A Data Calibration Method of Micro Air Quality Detector Based on LASSO Regression and NARX Neural Network Combined Model

Bing Liu (Liub1@niit.edu.cn)
Nanjing Vocational University of Industry Technology

Yueqiang Jin
Nanjing Vocational University of Industry Technology

Dezhi Xu
Nanjing Vocational University of Industry Technology

Yishu Wang
Nanjing Vocational University of Industry Technology

Chaoyang Li
Henan University of Technology

Research Article

Keywords: NARX Neural Network, influencing factors

Posted Date: September 3rd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-861855/v1

A data calibration method of micro air quality detector based on LASSO regression and NARX neural network combined model

Bing Liu[1*], Yueqiang Jin[1], Dezhi Xu[2], Yishu Wang[1] & Chaoyang Li[3]

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

[2]Organization Department, Nanjing Vocational University of Industry Technology, Nanjing 210023, China

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

Abstract

Studies have shown that there is a certain correlation between air pollution and various human diseases, especially lung diseases, so it is very meaningful to monitor the concentration of pollutants in the air. Compared with the national air quality monitoring station (national control point), the micro air quality detector has the advantage that it can monitor the concentration of pollutants in real time and grid, but its measurement accuracy needs to be improved. In this paper, the measurement data of the micro air quality detector is calibrated with the help of the LASSO regression and NARX neural network combination (LASSO-NARX) model using the data measured by the national control point. First, correlation analysis is used to test whether the correlation between the concentration of air pollutants and its influencing factors is significant. Second, LASSO regression is used to give the quantitative relationship between pollutant concentration and various influencing factors. Third, the predicted value of each pollutant concentration in the LASSO regression model and the measurement data of the micro air quality detector are used as input variables, and the LASSO-NARX model is constructed using the NARX neural network. Finally, several indicators such as Root Mean Square Error, goodness of fit, Mean Absolute Error and Relative Mean Absolute Percent Error are used to compare various air quality models. The results show that the prediction results of the LASSO-NARX model are not only better than the LASSO model alone and the NARX model alone, but also better than the commonly used multilayer perceptron and radial basis function neural network. The LASSO-NARX model performed equally well on the training set and test set, indicating that the model has excellent generalization capabilities. Using this model to calibrate the measurement data of the micro air quality detector can increase the accuracy by 61.3% to 91.7%.

1. Introduction

With the development of science and technology, the progress of industry and the rapid increase of the global population, the environment that people depend on has been greatly destroyed. Many areas have experienced environmental problems such as acid rain, species extinction, and land desertification. Environmental issues have become one of the common concerns of all countries in the world today, and they are also a major challenge facing mankind in the 21st century. Especially air pollution, which can easily lead to diseases of respiratory diseases, such as acute and chronic bronchitis, asthma, pneumonia, and even lung cancer [1-3]. According to estimates by the World Health Organization, 7 million people die each year from diseases caused by air pollution [4, 5].

The pollutants in the air are mainly inhalable particles, SO_2, NO_2 and other substances. The commonly used index to measure the quality of air is AQI, which is the Air Quality Index.

The larger the AQI value, the more serious the air pollution, and the greater the harm to human health. AQI is calculated based on six air pollutants: PM2.5, PM10, CO, NO_2, SO_2 and O_3 ("two dusts and four gases"). As air quality is getting more and more attention, it is particularly important to monitor air quality.

In order to monitor the air, several national air quality monitoring stations (national control points) are generally set up in a key environmental protection city. Multi-parameter automatic monitoring equipment is installed in the air quality monitoring station for continuous automatic monitoring, and the monitoring results are stored in real time and analyzed to obtain relevant data. The construction and maintenance costs of national control points are relatively high, so the number of national control points is very small, which makes it difficult to conduct comprehensive monitoring of an area. In addition, although the national control point data is relatively accurate, it is often not released in real time, so it is difficult to realize real-time monitoring of air quality. In order to overcome the shortcomings of grid monitoring and real-time monitoring of pollutant concentrations at national control points, some companies have developed micro air quality detectors and arranged them in grids (self-built point). The micro air quality detector has the advantages of low cost, easy installation, and convenient data reading. It can conduct real-time monitoring and grid monitoring of pollutant data [6-8]. However, since the electrochemical sensor used in the micro air quality detector is susceptible to external influences, the range drift and zero point drift will occur after a period of use, and the data measured by the self-built point will have a certain error. How to use the national control point data to calibrate the self-built point data is a problem worthy of study.

The commonly used pollutant concentration prediction models are mainly divided into two categories. The first type is the atmospheric chemistry transmission model, which uses the theory of the atmospheric system to simulate the physical and chemical processes of pollutants in a specific area, and uses the generated pollutant grid data to predict air quality [9, 10]. The mechanism of the atmospheric chemistry transmission model is complex, and is limited by the accuracy of the ground emission inventory, and its pollutant forecast effect is not very good.

Another commonly used pollutant concentration prediction model is a statistical model based on machine learning algorithms. The multiple linear regression model is a more classic statistical model, and its biggest advantage is its interpretability. Since the quantitative relationship between pollutants and other variables of the model can be intuitively known through the regression equation, the construction of a multivariate linear regression equation is still a commonly used air quality prediction modeling idea [11, 12]. Lei, M. T., et al used meteorological and air quality data from 2013 to 2017 for five years to establish a statistical model based on linear multiple regression (MR) and classification regression tree (CART) analysis. The model successfully predicted the concentrations of NO_2, PM10, PM2.5 and O_3 in Macau on the second day [13]. Liu, Bing, et al. successfully predicted the concentration of major pollutants in the air using a combined model of stepwise regression and multilayer perceptron neural network [14]. It is difficult for multiple linear regression models to detect the complex and potentially non-linear relationship between predictor variables and response variables, so machine learning algorithms such as artificial neural networks[15-18], support vector machines [19-22], random forest [23-26] and extreme gradient boosting [27-29] have become the mainstream of pollutant concentration prediction. The nonlinear auto-regressive with exogenous (NARX) input model increases the delay and feedback mechanism, so it enhances the ability to remember historical data. In recent years, it is often used for air quality prediction. Moursi, A. S. et al. used the PM2.5 concentration, cumulative wind speed and cumulative rainfall hours in the past 24 hours as independent variables, and successfully predicted the PM2.5 concentration in the next hour using the NARX model [30]. Mohebbi, M. R., et.al successfully simulated the carbon monoxide concentration in Shiraz using the NARX neural network model without traffic data. The results show that the dynamic neural network is better than the static neural network in the prediction accuracy of CO concentration in this area [31].

