Research on Traceability of Atmospheric Particulate Pollutants Based on Particle Size Data

Haonan Yu[1], Yunbao Zhou[1], Yuhuan Jia[2], Jingjin Ma[2], Benfeng Pan[2], Wei Zhou[1(✉)], and Yang Chen[1]

[1] Key Laboratory of Big Data and Artificial Intelligence in Transportation, Ministry of Education, School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China
wzhou@bjtu.edu.cn

[2] Hebei Sailhero Environmental Protection High-Tech Co. Ltd., Shijiazhuang 050035, Hebei, China

Abstract. Particle pollution is one of the important sources of air pollution. With the maturity and portability of domestic particle size monitoring equipment, a large amount of air particle size monitoring data has been generated. How to use these data to analyze pollution sources and propose corresponding prevention measures is currently an urgent problem to be solved. Firstly, this study analyzed particle size data and determined the distribution characteristics of particle size in different pollution sources; Secondly, a traceability model based on random forest and factor analysis was constructed to achieve the problem of analyzing air pollution sources using particle size spectrum data; Finally, through experimental comparison, case analysis, and expert experience, the model was verified its effectiveness in the actual situation. This study is the first to use pollutant particle size monitoring data, which achieves high-resolution pollutant traceability compared to traditional chemical composition methods. This study introduces machine learning methods into traditional factor analysis method to improve processing efficiency and accuracy, providing a reference basis for air pollution control.

Keywords: Random Forest · Factor Analysis · Particle Size Data · Pollutant Traceability

1 Introduction

1.1 Research Background

Air quality is related to people's well-being. Quickly identifying the location, time, and category of atmospheric pollution sources has important research value for the formulation of prevention and control measures. At present, pollutant traceability methods are mainly carried out through chemical component analysis. These methods have high recognition, but the detection instruments are expensive, the timeliness is poor, and the

W. Hong and G. Kanaparan (Eds.): ICCSE 2023, CCIS 2023, pp. 429–441, 2024.
https://doi.org/10.1007/978-981-97-0730-0_38

time resolution is low. In recent years, some companies have developed fine-grained particle size monitoring equipment and piloted it in some regions, aiming to analyze pollution sources through atmospheric particle size data. Therefore, this study conducts research on pollutant traceability based on particle size monitoring data.

The physical and chemical properties of particles mainly depend on their particle size [1]. We are familiar with air particle pollutants such as PM2.5 and PM10, in which 2.5 and 10 refer to particle size. The properties and sources of particles with different particle sizes may also be different. From Fig. 1, it can be seen that the transformation relationships between them are complex.

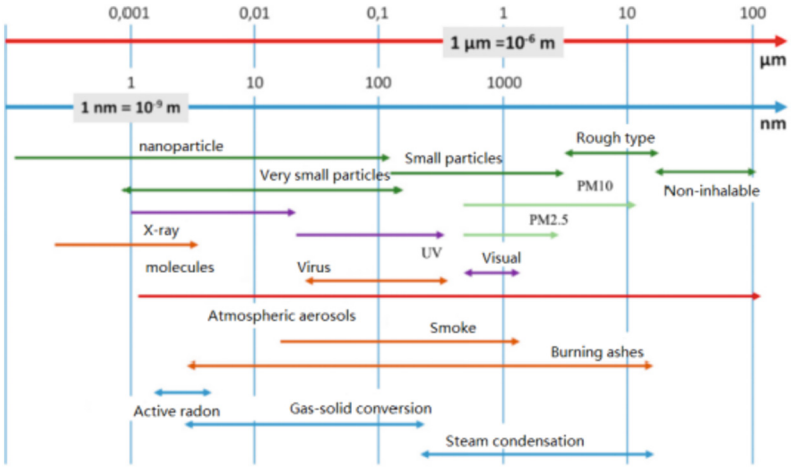

Fig. 1. Particle size fractionation of particulate matter [source: © J. Fontan].

Qualitatively or quantitatively identifying the source of atmospheric particulate pollutants in an environmental receptor through chemical, physical, and mathematical methods is called traceability of particulate pollutants [2, 3]. How to design a reliable and effective method for the traceability through particle size to achieve the purpose of environmental monitoring is a meaningful work at present.

