Research on $PM_{2.5}$ Integrated Prediction Model Based on Lasso-RF-GAM

Tingxian Wu[1], Ziru Zhao[1], Haoxiang Wei[2], and Yan Peng[1](✉)

[1] School of Management, Capital Normal University, Beijing, China
pengyan@cnu.edu.cn
[2] School of Engineering, University of Washington Seattle, Seattle, USA

Abstract. $PM_{2.5}$ concentration is very difficult to predict, for it is the result of complex interactions among various factors. This paper combines the random forest-recursive feature elimination algorithm and lasso regression for joint feature selection, puts forward a $PM_{2.5}$ concentration prediction model based on GAM. Firstly, the original data is standardized in the data input layer. Secondly, features were selected with RF-RFE and lasso regression algorithm in the feature selection layer. Meanwhile, weighted average method fused the two feature subsets to obtain the final subset, RF-lasso-T. Finally, the generalized additive models (GAM) is used to predict $PM_{2.5}$ concentration on the RF-lasso-T. Simulated experiments show that feature selection allows GAM model to run more efficiently. The deviance explained by the model reaches 91.5%, which is higher than only using a subset of RF-RFE. This model also reveals the influence of various factors on $PM_{2.5}$, which provides the decision-making basis for haze control.

Keywords: $PM_{2.5}$ concentration · Random Forest-RFE · Lasso method · Feature fusion · GAM model

1 Introduction

In recent years, PM $_{2.5}$ concentration has increased astronomically on almost every continent, and studies show that the damage done are catastrophic and some are even irreversible. Not long-ago the data collected by the Chinese Ministry of Environmental protection presented a sharp increase in both $PM_{2.5}$ concentration and cardiovascular disease. Researchers have shown that high $PM_{2.5}$ concentration is the main cause of lung cancer, respiratory disease, and metabolic disease [1].

Predicting $PM_{2.5}$ concentration has been done in article [2], Professor Joharestani and Professor Cao collected $PM_{2.5}$ air pollution data and climatic features. With Random Forest Modeling and Extreme Gradient Boosting, they were able to form a model to predict $PM_{2.5}$ in the certain areas. But unfortunately, meteorological phenomes caused their data inaccuracy. The data in our experiment was professionally measured and provided by China's National Population and Health Science Data Sharing Service Platform.

The rest of the paper is arranged as the following: In the second part of this paper, the relevant research work is introduced, and in the third part, a $PM_{2.5}$ concentration

Y. Tan et al. (Eds.): DMBD 2020, CCIS 1234, pp. 83–94, 2020.
https://doi.org/10.1007/978-981-15-7205-0_8

prediction model is proposed. This model combines RF-RFE method and lasso model for joint feature selection. Then, the $PM_{2.5}$ concentration prediction model based on GAM is constructed. The model is verified by using the meteorological monitoring data and pollution index monitoring data of the Ministry of Environmental Protection. The experimental results show the effectiveness of the integrated model. In the fourth part, it is summarized that the model can be used to predict the concentration of $PM_{2.5}$, which is helpful for the relevant departments to better understand the influencing factors of $PM_{2.5}$ concentration and provide auxiliary decision-making basis for air pollution control.

2 Related Research Work

2.1 RF-RFE Method

RF(Random forest) and RF-RFE(Random Forest-Recursive Feature Elimination algorithm) are both descendants of machine-learning. RF is compatible with high-dimensional problems and can predict nonlinear relationships with the downside of its frequent inability to identify strong predictors in the presence of other correlated predictors [3].

Random forest can be used for parallel operation, regression analysis, classification, unsupervised learning, and other statistical data analysis; even when there is a large proportion of data missing from the data set, it has the ability to estimate the absent data value, and keep the accuracy unchanged. Based on the above characteristics, this study will select random forest for RFE and RF-RFE for feature selection of PM2.5 concentration influencing factors.

