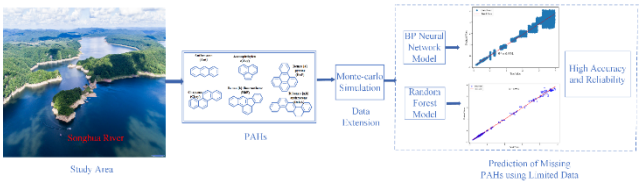
Global NEST Journal, Vol 27, No 2, 06802
Copyright© 2024 Global NEST
Printed in Greece. All rights reserved

Prediction of missing PAHs in the songhua river using monte carlo expansion, BP neural network, and random forest models

Zhangwei Cao[1], Xiurong Si[2], Feng Chen[2], Min Li[2]* and Fansheng Meng[3]
[1]Langfang City Human Resources and Social Security Bureau, Langfang 065000, China
[2]North China Institute of Aerospace Engineering, Langfang 065000, China
[3]Chinese Research Academy of Environmental Sciences, Beijing 100012, China
Received: 11/09/2024, Accepted: 18/11/2024, Available online: 25/11/2024
*to whom all correspondence should be addressed: e-mail: ht00134@nciae.edu.cn
https://doi.org/10.30955/gnj.06802

Graphical abstract



Abstract

Polycyclic aromatic hydrocarbons (PAHs) are toxic and ubiquitous in the water environment. The content of PAHs needs to be systematically and comprehensively detected to achieve the environmental quality assessment. However, due to the constraint of testing cost, detection data in the Songhua river are often missing in polluted sites. In this study, the contents of 6 PAHs in the Songhua River were predicted by utilizing limited data. Data were expanded by Monte Carlo, and BP neural network and random forest models were built to predict the contents of PAHs in missing samples. Both models showed good training effects, with high accuracy and reliability. BP neural network was slightly better than random forest in prediction of the contents of PAHs based on several performance indicators. However, when dealing with large amounts of data, random forest still performed well due to its inclusion of numerous decision trees. The two models could provide important decision support for the spatial analysis and evaluation of water pollution. This approach could not only help to understand the distribution patterns of pollutants in water bodies, but also improve the targeting and effectiveness of management and intervention measures.

Keywords: Monte Carlo, BP neural network, random forest, simulation prediction, PAHs

1. Introduction

Polycyclic aromatic hydrocarbons (PAHs) are long-term toxic substances commonly found in the environment, consisting of more than two benzene rings in curved, clustered or linear arrangement. PAHs are important causes of human cancer and enter the human body through the atmosphere, water, food and other ways (Mallah *et al.* 2022). From the perspective of ensuring the safety of human production and life, the content of PAHs in the water environment needs to be systematically and comprehensively detected to achieve the environmental quality assessment and safety management of surface water bodies. However, due to the constraints of testing cost and project cycle, detection data are often missing. Therefore, how to obtain more comprehensive information by using limited detection data has become a current research focus.

Artificial neural network is an operation model that imitates the structure and function of human brain. It has the basic functions such as associative memory, classification recognition, optimization calculation and nonlinear mapping (Wu & Feng, 2018). BP neural network is a kind of artificial neural network based on error inverse propagation, which can process data with no clear relationship between known conditions and results (Zhang *et al.* 2018). By establishing a certain mapping relationship between conditions and results, it is not necessary to determine the mathematical equation of mapping before constructing the network, and extract information features from incomplete samples for prediction and evaluation. BP neural network has achieved good results in predicting the spatial distribution of pollutants (Wang *et al.* 2024). In order to ensure the integrity of the detection data, the missing data was predicted by constructing BP neural network. The factor without missing data in the sample was taken as the input condition of the model, and the factor with missing data was taken as the output condition.

Random forest is an integrated learning method that mimics the structure and function of natural forests (Speiser *et al.* 2019) . Multiple decision trees are used to train, predict data and improve the accuracy and robustness of the model, which has strong classification, regression and feature selection capabilities. Random forest can handle data sets with complex and unclear relationships between variables (Caglar & Eker, 2021). Random forest model also shows good effect in

environmental science research, especially in predicting the concentration distribution of pollutants in complex environmental systems (Wang *et al.* 2021). The factors without missing data in the sample were taken as the input conditions of the model, and the factors with missing data were taken as the output conditions.

