PLANNING MALAYSIA:
Journal of the Malaysian Institute of Planners
VOLUME 20 ISSUE 4 (2022), Page 74 – 86

INCORPORATING ANN WITH PCR FOR PROGRESSIVE DEVELOPING OF AIR POLLUTION INDEX FORECAST

Ang Kean Hua[1], Mohamad Pirdaus Yusoh[2], Junaidah Yusof[3], Mohd Fadzil Ali Ahmad[4], Syazwani Yahya[5], Sazwan Syafiq Mazlan[5], Munaliza Jaimun[6]

[1]*Faculty of Social Sciences and Humanities,*
[2]*Borneo Institute for Indigenous Studies,*
UNIVERSITI MALAYSIA SABAH (UMS)
[3]*Faculty of Social Sciences and Humanities*,
UNIVERSITI TEKNOLOGI MALAYSIA (UTM)
[4]*Kolej Kejuruteraan (KKEJ),*
UNIVERSITI MALAYSIA PAHANG (UMP)
[5] *Faculty of Science and Technology*,
[6] *Faculty of Business and Management*,
QUEST INTERNATIONAL UNIVERSITY (QIU)

Abstract

This study circumscribes the modelling for concentration of Air Pollutant Index (API) in Selangor, Malaysia. The five monitored environmental pollutant concentrations (O_3, CO, NO_2, SO_2, PM_{10}) for ten years (2006 to 2015) data are used in this study to develop the prediction of API. The selected study area is located in rapid urbanised areas and surrounded by a number of industries, and is highly influenced by congested traffic. The principal component regression (PCR) for the combination of the principal component analysis together with multiple regression analysis, and artificial neural network (ANN), are used to predict the API concentration level. An additional approach using a combination method of PCR and ANN are included into the study to improve the API accuracy of prediction. The resulting prediction models are consistent with the observed value. The prediction techniques of PCR, ANN, and a combination method of R^2 values are 0.931, 0.956, and 0.991 respectively. The combination method of PCR and ANN are detected to reduce the root mean square error (RMSE) of API concentration. In conclusion, different techniques were used in the combination method of API prediction which had improved and provided better accuracy rather than being dependent on the single prediction model.

***Keywords**:* Air Pollutant Index concentration, principal component regression, artificial neural network, combine prediction model

[2] Lecturer at Universiti Malaysia Sabah. Email: pirdaus@ums.edu.my

INTRODUCTION

Today, developed and developing countries are unavoidable in fronting the issues regarding the ambient air quality. Biological, chemical and physical properties have become the main determination to the air quality status, which can be manifested through the air pollutant index (API). Precisely, the API is a numerical index which is measured based on the combination of selected air quality variables. This index is vital in assessing the differences in the sources of air quality (Hua, 2018; Azid et al, 201a). After the compliance has been decided, the statistical data can be beneficial by carefully warning the public regarding the health effects (Azid et al, 2014). When air pollutants have exceeded the normal condition of the ambient level of air quality status, this could bring chronic effects to human health (Moustris et al, 2010). Hence, air pollution required serious attention by authorities, especially in highly populated and manufacturing industries in the urban areas (Mutalib et al, 2013). In Malaysia, API is an indicator used by the government to monitor the air quality status since 1989 (Hua, 2018; Azid et al, 2014). By combining the five categories of sub-indexes of air pollutants such as ozone (O_3), carbon monoxide (CO), nitrogen dioxide (NO_2), sulphur dioxide (SO_2), and particulate matters under 10 µm (PM_{10}) (Mutalib et al, 2013) which provide the total value of API. The increment of one unit value of API will cause the air pollution level to increase by one unit and this circumstance will affect the human health greater. The level of air quality status has been monitored based on the Recommended Malaysian Air Quality Guidelines (RMAQG) since 1989 by the Department of Environment (DOE), Malaysia (Hua, 2018; Azid et al, 2014a; Mutalib et al, 2013).

