VNU Journal of Science: Earth and Environmental Sciences

Journal homepage: https://js.vnu.edu.vn/EES

Original Article

Forecast of Hourly Tropospheric Ozone Concentration in Quang Ninh using MLP and SVM

Nguyen Thi Thu Phuong[1,2], Mac Duy Hung[1,2,*], Duong Thanh Nam[3], Nghiem Trung Dung[1]

[1]*Hanoi University of Science and Technology, 1 Dai Co Viet, Hanoi, Vietnam*

[2]*Thai Nguyen University of Technology, 666, 3/2 street, Thai Nguyen, Vietnam*

[3]*Center for Research and Technology Transfer, Vietnam Academy of Science and Technology, 18 Hoang Quoc Viet, Hanoi, Vietnam*

Received 06 April 2020
Revised 15 July 2020; Accepted 27 July 2020

Abstract: Support vector machine (SVM) and multilayer perceptron (MLP) were used to forecast hourly tropospheric ozone concentration at three locations of Quang Ninh, namely Cao Xanh, Uong Bi and Phuong Nam. Data used to train the models are the hourly concentrations of gaseous pollutants (O_3, NO, NO_2, CO) and meteorological parameters including wind direction, wind speed, temperature, atmospheric pressure, relative humidity measured in the 2016. Both models accurately forecast tropospheric ozone levels compared to the observation data. The correlation coefficients (r) of the models applied for the three locations range from 0.85 to 0.91. In addition, SVM exhibits a more accurate prediction than MLP, especially for those with large variations, i.e. high standard deviations.

Keywords: Tropospheric ozone, SVM, MLP, machine learning, Quang Ninh.

* Corresponding author.

E-mail address: mduyhung@gmail.com

https://doi.org/10.25073/2588-1094/vnuees.4604

1. Introduction

Ozone is found primarily in two layers of the atmosphere: the stratosphere and the troposphere. Ozone in the troposphere is called tropospheric ozone or ground level ozone. Ozone in the stratosphere shields to protect Earth's surface from the sun's harmful ultraviolet radiation. Conversely, tropospheric ozone can be harmful to human and the ecosystem [1-3].

The majority of tropospheric ozone formation is occurred when ozone precursors such as nitrogen oxides (NO_x), carbon monoxide (CO) and volatile organic compounds (VOCs) react in the atmosphere in the presence of sunlight [1, 3]. If acute ozone exposure ranges from hours to a few days, it directly affects the lungs and the entire respiratory system. By the negative impacts on human health, ecosystem and climate, it is necessary to provide with information on the variation of tropospheric ozone to the community as well as to forecast tropospheric ozone concentration [3].

This issue engages environmental modelers in the development of forecasting models. More and more techniques have been being used to forecast air quality, of which the most widely used method is machine learning, and of course, the forecast of tropospheric ozone levels has made great success [4]. This method can quickly process big data and through forecasting algorithms, the results are delivered faster and more accurately. In particular, the greater the amount of training data, the more accurate the forecast results. This is especially important in air quality management, typically to predict pollutants that are highly toxic for human [1-4].

Techniques used to predict tropospheric ozone concentration are the decision tree algorithm (CART, M5), regression algorithm (LR), bagging, especially, support vector machine (SVM), the multilayer perceptron (MLP). In which, the last two techniques are popular learning machines in present [4 - 7]. Forecasting results depend on many factors such as precursors, meteorological conditions, advantages and disadvantages of each method such as inherent local minima, "black-box" property and over-fitting, parameters identification [5]. Studies on the forecast of tropospheric ozone in Vietnam using artificial intelligence have been initiated; however, they are often focused on big cities like Hanoi, Can Tho, Ho Chi Minh City [8, 9, 10]. In Vietnam, most prediction of tropospheric ozone uses photochemical models and the use of machine learning to predict this pollutant is quite new [8, 9, 10, 11]. Moreover, there are few studies using SVM and MLP algorithms to predict tropospheric ozone. Therefore, this study is aimed to apply machine learning to predict tropospheric ozone in mountain/remote areas for air quality management. This study used SVM and MLP to predict tropospheric ozone in Quang Ninh, Vietnam.

