International Journal of Applied Engineering Research ISSN 0973-4562 Volume 15, Number 5 (2020) pp. 436-444

Air Quality Index Prediction in the Eastern Regions of Thailand with Accuracy of Neural Networks

Jatupat Mekparyup[1] and Kidakan Saithanu[1,*]

[1]*Department of Mathematics, Faculty of Science, Burapha University, 20131, Thailand.*

Abstract

Air pollution has turned to a critical environmental problem nowadays. Prediction of air quality then plays a significant role in notifying or warning people about and controlling air pollution in every countries including Thailand. Based on data measured by the eight monitoring stations located in Rayong, Chon Buri and Chachoengsao, the simple neural network design, Multi-Layer Perceptron or MLP, was built to appraise and predict the air quality index or AQI in the eastern regions of Thailand. The study results indicate that O_3 and PM_{10} respectively play the dominant role in AQI value while NO_2, SO_2 and CO together account for less than 2% importance. The period effected levels of AQI is classified into three groups. The low AQI is at the end of summer and in rainy season (April, June to September). The medium AQI is in summer and at the beginning of rainy season (February, March and May). The high AQI is in winter (October to January). Additionally, the obtained neural network models are able to rather perfectly predict and classify the AQI groups, as seeing of the accuracy of high percentage for correct classification rate or CCR with approximately 90% in training data set as well 88% in validation data set.

Keywords: Air Pollutant, Air Quality Index, Neural Networks, Multi-Layer Perceptron, Eastern Regions of Thailand

INTRODUCTION

Air pollution is a condition which hazardous or excessive quantities of substances are emitted into the air. Generally, air pollution is caused by natural source; dust, fume or gases from volcanic eruption, etc., and human activities which lead to harmful gas such as carbon monoxide released from vehicles, etc. Because of urbanization growth and industrial development, air pollution problem then becomes seriously environmental threats to human beings specifically health effects of respiratory disease. Exposure to outdoor air pollution brings about millions of people die in each year from diseases [1]. The objective of measuring air quality is to provide the information which is easy to understand and to be used for particular actions. The Air Quality Index or AQI can be the explanation. The scales of AQI are different from country to country because the air quality standards differ and organizations choose varying levels of categories, which presents an obstacle for comparison and diminishes its usability [2]. The Air Quality and Noise Management Bureau, Pollution Control Department is Thai government agency which proceeds and monitors air quality of any day based on the five main air pollutants; SO_2, NO_2, CO, O_3 and PM_{10}. The standard level of AQI specified by the Thai Environment Protection Department as the satisfactory air is 100. In 2010, $PM_{2.5}$ had just been auxiliary counted as the sixth principal air pollutant for calculating AQI yet it is collected by only a few monitoring stations in Thailand [3]. The eastern regions of Thailand compose of seven provinces; Chachoengsao, Chanthaburi, Chon Buri, Prachin Buri, Rayong, Sa Kaeo and Trat. Both Chon Buri and Rayong Provinces are key locations of many leading developer industrial estates. Chon Buri Province is the site of four big important industrial estates; Laem Chabang, Amata Nakorn, Hemaraj Chonburi and Pinthong. Map Ta Phut is a town in Rayong Province which is the location of Thailand's biggest industrial park, The Map Ta Phut Industrial Estate. Moreover, the eastern seaboard of Thailand usually known as the "Eastern Economic Corridor" or EEC is an advanced economic area which plays a major role in Thailand's economic covered three provinces; Chon Buri, Chachoengsao and Rayong. Therefore, these areas are the center of various industries such as petrochemicals industry and the automotive and electronics sectors. That is why population has grown rapidly and then air pollution problem has become much more increase at present according to the air quality report of [4] which informed Chon Buri is fourteen of the Southeast Asia's 15 cities with the highest $PM_{2.5}$. The basic monitoring stations not only cost expensive but also Chanthaburi and Trat have no monitoring stations thus creating a supplementary index tool other than the basic one is very beneficial. Solving problem of air pollution is a long-term process. The AQI prediction then can assist for planning or managing of air quality also preventing of damage caused by air pollution. Different techniques have been applied to forecast AQI, for example; [5] used stepwise of multiple regression to create the map of AQI model and [6] applied discriminant analysis to assess whether AQI exceeds the standard level or not. In addition, neural network is one of universal algorithms widely utilized for AQI classification since it is able to evaluate without any statistical assumptions, for example; [7], [8] and [9] used Multi-Layer

perceptron or MLP, [10] employed Radial Basis Function or RBF and [11] used Support Vector Regression. This study hence aims to present how well neural network models can efficiently allocate the air quality in to the correct group in the eastern areas of Thailand.