In this study, BP neural network and random forest models were constructed to predict contents of 6 PAHs in the Songhua River, which took the spatial parameters of water monitoring stations and the known content of PAHs as input, and predicted the content of PAHs in a specific time period, resulting in the effective filling of data gaps. Furthermore, the two models were compared in several performance indicators. The evaluation of the prediction effect of BP neural network and random forest model can provide important decision support for the spatial analysis and evaluation of water pollution. This approach not only helped to understand the distribution patterns of pollutants in water bodies, but also improved the targeting and effectiveness of management and intervention measures.

2. Methodology

2.1. Overview of the study area and layout of sampling points

The Songhua River is one of the seven major rivers in China, and its total water resources rank third in China with 556,800 km^2 drainage area (Yang *et al.* 2023). In this study, 6 priority of PAHs in the Songhua River was investigated, including Acenaphthylene (Acy), Anthracene (Ant), Chrysene (Chry), Benzo [b] fluoranthene (BbF), Benzo [a] pyrene (BaP), and Dibenzo [a,h] anthracene (DahA), and the data of PAHs in the main existing national and provincial control monitoring sections of the Songhua River were detected during the wet season (Zheng & Ni, 2024).

2.2. Data pre-processing

As an input item, the detection data had an important impact on the accuracy of the calculation result of the model, because the sampling time, sampling period, measuring instruments and analysis methods would introduce uncertainty errors (Mishra *et al.* 2020). Data pre-processing included data auditing and data standardization.

In order to eliminate the impact of uncertain values on test data, formula (1) was adopted for data review:

$$\begin{cases} D_{ij}^{k} = v_{ij}^{k}, & \sigma_{ij}^{k} = u_{ij}^{k} + d_{ij}^{k}/3 \\ D_{ij}^{k} = d_{ij}^{k}/2, & \sigma_{ij}^{k} = \bar{d}_{ij}^{k}/2 + d_{ij}^{k}/3 \\ D_{ij}^{k} = \tilde{v}_{ij}^{k}, & \sigma_{ij}^{k} = 4\tilde{v}_{ij}^{k} \end{cases} \quad (1)$$

Where: the measured concentration, analysis uncertainty, and method detection limit were represented by v_{ij}^{k}, u_{ij}^{k}, and d_{ij}^{k} respectively. The arithmetic mean of the method's detection limits was denoted as $\bar{d}_{ij}^{k}$, while the geometric mean of the measured concentration was represented by $\tilde{v}_{ij}^{k}$.

In order to solve the problem of mismatch between the concentration dimensions of different pollutants and eliminate some unreasonable effects that may be caused by the large difference between the values, it is necessary to standardize the concentration values of each sampling point (where the undetected values are replaced by the mean value). Standardized calculation of raw data was calculated according to formula (2) :

$$Z_i = \frac{X_i - \bar{X}}{S} \quad (2)$$

Where: X_i was the i-th observation of the variable x, and $\bar{X}$ was the average of the variable x, and S was the standard deviation.

The coefficient of variation (CV) can reflect the degree of variation of pollutants, and a larger value indicates that the pollutant is more affected by human activities (Karami & Mazaheri, 2021). From **Table 1**, it could be seen that the CV of PAHs in the study area was generally greater than 1, indicating strong variability. This showed that the content of PAHs was strongly affected by human pollution sources.

The prediction models of BP neural network and random forest were constructed by Python software. The model accuracy was analyzed using OriginPro 9 software.

2.3. Monte Carlo simulation

2.3.1. Artificial simulation data of recipient samples

Using a computer program to randomly generate pollution source spectrum data, the pollution source spectrum array generated by artificial simulation were represented by SP1, SP2, ... SP10…. Each type of pollution source contained 6 kinds of pollutants, and 14 sample data were generated in each group of receptor sample data, and each group of receptor samples were a 14×6 data matrix. Considering the non-collinearity requirement of the receptor model, the pollution source spectra of each group were quite different and there was no collinear feature.

