Research on data correction method of micro air quality detector based on combination of partial least squares and random forest regression

Bing Liu (Liub1@niit.edu.cn)
Nanjing Vocational University of Industry Technology
Wangwang Yu
Nanjing Vocational University of Industry Technology
Yishu Wang
Nanjing Vocational University of Industry Technology
Qibao Lv
Nanjing Vocational University of Industry Technology
Chaoyang Li
Henan University of Technology

Research Article

Keywords: Electrochemical Sensor, Correlation Analysis, Influencing Factors, Vector Machine Model, Multilayer Perceptron Neural Network

Posted Date: February 26th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-241776/v1

Research on data correction method of micro air quality detector based on combination of partial least squares and random forest regression

Bing Liu[1], Wangwang Yu[2], Yishu Wang[1], Qibao Lv[1] & Chaoyang Li[3]

[1]Public Foundational Courses Department, Nanjing Vocational University of Industry Technology, Nanjing 210023, China

[2]School of Mechanical Engineering, Nanjing Vocational University of Industry Technology, Nanjing 210023, China

[3]College of Management, Henan University of Technology, Zhengzhou 450001, China

Abstract

The issue of air quality has attracted more and more attention. The main methods for monitoring the concentration of pollutants in the air include national monitoring station monitoring and micro air quality detector testing. Since the electrochemical sensor of the micro air quality detector is susceptible to interference, the monitored data has a certain deviation. In this paper, the combined model of partial least square regression and random forest regression (PLS-RFR) is used to correct the detection data of the micro air quality detector. First, correlation analysis is used to find out the factors that affect the concentration of pollutants. Second, partial least squares regression is used to give the quantitative relationship of the influence of each influencing factor on the concentration of pollutants. Finally, the predicted value of partial least squares regression and various influencing factors are used as independent variables, and the pollutant concentration monitored by the national monitoring station is used as the dependent variable, and the PLS-RFR model is obtained with the help of random forest software package. Relative Mean Absolute Percent Error (MAPE), Mean Absolute Error (MAE), goodness of fit (R^2), and Root Mean Square Error (RMSE) are used as evaluation indicators to compare PLS-RFR model, support vector machine model and multilayer perceptron neural network. The results show that no matter which evaluation index, the prediction effect of the PLS-RFR model is the best, and the model has a good prediction effect in the training set or the test set, indicating that the model has good generalization ability. This model can play an active role in the promotion and deployment of micro air quality detectors.

1. Introduction

Air quality is a problem that cannot be ignored. In addition to the harm to the climate, the most important thing is the harm to the human body. According to statistics, 3 million people die every year due to air quality problems. Many cardiovascular diseases, respiratory diseases, lung cancer and other diseases have a certain relationship with air pollution [1-3]. Although government departments are working hard to curb the adverse effects caused by air pollution, the air quality in many cities is still very severe. Therefore, air quality should be monitored by relevant national departments in real time, so that corresponding countermeasures can be taken in time to reduce the harm to humans and the environment.

In order to monitor air quality, many countries have set up national monitoring stations (national control points) in some key cities. Multi-parameter automatic monitoring equipment is installed in the national control point for continuous automatic monitoring, and the monitoring results are stored in real time and analyzed to obtain relevant data. However, the cost of building and maintaining national monitoring stations is high, so the number of national control points is very small. In addition, due to the lag in the release of national

control point data, it is difficult to conduct grid-based monitoring and real-time monitoring of air quality in a certain area.

Another commonly used monitoring method is to monitor air quality by using a micro air quality detector (self-built point). There are many advantages to using a micro air quality detector for air quality monitoring. First, the production and maintenance cost of the micro air quality detector is very low, and the installation is very simple. This advantage makes it very convenient for grid deployment of micro air quality detectors in specific areas. Second, it is easy to read the readings and can monitor the concentration of pollutants in the air in real time. Third, it can not only monitor the concentration of pollutants, but also other meteorological parameters such as temperature, wind speed, precipitation and other meteorological parameters in the area can also be monitored in real time. However, because the electrochemical sensor in the micro air quality detector is susceptible to interference from unconventional gaseous pollutants, and it is prone to range and zero offset during a period of time in use, there is a certain error in its measurement data [4, 5]. Therefore, the data of the national control point needs to be used to correct the self-built point data.

The commonly used methods for predicting the concentration of pollutants mainly include the prediction of pollution weather conditions, numerical prediction and statistical prediction. The prediction of polluted weather conditions can conduct semi-quantitative or qualitative analysis on the prediction model of future pollutant concentration through meteorology and other related theories. The reliability of the forecast of polluted weather conditions is mainly based on the forecast of pollutant weather factors and a large amount of historical data [6]. Because the pollution source is not taken into consideration in the process of predicting polluted weather conditions, this leads to low prediction accuracy of polluted weather conditions.

Numerical prediction is based on atmospheric dynamics, fluid dynamics, and thermodynamics. The prediction is a method of predicting the concentration of pollutants in the future through computer high-speed numerical calculation. Numerical prediction is a quantitative and objective method, which is different from previous predictions of polluted weather conditions based on experience and meteorology [7, 8]. Although the accuracy of numerical prediction is high, there are still some shortcomings. It is computationally complex and time-consuming, and requires a huge amount of data. Therefore, the method needs further improvement and innovation.

The statistical prediction model does not consider the physical and chemical changes of pollutants. It builds models by analyzing characteristic factors related to changes in pollutant concentration, analyzing historical statistical data [9]. A series of models have been proposed for forecasting air quality using statistical forecasting methods.

Traditional methods such as linear regression [10, 11], time series [12, 13], gray model [14], support vector machine [15-18], etc. are often used to predict the concentration of pollutants. The traditional method has a simple structure and is often easy to identify, and the model interpretation ability is relatively strong. However, the reasons for the formation of pollutants are complex. It is a complex physical process with obvious temporal and spatial differentiation and nonlinear characteristics. Therefore, traditional methods are often not particularly good in terms of prediction accuracy.

In addition, the recently popular artificial intelligence method dominated by neural network algorithms is often used to predict the concentration of pollutants [19-21]. However, the neural network algorithm has a slow convergence rate for the prediction of high-dimensional features, which is easy to cause over-fitting problems, and requires a large number of parameters to determine the network structure. In recent years, random forest regression has also been applied to air quality forecasting models [22-25]. Random forest algorithm is a highly accurate algorithm among machine learning algorithms. It can overcome the shortcomings of a single prediction (or classification) model and is widely used in various fields.

2. Material and methods

2.1. Data source and preprocessing

The data from the 2019 question D of Chinese college students' mathematical modeling is selected to build the data correction model of the micro air quality detector. Two sets of data are contained in it. The first set of data is the hourly concentration data of six pollutants ("two dusts and four gases" includes PM2.5, PM10, CO, NO_2, SO_2, O_3) from November 14, 2018 to June 11, 2019 provided by the national monitoring station. It has a total of 4200 groups and is considered accurate data in this study. The other set of data is data provided by a micro air quality detector juxtaposed with the national monitoring station. It has a total of 234,717 sets of data, and the interval between each set of data does not exceed 5 minutes. The self-built site not only provides data on the concentration of six types of pollutants, but it also provides five meteorological parameters [16]. Because the data detected by the electrochemical sensor in the micro air quality detector has errors, it needs to be corrected by the data of the national monitoring station.

Before establishing the air quality correction model, the test data must be preprocessed. The first step is to delete the data that the self-built point cannot correspond to the time of the national control point. In the second step, the measurement data that is more than 3 times the mean of the left and right neighbors are regarded as outliers and deleted. The abnormal value may be caused by a sensor failure or a sudden change in the concentration of pollutants caused by nearby human activities. The third step is to average the data in each hour of the self-built point to make it correspond to the measured data of the national control point. After data preprocessing, a total of 4135 sets of corresponding data were obtained for this study. Table 1 shows all the variables in this study.

Table 1
Descriptive statistics of air quality variables from data from national control points and self-built points.

Input variable	Ranges	Mean	Standard deviation	Skewness	Kurtosis
PM2.5/(μg/m^3)	1~216.883	64.127	37.328	0.988	0.701
PM10/(μg/m^3)	2~443.25	102.391	65.267	1.476	2.862
CO/(μg/m^3)	0.05~3.895	0.863	0.452	1.463	3.136
NO_2/(μg/m^3)	0.947~157.136	45.209	28.403	0.653	-0.259
SO_2/(μg/m^3)	1~651.3	19.397	18.723	12.781	342.11
O_3/(μg/m^3)	0.579~259	61.586	40.941	1.091	2.035
Wind speed/(m/s)	0.133~2.387	0.7	0.346	0.862	0.748
Pressure /(Pa)	996.871~1039.8	1018.8	8.889	-0.093	-0.599
Precipitation /(mm/m^2)	0~312.1	132.084	87.004	0.245	-0.728
Temperature /(℃)	-3.882~37.944	11.882	8.603	0.625	-0.399
Humidity /(rh%)	10.667~100	68.903	21.931	-0.487	-0.756

2.2. Data exploratory analysis

After preprocessing the experimental data, an exploratory analysis of the data is needed. Due to the complicated analysis of pollutant concentration data per hour, it is not conducive to visually reflect the changes in pollutant concentration. In this study, the pollutant concentrations of national control points and self-built points were averaged daily and then compared. The concentration of PM10 is selected as the research object for detailed analysis in this paper, and similar methods are used for the analysis of other pollutant concentrations.

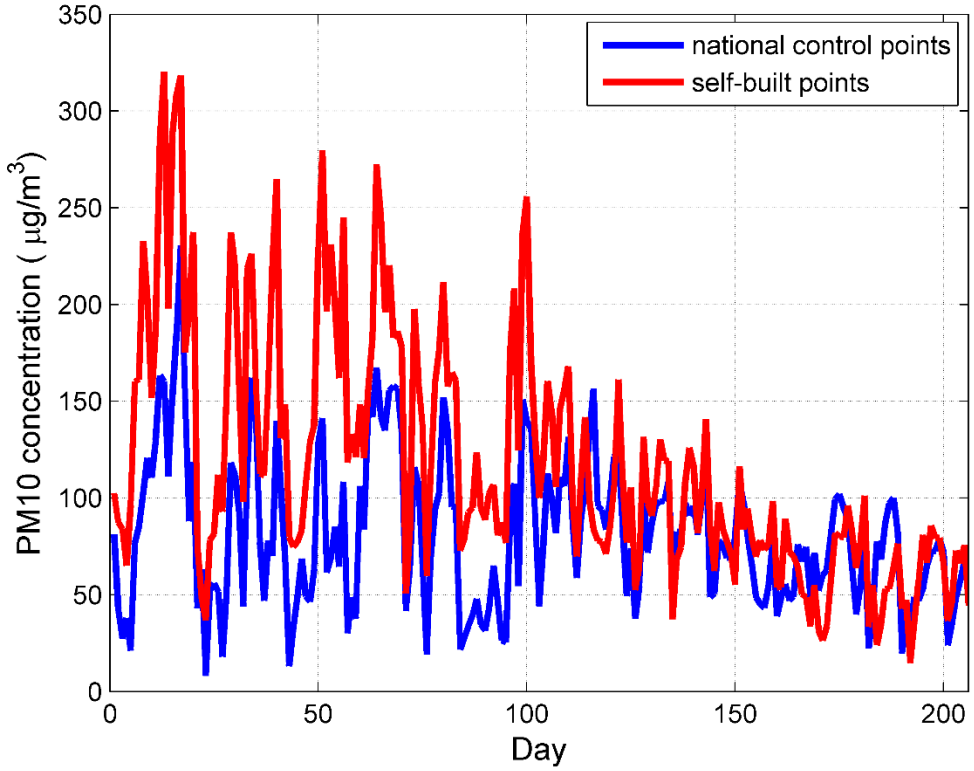

Fig.1. Comparison of daily average PM10 concentration data between national control points and self-built points. Figures are generated using Matlab (Version R2016a, https://www.mathworks.com/) [Software].

