Prediction of PM2.5 Concentration Based on Time Series

Qirui Xu
Department of Information Science, Beijing University of Technology, Beijing, China
xuqirui@bjut.edu.cn

Abstract. PM$_{2.5}$ is an important component of air pollution, its chemical composition and source are relatively complex, and it stays in the atmosphere for a long time, which has seriously endangered public health, damaged the ecosystem and even affected the climate and environment. Therefore, the prediction of PM$_{2.5}$ is of great significance for public health protection and environmental management. One of the reasons why it is difficult to predict PM$_{2.5}$ is the interaction between air pollutants, so a method that can not only accurately predict the change of PM$_{2.5}$ concentration, but also explain the results is needed. In this paper, different surface meteorological information is used as time series data to predict the hourly PM$_{2.5}$ concentration in Wuhan. The prediction can explain the influence of input variables, and use random forest to learn and predict the data. In addition, the meteorological information, air pollutant information and dust value of Zhengzhou City grouped by time series are used as input variables to reflect the air characteristics of North China. This method can learn all the specific time information of the input variables processed by time series in a balanced way, and can also explain the related effects of the input variables. In addition, it also shows that the related variables in North China have an important impact on the generation of PM$_{2.5}$ in Wuhan.

Keywords: Air quality prediction, PM$_{2.5}$, random forest, time series data, machine learning

1. Introduction
Since the 20th century, with the rapid development of the world economy, the growing population and industrial production scale, as well as the sharp increase of energy consumption based on fossil energy, such as coal and oil, have led to a sharp increase in the content of gas and solid pollutants in the air, which eventually led to air pollution. Studies have shown that if people are exposed to air pollution for a long time, they will have symptoms such as boredom, low spirits, dizziness and inattention [1]. Short term exposure to high concentration of polluted air can lead to the increase of morbidity, mortality and total mortality of cardiopulmonary system [2]. Air pollutants mainly include particulate matter (PM$_{2.5}$, PM$_{10}$, dust, etc.), carbon monoxide (CO), nitrogen oxides (NO$_2$, NOx, etc.), sulfur oxides (SO$_2$, etc.), ozone (O$_3$), hydrocarbons (methane, ethylene, etc.), mercury, volatile organic compounds, lead and toxic substances, etc. PM$_{2.5}$ (fine particle) refers to the particles in the ambient air whose aerodynamic equivalent diameter is less than or equal to 2.5 microns. It is an important component of air pollution and one of the main bottlenecks to improve the air environmental quality. The chemical composition and
source of PM$_{2.5}$ are more complex, the retention time in the atmosphere is longer, and the impact on health, environment and climate is more significant [3].

A large amount of PM$_{2.5}$ will be produced when the atmosphere stagnates, so the weather condition is very important, such as temperature, wind speed, precipitation and other weather information needs to be taken as variables. In addition, PM$_{2.5}$ is caused by the chemical reaction between gases produced by fossil fuel combustion, so the information about air pollutants such as CO, NO$_2$ and O$_3$ should be taken as variables. In addition, when a large number of air pollutants or fine dust (raw materials used to produce PM$_{2.5}$) enter Wuhan from other areas, a large amount of PM$_{2.5}$ will be produced. Therefore, we need to take the weather information, air pollutant information and dust level near Wuhan as variables. Since the generation of PM$_{2.5}$ takes more than one hour, it is necessary to group the input variables as a time series to reflect the atmospheric trend.

There is a nonlinear interaction between air pollutants, and one of the reasons for the difficulty of PM$_{2.5}$ prediction is the interaction between input variables. Therefore, when predicting PM$_{2.5}$, we need to explain the related problems of the impact of input variables. In this paper, different surface meteorological information is used as time series data to study the hourly PM$_{2.5}$ concentration prediction in Wuhan city. The prediction can explain the influence of input variables, and the adjusted bootstrap number random forest is used to learn and predict the data. In addition, the meteorological information, air pollutant information and dust value of Zhengzhou City grouped by time series are used as input variables to reflect the air characteristics of North China.