2.3.2. Monte Carlo simulation

Considering the actual environment, spectrum data of pollution source and sample data of acceptor had a certain degree of uncertainty, and there were some changes in the migration process of pollutants from emission source to acceptor (Li *et al.* 2024). In order to simulate practical applications more realistically, Monte Carlo simulation method were used here to generate receptor sample data:

$$D_{ij} = A_{ij} + C_{ij}A_{ij}\sqrt{2}\,\mathrm{erf}^{-1}(2R_{ij} - 1) \quad (3)$$

Where, D_{ij} was the concentration of Class i pollutants in the JTH sample considering the error of the acceptor sample; A_{ij} was the initial acceptor sample concentration matrix, which was obtained from pollution source spectrum and set source contribution matrix according to formula (3). C_{ij} was the concentration variation coefficient of Class i pollutant in sample j. erf^{-1} was the inverse Gaussian error function; R_{ij} was a random number from 0 to 1.

Based on Monte Carlo simulation, the 1 group of receptor samples were expanded into a 14×6 data matrix of 1000 group of receptor samples. Furthermore, BP artificial neural network model and random forest model were constructed.

Table 1. Prediction of accuracy and error analysis of BP neural network

Substance	The minimum (ng/L)	The maximum (ng/L)	The average (ng/L)	Standard deviation (ng/L)	Coefficient of variation
Acy	1.0	32.5	7.5	8.7	1.2
Ant	0.9	1093.5	118.2	286.6	2.4
Chry	2.1	306.2	46.0	82.2	1.8
BbF	2.7	3240.8	287.9	829.2	2.9
BaP	1.2	2362.7	190.8	604.7	3.2
DahA	0.8	354.8	52.1	115.4	2.2

2.4. BP artificial neural network

The neural network was trained by using the known data in the sample. After the training met the requirements, the known data of the missing sample was input into the model, and the output value was the predicted value of the missing data (Wright *et al.* 2022). Fifty groups were randomly selected as validation samples, and 50, 500 and 950 groups were randomly selected as training samples, in which c (Ant), c (Chry), c (BbF), c (Acy), c (BaP) and c (DahA) were missing data. The missing data were predicted by the constructed neural network and compared with the measured data of the validation sample to evaluate the prediction accuracy of the model.

2.5. Random Forest

Random forest was trained by using the known data in the sample. After the training met the requirements, the known data of the missing sample was input into the model, and the output value was the predicted value of the missing data. 20% of the data set were randomly selected as validation samples, and the remaining 80% were selected as training samples, in which c (Ant), c (Chry), c (BbF), c (Acy), c (BaP) and c (DahA) were taken as missing data. The missing data were predicted by the constructed random forest, and the prediction accuracy of the model was evaluated by comparing with the measured data of the validation sample.

3. Results and discussion

3.1. The construction of BP neural network model and random forest model

The topology structure of BP neural network was shown in **Figure 1**. Each node of the neural network represented a specific output function, which was called the activation function, and the connection of each two nodes represented the weight. The weight value was adjusted through continuous learning. The number of double-layer hidden layer units constructed was 20 and 15, respectively. The tangent function tansig was selected as the transfer function of the hidden layer, the linear function purelin was selected as the transfer function of the output layer, and the conjugate gradient function trainscg was selected as the training function for the sample. The maximum iteration times of PAHs index were set to 5000 and 10000 times respectively, and the weight and threshold were finally determined through repeated iteration, and the prediction model was established.

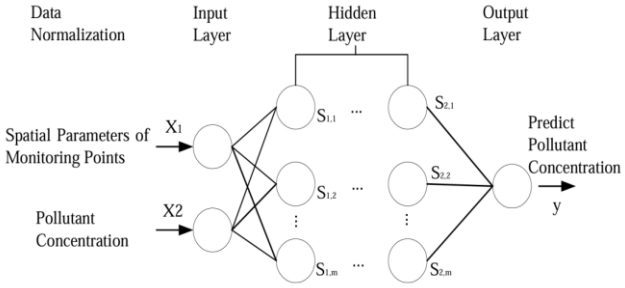


Figure 1. The topology of BP neural network model

The random forest model was constructed as shown in **Figure 2**. The random forest contained a large number of decision trees, each of which was a classification or predictor (Kovács *et al.* 2020). Every decision tree nodes represented a specific point feature selection and division, the choice of connection on behalf of the characteristics of the decision tree path. Through the bootstrap sampling method, the construction of decision tree was adjusted by continuous learning, so that the model had better generalization ability.

