Prediction of household dust mite concentration based on machine learning algorithm

Chanjuan Sun[1,*], *Leyang* Li[1], *Shijie* Hong[2], *Chen* Huang[1], *Jingguang* Li[3], and *Zhijun* Zou[1]

[1]University of Shanghai for Science and Technology, China.

[2]Shanghai Tenth People's Hospital

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

Abstract. Household dust mites (HDMs) are the important allergens causing allergic diseases in children. A predictive model can help us understand the concentration of HDMs in different areas of China to better prevent and control this kind of allergen. This study used 454 household inspection samples in childrens' room obtained from China, Children, Homes, Health (CCHH) phase 2 study, conducted during 2013-2014. Spearman correlation and multiple logistic regression were used to explore the influencing factors of HDMs concentrations, by comprehensively considering residents' lifestyle, building characteristics, environmental exposure, especially dampness-related exposures. This study used the Gradient Boosting Decision Tree(GBDT) algorithm to build the prediction model. The data from CCHH were used to established the prediction model. It was found that there were some differences in the influencing factors between two types of HDMs. The concentration of HDMs were found a significant correlation ($p<0.05$)with the number of indoor moisture indicators. 17 influencing factors of HDMs concentrations from four 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 predicting the HDMs concentrations in children's bedrooms and the influence of the influencing factors.

1 Introduction

In recent decades, the incidence of allergic diseases in children in China has shown an upward trend. Household dust mites(HDMs) are one of the common inhaled allergens which caused allergic diseases[5-7,12-16]. Epidemiological surveys worldwide have shown that HDMs are important allergens. The most common mite species producing allergens are Der f1(D. farinae) and Der p1(Dermatophagoides pteronyssinus). The influencing factors of dust mites concentration are very necessary to study. In this paper, we determine the influencing factors from four aspects : residents' lifestyle, building characteristics, environmental exposure and dampness-related exposures. At present, the acquisition method of HDMs concentration is mainly on-site sampling and detection. This method is costly and time-consuming, so it is of certain significance to build a prediction model of HDMs concentration.

Studies have shown that the concentration of HDMs and indoor humidity, people ' s living habits, and architectural features have a certain relationship, and found some factors affecting the concentration of dust mite, such as our ear familiar wash sheets, etc., the article found 17 kinds of dust mite concentration factors through correlation analysis. The prediction model based on influencing factors and dust mites concentration can reflect the HDMs concentration level in the house.

In recent years, machine learning algorithms have been applied to many occasions to do prediction models. Zhao et al. used XGBoost algorithm to predict the risk level of expressway interchange exit, and obtained accurate prediction results(93.69%)[1]. Huang et al. used machine learning algorithm to build the prediction model of summer precipitation in Hunan and achieved good prediction results[2]. Li et al. predicted indoor PAEs concentration by BP neural network model, and got better prediction results[3]. The research shows that the prediction accuracy of ensemble learning algorithm is better than that of support vector machine 、 BP neural network algorithm etc.[2,19-22], so the Gradient Boosting Decision Tree (GBDT) algorithm is used to build the prediction model, and the dust mite concentration is used as the prediction target.

2 Features selection

2.1 Data collection

This study used 454 household inspection samples in children's room from Shanghai obtained from China, Children, Homes, Health (CCHH) phase 2 study[4]. The dust samples of children's bedroom were collected.

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

The concentrations of house dust mites and dust mites in the samples were determined by enzyme-linked immunosorbent assay (ELISA). Finally, 382 samples and corresponding questionnaire contents were screened out (see Table 1). These 382 samples were used to build prediction model.

Table 1. Questionnaire on related influencing factors.