2. Related work

Farah et al. [4] applied the time series method to the urban air quality of Beirut, and introduced the autocorrelation function of time series to realize the relationship between other factors except SO$_2$. However, this study does not reflect the relevant information of air pollutants that cause secondary fine dust. Huang et al. [5] predicted monthly PM$_{2.5}$ concentrations in northern China using a random forest model based on aerosol optical thickness data and ground observation data.

Durbin et al. Uses ARMA to predict the fine dust in unit time. However, these methods only calculate the change of fine dust [6], which is easily affected by the rapid change of fine dust nonlinearity, and its performance is poor compared with the deep learning or mixing model (learning multiple input variables except dust). Wang et al. [7] used Arima, a single variable time series prediction model, to model PM$_{2.5}$, and extract the timing characteristics of PM$_{2.5}$ data, but did not consider the correlation between other pollutants and PM$_{2.5}$. The model is relatively simple and the prediction accuracy is not high.

3. Research method

3.1. Input variables

| Wuhan (Numeric) | Variable Type | Description |
|-----------------|---------------|-------------|
| Weather         | Temperature, pressure, humidity, wind speed, rain(5 Variables) |
| Air Pollutant   | NO$_2$, CO, O$_3$, SO$_2$(4 Variables) |
| zhengzhou (Numeric) | Weather | Temperature, pressure, humidity, wind speed, rain(5 Variables) |
| Air Pollutant   | NO$_2$, CO, O$_3$, SO$_2$(4 Variables) |
| particulate Matter | PM$_{10}$, PM$_{2.5}$(2 Variables) |
| Target (Category) | particulate Matter | PM$_{2.5}$ of WuHan |

In this paper, the hourly air quality monitoring data is used to predict the hourly PM$_{2.5}$ concentration in Wuhan. In the monitoring data, the weather information and air pollutants in Wuhan are used as input variables. In addition, the weather information, air pollutant information and dust value of Zhengzhou are used as input variables to reflect the air characteristics of North China north of Wuhan. The purpose
of this study is to reflect the various meteorological flows used for PM$_{2.5}$ prediction in Wuhan by learning the surface observation data of Wuhan and Zhengzhou at the same time. There are 20 input variables in this study, and their composition is shown in Table 1 below.

### 3.2. Data time series preprocessing

In this study, time series preprocessing was carried out, in which input variables and predicted targets were grouped by time step $n$ to predict hourly PM$_{2.5}$ concentration in Wuhan. In this way, all input variables of a certain time unit can be trained in the prediction model.

When all 20 input variables of time $t_i$ are $X(t_i)$, the formula (1) for predicting PM$_{2.5}$ concentration in Wuhan at $t_{n+1}$ by learning the n hour pretreatment data from $t_1$ to $t_n$ in the model is as follows

$$\text{Predicted PM}_{2.5}(t_{n+1}) = \text{Model}(X(t_1), X(t_2), ..., X(t_n))$$

The input value (excluding the predicted PM$_{2.5}$ concentration in Wuhan) was normalized to a value between 0 and 1. In some studies, if there is a missing value in a variable, the missing value is corrected by the average value of the variable. However, in this study, when a variable value is missing during time step preprocessing, it is removed from the dataset.

### 3.3. Learning model

In this paper, random forest classifier [8] is used as a learning model. In the case of using LSTM in recent dust prediction research, the longer the input data series or time series, the more the previously input information is lost [9], which is difficult to explain the related influence of input variables. In addition, in the case of hybrid model, the ability to explain the influence of input variables may be worse than that of deep learning model. On the other hand, because random forest can learn all input variables in a balanced way, it can learn without losing the time-specific information of all input variables when input time series preprocessing data. Compared with the previous model, the advantage of this model is that the explanatory power of the results is relatively good. These characteristics of random forest can bring synergistic effect with the input data of this study.