For Malaysia to be a developed country by 2020 in the industrialised sectors, it is important to be linked with rapid economic growth. These circumstances cause degradation in urban environments through tremendous industrial pollution. Air pollution becomes a major issue in rapid economic growth which negatively affects human quality of living, agriculture, animals as well as the ecosystems (Azid et al, 2015; Mutalib et al, 2013). Simultaneously, non-living beings such as buildings, monuments, statues, etc., are suspected to be included in the damages list due to air pollution issues. According to past studies, the major air pollutant sources detected in Malaysia were derived from mobile, stationary, and trans-boundary pollution (Dominick et al, 2012; Jamal et al, 2004; Afroz et al, 2003; Awang et al, 2000). Specifically, mobile pollution can be referred to as motor vehicle emissions (Awang et al, 2000); while stationary pollution are referred to as the dust emitted from urban construction works and quarry, open burning, power plants, etc. (Dominick et al, 2012); and trans-boundary pollution originated from air pollution of forest burning or volcanic eruptions which were transported from neighbouring countries (Jamal et al, 2004). The major issues highlighted are involved with trans-boundary pollution which occurred in Southeast Asian countries like Southern Thailand, Brunei, Singapore, as well as Malaysia, where these catastrophic events happened due to the open and uncontrolled burning of forest by farmers to clear the land for other activities (Torre, 2013). A greater impact of air pollution has increased when El-Nino phenomenon takes place to create extreme dry weather conditions (Queh, 2002); which has already been experienced by Malaysians in 1997 (Brauer and Hisham-Hashim, 1998). Therefore, frequent assessment and monitoring of the air quality data by the majority of the regions denoting the pollutant sources have increased with time which have inconsistently satisfied the national standard levels.

Air pollutant sources are uncontrollable major tasks globally. According to DOE Malaysia, the air quality is collected consistently which monitors the data daily to acknowledge the status as well as to inform the people about the pollutant concentrations at the exact time (Dragomir, 2010). The major source of air pollution which is suspected to occur in highly built-up areas, like urban and suburban areas, are O_3 and PM_{10} (Hua, 2018; Azid et al, 2015; Azid et al, 2014; Mutalib et al, 2013), which are also suspected to become the main cause of deterioration on human health (Azid et al, 2014; Mutalib et al, 2013). Therefore, DOE Malaysia had provided important data regarding air pollution through various programmes such as using environmetric techniques, which are involved with cluster analysis (CA), discriminant analysis (DA), factor analysis (FA), and principal component analysis (PCA) (Hua, 2018; Hua, 2017; Hua et al, 2016; Azid et al, 2015; Azid et al, 2014; Mutalib et al, 2013). These techniques have the advantage in understanding the air quality status in the studied area by interpreting the complex databases, as well as benefits in the programmes for monitoring the air quality with efficient management. Nevertheless, the prediction on air quality had become a challenge when using a simple mathematical formula especially involved with complex data (Mutalib et al, 2013) which are incapable of relating the non-linear with the different variables (Afzali et al, 2012). Since the air quality exists in the stochastic time series and are able to be predicted based on the historical data (Giorgio and Piero, 1996), therefore, it is vital to help by reducing human health issues by planning and controlling strategies for proper actions.

To track and predict the air quality status, it is important to understand the methods used for modelling in this study. The aim for this paper is to select the explanatory variables by applying principal components before including it into the multiple regression analysis, and continue inserting it into the artificial neural network of the resulting residuals. Multivariate Linear Regression (MLR) analysis is considered as one of the methods which are widely used to express the dependence variable onto various independent variables. Nevertheless, these techniques seem to confront an issue when independent variables are correlated between each other (McAdams et al, 2000). Due to the correlation occurring among the independent variable, which is also known as multicollinearity, which could hardly assist in providing accurate contributors of physical process in an equation. Therefore, the method to avoid multicollinearity and overlapping information of independent variables is by using the principal component analysis (PCA) of environmetric techniques. PCA is popular in identifying the variations of environmental pollution and investigating the factors that influence the quality concentration (Lengyel et al, 2004; Klaus et al, 2001). Simultaneously, PCA has the ability to resolve the issue regarding multicollinearity and examine the relationships within independent variables, which could react as predictors. An advantage from this technique in optimising spatial patterns as well as removing multicollinearity problems, is the newly provided variables as predictors by PCA which are considered ideal and appropriate for input into the regression equation (PCR) (Abdul-Wahab et al, 2005). The specific procedure for applied PCR regarding the pollution can be found elsewhere (Hua et al, 2018; Azid et al, 2014; Dominick et al, 2012).