Fig. 1 shows the changes in the daily average PM10 concentration of national control points and self-built points. It can be seen that the overall trend of the two is roughly the same, but there are also certain differences. The difference between the two presents a characteristic that the greater the concentration, the more obvious the difference is, which is the same as the characteristic presented by an ordinary measuring device. The phenomenon that the PM10 concentration varies in different months is also reflected in Fig. 1. In order to intuitively reflect the changes of PM10 concentration over the months, this paper draws a PM10 concentration box plot [9, 11].

It can be seen from Fig.2 that the average monthly concentration of PM10 was the highest in November, with a concentration of 108.15μg/m^3; the monthly average concentration of PM10 was the lowest in June, with a concentration of 58.58μg/m^3 (no data from July to October). It is cold in winter and coal-fired for heating, which emits a large amount of smoke and dust is one of the main reasons for the high concentration of PM10 in winter. In addition, due to the low air humidity in winter, the weak sedimentation of suspended particles in the atmosphere is also the reason for the high concentration of PM10 in winter. In summer, the precipitation is large, which has a strong sedimentation effect on the suspended particles in the atmosphere, resulting in lower PM10 concentration in summer.

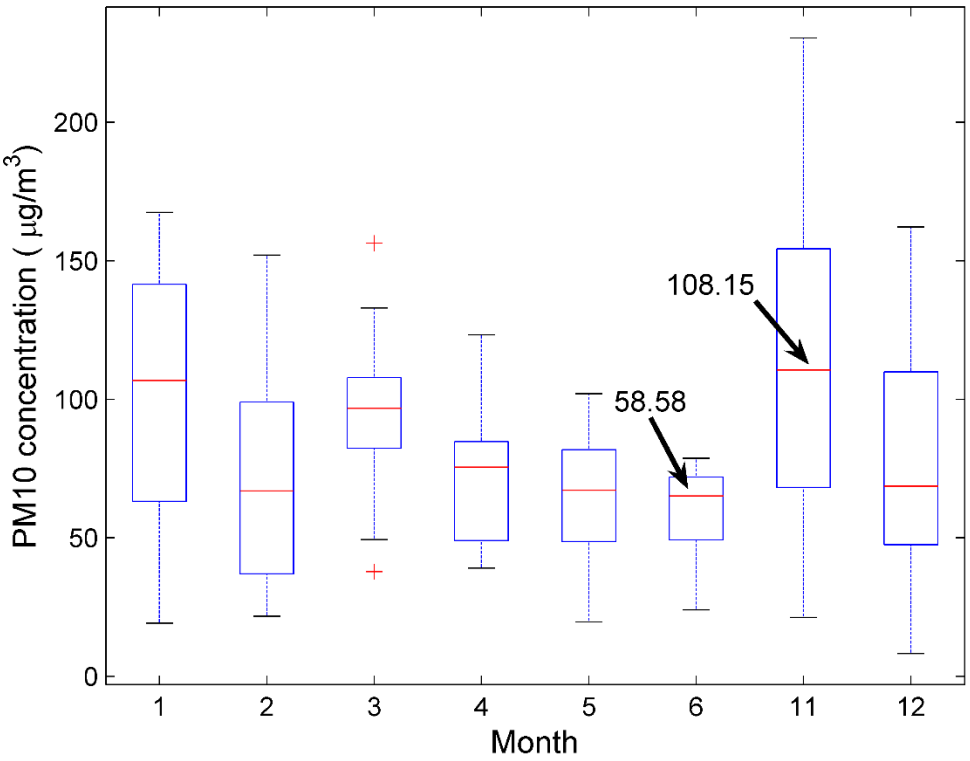

Fig.2. Compare the concentration of PM10 in national control points monthly. Note that there is no data from July to October.

2.3. Correlation analysis

Air pollutants are causing more and more harm to humans and the environment. There are many factors that affect the concentration of pollutants in the air, and there are also correlations between them. Pearson correlation coefficient (Eq. (1)) is often used as a statistical indicator to reflect the close degree of correlation between variables [21, 26]. It is used in this study to measure the correlation between six types of pollutants and five meteorological parameters. It can be seen from Table 2 that only the correlation between NO_2 concentration and temperature is not significant, and the correlation coefficients between other variables have passed the significance test. In order to visualize the correlation between variables, this paper draws a matrix color block diagram as shown in Fig. 3. The area of the circular color block in the matrix color block diagram represents the absolute value of the correlation coefficient and the darker the color, the stronger the correlation between the variables.

$$r = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\cdot\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}} \quad (1)$$

Table 2
Pearson linear correlation coefficients between six types of air pollutant concentrations and climate (Band * indicates significant correlation at a significant level of 0.05).

Variable	PM2.5	PM10	CO	NO_2	SO_2	O_3	Wind speed	Pres-sure	Precip-itation	Temp-erature	Humi-dity
PM2.5	1.00	0.89*	0.66*	0.26*	0.29*	-0.26*	-0.23*	0.89*	-0.70*	-0.16*	0.18*
PM10		1.00	0.63*	0.34*	0.35*	-0.19*	-0.18*	0.38*	-0.10*	-0.03*	-0.09*
CO			1.00	0.30*	0.31*	-0.27*	-0.31*	-0.07*	0.08*	-0.05*	0.22*
NO_2				1.00	-0.34*	-0.26*	-0.36*	-0.10*	-0.14*	-0.02	-0.11*
SO_2					1.00	-0.28*	-0.19*	0.19*	0.27*	-0.10*	0.11*
O_3						1.00	0.39*	-0.45*	-0.12*	0.68*	-0.62*
Wind speed							1.00	0.09*	0.06*	0.07*	-0.32*
Pressure								1.00	0.23*	-0.85*	0.15*
Precipitation									1.00	-0.14*	0.86*
Temperature										1.00	-0.49*
Humidity											1.00

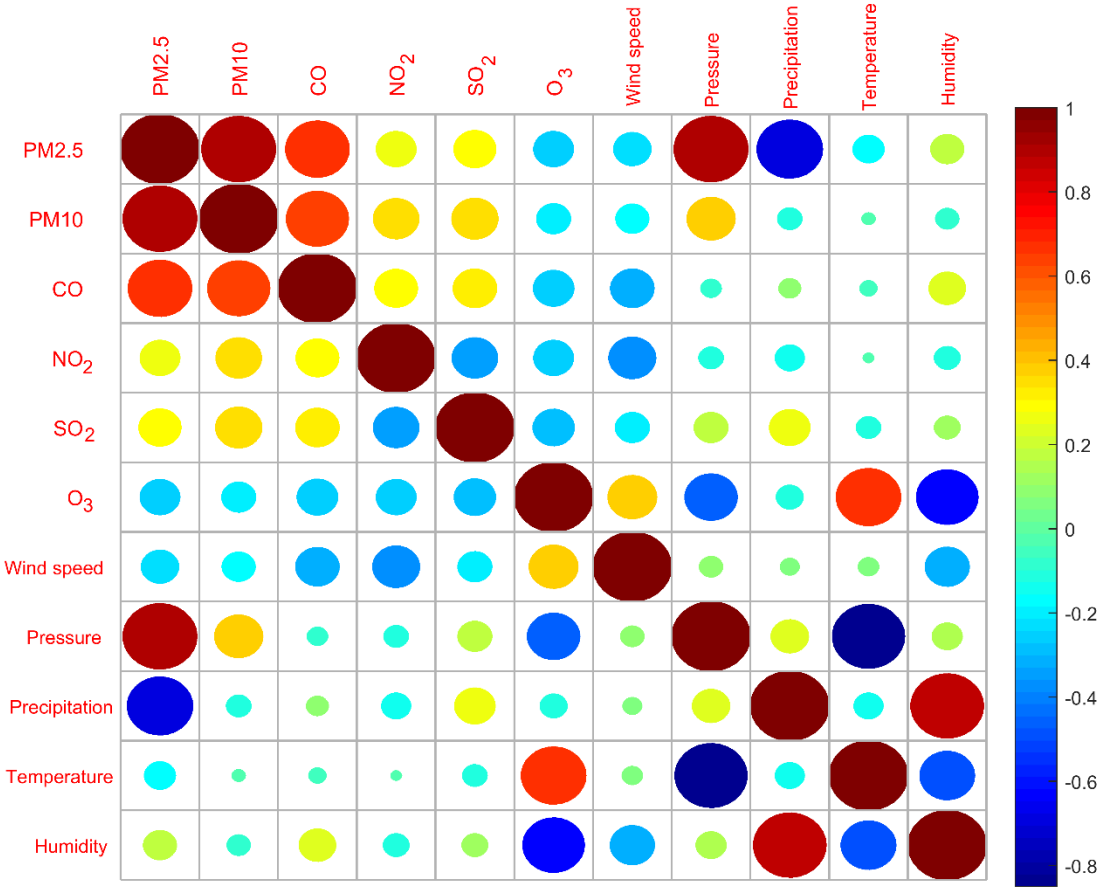

Fig.3. Matrix color block diagram of the correlation coefficient matrix between the concentration of six air pollutants and five meteorological parameters.

3. Establishment of sensor calibration model

3.1. Introduction to basic principles

The partial least squares (PLS) regression method was first proposed by Wold and Albano in 1983. In the past 40 years, it has developed rapidly in theories, methods and applications. In the research of actual problems, the regression problem often encountered is that the number of independent variables $x_1, x_2, \cdots, x_k$ is relatively large, and the number of samples

n is not large. The independent variables often have a strong correlation, which is called multicollinearity. If the multicollinearity problem is serious, the diagonal elements of the variance matrix $D(\hat{\beta}) = \sigma^2(X'X)^{-1}$ of the regression coefficient $\hat{\beta}$ are very large, that is, $Var(\hat{\beta}_0), Var(\hat{\beta}_1), \cdots, Var(\hat{\beta}_k)$ is very large. Therefore, the validity of the estimation of $\beta_0, \beta_1, \cdots, \beta_k$ is lost, and the degree of influence of the independent variable on the dependent variable cannot be correctly judged.

If the number of independent variables is more than the number of samples, the classic least square method (OLS) cannot be applied. Although the principal component regression solves the contradiction of $k > n$, its method of selecting F_i has nothing to do with the dependent variable y. It only looks for a representative $F_1, F_2, \cdots, F_r$ in the independent variable $x_1, x_2, \cdots, x_k$. PLS is different from PCR in this point. It considers the correlation with y when looking for a linear combination of $x_1, x_2, \cdots, x_k$. It chooses the latent factors $Z_1, Z_2, \cdots, Z_r$ that has a strong correlation with y and can be easily calculated [27, 28].