Random forest is a method that can be used for both classification and regression. By selecting multiple results from the prediction results of multiple decision trees with smaller deviation, a single learning result with smaller variance and deviation can be obtained, which is an ensemble tree technology. In this case, bootstrap bootstrap (a method of sampling multiple training data by recovering and extracting as much training data as the whole data size) is used to create a decision tree with smaller deviation. In addition, in order to prevent intensive learning of specific variables, when dividing the nodes of the decision tree, $m$ new input variables selected each time among the total $q$ input variables are considered as the goal of node division. In this case, the number of variables $m$ is usually determined as $m \approx \sqrt{q}$ or $m \approx \log_2 q$. Suppose that the total number of bootstrap samples (decision trees) for random forest training is $B$, and each bootstrap sample set is $b_j$. The final prediction result $\text{Pred}(b_j)$ of single tree of random forest classifier is determined as formula (2).

$$\text{Predicted PM}_{2.5}(t_{n+1}) = \text{Majority Vote}[\text{Pred}(b_1), \text{Pred}(b_2), ..., \text{Pred}(b_B)]$$

Random forest classifier calculates the importance of variables as the average value of the amount of information that can be obtained when a specific variable divides the data in each decision tree. Let $\hat{p}(b_j, k, x_q(t_i))$ represents the probability that the data of the sample $b_j$ guided by the specific variable $x_q$ based on time $t_i$ is marked as $k$. when the information standard is set to "entropy", the expected value of $I_{b_j}$ is the information amount in the sample $b_j$ of the variable $x_q(t_i)$ as shown in formula (3).

$$I_{b_j}(x_q(t_i)) = - \sum_{k=1}^{K} \hat{p}(b_j, k, x_q(t_i)) \times \log \hat{p}(b_j, k, x_q(t_i))$$
In this case, the variable $x$ in the random forest classifier with a total of $B$ decision trees is used. The importance of is shown in formula (4).

$$I_B(x_q(t_i)) = \frac{1}{B} \sum_{b=1}^{B} I_B(x_q(t_i))$$

(4)

Since the preprocessed data in this study has input variables every time, in order to measure the importance of specific variables from the overall perspective, the importance of specific variables is summarized by time, as shown in formula (5).

$$TotalI_B(x_q) = \sum_{i=1}^{n} I_B(x_q(t_i))$$

(5)

The importance of each variable obtained from equation (5) is converted to the relative percentage of all variables.

4. Experiment

4.1. Data

In this study, the data came from the real-time air quality release platform of the Ministry of environmental protection, and was released by Qingyue open environmental data center (http://data.epmap.org). It includes hourly weather information data from April 29, 2018 to April 29, 2019. The hourly data of air pollutants and dust in the same period are used for learning, verification and testing. The total number of data of each preprocessing is slightly different according to the time step, and the training, validation and test data are randomly allocated in the ratio of 6:2:2.

4.2. Prediction labels

In this study, the prediction labels were classified according to the ambient air PM$_{2.5}$ classification standard currently used in China. According to the classification criteria, when PM$_{2.5}$ (average $\mu g/m^3$ per hour) is 0-50, it is "excellent"; when PM$_{2.5}$ is 51-100, it is "good"; when PM$_{2.5}$ is 101-150, it is "mild pollution"; when PM$_{2.5}$ is 151-200, it is "moderate pollution"; when PM$_{2.5}$ is 201-300, it is "severe pollution"; when PM$_{2.5}$ is greater than 300, it is "severe pollution". However, due to the large number of grades and the small percentage of "heavy pollution" labels, in this study, "light pollution" and "moderate pollution" were classified as "medium pollution", and "heavy pollution" and "serious pollution" were classified as "bad pollution". The data quantity of each tag after preprocessing is shown in Table 2.

| Table 2. Number of Data for Each Target Label (Timestep=2) |
|-------------------------------|----------------|----------------|-----------|--------|-------|
| Train                         | excellent     | good           | medium    | bad    | total |
| 662                           | 3096          | 1440           | 58        | 5256   |
| Validation                    | 217           | 1029           | 487       | 19     | 1752  |
| Test                          | 231           | 1041           | 460       | 20     | 1752  |