The models of artificial neural networks (ANN) have the capability in handling the problems of multiple variables as well as determining the non-linear relationship involved with air quality index. Hence, ANNs is considered having flexibility, efficiency, and accuracy, to practice non-linear patterns between input and output values as well as defining the solution towards the complex data accurately (Rahman et al, 2013) including additional new data that was presented (Azid et al, 2015; Azid et al, 2014; Mutalib et al, 2013). This technique undergoes a training process or early experience to provide its own weight distribution for the linkage through learning procedure, which does not require an algorithm to determine the solution of the problem (Garcia et al, 2011). Therefore, past studies had proven that ANN have the potential in solving environmental problems, particularly involved with air quality pollution (Brunelli et al, 2007; Tecer, 2007; Perez and Reyes, 2006; Niska et al, 2005; 2004). To provide better performance, a combination of methods of prediction had been introduced since 1969 by Bates and Granger, whereby this technique includes bootstrapping, gagging, stacking and boosting, rather than being dependent only on the single prediction technique. A combination of methods suggested for use in this study is PCR and ANN, which were supported by past literature in various fields of studies that have the capability to increase the prediction performance (Chia et al, 2012; Noori et al, 2010; 2009). In general, the selected combination methods indicate a better performance than the possess individual prediction. Consequently, although there is a minor debate regarding the inconsistency of ANNs' performance (Khashei and Bijari, 2011) which are involved with a large number of factors such as network structure, training methods, as well as sample data (Chia et al, 2012;

Noori et al, 2010; 2009), these weaknesses can be overcome by getting involved with the idea of combination prediction methods.

As stated earlier, this study aims to predict the air quality status by using principal component regression (PCR) by including the techniques of principal components and multiple regression analysis. Apart from the prediction methods involved with PCR and ANN individually, a combination of both methods are also suggested in this study, not only to identify the concentration of pollutant sources, but also to investigate the relationship between air quality variables in the study area.

MATERIALS AND METHODS

Study Area

Eight (8) sampling stations have been set up for monitoring in the study area of Selangor state (Figure 1), with the geographical coordinate of latitude between 2°35’23.53”N to 3°47’55.09”N and longitude between 100°56’25.09”E to 101°57’58.50”E (Table 1). With approximately 8104 km^2 area, the Selangor state is placed in the western part of Peninsular Malaysia with the boundary between Perak in the north, Pahang in the east, as well as Negeri Sembilan in the south, and facing the seaside of the Straits of Malacca in the west. Specifically, selected sampling stations at different locations are due to highly built industrial areas, residential areas, and heavy traffic congestion in the study area. According to the DOE (2012) report, the Malaysia air quality status was considered between good to moderate levels for all the time being. Generally, majority of the studies reported that Malaysia’s API is affected by particulate matter (PM_{10}) (Awang et al, 2000) due to the PM_{10} concentration which has exceeded the minimum good quality level and is unceasingly greater than the other pollutants (Othman et al, 2010). Although Malaysia are cleared from typhoon, volcanic eruptions, as well as earthquakes, which keeps maintaining the air quality status; however, high population and rapid economic growth have become the main factor to worsen the level of air quality and it’s vital to conduct a prediction for early preparation by consistently assessing the data.

Table 1: The details of 8 monitoring stations in Selangor, Malaysia.