There are four main steps in partial least squares regression. The first step is to standardize variables. For the convenience of writing, we also mark the standardized vector $y^*$and the design matrix X^* as y and X. The following calculations are for y and X after standardization. The second step is to extract the first latent factor Z_1, where $Z_1 = XW_1$, and W_1 is shown in Eq. (2). After obtaining the first latent factor Z_1, respectively implement the regression of y and X on Z_1. That is $y = r_1 Z_1 + y_1$, where $r_1 = \frac{y'Z_1}{\|Z_1\|^2}$ is the regression coefficient and y_1 is the residual vector; $X = Z_1 P_1' + X_1$, where $P_1 = \frac{X'Z_1}{\|Z_1\|^2}$ is the regression coefficient vector, and X_1 is the residual matrix. The third step is loop iteration. Replace y and X with the residual vector y_1 and the residual matrix X_1, and continue to use the same method as the first latent factor Z_1 to obtain the remaining latent factors $Z_1, Z_2, \cdots, Z_a$. The fourth step is to determine the number of latent factors a. The number of latent factors can be determined by cross-validation. Let $Q_a^2 = 1 - \frac{PRESS_a}{SS_{a-1}}$, where $SS_a = \sum_{i=1}^{n}(y_i - \hat{y}_{ai})^2$, $PRESS_a = \sum_{i=1}^{n}(y_i - \hat{y}_{a(-i)})^2$, where $\hat{y}_{ai}$ is the fitted value of y_i using the first a potential factors $Z_1, Z_2, \cdots, Z_a$ for regression calculation. In addition, $\hat{y}_{a(-i)}$ is the fitted value of y_i which is calculated by removing the *i*-th sample point, using $n-1$ sample points and a latent factors. It is generally stipulated that when $Q_{a+1}^2 < 1 - 0.95^2 = 0.0975$, the introduction of a new latent factor Z_{a+1} does not make much sense for the prediction of the model. At this time, the number of comprehensive variables is taken as a.

$$W_1 = \frac{1}{\sqrt{\sum_{i=1}^{k} Cov^2(y, x_i)}} \begin{bmatrix} Cov(y, x_1) \\ \vdots \\ Cov(y, x_k) \end{bmatrix} \tag{2}$$

The random forest algorithm was proposed by Breman in 2001. It is a combination model based on classification regression tree. The random forest regression algorithm uses bootstrap resampling technology. It extracts a certain amount of samples from the original samples with replacement, and uses them as a new training sample set for decision tree modeling, thereby combining them into a multiple decision tree model for regression prediction. The structure framework of the random forest regression prediction model is shown in Fig. 4. Suppose *S*=(*X, Y*) is the original data set, where *X* is the original input data and *Y* is the original output data. Randomly sample k original data from S (the sample size of each sample is consistent with the original data set) to form a training sample set. A set of *k* decision trees $h(X, \theta_k), k = 1,2, \cdots, K$ is generated through training samples to form a random forest, where *X* is the input vector and θ_k is the random vector used by the *k*-th decision tree to select sample points. Each decision tree produces a prediction value y_k, and the final prediction result is the average of *k* regression prediction results.

Assuming that the original data set contains n records, the probability of each sample not being selected is $(1-1/n)^n$. When n is large enough, $(1-1/n)^n$ converges to $1/\mathrm{e} \approx 0.368$, which means that only $2/3$ of the original data is selected to generate the training sample set. The unselected data samples can be used as test data to evaluate the generalization performance of the regression tree [22]. This part of the data is called out of bag data (OOB). Assuming that the training set is drawn from the independent distribution set of random variables X and Y, the mean square generalization error of the random forest predicted value is $E_{XY} = [Y - \hat{y}(X)]^2$.

Each classification tree in the random forest is a binary tree, and its generation follows the principle of top-down recursive splitting, that is, the training set is divided sequentially from the root node. The root node contains all training data, split into left and right nodes, which contain a subset of training data respectively. Splitting is carried out in accordance with the principle of minimum node impurity, until the stopping rule is met and the growth stops. The measure of impurity generally follows the Gini criterion. Assuming that the data set H contains m categories, and P_j is the frequency of the occurrence of j-type elements, the Gini index is $G_H = 1 - \sum_{j=1}^{m} P_j^2$.

Random forest method can measure the importance of variables. Randomly change the value of a characteristic variable (increase the noise), and use the generated random forest model to calculate the OOB fitting error. The greater the increase in OOB error after the characteristic variable is added to the noise, the higher the importance of the variable. Suppose the OOB test error of variable X is e, and the trial error of the OOB data after adding noise is e', then the importance of X is $\phi = (e' - e)/e$.

There are two main parameters that the random forest algorithm needs to determine, and they have a great influence on the prediction effect of the model [23, 29]. One of the parameters is the number of decision trees ntree. Only when the number of trees is large enough, the error of the model can stabilize, but the number of numbers should not be too much, otherwise there will be overfitting, and the error will increase instead. The second parameter is the leaf node mtry, which is the number of variables used for splitting randomly selected features from all feature sets. For regression prediction, generally mtry = s/3 (s is the total number of feature variables). In order to obtain the best mtry value, it can also be determined by calculating the OOB error.

Random forest regression increases the difference between regression models by constructing different training sets, thereby improving the extrapolation prediction ability of the combined regression model. Suppose that after K rounds of training $h(X, \theta_1), h(X, \theta_2), \cdots, h(X, \theta_k)$ decision trees are obtained, their corresponding regression prediction values are$\{ y_1(X), y_2(X), \cdots, y_k(X)\}$. These training sequences construct a combined prediction model, and the final regression prediction result is $\overline{y}(X) = (1/k) \sum_{k=1}^{K} y_k(X)$.

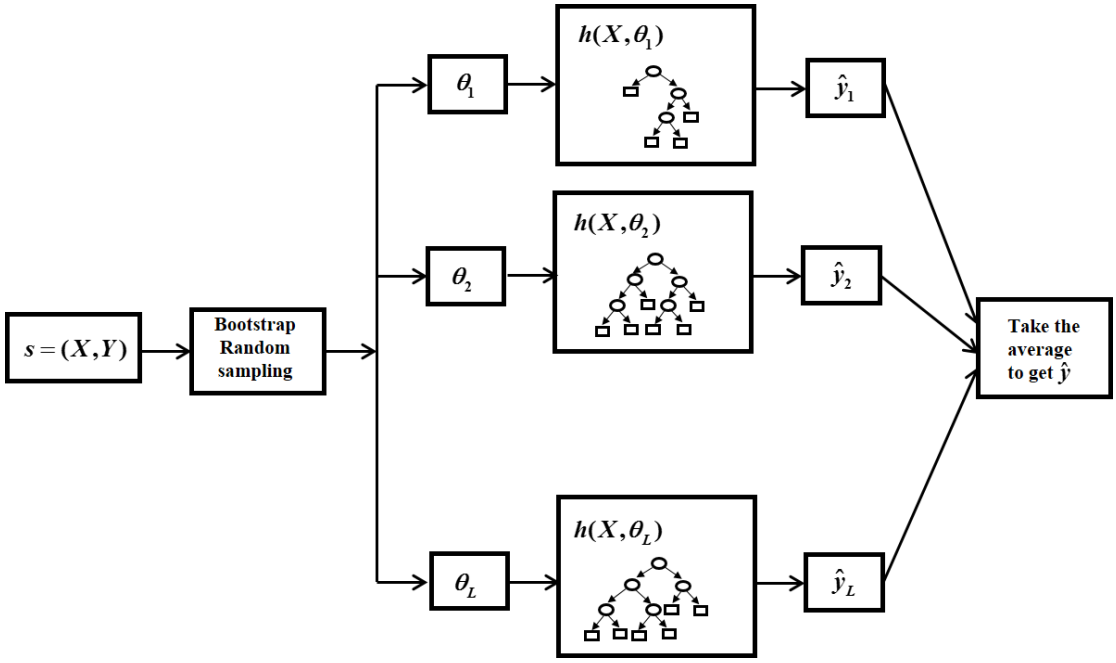

Fig.4. Frame structure of random forest regression prediction model.

3.2. Partial least squares regression model construction

It is of great significance to use micro air quality detectors to monitor the concentration of pollutants in grids and in real time. However, since the sensor in the micro air quality detector is easily affected by other factors, it is necessary to calibrate its measurement data.

The classic multiple linear regression model is often used to explain the quantitative relationship between the dependent variable and the independent variable [30]. Correlation analysis shows that the various variables that affect the concentration of pollutants affect each other. Through the diagnosis of multicollinearity, it is found that the maximum variance inflation factor (VIF) between them is 28.71, which is much greater than 10, indicating that there is serious multicollinearity between variables. The classic multiple linear regression model is no longer applicable. Principal component regression and partial least squares regression can be used to solve the multicollinearity problem in the model [10, 31]. However, because principal component regression did not take into account the value of the dependent variable when extracting principal components, this study uses partial least squares regression for modeling. In this study, the pollutant concentration at the national control point was used as the dependent variable *y*, the pollutant concentration at the self-built point and the meteorological parameter monitoring data were used as the independent variable *X*, and a partial least square regression model was established with the help of SPSS20.0.

Table 3
The proportion of variance explained by the first seven latent factors.

Latent Factors	*X* Variance	Cumulative *X* Variance	*Y* Variance	Cumulative *Y* Variance	Adjusted *R-square*
Z_1	0.254	0.254	0.577	0.577	0.577
Z_2	0.183	0.437	0.163	0.74	0.74
Z_3	0.121	0.559	0.052	0.792	0.792
Z_4	0.083	0.641	0.008	0.8	0.8
Z_5	0.098	0.739	0.003	0.803	0.803
Z_6	0.052	0.792	0.004	0.807	0.807
Z_7	0.047	0.839	0.002	0.809	0.808

It can be seen from Table 3 that the first 7 latent factors already contain more than 80% of the independent variable and dependent variable information, and the adjusted R^2=0.808, combined with the principle of selecting latent factors, this article decided to select 7 latent factors. Columns 2-8 of the second row give the regression results of the standardized dependent variable on the latent factors. The remaining rows in columns 2-8 are the linear combination results of the latent factors with respect to the standardized independent variables. It can be seen that for factor Z_1, PM2.5*, PM10* and CO* have the largest weights; for factor Z_2, PM2.5*, PM10* and Humidity* are the largest weights. For other factors, the weight of its influencing factors can be discussed similarly. The last two columns of Table 4 give the regression equations between the predicted PM10 concentration and the monitoring data of self-built points. Using this regression equation, the predicted value of PM10 concentration in the PLS model can be obtained.

Table 4
Coefficient estimation, cumulative variable importance and factor weight results (The variable name band * indicates the corresponding standardized variable).