4.3. Setting learning model parameters

In this study, in order to determine the number of bootstrap samples (number of decision trees) $B$ to be used in the random forest, the parameters were changed, and the changes of model performance were observed on the validation data. Ubuntu 16.04, python 3.5 and sklearn libraries are used to implement random forest. One hour after learning is predicted by dividing the time step into two hours. The results are shown in Table 3 and Figure 2.
Table 3. F1-Scores by Tree Number on Validation Data

| B   | PM2.5 Label | excellent | good | medium | bad  |
|-----|-------------|-----------|------|--------|------|
| 1   |             | 0.701     | 0.741| 0.710  | 0.676|
| 10  |             | 0.776     | 0.817| 0.788  | 0.756|
| 30  |             | 0.789     | 0.824| 0.801  | 0.773|
| 100 |             | 0.796     | 0.828| 0.806  | 0.780|
| 200 |             | 0.797     | 0.828| 0.807  | 0.781|
| 300 |             | 0.800     | 0.829| 0.808  | 0.782|

It can be seen from table 3 that when B exceeds 200, the improvement of prediction accuracy is very small. With the increase of B, the learning time also increases, so in this study, the value of B is set to 200.

In the case of using tree technology, pruning can prevent the model from over fitting the training data. Due to the results of Pruning Experiment, the learning model of this study trains the nodes of the tree to the end on the uncut training data, and the performance of the verification data is very good. This can be explained as the characteristics of repeating and cycling weather information of similar patterns in the learning and validation data.

In addition, the information index of the model node partition standard is entropy, not Gini coefficient, and the number of input variables m should not be regarded as a random number when dividing each node. When it is set to an approximate value, it shows higher performance. Not approximations. Table 4 shows the super parameters of the learning model determined by validation data.

Table 4. Hyperparameters for the Proposed Random Forest

| Name                  | Value    |
|-----------------------|----------|
| n_estimators          | 200      |
| bootstrap             | True     |
| criterion             | ‘entropy’|
| max_depth             | None     |
| max_features          | ‘log2’   |
| min_impurity_decrease | 0        |

4.4. Next time prediction results by time step
This study compared the learning performance of the model in time series by gradually adjusting the time step of input data from 1 hour to 12 hours, so as to compare the performance of PM2.5 concentration prediction after 1 hour. The results are shown in Figure 1.
As shown in Figure 2, the prediction accuracy of the "good" tag increases until the time step is between 4 and 5, and then decreases. The label "excellent", "medium" and "bad" have fluctuations, but when the time step is 11, the prediction accuracy is the highest. Generally, when the time steps in all tags have a certain value instead of a small timestep, the prediction accuracy will be improved, which seems to reflect the characteristics of PM$_{2.5}$ and takes some time to generate. Similarly, since the accuracy of "medium" and "bad" labels is considered to be the highest at time step 11, it can be seen that it takes a long time to predict high concentration PM$_{2.5}$, and the decision boundary of "bad" label is the most complex.

4.5. Importance of variables
In this study, in order to explain the related effects of input variables, the importance of each variable is set as the relative percentile score, and the top 10 are extracted based on the specific variables and the amount of information that can be obtained during data segmentation. The results are shown in Table 5.

| Rank | Variable | Timestep=1 | Importance | Variable | Timestep=12 | Importance |
|------|----------|------------|------------|----------|-------------|------------|
| 1    | O$_3$    |            | 14.25      | O$_3$    |             | 15.13      |
| 2    | NO$_2$   |            | 11.18      | NO$_2$   |             | 11.27      |
| 3    | zPM$_{2.5}$ |          | 9.83       | zPM$_{2.5}$ |           | 8.76       |
| 4    | SO$_2$   |            | 7.94       | SO$_2$   |             | 7.76       |
| 5    | zPM$_{10}$ |           | 7.12       | zPM$_{10}$ |            | 7.01       |
| 6    | CO       |            | 5.13       | CO       |             | 5.47       |
| 7    | wind speed |           | 4.91       | zNO$_2$  |             | 5.01       |
| 8    | wind speed |           | 4.50       | wind speed |          | 4.92       |
| 9    | rain     |            | 4.13       | zwind speed |         | 4.23       |
| 10   | humidity |            | 4.07       | humidity |             | 4.01       |

It can be seen from table 5 that O$_3$, NO$_2$, SO$_2$ and co related to PM$_{2.5}$ production rank higher, while fine dust (zPM$_{2.5}$, zPM$_{10}$) in Zhengzhou also ranks at the top. Regardless of the time step, it will show this trend. In other words, in the PM$_{2.5}$ prediction, the variables related to PM$_{2.5}$ are the main, and the dust variables related to Zhengzhou are also very important, so it is also important to consider the impact of North China on air quality pollution in Wuhan.