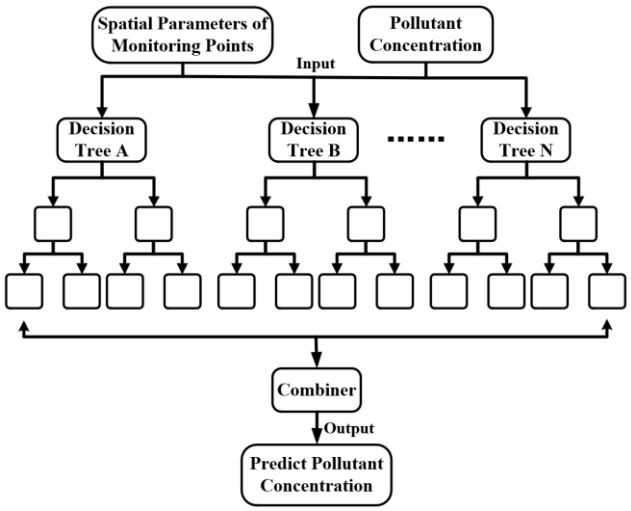


Figure 2. The topology of random forest model

In the constructed random forest, the number of decision trees was set to 200. When the decision node was binary, squared error as the splitting function of the node was selected, so that each splitting of the decision tree could reduce the sample impurity as much as possible. In terms of loss function, the mean squared error (MSE) to measure the forecasting performance of the model was

chosen. In the training process, Bagging strategy was used to train the samples. For feature selection, a random subspace approach was taken, that is, at each node, partially randomly selected features were considered. After model training, the final prediction category or prediction value was determined by voting.

3.2. Influence of numbers of training samples on the accuracy of the two models

50, 500 and 950 groups were randomly selected as training samples under 5000 and 10000 iterations to investigate the influence of different training samples on the training accuracy, as shown in **Figure 3 and Figure 4**. As can be seen, with the increase of the number of training samples, the MSE increased gradually, indicating that the training accuracy decreased gradually with the increase of the number of training samples. When the training sample data was reduced to 50, the MSE was the smallest, which may be due to the phenomenon of overfitting. As the number of training samples increased, the MSE tended to stabilize faster and faster, indicating that the more data, the easier it was to find characteristics. Therefore, appropriately increasing the number of measured samples and decreasing the number of training samples could effectively improve the prediction accuracy of the model. By comparing **Figure 3 and Figure 4**, it could be found that the MSE of neural network gradually increased with the increase of iterations, and MSE was closer to the set target error at the end of training, indicating that the training accuracy was gradually improved with the increase of iterations, so the number of iterations was gradually increased until MSE was no longer reduced. The training precision cannot be improved, and the optimal number of iterations was obtained.

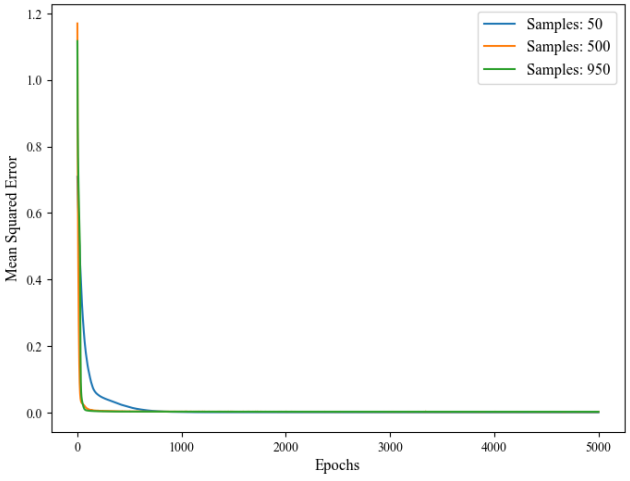


Figure 3. The influence of numbers of training samples on the accuracy of BP neural network model under 5000 iterations

The constructed random forest model as above was trained, and the influence of numbers of decision trees on the training accuracy was shown in **Figure 5**. As can be seen, the MSE gradually decreased with the increase of the number of decision trees, indicating that the training accuracy gradually increased with the increase of the number of decision trees. MSE tended to be stable until the number of decision trees reached 200. The training accuracy cannot be improved, and the number of optimal decision trees was obtained.