2.2 Lasso Regression

The lasso method (Least Absolute Shrinkage and Selection Operator) was first introduced by Professor Robert Tibshirani. The Lasso method which include the regression shrinkage and selection. It is said to "minimize the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant." [4] Ultimately, Lasso has the ability to achieve two main tasks, regularization and feature selection. In a recent study with Professor Lingzhen Dai and his colleagues, they used the adaptive Lasso method to identify PM2.5 components associated with Blood Pressure in Elderly man. [5] Adaptive Lasso is a more recent version of Lasso, it uses weight to penalize different predictors and allows the researchers in this study to identify the right subset model and satisfy asymptotic normality.

For the given data $\left(X^i, y_i\right), i = 1, 2, \ldots N, X^i = \left(X_{i1}, \ldots X_{ip}\right)^T$ are the predictor variables and y_i are the responses. The mathematical expression of lasso method is

$$\left(\hat{\alpha}, \hat{\beta}\right) = arg\ min\left\{\sum_{i=1}^{N}(y_i - \alpha - \sum_{j}\beta_j x_{ij})^2\right\}, subject\ to \sum_{j}|\beta_j| \leq t. \quad (1)$$

Where, $t \geq 0$ is a tuning parameter. α is the penalty parameter, which has a negative correlation with the number of features finally selected; β is the coefficient corresponding to the characteristic words in each column of the independent variable x. L1 regularization adds the L1 norm of coefficient β as the penalty term to the loss function, because the regular term is non-zero, thus forcing the coefficient corresponding to the weak feature to become 0.

2.3 Generalized Additive Models (GAM)

In order to reach a better understanding of GAM (Generalized Additive Models), one should be familiar with its close relative GLM (Generalized Linear Models). GLM is a generalization for logistic regression in a linear regression.

The general form of GAM is:

$$\log(\lambda) = \alpha + \sum_{i=1}^{P} f_i(x_i) \tag{2}$$

Where f_i represents smoothing functions such as smooth spline, natural cubic spline and local regression [6].

3 $PM_{2.5}$ Integrated Prediction Model Based on Feature Selection and GAM Model

The RF-RFE-lasso and GAM algorithms are used to form the $PM_{2.5}$ integrated prediction model.

The model structure is shown as Fig. 1.

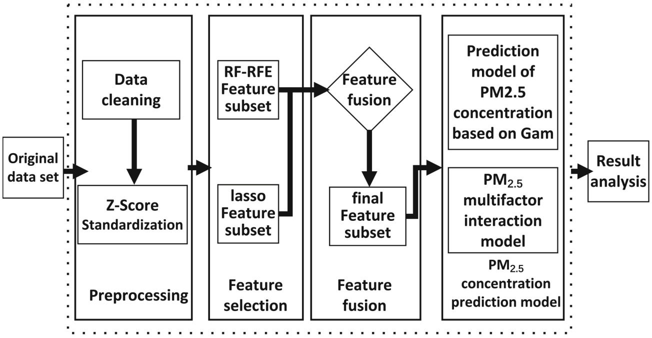

Fig. 1. $PM_{2.5}$ concentration integration model

3.1 Data Preprocessing

3.1.1 Data Source

The experimental data are provided by China's National Population and Health Science Data Sharing Service Platform and released by the Ministry of Environmental Protection. (http://www.ncmi.cn/). Data Set 1 is the air pollution index of provincial capitals from January 1, 2015 to December 31, 2016. Data set 2 is the surface meteorological data from January 1, 2015 to December 31, 2016. The variables selected for this specific experiment are listed in Table 1.