4.6. Contrast experiment
In this study, a comparative experiment between the proposed method and the existing model is carried out. In the contrast experiment, logistic regression and LSTM (long short memory), which are commonly used for classification and prediction, are used. The implementation uses sklearn and tensorflow 1.14-gpu respectively, and the graphics card specification is geforce GTX 10606gb. In these two models, the best super parameters are determined by grid search method in the validation data.

Saga [10], an effective method for learning large amounts of data, is used as an optimization method in logistic regression, and the prediction performance is verified by tolerance (tol), normalized value (c) and penalty parameter. The best super parameters of logistic regression are shown in Table 6, and the verification performance is shown in Table 7.

| Name | Hyperparameters                      | Optimal |
|------|--------------------------------------|---------|
| penalty | ['l1', 'l2', 'elasticnet']          | 'l1'    |
| tol   | [0.00001, 0.0001, 0.001, 0.01, 0.1]  | 0.1     |
| C     | [5.0, 4.0, 3.0, 2.0, 1.0, 0.1]       | 1.0     |
When using LSTM, the optimizer uses Adam [11], which is used in many recent studies, and verifies the prediction performance with the recommended learning rate of 0.001 and other learning rates. In addition to the learning rate, also through the batch_size, num_units and number of cell verify the prediction performance. The maximum number of epoch is set to 200, but when the loss value of the model in the validation data increases, it can be automatically determined by early stopping. The optimal super parameters of LSTM are shown in Table 8, and the verification performance is shown in Table 9.

Table 7. F1-Score of Hyperparameters for Logistic Regression

| penalty   | tol | C  | F1-Score for ‘Bad’ Label |
|-----------|-----|----|-------------------------|
| L1        | 0.1 | 1.0| 0.4864                  |
| L1        | 0.1 | 2.0| 0.4863                  |
| elasticnet| 0.1 | 5.0| 0.4859                  |
|           |     |    |                         |
| L2        | 0.1 | 1.0| 0.4811                  |
|           |     |    |                         |
| L1        | 0.00001 | 2.0| 0.4459                  |
| L1        | 0.00001 | 2.0| 0.4459                  |

Table 8. Hyperparameters for LSTM

| Name               | Hyperparameters       | Optimal |
|--------------------|-----------------------|---------|
| batch_size         | [512, 256, 128]       | 256     |
| num_units(cell output) | [512, 256, 128] | 512     |
| number of cell     | [3, 2, 1]             | 2       |
| learning_rate      | [0.01, 0.001, 0.0001, 0.00001] | 0.0001 |
| epochs             | -                     | Early Stop |

Table 9. F1-Score of Hyperparameters for LSTM

| number of cell | batch size | num units | learning rate | F1-Score for ‘Bad’ Label |
|----------------|------------|-----------|---------------|-------------------------|
| 2              | 256        | 512       | 0.0001        | 0.6771                  |
| 3              | 256        | 256       | 0.0001        | 0.6754                  |
| 1              | 256        | 512       | 0.0001        | 0.6696                  |
| ...            | ...        | ...       | ...           | ...                     |
| 1              | 128        | 256       | 0.001         | 0.6641                  |
| 3              | 512        | 256       | 0.001         | 0.6640                  |
| ...            | ...        | ...       | ...           | ...                     |
| 1              | 512        | 256       | 0.01          | 0.6354                  |
| ...            | ...        | ...       | ...           | ...                     |

In the process of LSTM super parameter verification, if the learning rate is between 0.01 and 0.0001, the loss value in epoch 200 increases, and the learning is terminated automatically. However, if the learning rate is 0.00001, the learning cannot be completed within epoch 200. After that, the learning rate is set to 0.00001, and then several pairs of super parameters are randomly selected to train and learn without epoch restriction. Each group of super parameter combination needs more than 1000 epoch and more than 11 hours, but it does not show higher performance.

