Archives of Environmental Protection
Vol. 47 no. 3 pp. 98–107

PL ISSN 2083-4772
DOI 10.24425/aep.2021.138468

Prediction of PM2.5 hourly concentrations in Beijing based on machine learning algorithm and ground-based LiDAR

Zhiyuan Fang[1,2,3], Hao Yang[1,2,3,] Cheng Li[1,2,3], Liangliang Cheng[1,2,3], Ming Zhao[1,2]*, Chenbo Xie[1,2]*

[1]Key Laboratory of Atmospheric Optics, Anhui Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Hefei 230031, China
[2]Science Island Branch of Graduate School, University of Science and Technology of China, Hefei 230026, China
[3]Advanced Laser Technology Laboratory of Anhui Province, Hefei 230037, China

*Corresponding author's e-mail: zhaom@aiofm.ac.cn, cbxie@aiofm.ac.cn

Keywords: PM2.5; LiDAR; Machine Learning; Air pollution monitoring.

Abstract: The prediction of PM2.5 is important for environmental forecasting and air pollution control. In this study, four machine learning methods, ground-based LiDAR data and meteorological data were used to predict the ground-level PM2.5 concentrations in Beijing. Among the four methods, the random forest (RF) method was the most effective in predicting ground-level PM2.5 concentrations. Compared with BP neural network, support vector machine (SVM), and various linear fitting methods, the accuracy of the RF method was superior by 10%. The method can describe the spatial and temporal variation in PM2.5 concentrations under different meteorological conditions, with low root mean square error (RMSE) and mean square deviation (MD), and the consistency index (IA) reached 99.69%. Under different weather conditions, the hourly variation in PM2.5 concentrations has a good descriptive ability. In this paper, we analyzed the weights of input variables in the RF method, constructed a pollution case to correspond to the relationship between input variables and PM2.5, and analyzed the sources of pollutants via HYSPLIT backward trajectory. This method can study the interaction between PM2.5 and air pollution variables, and provide new ideas for preventing and forecasting air pollution.

Introduction

With the development of the economy and the acceleration of industrialization, environmental pollution has become a serious social problem. Amongst the main causes of environmental pollution are fossil fuels, biomass and dust generated during construction, which make fine particles less than or equal to 2.5 microns (Kaufman et al. 2002). Particles less than 2.5 microns in diameter (PM2.5) can remain in the atmosphere for a long period of time. These particles can cause hazy weather, interact with other substances when dispersed by wind, and also harmful for human health (Butt et al. 2017). Due to rapid economic development, PM2.5 pollution is severe in some areas of China, which has caused widespread concern among the government and the public. The Chinese government has developed a series of ambient air quality standards that include PM2.5 and other pollutants in the list of those to be monitored and currently, the larger cities have ground-based monitoring networks for PM2.5 monitoring (Zhenyi et al., 2014, Gui et al. 2016)

The main methods currently used to monitor PM2.5 are simulation prediction, measurement, and statistical analysis (Gui et al. 2016) The simulation prediction method considers atmospheric transport models and combines physical, chemical and meteorological models to predict the evolution of aerosols and PM2.5. However, due to the complex interactions between meteorological conditions, pollutant emissions and the actual atmosphere, approximations and simplifications are inevitable in the models, leading to errors in monitoring and predicting PM2.5 concentrations. The measurement methods are divided into weight, β-ray absorption and microbalance methods. Due to measurement errors and equipment limitations, these methods can only be conditionally used to monitor the mass concentration of atmospheric PM2.5 (Chu et al. 2016). In contrast, under various atmospheric conditions, statistical methods combine simulation prediction methods and measurement methods to accumulate models and historical data pertaining to meteorology and PM2.5, which then help to predict future PM2.5 concentrations based on a large amount of data (Yan et al. 2016). In recent years, due to the wide application of machine learning, a series of advances has been made in applying machine learning algorithms to PM2.5 statistical methods. To explore the quantitative relationship

between AOD and PM2.5, multiple linear regression models, neural networks, nonlinear regression models, mixed-effects models, and hidden Markov models in machine learning have been applied (Berdnik and Loiko 2016, Jones 2008, Nabavi et al. 2018). "These methods have been currently widely used mainly on satellites, such as the Moderate Resolution Imaging Spectroradiometer (MODIS) and CALIPSO's AOD products, which are used for PM2.5 predictions over large ground areas (Belle and Liu 2016, Hutchison et al 2008, Toth et al. 2018). Although reflecting the pollution distribution over the entire atmosphere, this approach is limited by space and time, and can only reflect the pollution situation at the moment of transit, precluding continuous weather conditions and hourly PM2.5 concentrations. Li et al. (2019) proposed a new multiple regression model to predict daily average PM10 concentrations using AOD and meteorological data from satellite observations (Li et al. 2019). However, this method has 30% error in the results."

Ground-based LiDAR is favored as a detection tool for its continuity and accuracy as it can facilitate the real-time continuous monitoring of pollution in a designated area and obtain more accurate AOD values compared to satellite-based LiDAR to make up for the shortcomings of satellite-based LiDAR (Chan 2009). In this work, we combine the AOD values with meteorological data and construct four empirical models using machine learning algorithms to predict the hourly PM2.5 distributions in Beijing. We then compare and analyze the performances of the four models, and analyze the influential factors in the pollution process.

Data and Method

Study area

Severe pollution caused by the fine particulate matter dominated by PM2.5 can bring about a series of economic, environmental and health problems. Beijing, as the capital of China and the core economic zone of northern China, is frequently subjected to hazy weather, so the study of PM2.5 is very important. For the monitoring of PM2.5 and other pollutants, environmental monitoring stations around the world take daily measurements (Zhenyi et al. 2015, Yan et al. 2016). In this study, the Huairou monitoring station (116.644°E, 40.3937°N) was selected to obtain the hourly average PM2.5 and air quality monitoring data. The location of the Huairou monitoring station is shown in Fig. 1.