1.2 Related Studies

Currently, there is less research literature on particulate matter, and higher resolution particle size spectrum monitoring data are difficult to obtain publicly in large quantities. Moreover, most of the current particle size data have fewer features, there are no more comprehensive datasets of particle size, and most of the time periods are relatively short.

Traditional air particulate pollutants traceability methods [4, 5] mainly include the source inventory method, source diffusion model method, and receptor model method, and there are fewer studies on traceability based on particulate particle size data.

The source inventory method refers to the real-time and accurate monitoring of the pollution emissions of each pollution source, and then the pollution traceability

according to the emissions of each source. This method requires the monitoring of the specific emissions of each pollution source in the study area, so it requires a lot of financial and human resources and is difficult to implement.

The source diffusion model method [6] is a mathematical modeling approach to simulate the processes of diffusion, transmission, and deposition of pollutants, which fully considers external influences such as meteorology and topography. The disadvantages of the diffusion model method mainly include errors in the simulation process of the meteorological field.

The receptor model method refers to the sampling of discrete sampling points in the study area, and then analyzing the pollutant composition of the sampling points by physical or chemical methods, so as to obtain the proportion of the contribution of each pollution source to the points. Common methods for source analysis based on receptor models include chemical mass balance (CMB), probabilistic matrix factorization (PMF), principal component analysis (PCA), factor analysis (FA), and multi-source linear analysis (UNMIX), etc.

The study of traceability of particulate pollutants means analyzing the contribution proportion of pollutants by mathematical or physicochemical methods. Huang et al. [7] summarized the receptor model methods in the field of pollution source tracing and introduced the commonly used receptor models and development overviews at domestically and abroad. With the improvement of technology in recent years, Zheng et al. [8] introduced the source analysis methods for PM2.5 in China and proposed the future development direction of particulate matter source analysis research, which is an important scientific reference and reference value for future PM2.5 source analysis work. Wang [9] carried out a source analysis of atmospheric VOCs and used a factor analysis model to obtain the contribution ratio of each component.

Zhang et al. [10] used a positive definite factor model to analyze the source of winter road particulate matter particle size distribution in Tianjin and obtained three main sources of road dust, tire wear, and motor vehicle tailpipe emission aging. Huang et al. [11] analyzed the heavy metal pollution sources in agricultural soil at different scales and established a soil heavy metal pollution source analysis method adapted to different scales on the basis of the traditional source analysis work.

Chen et al. [12] used machine learning methods to study the source analysis of air pollution in Shanghai, using GBRT and Random Forest algorithm to establish a source analysis model and put forward some suggestions and measures for the prevention and control of air pollution in Shanghai. Chen et al. [13] compared hour-by-hour PM2.5 concentration prediction using multiple machine learning models, and XGBoost model, LightGBM model and random forest model were comparatively analyzed.

For the traceability of particulate pollutants, there are relatively few studies domestically and abroad, and most of them are based on the methods of factor analysis. While the existing factor analysis analytical methods have low temporal resolution, consume computational resources and require experts' experience, etc. Therefore, this study will explore the problem of pollution traceability from two technical routes factor analysis and machine learning.

1.3 Purpose

This study aims to quickly obtain the proportion of pollutants and achieve traceability through particle size data, based on factor analysis and machine learning methods. This study has two originalities: it is the first to use pollutant particle size data, which achieves high-resolution pollutant traceability compared to traditional chemical composition methods; it introduces machine learning methods into traditional factor analysis methods to improve processing efficiency and accuracy, providing a reference basis for air pollution control.

Traditional air pollution source methods are seldom analyzed by using particle size spectrum data, so this study has some practical significance in engineering; due to less research on machine learning in this field [14], this study has some theoretical research significance.

2 Data Analysis

2.1 Particle Size Data Description

The data used in this study were collected by particle size monitoring equipment developed by Hebei Advanced Environmental Protection Industry Innovation Center Co. The equipment is designed based on the principle of high-precision light scattering particle counting. The core element of the equipment is an optical particle sensor, which analyzes the Lorenz-Mie scattered light of individual particles in order to determine the particle size. During the measurement, individual particles will pass through an optically differentiated measurement space that is uniformly illuminated using a multi-colored LED light source. Each particle generates scattered light pulses detected at an angle of 85°-95°, and the number of particles is determined based on the number of scattered light pulses, with the level of the measured scattered light pulses being the level of the particle size.