2. Material and methods

2.1. Data source and preprocessing

The appearance of the micro air quality detector makes it possible to monitor the concentration of pollutants in real time, but the accuracy of its measurement needs to be improved. The two sets of data are collected in this paper to build the data calibration model of the micro air quality detector. The first set of data is measured by a national monitoring station in Nanjing, which provides the concentration of two dusts and four gases from November 14, 2018 to June 11, 2019. It has a total of 4200 pieces of data, and the interval of each group of data is mostly 1 hour. The second set of data is measured by a self-built point equipped with a micro air quality detector. It contains 234,717 pieces of data whose time interval does not exceed 5 minutes. The location of the self-built point is juxtaposed with the national control point. It not only measures the concentration of the two dust and four gases in the same period, but also provides five meteorological parameters of wind speed, pressure, precipitation, temperature and humidity.

Preprocessing of data is a prerequisite for building statistical models. The first step is to delete duplicate data and obviously abnormal data (greater than 3 times the average value of the left and right neighbors) in the data. In the second step, the self-built point data is averaged on an hourly basis, and the averaged self-built point data is used to correspond to the national control point data, and the data that cannot be corresponding is deleted [14]. The preprocessed data is shown in Table 1.

Table 1
Descriptive statistics of pollutant concentrations and meteorological parameters measured by 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

Due to the influence of internal factors and external factors, there are certain errors in the data measured by the micro air quality detector. This article draws a time series chart to show the difference between self-built point and national control point [20, 32]. The discussion method of the two dusts and four airs is similar. We randomly select O_3 for analysis.

It can be seen from Fig. 1 that the change trend of O_3 concentration at the self-built point is roughly the same as that at the national control point. However, there is a certain difference between the O_3 concentration of the self-built point and the national control point. In the first 1500 hours, the O_3 concentration of self-built point was generally higher than that of national control points. After 1500 hours, the fluctuation degree of O_3 concentration at the national control point is generally greater than the fluctuation degree of the O_3 concentration at the self-built point.

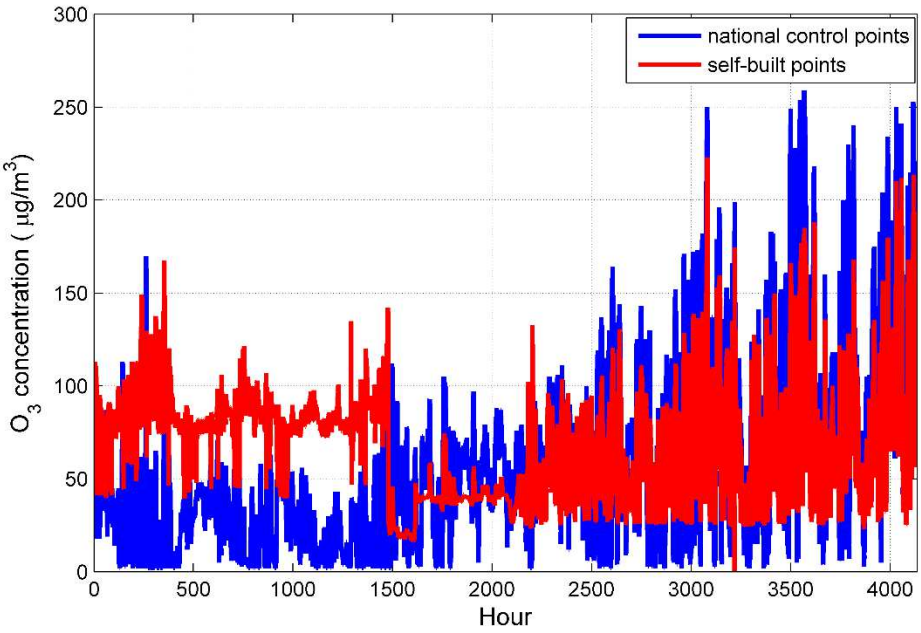

Fig.1. Comparison of hourly average O_3 concentration data between national control points and self-built points. Figures are generated using Matlab (Version R2016a, https://www.mathworks.com/) [Software].

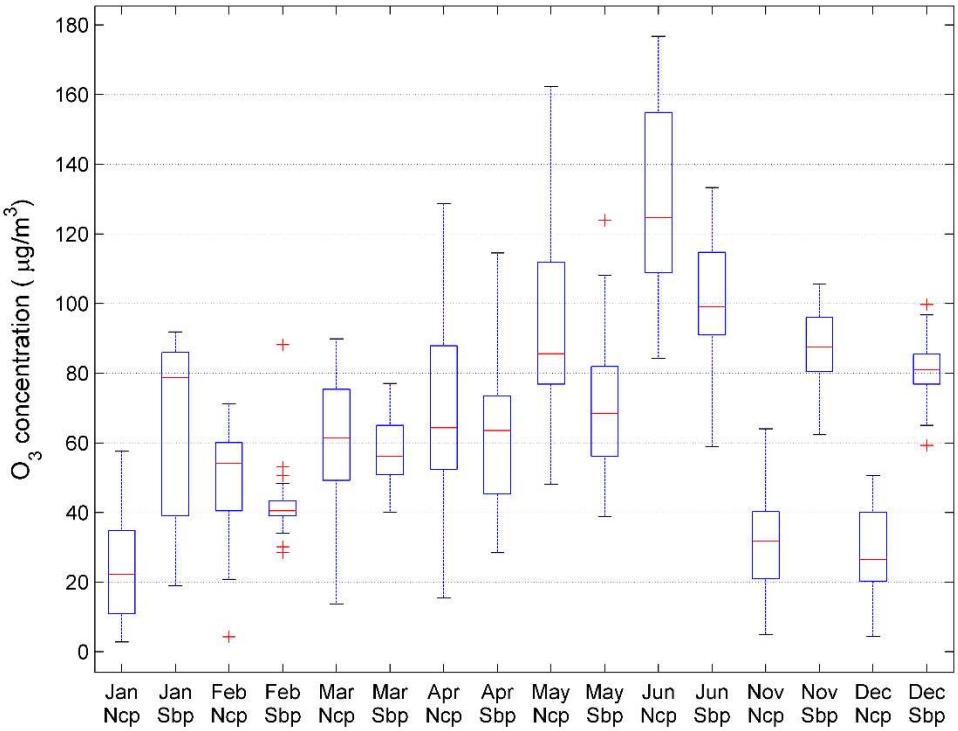

Fig.2. Compare the O_3 concentration of national control points (Ncp) and self-built points (Sbp) on a monthly basis. Note that there is no data from July to October.