Meteorological data

Meteorological conditions affect the formation and transportation processes of atmospheric pollutants, which are key factors in estimating PM2.5. Meteorological variables include air temperature (AT), relative humidity (RH), wind speed (WS), and pressure (P). The meteorological data are obtained by releasing meteorological balloons at the Huairou meteorological site at 7:15 a.m. and 7:15 p.m. daily. Derived from the monitoring and meteorological data, the air quality measurements are shown in Table 1. In this paper, the data set includes the same period of AOD (from ground-based LiDAR), air pollutants (PM2.5, NO_2, SO_2, O_3, and CO), and meteorological factors (wind speed, temperature, relative humidity, and pressure) (Zhenyi et al. 2015, Gui et al. 2016, Li and Zhang 2019). Table 1 shows the range of pollution data and meteorological data for the Huairou station.

LiDAR data

The ground-based LiDAR data were obtained from the aerosol water vapor detection LiDAR developed by the Key Laboratory of Atmospheric Optics, Anhui Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, as shown in Fig. 1. The measurement experiment was conducted continuously from November to December at the Yanqi Lake campus of the University of Chinese Academy of Sciences (40.41°N 116.68°E). As can be seen from Fig. 1, the distance between the LiDAR observation site and the monitoring site in Huairou is only 3.56 km, which has little effect on the experimental results.

In this study in order to adjust the laser working status, the lidar system is interrupted for 4 minutes after every 15 minutes of continuous operation. During the experiment, the measurement was stopped when rain or snow was encountered. Considering the strong influence of low-altitude airflow on surface PM2.5, the starting height of the processing of the LiDAR signal in this study was 225 m. The specific LiDAR system's parameters are shown in Table 2. In the process of the inversion of AOD, the atmospheric extinction coefficient is first inverted by the Fernald method, and the LiDAR constant is calibrated with a sun photometer, then the AOD is inverted (Fernald 1984, Hu et al. 2006). The accuracy of the extinction coefficients is improved by segmenting the calibration points. For machine learning, the input variables are divided into two groups, including atmospheric pollutants and weather variables, and AOD is classified as a range of atmospheric pollutants. The

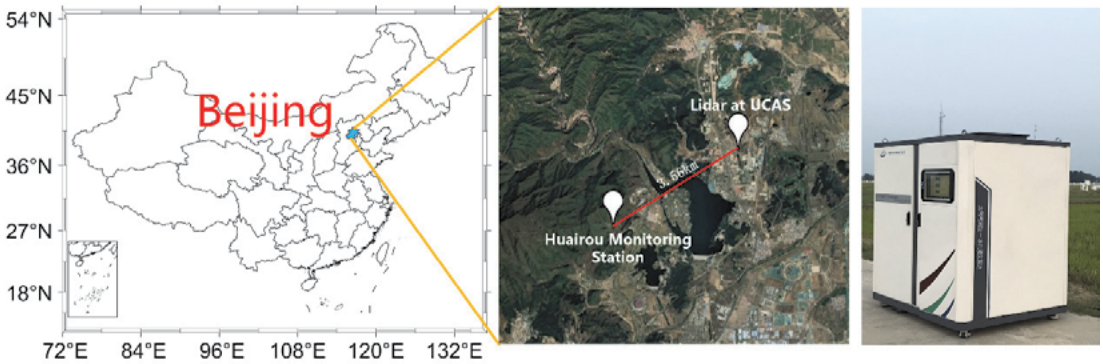

Fig. 1. Topographic and Location map of Beijing, the location of Huairou Monitoring Station, and ground-based LiDAR at UCAS

data were normalized to help improve the speed and accuracy of the model before the training.

Method of Experiment

In recent years, machine learning algorithms(MLAs) have been widely used in weather inversion. Since machine learning requires a large amount of observation data for training, the first task is to collect a large amount of data and construct a dataset. In this study, the machine learning components used in this research include multiple linear regression (MLR), support vector machine (SVM), backpropagation neural network (BP neural network), and ground-based LiDAR random forest (RF) (Bishop 1995, Mao et al. 2017, Breiman 1996). All data sets were divided into a test set and a training set; the model used the data from January 2015 as the test set and the data from November and December 2014 as the training set. To avoid overfitting, the training set was divided into training data and validation data.

In regression analysis, if there are two or more independent variables, they are called multiple regression. When influenced by multiple factors, multiple linear regression analysis can be used. Multiple linear regression attempts to predict the outcome by describing the relationship between two or more independent variables, and the output by using a linear equation. The equation is expressed as:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + ... + b_nx_n + \varepsilon(n = 1,2,3....) \quad (1)$$

$$z = \min \sum (y_i - \hat{y}_i)^2 \quad (2)$$

where b_0, b_1, b_2, b_3 ... b_n are the parameter values, ε is the value of the output of the error, and the predicted value is z. The results were estimated using the least-squares method, which minimizes the square of the difference between the two values. This method is easier in multi-factor model analysis, and can accurately correlate the degree of each factor and improve the validity of the prediction equation and it can also accurately correlate the degree of correlation and regression fit between the factors. The disadvantage of this method is that some dependent variables are unpredictable in some analyses, making regression analysis limited.

Backpropagation neural network (BPNN) is one of the most widely used network models proposed by a group of scientists led by Rumelhart and McClelland in 1986 (Bishop 1995). It is a multilayer feed forward network trained based on an error backpropagation algorithm. The BP neural network method is the fastest descent method that uses backpropagation to continuously adjust the weights and thresholds of the network to minimize the sum of squared errors (Nabavi et al. 2018, Mao et al. 2017). The BP neural network method contains hidden and output layer as well as AOD, temperature, relative humidity, wind speed and pressure, CO, SO_2, NO_2, O_3 which were calculated into the network as meteorological factors. The hidden and output layers selected for this paper are shown in Fig. 2.

Support vector machines (SVMs) are a class of generalized linear classifiers that classify data in a supervised learning manner (Breiman 1996, Liu et al. 2017). Support vector machines obtain a classification function formally similar to a neural network, whose output is a linear combination of multiple intermediate nodes, each corresponding to an input sample and an inner product of support vectors. Support vector machine algorithms have significant advantages in solving nonlinear, small-sample, large-dimensional problems (Liu et al. 2017).