All particle size data captured by the equipment have 2 types of attributes: mass concentration (mass of particles per unit volume, $\mu g/m^3$) and number concentration (number of particles per unit volume, count/m^3) respectively. Because single type cannot fully capture the particle size distribution, the two attributes are used in the study.

The particle size data for this study includes two categories:

1. Pollution source spectrum data: about 20,000 items from seven types of pollution sources (categories are shown in Fig. 2), including 83 dimensions (11 mass concentration attributes and 72 number concentration attributes), collected in several places and specific pollution source situation.
2. Particle size monitoring data: about 30,000 items from Baoding City by continues monitoring every second from April 1st to 30th in 2023, which contains 83 dimensions of particle size data. During the period, there is an obvious dust weather event started at 21:00 on April 10th, with a PM10 concentration of 459 $\mu g/m^3$, peaked at 2084 $\mu g/m^3$ at 02:00 on the 11th, and ended at 19:00 on the 11th, with a PM10 concentration of 166 $\mu g/m^3$.

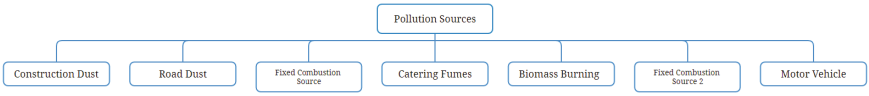

Fig. 2. Seven types of pollution sources

2.2 Characterization of Pollution Sources Based on Pollution Source Spectrum Data

Through the preliminary analysis of the pollution source spectrum data, it can be plotted to characterize the spectrum of different pollution sources, Fig. 3 is the distribution of the mass concentration attribute of different 7 pollution sources. It can be seen that the mass concentration distribution of different pollution sources has different feature, for example, construction dust has lower concentrations in small particle size segments and higher concentrations of PM7-10 in large particle size segments; road dust also has lower concentrations in small particle size segments and higher concentrations in large particle size segments, but the highest concentration occurs in PM2.5-4.

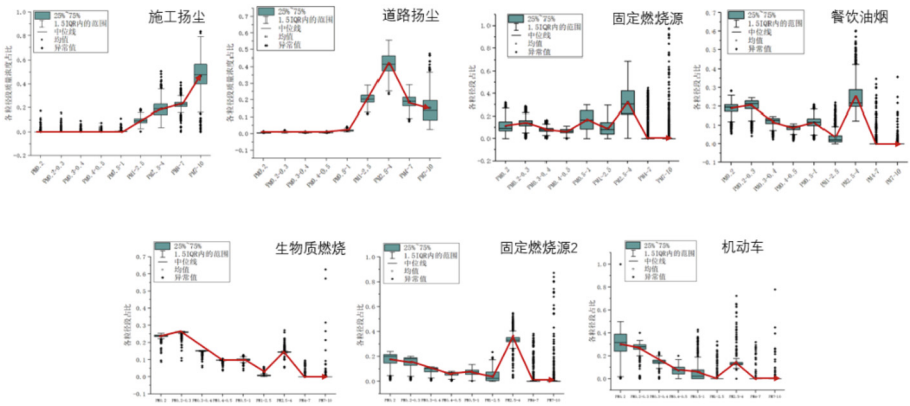

Fig. 3. Distribution of mass concentration from 7 pollution sources (Top (from left to right): 1. Construction dust, 2. Road dust, 3. Fixed combustion source, 4. Catering fumes. Bottom: 5. Biomass burning, 6. Fixed combustion source 2, 7. Motor vehicle)

2.3 Correlation Analysis of Particle Size Monitoring Data

Correlation analysis is one of the important statistical methods to study the correlation between attributes. The correlation coefficient can be used to describe the relationship between variables (attributes) and the degree of linear correlation between two variables by calculating the Pearson correlation coefficient, which takes a value between -1 and 1. Negative values indicate a negative correlation, positive values indicate a positive correlation, and absolute values closer to 1 indicate a stronger correlation, while values closer to 0 indicate a weaker correlation.