2. Methods

2.1. Site Characterization and Data

The study was conducted based air quality monitoring data of one year, from January 1st, 2016 to December 31st, 2016, at three monitoring stations of Quang Ninh, Vietnam, namely Cao Xanh, Uong Bi and Phuong Nam. Data used are hourly concentrations of tropospheric ozone and other gaseous pollutants (NO, NO_2, CO); and meteorological parameters (wind direction, wind speed, temperature, air pressure, humidity), which were monitored at these stations. The data were processed by excel and Rstudio and then, divided into two subsets, in which one would be used for training and the other would be for testing. The training dataset is the data from January 2016 to August 2016; the testing dataset is the data from September 2016 to December 2016.The research process is shown in Figure 1.

2.2. Data Processing

Raw data were processed before being used for training and testing by MLP and SVM algorithm. Firstly, any data point in the dataset

having its value ≤ 0 is detected and removed to make a data gap. Secondly, abnormal values (outliers) are also detected by Box and Whisker method (IQR method-Interquartile and removed to create data gaps.

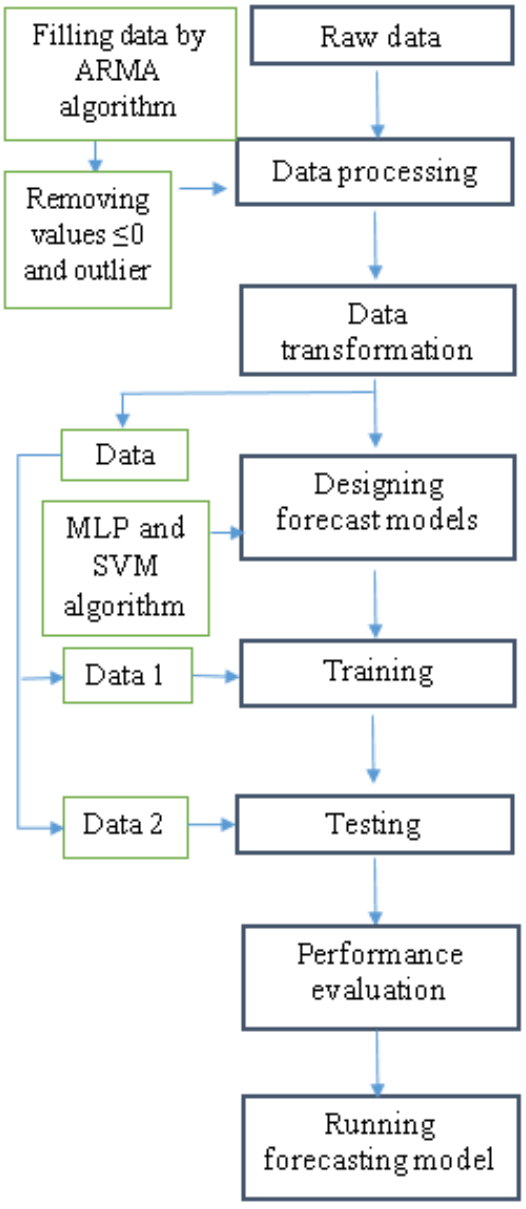


Figure. 1. Research process.

Raw data were processed before being used for training and testing by MLP and SVM algorithm. Firstly, any data point in the dataset having its value ≤ 0 is detected and removed to make a data gap. Secondly, abnormal values (outliers) are also detected by Box and Whisker method (IQR method-Interquartile and removed) to create data gaps. This method divides a data set into quartiles. The values that divide each part are called the first (Q1), second (Q2), and third (Q3) quartiles. Then, IQR=Q3-Q1 and the values beyond marginal values (Q1 - 1.5*IQR or Q3 +1.5*IQR) can be outliers. Finally, these data gaps are filled up by Autoregressive Moving Average algorithm (ARMA) in forecast package in Rstudio software.