Table 1. Statistics of measured variables at Huairou station from 1 November 2014 to 1 January 2015

Variable	Unit	Range	Mean	St. Dev
$PM_{2.5}$ (hourly)	μg/m^3	[5,280]	62.74	66.76
AOD (532 nm)	float	[0,2]	0.52	0.43
Windspeed (hourly)	m/s	[1,5]	2.12	1.53
Temperature (hourly)	°C	[-9,20]	1.76	4.61
RH (hourly)	%	[0,100]	42.61	25.03
Pressure (hourly)	hPa	[1007.3,1036]	1022.93	5.21
CO (hourly)	mg/m^3	[0,6]	1.35	0.98
NO_2 (hourly)	μg/m^3	[0,130]	45.26	31.07
O_3 (hourly)	μg/m^3	[0,84]	28.08	26.93
SO_2 (hourly)	μg/m^3	[0,121]	21.04	19.61

Table 2. System Parameters of the LiDAR System

Item	Technical Parameters
Laser Company	Continuum
Wavelength/nm	355/532/1064
Pluse energy/mJ	50/90/250
Pulse width/ns	20
Repetition rate/Hz	20
Receiving Telescope/mm	400 diameter, Cassegrain

Random forest(RF) is an effective statistical method for solving nonlinear relationships, which was proposed in 2001 and is mainly based on decision tree theory, with the basic principle of constructing each tree based on samples, and then dividing the samples by the best randomly selected partition points. Fig. 3 shows the structure of the decision tree. The advantage of this method is that it is robust to overfitting and the error converges to a limit as the number of forests increases.

Error assessment

In order to evaluate the performance needs of the model, the root mean square error (RMSE), mean deviation (MD) and index of agreement (IA) are used in this paper, as shown in Equations 3–5 (Yang et al. 2017).

$$\mathrm{MD} = \frac{1}{N}\sum_{i=1}^{N}\left|O_i - P_i\right| \tag{3}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(O_i - P_i\right)^2} \tag{4}$$

$$\mathrm{IA} = 1 - \frac{\sum_{i=1}^{N}\left(O_i - P_i\right)^2}{\sum_{i=1}^{N}\left(\left|O_i - \bar{O}\right| + \left|P_i - \bar{O}\right|\right)^2} \tag{5}$$

Where N is the number of time points; O_i and P_i represent the observed and predicted values, respectively; and $\bar{O}$ is the observation mean.(add)

Results and discussion

Comparison of Model Performance in Testing Set

Since machine learning algorithms require a large amount of observed data for training, the dataset should be divided into two groups: test data and training data. The training data set accounts for 80% of the total data volume and the test data accounts for 20% (Mao et al. 2017, Yang et al. 2020). To compare the results of all test data, the predicted values of PM2.5 are extracted from the results of four machine learning algorithms. In this paper, we use the data from January 2015 as the test set and the data from November 2014 and December 2014 as the training set.

The inputs to the model include the datasets of AOD, SO_2, NO_2, CO, O_3, AT, RH, WS, and P. The modeling results are showed in Fig.3. The testing data have showed the severe pollution in Beijing in January 2014. Although the four MLAs made higher predictions of PM2.5, the limited pollution data in the training set has a significant impact on the results of MLR, SVM and BPNN. The prediction accuracy of the RF model is higher than that of the other three models given the decision tree approach, which improves the learning efficiency. The MD and RMSE values of RF are lower than those of the other three MLAs. It can be seen that the use of random forest can improve the accuracy of sample prediction when the sample size is small.

Comparison of Four Different Models in Total Data Set

To compare four MLAs, the total data from Novemebr 2014 to January 2015 were used in this experiment. All data sets were

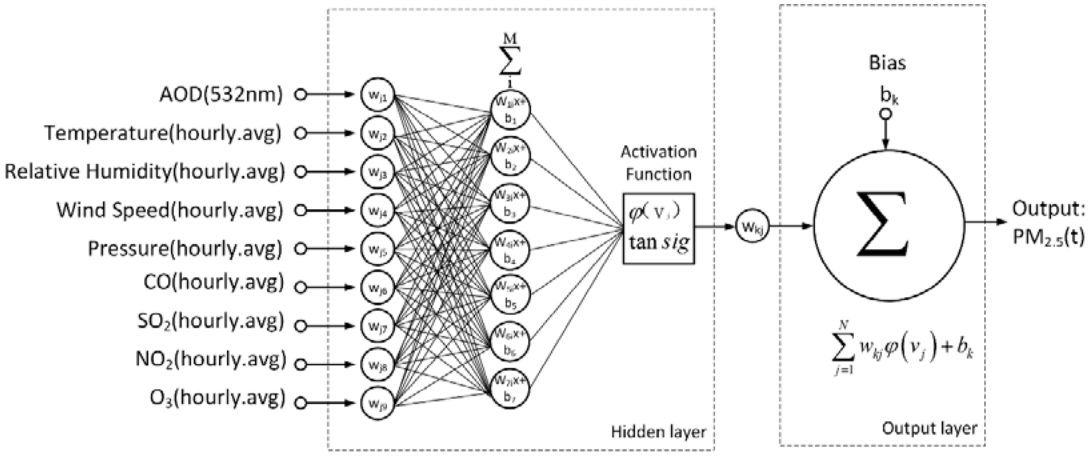

Fig. 2. Architecture of the BP neural network

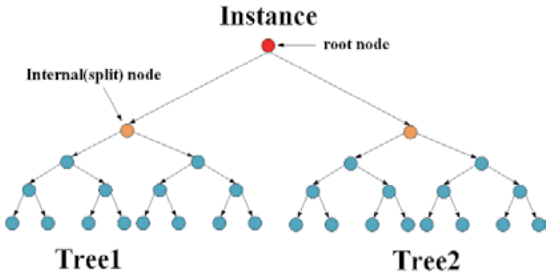

Fig. 3. A common structure of a decision tree. The red circle is the root node; and there is only one root node in the example. The yellow circles are the internal nodes used to perform the segmentation. The blue circles are the terminal nodes, also known as the leaf nodes, which are the predicted outcome.

studied via machine learning and the initial conditions which were re-entered for four MLAs. The results are analyzed and compared in Fig. 4.