Station ID	Location	Latitude	Longitude
Station 1	Klang, Selangor	3°0’53.72”N	101°24’47.02”E
Station 2	Petaling Jaya, Selangor	3°7’59.37”N	101°36’28.53’E
Station 3	Shah Alam, Selangor	3°6’17.03”N	101°33’21.66”E
Station 4	Kuala Selangor, Selangor	3°19’16.13”N	101°15’22.61”E
Station 5	Putrajaya, Wilayah Persekutuan	2°54’52.49”N	101°41’23.69”E
Station 6	Cheras, Kuala Lumpur	3°6’22.62”N	101°43’5.00”E

Station 7	Batu Muda, Kuala Lumpur	3°12’45.08”N	101°40’56.47”E
Station 8	Banting, Selangor	2°48’59.98”N	101°37’23.21”E

Data Collection and Data Treatment

The data of air pollutant sources include ozone (O_3), carbon monoxide (CO), nitrogen dioxide (NO_2), sulphur dioxide (SO_2), and particulate matters under 10 µm (PM_{10}) which provides the status of API which was employed by DOE Malaysia since 1989, and obtained from the department from January 01, 2006 to December 31, 2015. Teledyne Technologies Inc., USA and Met One Instrument Inc., USA, are the instruments used to monitor the data of air quality by DOE. Based on the Standard Operating Procedures for Continuous Air Quality Monitoring (2006), PM_{10} was assessed using BAM-1020 Beta Attenuation Mas Monitor analyser from the apparatus of Met One Instrument Inc. USA. The equipment used in capturing the data is capable of detecting the high level of resolution of 0.1 µg m^{-3} at a 16.7-L min^{-1} flow rate, whereby the limit of detection for the lower levels are <4.8 µg m^{-3} and <1.0 µg m^{-3} for one (1) and twenty-four (24) h, respectively. Simultaneously, the other variables include a monitor using Teledyne Technologies Inc., USA, which are involved with Teledyne API Model 100A/100E, Teledyne API Model 200A/200E, Teledyne API Model 300A/300E, as well as Teledyne API Model 400A/400E to evaluate the NO_2, SO_2, CO and O_3, respectively. The analyser used to detect the lowest level are at 0.4 ppb in SO_2, NO_2, and O_3 using UV fluorescence method, chemiluminescence method, as well as UV absorption (Beer-Lambert) method, respectively; while the CO can only be measured by using the infrared absorption (Beer-Lambert) method with the lowest level of detection at 0.4 ppm. All variables were tested at the precision level of 0.5%.

This study uses the statistical data of 960 dataset (12 data per stations x 8 stations x 10 years) and total number of 4800 dataset (12 data per stations x 8 stations x 10 years x 5 variables). The missing data for the overall data point is very small (approximately less than 5%). By using software SPSS Ver. 23, the nearest neighbour method was included to facilitate the data analysis (Hua, 2018; Azid et al, 2015). To be specific, the nearest neighbour method determines the distance between the two and the closest point, where the gaps of the end points are employed towards all missing values (Azid et al, 2014; Dominick et al, 2012). The equation is shown as Eq. 1;

$$y = y_1 \; if \; x \leq x_1 + [\frac{x_2 - x_1}{2}] \text{or } y = y_1 \; if \; x > x_1 + [\frac{x_2 - x_1}{2}] (1)$$

where, y refers to interpolant, x refers to interpolant's time point, y_1 and x_1 refer to coordinate of the gap for beginning points, and y_2 and x_2 refer to coordinate of the gap for endpoints.

The Hybrid Approach of Methodology

Hybrid approach had been reported in previous literature, which indicated that although the linear model has successfully resulted in the data that exists in linear relationships, difficulties arise during the involvement with several variables, multicollinearity, as well as outliers. Meanwhile, ANN has the ability to model the data which are involved with non-linear variables. Hence, the alternative to improve the prediction performance is by the applied hybrid approach that has the capability to enhance by capturing the possibility of different patterns in the data especially involved with linear and non-linear empirical data (Noori et al, 2010; Al-Alawi et al, 2008). Consequently, this study uses a combination method to determine between the components for linear and non-linear as in Eq. 2;

$$y_1 = G_1 + N_1 (2)$$

where, G_1 is the component of the linear and N_1 is the component of non-linear; whereby both components are evaluated by employing the subsequent techniques.