George Box and Gwilym Jenkins (1976) studied ARMA model to apply to the analysis and prediction of time series. This method is also called Box-Jenkins method, which consists of four steps: identifying test models, estimating, verifying and predicting tests. This method is a combination of moving average and autoregressive process, this model can be understood by the following equation [12]:

AR:$x_t = \alpha_1 x_{t-1} + \ldots + \alpha_p x_{t-p} + z_t$;

MA: $x_t = \beta_0 z_t + \beta_1 x_{t-1} + \ldots + \beta_q x_{t-q}$

And ARMA model:

$$x_t = \alpha_1 x_{t-1} + \ldots + \alpha_p x_{t-p} + z_t + \beta_1 x_{t-1} + \ldots + \beta_q x_{t-q}$$

(2-1) Where $\alpha_1, \ldots, \alpha_p$ and $\beta_1, \ldots, \beta_p$ are corresponding coefficients.

2.3. Data transformation

Raw data were transformed to eliminate the disruption of the wind direction angle (WD) at 360°, the wind direction index (WDI) is used to denote the wind direction, calculated using the following equation:

$$WDI = 1 + \sin(WD + \pi / 4) \quad (2\text{-}2)$$ [1]

where WD is the wind direction (with 0° corresponding to the north). Therefore, WDI has a minimum of 0.07 for the south wind (180°) and a maximum of 1.96 when the WD is 315°.

2.4. Forecasting models

MLP and SVM were used in this study, with the dataset divided as data 1 with 75% (6567 lines) for training and data 2 with 25% (2189 lines) for testing.

Support vector machine (SVM)

Support vector machine (SVM) has been proposed by V. N. Vapnik for data classification. SVM creates a hyperplane in multidimensional space, related to classification and regression algorithms [2].

The function can be presented as the following equation:

$$\hat{y} = \hat{F}(x) = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) K(x, z_i) + b \quad (2\text{-}3)$$

where, and α and α* are Lagrangian parameters; K (x, zi) is called kernel function. In this study, the number of input variables are nine with two hidden layers and having five neural in each layer and training epochs are 4000.

Multilayer Perceptron (MLP)

MLP is one of the neural network architectures with three layers of neurons: input layer, hidden layer and output layer. Each neuron in the layer links with all neurons in the previous layer. The output of the previous layer neuron is the input of the neuron in the next layer [3]. Each layer uses a linear combination function. These networks create models and connect the input with the output using historical data. The MLP algorithm performs the following form [3]:

f: X⊂Rd → Y⊂Rc

$$f(x) = \sum_{j=0}^{h} c_j \psi(w_j^T x + w_{jo}) + c_0 \quad (2\text{-}4)$$

In which: $\psi(w_j^T x + w_{jo})\psi$ is the activation function of the hidden neuron layer; w_j^T is the parameter vector of separate neurons; w_{jo} is a threshold value; c_j is the weight vector of the nerve cell and c_{j0} is the threshold value. In this study, important setting parameter is epsilon with the range from 0 to 0.2 and the step change is 0.01.

Performance evaluation

The performance of the models was assessed based on statistical indicators including average absolute error (MAE), mean square error (RMSE), and correlation coefficient (r) [4]. MAE and RMSE measure residual errors, which give a global idea of the difference between the observed and forecasted values. The lower the values of MAE and RMSE indicate that the model is better. They are calculated as follows:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} \left|\hat{Y}_i - Y_i\right| \quad (2\text{-}5)$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \left(\hat{Y}_i - Y_i\right)^2} \quad (2\text{-}6)$$

Y_t is the true target metric value for observation i, Y_i is the target metric value for observation i as predicted by the model, and n is the number of data.