In Fig. 4, the fitted R2 decreased compared to Fig. 3 when analyzed in large number of samples. Weather which decreased the accuracy of the prediction had a greater influence here. However, the predicted R2 of MLR, SVM, and BPNN was still in the range from 0.83 to 0.86 after the study of a large number of samples. Due to their higher accuracy, the predicted values can be important tools in assessing PM2.5. Besides, the overall prediction accuracy of the RF model is superior to those three models, as its R2 reached 0.97.

Fig. 5 shows the predicted and observed ground-level PM2.5 at different pollution levels; the color bar indicates probability of the number of data points, the red solid line is linear regression. With clean weather, the values of the correlation coefficients (R) were 0.86–0.97, and the RMSE was between 5.37 and 15.51. On more moderate days, MLR, SVM

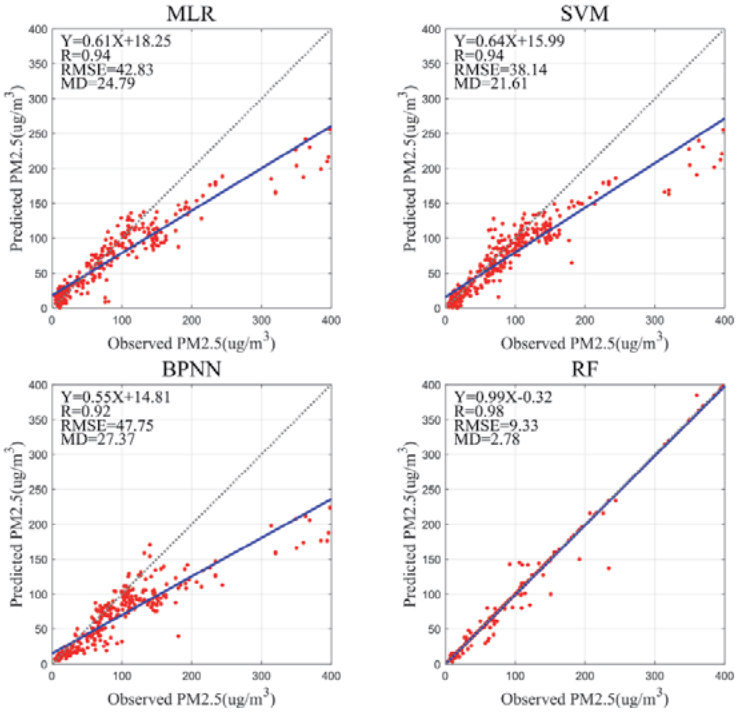

Fig. 4. Scatter plots of predicted vs. observed ground-level PM2.5 in the testing set in Beijing during Nov. 2017–Jan. 2015 in four MLAs.

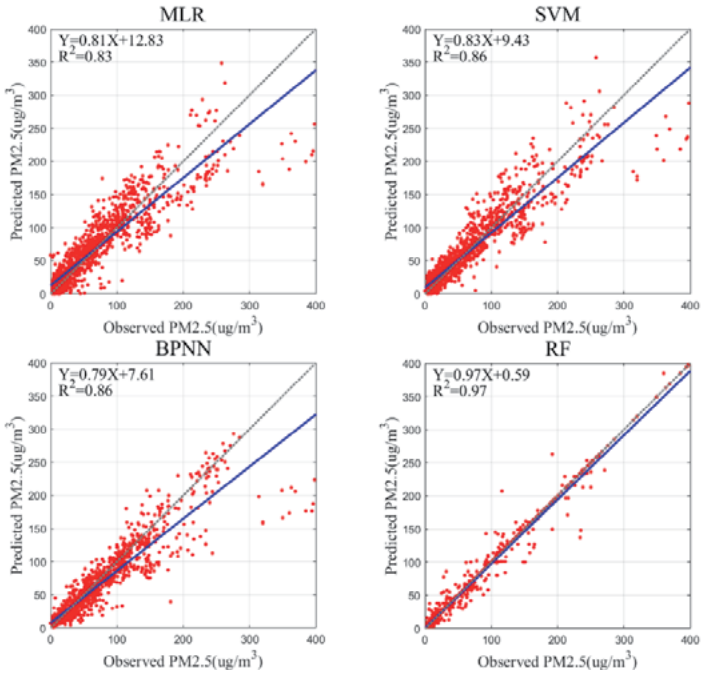

Fig. 5. Scatter plots of predicted vs. observed ground-level PM2.5 in total set in Beijing during Nov. 2017–Jan. 2015 in four MLAs.

and BPNN displayed lower R than on clean days, reaching 0.48–0.52, while the RF reached a value of 0.91 in R. The value of RMSE also increased in this example. The accuracy of the prediction was reduced on moderate days due to the presence of clouds and various complex weather characteristics. Under heavy pollution, the value of R was 0.52–0.64, except for when using the RF method. Additionally, the value of RMSE was higher contrasting with the other two days. The largest RMSE was 75.84 for BPNN, which means that the predicted value was unstable. There are possibly two reasons for this. The error in the LiDAR observation was higher on these days because of the complicated weather conditions. The experiment extracted about 129 retrieval counts under heavy air pollution, and the products were relatively low.

In conclusion, although the accuracy of the prediction for different weather conditions was inevitably lower in some cases, the results achieved by combining different variables with the MLAs were a good way to predict PM2.5 concentrations. Particularly for RF, the accuracy and mean-variance of the predicted results were very stable.

Error Analysis

The comparison of the MLR, SVM, BPNN, and RF is clearly delineated in Table 3. Generally, the R^2 was changed slowly as the amount of data derived from all the MLAs increased (Nabavi et al. 2018). However, the MD and RMSE decreased to some degree in MLR, SVM, and BPNN, which means that the degree of deviation between predicted and measured values falls as the quantity of data increases. RF offered stable values, as the samples increased in terms of absolute errors (measured by RMSE and MAE). However, as concerns IA as the relative measure in the four methods, the total set has higher results than the testing set. For BPNN, the result was quite satisfactory, with the value ranging from 86.31% to 94.11%.