Table 1. Parameter information

Variable	Parameter	Unit	Period
X1	PM_{10}	$\mu g/m^3$	2015.1–2016.12
X2	SO_2	$\mu g/m^3$	
X3	NO_2	$\mu g/m^3$	
X4	O_3	$\mu g/m^3$	
X5	Pressure	0.1 hPa	
X6	Pressure max	0.1 hPa	
X7	Pressure min	0.1 hPa	
X8	Temperature	0.1 °C	
X9	Temperature max	0.1 °C	
X10	Temperature min	0.1 °C	
X11	Relative humidity	1%	
X12	Relative humidity min	1%	
X13	Daily rainfall	0.1 mm	
X14	Wind speed	0.1 m/s	
X15	Wind speed max	0.1 m/s	
X16	Direction of Wind speed max	–	
X17	Wind speed extreme max	0.1 m/s	
X18	Direction of Wind speed max	–	
X19	sun	0.1 h	

3.1.2 Data Standardization

As presented in Table 1, each data item has different dimensions. When joint analysis is carried out, Z-Score standardization method is selected for dimensionless standardization processing, as shown in Formula (3).

$$Znorm(x_i) = \frac{x_i - \overline{X}}{\sigma(X)}. \tag{3}$$

Where $\boldsymbol{x_i}$ represents the original value, $\overline{\boldsymbol{X}}$ represents the mean value of the original data, $\boldsymbol{\sigma(X)}$ is the standard deviation of the original data, and $\boldsymbol{Znorm(x_i)}$ is the standardized result.

3.2 Feature Selection

The complex nonlinear relationship between various indexes of air quality monitoring and $PM_{2.5}$ concentration and the multicollinearity among air quality indexes affects the performance of the model. For our specific experiment, the RF-RFE algorithm is utilized alongside Lasso algorithm to select a feature subset with stronger correlation with the results, so as to improve the prediction accuracy and efficiency of the model.

3.2.1 Feature Selection Based on RF-RFE

The RF-RFE algorithm is used to filter the 19 variables in Table 1, and the resulting feature subset is the subset corresponding to the minimum root mean square error (RMSE).

The implementation process of the RFE method is as follows:

(1) Construct a feature matrix with the given feature vectors in the data set, each row represents a sample and each column corresponds to a feature;
(2) Set the control parameters for constructing RFE function and adopt random forest function as well as the 20-fold cross-validation sampling method;
(3) RFE algorithm is used to sort these features according to their correlation with $PM_{2.5}$ concentration;
(4) Based on the ranking results, the first N (N is the number of features meeting the user's needs) features with the highest correlation with $PM_{2.5}$ concentration are selected as the featured subset.

The final number of features is 9, namely PM_{10}, NO_2, RH_ave, SO_2, RH_min, wind_ex, sun, O_3, and wind_max. As shown in Fig. 2.

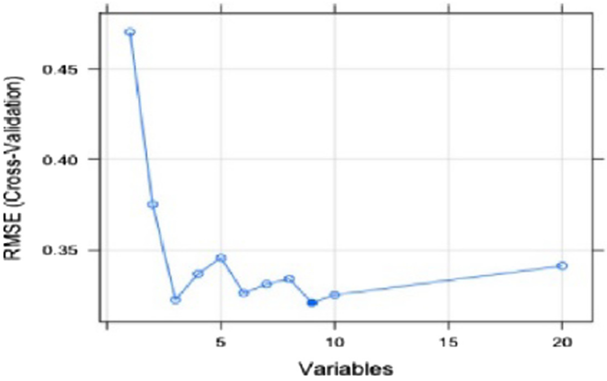

Fig. 2. RF-RFE result graph

3.2.2 Feature Selection Based on Lasso

Lasso estimation is carried out on all independent variables, and the change process of regression coefficient is shown in Fig. 3.