MATERIALS AND METHODS

The study was represented by measurements from eight monitoring stations in the eastern regions of Thailand; Ta Sit Subdistrict Administrative Organization, Pluak Daeng District (28T), Map Ta Phut Health Center, Mueang District (29T), Rayong Provincial Agricultural Extension Office, Mueang District (30T), Rayong Field Crops Research Center, Mueang District (31T), Thung Khru Subdistrict Region, Si Racha District (32T), Bo Win Subdistrict Region, Si Racha District (33T), Regional Environment Office 13, Ban Suan Subdistrict, Mueang District (34T) and Wang Yen Subdistrict Municipality, Plaeng Yao District (60T). The data of 1,077 observations was collected since 2008 to June, 2019 from the Air Quality and Noise Management Bureau, Pollution Control Department [12] in concentration terms of the monthly average of maximum emission for the five main air pollutants; SO_2, NO_2, CO, O_3 and PM_{10}. The value of Air Quality Index or AQI is subsequently reported for the air pollutant with the highest I_p value which is computed from Equation 1.

$$I_p = \left(\frac{I_{Hi} - I_{Lo}}{BP_{Hi} - BP_{Lo}} \right) \left(C_p - BP_{Lo} \right) + I_{Lo} \qquad (1)$$

where I_p is the index for pollutant p, C_p is the concentration of pollutant p, BP_{Hi} is the concentration breakpoint that is greater than or equal to C_p, BP_{Lo} is the concentration breakpoint that is less than or equal to C_p, I_{Hi} is the AQI value corresponding to BP_{Hi}, I_{Lo} is the AQI value corresponding to BP_{Lo}. The index is rounded to the nearest integer and the largest I_p for each pollutant is reported as AQI which is in the scale from 0 to 500 [13].

The steps of procedure for appraising and predicting AQI in the eastern regions of Thailand with accuracy of neural networks were as follows.

1. Identifying pattern or characteristic of the five main air pollutants for each monitoring station and overall with the basic descriptive statistics like mean, standard error (se.) of mean including Pearson correlation coefficients which examine the correlation tests between each pair of pollutants as well as the associations between AQI and each pollutant.
2. Identifying pattern or characteristic of AQI with line graph and pie chart of the monthly average AQI of each monitoring station and overall.
3. Classifying the five main air pollutants and the eight monitoring stations with cluster analysis.

   To join variables with similar characteristics into respective groups in a way that the degree of association between two variables is maximal if they belong to the same group and minimal otherwise, cluster of variables in cluster analysis is employed. Because of no prior information on the number of the particular patterns in this study, the agglomerative hierarchical technique was utilizable for both of clustering five main air pollutants and eight monitoring stations. Groups of variables are pictured from the individual entities by combining the smallest distance ($D = \{d_{ik}\}$) and joining the corresponding variables, say, U and V, to get cluster (UV). Then, cluster U and V are merged. The distance between (UV) and any other cluster W are computed as of Equation 2 [14].

   $$d_{(UV)W} = \min\{d_{UW}, d_{VW}\} \qquad (2)$$

   where d_{UW} and d_{VW} are the distances between the smallest distance of clusters U and W and clusters V and W, respectively.
4. Predicting AQI in the given localities through the utilization of neural network models, it meant to design and verify a classification model with the simple neural network architecture. Three following steps were realized.

   4.1 The data was separated into 2 sets for analysis. The training data set was first randomly selected 70% of all data using for model training (754 observations). The remainder called validation data set (323 observations) was counted for suitability of model validation.

   4.2 Building the simple neural network models applied for classification of air quality. The Multi-Layer perceptron or MLP was then trained with two architectures of 3 or 5 hidden nodes in a hidden layer following a suggestion of [15], [16]. The class of air quality was simply classified only 3 groups because the AQI value in the eastern regions of Thailand had not ever beyond 200 in the past. For this reason, the number of output nodes in an output layer was 3 which were consistent with the three predefined groups of air quality. The first group was good air quality counting on the AQI ranged 0-50. The second group was moderate air quality counting on the AQI ranged 51-100. Otherwise, the AQI ranged 101-200 was distributed to the group of unhealthy for sensitive air quality. The number of input nodes was 5 which were equal to the five main air pollutants.

4.3 Measuring the accuracy of neural network models with a correct classification rate (CCR), one of popular criterions applied in classification problem [17], [18], which is calculated as of Equation 3.

$$CCR = \frac{\sum_{k=0}^{C-1} CC_k}{n} \quad (3)$$

where CC_k is the number of correctly designated observations and n is the number of observations in the considered group.

RESULTS

The results of study are particularly described in this section.

1. The basic descriptive statistics of monitored pollutants for each monitoring station and overall are presented in Table 1. All pollutants do not exceed the average based standard but each station also overall face moderate and rather high levels of O_3 and PM_{10}. The highest levels of O_3 and PM_{10} are respectively shown in station 34T and 29T.