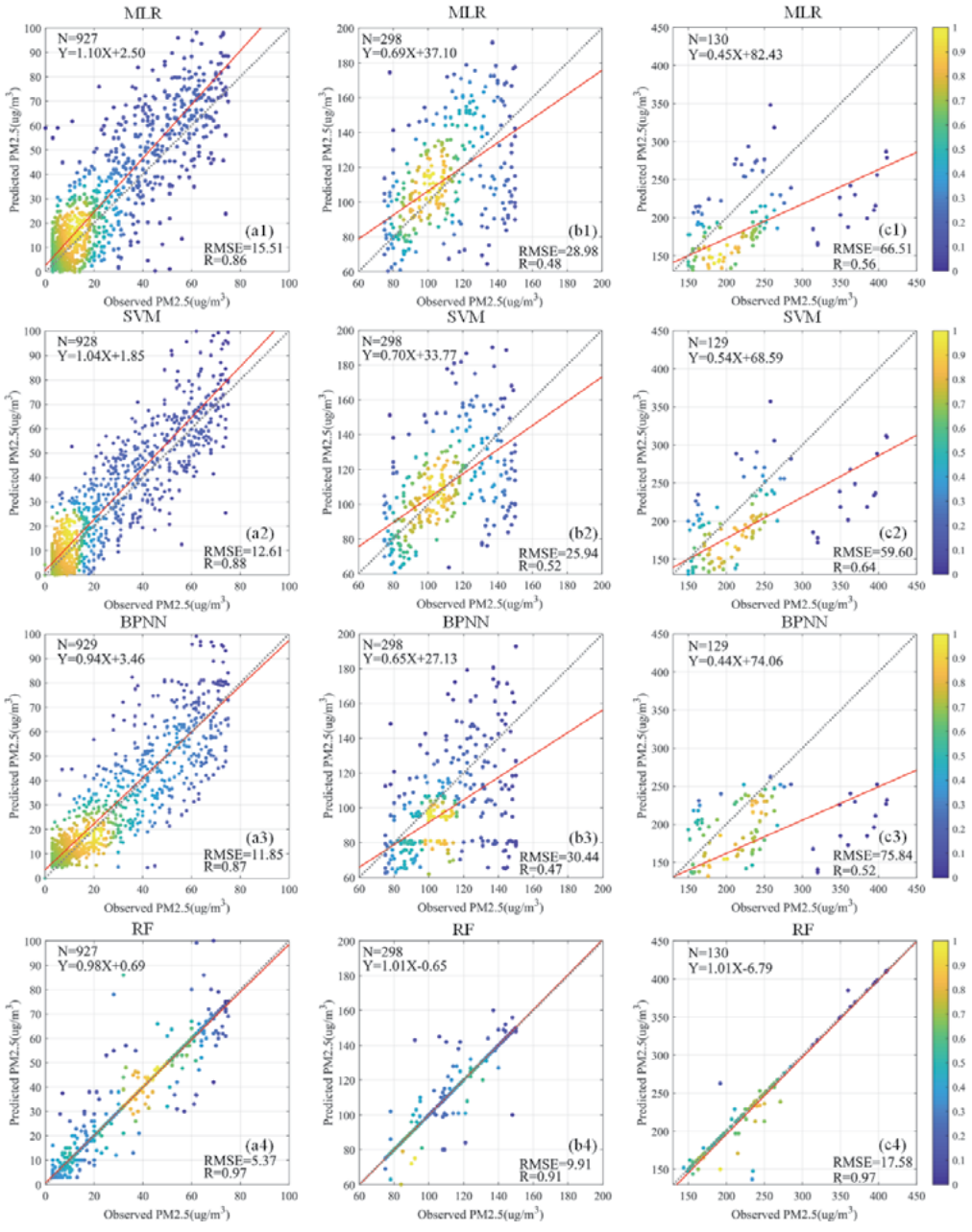

Fig. 6. Scatter plots of predicted vs. observed ground-level PM2.5 under (a) clean weather, (b) moderate weather, and (c) heavy pollution

High concentrations of PM2.5 can lead to more errors when air pollution is very severe in Beijing. To decrease the errors on polluted days, the number of statistics days used should be increased, and the number of weather conditions should be assessed.

The distribution of prediction errors was calculated according to Equations 3–5, where the horizontal axis represents the absolute error values and the vertical axis represents the frequency of occurrence, as shown in Fig. 6. For all MLAs, more than 70% of the absolute errors were in the region of 10 μg/m^3, with fewer values exceeding 50 μg/m^3. It can be seen from Fig.6 that BPNN and RF are superior to other methods, with more than 80% of 0–10 μg/m^3. Additionally, RF achieved the highest results among all the machine learning methods, which means that it is the most reliable.

The RF model was relatively weak in predicting the peak concentration of PM2.5. This may be due to the mean effect of the regression analysis. The presence of large residuals was affected by heavy PM2.5 pollution due to different weather conditions, such as precipitation and cloudiness, and was limited by a weak predictive ability in the regression analysis.