Input variable	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	y	coefficient
y^*	0.475	0.303	0.199	0.098	0.082	0.092	0.066	constant	1315.199
$PM2.5^*$	2.135	1.991	1.951	1.94	1.937	1.933	1.933	PM2.5	0.632
$PM10^*$	2.043	1.854	1.802	1.793	1.791	1.789	1.787	PM10	0.197
CO^*	1.083	0.96	1.017	1.012	1.01	1.012	1.011	CO	20.798
NO_2^*	0.995	0.879	0.854	0.881	0.88	0.88	0.879	NO_2	0.344
SO_2^*	0.417	0.372	0.366	0.364	0.363	0.364	0.364	SO_2	0.089
O_3^*	0.498	0.473	0.557	0.556	0.561	0.561	0.561	O_3	-0.044
Wind speed*	0.526	0.499	0.534	0.532	0.534	0.533	0.533	Wind speed	-0.888
Pressure *	0.108	0.447	0.493	0.491	0.508	0.572	0.587	Pressure	-1.209
Precipitation *	0.281	0.671	0.69	0.702	0.702	0.7	0.703	Precipitation	-0.071
Temperature*	0.097	0.581	0.611	0.609	0.613	0.641	0.65	Temperature	-0.665
Humidity *	0.265	1.159	1.273	1.288	1.287	1.294	1.293	Humidity	-1.133
Time*	0.488	0.697	0.674	0.671	0.673	0.707	0.706	Time	-0.003

3.3. PLS-RFR model construction

Using the PLS model can not only predict the concentration of pollutants, but also show the quantitative relationship between independent variables and dependent variables. However, the factors that affect the concentration of pollutants are very complex, and they have a nonlinear effect on the concentration of pollutants. Random forest regression model is used to find the nonlinear relationship between pollutant concentration and various factors. In the random forest regression model, the pollutant concentration is the dependent variable, and the self-built point measurement value and the PLS model prediction value are the independent variables. We call the model combining partial least squares regression and random forest regression the PLS-RFR model. The specific modeling process is shown in Fig. 5.

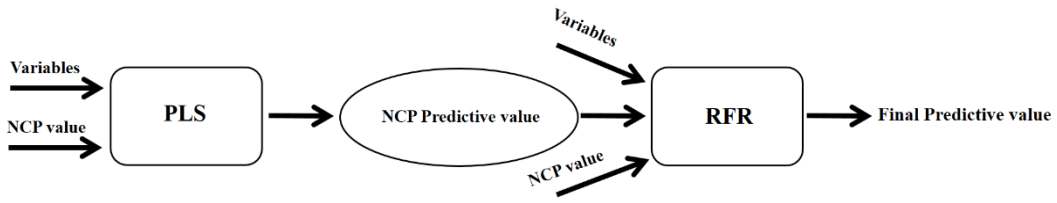

Fig.5. The flux diagram of the regression process, where ncp represents the concentration of pollutants measured at the national control point.

In the random forest regression model, 4135 samples are randomly divided into training samples and testing samples at a ratio of 8:2. There are 4135 sets of sample data in this study, the training sample set is 3309 sets, and the test sample set is 826 sets. In the regression model, the number of leaf nodes mtry generally defaults to 1/3 of the number of variables in the data set. The pollutant concentration prediction model contains a total of 13 influencing factors, so the number of leaf nodes mtry = 4. According to theoretical research, the generalization error of the random forest regression model will gradually converge as ntree increases, so a larger ntree value can be selected to build the model. Using OOB error rate estimation method to determine the ntree value, the result is shown in Fig. 6. It can be seen that when the number of decision trees ntree >100, the OOB error rate tends to stabilize, so ntree =180 is selected in this study.

Through importance evaluation, the importance of influencing factors to the concentration of pollutants can be determined. The Matlab random forest software package is used to realize the importance score calculation process, and the importance scores of different influencing factors are arranged in descending order to obtain the importance ranking of each variable in the training model [24, 25]. It can be seen from Fig. 6 that the PLS predictive value has the largest importance metric, indicating that the partial least squares model plays an important role in the prediction of pollutants. The PM10 concentration importance metric is the smallest, because the main contribution of the PM10 concentration measurement data at the self-built point has been included in the partial least squares model, which also reflects the correctness of the partial least squares model.

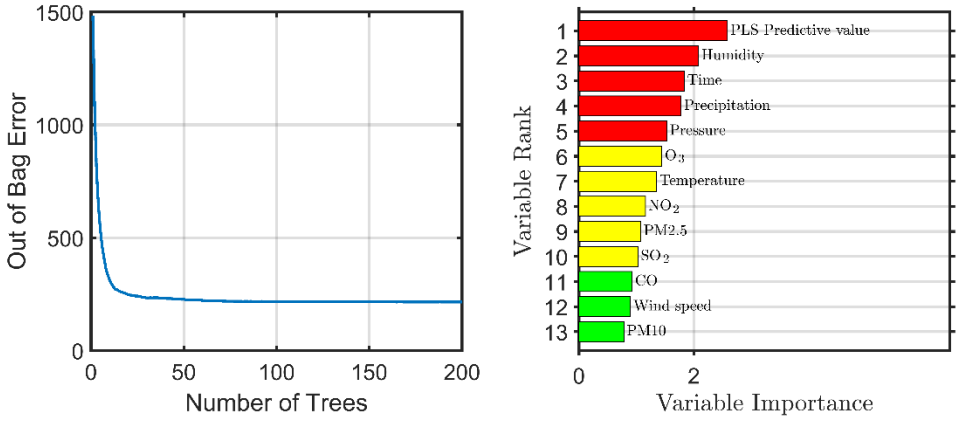

Fig.6. Random forest regression modeling process. The number of decision tree choices is seen on the left. The order of importance of the variables is seen on the right.

In this study, the concentration of pollutants was predicted by the partial least squares model, and the prediction results and self-built point measurement data were used as independent variables, combined with the random forest regression model to construct the PLS-RFR model. Fig. 7 shows the fitting results of the PLS-RFR model in the training set and the test set. It can be seen that the correlation between the predicted value of the pollutant concentration and the true value of the national control point in both the training set and the test set exceeds 0.95. Moreover, the linear regression coefficient of the predicted value of pollutant concentration and the true value of the national control point are close to 1, indicating that the PLS-RFR model has achieved good results in predicting the concentration of pollutants.

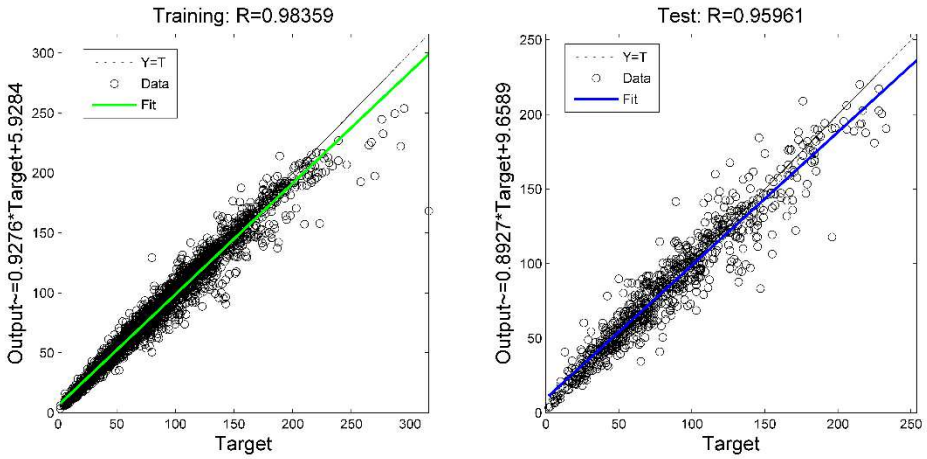

Fig.7. The prediction effect of PM10's PLS-RFR model on the training set and test set.

In order to visually show the deviation between the predicted value of the pollutant concentration and the true value of the national control point, the residual analysis diagram is drawn as Fig. 8. It can be seen that most of the residual values are randomly distributed in [-50, 50]. The 1412th residual is the largest among all samples, which is 147.73μg/m^3. Checking the original data, the true value of PM10 concentration at this point is 316μg/m^3, which is much higher than the neighboring values, indicating that human activities have a great impact on the pollutant concentration during this period. In order to visually display the residual distribution of the model, this paper deletes the residual of this point and draws the residual histogram of other samples. It can be seen that the overall residuals are roughly distributed normally, and most of the sample residuals are near zero.

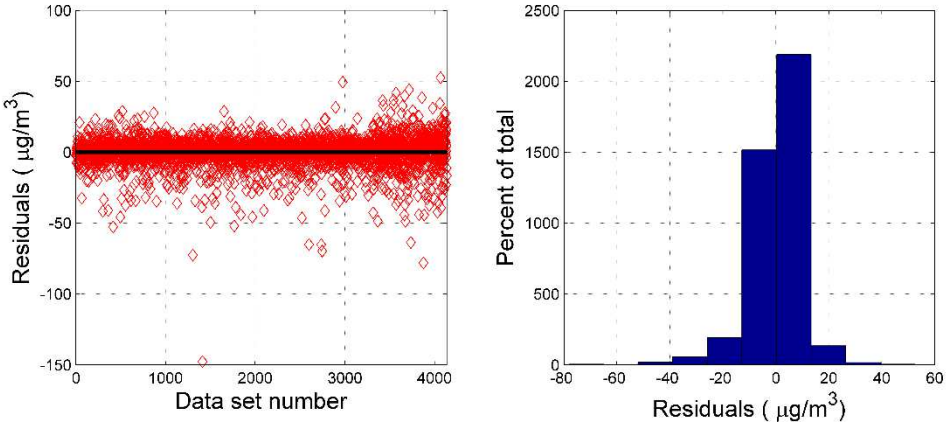

Fig.8. Residual test of PLS-RFR model. The residuals vs. data set number plot is seen on the left. The histogram of the residuals is seen on the right.

4. Discussion

In order to further verify the reliability of the PLS-RFR model, support vector machines (SVR) and multilayer perceptron (MLP) neural networks were selected to predict the concentration of pollutants, and the prediction results were compared with the PLS-RFR model. This study uses relative Mean Absolute Percent Error (MAPE), Mean Absolute Error (MAE), goodness of fit (R^2), and Root Mean Square Error (RMSE) to compare the prediction results of the model. R^2 can measure the degree of fit of the model, and the closer to 1, the better the degree of model fit. The closer the other three indicators are to 0, the smaller the error between the predicted value and the true value. The specific formula is Eqs. (3)–(6), where y_t is the real value measured at the national control point, w_t is the predicted value of the model, and $\bar{y}$ is the average value of the measured data at the national control point.

It can be seen from Table 5-8 that no matter which evaluation index, the measurement data error of the self-built point is the largest, and several other models can be used to correct it. It should be noted that the negative value of the goodness of fit of the self-built points is caused by the large measurement error of the self-built points. The traditional partial least squares regression model can improve the measurement error of self-built points, but the effect is very ordinary. The correction effect of multilayer perceptron neural network and support vector machine on self-built point data is better than that of partial least square regression. The random forest regression model works well, but the PLS-RFR model has the best predictive effect in the six pollutant concentrations regardless of the evaluation index.

Air quality has a strong correlation with human activities. Human activities are cyclical. This article selects one week as a cycle, and plots the data of PM10 national control points, self-built points and forecast points into a line chart [32, 33]. It can be seen from Fig. 9 that the red self-built point curve and the blue national control point curve have a certain deviation, indicating that the data monitoring accuracy of the micro air quality detector needs to be improved. The black model fitting curve almost coincides with the blue national control point curve, indicating that the PLS-RFR model performs better in predicting the concentration of pollutants. Using this model to calibrate the detection data of the micro air quality detector can achieve better results.