- Pearson correlation coefficient (r)

$$r = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y}_i)^2 - \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y}_i)^2}} \quad (2\text{-}7)$$

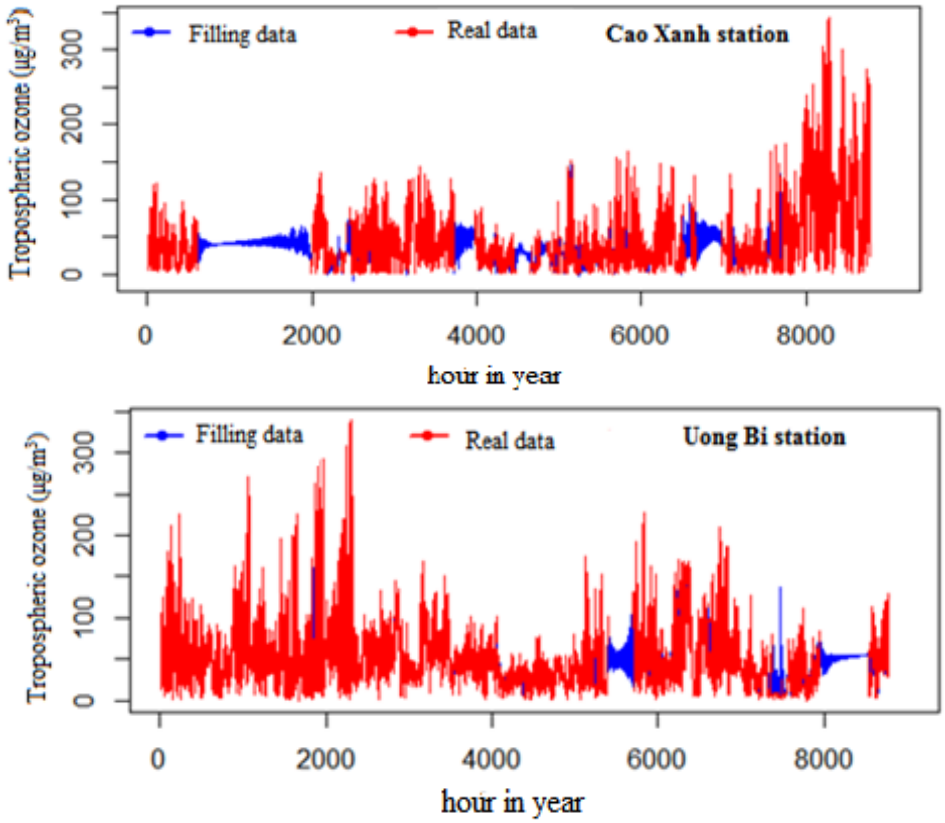
3. Results and Discussion

3.1. Filling up the Missing Data Using ARMA Algorithm

The dataset is processed to remove zero values, negative values and outliers to make data gaps (blank data).The summary of data on tropospheric ozone, precursors and meteorological parameters after removing these values (but before filling up)in the three stations is shown in Table 1.

Table 1. Summaryof data at three stations before filling up

Parameters	Temperature	Humidity	Wind speed	Wind direction	Solar Radiation	O_3	CO	NO	NO_2
Uong Bi station									
Existing number of data points	8363	8364	8364	8364	8364	8364	7439	5960	6822
Missing number of data points	423	422	422	422	422	422	1347	2826	1964
Missing rate (%)	4.8	4.8	4.8	4.8	4.8	4.8	15.3	32.2	22.4
Cao Xanh station									
Existing number of data points	6590	6590	6590	6590	8308	6785	6101	5943	5565
Missing number of data points	2196	2196	2196	2196	478	2001	2685	2843	3221
Missing rate (%)	25	25	25	25	5.4	22.8	30.6	32.4	36.7
Phuong Nam station									
Existing number of data points	7934	7934	7935	7935	7935	7935	7134	5750	7406
Missing number of data points	852	852	851	851	851	851	1652	3036	1380
Missing rate (%)	9.7	9.7	9.7	9.7	9.7	9.7	18.8	34.6	15.7