Prediction contribution

To study the influence of each parameter on the results, we applied the method of importance analysis of the variables to quantitatively describe the input parameters. By calling this module the importance of the features trained by the RF method can be extracted and the input variables can be ranked according to their importance, so this module can analyze the importance of the input variables in predicting the effect of PM2.5 concentrations. Using the weight analysis module in the RF method allows for analyzing the influence of the input variables on the predicted PM2.5 concentrations (Yang et al. 2020). The final results obtained are shown in Fig. 7. The horizontal coordinates indicate the content of the input variables, and the vertical coordinates indicate the first-order indices. From Fig. 7 it can be seen that AOD, NO_2, SO_2, and CO display high first-order sensitivity, indicating a strong influence on PM2.5. Among them, AOD is the most influential predictor, indicating that aerosol concentration can reflect PM2.5 content to some extent. Other variables influence the concentration of PM2.5 to varying degrees through their interaction, so it is necessary to further investigate the four

Table 3. The comparison of the MLR, SVM, BPNN, and RF

Methods	Data set	R^2	MD (μg/m^3)	RMSE (μg/m^3)	IA (%)
MLR	testing set total set	0.83 0.88	24.79 17.57	42.83 27.73	91.79 94.96
SVM	testing set total set	0.86 0.88	21.61 17.28	38.14 27.78	96.31 96.65
BPNN	testing set total set	0.86 0.85	27.37 16.87	47.75 30.79	86.31 94.11
RF	testing set total set	0.97 0.96	2.78 2.71	9.33 10.25	99.40 99.69

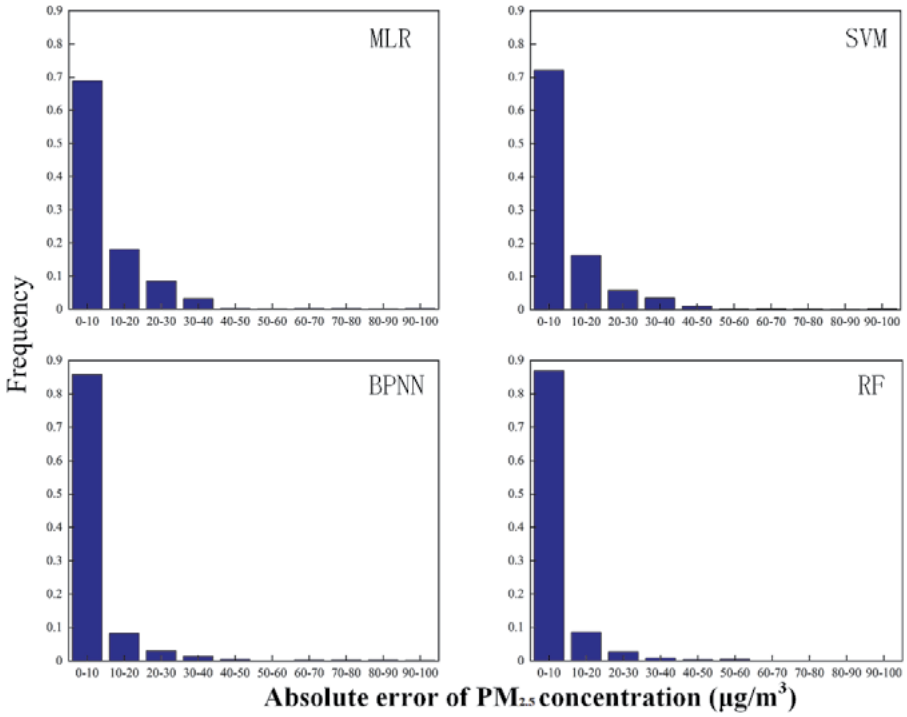

Fig. 7. Error in PM2.5 Concentration Prediction

input variables, and we selected a pollution process through which we could study and analyze their relationship.

Estimation of PM2.5 concentrations during a haze period

We extracted the largest factors affecting PM2.5 predictions, including AOD, NO_2, SO_2 and CO, and these are shown in Fig. 8. The relationship between the above-mentioned factors and PM2.5 varies with time, so it is necessary to investigate these four factors. For this purpose, we selected one pollution process (November 21–24) for analysis. Its continuous variation over 3 days was observed through graphs. As can be seen from the graphs, the trends in AOD, NO_2, SO_2 and CO are basically the same: the changes in AOD and PM2.5 are the same but they decrease from 21:00 on the 21st to the early morning of the 22nd, while the concentration of PM2.5 remains high, and the fluctuations between the two are slightly different from the 22nd to the 24th. Although AOD and PM2.5 are highly correlated, these are not the only indicators of particulate pollution, given the complexity of the pollution situation. The weakened wind speeds in autumn and winter, due to heating and coal combustion etc., provide conditions for PM2.5 accumulation, which may lead to a strong predictive power of AOD in relation to the PM2.5 concentrations. During significant air pollution, NO_2 follows a similar trend to PM2.5, but its concentration decreases slowly between the early morning of the 22nd and around 12:00 a.m. on this day. This suggests that the emissions of air pollutants, especially NO_2, are responsible for PM2.5 in most cases, as revealed in Fig.7. PM2.5 shows a lagging trend with respect to SO_2, but the overall trend is the same, indicating that there is a correlation between the two in the case of severe air pollution. However, the sensitivity is less than that of NO_2, which may be due to the chemical nature of SO_2 being less active than NO_2. The trend of CO is almost the same as that of PM2.5, indicating that in the case of severe air pollution, there is a close correlation.

According to the model results, in order to understand the process of pollution and analyze its potential sources better, we analyzed the backward trajectories of air pollutants in combination with HYSPLIT, as shown in Fig. 9. According to the 48-hour backward trajectory analysis from November 20 to 23, 2014 (Stein et al. 2016), we found that most of the pollution on the 20th came from the area around Beijing, which is a densely populated and industrially concentrated area. The trajectory line of 100 m at low altitude was short and it did not extend upward, indicating that the air at lower levels did not exchange with the air at higher levels, but the pollution increased. The air mass mainly came from the Siberian plateau, but the trajectory line of 100m at a low altitude was still short, indicating the poor weather conditions, including poor diffusion and a low boundary layer height. The air mass remained in the local area for a long time and carried a large amount of air pollutants into the Beijing area, where the pollution was further strengthened. On the 22nd, the Beijing area was affected by an air mass from the northwest; the troposphere and stratosphere layers exchanged gases, the weather conditions improved, and the pollutants gradually spread, which means the gas concentration of the pollutants decreased, which is consistent with the content of Fig. 8. For the 23rd, the HYSPLIT shows that the gases mainly came from Beijing and the surrounding cities, and the shorter trajectory lines of the air masses in these areas and the poor diffusion conditions led to increases in pollution. As a result, pollutants started to accumulate again, and the concentration of PM2.5 increased.