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t - w_t}{y_t}\right| \quad (3)$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t - w_t| \quad (4)$$

$$R^2 = 1 - \frac{\sum_{t=1}^{n}(y_t - w_t)^2}{\sum_{t=1}^{n}(y_t - \bar{y})^2} \quad (5)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - w_t)^2} \quad (6)$$

Table 5
MAPE of six types of air pollutant concentrations between self-built points, model forecast values and national control point.

Input variable	Self-built points	PLS	PLS-RFR	RFR	SVR	MLP
PM2.5	0.447	0.171	0.081	0.087	0.133	0.185
PM10	0.887	0.222	0.087	0.095	0.107	0.210
CO	0.478	0.313	0.079	0.083	0.112	0.283
NO_2	2.129	0.566	0.115	0.121	0.170	0.471
SO_2	0.685	0.660	0.110	0.115	0.131	0.530
O_3	4.322	1.134	0.279	0.304	0.373	1.002

Table 6
MAE of six types of air pollutant concentrations between self-built points, model forecast values and national control point.

Input variable	Self-built points	PLS	PLS-RFR	RFR	SVR	MLP
PM2.5	18.181	7.189	3.380	3.485	5.821	7.763
PM10	50.151	13.747	5.964	6.299	7.080	13.184
CO	0.549	0.261	0.075	0.079	0.110	0.237
NO_2	29.838	11.702	3.421	3.515	4.658	9.991
SO_2	12.867	9.427	1.680	1.736	2.116	7.246
O_3	36.63	16.073	5.439	5.638	7.647	14.396

Table 7
R^2 of six types of air pollutant concentrations between self-built points, model forecast values and national control point.

Input variable	Self-built points	PLS	PLS-RFR	RFR	SVR	MLP
PM2.5	0.551	0.905	0.976	0.976	0.933	0.907
PM10	-1.076	0.809	0.957	0.953	0.938	0.827
CO	-0.929	0.508	0.938	0.932	0.872	0.708
NO_2	-1.333	0.601	0.944	0.942	0.899	0.752
SO_2	-0.726	0.558	0.970	0.969	0.958	0.786
O_3	0.094	0.810	0.969	0.969	0.945	0.864

Table 8
RMSE of six types of air pollutant concentrations between self-built points, model forecast values and national control point.

Input variable	Self-built points	PLS	PLS-RFR	RFR	SVR	MLP
PM2.5	22.436	10.312	5.169	5.207	8.649	10.777
PM10	66.263	20.110	9.562	9.940	11.656	19.126
CO	0.679	0.343	0.122	0.128	0.175	0.304
NO_2	37.183	15.375	5.766	5.847	7.725	13.216
SO_2	26.24	13.282	3.454	3.513	4.116	9.984
O_3	45.673	20.912	8.421	8.433	11.304	18.603

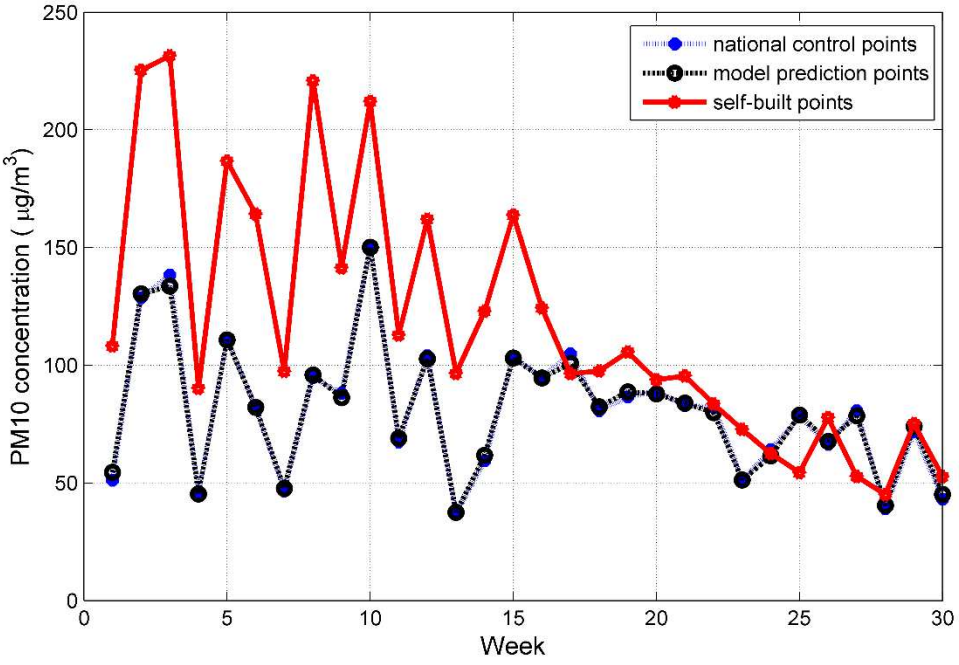

Fig.9. Comparison of the weekly average concentration of PM10 between national control points, PLS-RFR model calibration points and self-built points.

5. Conclusions

The air quality is judged by the concentration of pollutants in the air. Air pollution is a complex phenomenon, and the concentration of air pollutants at a specific time and place is affected by many factors. As a country's industry develops, its air pollution often becomes more and more serious. Severe air pollution not only greatly restricts visibility, but also seriously affects human health [1, 3]. Government departments and the people are generally

more and more concerned about the consequences of such severe air pollution. Therefore, it is very necessary to monitor air quality.

Many countries have realized the harm that air pollution brings to human society. They have established some national monitoring stations to monitor air quality and take corresponding measures to deal with air pollution based on the monitoring data. However, due to the high cost of construction and maintenance of national monitoring stations, it is difficult to realize real-time monitoring and grid monitoring of pollutant concentrations. Electrochemical sensors are used in micro air quality detectors, which greatly reduces the cost of air quality monitoring. The micro air quality detector is not only convenient for deployment and control, but also can release data in real time, which is helpful for government departments to take timely response measures to pollution sources [34]. However, electrochemical sensors are susceptible to interference from themselves and other factors, leading to certain deviations in monitoring data.

The pollutant correction model using partial least squares regression can correct the measurement data of the micro air quality detector. Its advantage is that it can give the quantitative relationship of each influencing factor to the concentration of pollutants, but the disadvantage is that the correction effect is not particularly good. Compared with partial least squares regression, the random forest regression model has a better advantage in data correction. The random forest regression model not only has a higher fitting effect than other algorithms, but also has a short training time and strong anti-overfitting ability. But like other machine learning algorithms, the random forest regression model cannot give the quantitative relationship between the various influencing factors and the concentration of pollutants.

The PLS-RFR model combining the partial least squares regression model and the random forest regression model given in this study combines the advantages of the two models and greatly improves the prediction of the six types of pollutant concentrations. The data of 4 seasons throughout the year are all included in the model, and the model time span is 206 days. 4135 sets of data are included in the model, and the accuracy of the model is very high regardless of the training set or the test set. All this shows that the model performs well in terms of stability and generalization. This model can play an active role in the promotion and deployment of micro air quality detectors. In future research, we can consider using a time series model to extract the information contained in the predicted residuals to further improve the prediction accuracy of the model.

Acknowledgements

This work was supported by the Youth Program of National Natural Science Foundation of China (no.71602051) and Key Scientific Research Fund Project of Nanjing Vocational University of Industry Technology (no. 901050617YK002).

Author information

Affiliations

1. Public Foundational Courses Department, Nanjing Vocational University of Industry Technology, Nanjing 210023, China

Bing Liu, Yishu Wang & Qibao Lv

2. School of Mechanical Engineering, Nanjing Vocational University of Industry Technology, Nanjing 210023, China

Wangwang Yu

3. College of Management, Henan University of Technology, Zhengzhou 450001, China

Chaoyang Li

Corresponding author

Correspondence to Bing Liu, E-mail address:Liub1@niit.edu.cn.

Author contributions

B.L., W.Y., Y.W. and Q.L. wrote the main manuscript text, and C.L. is responsible for data processing and model verification.

References

[1] Poloniecki, J. D., Atkinson, R. W., Deleon A. P. & Anderson, H. R. Daily time series for cardiovascular hospital admissions and previous day's air pollution in London, UK. *OCCUP. ENVIRON. MED.* **54**, 535-540 (1997).

[2] Akimoto, H. Global Air Quality and Pollution. *Science*, **302**, 1716-1719 (2004).

[3] Brauer, M. et al. Exposure Assessment for Estimation of the Global Burden of Disease Attributable to Outdoor Air Pollution. *ENVIRON. SCI. TECHNOL.* **46**, 652-660 (2012).

[4] Spinelle, L., Gerboles, M., Villani, M. G., Aleixandre, M. & Bonavitacola, F. Field calibration of a cluster of low-cost available sensors for air quality monitoring. part A: Ozone and nitrogen dioxide. *SENSOR. ACTUAT. B-CHEM.* **215**, 249-257 (2015).

[5] Masson, N., Piedrahita, R. & Hannigan, M. Approach for quantification of metal oxide type semiconductor gas sensors used for ambient air quality monitoring. *SENSOR. ACTUAT. B-CHEM*. **208**, 339-345 (2015).

[6] Zhang, L., Liu, Y. & Zhao, F. Singular value decomposition analysis of spatial relationships between monthly weather and air pollution index in China . *STOCH. ENV. RES. RISK A.* **3**, 733-748 (2018).

[7] Azid, A. et al. Assessing Indoor Air Quality Using Chemometric Models. *POL. J. ENVIRON. STUD.* **6**, 2443-2450 (2018).

[8] Tai, A. P. K., Mickley, L. J. & Jacob, D. J. Correlations between fine particulate matter (PM2.5) and meteorological variables in the United States: Implications for the sensitivity of PM2.5 to climate change. *ATMOS. ENVIRON.* **44**, 3976-3984 (2010).

[9] Lei, M. T., Monjardino, J., Mendes, L. & Ferreira, F. Macao air quality forecast using statistical methods. *AIR. QUAL. ATMOS. HLTH.* **2,** 249-258 (2019).

[10] Ayers, G. P. Comment on regression analysis of air quality data. *ATMOS. ENVIRON.* **35**, 2423-2425 (2001).

[11] Huang, Z. & Zhang, R. Efficient estimation of adaptive varying-coefficient partially linear regression model. *STAT. PROBABIL LETT.* **79,** 943-952 (2009).

[12] Elangasinghe, M. A., Singhal, N. , Dirks, K. N. , Salmond, J. A. , & Samarasinghe, S. Complex time series analysis of PM10 and PM2.5 for a coastal site using artificial neural network modelling and k-means clustering. *ATMOS. ENVIRON.* **94**, 106-116 (2014).

[13] Dong, M. et al. PM2.5 concentration prediction using hidden semi-Markov model-based times series data mining. *EXPERT. SYST. APPL.* **36**, 9046-9055 (2009).

[14] Dun, M., Xu, Z., Chen, Y., & Wu, L. Short-term air quality prediction based on fractional grey linear regression and support vector machine. *MATH. PROBL. ENG.* **2020**, 1-13(2020).