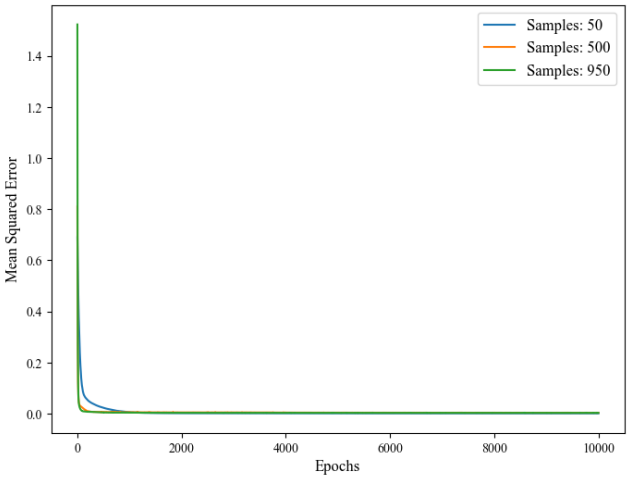


Figure 4. The influence of numbers of training samples on the accuracy of BP neural network model under 10000 iterations

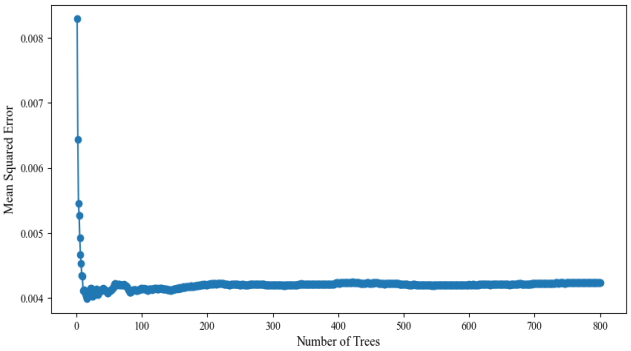


Figure 5. Effects of numbers of decision trees on the accuracy of random forest model

3.3. The precision and error analysis of the two models

All the training samples were selected to establish the BP neural network model and random forest model, and the results of training error curve and linear regression were shown in **Figure 6 and Figure 7**. The correlation coefficient (R^2) between the output value of training samples and the target value reached 0.991and 0.99716, which were close to 1, indicating that the training effect of the two models of PAHs samples was good.

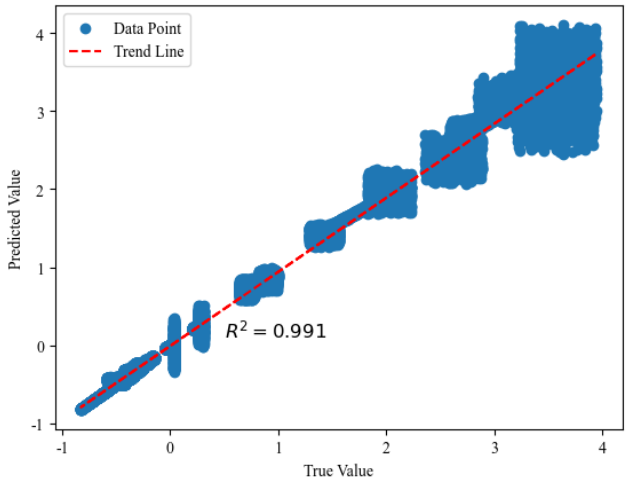


Figure 6. The training error curve and linear regression results of BP neural network model

Furthermore, the two models after training was used to predict the missing data c (Ant), c (Chry), c (BbF), c (Acy), c (BaP) and c (DahA), and the results were shown in **Figure 8 and Figure 9**. The R^2 value of the prediction result of missing data and the measured result was close to 1, indicating a good coincidence between the predicted curve and the measured curve.