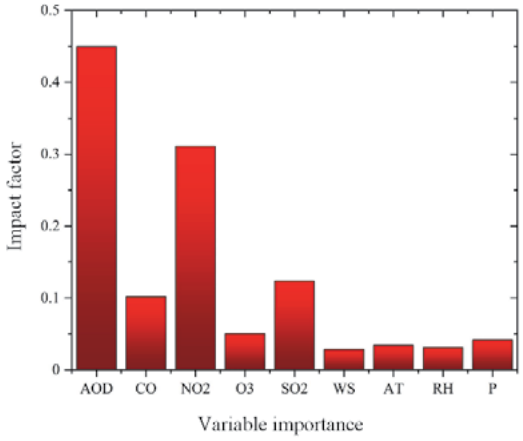

Fig. 8. Variable importance analysis of the RF model

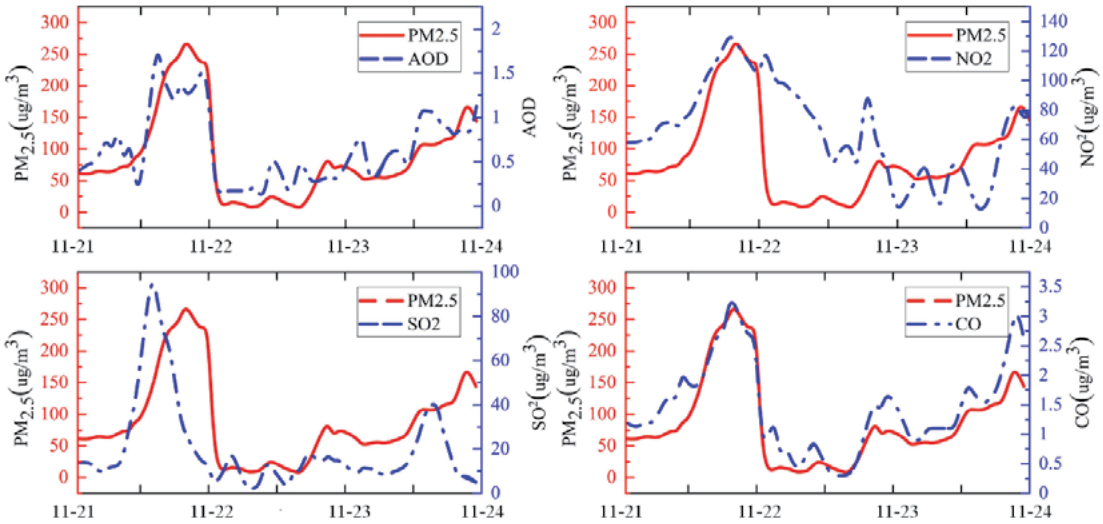

Fig. 9. Comparison of PM2.5 and AOD, NO_2, SO_2, CO at the surface during 21 Nov to 24 Nov 2014

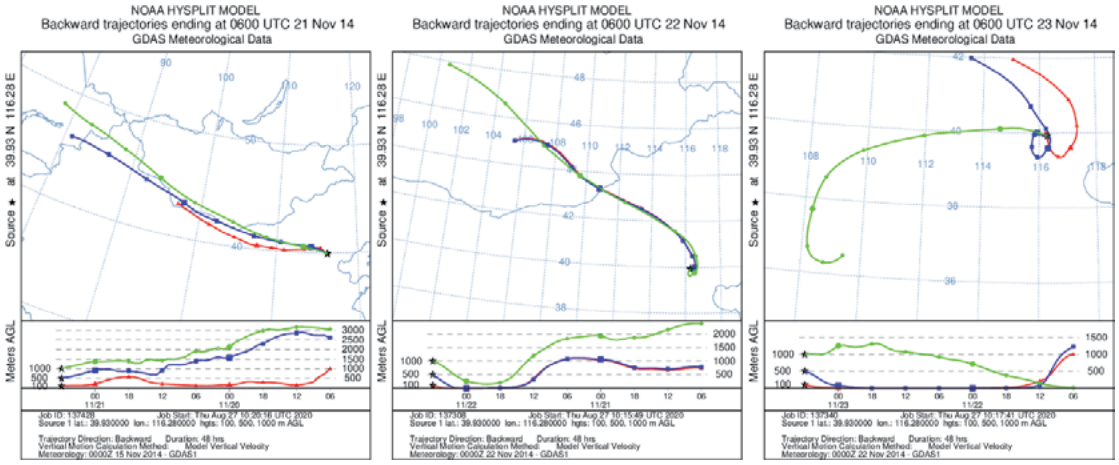

Fig. 10. HYPLIY backward trajectory generated by Hybrid Single Lagrangian Integrated Trajectory Model

Conclusion

In this paper, four MLA methods are used to combine high-quality LiDAR AOD, meteorological variables, and atmospheric pollution variables for PM2.5 estimations. The RF method has a high prediction accuracy, exceeding the other three machine learning methods, and this can predict PM2.5 concentrations in different weather with high quality. The variable control module of the RF model can be used to analyze the contribution of different predictors to PM2.5 formations. The results of the study show that the prediction accuracy of PM2.5 is higher on sunny days and lower on cloudy and polluted days. With the increase of sample size, the prediction accuracy can be improved, but there are more factors (such as wind, water and vapor, etc.) in the atmosphere that affect heavy air pollution days, which makes it difficult to perform further study on improving the prediction accuracy. However, more samples can accurately reflect the pollution factors in the environment, which can improve the accuracy of pollution prevention measures in Beijing. Combining ground-based LiDAR and meteorological factors, machine learning models can better predict and analyze air pollution. help the formation of pollution, pollution control measures, and pollution forecasting better.

Conflicts of Interest

The authors declare no conflict of interest.

References

Belle, J. & Liu, Y. (2016). Evaluation of Aqua MODIS Collection 6 AOD Parameters for Air Quality Research over the Continental United States. *Remote Sensing*, 8(10), pp. 815–820.