Influencing factors	Questions	Options
Building characteristics	Flooring in child's room Cement floor	0=”no”，1=”yes”
	Wall covering in child's room_Lime	0=”no”，1=”yes”
	Windows type	0 = “single-layer glass”, 1 = “double-layer glass”, 2 = “three-layer glass”
Residents' lifestyle	Frequency of opening the window -Spring	0 = “never”, 2 = “sometimes”, 3 = “frequent”
	Frequency of opening the window -Summer	0 = “never”, 2 = “sometimes”, 3 = “frequent”
	Frequency of opening the window -Fall	0 = “never”, 2 = “sometimes”, 3 = “frequent”
	Frequency of opening the window -Winter	0 = “never”, 2 = “sometimes”, 3 = “frequent”
	Use air purification equipment at home？	0=”no”，1=”yes”
	Do you ventilate bedroom in autumn and winter ?	0 = “open a little ”, 1 = “open small ”, 2 = “open every day”, 3 = “open weekly times ”, 4 = “open less than weekly times ”
Dampness-related exposures	Wet spots around exterior windows	0 = “no”, 1 = “weak”, 2 =” small”, 3 = “obvious”
	Internal wall wet spots	0 = “no”, 1 = “weak”, 2 =” small”, 3 = “obvious”
	Does the floor paint fall off?	0 = “no”, 1 = “weak”, 2 =” small”, 3 = “obvious”
	Does the window have condensation phenomenon?	0=”none”，1<=5cm, 2=5-25cm, 3>=25cm

2.2 Correlation analysis

In this paper, the influencing factors of HDMs concentrations were divided into four categories : residents' lifestyle, building characteristics, environmental exposure(average temperature and average relative humidity of children' s bedroom), and dampness-related exposures. Through normal distribution test, it was found that the concentrations of Derp1 and Derf1 are not normal distribution. When analyzing the correlation of variables, Spearman correlation coefficient was used to do correlation analysis. The results are shown in table 2. It was found that there were some differences in the influencing factors between two types of HDMs. The Der f1 had a significant negative correlation with temperature (sig: 0. 000,r: -0. 624), but had no significant correlation with humidity. However, the Der p1 had a correlation with humidity (sig: 0. 000,r:0. 172), but had no significant correlation with temperature.

The number of wet characterization items was quantified from 0 to 8 according to the eight wet characterizations in the phase 2 of the CCHH questionnaire. According to the concentration of dust mite, the family was divided into four subgroups by quartiles. The correlation results between dust mites concentration and the number of moisture characterization items are shown in Table 3 through multivariate logistic regression analysis. The sub-groups from the upper bound to the upper quartile were used for reference, and found a significant correlation ($p<0.05$) between the concentration of both of Der p1 and Der f1 in the second quartile(25%~50%)and the number of indoor moisture indicators by multivariate logistic regression.

The prediction model was established by using the influencing factors in Table 2 which are correlated with Der p1 and Der f1 and the number of wet characterization items as eigenvalues.

Table 2. Correlation analysis of influencing factors of dust mites concentration.

Influencing factors	Correlation coefficient (r)	
	Der p1	Der f1
Flooring in child's room_Cement floor	. 113*	0. 007
Wall covering in child's room_Lime	. 093*	0. 065
Windows type	. 145**	0. 065
Ventilation in autumn and winter	0. 068	0. 103
Use of air purification equipment	. 170**	0. 15
Window opening habit in spring	-. 136**	-0. 009
Window opening habit in summer	0. 029	. 097*
Window opening habit in autumn	-. 138**	-0. 006
Window opening is used to winter	-. 137**	-0. 027
Average temperature	-. 624**	0. 036
Average relative humidity	0. 002	. 172**
Wet spots around exterior walls	0. 035	-. 097*
Moisture spots around the inner wall	-0. 009	-. 099*
Floor paint shed	. 000	-. 123**
Water loss of windows	. 125**	0. 014

1)* * At 0. 01 level (double tails), the relevance is significant.

2)* At 0. 05 level (double tail), the correlation is significant.

Table 3. Multivariate logistic regression analysis of dust mites concentration and the number of indoor moisture indicators.

Items		OR,95%CI	
Concentration (Refer to the first quartile)		DP1	DF1
25%-50%	the number of indoor moisture indicators	0. 832，0. 705-0. 982*	0. 809，0. 693-0. 945**
50%-75%		0. 695，0. 452-1. 067	0. 826，0. 708-0. 962*
75%-100%		0. 974，0. 858-1. 105	0. 863，0. 746-0. 999*

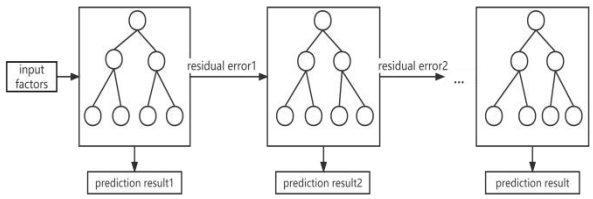
1)* * At 0. 01 level (double tails), the relevance is significant.