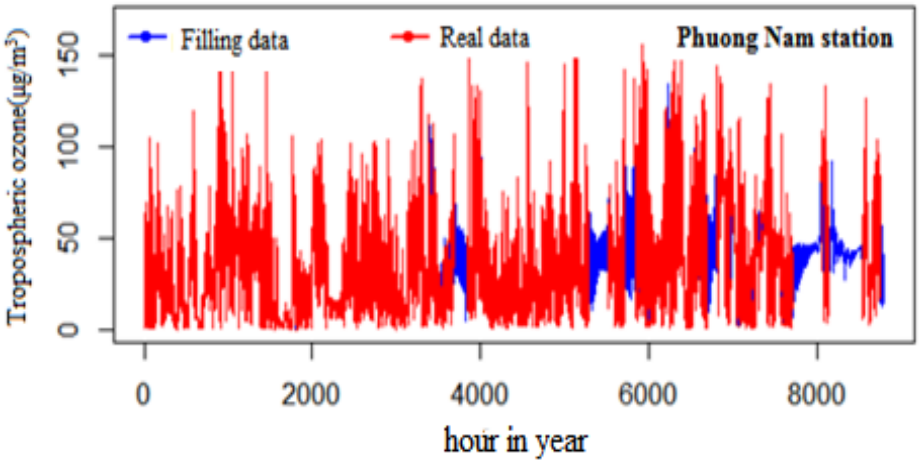


Fig.2. Filling missing data at three stations.

In Figure 2, the red line is the existing (observation) data and the blue line is filling data. The performance of ARMA algorithm in filling up the data of ozone tropospheric was evaluated as shown in Table 2.

Table 2. The performance of ARMA algorithm in filling ozone tropospheric data

Parameters	Cao Xanh	Uong Bi	Phuong Nam
RMSE (µg/m^3)	26.87	19.75	17.35
MAE (µg/m^3)	18.09	11.18	9.64
r	0.57	0.72	0.81

The correlation coefficients increase from Cao Xanh station (0.57) to Phuong Nam station (0.81), proposing that the algorithm can fill up data better when the missing rate is less.

It can be seen that the results of ARMA algorithm in Uong Bi and Phuong Nam stations better than Cao Xanh station, explained by the missing rates of Uong Bi station (4.8%), Phuong Nam station (9.7%) and Cao Xanh station (22.8%). However, the relatively high correlation coefficients indicate that this algorithm is suitable for filling up data and thereby, improving the forecasting results.

3.2. Forecasting results of tropospheric ozone for 1 hour

Results of forecasting of tropospheric ozone for 1 hour in three stations are presented in Figure 3. The performance of SVM and MLP models in forecasting at three stations was assessed as shown in Table 3.

Table 3. Performance of two models in forecasting tropospheric ozone levels at three stations

Parameter	Cao Xanh		Uong Bi		Phuong Nam	
	MLP	SVM	MLP	SVM	MLP	SVM
RMSE (µg/m^3)	28.54	28.20	11.87	10.75	11.24	10.51
MAE (µg/m^3)	15.09	14.33	7.18	6.37	6.75	6.06
r	0.85	0.86	0.88	0.91	0.86	0.88