According to the formula of Pearson correlation coefficient, it can be calculated to obtain the correlation coefficient matrix, which was plotted as a heat map as shown in Fig. 4. It can be observed that there is a positive correlation among all particle sizes. Some particle size monitoring data exhibit strong correlations with each other. For instance, the correlation coefficients between PM0.2, PM0.2-0.3, and PM0.3-0.4 are 0.68, 0.67, and 1.00, respectively. The correlation coefficient between PM1-2.5 and PM2.5-4 is 0.88, while the correlation coefficient between PM10-15 and PM15-20 is 0.92, indicating a strong correlation. On the other hand, there is relatively weak correlation between large particle pollutants and small particle pollutants.

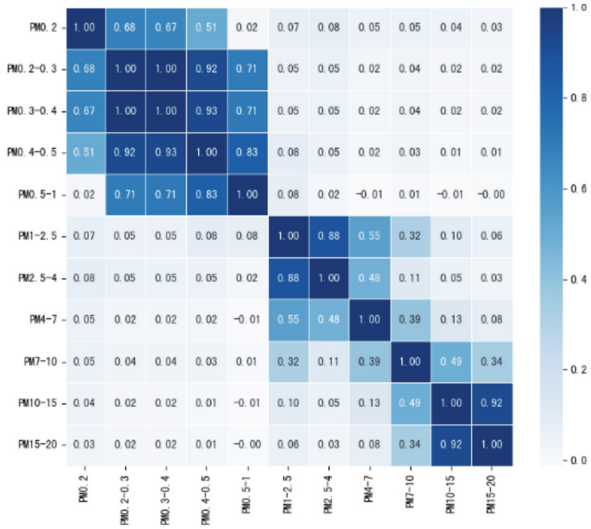

Fig. 4. Schematic diagram of correlation coefficient matrix

2.4 Similarity Analysis of Particle Size Monitoring Data

By analyzing the data feature between adjacent rows, we can gain insight into the similarity relationships between adjacent time data. To measure the similarity between adjacent data items, we calculate the Manhattan distance as a metric. Due to the large dataset, we applied this method to a dataset consisting of 2880 min-level data related to dust weather events (see Sect. 2.1). The calculated similarity relationships are presented in Fig. 5.

The lighter the color means the closer the data, and the higher the similarity, while the darker the color means the farther the data and the lower the similarity. It shows that the color between the data of adjacent time is very light, while the color turns very dark, especially April 10th 21: 00 to 11th 2: 00, means it enters the dusty time.

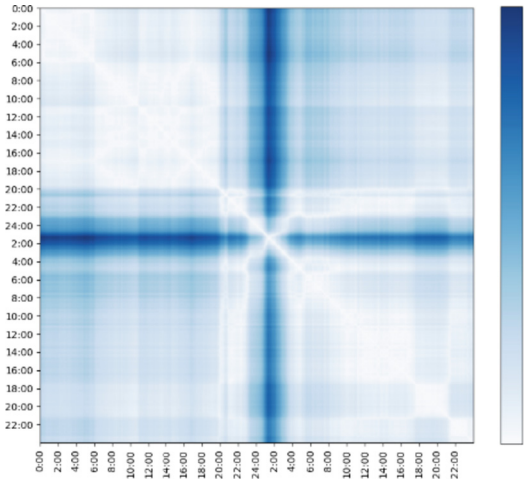

Fig. 5. Similarity analysis graph of dust event (from April 10^{th} to 11^{th})

3 Research on Traceability Model for Particulate Pollutants

3.1 Overall Design

Factor analysis is one of the commonly used methods in pollutant traceability, it can be used to statistically identify underlying structures or "factors" that explain the correlation between variables. It can be reduced a large number of influences to a few key factors, which helps to improve the efficiency and interpretability of subsequent analyses. Machine Learning (ML) is an algorithmic approach that allows computers to learn laws and patterns from data and can be applied to a variety of prediction and classification tasks. For example, ML model (Random Forest) is used to classify the particle size spectrum data into different pollution sources.

Therefore, the purpose of the traceability of atmospheric particulate pollutants will be performed by two technical methods, factor analysis and machine learning, as shown in Fig. 6, and validated by comparative experiment, case study and expert's experience.

