The gradient-boosting decision tree model can predict the concentration of PAEs in children bedroom

Chanjuan Sun [1,*], Qinghao Wang[1], Chen Huang[1], Jingguang Li[2], Jialing Zhang[1], Zhijun Zou[1], Lang Tian[1]

[1] University of Shanghai For Science And Technology, China.

[2] Shanghai Research Institute of Building Science Group Co., Ltd.

Abstract. Exposure of phthalate has adverse effects on child health. Currently, the field measurement on PAEs concentration in children's bedrooms were limited, and the test of PAEs is laborious. Based on the data of home detection in 454 residences from March 2013 to December 2014 in Shanghai, the association of PAEs in children's bedroom and building characteristics, residents' lifestyle and indoor environment characterization were built by Spearman correlation. According to the Spearman correlation coefficient method, the concentration of PAEs, such as residential area was significantly correlated with DMP, BBP and DiBP in children's bedroom (sig <0.05, sig <0.01, sig <0.01; r> 0), and the use of chemicals was significantly associated with DEP and DiBP in children's bedroom (sig <0.05, sig <0.05; r> 0). Then a gradient-boosting decision tree model with higher prediction accuracy is established. The influencing factors of the studied PAEs were determined by comprehensive consideration of the current study and literature review. 11 influencing factors of PAEs concentrations from three aspects were finally established in this study. The training model of GBDT has a reasonable accuracy(R^2 >0.9). This paper provides a reference for the prediction of PAEs concentration in the residential bedroom and the influence degree of influencing factors.

Introduction

Phthalate compounds (PAEs) are a class of widely used compounds, mainly used as plasticizer, softener, emollient, moisturizer, defoagent, etc., common in toys, personal care products, home decoration materials, etc. Studies have shown that PAEs in indoor dust fall are the major source of contact in humans. Some epidemiological evidence suggests that the emergence of indoor plasticizing products is associated with allergic symptoms in the respiratory tract (e. g., asthma), nose, and skin[1-2]. At present, the field measurement range of PAEs concentration in children's bedrooms is limited, and the test of PAEs is time-consuming and laborious[3]. By building predictive models, we can understand the concentration characteristics of PAEs in child bedrooms in different regions of the country, thus to better prevent and control contamination of PAEs.

The study found that mechanical ventilation in children's bedrooms, the use of chemicals (such as mosquito coil incense, incense, skin cream, etc.), indoor temperature and plastic toys are important factors affecting PAEs[4-6]. After determining the influencing factors, the trained prediction model was used to predict the pollutant concentration of six PAEs in DMP, BBP, DEP, DEHP, DiBP and DnBP. At present, in the prediction model of the concentration of major pollutants in indoor air, the existing multiple linear regression air pollutant concentration prediction model has weak applicability, and the BP neural network algorithm has the defect of easy local optimality[7-8]. While the gradient improvement decision tree (GBDT) can flexibly process various types of data because it does not need to scale the data, and the advantages of high prediction accuracy is greatly favored[9]. By constructing a GBDT model to predict the mass concentration of PAEs in children's bedroom dust fall, it provides a reference for predicting the concentration of PAEs and the degree of influencing factors in residential bedrooms.

1 Features selection

1.1 Date source

The CCHH research group (China, Children, Homes, Health research) began organizing research in 2010, targeting the relationship between the indoor environment and child health. From March 2013 to December 2014, Shanghai CCHH research group carried out the second stage of the study, visited 454 qualified families and conducted questionnaire survey and home detection. Using vacuum cleaner to collect dust on the door frame, window frame and floor surface of children's bedroom, and the concentration of PAEs was obtained through experimental steps such as miscellaneous removal, centrifugation and extraction. In this study, complete and effective field test samples and questionnaire results from 268 households were

*Corresponding author: Chanjuan Sun, sunchanjuan@usst.edu.cn

extracted from data from 454 households based on the detection of PAEs[10].

