Research on reverse identification algorithm of indoor pollution source based on machine learning

Xin Su, *Zhengwei* Long*, and *Yinyue* Xu

Tianjin Key Laboratory of Indoor Air Environmental Quality Control, School of Environmental Science and Engineering, Tianjin University, Tianjin, China

Abstract. Indoor sudden pollutant leakage brings environmental pollution and occupational exposure, so it is more and more important to obtain the location and identification of leakage sources. Through the forward method based on machine learning, this paper establishes a reverse traceability model for indoor multiple pollution sources. The POD method is used to obtain a large number of intermediate working condition data. The data pre-processing strategy of first normalization and then random forest feature screening can effectively improve the accuracy and generalization ability of the model. Based on a real environmental room case, model verification and sensor deployment optimization are carried out. The results show that the four sensors deployed in a specific location can achieve more than 95% positioning accuracy. In addition, the leakage possibility ranking component embedded in the model can effectively guide the staff to check the leakage points in turn, and the efficiency of three checks is as high as 99.91%.

1 background

In the real industrial production process, the aging and failure of the original sealing mechanism and the change of the medium condition lead to the abnormal leakage of toxic and harmful substances. It is necessary to control and eliminate pollution sources. The process of quickly obtaining the location information of the pollution source according to the sensor on-site monitoring data belongs to the inverse problem solving [1].

Existing backward traceability methods mainly include backward method and forward method [3]. The backward method has strong theoretical basis and physical significance, but its rapid application in practical pollution source identification is limited by the long reverse solution time, poor stability and fixed prior knowledge of the transmission model [2]. The forward method can respond in real time in practical application and is closely combined with the field monitoring data of sensors, but it often needs a lot of simulation as the data support of model training in the early stage. It can be seen that there is a lack of online source identification method that can simultaneously meet the requirements of no prior knowledge, low computational cost, fast solution speed and high traceability accuracy in the industrial field of reverse traceability tasks.

Vukovic et al. [5] first applied the forward method to the identification of multi-area building pollution sources. Bastani et al. [6] used a limited number of sensors to monitor the instantaneous concentration of different areas to achieve real-time rapid traceability. The above research inspired us to acquire and rapidly expand the data set through forward simulation, and further use the forward method to carry out reverse traceability [4]. In this paper, CFD technology is used as a tool for forward simulation data set preparation, and on this basis, pollution source location identification is carried out.

2 Machine learning traceability model framework

2.1 An overview of the framework

Figure 1 shows an overview of the reverse traceability model framework proposed in this paper.

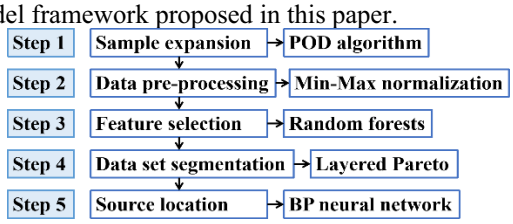


Fig. 1. An overview of the backward traceability model framework.

2.2 Algorithm introduction

Step 1 Numerical simulation and sample expansion

Star-CCM software with large scale parallel pre and post processing capability and computing capability is used to carry out numerical simulation, and the following assumptions are made: (1) The single pollution source problem. (2) Contaminants are released instantaneously. (3) The environment is constant.

* Corresponding author: longzw@tju.edu.cn

Traditional numerical simulation methods have low efficiency in acquiring datasets [4]. In order to make up for the shortcoming of the forward method, this paper creatively applies POD to the data set expansion of artificial neural network model. It can achieve the reduction and rapid acquisition of physical field [7].

Step 2 Data pre-processing strategy

The results show that the accuracy of the test set constructed with Norm-RF pre-processing strategy is higher than that constructed with RF-Norm pre-processing strategy. After the implementation of data pre-processing strategy, the data set is significantly reduced and the training time of the model is greatly shortened, which greatly improves the training efficiency of the model and reduces the calculation cost. The normalization method used linear function method (Min-Max Scaling) and the formula was as follows:

$$C_{norm} = \frac{C - C_{min}}{C_{max} - C_{min}} \quad (1)$$

Step 3 Random forest feature selection

Random forest algorithm can evaluate the contribution of each feature in classification and realize parallel training. The process is summarized as follows:(1) N_estimators time sampling is randomly put back from the original extended training set to generate 100 training sets with intersection. (2) Generate an independent decision tree model for each training set. (3) Each split of a single decision tree is based on the Gini index to select the optimal feature. (4) The importance score of features is normalized and the output is arranged in descending order. Through the above steps, the features with higher importance scores can be used as the input of the neural network classifier, which greatly improves the training efficiency of the model.