Since there are certain differences in meteorological parameters in each month, in order to reflect the influence of meteorological parameters on the concentration of pollutants, we have drawn a box plot [33] as shown in Fig. 2. It can be seen that the difference in O_3 concentration between self-built point and national control point is different every month. In November,

December, January and February, the O_3 concentration difference between the self-built point and the nationally controlled point is large. The reason is that the low temperature and low humidity during this period affect the accuracy of the electrochemical sensor. It can be seen that meteorological parameters are also factors that affect the concentration of pollutants.

2.3. Correlation analysis

The key to air quality prediction is the prediction of the concentration of pollutants such as two dusts and four gases. Predicting the concentration of pollutants must find out the main factors that affect it [10]. Because the factors that affect the concentration of pollutants in the air are more complex, and the factors themselves also affect each other, quantitative indicators are needed to describe them. Pearson correlation coefficient (Eq. (1)) is a statistical indicator used to reflect the degree of correlation between variables [13, 29].

Table 2 shows the correlation between the concentration of six types of pollutants and meteorological parameters. It can be seen that at a significant level of 0.05, all variables have a significant correlation with each other except for the NO_2 concentration and temperature. The absolute value of the correlation coefficient between many of these variables exceeds 0.8, indicating that they are highly correlated. The matrix color block diagram can intuitively show the correlation coefficient between the variables. In Fig. 3, the area of the sector represents the absolute value of the correlation coefficient, light color represents positive correlation, dark color represents negative correlation, and the lighter the color, the larger the correlation coefficient.

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

Table 2
Pearson linear correlation coefficients between six types of air pollutant concentrations and meteorological parameters (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

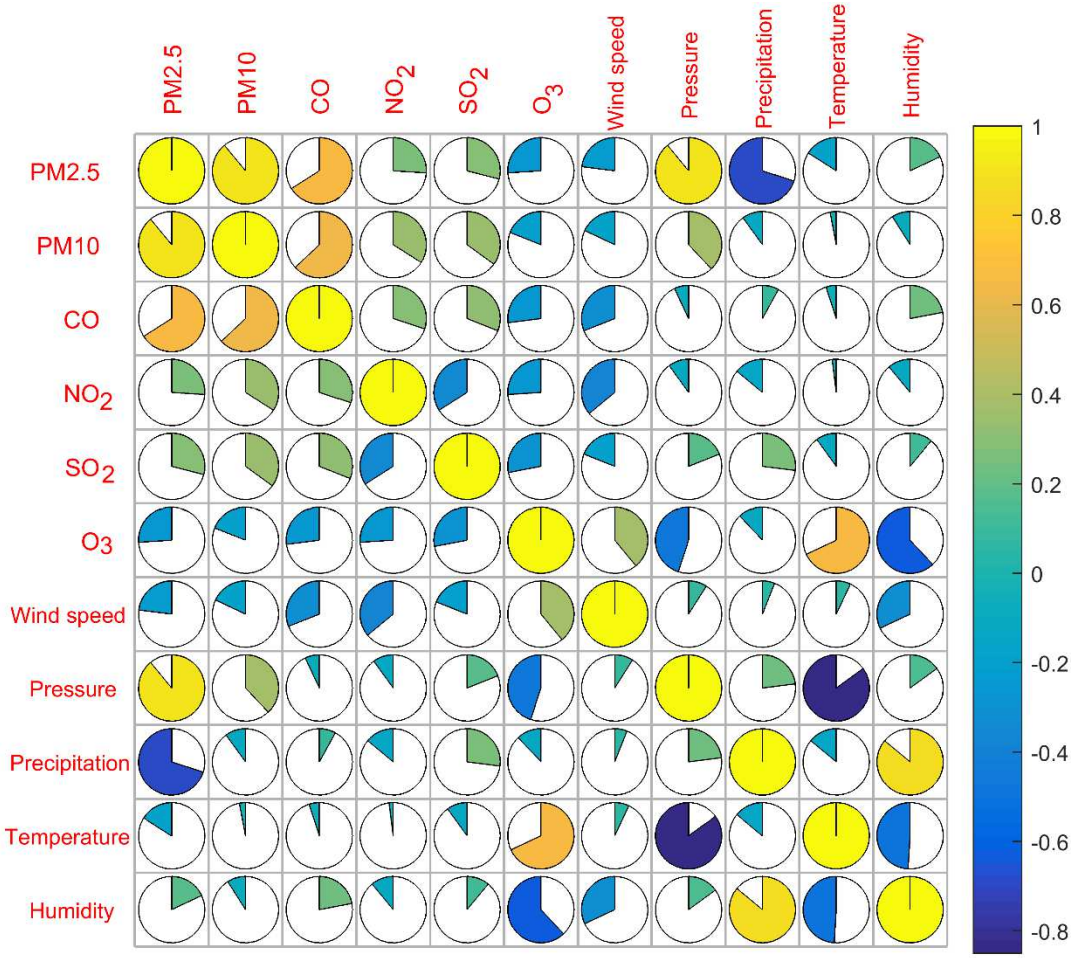

Fig. 3. Matrix color block diagram of the correlation coefficient matrix between the concentration of six air pollutants and five meteorological parameters.

3. Establishment of sensor calibration model

3.1. Introduction to basic principles

Least absolute shrinkage and selection operator (LASSO) was first proposed by Tibshirani in 1996. This method is a compression estimation. It constructs a penalty function to obtain a more refined model, so that it can compress some coefficients, and at the same time set some coefficients to zero, to achieve the effect of subset shrinkage [29, 34].