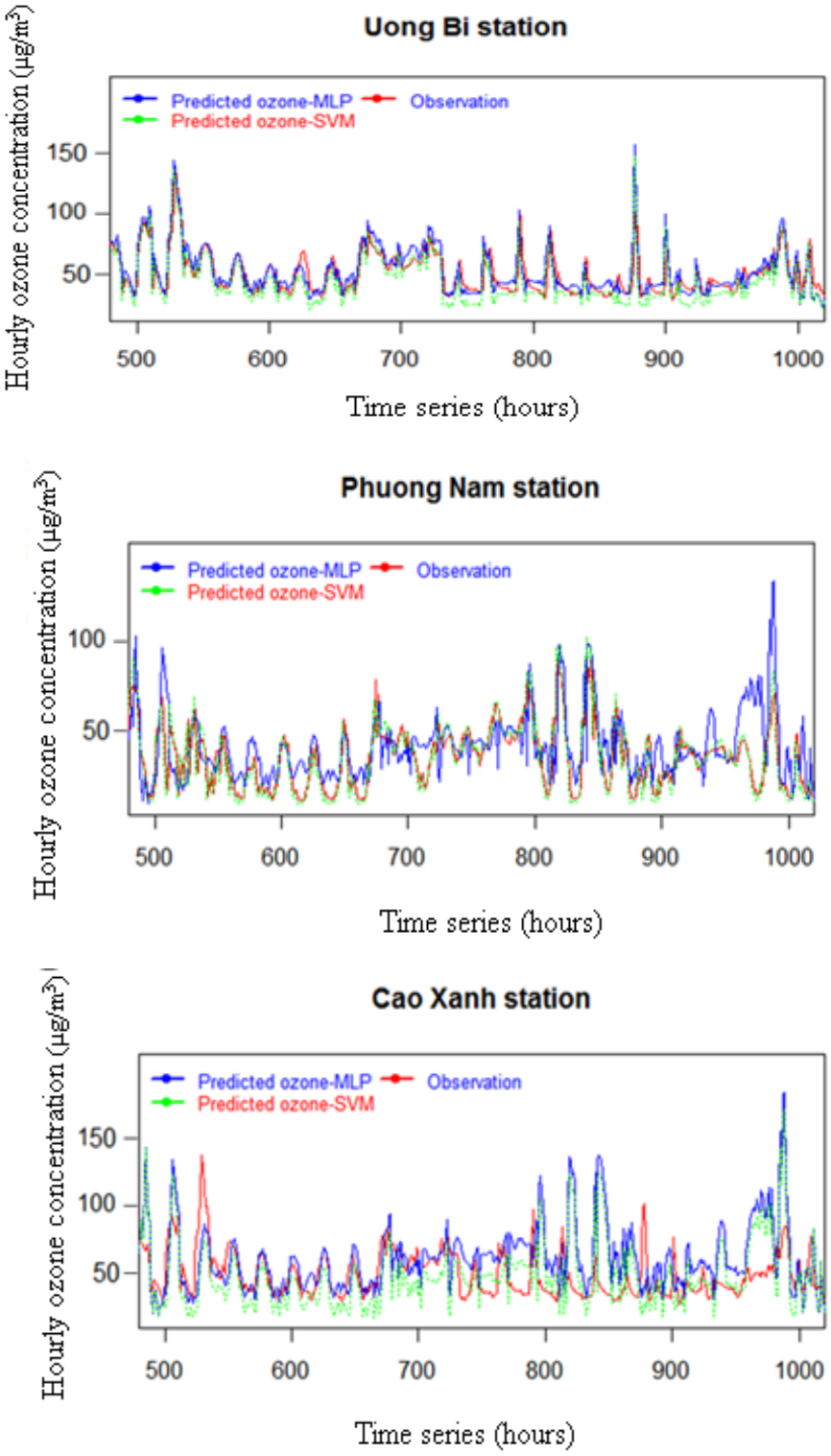


Figure. 3. Simulating ozone concentration forecast at three stations using MLP and SVM.

For both SVM and MLP models, the performance is not much different, with r ranging from 0.85 to 0.91. In particular, the correlation coefficient of MLP at three stations is lower than SVM.

In Table 2, both MLP and SVM in Cao Xanh station are lower than those in Uong Bi and Phuong Nam station are. This result can be explained by the fact that the accuracy of the forecasting of SVM or MLP models depends on the quality of the input data. In this study, the rate of missing data of the monitoring station in Cao Xanh is the largest, so this factor significantly affects the performance of the model. Table 2 shows that MAE and RMSE decrease gradually from Cao Xanh to Uong Bi and Phuong Nam station, showing the increasing the accuracy of forecasting at the respective stations. The smaller the values of MAE and RMSE, the higher the accuracy of the forecast results. MAE and RMSE of Uong Bi and Phuong Nam stations are quite similar and much lower than Cao Xanh station. This result confirms that the lack of data, especially the large gaps that have greatly affected the accuracy of the forecast. The values of MAE and RMSE also show that the accuracy of the model is gradually improved from MLP to SVM. SVM has the ability to not only predict the exact ozone concentration but also to predict the trend of ozone change. The results of this study are similar to those of Wei's in that MLP model may encounter localized, articular minimization problems, inherent in most artificial neural networks (ANN), while the SVM provides a solution to overcome these problems [13].

