The Nonlinear Autoregressive Exogenous Neural Network Performance in Predicting Malaysia Air Pollutant Index

Rosminah Mustakim, Mazlina Mamat#
Faculty of Engineering, Universiti Malaysia Sabah, Jalan UMS, 88400 Kota Kinabalu, Sabah, Malaysia.
#Corresponding author. E-Mail: mazlina@ums.edu.my; Tel: +6088-320000; Fax: +6088-435324.

ABSTRACT Predicting the air quality is important particularly in the areas where air pollution is becoming a major health problem. This paper presents and evaluates the Nonlinear Autoregressive Exogenous (NARX) Neural Network performance in predicting the Air Pollutant Index (API) at three industrial areas in Malaysia: Pasir Gudang, Larkin and TTDI Jaya. The NARX was implemented in an open loop feed-forward architecture and was trained to produce an hour ahead API prediction based on the past values of air quality and meteorological parameters. Six air quality parameters: CO, NO_2, O_3, $PM_{2.5}$, PM_{10}, SO_2, and three meteorological parameters: wind direction, wind speed and ambient temperature were used as input while the API was set as the output. The prediction performance was measured by using the Coefficient of Determination (R^2) and Root Mean Square Error (RMSE) tests. Results show that the performance of NARX model was encouraging with R^2 value above 0.97 and RMSE value around 1.21 based on the data collected in 2018 at the three monitoring stations.

KEYWORDS: Nonlinear Autoregressive Exogenous Neural Network; Malaysia Air Pollution; Prediction, Time Series
Received 19 November 2020 **Revised** 26 January 2021 **Accepted** 29 March 2021 **Online** 2 November 2021

Original Article

INTRODUCTION

The air quality prediction system has become as important as weather forecast in Malaysia due to the recurring haze and the raising of air pollution problems in the recent years (Regencia, 2019; Sentian *et al*., 2019). The absence of such system had caused some disruptions in daily activity such as emergency school closure and event cancellation during hazy days (Lavallee, 2015; Mohamed Radhi, 2019). However, up until the present day, a system for air quality prediction has not been officially established by the Department of Environment (DOE) Malaysia, who is responsible for managing the air quality control in the country (Department of Environment, Ministry of Energy, Science, Technology, Environment & Climate Change, 2019).

Efforts were made by Malaysian scholars to improve the air quality management in the country. Rahman *et al.* (2015) and Zakaria *et al.* (2018) studied the air pollution trend in Klang Valley and Shah Alam to establish the correlation between air pollutants and the meteorological parameters. Leong *et al.* (2019) performed air quality prediction in Penang and Perak using Support Vector Machine (SVM) model. A research by Koo *et al.* (2020) analysed and concluded that the Fuzzy Time Series (FTS) model outperformed the other prediction models namely the Autoregressive Integrated Moving Average (ARIMA), Trend and Seasonality (TBATS), Artificial Neural Network (ANN), ARMA errors, Box-Cox Transformation and Trigonometric Regressors. Fong *et al.* (2018) predicted the PM_{10} concentration in Terengganu using Multiple Linear Regression (MLR) and Principal Component Regression (PCR) models during Southwest Monsoon (SWM) and Northeast Monsoon (NEM) seasons. The prediction performance showed that MLR outperformed PCR in both SWM and NEM seasons with the R^2 of 0.626 and 0.715 respectively. Azid *et al.* (2014) on the other hand attempted to develop the air quality time series prediction model using the Multilayer Perceptron Neural Network (MLP) model. The prediction performance of the MLP model was 0.615 (R^2) and 10.026 (RMSE) respectively.

Apart from the mentioned models, a model known as the Nonlinear Autoregressive Exogenous (NARX) neural network was published to be more superior (Wang & Bai, 2014; Lin *et al*., 2017). The NARX model was employed in Wuhan and Beijing where the prediction performance was excellent with R^2 of 0.9701. In both researches, the NARX was also proven to outperform the other prediction models such as SVM, Linear Regression (LR), ANN and ARIMA. The excellent performance of NARX model has motivated us to evaluate its performance in predicting the Malaysia API. The following sections describes the air quality data, the NARX model and the analyses that were conducted to verify the claim.

