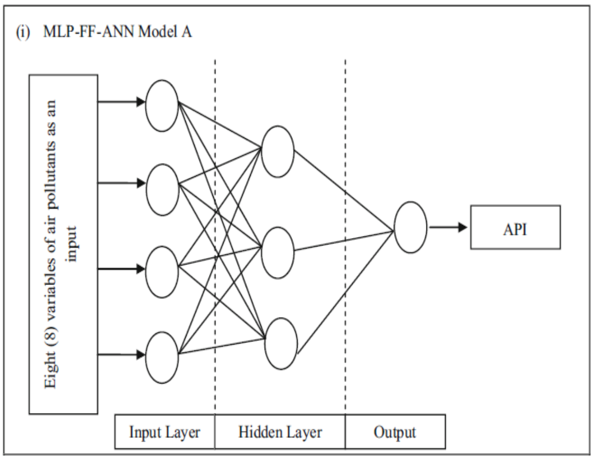

Figure 1: ANN model structure for five variable of air pollutants

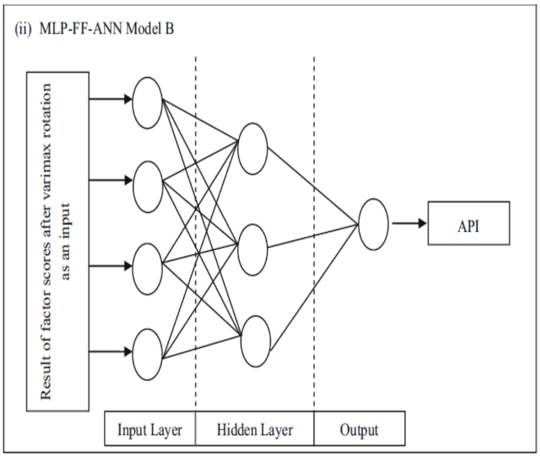

Figure 2: ANN model network structure after varimax rotation

3.6 Evaluate the Effectiveness

In order to determine the outcome of the ANN models, two performance features were reflected on the model evaluation, which were known as the coefficient of determination (R^2) and the root mean square error (RMSE). The highest value of R^2 and the lower value of RMSE were declared as the best linear models. After that, the expected amount of ANN models were contrasted to others in order to acquire a model that relied on a few variables for APIs estimation. These ANN models were implemented by using JMP15 software, which was a flexible and easy-to-use tool.

4. RESULTS AND DISCUSSION

4.1 Data Pre-Treatment

The nearest neighbor method is used to estimate the missing data set. The endpoints of the gaps are used as the estimation value for all the missing data sets.

4.2 Results for Principal Component Analysis (PCA)

After running the data using XLSTAT 2019, the result shows which pollutant is important to identify the air quality index. Table 1 shows the result of the Kaiser-Meyer-Olkin (KMO) test. From the result, it shows that all pollutants are larger than 0.5. This evaluates that all pollutants are sufficient and could be used for further study.

Table 1: Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy

Pollutant	**Result**
PM_{10}	0.558
SO_2	0.603
NO_2	0.599
O_3	0.677
CO	0.620

4.3 Identification of Air Pollution Source based on Varimax Rotation

Table 2 shows the result after varimax rotation was conducted. There are only two principal components (PCs) out of the five PCs that were selected. The two PCs were selected due to the eigenvalues which are greater than 1 which contain 61.528% of the total variance.

Table 2: Descriptive Statistics of the Selected Original PCs

Variable	**PC1**	**PC2**
Eigenvalue	1.987	1.090
Variability (%)	39.733	21.795
Cumulative (%)	39.733	61.528

However, the eigenvalues that are smaller than 1 were disregarded because of their similarity with more significant factors. This shows that multicollinearity was present among the original variables. Despite the total variance which was less than 70%, the cutoff point was calculated by using a scree plot. PC1 and PC2 were chosen since the eigenvalues are greater than 1 and it is considered significant in the varimax rotation analysis. In this study, the absolute values greater than 0.75 of the VF were put as the selection threshold because the values are strong. Two out of the five air pollutants fulfill the 0.75 VF threshold as shown in Table 3. The variables with values greater than 0.75 are PM_{10} and NO_2. These pollutants are the major contributing pollutants at the selected monitoring stations in Sabah and Sarawak.