[15] Sheng, J. et al. Prediction of dust fall concentrations in urban atmospheric environment through support vector regression. *J. CENT. SOUTH UNIV.* **17**, 307-315

(2010).
[16] Liu, B., Jin, Y. & Li, C. Analysis and prediction of air quality in Nanjing from autumn 2018 to summer 2019 using PCR-SVR-ARMA combined model. *SCI. REP-UK.* https://doi.org/10.1038/s41598-020-79462-0.
[17] Li, Y. & Tao, Y. Daily PM10 concentration forecasting based on multiscale fusion support vector regression. *J. INTELL. FUZZY SYST.* **6**, 3833-3844 (2018).
[18] A. Suárez Sánchez, P. J. García Nieto, P. Riesgo Fernández, J. J. del Coz Díaz, & F. J. Iglesias-Rodríguez. Application of an SVM-based regression model to the air quality study at local scale in the Avilés urban area (Spain). *MATH. COMPUT. MODEL.* **54,** 1453-1466 (2011).
[19] Wang, Z., Feng, J., Fu, Q. & Gao, S. Quality control of online monitoring data of air pollutants using artificial neural networks. *AIR QUAL. ATMOS. HLTH*. **12**, 1189-1196 (2019).
[20] Samia, A., Kaouther, N. & Abdelwahed, T. A Hybrid ARIMA and Artificial Neural Networks Model to Forecast Air Quality in Urban Areas: Case of Tunisia. *ADV. MATER.* **518,** 2969-2979 (2012).
[21] Liu, B., Zhao, Q., Jin, Y., Shen, J. & Li, C. Application of combined model of stepwise regression analysis and artificial neural network in data calibration of miniature air quality detector. *SCI. REP-UK.* https://doi.org/10.1038/s41598-021-82871-4.
[22] Zimmerman, N. et al. A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring. *ATMOS. MEAS. TECH.* **11,** 291-313 (2018).
[23] Nguyen, H. & Bui, X. N. Predicting blast-induced air overpressure: a robust artificial intelligence system based on artificial neural networks and random forest. *NAT. RESOUR. RES.* **3**, 893-907 (2019).
[24] Joanna A. Kamińska. The use of random forests in modelling short-term air pollution effects based on traffic and meteorological conditions: a case study in wrocaw. *J. ENVIRON. MANAGE.* **217**, 164-174 (2018).
[25] Yu, R., Yang, Y., Yang, L., Han, G. & Oguti, M. RAQ–A Random Forest Approach for Predicting Air Quality in Urban Sensing Systems. *Sensors.* **16**, 86-104 (2016).
[26] Song, Z., Deng, Q. & Ren, Z. Correlation and principal component regression analysis for studying air quality and meteorological elements in Wuhan, China. *ENVIRON. PROG. SUSTAIN.* **39**, 1- 11 (2020).
[27] Beyaztas, U. & Shang, H. L. On function-on-function regression: partial least squaresapproach. *ENVIRON. ECOL. STAT.* **1,** 95-114(2020).
[28] Camarrone, F. & Hulle, M. M. V.(2018). Fast multiway partial least squares regression. *IEEE T. BIO-MED. ENG.* **2,** 433-443 (2019).
[29] Ding, H. J., Liu, J. Y., Zhang, C. M. & Wang, Q. Predicting optimal parameters with random forest for quantum key distribution. *QUANTUM INF. PROCESS.* **2,** 1-8 (2020).
[30] Cordero, José María, Borge, Rafael & Narros, Adolfo. Using statistical methods to carry out in field calibrations of low cost air quality sensors. *SENSOR. ACTUAT. B-CHEM.* **267**, 245-254 (2018).
[31] Wu, Q. & Lin, H. A novel optimal-hybrid model for daily air quality index prediction considering air pollutant factors. *SCI. TOTAL ENVIRON.* **683**, 808-821 (2019).
[32] Wang, X. & Lu, W. Seasonal variation of air pollution index: Hong kong case study. *Chemosphere,* **63**, 1261-1272 (2006).
[33] Liu, Q., Liu, Y., Yang, Z., Zhang, T. & Zhong, Z. Daily variations of chemical properties in airborne particulate matter during a high pollution winter episode in beijing. *Acta Sci. Circumst.* **34**, 12-18 (2014).
[34] Castell, N. et al. Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates? *ENVIRON. INT.* **99**,293-302 (2017).

Figures

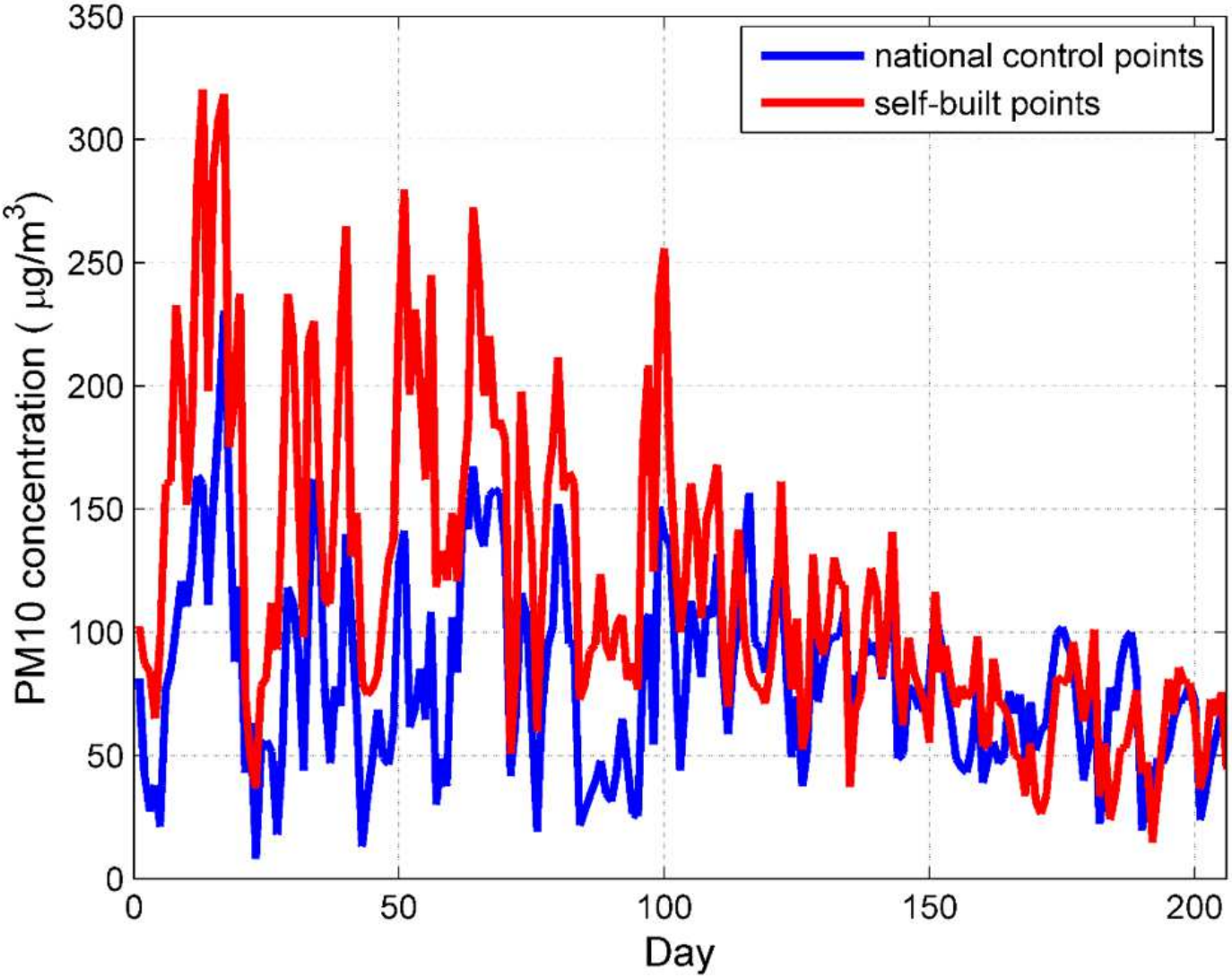

Figure 1

Comparison of daily average PM10 concentration data between national control points and self-built points. Figures are generated using Matlab (Version R2016a, https://www.mat- hworks.com/) [Software].

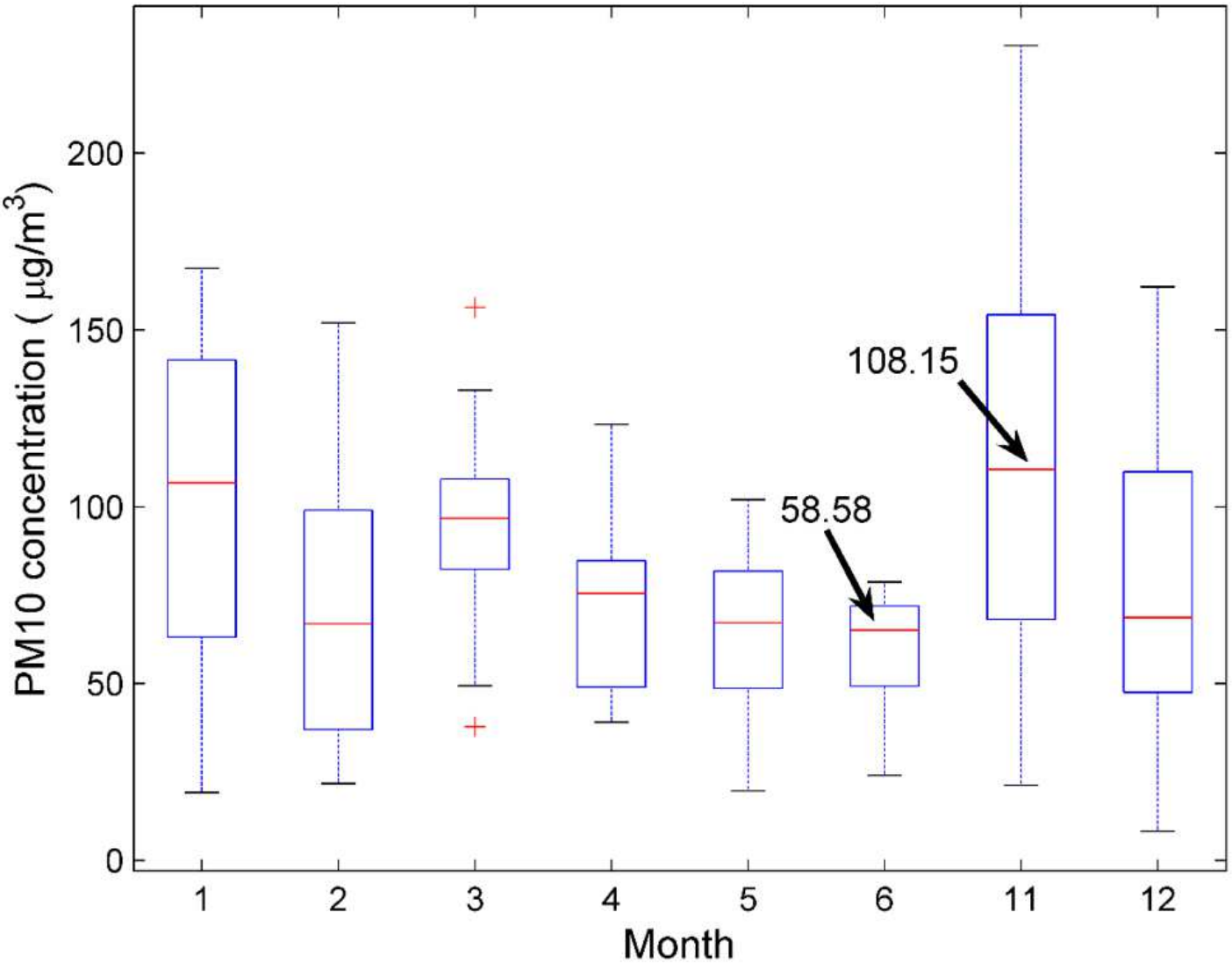

Figure 2

Compare the concentration of PM10 in national control points monthly. Note that there is no data from July to October.