In a general regression model, the observed values of each data are generally considered to be independent of each other. Because there are many variables in the model, their dimensions are often different. In order to eliminate the interference of dimensions, all independent variables $X_i = (x_{i1}, x_{i2}, \cdots, x_{im})$ need to be standardized transformation. The standardized $z_{i1}, z_{i2}, \cdots, z_{im}$ mean is 0, and the variance is 1. Eq. (2) is the LASSO estimate of the regression model, where $S \geq 0$ is the harmonic parameter. For any s, the estimate of α is $\hat{\alpha} = \bar{y}$.In order to reduce the overall regression coefficient, only the harmonic parameter s needs to be adjusted. When x = 0, the coefficients of some variables will decrease, or even close to 0 or equal to 0. These irrelevant or weakly correlated independent variables will be filtered out, thereby improving the accuracy and interpretability of the regression model.

$$\left(\hat{\alpha}, \hat{\beta}\right) = \arg\min_{(\alpha,\beta)} \sum_{i=1}^{n} (y_i - \alpha - \sum_{j=1}^{p} x_{ij}\, \beta_j)^2$$

$$\text{subject to } \sum_{j=1}^{p}|\beta_j| \leq s \tag{2}$$

Eqs. (3)-(4) is often used to solve LASSO regression coefficients and evaluate the pros and cons of regression models, where n represents the total number of samples; p represents the number of independent variables in the subset regression model; SSE_p represents the sum of squares of the residuals after the dependent variable Y is regressed; δ^2 represents the prediction of the mean value of the variance when all independent variables regress the dependent variable Y. According to this, the model when C_p is the minimum value is obtained, and the best subset of variables in the global scope is obtained, and the regression equation with the best effect is generated at the same time [35, 36].

$$C_p = \frac{\mathrm{SSE}_p}{\delta^2} - n + 2p \tag{3}$$

$$SSE_p = \sum_{i=1}^{n}(Y_i - Y_{pi})^2 \tag{4}$$

A typical NARX neural network is mainly composed of input layer, hidden layer, output layer and input and output delay. NARX neural network model is a kind of nonlinear discrete system, which can be represented by a nonlinear difference equation (Eq. (5)), where y represents the output variable; x represents the external input variable; d represents the delay step. Different delay steps can be set for output variables and input variables to control the time step of continuous prediction.

Eq. (6) is the calculation formula for the output of each layer, where x_i represents the input of each layer of neurons, that is, the output of the previous layer of neurons; $\omega_{i,j}$ represents the weight between layers; b_j represents the threshold of the layer; f represents the activation function. The activation function of the hidden layer of the NARX neural network uses the hyperbolic tangent function (Eq. (7)), and the output layer uses the linear function (Eq. (8)).

$$y(t) = f(x(t-1), x(t-2), \cdots, x(t-d), y(t-1), y(t-2), \cdots, y(t-d)) \tag{5}$$

$$H_j = f\left(\sum_{i=1}^{n} \omega_{i,j} x_i - b_j\right) \tag{6}$$

$$\tanh(\mathrm{x}) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{7}$$

$$\mathrm{linear}(\mathrm{x}) = \mathrm{x} \tag{8}$$

3.2. LASSO regression model construction

From the correlation analysis, we can see that there is a strong correlation between the concentration of various pollutants, and between the pollutants and meteorological parameters. In this paper, the pollutant concentration at the national control point is used as the dependent variable, and the pollutant concentration and meteorological parameters measured at the self-built point are used as independent variables to establish a multiple linear regression model. An important requirement of multiple linear regression models is that the independent variables are independent of each other. Through the multicollinearity diagnosis of the model, we can see that the maximum variance inflation factor of the multiple linear regression model is 26.631, which is greater than 10. Therefore, the multiple linear regression model has serious multicollinearity. Multicollinearity will make the air quality prediction model very unstable and cause over-fitting problems.

Commonly used methods to solve multicollinearity in practical problems are: (i) Selecting the independent variables, and the representative methods include forward regression,

backward regression and stepwise regression. (ii) Perform dimensionality reduction processing on independent variables. Representative methods include principal component regression and partial least squares regression. (iii) Biased estimation of regression coefficients, representative methods include ridge regression and LASSO regression. This study uses LASSO regression to solve the problem of multicollinearity. Compared with ridge regression, LASSO regression can select variables and eliminate some variables that have no significant influence on the dependent variable. Compared with stepwise regression, LASSO regression can retain those variables that are between significant and non-significant effects on the dependent variable, so the estimation deviation is not too large.

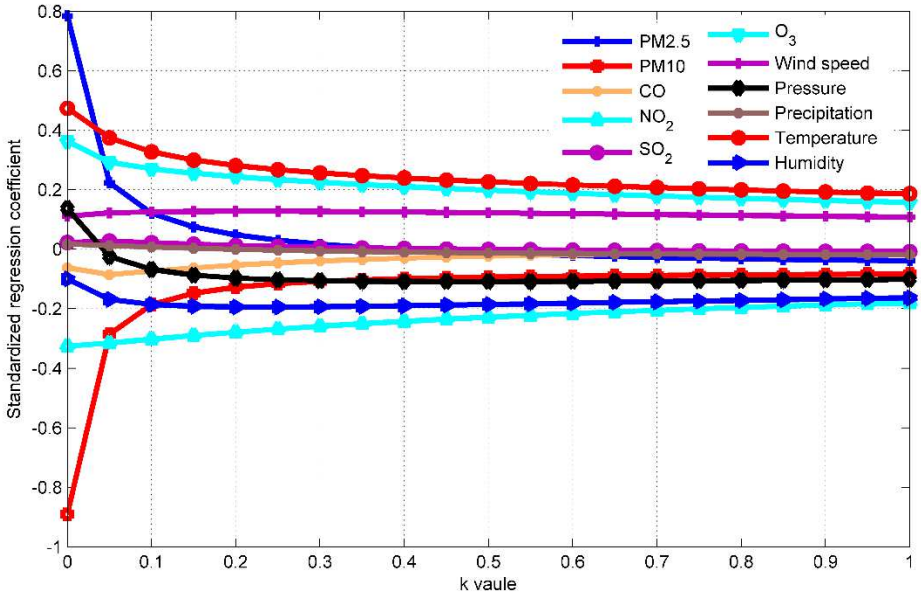

Fig. 4. The trace diagram of all input variables, where the dependent variable is the O_3 concentration measured by the national control point.

The analysis of LASSO regression using SPSSAU (https://spssau.com/) software is divided into two steps: (i) Find the best K value based on the trajectory graph. The selection principle of K value is the minimum K value when the standardized regression coefficient of each independent variable becomes stable. The smaller the K value, the smaller the deviation, when the K value is 0, it is an ordinary linear OLS regression. (ii) Manually input K value for regression modeling. For the K value, generally the smaller the better, and it is generally recommended to be less than 1.After determining the K value, we can manually enter the K value to get the Lasso regression model estimate.