Table 1: Basic descriptive statistics (mean ± se. of mean) of monitored pollutants

Station	Monitored pollutants				
	SO_2	NO_2	CO	O_3	PM_{10}
28T	13.49±1.00	28.28±1.06	1.36±0.06	74.72±2.26	65.44±2.18
29T	57.92±1.83	44.07±1.46	1.73±0.07	77.71±2.58	73.54±2.79
30T	15.96±0.88	47.53±1.63	2.06±0.05	79.01±2.91	56.85±2.14
31T	28.59±2.31	38.16±1.26	1.08±0.04	83.27±2.80	62.44±2.43
32T	32.06±1.54	45.09±1.59	1.57±0.11	76.03±2.34	65.49±2.29
33T	19.60±1.10	43.28±1.54	2.00±0.08	87.14±2.67	59.13±2.94
34T	13.49±0.57	53.30±1.82	1.73±0.06	93.26±2.75	41.11±2.26
60T	29.92±2.25	33.82±1.68	1.28±0.06	86.93±1.81	54.59±1.98
Overall	27.67±0.76	41.69±0.58	1.61±0.03	82.18±0.92	59.66±0.89
Average based standard	300 1 hour (ppb.)	170 1 hour (ppb.)	30 1 hour (ppm.)	100 1 hour (ppb.)	120 24 hour ($\mu g/m^3$)

Table 2 explains the correlation tests with Pearson correlation coefficients and corresponding p- values (number in parenthesis). There are no relationships between 3 pairs of monitored pollutants; SO_2 with NO_2, CO and O_3. Also, there is no association between AQI and SO_2 while the highest Pearson correlation coefficient (0.953) is found between AQI and O_3.

Table 2: Pearson correlation coefficients and corresponding p-values among the five main air pollutants and AQI

	SO_2	NO_2	CO	O_3	PM_{10}
NO_2	0.060 (0.102)				
CO	-0.013 (0.405)	0.721 (0.000)			
O_3	-0.016 (0.651)	0.476 (0.000)	0.236 (0.000)		
PM_{10}	0.264 (0.000)	0.289 (0.000)	0.133 (0.000)	0.363 (0.000)	
AQI	0.041 (0.255)	0.476 (0.000)	0.241 (0.000)	0.953 (0.000)	0.480 (0.000)

2. The monthly average AQI of each monitoring station and overall are separately illustrative in Figure 1 and Figure 2. The monthly averages AQI of station 28T and 60T are lower when compare with those of the other six stations in January. The station 60T and 34T depict the highest value of monthly averages AQI during May to September and October to December, respectively.

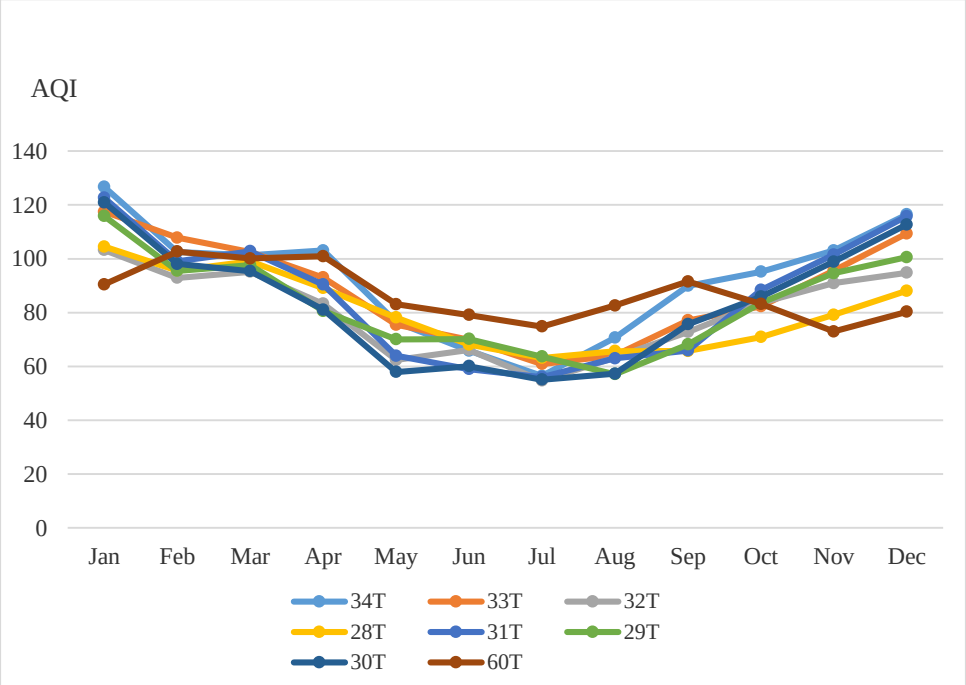

Figure 1: Line graph of monthly average AQI of each monitoring station

For overall, it can say that the monthly averages AQI are high during October to January, moderate during February to April and then low during May to September.