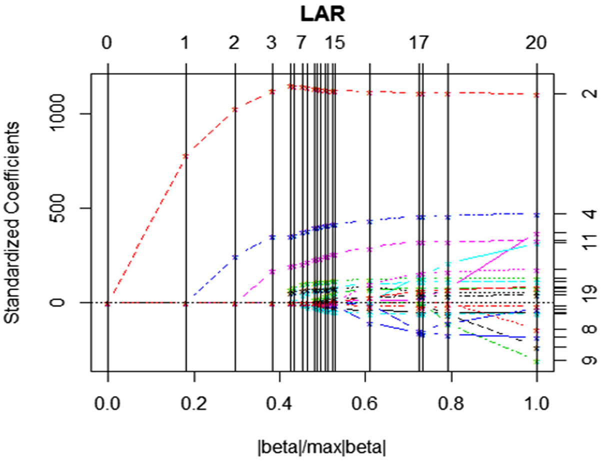

Fig. 3. Lasso estimation path map

In the figure, the abscissa serves as the value of β, or the number of steps, and the parameters represented by the ordinate appear as the vertical lines. The vertical line in the figure corresponds to the number of iterations in lasso.

The final selected feature subset is shown in Table 2.

Table 2. Feature selected with Lasso

Variable	PM_{10}	NO_2	RH_ave	RH_min	SO_2
Range	1.00	2.00	3.00	4.00	5.00
Variable	O3	wind_direction_max	season	wind_ave	wind_direction_ex
Range	6.00	7.00	8.00	9.00	10.00
Variable	precipitation	wind_max	pre_ave	wind_ex	temp_min
Range	11.00	12.00	13.00	14.00	15.00

3.2.3 Feature Subset Fusion

Common feature subset fusion methods include feature element intersection method and weighted fusion method. The general form of the weighted fusion formula is.

$$\frac{(a_1 * mathod1 + a_2 \ldots a_n * methodn)}{n} \tag{4}$$

a_1, a_2, etc. are the weighting coefficients of the feature subset, where $a_1 + a_2 + \ldots a_n = 1$.

The weighted average method is used to obtain the fused subset of the RF-RFE and the lasso feature subsets. The process is as follows: in line with the ordering of each feature in the RF-RFE feature subset and the lasso subset, the corresponding score is assigned in ascending order based on rank. The weights coefficients of both subsets are set to 50% and the total score of each feature is added and averaged. Finally, the top 12 features with the highest average score are selected as the final feature subsets and given the corresponding name. RF-lasso-T, as shown in Table 3:

Table 3. Joint feature subset

Serial number	Features	Score	Average value
1	PM_{10}	15	7.5
2	NO_2	14	7
3	RH_ave	26	13
4	SO_2	23	11.5
5	RH_min	23	11.5
6	O_3	18	9
7	wind_ex	12	6
8	wind_max	11	5.5
9	sun	9	4.5
10	wind_direction_max	9	4.5
11	season	8	4
12	wind_ave	7	3.5

3.3 Construction of $PM_{2.5}$ Concentration Prediction Model

3.3.1 Prediction Model of $PM_{2.5}$ Concentration and Influencing Factors

Based on the joint feature subsets obtained by RF-RFE and lasso methods, PM_{10}, SO_2, NO_2 concentration and other explanatory variables were introduced into the model. The effects of explanatory variables are eliminated as a result of a smoothing spline function while seasonal dummy variables are introduced to eliminate periodic effects. The degree of freedom is selected by determining the smallest sum of the absolute values of the model partial autocorrelation (PACF). The GAM model constructed by the joint feature subset RF-lasso-T is shown in formula (5):

$$Y(PM2.5) = \sum_{i=1}^{j} s(X) + \alpha. \tag{5}$$

Where Y is the concentration of $PM_{2.5}$, X refers to the variable, I is the serial number of each variable, J is the number of variables, and S is the smooth function of the model. The entire model adopts the Gaussian iteration method.

3.3.2 Analysis of Prediction Results

In agreement with the 80% and 20% proportion, the whole data set is divided into a training set and a test set.