3.2 Classification Model Based on Random Forest

Random Forest (RF) [15–17] is an extension of Bagging [18] (a parallel integrated learning method), which fixes the base learner as a decision tree. In the process of constructing a random forest, the training set is first randomly sampled several times, and each time a sample set is obtained. Then, each sampling set is used to train a decision tree, and each tree randomly selects a portion of sample features at nodes and divides the left and right subtrees according to the optimal values of these features [19]. The main advantages of the random forest algorithm include high parallelization, advantageous training speed for large samples in the big data era, and insensitivity to some missing features [20]. Therefore, the random forest is used in this study for classifying 7 pollution sources by using the particle size spectrum data. Compared to other methods such as SVM, Naive Bayes, RF has the highest accuracy 95%.

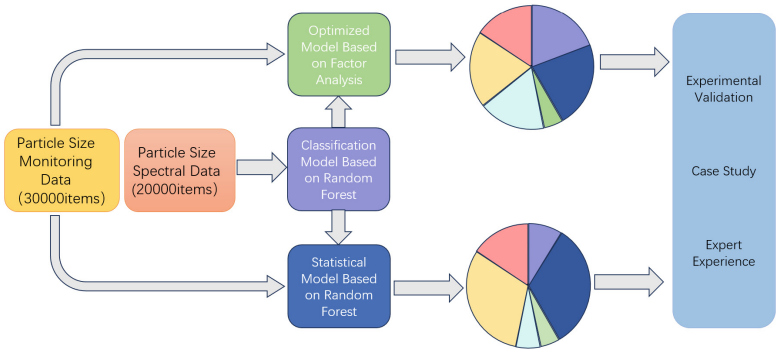

Fig. 6. Solution for the traceability of particulate pollutants

3.3 Method 1: Optimized Model Based on Factor Analysis

Factor Analysis Methods. Factor analysis is a method of multivariate statistical analysis that reduces a number of variables with intricate relationships to a few unrelated new composite factors. By grouping several variables that are more closely related in the same category, each category becomes a factor. Receptor models (in Sect. 1.2) are widely used in pollution source analysis, including positive definite matrix factorization (PMF), principal component analysis (PCA), and chemical mass balance (CMB), etc. PMF is a kind of factor analysis, which has non-negative constraints on the sources, so it can decompose the natural sources better. PMF often used to resolve pollution sources by chemical composition of particulate matter (water-soluble ions, organic fractions, carbon fractions, and metal elements). In the study method 1 is inspired by the PMF.

The PMF model first calculates the errors of different factors in the input data, and then finds the optimal source and contribution matrices by the least squares method. Input data matrix X can be defined as:

$$X = S \times F + E \tag{1}$$

The input data matrix is divided into S factor contribution matrix and F factor spectrum matrix. The Factor Contribution Matrix shows the average contribution of different factors to each component. The Factor Spectrum Matrix shows the percentage and concentration of different components in each factor.

Contribution Analysis of Pollution Sources Based on Random Forest. Firstly, the input data are processed in one hour time, by using factor analysis each hour will return the corresponding two matrices S and F, and the obtained S matrix size is 3600*7. Then, it will be averaged to get the percentage of each pollution source. However, since each pollution source does not correspond to the specific class of pollution source it belongs to, so it is necessary to judge its specific pollution source category. Obtained data on the percentage of pollution sources is input into the random forest model (in Sect. 3.2) to judge which specific category belongs to which pollution source. Finally, the obtained F matrix can be used to get the percentage of contribution of the pollution source.

3.4 Method 2: Statistical Model Based on Random Forest

Random Forest Based Pollution Source Contribution Analysis. By using the existing pollution source spectrum data random forest model is well trained (in Sect. 3.2). Then using the trained model to discriminate the categories of the monitoring data by using statistical method, that is counting the number of each type of pollution source in a certain time interval, and then calculate the percentage of each type of source as the contribution of that type of source to the pollution at that moment.

Calculation Method of Pollution Source Contribution Based on Probability Statistics. Due to the proportion value of categories with smaller contributions is very small, this study conducted a square root operation on the preliminary results and normalized them to optimize the results. To further optimize the fitting of the model to the real situation, weight is assigned to the results of each data, which is the ratio of the sum of the mass concentrations of the data to the sum of all data mass concentrations over the entire time interval.

4 Experiments and Discussions

4.1 Experimental Setup and Evaluation Indicators

Computer hardware uses CPU of i5-8300 and memory of 16GB, software is based on Anaconda and Visual Studio Code, a Python environment configuration software that makes it easy to manage different Python development environments without conflicts.