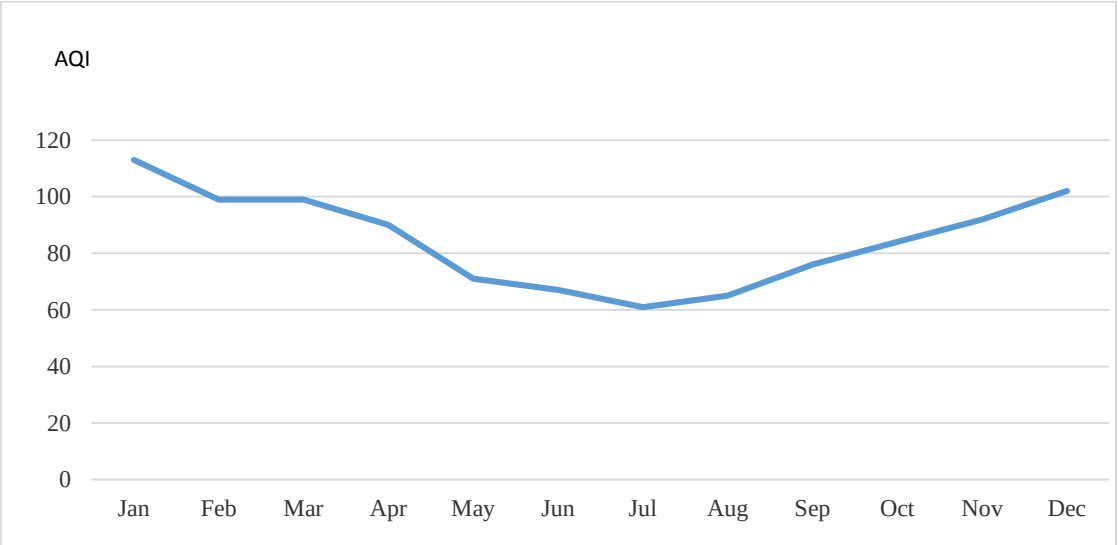

Figure 2: Line graph of monthly average AQI of overall

Figure 3 indicates percentage of pollutants which impact on AQI for each monitoring station and overall. O_3 and PM_{10} dominate AQI in 3 stations; 28T, 34T and 60T. There are 2 stations; 30T and 33T, which directed with O_3, PM_{10} and NO_2 while the two stations, 31T and 32T, are controlled with O_3, PM_{10} and SO_2. Only the station 29T displays 4 pollutants; O_3, PM_{10}, SO_2 and CO, influenced to AQI. For overall, O_3 and PM_{10} respectively play the

dominant role in AQI value while NO_2, SO_2 and CO together account for less than 2% importance.

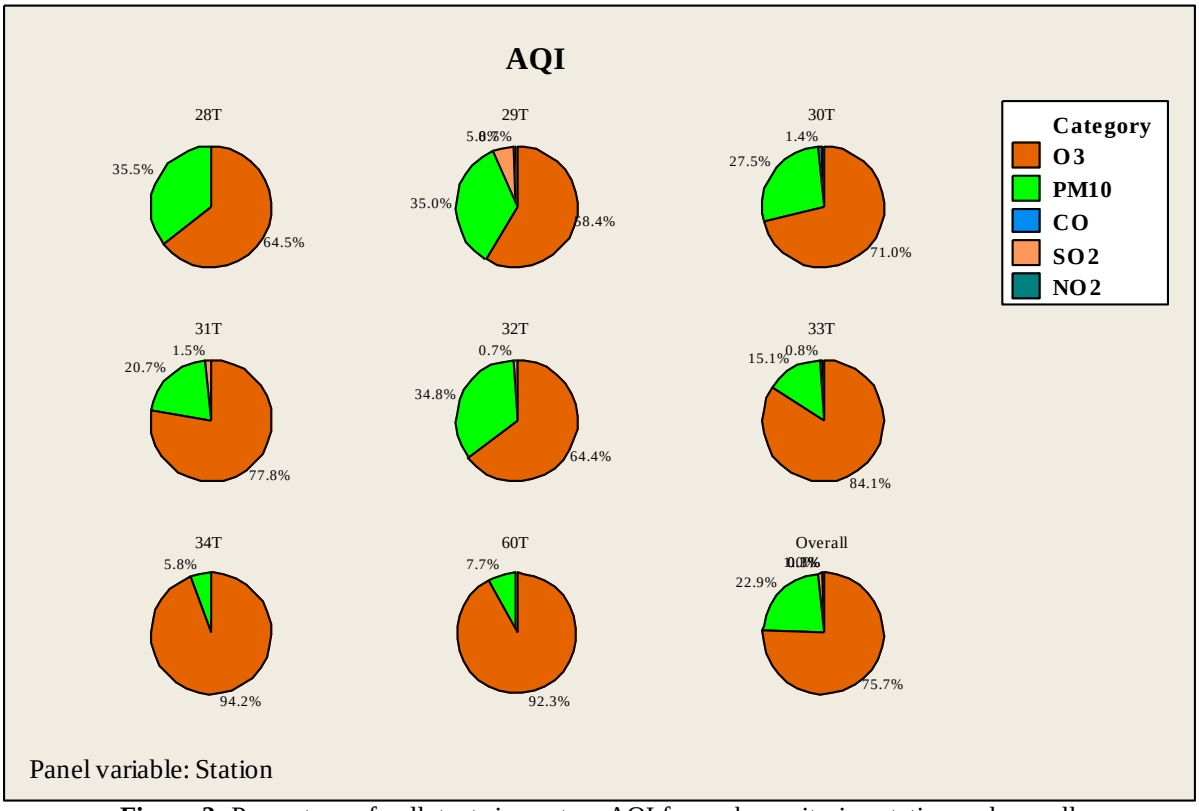

Figure 3: Percentage of pollutants impact on AQI for each monitoring station and overall

3. Based on the agglomerative hierarchical technique of cluster analysis, two dendograms are succeeding pictured for classifying air pollutants and monitoring stations in Figure 4 and Figure 5. Figure 4 demonstrates three pollutants; NO_2, O_3 and PM_{10}, were merged in the same group while CO and SO_2 were exclusively placed to some other groups.