For the LASSO regression model of O_3 concentration prediction, it can be seen from Fig. 4 that when k=0.05, the standardized regression coefficients of each independent variable tend to be stable, so this paper takes k=0.05 to establish the LASSO regression model. In the model, PM2.5 concentration, CO concentration, SO_2 concentration, pressure and precipitation have no effect on O_3 concentration, so they are excluded from the model. The F value in the model test is 1123.756, and the corresponding p value is less than 0.01, indicating that at the significance level of 0.01, the overall variables introduced into the model have a significant impact on the pollutant concentration. The coefficient of determination of the LASSO model is 0.750, indicating that 75% of the change in O_3 concentration can be explained by the change in the independent variables introduced into the model, and the model has a high degree of goodness of fit. The results of the remaining pollutants LASSO regression model are shown in Table 3.

Table 3
LASSO regression model of six types of air pollutant concentrations. In the model, the dependent variable is the concentration of the six pollutants at the national control point, and the independent variable is the variable monitored by the self-built point (— represents the variables eliminated in the model).

Independent variable	PM2.5	PM10	CO($\times 10^{-2}$)	NO_2	SO_2	O_3
Constant	8.663	47.475	2.127	174.759	-303.100	63.734
PM2.5	0.724	0.890	0.005	0.070	—	—
PM10	—	—	—	—	0.034	-0.0315
CO	1.022	24.0446	0.197	-10.787	31.255	—
NO_2	—	0.247	0.002	0.368	0.038	-0.550
SO_2	—	—	—	0.012	—	—
O_3	—	—	—	-0.148	0.081	0.264
Wind speed	—	—	-0.033	-14.472	-2.268	12.520
Pressure	—	—	-0.002	-0.111	0.289	—
Precipitation	—	-0.00467	—	-0.030	0.0015	—
Temperature	—	—	—	—	—	2.188
Humidity	-0.083	-0.760	—	-0.363	—	-0.375
k value	0.050	0.040	0.010	0.020	0.020	0.050
F value	2307.828	1339.744	284.478	308.185	237.27	1123.756
P value	0.000	0.000	0.000	0.000	0.000	0.000
R^2	0.860	0.781	0.431	0.451	0.388	0.750

3.3. LASSO-NARX model construction

The LASSO regression model gives a quantitative linear relationship between the pollutant concentration and various influencing factors [31]. However, there is a nonlinear relationship between pollutant concentration and influencing factors, and the accuracy of pollutant prediction needs to be improved. Taking into account the time sequence of pollutant concentration, this paper uses NARX neural network to improve the accuracy of pollutant concentration prediction. We take the predicted value of LASSO regression and the data measured by self-built points as input, and the concentration of six pollutants as output to establish the NARX neural network model. This combined model is called the LASSO-NARX model in this paper, and the specific process is shown in Fig. 5.

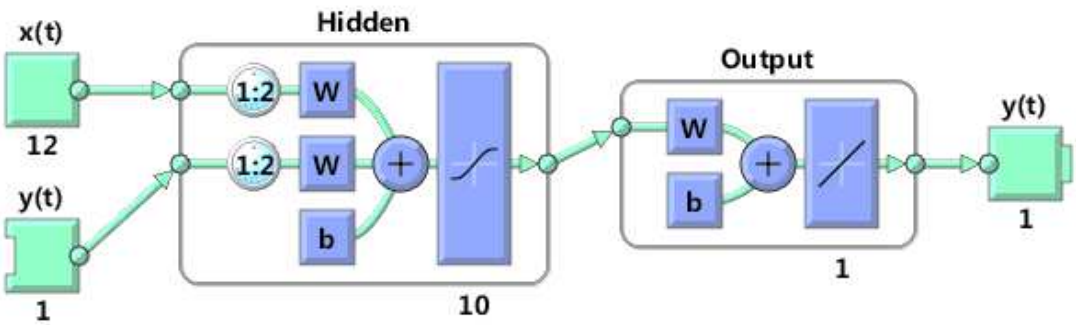

Fig.5. The frame structure of the LASSO-NARX model, where the input is the predicted value of the LASSO regression model and the measured value of the self-built point

In the NARX neural network, it can be known from the Kolmogorov theorem that at most two hidden layers can identify arbitrary nonlinear characteristics, so this paper selects the default one hidden layer in Matlab. The number of nodes in the hidden layer of the neural

network is determined by considering the training effect and training time. For the delay order in the model, determine the order change range based on experience, and find out the order when it no longer changes significantly as the model delay order according to the change of the mean square error of the model under different orders.

In the NARX model, 4135 samples are randomly divided into training set, validation set and test set at a ratio of 7:1.5:1.5. For comprehensive comparison, the input delay of NARX neural network is selected as 2, and the number of hidden layer nodes is 10. The training algorithm adopts the Levenberg-Marquardt algorithm with shorter training time, and the LASSO-NARX model is established with the help of Matlab software.