Firstly, the technique for principal component regression (PCR) is applied to calculate the component for linear data (G_1) which is involved with the five (5) variables of air quality in order to predict the concentration of API. The procedures with specific information to employ PCR for prediction on the API are available to be obtained elsewhere in Hua (2018) and Azid et al (2014).

Secondly, the linear model that produced the residuals contain the non-linear relationship which can be explained further in Eq. 3;

$$e_t = y_t - G_t (3)$$

where, e_t refers to the model for PCR (linear model) of the residual at the time t, y_t refers to the value which is observed at the time t, and G_t refers to the predict value for the time t which resulted from the PCR model (Abdul-Wahab et al, 2005).

Thirdly, to model the ANN can only be carried out based on the residuals resulted through the model of PCR which are included into Eq. 4;

$$e_t = f(e_{t-1}, e_{t-2}, \dots\dots, e_{t-n}) + \varepsilon_1 (4)$$

where, f refers to as the function for non-linear that are examined using a neural network, and ε_l can be referred to the error value of random. The prediction of API for the combination method between ANN and PCR can be expressed in Eq. 5;

$$\hat{y} = G_t + \hat{j_1}(5)$$

where, $\widehat{j_1}$ is referred to the value of ANN. Hence, in depth techniques used in the hybrid methods can be obtained from the studies of Al-Alawi (2008) and Zhang (2003); while the procedures to apply for the ANNs methods to predict API that originated from the studies are derived from Hua (2018) and Azid et al (2014).

RESULT AND DISCUSSION

The outcome for the analysis of predicted API using different methods of PCR, ANN and the combination method can be expressed in Table 2. According to Figure 2, the result which indicated from the method analysis of PCR onto API concentration shows that the adjusted coefficient of variations (R^2) is considered fairly satisfied at 0.931, whereby the concentration of API is under-estimated along the sampling stations. Nonetheless, the most significant notables are in sampling station 1 and sampling station 6. The air pollutant sources are suspected to originate from the shipping activities from Port Klang valley as well as motor vehicle emissions and construction works from highly built-up urban areas in the Kuala Lumpur district. Simultaneously, prediction on API concentration using the model of ANN shows a significant improvement to provide the R^2 about 0.956 (Figure 3). On the same note, the results from ANN have an overestimate in sampling station 2, 3, 6 and 7; which highlighted that the air quality is associated with the urban pollution which includes the meteorological factor which affects the surrounding environment. For instance, wind directions are considered as one of the consequences of meteorological state to influence directly on the emission effect of the atmosphere to enhance the air pollution status. Lastly, applying the model of PCR into ANN resulting in residuals for an almost perfect fit of API concentrations with R^2 are 0.991 (Figure 4). Therefore, applying the last model of the combination method and examining the residuals indicate that the results are almost normally distributed with approximately zero mean and non-multicollinearity issues on the correlation detected.

Table 2: The actual and predicted mean for API concentration using different prediction techniques.

Sampling Station	Mean			
	Actual API	**Predicted API by PCR**	**Predicted API by ANN**	**Predicted API by combine method**
1	10.90	8.05	10.8	10.85
2	8.43	7.99	9.71	8.38
3	8.88	8.55	9.62	8.67
4	8.37	7.76	7.26	8.41
5	7.44	6.75	7.82	7.07
6	8.51	6.11	9.76	8.48
7	5.73	4.8	6.25	5.65
8	5.50	4.28	4.87	5.55

RMSE or root mean square error plus correlation coefficients are used in this paper to determine a comparison between actual and predicted value in the models. Generally, RMSE can be expressed through Eq. 6;

$$RMSE = \sqrt{\frac{\sum \quad (A_i - E_i)^2}{n}} (6)$$

where, A_i is the API concentration for the actual value for every station, as well as E_i refers to each of the models used to estimate the concentration level. Table 3 shows the outcome of RMSEs for the total value of predicted API concentration based on the sampling stations for each model used in the study. The resulting errors for the range in model of PCR are 5.25 to 11.81, meanwhile the model of ANN error ranges between 3.89 to 4.16, as well as the combination method ranges between 2.97 to 4.73. In order to identify the ranks, it is suggested to use the Wilkoxon test, where the results indicated are between ANN and the combination models are not having any significant difference in statistical analysis (p-value = 0.257), yet, only the model of PCR has a significant difference for either of the two models used in the study (p-value < 0.01).