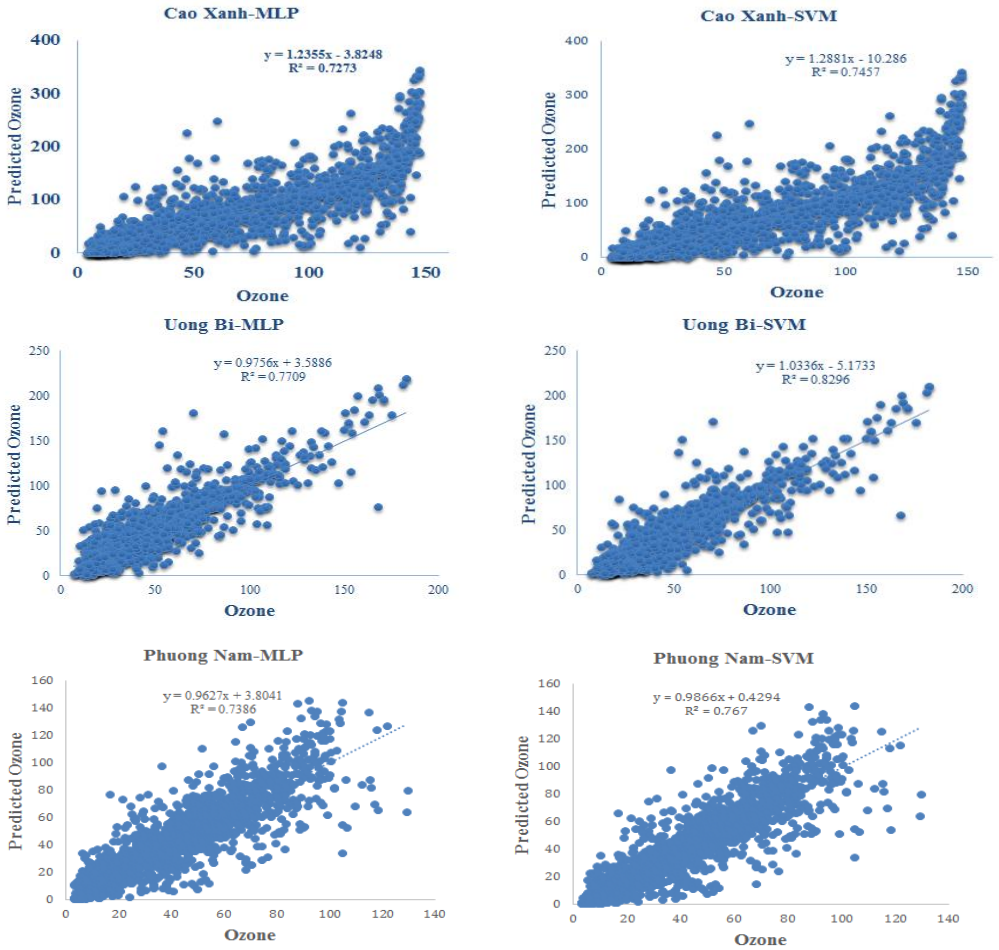


Figure.4. Scatter plots of the observation and predicted tropospheric ozone for two models.

Therefore, using SVM model to predict tropospheric ozone or other air pollutants is a promising tool. Both MLP and SVM models have shown their good ability in the forecasts of low concentrations of tropospheric ozone. However, they are not good enough in the forecast of high ozone concentrations and high variations.