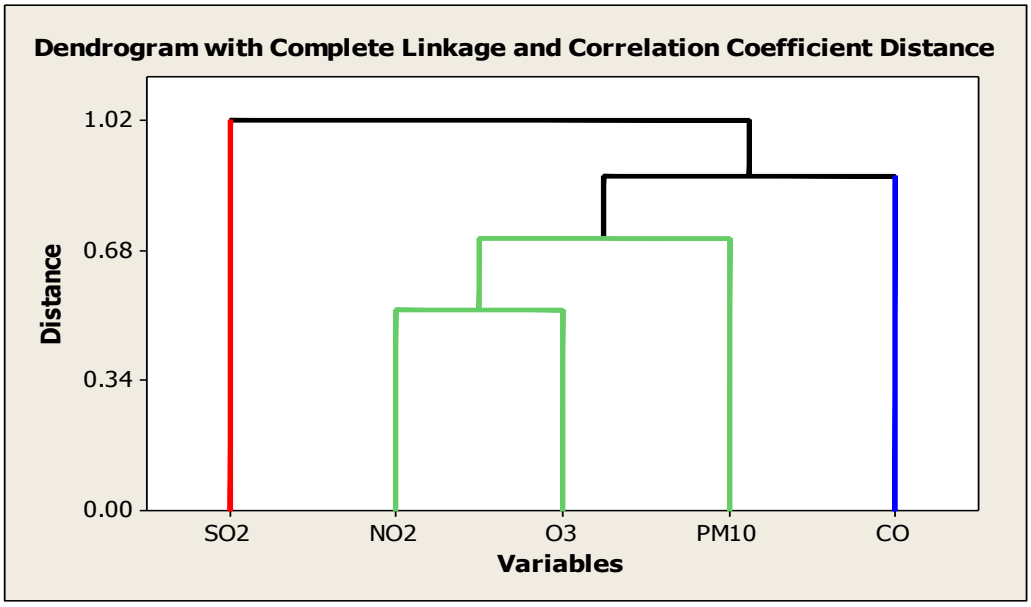

Figure 4: Dendogram of clustering the five main air pollutants

Figure 5 provides three different classes of monitoring stations were grouped by the AQI value. Cluster 1 contained 3 stations; 32T, 33T and 34T, with low AQI values are in Chon Buri Province. Cluster 2 held 4 stations; 28T, 29T, 30T and 31T, with moderate AQI values are in Rayong Province. Finally, only the station 60T with high AQI value is in Chachoengsao Province.