Table 3: The predicted API concentrations using RMSE for selected sampling stations of each model.

Sampling Station	RMSE of AQI		
	PCR method	**ANN method**	**Combine method**
1	5.25	3.10	3.48
2	6.58	3.33	3.65
3	7.49	3.62	4.05
4	11.81	4.16	4.11
5	7.70	3.51	3.78

6	5.78	2.89	2.97
7	6.26	3.05	3.06
8	10.65	3.68	4.73

Lastly, the logarithms predicted for the concentration of API for the three methods versus the logarithms of observed value are transformed into scatter plots, which are expressed in Figure 5 to Figure 7. The adequacy of the model can be determined through the clustered points that are close to the 45° tangent line. The scatter plot resulted from the regression line for PCR, ANN and the combination method are 0.865, 0.923, and 0.981 respectively. Although the slope for a perfect fit from the outcomes of the three models are not significantly different, PCR models with the value of R^2 are considered the lowest among three with 0.838; continued by the ANN model with 0.887, as well as the combined model with 0.982. Hence, since the comparison performances for RMSE results of ANN and the combination method are almost similar, nevertheless, the combination method provides a better outcome than the ANN model. The main reason to enable the combination method to be chosen for better performance is because of the applied principal component which has the capability to indicate variables of predictor to clarify the API concentration level and the variation naturally. In other words, to be more specific, the selected through principal component method to provide the variables of predictor are O_3 and PM_{10} which are considered as the main pollutant sources to affect the air quality level in majority of the sampling stations, while SO_2 and CO are suspected to increase in certain sampling stations.

CONCLUSION

This study illustrated that the prediction of API concentration is more accurate when using different methods to model by using PCR techniques to select the most appropriate variables to provide as explanatory variables before applying to the regression technique for model purposes. Simultaneously, the process has continued to input the resulted residuals into ANN technique for the combination method. In other words, the model of ANN used in the combination method has the ability to capture the undetected residuals of non-linearity in PCR to fit into the analysis. Despite the model of ANN have the capability to contribute a better fit compared to the model of PCR, nevertheless, the combination between ANN and PCR have proven to have a greater significant accuracy of prediction for API concentration level for the monitored sampling stations. Techniques involved with the root mean square error (RMSE) plus mean absolute percentage error (MAPE) are used in the study to substantially lower the prediction of the combination method, when compared to the individual method for prediction of

PCR or prediction of ANN. The modelling involved in this approach is significant and provides a better potential for other field of studies.

DATA AVAILABILITY

The data used to support the findings of the study are available from the corresponding author upon request.

ACKNOWLEDGEMENT

The authors would like to thanks to Research Management Centre (PPP), Universiti Malaysia Sabah (UMS) for financial support and fully funded to publish this article journal; continue by the Malaysian Department of Environment (DOE) and the Department of Town and Country Planning (JPBD) for providing air quality data and GIS map-based information, which including land use activities in Selangor State.

REFERENCES

Abdul-Wahab, S. A., Bakheit, C. S., & Al-Alawi, S. M. (2005). Principal component and multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations. *Environmental Modelling & Software*, *20*(10), 1263-1271.

Afroz, R., Hassan, M. N., & Ibrahim, N. A. (2003). Review of air pollution and health impacts in Malaysia. *Environmental research*, *92*(2), 71-77.

Afzali, M., Afzali, A., & Zahedi, G. (2012). The potential of artificial neural network technique in daily and monthly ambient air temperature prediction. *International Journal of Environmental science and development*, *3*(1), 33.