At Phuong Nam and Cao Xanh stations, SVM shows a more accurate forecast of ozone fluctuations compared to MLP, especially in high ozone concentrations. These two shortcomings of the MLP model are further improved at Uong Bi station; not almost all forecasts of SVM and MLP are much different, especially in areas with high ozone levels.

Figure 4 shows the comparison between the observed ozone concentration and the forecasted one for both SVM and MLP at three stations. It can be seen from Figure 4 that, both SVM and MLP have relatively high r^2, indicating that both models can predict well the hourly ozone concentration, data points are less dispersed. However, the SVM model has better predictability than the MLP model by comparing the r^2 coefficient between the two models, typically at Uong Bi station. From the results of all stations shown in this study, to predict tropospheric ozone concentration in Quang Ninh, the SVM model will be preferred for use due to its greater accuracy.

4. Conclusion

The prediction of hourly concentrations of tropospheric ozone at three locations of Quang Ninh province, namely Cao Xanh, Uong Bi and Cao Xanh was conducted using artificial intelligence with two models, MLP and SVM. The performance of these models in the forecast of tropospheric ozone was evaluated by RMSE, MAE and correlation coefficient. The results show that, for the dataset used in this study, SVM is better than MLP in the forecast of tropospheric ozone, especially in the situations of high fluctuations and high concentrations of ozone.

References

[1] O. Hov, Tropospheric Ozone Research: Tropospheric Ozone in the Regional and Sub-regional Context, Springer Science & Business Media, New York, 2012.

[2] I.S. Isaksen, Tropospheric Ozone: Regional and Global Scale Interactions, Springer Science & Business Media, New York, 2012.

[3] H.J. Seinfeld, N.S Pandis, Atmospheric chemistry and physics: from air pollution to climate, John Wiley & Sons Inc, New Jersey, 2016.

[4] S. Al-Alawi, S. Abdul-Wahab, Assessment and prediction of tropospheric ozone concentration levels using artificial neural networks, Environmental Modelling & Software 17 (2002) 219–228. https://doi.org/10.1016/S1364-8152 (01) 00077-9.

[5] A. Abri, S. Eman, Modelling Atmospheric Ozone Concentration Using Machine Learning Algorithms, Loughborough University, Loughborough, 2016.

[6] EPA, Guidelines for Developing an Air Quality (Ozone and $PM_{2.5}$) Forecasting Program, North Carolina, 2003.

[7] M. Awad, R. Khanna, Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers, New York Apress, Berkeley, CA, 2015.

[8] H.Q. Bang, H.D. Nguyen, K. Vu, V.T. Hien, Photochemical Smog Modelling Using the Air Pollution Chemical Transport Model (TAPM-CTM) in Ho Chi Minh City, Vietnam, Environ Model Assess 24 295–310 (2019). https://doi.org/10.1007/s10666-018-9613-7

[9] H.Q. Bang, N.T. Tam, V.H.N Khue, A study on the development of ozone pollution map and ozone pollution regime in Can Tho city to propose solutions to reduce ozone pollution, Journal of Science and Technology Development 1(6) (2017) 247- 257. https://doi.org/10.32508/stdjns.v1i6.635

[10] L.H. Nghiem,N.T.K. Oanh, Comparative analysis of maximum daily ozone levels in urban areas predicted by different statistical models, Science Asia 35(3) (2009) 276–283. https://doi.org/10.2306/scienceasia1513-1874.2009.35.276.

[11] M.D. Hung, N.T. Dung, H.X. Co, Application of machine learning to fill in the missing monitoring data of air quality, Vietnam Journal of Science and Technology 56 (2C) (2018) 104-110. https://doi.org/10.15625/2525-2518/56/2C/13036.

[12] A. Smola and S.V.N. Vishwanathan, Introduction to Machine Learning, Cambridge University Press, Cambridge, 2008.

[13] W.Z. Lu, D. Wang, Learning machines: Rationale and application in ground-level ozone prediction, Applied Soft Computing 24 (2014) 135-141. https://doi.org/10.1016/j.asoc.2014.07.008.