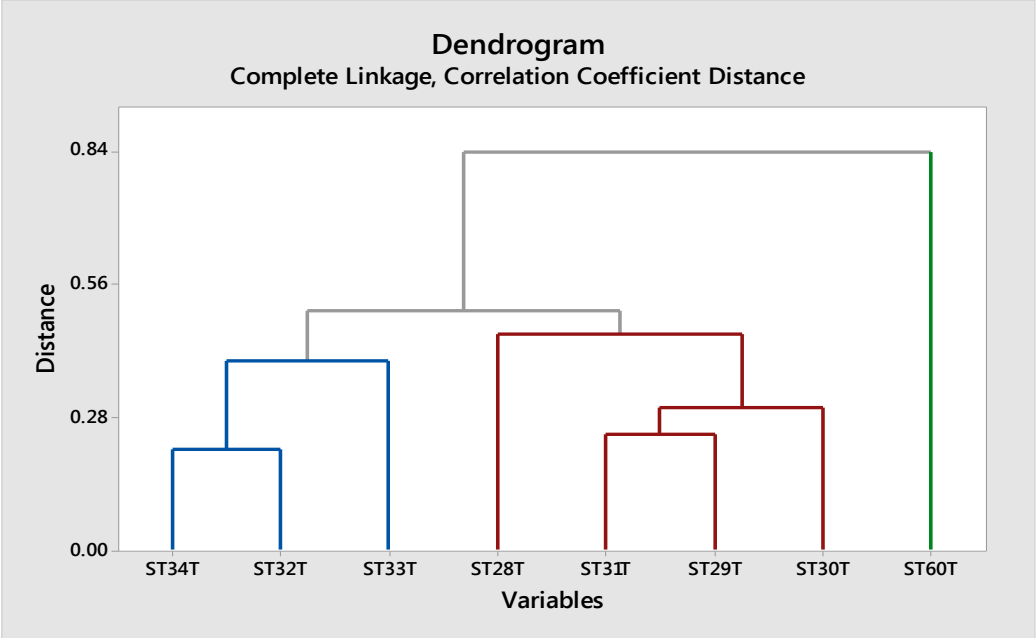

Figure 5: Dendogram of clustering the eight monitoring stations

Moreover, Figure 6 is dendogram utilized for explaining whether there is a seasonal impact on AQI or not. For merging months based on AQI, it exposed that there were three distinct clusters relating to a season. Cluster 1 is at the end of summer and in rainy season; April, June to September, denoted the low AQI. Cluster 2 is in summer and beginning of rainy season; February, March and May, presented the medium AQI. Cluster 3 is in winter; October to January, stand for the high AQI.

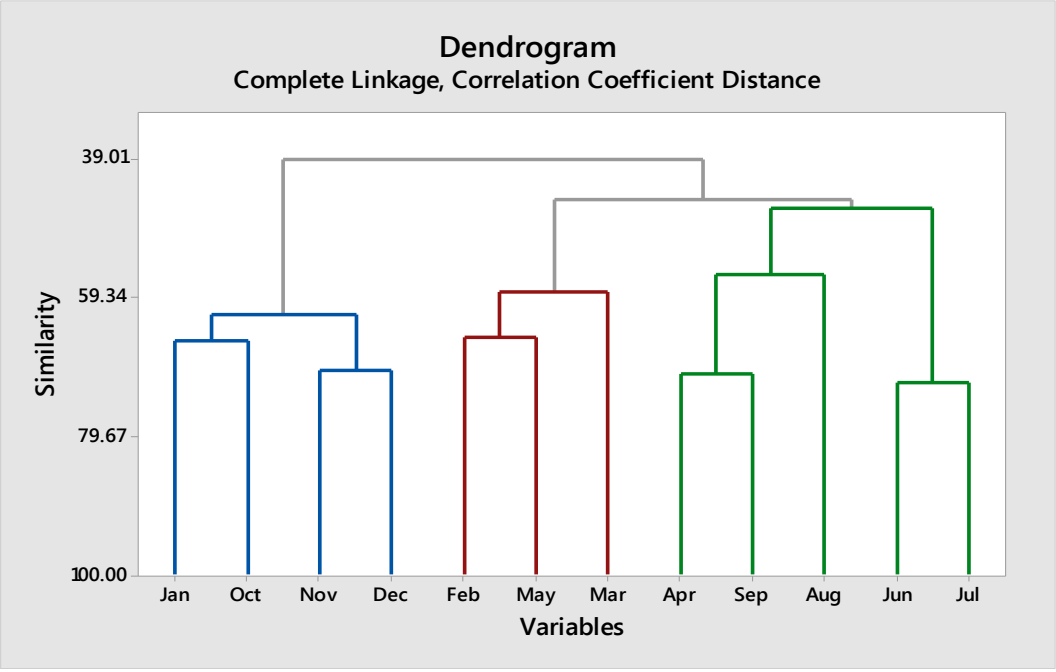

Figure 6: Dendogram of clustering the months

4. Once the MLP networks were built with 5 input nodes, 3 or 5 hidden nodes and 3 output nodes. The two simple architectures therefore investigated for this study are MLP 5-3-3 in Figure 7 and MLP MLP5-5-3 in Figure 8.