METHODOLOGY

Air Quality Data

Modelling the API time series using NARX utilizes the past values to predict the future value of the API. For this, six air quality parameters (Carbon Monoxide (CO), Nitrogen Dioxide (NO_2), Ozone (O_3), Sulphur Dioxide (SO_2), Particulate Matter 10 (PM_{10}), and Particulate Matter ($PM_{2.5}$)) and three meteorological parameters (Wind Speed (WS), Wind Direction (WD) and Temperature (T)) were used. These data were monitored and collected by a private agency appointed by the Malaysian DOE namely the Alam Sekitar Malaysia Sdn. Bhd. (ASMA). Three monitoring stations located in the Malaysian industrial zone: Larkin, Pasir Gudang and TTDI Jaya as shown in Figure 1 were selected among others in account of data availability.

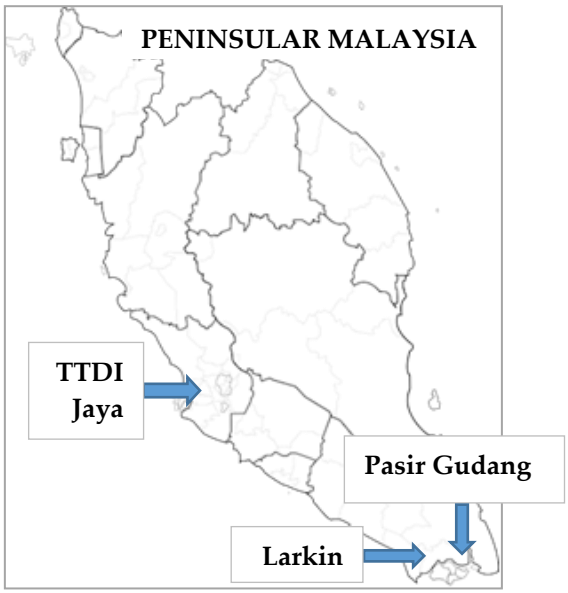

Figure 1. The monitoring station location map.

Pasir Gudang and Larkin industrial areas are located in Johor Bahru district, Johor. Pasir Gudang industrial area is mostly occupied by heavy industries such as shipbuilding, transportation and logistics, petrochemical and palm oil storage and distribution. The Larkin industrial area is occupied by manufacturers and factories from various kind of industry such as mechanical component manufacturer, electronic manufacturer, plastic fabrication factory, food product factories, glass manufacturer, metal fabrication factory and many others. Larkin is also surrounded by several other industrial areas such as Tebrau 1, Tampoi and Dewani. The TTDI Jaya is located nearby many industrial areas in Shah Alam, Selangor. Saujana Indah and Hicom-Glenmarie are the nearest industrial park to TTDI Jaya with the distance of within 2km radius away. Among of the occupiers in these industrial parks are foods, cosmetics and machinery products manufacturers. The existence of

various factories and plants in these areas contributed to the air pollution (Oliver *et al.*, 2014) and economic losses due to the haze related outpatients cases (Hanafi *et al.*, 2019).

Data Pre-processing

The hourly time step data for the year 2016 and 2018 were utilized in this research. The inclusion of $PM_{2.5}$ as a new API parameter in the middle of 2017 had caused data instability in which a fairly significant number of missing data-points was evidenced hence the 2017 data were excluded. For year 2016 and 2018, the missing data-points in percentage for each monitoring station were given in Table 1. To replace the missing data, the mean of nearby points imputation method was applied where the missing data-point was substituted with the mean of its two nearby data-points (Ali & Dacey, 2017).

Table 1. The percentage of missing data-points for each monitoring station.

Station	2016	2018
Pasir Gudang	9.6	2.3
Larkin	6.2	2.7
TTDI Jaya	11.3	3.3

The collected API data were also processed to remove the outliers using the Mahalanobis distance analysis described in Equation (1). The Mahalanobis distance analysis is a statistical method used to identify and remove multivariate data outliers hence improving the prediction accuracy (Leong *et al.*, 2019).

$$d = \sqrt{(x - \bar{x})^T \cdot C^{-1} \cdot (x - \bar{x})} \quad (1)$$

Based on Equation (1), the Mahalanobis distance is denoted as d, x is the row vector for each time step, $\bar{x}$ is the row vector of the parameters mean values and C^{-1} is the parameters inverse covariance matrix. A new column was added in the data to record the Mahalanobis distance for each timestep. Another column was added to calculate the p-value of the chi-square right tail distribution of the Mahalanobis distance with the degree of freedom equal to the number of parameters to predict API. The p-value is then compared to 0.001 (Tabachnick & Fidell, 2019). The p-value which is lower than 0.001 is considered to be an outlier. The time steps of the outliers were removed from the data. The percentage of outliers removed from the data in each monitoring station were recorded in Table 2.