Al-Alawi, S. M., Abdul-Wahab, S. A., & Bakheit, C. S. (2008). Combining principal component regression and artificial neural networks for more accurate predictions of ground-level ozone. *Environmental Modelling & Software*, *23*(4), 396-403.

Awang, M.B., Jaafar, A.B., Abdullah, A.M., Ismail, M.B., Hassan, M.N., Abdullah, R., Johan, S., & Noor, H. (2000). Air quality in Malaysia: impacts, management issues and future challenges. *Respirology*, *5*(2), 183-196.

Azid, A., Juahir, H., Toriman, M. E., Endut, A., Kamarudin, M. K. A., Rahman, A., & Nordin, M. (2015). Source apportionment of air pollution: A case study in Malaysia. *Jurnal Teknologi*, *72*(1), 83-88.

Azid, A., Juahir, H., Toriman, M.E., Kamarudin, M.K.A., Saudi, A.S.M., Hasnam, C.N.C., Aziz, N.A.A., Azaman, F., Latif, M.T., Zainuddin, S.F.M., & Osman, M.R. (2014). Prediction of the level of air pollution using principal component analysis and artificial neural network techniques: A case study in Malaysia. *Water, Air, & Soil Pollution*, *225*(8), 2063.

Bates, J. M., & Granger, C. W. (1969). The combination of forecasts. *Journal of the Operational Research Society*, *20*(4), 451-468.

Brauer, M., & Hisham-Hashim, J. (1998). Peer reviewed: fires in Indonesia: crisis and reaction. *Environmental science & technology*, *32*(17), 404A-407A.

Bruelli, U., Piazza, V., Pignato, L., Sorbello, F., & Vitabile, S. (2007). Two days ahead prediction of daily maximum concentration of SO2, O3, PM10, NO2, CO in the urban area of Palermo, Italy, Atom. *Journal of Environmental engineering*, *41*(14), 2967-2995.

Chia, K. S., Rahim, H. A., & Rahim, R. A. (2012). Neural network and principal component regression in non-destructive soluble solids content assessment: a comparison. *Journal of Zhejiang University SCIENCE B*, *13*(2), 145-151.

Department of Environmental (DOE) Malaysia (2012). *Malaysia Environmental Quality Report*. Kuala Lumpur: Department of Environment, Ministry of Natural Resources and Environment.

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.

Dragomir, E. G. (2010). Air quality index prediction using K-nearest neighbor technique. *Bulletin of PG University of Ploiesti, Series Mathematics, Informatics, Physics, LXII*, *1*(2010), 103-108.

Garcia, I., Rodriguez, J. G., & Tenorio, Y. M. (2011). Artificial neural network models for prediction of ozone concentrations in Guadalajara, Mexico. In *Air Quality-Models and Applications*. InTech.

Giorgio, F. & Piero, M. (1996). Mathematical models for planning and controlling air quality. *Proceedings of IIASA Workshop, 17*.

HUA, A. K. (2018). Applied Chemometric Approach in Identification Sources of Air Quality Pattern in Selangor, Malaysia. *Sains Malaysiana*, *47*(3), 471-479.

Hua, A. K. (2017). Analytical and Detection Sources of Pollution Based Environmetric Techniques in Malacca River, Malaysia. *Applied Ecology and Environmental Research*, *15*(1), 485-499.

Hua, A. K., Kusin, F. M., & Praveena, S. M. (2016). Spatial Variation Assessment of River Water Quality Using Environmetric Techniques. *Polish Journal of Environmental Studies*, *25*(6).

Jamal, H.H., Pillay, M.S., Zailina, H., Shamsul, B.S., Sinha, K., Zaman Huri, Z., Khew, S.L., Mazrura, S., Ambu, A., Rahimah, A., & Ruzita, M.S. (2004). *A study of health impact and risk assessment of urban air pollution in Klang valley*, UKM Pakarunding Sdn Bhd, Malaysia, Kuala Lumpur.