Because there are no evaluation indicators for traceability models, this study considers that the changes in the contribution of particulate air pollution in adjacent time periods are correlated, and therefore introduces the concept of similarity to analyze the experimental results. The Manhattan distance between the experimental results of two neighboring data is calculated as a measure of their similarity. The Manhattan distance (Eq. 2) is a geometric term used in geometric metric space to label the sum of the absolute axial distances of two points on a standard coordinate system. The shorter the Manhattan distance, the better we consider the algorithm fitting.

$$d(i,j) = |x_i - x_j| + |y_i - y_j| \tag{2}$$

4.2 Comparative Experiment of the Two Methods

Comparative experiment between method 1 (in Sect. 3.3) and method 2 (in Sect. 3.4) are conducted. In Fig. 7, we can see that the results of the two pollutant traceability methods by randomly selecting a day are roughly the same. Fixed combustion sources and fixed combustion sources 2 account for a large proportion, while the rest of the pollution sources of biomass combustion, construction dust, road dust, motor vehicles, and catering fumes do not contribute significantly. The two methods are consistent in the overall results.

By calculating the Manhattan distance of neighboring samples of the results of the two methods, it can be found that the average distance of method 2 is 0.0012 and method

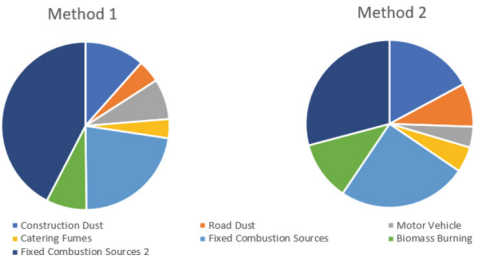

Fig. 7. Method1 (Optimized Model Based on Factor Analysis) vs. method2 (Statistical Model Based on Random Forest) (a day randomly selected)

1 is 0.0030, so the method 2 is a better fit for pollution over time in real data. Based on the expert experience for this period of data, it can be concluded that particulate pollutants were significantly affected by fixed combustion sources at that time. Therefore, we can believe that both methods have a certain degree of accuracy.

4.3 Case Study

In the particle size monitoring data during April 10th to 11th are typical dust weather. By using method 1, the experimental results are shown in Fig. 8. The contribution of road dust (green part) started to rise significantly from 20:00 on April 10th, which is in line with the weather observation results (the dust weather started at 21:00 on April 10th, with PM10 concentration of 459 $\mu g/m^3$, peaked at 2084 $\mu g/m^3$ at 2:00 on April 11th, and ended at 19:00 on April 11th with PM10 concentration dropping to 166 $\mu g/m^3$, and lasted for a total of 22h; then by the influence of sand and dust reflux, PM10 concentration exists to rise again.) The modeling of the PM10 concentration, which was carried out in the past, proved that the model has some feasibility.

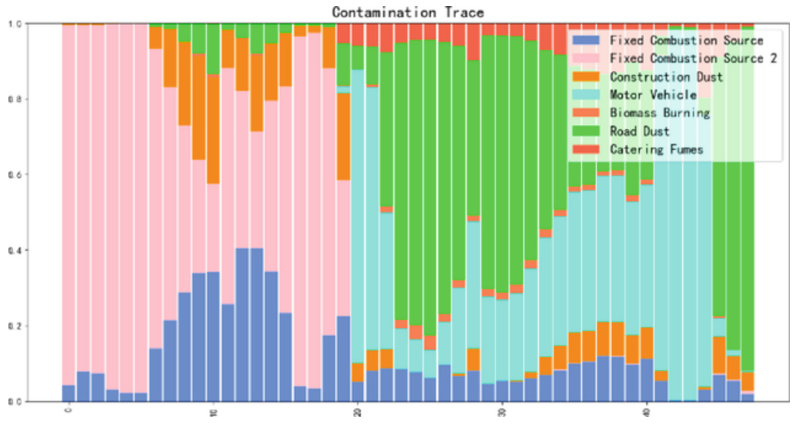

Fig. 8. Pile-up histogram of 7 pollution sources (from April 10th to 11th)

Figure 9 shows the contribution of various pollution sources in all April, the horizontal axis corresponds from April 1st to 30th, and the vertical axis is the contribution of pollution sources. The black box in the figure is the period when the dust weather is more serious, which is consistent with meteorological observations. The figure shows that the contribution of various pollution sources fluctuates greatly, and the model needs to be further optimized.