When the time step is 2, the experimental comparison results between the proposed method and the best super parameter logistic regression and the prediction performance of LSTM test data are shown in table 10.
Table 10. F1-Scores of the Proposed Method and Other Models (Timestep=2)

| Model         | excellent | PM$_{2.5}$ Label | good | medium | bad  |
|---------------|-----------|------------------|------|--------|------|
| Logistic Regression | 0.502     | 0.701            | 0.543| 0.488  |      |
| LSTM          | 0.734     | 0.792            | 0.601| 0.675  |      |
| Logistic Regression | 0.779     | 0.808            | 0.624| 0.679  |      |

It can be seen from table 10 that the prediction performance of the proposed method is the best for all tags. In particular, this method has higher performance than LSTM, which has advantages in time series learning. In addition, this method can also calculate the importance of variables that are difficult to understand in deep learning model.

Secondly, taking the important difference of "bad" label in PM$_{2.5}$ prediction as the object, a comparative experiment is carried out by changing the time step. The results are shown in Figure 2.

![Figure 2. F1-Scores for ‘bad’ Label of Proposed Method and Other Models by Timestep](image)

It can be seen from Figure 2 that the prediction performance of the proposed method shows the highest performance in all time steps, and the trend of equalization.

In the case of LSTM, it fluctuates according to the time step, and shows similar performance in time steps 2, 4 and 5. In the case of logistic regression, there are some problems, the prediction accuracy increases with the increase of time step, but the fluctuation range is large, and the overall performance is the lowest. The results show that the proposed method has the best performance in all tags and time steps.

5. Conclusion

This study proposes a method to predict PM$_{2.5}$ concentration per hour in Wuhan by using random forest. The method adjusts the number of bootstrap programs after preprocessing different ground observation data into a time series. The experimental results show that the proposed method has better performance than the existing models in all tags. In addition, in order to explain the related influence of input variables, the importance of variables is calculated by summing the importance of each variable according to the time step. The results show that the generation of PM$_{2.5}$ and the variables related to the influence of North China are very important.

In the future, by expanding the research scope, the research scope can be extended not only to other domestic cities, but also to other countries and regions. In addition, it can be expected that the proposed method combined with deep neural network can provide additional performance improvement.
References

[1] Xavier Morelli, Camille Rieux, Josef Cyrys, Bertil Forsberg, Rémy Slama. Air pollution, health and social deprivation: A fine-scale risk assessment [J]. Environmental Research, 2016, 147

[2] Victoria N Likhvar, Mathilde Pascal, Konstantinos Markakis, et al. A multi-scale health impact assessment of air pollution over the 21st century [J]. Science of the Total Environment, 2015, 514: 439-449.

[3] Huang, J., Pan, X., Guo, X. and Li, G. (2018). Impacts of air pollution wave on years of life lost: A crucial way to communicate the health risks of air pollution to the public. Environ. Int. 113: 42–49

[4] A. Charbel, C. Carine, B. Agnes et al., SO2 in Beirut: air quality implication and effects of local emissions and long-range transport, Air Quality Atmosphere and Health, 1, 2008, 167-178.

[5] K. Huang, Q. Xiao, X. Meng, G. Geng, Y. Wang, A. Lyapustin, D. Gu, and Y. Liu, “Predicting monthly high-resolution PM2.5 concentrations with random forest model in the North China Plain,” Environmental Pollution, Vol. 242, No.A, pp.675-683, 2018.

[6] Y. Lin, N. Mago, Y. Gao, Y. Li, Y. Y. Chiang, C. Shahabi, and J. L. Ambite, “Exploiting Spatiotemporal Patterns for Accurate Air Quality Forecasting using Deep Learning,” ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL), pp.359-368, Nov. 2018.

[7] Wang T J, Jiang F, Deng J J, et al. Urban air quality and regional haze weather forecast for Yangtze River Delta [J]. Atmospheric environment, 2012, 58 (15): 70-83.

[8] G. James, D. Witten, T. Hastie, and R. Tibshirani, “An Introduction to Statistical Learning with Applications in R,” Springer, 2017.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), pp.5998-6008, Dec. 2017

[10] A. Defazio, F. Bach, and S. Lacoste-Julien, “SAGA: A Fast Incremental Gradient Method With Support for NonStrongly Convex Composite Objectives,” Advances in Neural Information Processing Systems 27 (NIPS 2014), Jul.2014.

[11] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015), May.2015.