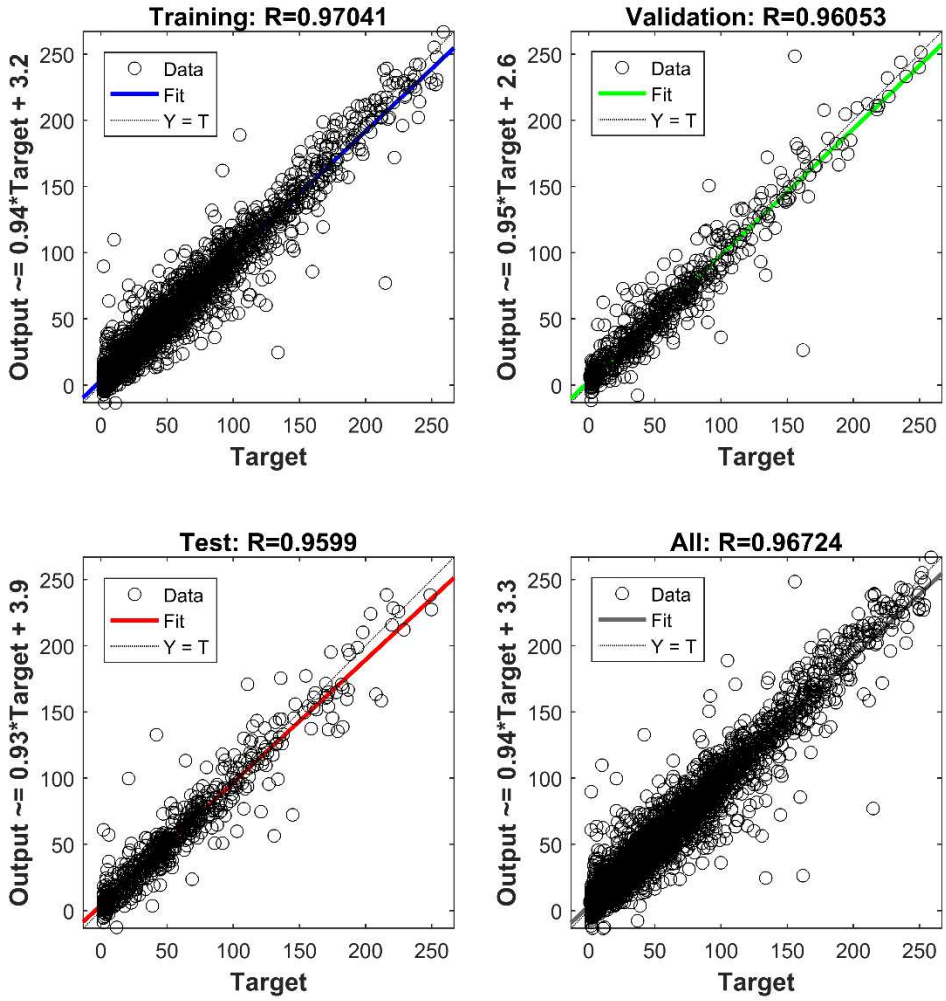

Fig.6. The prediction effect of O_3's LASSO-NARX model on the training set, validation set, test set and all sets.

In order to visually show the prediction effect of the LASSO-NARX model, we have drawn the O_3 concentration regression effect diagram. It can be seen from Fig. 6 that whether it is the training set, the validation set or the test set, the correlation coefficient between the predicted value of the model and the true value of the national control point exceeds 0.95, and the coefficients of each regression model are close to 1. It shows that the LASSO-NARX model has achieved good results in prediction. Fig. 7 is a time series diagram of the model. It can be seen that the predicted value of the model almost coincides with the measured value of

the national control point, and the residuals are mostly randomly distributed between [-40, 40], which also verifies that the LASSO-NARX model has good accuracy.

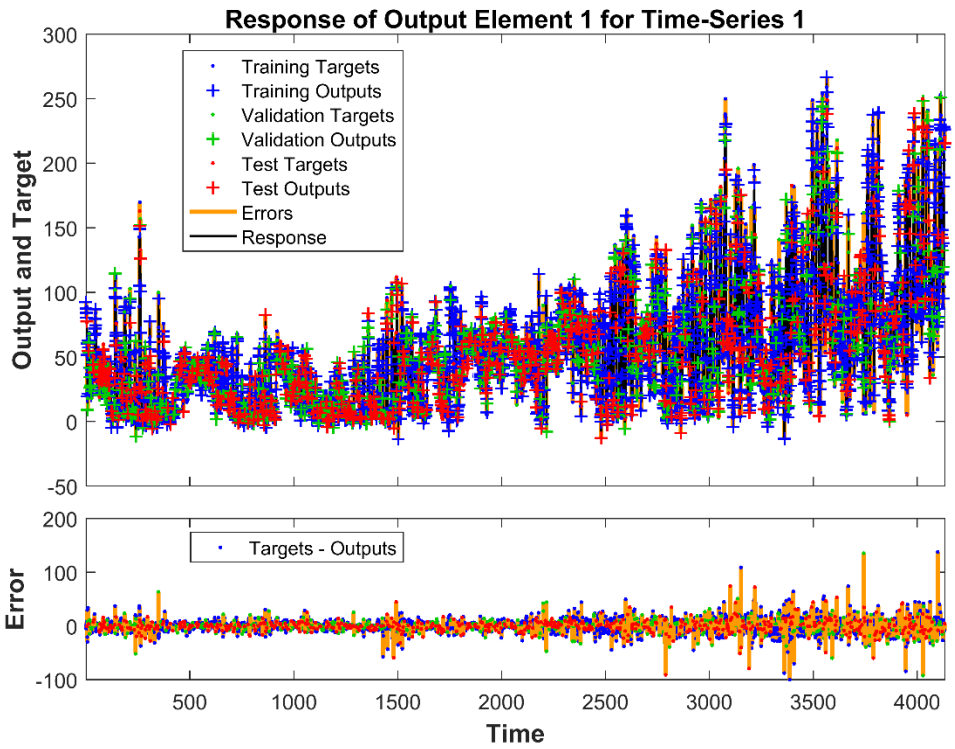

Fig.7. Residual test of LASSO-NARX model. The comparison between the predicted value of the model and the measured value of the national control point is at the top. The error vs. time chart is shown below.

4. Discussion

In the data calibration problem of the micro air quality detector, the LASSO model alone and the NARX neural network model alone can predict the concentration of pollutants. This paper also chooses a multilayer perceptron (MLP) and a radial basis function (RBF) neural network to compare with the LASSO-NARX model. Taylor diagrams are often used to visually compare the accuracy of various models [8]. The scattered points in the Taylor diagram represent the model, the radial line represents the correlation coefficient (Eq. (1)), the horizontal and vertical axis represents the standard deviation (Eq. (9)), and the dashed line represents the center root mean square error (Eq. (10)).

$$\sigma = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(w_t - \overline{w})^2} \tag{9}$$

$$E' = \sqrt{\frac{1}{n}\sum_{t=1}^{n}[(y_t - \overline{y}) - (w_t - \overline{w})]^2} \tag{10}$$

Fig. 8 is a Taylor analysis chart of O_3 concentration. It can be seen that compared with the O_3 concentration measured by the national control point, the O_3 concentration measured by the self-built point has the lowest accuracy, the LASSO model and the RBF neural network model have good accuracy, and the MLP neural network and NARX model have

higher accuracy. The LASSO-NARX model proposed in this article performs best in comparison with other models.