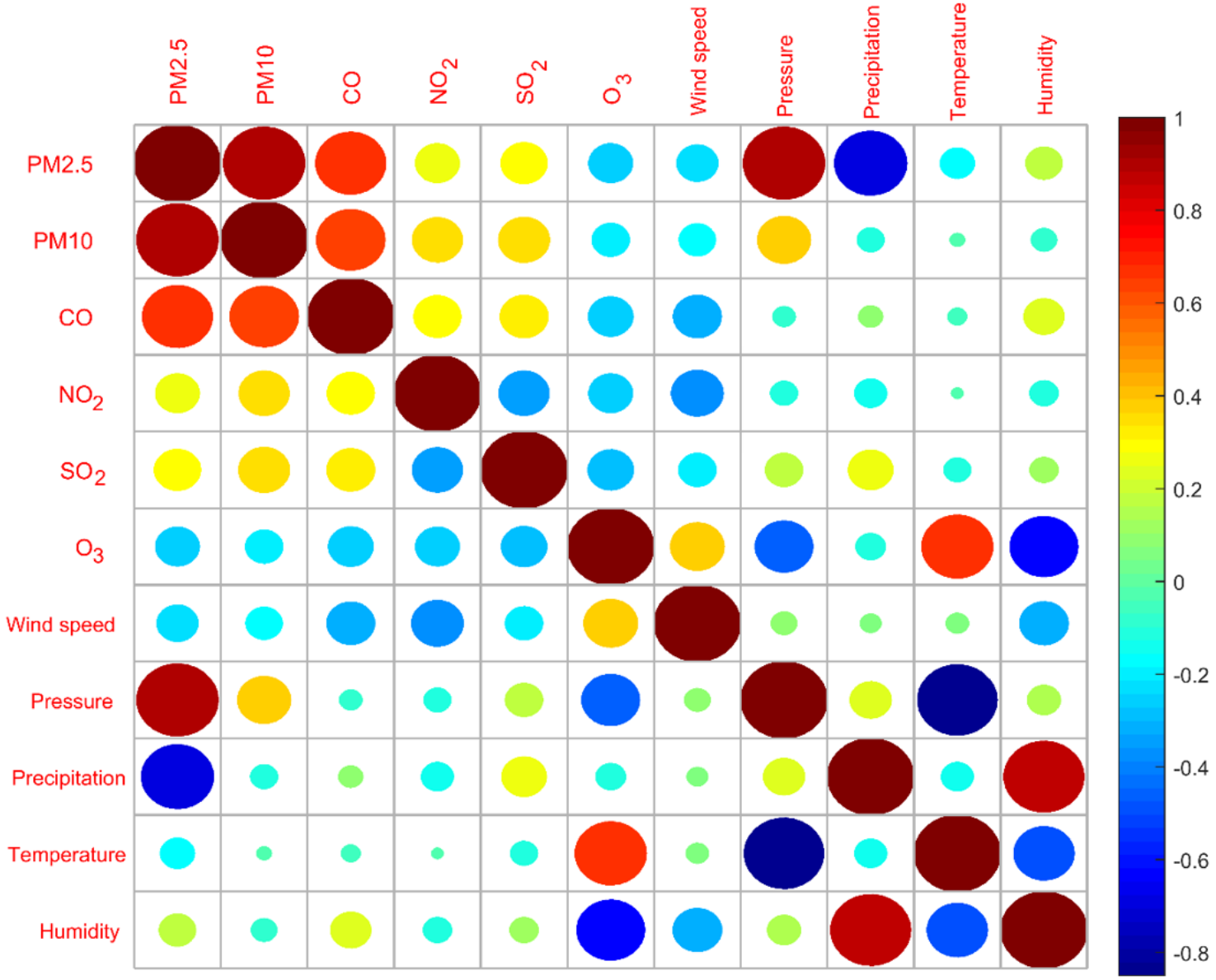

Figure 3

Matrix color block diagram of the correlation coefficient matrix between the concentration of six air pollutants and five meteorological parameters.

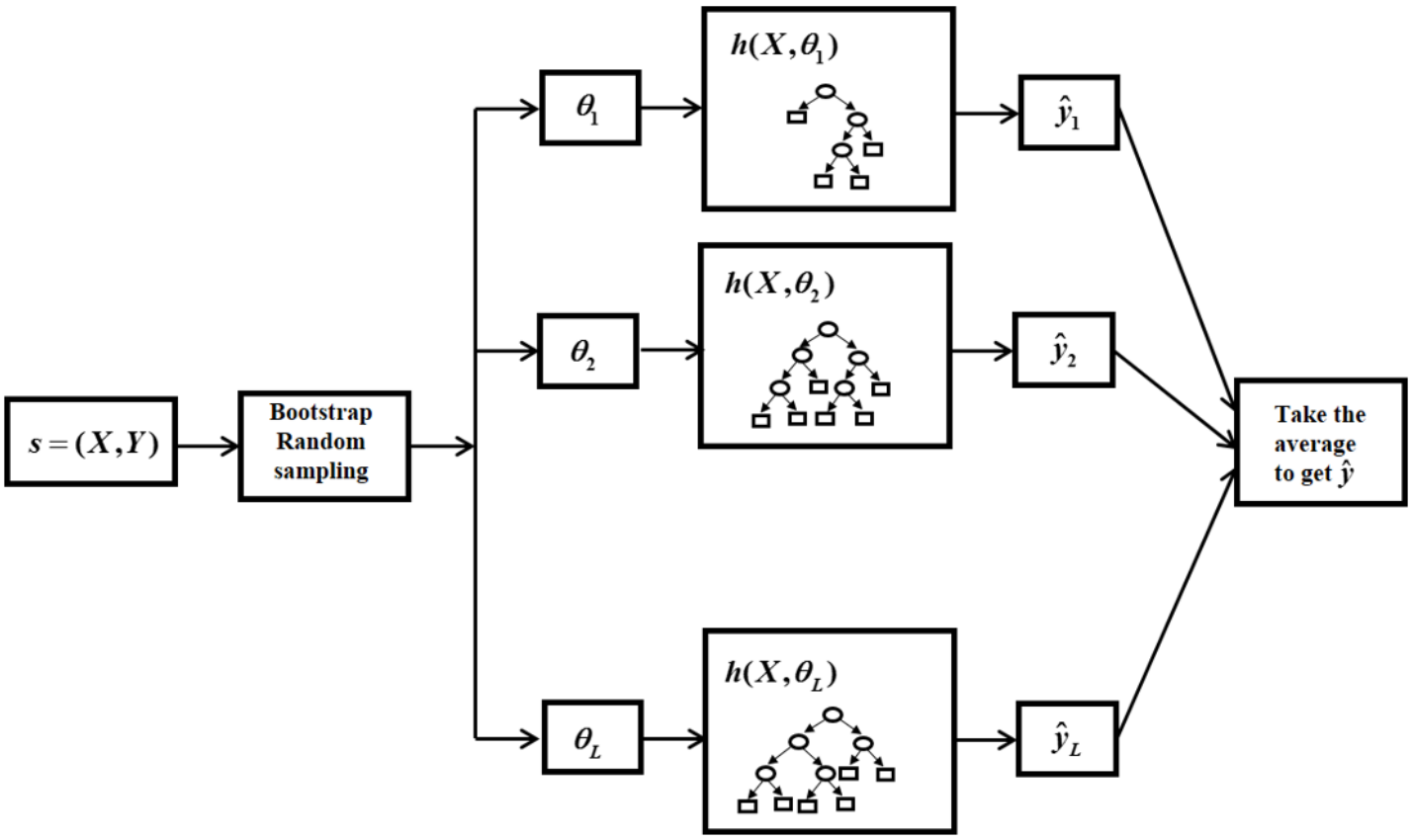

Figure 4

Frame structure of random forest regression prediction model.

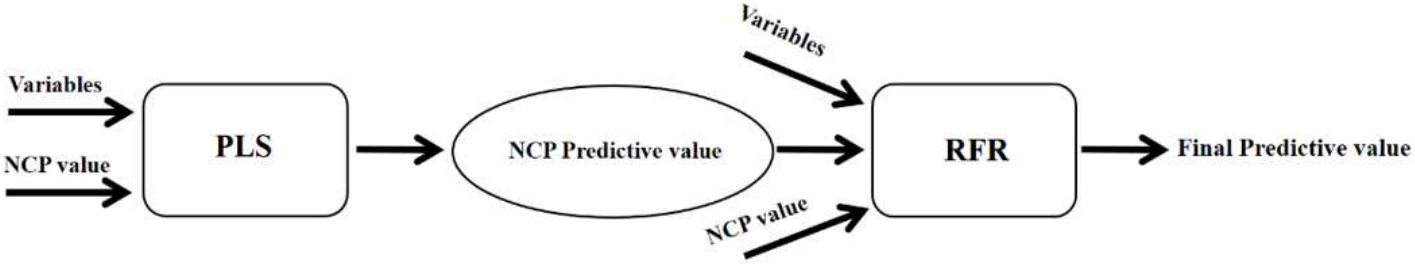

Figure 5

The flux diagram of the regression process, where ncp represents the concentration of pollutants measured at the national control point.

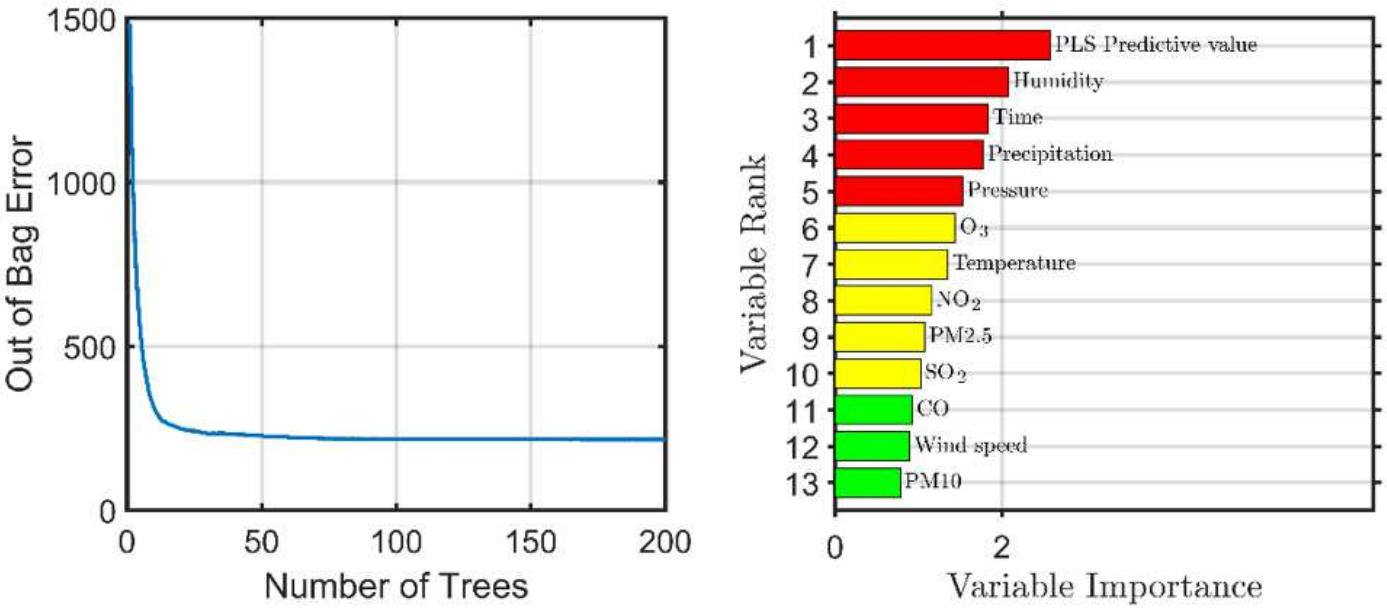

Figure 6

Random forest regression modeling process. The number of decision tree choices is seen on the left. The order of importance of the variables is seen on the right.