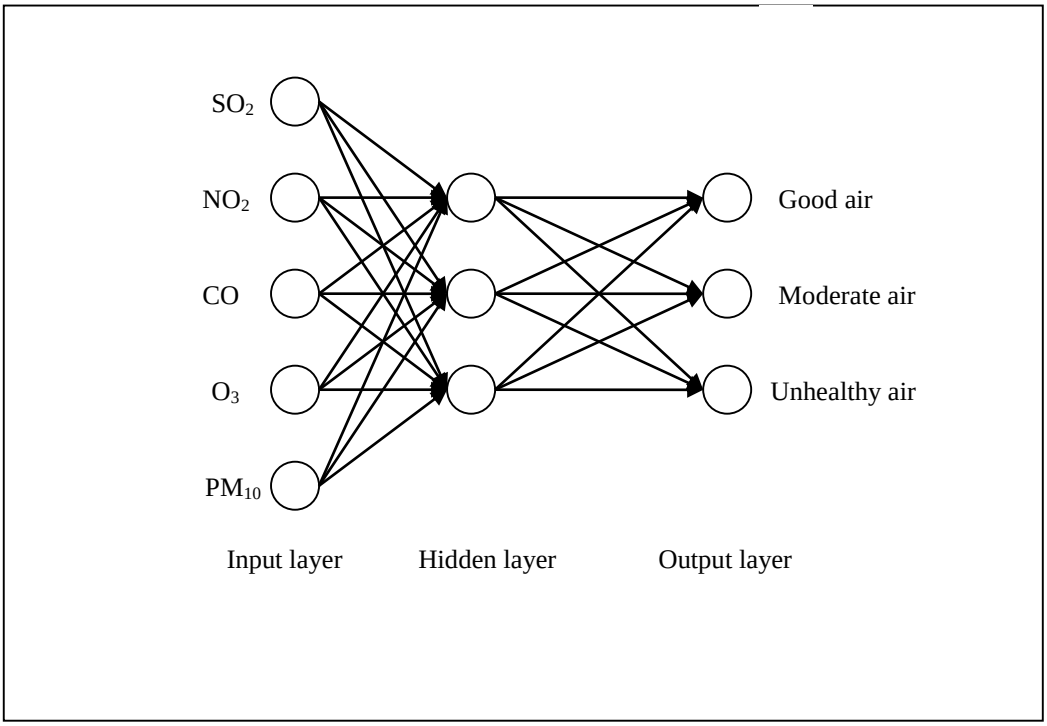

Figure 7: MLP5-3-3

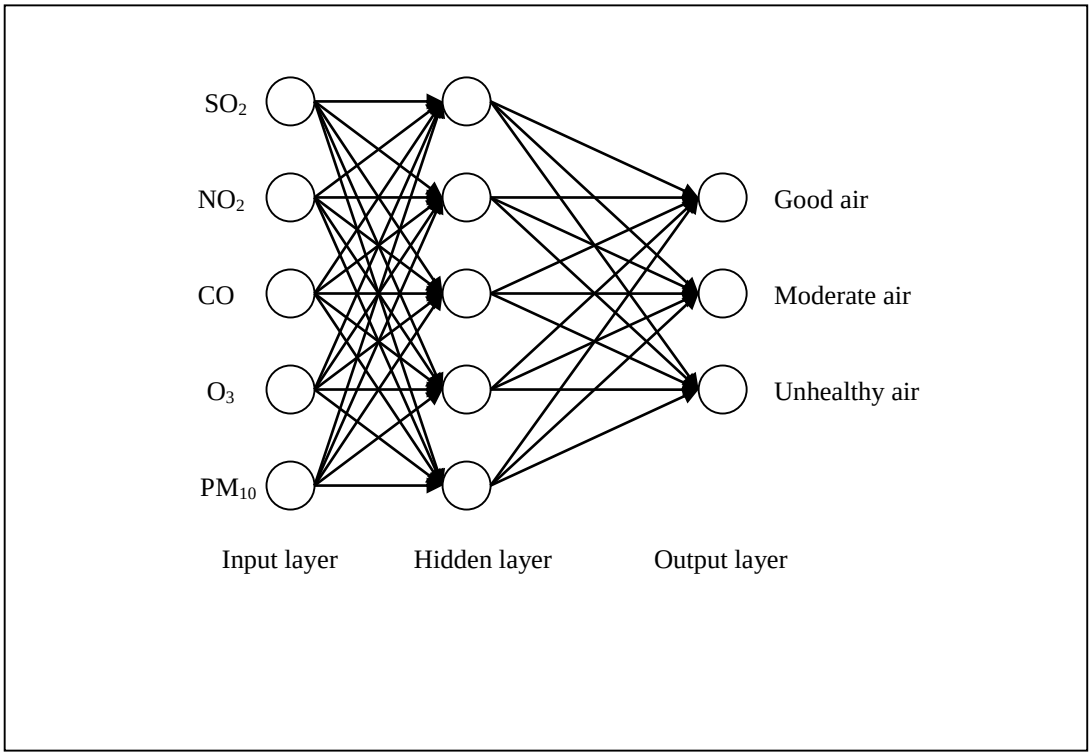

Figure 8: MLP5-5-3

After training the two MLP networks, the classification accuracy of the prediction models is displayed in Table 3. It indicates that both MLP models effectively predict and allocate AQI in the right groups, as seeing of the accuracy of high percentage of CCR with approximately 90% in training data set as well 88% in validation data set.

Table 3: Comparison of AQI classification between MLP5-3-3 and MLP5-5-3

Model	Data Set	AQI Group	Into AQI Group			
			Good	Moderate	Unhealthy	CCR (%)
MLP5-3-3	Training	Good	40	9	0	
		Moderate	2	514	0	
		Unhealthy	0	65	124	
		CCR (%)	81.63	99.61	65.61	89.92
	Validation	Good	23	3	0	
		Moderate	0	196	0	
		Unhealthy	0	33	68	
		CCR (%)	88.46	100.00	67.33	88.85
MLP5-5-3	Training	Good	39	10	0	
		Moderate	4	512	0	
		Unhealthy	0	60	129	
		CCR (%)	79.59	99.22	68.25	90.19
	Validation	Good	21	5	0	
		Moderate	1	195	0	
		Unhealthy	0	32	69	
		CCR (%)	80.77	99.49	68.32	88.24

CONCLUSIONS AND DISCUSSIONS

The studies are concluded and discussed as follows.