Khashei, M., & Bijari, M. (2011). A novel hybridization of artificial neural networks and ARIMA models for time series forecasting. *Applied Soft Computing*, *11*(2), 2664-2675.

Klaus, D., Poth, A., Voss, M., & Jáuregui, E. (2001). Ozone distributions in Mexico City using principal component analysis and its relation to meteorological parameters. *Atmosfera*, *14*(4), 171-188.

Lengyel, A., Héberger, K., Paksy, L., Bánhidi, O., & Rajkó, R. (2004). Prediction of ozone concentration in ambient air using multivariate methods. *Chemosphere*, *57*(8), 889-896.

McAdams, H. T., Crawford, R. W., & Hadder, G. R. (2000). A vector approach to regression analysis and its application to heavy-duty diesel emissions. Society of Automotive Engineers, Inc, Contract with the Energy Division of Oak Ridge National Laboratory (ORNL). *Contract No. DE-AC05-00OR22725*.

Moustris, K. P., Ziomas, I. C., & Paliatsos, A. G. (2010). 3-Day-ahead forecasting of regional pollution index for the pollutants NO 2, CO, SO 2, and O 3 using artificial neural networks in Athens, Greece. *Water, Air, & Soil Pollution*, *209*(1-4), 29-43.

Mutalib, S.N.S.A., Juahir, H., Azid, A., Sharif, S.M., Latif, M.T., Aris, A.Z., Zain, S.M., & Dominick, D. (2013). Spatial and temporal air quality pattern recognition using environmetric techniques: a case study in Malaysia. *Environmental Science: Processes & Impacts*, *15*(9), 1717-1728.

Niska, H., Rantamäki, M., Hiltunen, T., Karppinen, A., Kukkonen, J., Ruuskanen, J., & Kolehmainen, M. (2005). Evaluation of an integrated modelling system containing a multi-layer perceptron model and the numerical weather prediction model HIRLAM for the forecasting of urban airborne pollutant concentrations. *Atmospheric Environment*, *39*(35), 6524-6536.

Niska, H., Hiltunen, T., Karppinen, A., Ruuskanen, J., & Kolehmainen, M. (2004). Evolving the neural network model for forecasting air pollution time series. *Engineering Applications of Artificial Intelligence*, *17*(2), 159-167.

Noori, R., Khakpour, A., Omidvar, B., & Farokhnia, A. (2010). Comparison of ANN and principal component analysis-multivariate linear regression models for predicting the river flow based on developed discrepancy ratio statistic. *Expert Systems with Applications*, *37*(8), 5856-5862.

Noori, R., Abdoli, M. A., Ghazizade, M. J., & Samieifard, R. (2009). Comparison of neural network and principal component-regression analysis to predict the solid waste generation in Tehran. *Iranian Journal of Public Health*, *38*(1), 74-84.

Othman, N., Jafri, M. Z. M., & San, L. H. (2010). Estimating particulate matter concentration over arid region using satellite remote sensing: A case study in Makkah, Saudi Arabia. *Modern Applied Science*, *4*(11), 131.

Perez, P., & Reyes, J. (2006). An integrated neural network model for PM10 forecasting. *Atmospheric Environment*, *40*(16), 2845-2851.

Quah, E. (2002). Transboundary pollution in Southeast Asia: the Indonesian fires. *World Development*, *30*(3), 429-441.

Rahman, N. H. A., Lee, M. H., & Latif, M. T. (2013). Forecasting of air pollution index with artificial neural network. *Jurnal Teknologi (Sciences and Engineering)*, *63*(2), 59-64.

Tecer, L. H. (2007). Prediction of SO 2 and PM Concentrations in a Coastal Mining Area (Zonguldak, Turkey) Using an Artificial Neural Network. *Polish Journal of Environmental Studies*, *16*(4).

Torre, F.D.L. (2006). *Indon haze spreads to NML.*Retrieved from http://www.saipantribune.com/newsstory.aspx?cat¼1&newsID¼61706

Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, *50*, 159-175.

Received: 23rd September 2022. Accepted: 23rd November 2022