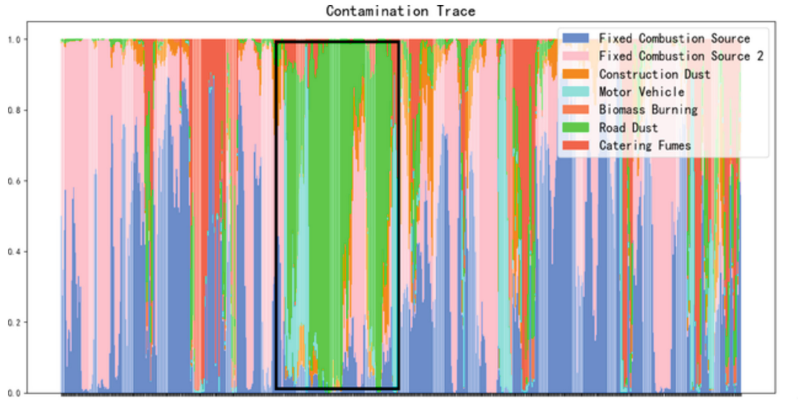

Fig. 9. Pile-up histogram of 7 pollution sources (from April 1st to 30th)

Analysis of the similarity: the results of calculating the Manhattan distance for the contribution results of the pollution sources are shown in Fig. 10, from which we can see that the overall similarity is consistent with the particle size spectral data, which indicates that the Random Forest-based particle pollutant traceability algorithm proposed in this study can fit the real dataset very well.

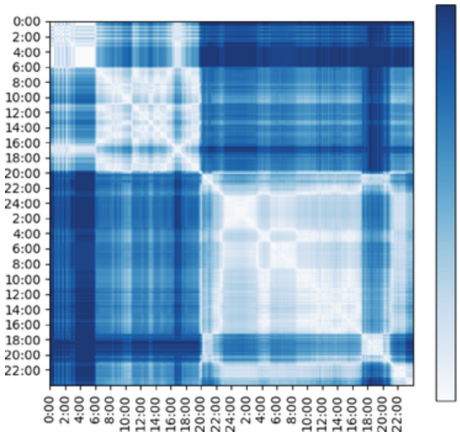

Fig. 10. Similarity analysis of the results of pollution source contribution

4.4 Expert Experience

Since the raw data could not provide accurate reference values for the contribution of pollution sources, this study also got confirmation and advices from experts in the field of environmental protection, who considered that the relevant experimental results were basically in line with the actual situation.

Through the case study on April 10, the experts believe that the conclusions of the model presented in this paper are consistent with the actual environment, the contribution ratio of different pollution sources is reasonably distributed, and the main pollution sources obtained are also consistent with the expert experience. The results of the proposed method are basically consistent with the traditional physicochemical methods, and our method has higher resolution and higher cost performance. At present, there is no particularly accurate numerical prediction, and it is hoped that more reasonable evaluation indicators can be explored through big data and machine learning.

5 Conclusion

By using of particle size data, this study combines machine learning and factor analysis methods to conduct in-depth research on traceability of atmospheric particulate pollutants, and the main contributions of this study are as follows: Firstly, this study analyzed particle size data and determined the distribution characteristics of particle size in different pollution sources; Secondly, a traceability model based on random forest and factor analysis was constructed to achieve the traceability of particulate pollutants using particle size data; Finally, through experimental comparison, case study and expert experience, the model was verified its effectiveness in the actual situation. This study is the first to use pollutant particle size data, which achieves high-resolution pollutant traceability compared to traditional chemical composition methods. This study introduces machine learning methods into traditional factor analysis method to improve processing efficiency and accuracy, providing a reference basis for air pollution control.

Due to the limitation of data acquisition, this study cannot obtain accurate pollution source contribution proportion data, so only pollution source spectrum data can be used for analysis, if many pollution source contribution proportion data can be obtained, the accuracy of the model can be further improved and the dependence on expert experience can be reduced.

Acknowledgements. Funding by Science and Technology Plan Project in Xinji City, Hebei Province: Development and Application of Particle Size Analysis Traceability and Decision Support System.