On the training set, two GAM models were constructed, one being with the feature fusion subset RF-lasso-T and the other using the full feature subset. The execution duration of the two models is compared shown as in the following figure (Fig. 4):

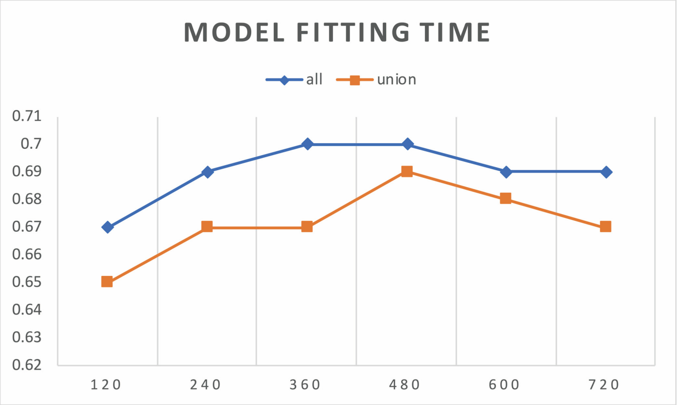

Fig. 4. Model time-consuming comparison chart

Evidently, the prediction time of the integrated GAM model is lower than that of the GAM model only whereas the deviance explained by GAM model based on RF-lasso-T is 91.5%, which is higher than that of GAM model based on RF-RFE (90.9%). The test results of RF-lasso-T-GAM model are shown in Table 4:

Table 4. RF-lasso-T GAM model hypothesis test results

Smooth factor	Coefficient	F	P
s(PM10)	**7.932**	**77.965**	**<2.00E−16**
s(NO2)	**6.81**	**13.437**	**<2.00E−16**
s(RH_ave)	**5.745**	**6.716**	**0.000000154**
s(SO2)	**1.768**	**6.836**	**0.000861**
s(RH_min)	6.637	2.26	0.019107
s(O3)	**1.524**	**23.223**	**7.36E−10**
s(wind_ex)	**8.395**	**2.62**	**0.004357**
s(wind_max)	2.816	1.226	0.420072
s(sun)	1	0.161	0.688202
s(wind_direction_max)	1	1.297	0.255324
s(wind_ave)	6.39	1.776	0.072678

Note: the bold ones are significant influence factors.

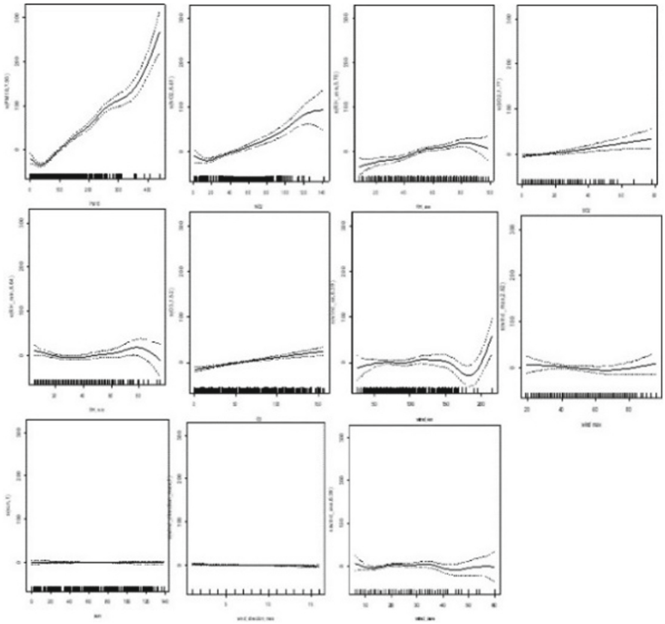

Fig. 5. Effect chart of $PM_{2.5}$ influencing factors

Figure 5 displays the effect of the influencing factors on $PM_{2.5}$. Dotted line in the figure shows the point by point standard deviation of fitting variables, i.e. the upper and the lower limits of signal interval; solid line portrays the smooth fitting curve of $PM_{2.5}$ concentration. Abscissa represents the measured value of each influence factor while ordinate denote the smooth fitting value of influence factors on $PM_{2.5}$ concentration. The value in ordinate brackets represents the estimated freedom value. The results show that the concentrations of PM_{10}, NO_2, RH_ave, RH_min, wind_ex, wind_max, wind_ave and $PM_{2.5}$ boast a non-linear relationship, while the concentrations of SO_2, O_3, sun, wind_direction_max and $PM_{2.5}$ consists of a linear relationship.