1. To identify pattern or characteristic of the five main air pollutants: Cluster analysis capably categorizes three pollutants; NO_2, O_3 and PM_{10}, in the same group corresponding to the result of correlation tests among these pollutants.
2. To identifying pattern or characteristic of AQI: The monthly average of AQI are low at the end of summer and in rainy season (April, June to September), medium in summer and beginning of rainy season (February, March and May), and high in winter (October to January) in accordance with clustering the months to explain whether or not there is a seasonal effect on AQI. Additionally, O_3 and PM_{10} are two key pollutants influenced to AQI as seeing of the high Pearson correlation coefficients between AQI and O_3 (0.953) also between AQI and PM_{10} (0.480)
3. To predict AQI with MLPs: Neural network models can be potentially applied in classification of air quality as seeing of the average CCR greater than 88% for both of MLP5-3-3 and MLP5-5-3 models. In addition, these two neural network architectures do fairly allocate for classification in the unhealthy group because it might be the complicated case. A better performance will be then performed with the advanced architecture like the RBF or adjusting some parameters of neural network, for examples, with more hidden layers or more hidden nodes in a hidden layer.

ACKNOWLEDGEMENT

This work was financially supported by the Research Grant of Burapha University (Grant no. 005/2562). The authors also were grateful to the Air Quality and Noise Management Bureau, Pollution Control Department for kindly supporting all data.

REFERENCES

[1] Zhang, Q., Jiang, X., Tong, D., Davis, S. J., Zhao, H., Geng, G., Feng, T., Zhang, B., Lu, Z., Streets, D. G., 2017, "Transboundary health impacts of transported global air pollution and international trade," Nature, 543, pp. 705-709.

[2] Lemes, S., 2018, "Air Quality Index (AQI) – Comparative Study and Assessment of An Appropriate Model for B&H," 12th Scientific/Research Symposium with International Participation.

[3] Pollution Control Department, 2019, "Situation and management in air and noise pollution in Thailand 2019," Retrieved from http://www.pcd.go.th/public/publications /print_report.cfm?task=air_noise61

[4] Greenpeace, 2019, W2019 World air quality reports," Retrieved from https://www.greenpeace.or.th/s/right-to-clean-air/2018-world-air-quality-report.pdf

[5] Intarat, T., 2011, "Geoinformatics Application on Air Quality Assessment: A Case study in Chonburi Province," Burapha Sci. J., **16**(1), pp. 32-40.

[6] Saithanu, K. and Mekparyup, J., 2014, "Air Quality Assessment in the Urban Areas with Multivariate Statistical Analysis at the East of Thailand," International Journal of Pure and Applied Mathematics, 91(2), pp 169-177.

[7] Wang, K., Wang, W. S., Zhang, X., and Sun, L. X., 2010, "Evaluation of forecast air quality based on BP neural network," J. Harbin Inst. Technol, 42, pp. 1278-1281.

[8] Durao, R., and Pereira, M. J., 2012, "MLP based models to predict PM_{10} and O_3 concentrations, in Sines industrial area," Geophysical Research Abstract, 14, EGU2012-13448.

[9] Kyriakidis, L., Karatzas, K., Kukkonen, J., Papadourakis, G., and Ware, A., 2013, "Evaluation and analysis of artificial neural networks and decision trees in forecasting of common air quality index in Thessaloniki, Greece," Eng. Intell. Syst., 21, pp. 111-124.

[10] Hajek, P., and Olej, V., 2015, "Predicting common air quality index – The case of Czech microregions," Aerosol and Air Quality Research, 15, pp. 544-555.

[11] Liu, H., Li, Q., Yu, D., and Gu, Y., 2019, "Air quality index and air pollutant concentration prediction based on machine learning algorithms," Applied Science, 9(4069), pp. 1-9.

[12] Pollution Control Department, 2019, "Air Pollution Reports in Thailand 2019," Retrieved from http://www.pcd.go.th

[13] U.S. Environmental Protection Agency, Office of Air Quality Planning and Standards Research Triangle Park, North Carolina 27711. (2006). Guideline for Reporting of Daily Air Quality – Air Quality Index (AQI). Retrieved October 23. 2012, from http://www.epa.gov/ttn/oarpg/t1/memoranda/rg701.pdf

[14] Johnson, R. A., and Wichern, D. W., 2007, "Applied multivariate statistical analysis. 5th ed.," Prentice-Hall Press, New Jersey.

[15] Cybenko, G., 1989, "Approximation by superpositions of a sigmoidal function," Mathematics of Control, Signals, and Systerms, 2(4), pp. 303-314.

[16] Hornik, K., Stinchcombe, M., and White H., 1989, "Multilayer feed forward networks are universal Approximators," Neural Networks, 2(5), pp. 359-366.

[17] El-Sebakhy, E. A., Hadi, A. S., and Faisal, K. A., 2007, "Iterative Least Squares Functional Networks Classifier," IEEE Transactions on Neural Networks, 18(3), pp. 844-850.

[18] Oh, C., and Ritchie, S. G., 2007, "Recognizing vehicle classification information from blade sensor signature," Pattern Recognition Letters, 28(9), pp. 1041-1049.