References

1. Yang, G.: Establishment of traceability system for aerosol particle size spectrometer. China University of Petroleum (Beijing), Master (2020)
2. Li, Y., et al.: Real-time chemical characterization of atmospheric particulate matter in China: a review. Atmos. Environ. **158**, 270–304 (2017)

3. Lv, M., Li, Y., Chen, L., Chen, T.: Air quality estimation by exploiting terrain features and multi-view transfer semi-supervised regression. Inf. Sci. **483**, 82–95 (2019)
4. Miao, Q., Jiang, N., Zhang, R., Zhao, X., Qi, J.: Characteristics and sources of atmospheric PM2.5 pollution in typical cities of the Central Plains Urban Agglomeration in fall and winter. Environ. Sci. **42**(01), 19–29 (2021)
5. Cao, J., Zhao, H.: Research on accurate enforcement of PM2.5 traceability in Beijing-Tianjin-Hebei region. Environ. Sustainable Dev. **44**(02), 57–61 (2019)
6. Zhu, S., Dong, W, Xu, J.: Characterization of PM2.5 pollution and its traceability and tracking in Urumqi. Environ. Protection Xinjiang **34**(03), 6–11 (2012)
7. Huang, S., Liu, F., Sheng, L., Cheng, L., Wulin, L.J.: Traceability of air pollution based on concomitant methods. Chin. Sci. Bull. **63**(16), 1594–1605 (2018)
8. Zheng, M., Zhang, Y., Yan, C., Zhu, G., James, J.S., Zhang, Y.: A review of source analysis methods for PM2.5 in China. Acta Scicentiarum Naturalum Universitis Pekinesis **50**(06), 1141–1154 (2014)
9. Wang, Q., et al.: Contribution of atmospheric VOCs to the generation of secondary organic aerosols and their sources in autumn in Shanghai. Environ. Sci. **34**(02), 424–433 (2013)
10. Zhang, G., Yin, B., Bai, W.: Particle size distribution and source analysis of roadway particles in Tianjin in winter. Environ. Sci. **43**(09), 4467–4474 (2022)
11. Huang, Y.: Research on the source analysis of heavy metal pollution in farmland soil at different scales. Zhejiang University, Master (2018)
12. Chen, Y.: Research on air pollution source analysis in Shanghai based on machine learning. East China Normal University, Master (2018)
13. Chen, J., Mou, F., Zhang, Y., Tian, T., Wang, J.: Comparison of hour-by-hour PM2.5 concentration prediction based on multiple machine learning models. J. Nanjing Forestry Univ. (Natural Science Edition) **46**(05), 152–160 (2022)
14. Wang, X., Huang, R., Zhang, W.: Ozone and PM2.5 pollution potential forecasting model based on machine learning method - a case study of Chengdu City. Journal of Acta Scicentiarum Naturalum Universitis Pekinesis **57**(05), 938–950 (2021)
15. Breiman: Random forests. Machine Learning **45**(1), 5–32 (2001)
16. JL Speiser, Michael, E., Miller, J.T., Edward Ip: A comparison of random forest variable selection methods for classification prediction modeling. Expert Systems with Applications 134, 93–101 (2019)
17. Wang, Y., Xia, S.: A review of random forest algorithm for integrated learning. Inf. Commun. Technol. **12**(01), 49–55 (2018)
18. Sagi, O., Rokach, L.: Ensemble learning: a survey. Wiley Interdisciplinary Rev. Data Mining Knowl. Discovery **8**(5), e1249 (2018)
19. Cao, Z.: Optimization Research on Random Forest Algorithm. Capital University of Economics and Business, Master (2014)
20. Li, G., Li, J., Zhang, L.: A feature selection method fusing ant colony algorithm and random forest. Comput. Sci. **46**(S2), 212–215 (2019)
21. Chen, G., Li, S., Knibbs, L.D.: A machine learning method to estimate PM2.5 concentrations across China with remote sensing, meteorological and land use information. Sci. Total Environ. **636**, 52–60 (2018)
22. van Aaron, D., Martin Randall, V., Michael, B., Winker David, M.: global estimates of fine particulate matter using a combined geophysical-statistical method with information from satellites, models, and monitors. Environ. Sci. Technol. **50**(7), 3762–3772 (2016)