Berdnik, V.V. & Loiko, V.A. (2016). Neural networks for aerosol particles characterization. *Journal of Quantitative Spectroscopy & Radiative Transfer*, 184.

Bishop, C.M., (1995). Neural Networks for Pattern Recognition. *Agricultural Engineering International the Cigr Journal of Scientific Research & Development Manuscript Pm*, 12(5), pp. 1235 – 1242.

Breiman & Leo, (1996). Bagging Predictors. *Machine Learning*, 24(2), pp. 123–140.

Butt, E.W., Turnock, S.T., Rigby, R., Reddington, C.L., Yoshioka, M., Johnson, J.S., Regayre, L.A., Pringle, K.J., Mann, G.W. & Spracklen, D.V. (2017). Global and regional trends in particulate air pollution and attributable health burden over the past 50 years. *Environmental Research Letters*. 10 (12), DOI: 10.1088/1748-9326/aa87be.

Chan, P.W. (2009). Comparison of aerosol optical depth (AOD) derived from ground-based LIDAR and MODIS. *Open Atmospheric Science Journal*, 3(1), pp. 131–137.

Chu, Y., Liu, Y., Li, X., Liu, Z., Lu, H., Lu, Y., Mao, Z., Chen, X., Li, N., Ren, M., Liu, F., Tian, L., Zhu, Z. & Xiang, H. (2016). A Review on Predicting Ground PM2.5 Concentration Using Satellite Aerosol Optical Depth. *Atmosphere*, 7(10), p. 129. DOI: 10.3390/atmos7100129.

Fernald, F.G. (1984). Analysis of atmospheric lidar observations: some comments. *Applied optics*, 5, pp. 652–653.

Gui, K., Che, H., Chen, Q., An, L., Zeng, Z., Guo, Z., Zheng, Y., Wang, H., Wang, Y., Yu, J. & Zhang, X. (2016)., Aerosol Optical Properties Based on Ground and Satellite Retrievals during a Serious Haze Episode in December 2015 over Beijing. *Atmosphere*, 7(5), pp. 70, DOI: 10.3390/atmos7050070.

Hu, S, Wang, Z., Xu, Q., Zhou, J. & Hu. H. (2006). Study on Lidar Measurement of Atmospheric Aerosol Optical Thickness. *Journal of Quantum Electronics*, 3, p. 307–310. (in Chinese)

Hutchison, K.D., Faruqui, S.J. & Smi, S. (2008). The Improving correlations between MODIS aerosol optical thickness and ground-based PM2.5 observations through 3D spatial analyses. *Atmosphere Environment*, 3(42), pp. 530–554, DOI: 10.1016/j.atmosenv.2007.09.050.

Jones, R.M. (2008). Experimental evaluation of a Markov model of contaminant transport in indoor environments with application to tuberculosis transmission in commercial passenger aircraft. Dissertations & Theses – Gradworks, 2008.

Kaufman, Y.J., Tanré, D. & Boucher, O. (2002). A satellite view of aerosols in the climate system. *Nature*, 419(6903), pp. 215–23.

Li, X. & Zhang, X. (2019). Predicting ground-level PM 2.5 concentrations in the Beijing-Tianjin-Hebei region: A hybrid remote sensing and machine learning approach. *Environmental Pollution*, 249, pp. 735–749, DOI: 10.1016/j.envpol.2019.03.068.

Bing,-C.L., Binaykia, A., Chang, P-C., Tiwari, M.K. & Tsao, C-C. (2017). Urban air quality forecasting based on multi-dimensional collaborative Support Vector Regression (SVR): A case study of Beijing-Tianjin-Shijiazhuang. *Plos One*, 12(7), pp. e0179763, DOI: 10.1371/journal.pone.0179763.

Mao, X., Shen, T. & Feng, X. (2017). Prediction of hourly ground--level PM2.5 concentrations 3 days in advance using neural

networks with satellite data in eastern China. *Atmospheric Pollution Research*, 6(8), pp. 1005–1015. S1309104217300296.

Nabavi, S.O.(2018). Prediction of aerosol optical depth in West Asia using deterministic models and machine learning algorithms. Aeolian Research, 35C: p. 69–84.

Stein, A.F. (2016). NOAA's HYSPLIT atmospheric transport and dispersion modeling system. Bulletin of the American Meteorological Society, p. 150504130527006, DOI: 10.1016/j.apr.2017.04.002.

Toth, T.D., Campbell, J.R., Reid, J.S., Tackett, J.L., Vaughan, M.A., Zhang, J. & Marquis, J.W. (2018). Minimum aerosol layer detection sensitivities and their subsequent impacts on aerosol optical thickness retrievals in CALIPSO level 2 data products. *Atmospheric Measurement Techniques*, 11, p. 499–514, DOI: 10.5194/amt-11-499-2018.

Yan, D., Lei, Y., Shi, Y., Zhu, Q., Li, L.& Zhang, Z. (2018). Evolution of the spatiotemporal pattern of PM2.5 concentrations in China – a 2 case study from the Beijing-Tianjin-Hebei region. *Atmosphere Environment*. 183, pp. 225–233, DOI: 10.1016/j.atmosenv.2018.03.041.

Yang, G., Lee, H. & Lee, G. (2020). A Hybrid Deep Learning Model to Forecast Particulate Matter Concentration Levels in Seoul, South Korea. *Atmosphere*, 11(4): pp. 348, DOI: 10.3390/atmos11040348.

Wang, Y., Chen, L., Li, S., Wang, X., Yu, C., Si, Y. & Zhang, Z. (2017). Interference of Heavy Aerosol Loading on the VIIRS Aerosol Optical Depth (AOD) Retrieval Algorithm. *Remote Sensing*, 2017. 9(4): p. 397, DOI: 10.3390/rs9040397.

Chen, Z., Zhang, J., Zhang, T., Liu, W. & Liu, J. (2015). Haze observations by simultaneous lidar and WPS in Beijing before and during APEC, 2014. *Science China(Chemistry)*, 2015. 09(v.58): p. 33–40, DOI: 10.1007/s11426-015-5467-x.