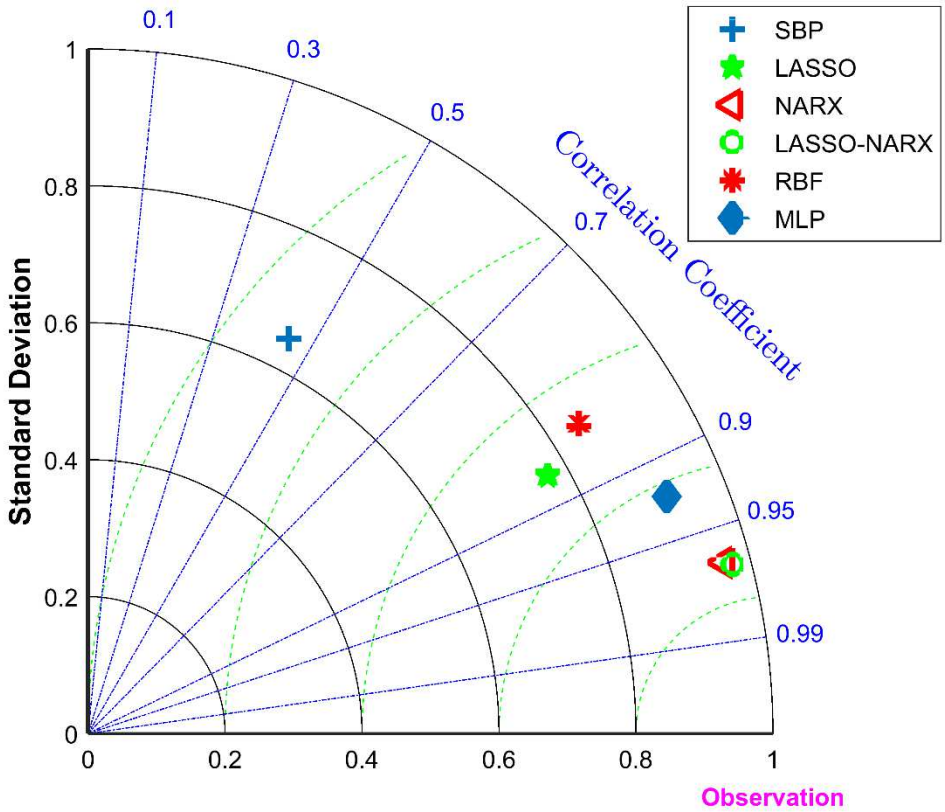

Fig. 8. Taylor diagrams of predicted values of five models and measured values of self-built points, where SBP stands for self-built points.

Root Mean Square Error (RMSE), goodness of fit (R^2), Mean Absolute Error (MAE) and Relative Mean Absolute Percent Error (MAPE) can also be used to compare various air quality prediction models. Eqs. (11)-(14) are specific formulas, where y_t is the measured value at the national control point, $\bar{y}$ is the average value of the national control point, and w_t is the regression value of the model [25, 28].

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - w_t)^2} \tag{11}$$

$$R^2 = 1 - \frac{\sum_{t=1}^{n}(y_t - w_t)^2}{\sum_{t=1}^{n}(y_t - \bar{y})^2} \tag{12}$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t - w_t| \tag{13}$$

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t - w_t}{y_t}\right| \tag{14}$$

It can be seen from Table 4-7 that in the comparison with the data of the national air quality monitoring station, the measurement data of the micro air quality detector has a large error, so it needs to be calibrated. The LASSO regression model and RBF neural network model can calibrate self-built point data, but the effect needs to be improved. The MLP neural network and NARX model have a good effect on the calibration of self-built point data, and

the LASSO-NARX model given in this article is the best in each evaluation index. In the index of goodness of fit, several self-built points are negative, which is caused by the large error of self-built points. Among the other three indexes, the most improved is the MAPE of NO_2, which is an increase of 91.7%, and the least improved is the RMSE of PM2.5, which is an increase of 61.3%.

Table 4
RMSE of six types of air pollutant concentrations between self-built points, model forecast values and national control point.

Input variable	Self-built points	LASSO	NARX	LASSO-NARX	RBF	SVR	MLP
PM2.5	22.436	12.515	8.800	8.687	19.323	8.649	10.777
PM10	66.263	21.495	13.911	13.208	30.570	11.656	19.126
CO	0.679	0.344	0.158	0.156	0.385	0.175	0.304
NO_2	37.183	18.035	8.081	7.715	19.029	7.725	13.216
SO_2	26.24	15.627	5.104	4.874	15.449	4.116	9.984
O_3	45.673	24.003	12.477	12.190	25.638	11.304	18.603

Table 5
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	LASSO	NARX	LASSO-NARX	RBF	SVR	MLP
PM2.5	0.551	0.860	0.931	0.933	0.667	0.933	0.907
PM10	-1.076	0.781	0.909	0.918	0.558	0.938	0.827
CO	-0.929	0.507	0.895	0.899	0.380	0.872	0.708
NO_2	-1.333	0.451	0.890	0.900	0.389	0.899	0.752
SO_2	-0.726	0.388	0.935	0.941	0.402	0.958	0.786
O_3	0.094	0.750	0.932	0.936	0.715	0.945	0.864

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	LASSO	NARX	LASSO-NARX	RBF	SVR	MLP
PM2.5	18.181	9.193	6.070	5.951	13.709	5.821	7.763
PM10	50.151	15.037	9.218	8.981	22.349	7.080	13.184
CO	0.549	0.263	0.100	0.098	0.288	0.110	0.237
NO_2	29.838	13.877	4.924	4.806	14.166	4.658	9.991
SO_2	12.867	10.421	2.684	2.464	9.998	2.116	7.246
O_3	36.63	18.683	7.948	7.788	18.930	7.647	14.396

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

Input variable	Self-built points	LASSO	NARX	LASSO-NARX	RBF	SVR	MLP
PM2.5	0.447	0.242	0.151	0.146	0.370	0.133	0.185
PM10	0.887	0.264	0.147	0.146	0.428	0.107	0.210
CO	0.478	0.317	0.096	0.095	0.379	0.112	0.283
NO_2	2.129	0.760	0.1816	0.177	0.737	0.170	0.471
SO_2	0.685	0.737	0.161	0.131	0.735	0.131	0.530
O_3	4.322	1.487	0.428	0.397	1.446	0.373	1.002

5. Conclusions

The human body needs to inhale 10-12 cubic meters of air every day, so the quality of air is closely related to human survival. The main pollutants that affect air quality are PM2.5, PM10, CO, NO_2, SO_2 and O_3 [1, 3]. Only by real-time monitoring of the concentration of pollutants, the government and relevant departments can take appropriate measures to the pollution source in a timely manner.

The national control points established by some countries can measure the concentration of pollutants more accurately. However, due to various reasons, the number of national control points is too small, and it is difficult to form grid monitoring. Another disadvantage of the national control point is that the release of data is lagging, so it is difficult to form real-time monitoring. The appearance of the micro air quality detector overcomes these shortcomings of the national control point, but because the electrochemical sensor used is too sensitive, the accuracy of the measurement needs to be improved.

The LASSO regression model can calibrate the data measured by the micro air quality detector and give the quantitative relationship between the pollutant concentration and each influencing factor, but it cannot find the nonlinear relationship between the pollutant concentration and each influencing factor. The NARX model can find the nonlinear relationship between the pollutant concentration and various influencing factors, and the prediction accuracy is significantly higher than the LASSO regression model. However, it cannot give a quantitative relationship between pollutant concentration and various influencing factors.