Step 4 Data segmentation and stratified sampling

Data set segmentation should follow stratified sampling. The data set can be divided into different types according to the location of pollution sources. In addition, the segmentation of training set and data set is based on Pareto criterion [8], that is, 80% samples are randomly selected as training set and 20% samples as test set. The Python data segmentation program automatically stratified sampling and segmentation of data sets, and fixed random seeds to ensure the randomness of sample extraction and reproducibility of the program.

Step 5 Neural network source location

In this paper, pollutant concentration distribution is taken as the input and source location as the output to build a reverse traceability model. BP neural network enables ANN to obtain minimum mean square error (MSE) under training samples by constantly modifying weights between connections and updating parameters during back propagation. The optimal value matching of neural network parameters can be determined by grid search.

In order to optimize the sensor deployment and guide the leak investigation plan, the incomplete enumeration method is used to carry out detailed pruning, and the number of sensors is reduced in turn on the premise of ensuring the accuracy of the source location identification test set. In addition, an intelligent sorting algorithm module is embedded in the output part of the neural network model, and the intelligent recommendation for the leak source location of each input sample is completed according to the possibility score output in the BP neural network classifier.

2.3 Traceability process

First, the POD algorithm is used to quickly obtain forward simulation data of a large number of intermediate working conditions, and the original augmented data set is further normalized. Then, the random forest method is used to carry out feature selection and overall pruning, then carry out stratified sampling and data segmentation, and finally use BP neural network to identify the possible locations of pollution sources and expand the model's leakage inspection sequence guidance function. Figure 2 shows the reverse traceability process in this document.

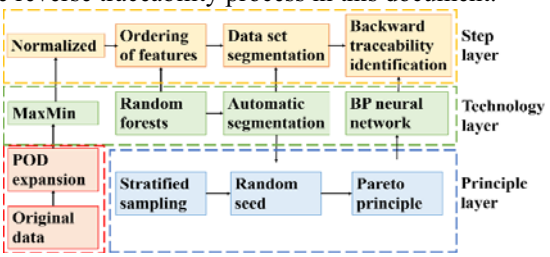


Fig. 2. Overall process of reverse traceability.

3 Case experiments and model validation

3.1 Case simulation and data preparation

In this paper, experiments and model verification are carried out based on a case study of a typical multi-source environmental laboratory [9]. See Figure 3 for indoor pollution source distribution and facility layout. Lin et al. [10] carried out a large number of ventilation experimental studies in this laboratory, which can be used to verify the accuracy of numerical simulation in this paper.

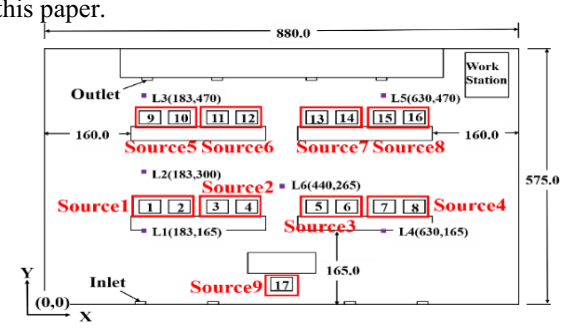


Fig. 3. Geometric model of typical environmental laboratory case.

In this paper, 20 groups of independent identically distributed sampling under air supply parameters were carried out for the above typical case [11]. It is proved that the numerical simulation results can be highly fitted with the experimental data. POD interpolation method was used to expand 180 simulated conditions from 9 source locations to 5643 conditions. The data set segmented by automatic random sampling was divided into a training set containing 4518 samples and a test set containing 1125 samples, and the 9 types of label data were evenly distributed.

The results of the forward numerical simulation show that the pollutant concentration distribution near the wall surface can be used as the analysis and training basis of multi-pollution source identification model. In this study, a total of 415 initial sampling points were set evenly on the left wall, right wall, back wall and near ceiling wall. The sensor layout will be further optimized through feature selection in the future.