The overall data set was divided into training and test sets based on the ratio of 80% and 20%. The training model was used to simulate a 146-day data in the test set, and the $PM_{2.5}$ daily concentration value was predicted. The model prediction fit is shown in Fig. 6. The average value of $PM_{2.5}$ is 76.25, the average value of the predicted $PM_{2.5}$ is 76.27, and the root mean square error is 0.377.

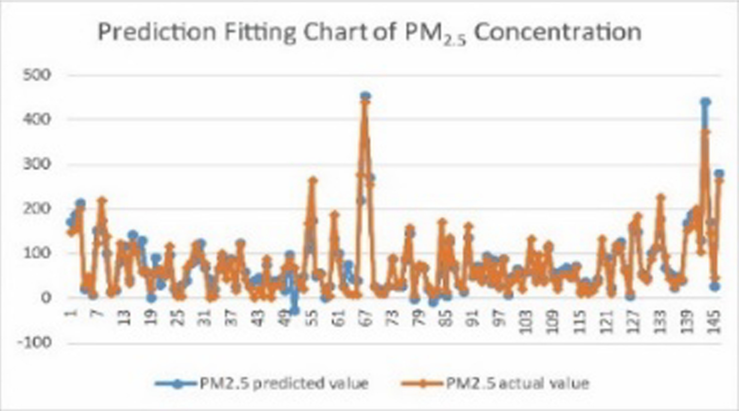

Fig. 6. Prediction fitting chart of $PM_{2.5}$ concentration

3.3.3 Analysis of Interaction Model of Influencing Factors

The change in $PM_{2.5}$ concentration can be affected by the interactions between influencing factors. In pursuance of the relationship between interactions and the concentration of $PM_{2.5}$, the smooth spline function is used to connect the influencing factors. With the combination of various pollutants and meteorological factors, an RF-lasso-T-GAM model was established. The deviance explained by the interaction model was 95.7% with an adjustment decision coefficient of 0.947. The results of the model show that cross variables such as PM_{10} and NO_2, PM_{10} and RH_ave, PM_{10} and O_3, PM_{10} and sun, NO_2 and SO_2, NO_2 and RH_min, NO_2 and O_3, NO_2 and wind_ex, RH_ave and SO_2, RH_min and O_3, RH_ave and O_3, RH_ave and wind_ex are significant at the level of $P < 0.001$. The cross terms include the interaction between air pollutants SO_2, NO_2, PM_{10} and meteorological elements as well as the interaction among air pollutants. All the above observations indicate that the change in $PM_{2.5}$ concentration is affected by the interaction between air pollutants and meteorological elements.

Taking the remarkable interaction between NO_2 and various factors as an example, the interaction model is visually plotted and shown in Fig. 7.