Table 2. The percentage of outliers removed for each monitoring station.

Station	2016	2018
Pasir Gudang	1.9	3.2
Larkin	2.0	3.3
TTDI Jaya	1.3	2.3

The Nonlinear Autoregressive Exogenous Neural Network (NARX) API predictor

In this research, a purely feedforward two-layer network NARX architecture with a sigmoid transfer function hidden layer and a linear transfer function output layer was used. A purely feedforward network uses real output, y together with other parameters, x at time t-1 as its input to predict the API $\hat{y}$ at time t. The x parameters are the air quality parameters CO, NO_2, O_3, SO_2, PM_{10}, $PM_{2.5}$ (for 2018 data), and the meteorological parameters WS, WD and T. Figure 2 illustrates the input and output of the NARX model in purely feedforward architecture and Equation (2) represents the relation in more detail. In Equation (2), $\hat{y}(t)$ is the predicted API at time t, f is the mapping function approximated by the feedforward network, $x_1(t-1), x_2(t-1), \ldots . x_n(t-1)$ are the air quality and

meteorological parameters at time *t*-1 where *n* is the total number of parameters used, while $y(t-1)$ is the real API at time *t*-1.

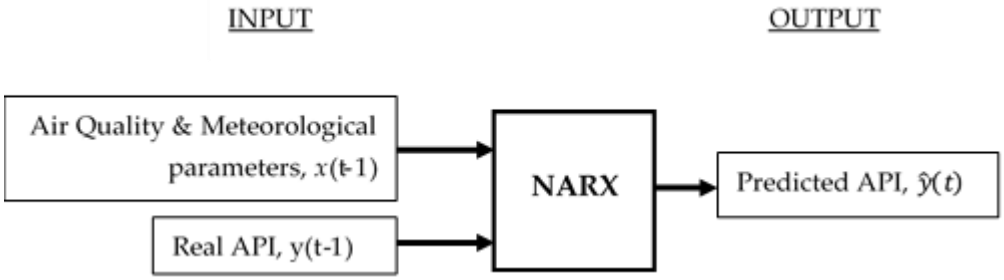

Figure 2. The NARX's input and output.

$$\hat{y}(t) = f\big(x_1(t-1), x_2(t-1), \ldots x_n(t-1), y(t-1)\big) \quad (2)$$

The air quality data were divided into three groups where 70% was allocated for training, 15% was allocated for validation and another 15% was reserved for testing. During training, the NARX was assigned a continuous stream of data that were arranged as in Equation (2). The Levenberg-Marquardt algorithm was employed as the learning algorithm and the prediction performance using different numbers of hidden neuron was analyzed. The prediction performance was analyzed by the coefficient of determination (R^2) and Root Mean Square Error (RMSE) values. The R^2 and RMSE are presented by Equation (3) and Equation (4), respectively.

$$R^2 = \left(\frac{1}{N}\frac{\sum_{t=1}^{N}(P_t-\bar{P})(T_t-\bar{T})}{\sigma_P \sigma_T}\right)^2 \quad (3)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(P_t - T_t)^2} \quad (4)$$

In Equations (3) and (4), P_t is the API predicted value, T_t is the API real value, $\bar{P}$ is the mean of API predicted values, $\bar{T}$ is the mean of API real values, N is the number of time steps of the API value, σ_T is the standard deviation of API real values and σ_P is the standard deviation of API predicted values. The R^2 value indicates the correlation between the real and the predicted API values. The value of R^2 is always between -1 to 1. The closer the value to 1 means the higher the correlation between the real and predicted API values. The RMSE on the contrary indicates the prediction errors standard deviation. The higher the value of RMSE means the higher the API prediction error.