The influence of PAEs concentration is divided into three categories: building characteristics (residence area, the age of architecture, metope decoration materials), lifestyle (the proportion of children's plastic toys in the bedroom, smoking, the use of chemicals), indoor environment characterization (residential damp characterization, children bedroom mechanical ventilation, 24 hours average temperature)[11-13]. The 24-hour average air temperature was obtained by measuring the 24-hour average air temperature twice per hour, and the remaining influencing factors were obtained by questionnaire.

Table1.Questionnaire survey of related influencing factors

Influencing factor	Question	Case selection
Address area	What is the area of the current residence?	1="≤40m²" 2="41-60m²" 3="61-75m²" 4="76-100m²" 5="101-150m²" 6="≥150m²"
The use of chemicals	Frequency of using mosquito coil incense	0="no" 1="yes，regularly" 2="yes，sometimes"
	Frequency of the incense being used	0="no" 1="yes，regularly" 2="yes，sometimes"
	Use a skin cream?	0="no" 1="yes"
Wet representation of the current residence	Is there a wet representation (such as window condensation, wet clothes, mildew, wet spots, water damage)?	0="否" 1="是"

1.2 correlation analysis

In this paper, IBM SPSS Statistics 25 was used to test the normal distribution of PAEs and influencing factors, and the data types were non-normal distributed. Therefore, Spearman's correlation coefficient method was used for the correlation analysis, and the influencing factors of PAEs were determined by considering this study and literature review[14]. The results are shown in Table 2.

Table2.Correlation analysis of the concentration of PAEs and the influencing factors

Influencing factor	correlation coefficient（r）		
	DMP	DEP	DiBP
	DBP	BBP	DEHP
Address area	.125*	0.010	.167**
	.159**	0.103	0.097
The age of architecture	0.090	-0.056	.199**
	0.041	0.073	.134*
The proportion of the plastic toys	-0.043	0.087	-0.033
	-0.023	.169*	0.002
Smoking	-.151*	-.149*	-0.015
	-0.022	-.121*	0.022
Mechanical ventilation in the children's bedroom	0.075	0.101	0.093
	.158*	-0.081	0.038
Skin cream	-0.048	-0.039	-0.049
	-0.052	-0.067	.137*
Humidity representation in the current residence	-0.032	0.008	-0.058
	.129*	0.049	0.065
Wallpaper	.151*	0.082	0.006
	0.039	0.098	-0.015
Incense	0.030	.135*	0.017
	-0.001	-0.011	-0.043
Mosquito-repellent incense	0.066	.131*	0.073
	0.100	0.027	0.097
The 24-hour average temperature was obtained	-0.01	-.162**	0.023
	-0.065	-0.05	0.079

Note: **Level 0.01 (double tail) with significant correlation.
* Level 0.05 (double tail) with significant correlation.

Since the concentration value of PAEs in the measured value varies greatly, such as DEHP, the minimum concentration is 8.43mg/m³, and the maximum value reaches 5306.99mg/m³, the measured concentration value of PAEs is logarithmically and will not change the nature and the correlation, but the scale of the variable is compressed, such as $\lg^{8.43} = 0.9258$, $\lg^{5306.99} = 3.7248$. So in the prediction model, there were negative values for the concentration value. After taking logarithmically, the data stabilizes and weakens the collinearity of the model.

2 Gradient-boosting decision tree model

2.1 Construction of the gradient improvement decision tree model

The essence of the improvement algorithm (Boosting) is to train the sample data, iteratively generate multiple weak learners, constantly learn from errors, generate a better performance learner, and solve the state that a single weak learner is in an underfit to the data set. GBDT is a kind of lifting algorithm, using the gradient lifting algorithm to train the decision tree model. The model consists of multiple classification regression trees, which forms a high-performance learning method by training the weak learners decision tree, and finding the optimal division of the decision tree, so as to optimize the model prediction accuracy[15-17].