Table 3: Rotated Factor Loadings Using Two PCs

Variable	**VF1**	**VF2**
PM_{10}	0.415	**0.758**
SO_2	0.725	-0.449
NO_2	**0.822**	-0.244
O_3	0.521	-0.139
CO	0.585	0.485
Eigenvalue	1.987	1.090
Variability (%)	39.733	21.795
Cumulative (%)	39.733	61.528

The VF1 contributed about 39.733% of the variation in the air quality data. Only one pollutant that had absolute value greater than 0.75 which was NO_2 with 0.822. While the VF2 contributed about 21.795% of the variation in the air quality data. There was one pollutant that contained absolute value greater than 0.75 which was PM_{10} with 0.758.

4.4 Design of the Neural Network

ANN models were used for prediction purposes. The process of trial-and-error numbers of hidden nodes in the network structured was tested. The data that was done by using pre-treatment was divided into three classes for both models. The three classes are training (60%), testing (20%), and validating (20%) of the data.

4.5 Result of the Effectiveness

To develop the ANN models, about 20 structured networks were tested. R^2 and RMSE are the prediction performance results of both models as shown in Table 4. In Model A, five variables were used as input variables. The values of R^2 and RMSE for Model A are 0.4454 and 10.5986, respectively which are nine hidden nodes. The value of R^2 is extremely poor fit since the value is closer to 0. Besides, in Model B, only two variables were used as input variables which are PM_{10} and NO_2. The value of R^2 in Model B is 0.3293 and 11.6552 for RMSE. For Model B, there are ten hidden nodes. The value of R^2 shows that it is also extremely poor fit because the value is not closer to 0. However, Model B shows the best result compared to Model A. According to Azid et al. (2013), Model B is selected as the best model due to the majority of predicted data which is not significantly different from the actual data. Although the R^2 values for Model A are more accurate than Model B, Model B uses fewer variables and is less complex than Model A which are the advantage of this model. Therefore, Model B not only saves time but the expense of monitoring purposes can also be saved. So, it is proven that ANN models can be used to predict the API values with negligible accuracy from all available inputs.

Table 4 : Prediction Performance for Model A and Model B by Using ANN

Model	Hidden node	R^2	RMSE
Model A	1	0.3113	11.8104
	2	0.3586	11.3972
	3	0.4203	10.8354
	4	0.4298	10.7461
	5	0.4034	10.9924
	6	0.4290	10.7540
	7	0.4351	10.6958
	8	0.4072	10.9567
	9	**0.4454**	**10.5986**
	10	0.4333	10.7129
Model B	1	0.2476	12.3445
	2	0.3116	11.8077
	3	0.3050	11.8639
	4	0.3139	11.7879
	5	0.3162	11.7677
	6	0.2927	11.9685
	7	0.3227	11.7121
	8	0.3177	11.7550
	9	0.3133	11.7931
	10	**0.3293**	**11.6552**

5. CONCLUSION

Principal Component Analysis (PCA) can identify the most important air pollutants in Sabah and Sarawak. The Kaiser-Meyer-Olkin (KMO) test shows the sampling adequacy is greater than 0.5 and it is considered that all pollutants can be applied for further analysis. The two PCs that have been generated by the rotated PCA indicate that only two out of five pollutants are the most important pollutants that have caused air pollution in Sabah and Sarawak which are PM_{10} and NO_2. Besides, for the Artificial Neural Network (ANN), it was divided into two models which are Model A and Model B where Model A contains five pollutants as input variables which are PM_{10}, SO_2, NO_2, O_3, and CO, and Model B only uses PM_{10} and NO_2.

The finding shows that Model A gives a larger value of R^2 than Model B. But, Model B gives better prediction compared to Model A in the terms of R^2 value. However, the prediction performance for Model B is lower than Model A, but the models can predict the API within acceptable accuracy. This shows that the rotated PCs are more effective and efficient due to the reduction of air pollutant without losing any important information. Future research may use other methods of prediction that is not used in this study. An example is the fuzzy time series. Researchers can also use the method of mathematical analysis and others that are suitable for the data. This method can also be used in another field such as marketing, insurance, and education. Besides, the Ministry of Environment, Malaysia can use this study as a reference to analyze and predict the air quality in Peninsular Malaysia. Lastly, future researchers can use other air pollutants such as wind speed, humidity, methane (CH_4), and others.

ACKNOWLEDGEMENT

The authors acknowledge the Air Quality Division of the Department of Environment (DOE) under the Ministry of Natural Resource and Environment, Malaysia, for permitting us to utilize air quality data, and supporting this study.