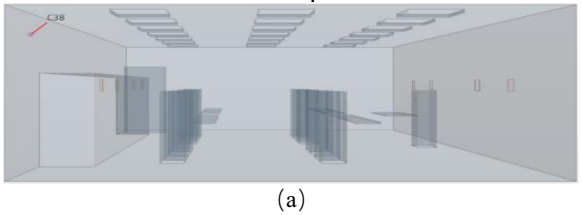
3.2 Experimental analysis and model application

The search objective of gradient anomaly grid search is to find the best parameter combination scheme to optimize model effect [12]. Based on the grid search optimization model parameters, this paper analyzed the model training time cost and source identification effectiveness before and after data pre-treatment, as shown in Table 1. The results show that the data pre-processing strategy can greatly improve the training efficiency of the model and reduce the computational cost. The redundant features of the extended data set are removed, which effectively avoids the over-fitting of the training set in the process of model training, and the generalization ability of the neural network source recognition model is strengthened.

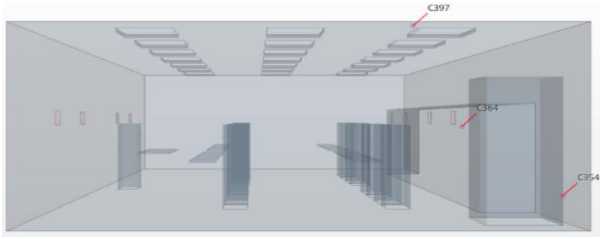
Table 1. Comparison table of cost and validity of model calculation before and after Norm-RF data pretreatment.

comparision project	Before data pre-processing	After data pre-processing
Number of features	415	10
Data set size /M	26.65	0.74
Training time /min	1020	25
Accuracy of training set /%	100.00	99.29
Accuracy of test set /%	97.20	99.38

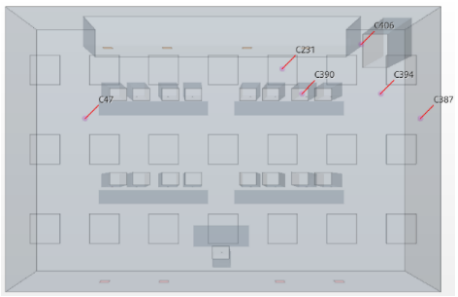
BP neural network model is used to carry out reverse tracing. After random forest feature selection, sensor deployment positions are reduced from the initial 415 sampling points to 10 important feature positions as shown in Figure 4. The accuracy of the test set is up to 99.38%, which can meet the actual requirements of rapid and accurate identification of pollution sources.



(a)



(b)



(c)

Fig. 4. Random forest feature selection location distribution map of 10 important features (a) One sensor on the left wall of the left view; (b) Three sensors on the left wall of the right view Top view; (c) 6 sensors on the ceiling.

3.3 Function expansion and effect verification

3.3.1 Sensor Deployment optimization

This paper proposes to use the incomplete enumeration method for detail pruning to gradually reduce the number of features.

In this paper, the sensor layout scheme is optimized under the premise that the accuracy threshold of source location identification is 95%. The accuracy of the source location recognition model is cross-verified for each layout scheme, and then the layout scheme with the highest accuracy is selected as the optimization strategy. It can be seen from Figure 5 that the reduction of the number of sensors is an important factor to reduce the accuracy of the source location recognition model, and grid search can improve the accuracy of the model to the maximum extent in the process of model training.

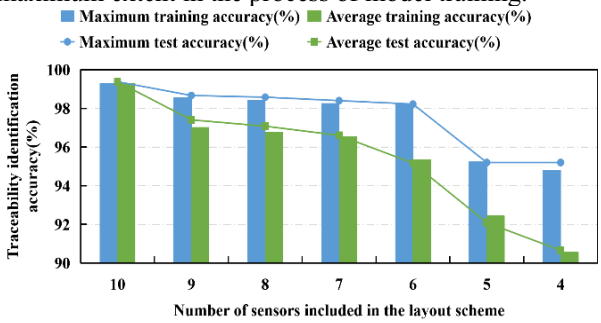


Fig. 5. Accuracy of source position recognition model with different number of sensors in detail pruning process.

When the number of sensors arranged in the room is reduced to 4, the accuracy of the source identification model test set based on layout scheme 8 can still remain above 95%. This extended function can provide effective guidance for the deployment of on-site traceability sensors. Figure 6 shows the effect difference between 21 layout schemes of four sensors.