Select the above factors and the concentration of PAEs as sample data （x_i, y_i） , where x_i is the input variables (residence area, The age of architecture, wall decoration materials, the proportion of plastic toys in children's bedroom, smoking, chemicals, wet characterization, mechanical ventilation, 24 hours average temperature), y_i is the corresponding concentration of PAEs. In order to achieve the best fitting results of the prediction model, the model parameters should be adjusted before the prediction model is established. The fitting results can be effectively improved by adjusting the hyperparameters, which are shown in Table 3.

Table3. Hyperparameters of GBDT

hyperparameter	Hyperparameter name	note
n-estimators	Maximum number of iterations of the weak learner	Too small n-estimators easily leads to underfitting, and too large n-estimators easily leads to overfitting.
Learning-rate	Step length (learning rate)	The combined results of learning-rate and n-estimators debugging determine the fitting effect of the algorithm
subsample	sub sampling	The value range is (0,1], in order to prevent overfitting and underfitting, generally take [0.5,0.8]
loss	loss function	With little noise, the default mean variance (ls) can be used.
max-depth	Decision tree of maximum depth	Depending on the data feature complexity, generally take the default value of 3

After adjusting the number of basic models (n_estimators),the predictable coefficient (R^2)is increased from 0.6 to 0.9.

2.2 Prediction results of the gradient improvement decision tree model

The predicted concentrations of PAEs were compared to the measured data and plotted as figures. The predictive performance of the model is often used by the predictable coefficient (R^2), Mean Squared Error(MSE) and Mean Absolute Error(MAE) to measure the deviation between the predicted value and the measured value[18-20]. To test the learning effect of the model, the true value is fitted to the predicted value, and the fitting results are shown in the following figures.

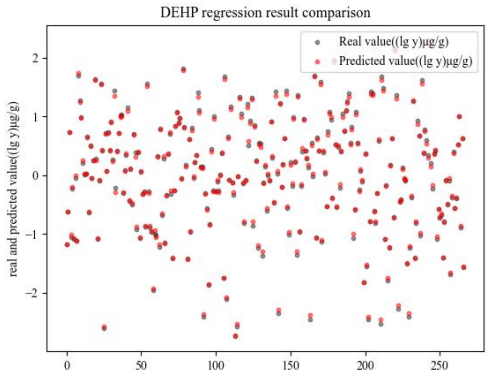

Fig. 1. Comparison of DEHP concentrations

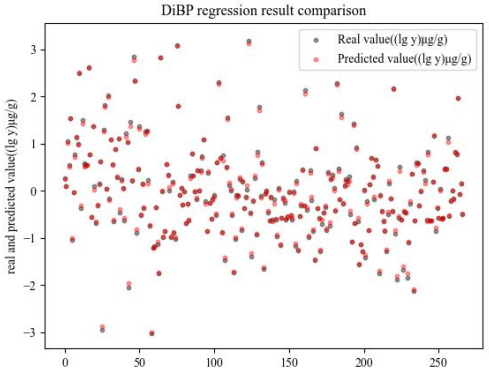

Fig. 2. Comparison of DiBP concentrations

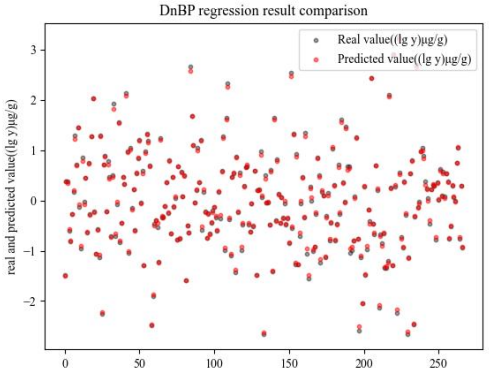

Fig. 3. Comparison of DnBP concentrations

As can be seen from the figure, the test dataset predicts the concentration values to the real values. The evaluation indicators of the regression are shown in Table 4.