The LASSO-NARX air quality combined model proposed in this study combines the advantages of the two models. It can not only reflect the quantitative relationship between the pollutant concentration and the influencing factors, but also has a higher prediction accuracy than the NARX neural network model alone. The LASSO-NARX model performs very well on the training set and test set, indicating that it has a strong generalization ability. The model uses a total of 4135 sets of data, and the data of the four seasons are all covered in the model, which also shows that the model is relatively stable. However, due to the different climatic conditions in different regions, this model may not be applicable to other regions. In the future, our team will try to collect data from other regions to further validate the model.

Acknowledgements

This work was supported by the Youth Program of National Natural Science Foundation of China (No.71602051) and Research Project of Higher Vocational Education in Nanjing Vocational University of Industry Technology (No. GJ20-30).

Author information

Affiliations

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

Bing Liu, Yueqiang Jin & Yishu Wang

2. Organization Department, Nanjing Vocational University of Industry Technology, Nanjing 210023, China

Dezhi Xu

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

Chaoyang Li

*email: Liub1@niit.edu.cn

Author contributions

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

Additional information

Correspondence and requests for materials should be addressed to B.L.

References

[1] Qiu, H. et al. Differential effects of fine and coarse particles on daily emergency cardiovascular hospitalizations in Hong Kong. *Atmos. Environ.* **64**, 296-302 (2013).

[2] 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).

[3] Johanna, L., Francine, L., Douglas, D. & Joel, S. Chronic Exposure to Fine Particles and Mortality: An Extended Follow-up of the Harvard Six Cities Study from 1974 to 2009. *Environ. Health. Persp.* **120**, 965-970 (2012).

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

[5] 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).

[6] 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. Actuator B-Chem.* **215**, 249-257 (2015).

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

[8] Cordero, José María, Borge, Rafael & Narros, Adolfo. Using statistical methods to carry out in field calibrations of low cost air quality sensors. *Sensor. Actuator. B-Chem.* **267**, 245-254 (2018).

[9] Azid, A. et al. Assessing Indoor Air Quality Using Chemometric Models. *Pol. J. Environ. Stud.* **6**, 2443-2450 (2018).

[10] 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).

[11] Spinelle, L., Gerboles, M., Villani, M. G., Aleixandre, M. & Bonavitacola, F. Field calibration of a cluster of low-cost commercially available sensors for air quality monitoring. Part B: NO, CO and CO2. *Sensor. Actuator B-Chem.* **238**, 706-715 (2016).

[12] Elbayoumi, M., Ramli, N. A., & Faizah, F. M. Y. N. Development and comparison of regression models and feedforward backpropagation neural network models to predict seasonal indoor $PM_{2.5–10}$ and $PM_{2.5}$ concentrations in naturally ventilated schools. *Atmos. Pollut. Res.* **6**, 1013-1023 (2015).

[13] Lei, M. T., Monjardino, J., Mendes, L. & Ferreira, F. Macao air quality forecast using statistical methods. *Air. Qual. Atmos. Hlth.* **2,** 249-258 (2019).

[14] 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.* **11,** 1-12 (2021).

[15] Feng, X. et al. Artificial neural networks forecasting of PM2.5 pollution using air mass trajectory based geographic model and wavelet transformation. *Atmos. Enviro.* **107**, 118-128 (2015).
[16] 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).
[17] Reich, S. L., Gomez, D. R., & Dawidowski, L. E. Artificial neural network for the identification of unknown air pollution sources. *Atmos. Environ.* **33,** 3045-3052 (1999).
[18] 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).
[19] 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).
[20] 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* **11,** 1-14 (2021).
[21] Deo, R. C., Wen, X., & Qi, F. A wavelet-coupled support vector machine model for forecasting global incident solar radiation using limited meteorological dataset. *Appl. Energ.* **168,** 568-593 (2016).
[22] Liu, B. et al. Urban air quality forecasting based on multi-dimensional collaborative Support Vector Regression (SVR): A case study of Beijing-Tianjin-Shijiazhuang. *Plos One*, **7,** 1-17 (2017).
[23] Kamińska, J. A. 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).
[24] 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).
[25] Liu, B., Yu, W., Wang, Y., Lv, Q. & Li, C. Research on data correction method of micro air quality detector based on combination of partial least squares and random forest regression. *Ieee Access,* **9,** 99143-99154 (2021).
[26] 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).
[27] Joharestani, M.Z., Cao, C., Ni, X., Bashir, B. & Talebiesfandarani, S. PM2.5 Prediction Based on Random Forest, XGBoost, and Deep Learning Using Multisource Remote Sensing Data. *Atmosphere* **10,** 373 (2019).
[28] Liu, B., Tan, X., Jin, Y. & Li, C. Application of RR-XGBoost combined model in data calibration of micro air quality detector. *Sci. Rep-UK.* **11,** 1-14 (2021).
[29] Zhai, B., & Chen, J. Development of a stacked ensemble model for forecasting and analyzing daily average $PM_{2.5}$ concentrations in Beijing, China. *Sci. Total Environ.* **635,** 644-658 (2018).
[30] Moursi, A. S., El-Fishawy, N., Djahel, S. & Shouman, M. A. An IoT enabled system for enhanced air quality monitoring and prediction on the edge. C*omplex Intell. Syst.* https://doi.org/10.1007/s40747-021-00476-w.
[31] Mohebbi, M. R., Jashni, A. K., Dehghani, M. & Hadad, K. Short-Term Prediction of Carbon Monoxide Concentration Using Artificial Neural Network (NARX) Without Traffic Data: Case Study: Shiraz City. *IJST-T. Civ. Eng.* **3,** 533-540 (2019).
[32] 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).
[33] Wang, X. & Lu, W. Seasonal variation of air pollution index: Hong kong case study. *Chemosphere,* **63**, 1261-1272 (2006).

[34] Tibshirani, T. The lasso method for variable selection in the Cox model. *STAT. MED.* **4,** 385-395(1997).
[35] Sethi, J. K. & Mittal, M. An efficient correlation based adaptive lasso regression method for air quality index prediction. *Earth Sci. Inform.* https://doi.org/10.1007/s12145-021-00618-1.
[36] Sun, H., Cui, Y., Gao, Q & Wang, T. Trimmed lasso regression estimator for binary response data. *Stat. Probabil. Lett.* https://doi.org/10.1016/j.spl.2019.108679