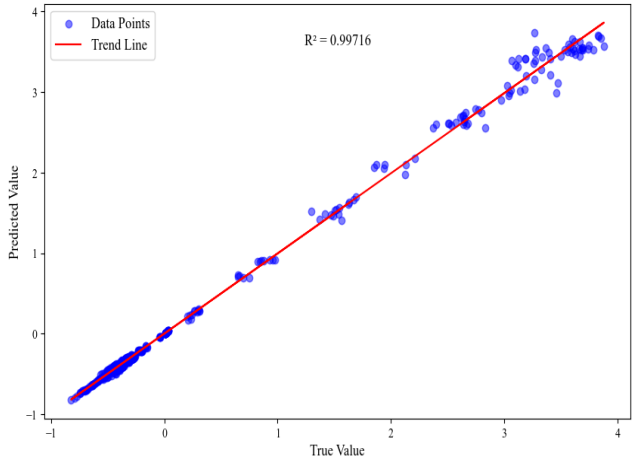


Figure 7. The training error curve and linear regression results of random forest model

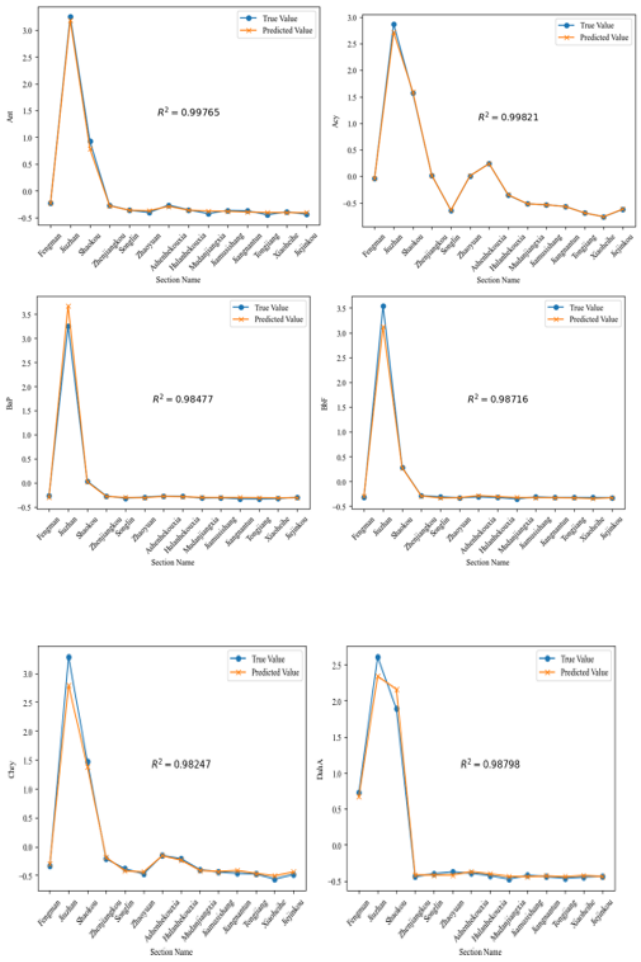


Figure 8. Scatter plot of predicted and measured values of BP neural network model

3.4. Comparison between BP neural network model and random forest model

As shown in **Figure 10**, the two models were compared from several aspects. Nash-Sutcliffe efficiency coefficient (NSE) is usually an important index to measure the reliability of simulation results, and its value ranges from (-∞, 1]. The NSE value of BP neural network was close to 1, ranging from 0.998 to 1.000. The simulation results were highly reliable and accurate, indicating that BP neural network could predict PAHs content perfectly and fit the actual observed data very well. The NSE values of random forest range from 0.9904 to 1.000, which was slightly lower than the BP neural network. Both the root mean square error (RMSE) and the mean absolute error (MAE) of BP neural network were smaller, which indicted that its prediction error was lower.