RESULT AND DISCUSSION

The prediction performance of the NARX model measured in RMSE and R^2 values for different numbers of hidden neuron is given in Table 3. It can be seen that the number of hidden neuron has not given much impact on the prediction performance. This can be observed from the RMSE and R^2 values that vary slightly across the tested hidden neurons. For Pasir Gudang, the R^2 values vary from 0.9731 to 0.9874 for 2016 data and from 0.9877 to 0.9925 for 2018 data. For Larkin, the 2016 data have R^2 variation from 0.9722 to 0.9826 while the 2018 data have R^2 variation from 0.9818 to 0.9919. Similar observation was recorded for TTDI Jaya where the R^2 values vary from 0.8821 to 0.9025 for 2016 data and from 0.9517 to 0.9709 for 2018 data. These observations show that a single digit hidden neuron is adequate to produce a reliable and accurate one step ahead API prediction. Apart from that, an improved prediction performance was observed in 2018 data compared to 2016 data for all three monitoring stations. This indicates that a valid and continuous data are important to maintain high prediction accuracy. As can be seen in Table 1, the percentage of missing data is lesser in 2018 than in 2016 in which this contributes to a better prediction accuracy. Further evaluation shows that the highest prediction performance was achieved by the data with the lowest percentage of missing data

and vice versa. For instance, the 2018 Pasir Gudang data with 2.3% missing data-points scored 0.9925 R^2 value while the 2016 TTDI Jaya data with 11.3% missing data-points only scored R^2 of 0.9025.

Table 3. The NARX's prediction performance.

Hidden Neuron	Pasir Gudang				Larkin				TTDI Jaya			
	2016		2018		2016		2018		2016		2018	
	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
2	1.1723	0.9862	0.7312	0.9917	1.2667	0.9810	0.6389	0.9906	5.0634	0.8866	1.2542	0.9671
4	1.1949	0.9850	0.7897	0.9911	1.3669	0.9777	0.6030	0.9912	4.9831	0.8884	1.4811	0.9599
6	1.1285	0.9874	0.6582	0.9925	1.4074	0.9772	0.6450	0.9913	4.8219	0.8991	1.2005	0.9709
8	1.2488	0.9845	0.8673	0.9890	1.3025	0.9778	0.6307	0.9915	4.6080	0.9064	1.3144	0.9613
10	1.6220	0.9731	0.7051	0.9926	1.3053	0.9773	0.6033	0.9915	5.0867	0.8821	1.5614	0.9517
12	1.6100	0.9741	0.7754	0.9916	1.2661	0.9819	0.6410	0.9919	4.7808	0.8963	1.2608	0.9649
14	1.2911	0.9841	0.7957	0.9898	1.3480	0.9782	0.6576	0.9903	5.0108	0.8887	1.3931	0.9597
16	1.1044	0.9880	0.7210	0.9923	1.3531	0.9780	0.6347	0.9918	4.5634	0.9025	1.4426	0.9602
18	1.1940	0.9857	0.7764	0.9917	1.2799	0.9795	0.6625	0.9911	4.8189	0.8822	1.2205	0.9754
20	1.0610	0.9885	0.9025	0.9877	1.2725	0.9796	0.6497	0.9902	4.8954	0.8908	1.2353	0.9719
22	1.1589	0.9867	0.7318	0.9919	1.1995	0.9826	0.6571	0.9906	4.9586	0.8866	1.3918	0.9550
24	1.1796	0.9866	0.7572	0.9910	1.5101	0.9722	0.6114	0.9920	4.8547	0.8967	1.5683	0.9540
26	1.2058	0.9856	0.7410	0.9919	1.4031	0.9762	0.6692	0.9910	4.7830	0.8954	1.4162	0.9592
28	1.0609	0.9874	0.7839	0.9904	1.3866	0.9768	0.9522	0.9816	5.0662	0.8862	1.4771	0.9650
30	1.4939	0.9781	0.8975	0.9877	1.2392	0.9811	0.6962	0.9898	4.8849	0.8945	1.2573	0.9657

CONCLUSION

A NARX based API predictor was developed by utilizing the air quality data collected at three air quality monitoring stations in Malaysia and its one step ahead prediction performance was evaluated. Results show that the one step ahead prediction of NARX predictor was encouraging with R^2 value above 0.97 and RMSE value around 1.21 for the 2018 data. This performance is considered very good and outperformed the other predictors in the previous API prediction modelling attempts in Malaysia. So far, the NARX predictor was proven superior than the MLP, SVM, FTS and MLR predictors (Azid *et al.*, 2014; Leong *et al.*, 2019; Koo *et al.*, 2020; Fong *et al.*, 2018). Despite this, more analyses should be done to extend the prediction to multiple steps ahead. Besides, different approaches to replace the missing data-points should be explored to further improve the prediction performance as it was evidenced that the prediction accuracy was reduced when the number of missing data-points was high.

ACKNOWLEDGEMENTS

We thank Malaysian Department of Environmental (DOE) for providing the air quality data for the research.