REFERENCES

Abbot, J., & Marohasy, J. (2012). Application of artificial neural networks to rainfall forecasting in Queensland, Australia. *Advances in Atmospheric Sciences*, *29*(4), 717–730. https://doi.org/10.1007/s00376-012-1259-9

Abd Rahman, N. H., Lee, M. H., Latif, M. T., & Suhartono. (2013). Forecasting of air pollution index with artificial neural network. *Jurnal Teknologi*, *63*(2), 59–64. https://doi.org/10.11113/jt.v63.1913

Abd Rahman, N. H., Lee, M. H., Suhartono, & Latif, M. T. (2015). Artificial neural networks and fuzzy time series forecasting: An application to air quality. *Quality and Quantity*, *49*(6), 2633–2647. https://doi.org/10.1007/s11135-014-0132-6

Alonso, D. N., Arribas, L. V. P., Manzoor, S., & Cáceres, J. O. (2019). Statistical tools for air pollution assessment: Multivariate and spatial analysis studies in the Madrid region. *Journal of Analytical Methods in Chemistry*, *2019*(6), 1-9. https://doi.org/10.1155/2019/9753927

Azid, A., Juahir, H., Latif, M. T., Zain, S. M., & Osman, M. R. (2013). Feed-forward artificial neural network model for air pollutant index prediction in the southern region of Peninsular Malaysia. *Journal of Environmental Protection*, *04*(12), 1–10. https://doi.org/10.4236/jep.2013.412a1001

Azid, A., Juahir, H., Toriman, M. E., Kamarudin, M. K. A., Mohd Saudi, A. S., Che Hasnam, C. N., et al. (2014). Prediction of the level of air pollution using principal component analysis and artificial neural network techniques: A case study in Malaysia. *Water, Air, and Soil Pollution*, *225*(8). https://doi.org/10.1007/s11270-014-2063-1

Chung, C. (2019, September 9). More efforts needed to 'clear the air'. The Star. Retrieved from https://www.thestar.com.my/news/nation/2019/09/09/more-efforts-needed-to-clear-the-air

Dominick, D., Juahir, H., Latif, M. T., Zain, S. M., & Aris, A. Z. (2012). Spatial assessment of air quality patterns in Malaysia using multivariate analysis. *Atmospheric Environment*, *60*, 172-181. http://dx.doi.org/10.1016/j.atmosenv.2012.06.021

Guresen, E., Kayakutlu, G., & Daim, T. U. (2011). Using artificial neural network models in stock market index prediction. *Expert Systems with Applications*, *38*(8)

Ku Yusof, K. M. K., Azid, A., Abdullah Sani, M. S., Samsudin, M. S., Muhammad Amin, S. N. S., Abd Rani, N. L., & Jamalani, M. A. (2019). The evaluation on Artificial Neural Networks (ANN) and Multiple Linear Regressions (MLR) models over particulate matter (PM10) variability during haze and non-haze episodes: A decade case study. *Malaysian Journal of Fundamental and Applied Sciences*, *15*(2), 164–172. https://doi.org/10.11113/mjfas.v15n2.1004

Mohamad, N. D., Ash'aari, Z. H., & Othman, M. (2015). Preliminary assessment of air pollutant sources identification at selected monitoring stations in Klang Valley, Malaysia. *Procedia Environmental Sciences*, *30*, 121-126. https://doi.org/10.1016/j.proenv.2015.10.021

Nejadkoorki, F. (2011). Air Pollution Monitoring Using Fuzzy Logic in Industries. In Vahdat, S. E., Nakhaee, F. M, *Advanced Air Pollution* (pp. 21-30). Iran. https://doi.org/10.5772/16947

Ong, H. S. (2019, September 13). Respiratory and eye-related illnesses on the rise due to haze. The Star. Retrieved from https://www.thestar.com.my/news/nation/2019/09/13/respiratory-and-eye-related-illnesses-on-the-rise-due-to-haze

Rani, N. L. A., Azid, A., Khalit, S. I., Juahir, H., & Samsudin, M. S. (2018). Air Pollution Index Trend Analysis in Malaysia, 2010-15. *Polish Journal of Environmental Studies, 27(2).*

Zhu, Y., Zhang, T., & Chen, C. (2017). Study on probability estimation of haze in Beijing based logistic regression model. *Journal of Geoscience and Environment Protection, 5*(06), 37. https://doi.org/10.4236/gep.2017.56005

Cawangan Kedah
Kampus Sungai Petani

e-ISSN: 2682-7840