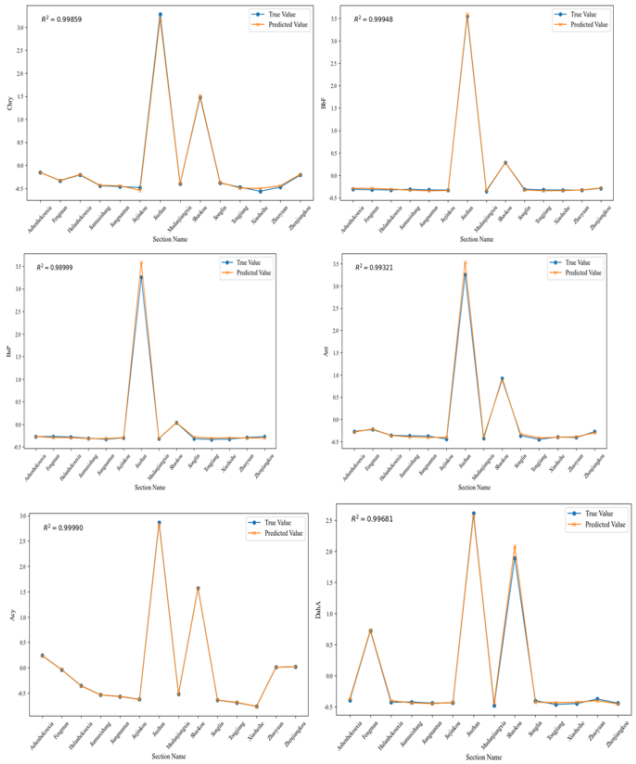


Figure 9. Scatter plot of predicted and measured values of random forest model

In summary, BP neural network was slightly better than random forest in prediction of the contents of PAHs based on several performance indicators (**Figure 10**), and the BP neural network could predict the PAHs content almost perfectly, showing an extremely high degree of fit and simulation effect. The reliability and accuracy of its simulation were close to the ideal state, which was suitable for scenarios that required high-precision prediction. Although the random forest model was slightly lower than the BP neural network in some performance metrics, when dealing with large amounts of data, it still performed well due to its inclusion of numerous decision trees. The robustness and processing speed of the random forest model was also a very important advantage for large-scale environmental data analysis.

Overall, the BP neural network model and random forest model showed their potential applications in environmental science and pollution monitoring, especially in the prediction the of content of PAHs. When choosing a suitable model, researchers should take into account the characteristics of the data, the accuracy requirements of the prediction, and the availability of

computational resources. BP neural network model may be a better choice when extremely high accuracy and complex data processing are require; random forest model may provide a very effective solution when large amounts of data need to be processed quickly.

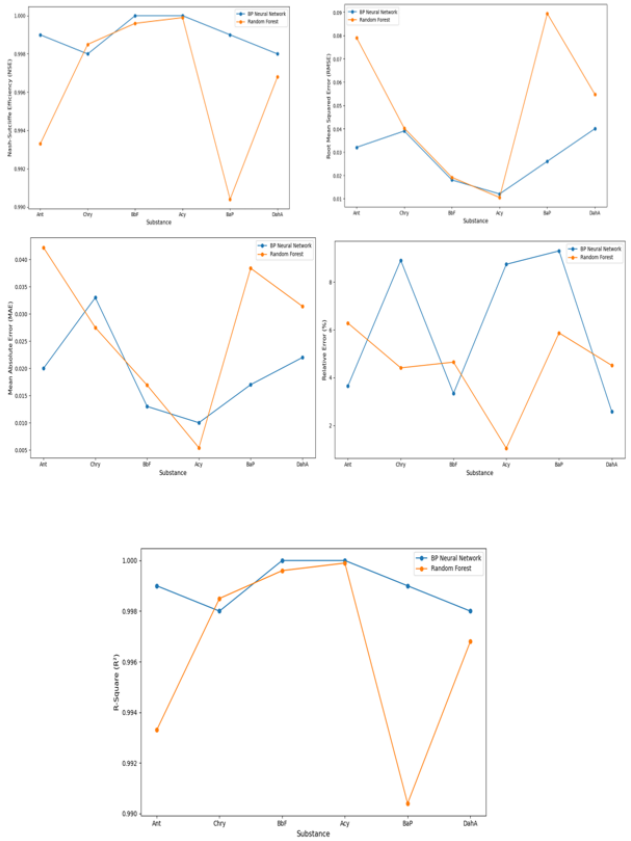


Figure 10. Comparison between BP neural network model and random forest model

4. Conclusion

BP neural network model and random forest model had good training effect. The error analysis between the predicted value and the measured value of the verification sample showed that the coefficient of determination and the simulated efficiency coefficient of each pollutant content were close to 1. The constructed models were accurate and reliable, and could predict the content of PAHs in the Songhua River well, and the input of weak correlation factors could further improve the accuracy of the prediction model. Furthermore, the two models were compared. The BP neural network was slightly better than the random forest model in several performance indicators, and could predict the PAHs content almost perfectly, showing an extremely high degree of fit and simulation effect. Although the random forest model was slightly lower than the BP neural network in some performance metrics, when dealing with large amounts of data, it still performed well due to its inclusion of numerous decision trees. When choosing an appropriate model, researchers should comprehensively consider the characteristics of the data, the accuracy requirements of the prediction, and the availability of computational resources. BP neural network model is suitable for the task that requires extremely high accuracy prediction, especially for scenarios with simple data structures and sensitive to errors. In contrast, random forest model performs better in dealing with large datasets that may contain more noise or more complex data structures, and its robustness and processing speed make it the preferred model for large-scale environmental data analysis.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (No. 51808008). National Key Research and Development Program of China (2021YFC3200103).