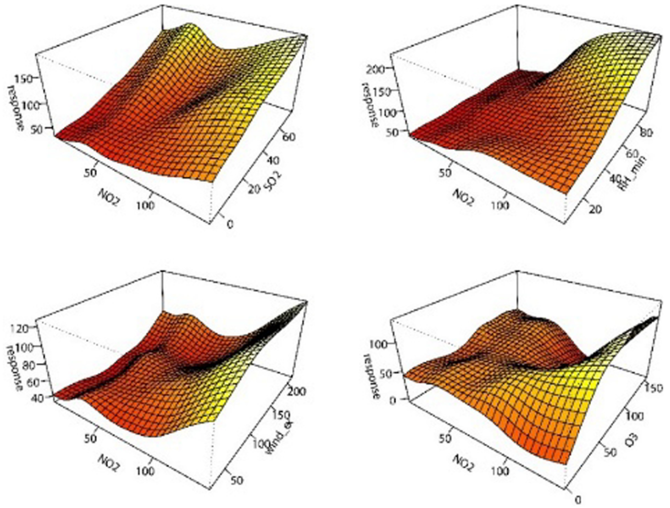

Fig. 7. Effect of interaction of influencing factors on $PM_{2.5}$ concentration

As can be seen from Fig. 7(1), when the NO_2 concentration is constant, with the increase of SO_2, the $PM_{2.5}$ concentration increases first, then decreases and then increases, which demonstrates a wave-like upward trend. When SO_2 is constant, with the increase of NO_2 concentration, the concentration of $PM_{2.5}$ shows a trend of increasing first, then decreasing and then increasing. From Fig. 7(2), when NO_2 concentration is constant, with the increase of RH_min, the concentration of $PM_{2.5}$ increases first, then decreases and then increases, but the overall trend is relatively stable. When RH_min is constant, with the increase of NO_2 concentration, the concentration of $PM_{2.5}$ shows a wave-like upward trend of increasing first, then decreasing and then increasing. From Fig. 7(3), when NO_2 concentration is constant, with the increase of wind_ex, the concentration of $PM_{2.5}$ decreases first and then increases, then increases again after a large decrease. When wind_ex is constant, with the increase of NO_2 concentration, the concentration of $PM_{2.5}$ decreases first and then increases, then increases again after a large decrease. As can be seen from Fig. 7(4), when the NO_2 concentration is constant, with the increase of O_3, the $PM_{2.5}$ concentration increases sharply first and then decreases, and then gradually increases in a wave shape. When O_3 is constant, with the increase of NO_2 concentration, the concentration of $PM_{2.5}$ shows a wave trend of decreasing first and then increasing, then there is a large decrease, and then it continues to rise.

4 Conclusion

In this study, RF-RFE and Lasso method are used to select the characteristics of the air quality data set. The runtime efficiency of the RF-lasso-T based on GAM model is higher than the GAM model constructed directly on data sets without feature selection and the deviance explained by this model is higher than the GAM model using a single feature subset of RF-RFE. The fitting results of this model can be further analyzed to obtain the meteorological and pollution factors with significant influence on $PM_{2.5}$, as well as the linear and nonlinear relationship between meteorological factors and $PM_{2.5}$ concentration. Visual results are provided to support auxiliary decision-making basis for $PM_{2.5}$ prediction and air pollution control.

Funding. This work was supported by the National Education Sciences Planning-Key Program of Ministry of Education (DLA190426).

References

1. Amsalu, E., Wang, T., Li, H., et al.: Acute effects of fine particulate matter ($PM_{2.5}$) on hospital admissions for cardiovascular disease in Beijing, China: a time-series study. Environ. Health **18**(70), 1–12 (2019)
2. Joharestani, M.Z., et al.: $PM_{2.5}$ prediction based on random forest, xGBoost, and deep learning using multisource remote sensing data. Atmosphere **10**(7), 364–373 (2019)
3. Darst, B.F., Malecki, K.C., Engelman, C.D.: Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. BMC Genet. **19**(65), 1–6 (2018)
4. Tibshirani, Robert: Regression shrinkage and selection via the lasso. J. Roy. Stat. Soc.: Ser. B (Methodol.) **58**(1), 267–288 (1996)
5. Dai, L., Mehta, A., Mordukhovich, I., Just, A., et al.: Differential DNA methylation and PM $_{2.5}$ species in a 450 K epigenome-wide association study. Epigenetics **12**(2), 139–148 (2016)
6. Hastie, T., Tibshirani, R.: Generalized additive models. Stat. Sci. **1**(3), 297–310 (1986)