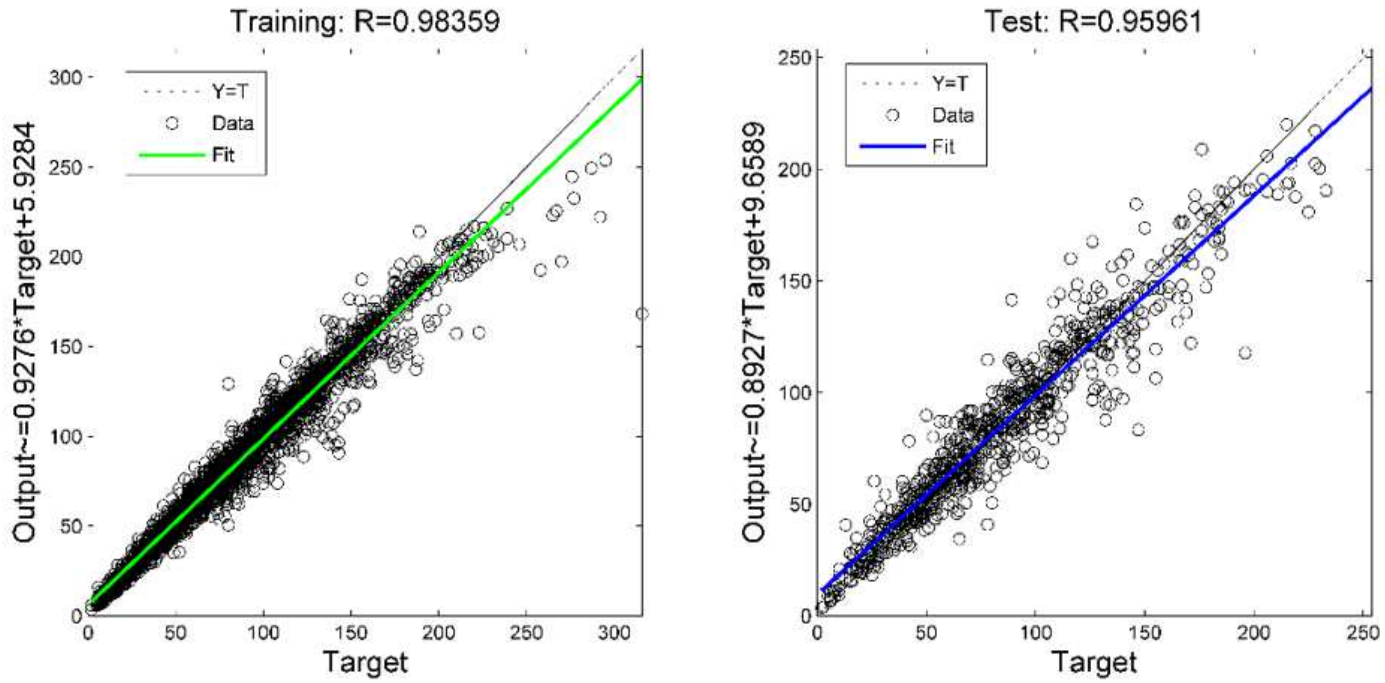

Figure 7

The prediction effect of PM10's PLS-RFR model on the training set and test set.

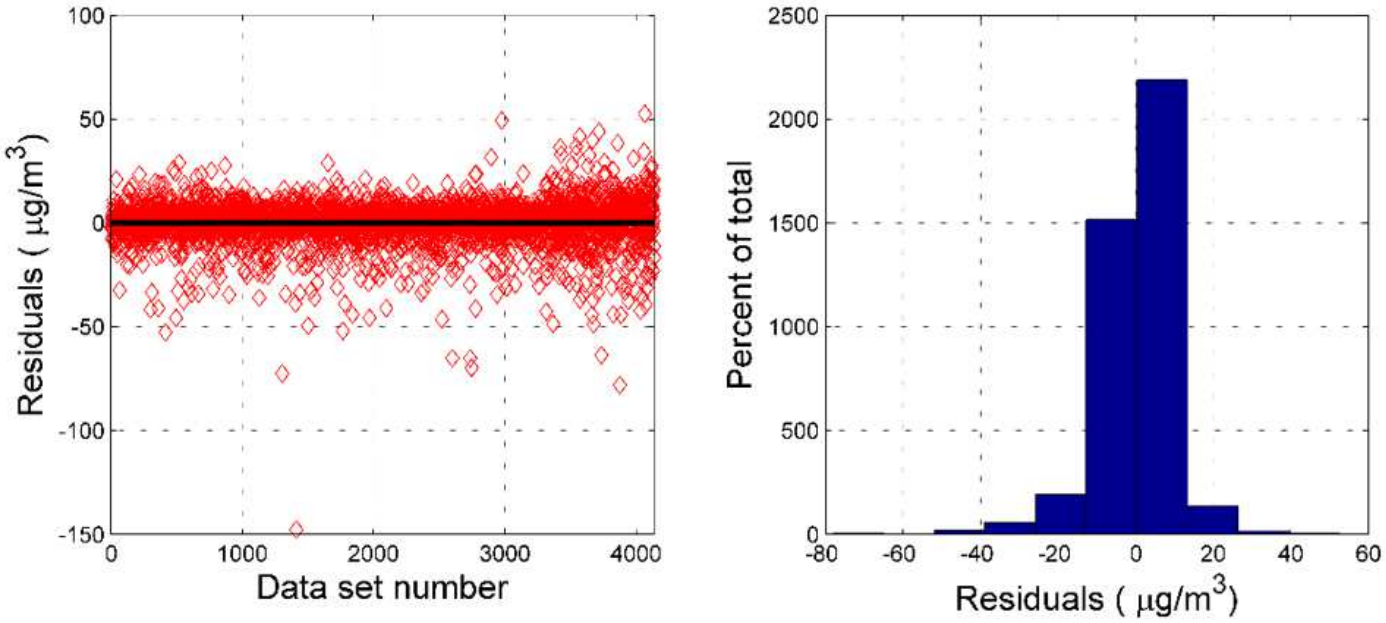

Figure 8

Residual test of PLS-RFR model. The residuals vs. data set number plot is seen on the left. The histogram of the residuals is seen on the right.

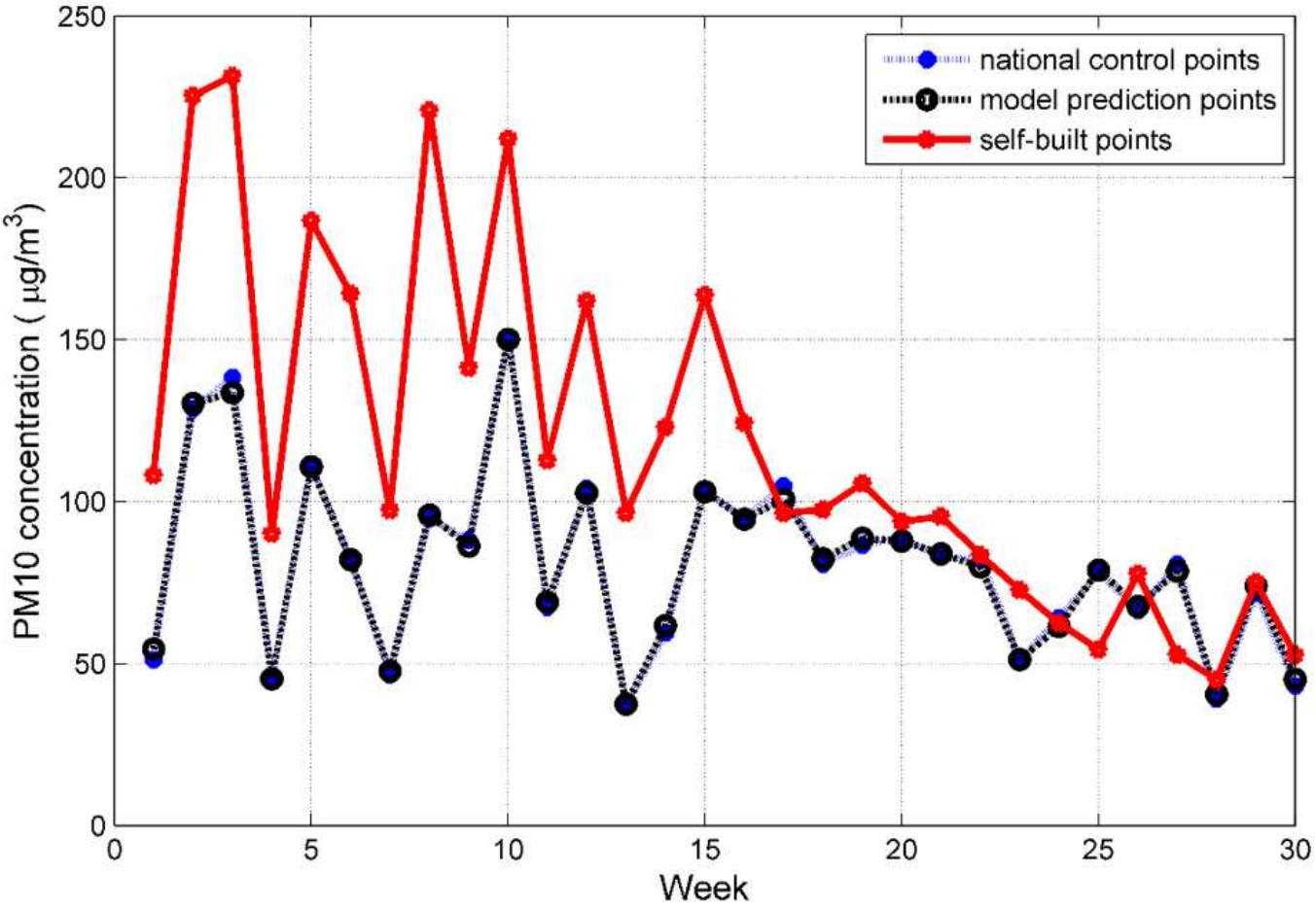

Figure 9

Comparison of the weekly average concentration of PM10 between national control points, PLS-RFR model calibration points and self-built points.