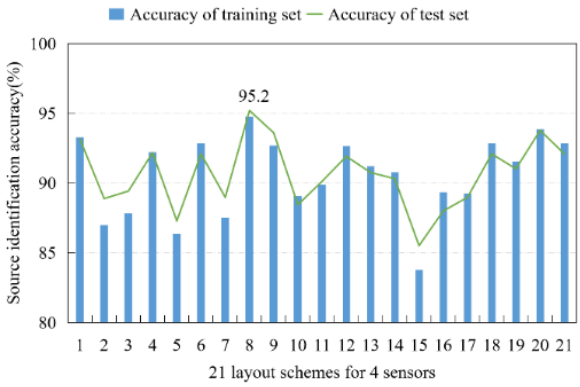


Fig. 6. Source position recognition model accuracy of 21 layout schemes with 4 sensors.

3.3.2 Sequence guidance for leak detection

This paper expects to further give the possible ranking of the location of the pollution source, so as to guide the staff to investigate.

This paper compares the source location likelihood ranking with the location label of the real leak source in turn. When the source position with the leak possibility ranked first is the real leak position, the one-time identification of the representative model is correct. We expect to obtain a source location identification model with high accuracy, and obtain the exact location of the source as much as possible under the premise of reducing the number of inspections. Table 2 shows the validation results of the intelligent recommendation algorithm.

Table 2. Summary of validity verification results of intelligent recommendation algorithm.

Number of screening	Once	Twice	Three times
Accuracy /%	95.20	99.82	99.91

The results show that the intelligent positioning algorithm proposed in this paper has a probability of more than 95% of accurately finding the location of the leak in one time, and the probability of finding the location of the pollution source within three times is close to 100%. It can meet the on-site application requirements of pollution source location identification and improve the efficiency of investigation.

4 Conclusion

This study proposes a set of pollution source identification framework, which can realize on-site rapid source traceability and leakage investigation guidance. The specific conclusions are as follows:

(1) The use of POD method for data set expansion can replace traditional numerical simulation to a certain extent, breaking the limitation of high computational cost of forward method.

(2) The random forest algorithm is suitable for the simplification and screening of the input data set of the source identification model. The data pre-processing process should follow the order of normalization first and then filtering features.

(3) The model expansion function realizes optimized sensor deployment, which can guide leak investigation and effectively improve investigation efficiency.

This study was supported by the National Key R&D Program of China through Grant 2021YFC2600500.

References

1. Zhang T, Li H, Wang S, *An inverse Lagrange inverse problem model for identification of indoor particle pollution sources, Journal of Civil* (in Chinese), Architectural & Environmental Engineering, **33**,112 ~ 116(2011)
2. Kathirgamanathan P, Mckibbin R, Mclachlan R, *Source Term Estimation of Pollution From an Instantaneous Point Source*, Research Letters in the Information and Mathematical Sciences, **3** (2002)
3. Shankar Rao K, *Source Estimation Methods for Atmospheric Dispersion*, **41**: 6964~6973(2007)
4. Liu X, Zhai Z, *Inverse Modeling Methods for Indoor Airborne Pollutant Tracking: Literature Review and Fundamentals*, **17**: 419~438(2007)
5. Vukovic V, Tabares-Velasco P C, Srebric J, *Real-Time Identification of Indoor Pollutant Source Positions Based On Neural Network Locator of Contaminant Sources and Optimized Sensor Networks*, **60**: 1034~1048(2010)
6. Bastani A, Haghighat F, Kozinski J A, *Contaminant Source Identification within a Building: Toward Design of Immune Buildings*, **51**: 320~329(2012)
7. Liu Y, Pan W, Long Z, *Optimization of Air Supply Parameters for Stratum Ventilation Based On Proper Orthogonal Decomposition*, **75** (2021)
8. Matei S A, Bruno R J, *Pareto's 80/20 Law and Social Differentiation: A Social Entropy Perspective*, **41**: 178~186(2015)
9. Yao T, Lin Z, *An Experimental and Numerical Study On the Effect of Air Terminal Layout On the Performance of Stratum Ventilation*, **82**: 75~86(2014)
10. Lin Z, *Effective Draft Temperature for Evaluating the Performance of Stratum Ventilation*, **46**: 1843~1850(2011)
11. Shao X L, Wang K K, Li X T et al. *Potential of Stratum Ventilation to Satisfy Differentiated Comfort Requirements in Multi-Occupied Zones*, **143**: 329~338(2018)
12. Yu Y, Zhu Y, Wan D et al. *Anomaly detection of hydrological time series based on sliding window prediction* (in Chinese), **34**: 2217~2220(2014)