Table4.Evaluation indicators of the regression

Evaluation indicators of the regression	DEP	DMP	BBP
	DiBP	DEHP	DnBP
MAE	0.0009	0.0015	0.0011
	0.0015	0.0011	0.0012
MSE	0.0238	0.0298	0.0259
	0.0319	0.0278	0.0272
R^2	0.9990	0.9985	0.9989
	0.9984	0.9988	0.9987

The closer R^2 is to 1, the closer the mean and standard deviation of the predicted and measured values are, and the smaller the MAE and MSE are, and the more accurate the prediction results are. From Table 3, the predicted R 2 of the six PAEs is greater than 0.99, and the MAE and MSE are close to 0, the predicted concentration values are clearly of high accuracy.

3 Conclusion

A gradient ascending decision tree was used to establish the prediction model of PAEs concentration in the child bedroom, which improved the accuracy of the model by adjusting the parameters. The results show that the prediction results are ideal and have high accuracy, which can provide a reference for predicting the influence of PAEs concentration and the degree of influencing factors in residential bedrooms, and provide data support for residential health risk analysis and ventilation design.

However, there are still shortcomings. This paper only adopts some factors that are significantly different in the correlation analysis, but some factors confirmed by a large number of studies, such as children's bedroom ash cleaning frequency and the use of electronic products, are not included in the features, and the prediction model needs to be further extended and improved.

References

1. Chen Huang, Zhiyuan Qing, Wei Liu , Xueying Wang, Jiao CAI, Zhijun Zou, Chanjuan Sun, Energy Research and Information, 34, 35 (2018).
2. Kexiu Li, Chanjuan Sun, Jialing Zhang, Zhijun Zou, Chen Huang , Management and Technology, 34, 56 (2022).
3. Xiaoni Yang, Xuguang Shang, Yangfan Xu , Dan Wang. Environmental Monitoring management and Technology, 32, 17 (2020).
4. Fumei Wang , Li Chen, Jiao Jiao, Leibo Zhang, Yaqin Ji, Zhipeng Bai, Liwen Zhang, Zengrong Sun, Chinese Environmental Science, 32, 780 (2012).
5. Jingjing Pei,Yahong Sun,Yihui Yin, Science of the Total Environment, 639,760 (2018).
6. Kang Hu,Qun Chen, Building and Environment, 94, 676 (2015)
7. Baolei Sun, Hao Sun, Jianwu Shi, Yaoqian Zhong, Journal of Environmental Science, 37, 1864 (2017).
8. Song Li, Ji Wang, Danchuang Zhang, Wei Xia, Computer Simulation, 32, 404 (2015).
9. Jerome H. Friedman, The Annals of Statistics, 29, 5 (2001).
10. Chenxi Liao , Shanghai: School of Environment and Architecture, University of Shanghai for Science and Technology, (2018).
11. Yirui Liang,Ying Xu. Atmospheric Environment .103, 147 (2015).
12. Lei Huang, Nicholas Anastas, Peter Egeghy, Daniel A, Vallero, Olivier Jolliet, Jane Bare, The International Journal of Life Cycle Assessment. 24, 1009 (2019).
13. Xingzi Ouyang, Meng Xia, Xueyou Shen, Yu Zhan, Journal of Environmental Sciences, 86, 107 (2019).
14. Juan Sun, Shuying Yang, Hasong Shen, Yanqiu Leng, Han Yan, Journal of China Executive Academy of Environmental Management, 26, 3 (2016).
15. Jerome H, Friedman, The Annals of Statistics, 29, 1189 (2001).
16. Heng Xia, Jian Tang, Junfei Qiao, Aijun Yan;Zihao Guo, Proceedings of the 32nd China Control and Decision-making Conference（5）[C], 221(2020).
17. Ning Zhou, Pengyu Xu, Jianxin Zhou, Proceedings of 2020 International Conference on Data Science and Information Technology （DSIT 2020）[C]. 6 (2020).
18. Wanhu Zhang, Junqi Yu, Anjun Zhao, Xinwei Zhou, Energy Reports, 7, 1588 (2021).
19. Haotian Xu, Frontiers in Economics and Management, 2, 3 (2021).
20. Yaofang Zhang, Jian Chen , Highway, 67, 221 (2022).