REFERENCES

[1] Rahman, S. R. A., Ismail, S. N. S., Ramli, M. F., Latif, M. T., Abidin, E. Z. & Praveena, S. M. 2015. The assessment of ambient air pollution trend in Klang valley. *World Environment*, 5(1), 1–11.

[2] Ali, S. & Dacey, S. 2017. Technical Review: Performance of existing imputation methods for missing data in SVM Ensemble creation. *Journal of Data Mining & Knowledge Management Process (IJDKP)*, 7(6), 75–91.

[3] APIMS. 2020. *Air Pollutant Index of Malaysia* (http://apims.doe.gov.my/public_v2/home.html). Last accesssed on 5 March 2020.

[4] Azid, A., Juahir, H., Toriman, M. E., Kamarudin, M. K. A., Saudi, A. S. M., Hasnam, C. N. C. & Yamin, M. 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), Article ID 2063.

[5] Department of Environment, Ministry of Energy, Science, Technology, Environment & Climate Change. 2019. *Official Portal of Department of Environment* (https://www.doe.gov.my/portalv1/en/). Last accesssed on 21 December 2019.

[6] Fong, S. Y., Abdullah, S. & Ismail, M. 2018. Forecasting of particulate matter (PM10) concentration based on gaseous pollutants and meteorological factors for different monsoons of urban coastal area in Terengganu. *Journal of Sustainability Science and Management*, (5), 3–17.

[7] Hanafi, N. H., Hassim, M. H., Noor, Z. Z., Ten, J. Y., Aris, N. M. & Jalil, A. A. 2019. Economic losses due to health hazards caused by haze event in Johor Bahru, Malaysia. *E3S Web of Conferences*, 90, Article ID 01009.

[8] Koo, J. W., Wong, S. W., Selvachandran, G., Long, H. V. & Son, L. H. 2020. Prediction of air pollution index in Kuala Lumpur using fuzzy time series and statistical models. *Air Quality, Atmosphere and Health*, 13(1), 77–88.

[9] Lavallee, G. 2015. *Sports events in Singapore, Malaysia cancelled due to haze* (https://www.rappler.com/world/asia-pacific/sports-events-cancel-haze). Last accesssed on 4 November 2020.

[10] Leong, W. C., Kelani, R. O. & Ahmad, Z. 2019. Prediction of air pollution index (API) using support vector machine (SVM). *Journal of Environmental Chemical Engineering*, 103208.

[11] Lin, K., Jing, L., Wang, M., Qiu, M. & Ji, Z. 2017. A novel long-term air quality forecasting algorithm based on kNN and NARX. Proceedings of the *12th International Conference on Computer Science and Education (ICCSE)*. 22-25 August, 2017, Houston, USA. pp 343–348.

[12] Mohamed Radhi, N. A. 2019. 671 schools closed due to worsening haze. *New Straits Times (*https://www.nst.com.my/news/nation/2019/09/522165/671-schools-closed-due-worsening-haze). Last accesssed on 10 September 2020.

[13] Oliver, L. H. L., Mohamed Musthafa, S. N. A. & Mohamed, N. 2014. Air quality and land use in urban region of Petaling Jaya, Shah Alam and Klang, Malaysia. *Environment Asia*, 7(1), 134–144.

[14] Regencia, T. 2019. Haze blankets Kuala Lumpur, Singapore as fires rage in Indonesia. *Al Jazeera: Indonesia News* (https://www.aljazeera.com/news/ 2019/09/haze-blankets-kuala-lumpur-singapore-fires-rage-indonesia-190910030159845.html). Last accessed on 21 December 2019.

[15] Sentian, J., Herman, F., Yih, C. Y. & Hian Wui, J. C. 2019. Long-term air pollution trend analysis in Malaysia. *International Journal of Environmental Impacts: Management, Mitigation and Recovery*, 2(4), 309–324.

[16] Tabachnick, B. G. & Fidell, L. S. 2019. *Using Multivariate Statistics* (7th Edition). New York: Pearson.

[17] Wang, L. & Bai, Y. 2014. Research on prediction of air quality index based on NARX and SVM. *Applied Mechanics and Materials*, 602–605, 3580–3584.

[18] Zakaria, U. A., Saudi, A. S. M., Abu, I. F., Azid, A., Balakrishnan, A., Amin, N. A. & Rizman, Z. I. 2018. The assessment of ambient air pollution pattern in Shah Alam, Selangor, Malaysia. *Journal of Fundamental and Applied Sciences*, 9(4S), 772-788.