References

Caglar Gencosman, B., & Eker Sanli, G. (2021). Prediction of polycyclic aromatic hydrocarbons (PAHs) removal from wastewater treatment sludge using machine learning methods. Water, Air, & Soil Pollution, 232(3), 87. doi: 10.1007/s11270-021-05049-8.

Karami Cheme, E., & Mazaheri, M. (2021). The effect of neglecting spatial variations of the parameters in pollutant transport modeling in rivers. Environmental Fluid Mechanics, 21, 587-603. doi: 10.1007/s10652-021-09787-5.

Kovács, P., Zhu, X., Carrete, J., Madsen, G. K., & Wang, Z. (2020). Machine-learning prediction of infrared spectra of interstellar polycyclic aromatic hydrocarbons. The Astrophysical Journal, 902(2), 100. doi: 10.3847/1538-4357/abb5b6.

Li, Y., Tian, F., Zhong, R., & Zhao, H. (2024). Source characteristics of polycyclic aromatic hydrocarbons and polychlorinated biphenyls in surface soils of Shenyang, China: A comparison of two receptor models combined with Monte Carlo simulation. Journal of Hazardous Materials, 462, 132805. doi: 10.1016/j.jhazmat.2023.132805.

Mallah M A, Changxing L, Mallah M A *et al.* (2022). Polycyclic aromatic hydrocarbon and its effects on human health: An overeview. Chemosphere, 296, 133948. doi: 10.1016/j.chemosphere.2022.133948.

Mishra, P., Biancolillo, A., Roger, J. M., Marini, F., & Rutledge, D. N. (2020). New data preprocessing trends based on ensemble of multiple preprocessing techniques. TrAC Trends in Analytical Chemistry, 132, 116045. doi: 10.1016/j.trac.2020.116045.

Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. Expert systems with applications, 134, 93-101. doi: 10.1016/j.eswa.2019.05.028.

Wang, C., Mao, G., Liao, K., Ben, W., Qiao, M., Bai, Y., & Qu, J. (2021). Machine learning approach identifies water sample source based on microbial abundance. Water Research, 199, 117185. doi: 10.1016/j.watres.2021.117185.

Wang, W., Chen, S., Chen, L., Wang, L., Chao, Y., Shi, Z., ... & Yang, K. (2024). Drivers distinguishing of PAHs heterogeneity in surface soil of China using deep learning coupled with geo-statistical approach. Journal of Hazardous Materials, 133840. doi: 1 0.1016/j.jhazmat.2024.133840.

Wright, L. G., Onodera, T., Stein, M. M., Wang, T., Schachter, D. T., Hu, Z., & McMahon, P. L. (2022). Deep physical neural networks trained with backpropagation. Nature, 601(7894), 549-555.doi: 10.1038/s41586-021-04223-6.

Wu, Y. C., & Feng, J. W. (2018). Development and application of artificial neural network. Wireless Personal Communications, 102, 1645-1656. doi: 10.1007/s11277-017-5224-x.

Yang, Y., Zhao, Z., Chang, Y., Wang, H., Wang, H., Dong, W., & Yan, G. (2023). PAHs and PAEs in the surface sediments from Nenjiang River and the Second Songhua River, China: Distribution, composition and risk assessment. Process Safety and Environmental Protection, 178, 765-775. doi: 10.1016/j.psep.2023.08.037.

Zhang, L., Wang, F., Sun, T., & Xu, B. (2018). A constrained optimization method based on BP neural network. Neural Computing and Applications, 29, 413-421. doi: 10.1007/s00521-016-2455-9.

Zheng, Z. Y., & Ni, H. G. (2024). Predicted no-effect concentration for eight PAHs and their ecological risks in seven major river systems of China. Science of The Total Environment, 906, 167590. doi: 10.1016/j.scitotenv.2023.167590.