2)* At 0. 05 level (double tail), the correlation is significant.

3 Results and discussion

3.1 Construction of GBDT prediction model

The ensemble learning algorithm has the advantages of preventing overfitting, strong generalization ability and flexible processing of various types of data. GBDT (Gradient Boosting Decision Tree) is an integrated learning algorithm based on the regression tree and the boosting algorithm obtains the final result by iterative accumulation of multiple decision trees[18-22]. The process of GBDT algorithm is as follows: At first, initializing the weak learner; then, in the model, the negative gradient value of the loss function is calculated for all samples and used as an estimate of the residual which is trained as the true value of the new sample, and a new regression tree is obtained. The best fitting residual approximation is calculated according to the area of the regression leaf node; next, used the result update the strong learner and then repeat the steps for M times; lastly to get the prediction model.



1) Residual error means the difference between real value and prediction value.

2) Input all the factors we found to train the decision tree.

Fig1. GBDT Process Diagram.

3.2 Prediction results

The predicted concentrations of DP1 and DF1 were compared with the measured data. Mean Square Error (MSE), Mean Absolute Error (MAE) and R Squared (R^2) are commonly used to evaluate the prediction performance of prediction models by measuring the deviation between the predicted and measured values. These ratios were used in this paper to evaluate the results. When MSE and MAE closer to 0, R^2 closer to 1, the prediction result is more ideal. The Table 4 shows the accuracy predicted by the prediction model,which reaching a good fitting results.

In this paper, only some influencing factors with significant differences in correlation analysis are adopted, some influencing factors proved by previous studies[4,12], such as the frequency of washing bed supplies, are not included in the characteristics, so the prediction model needs to be further expanded.

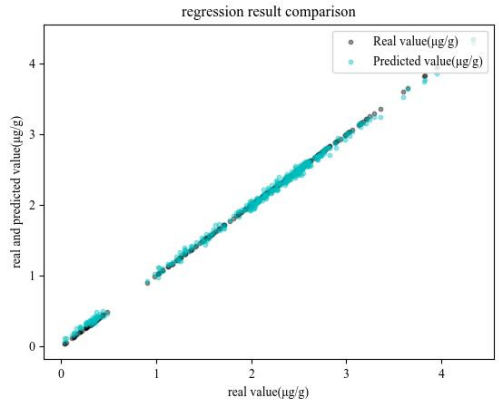


Fig2. Comparison of DP1 real value and predicted value

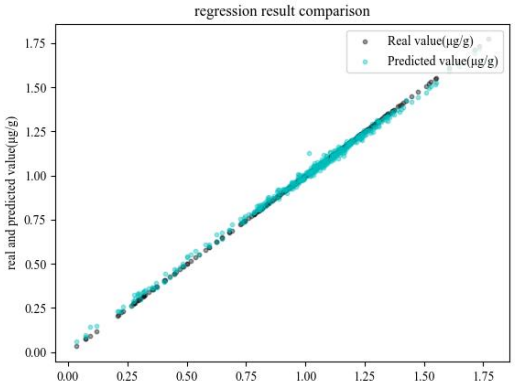


Fig3. Comparison of DF1 real value and predicted value

Table 4. The prediction accuracy of two kinds of HDMs.

HDMs	MAE	MSE	R^2
DP1	0.037	0.002	0.997
DF1	0.046	0.004	0.996

4 Conclusions

In this paper, the correlation analysis was used to find the influencing factors of HDMs concentration and to build the subsequent prediction model. GBDT algorithm establishes the prediction model of dust mite concentration in children's bedroom, and proves the good prediction results of the model through the accuracy evaluation standard, which provides a certain

way to control the dust mite concentration in children' s bedroom.

References

[1] Zhao Xiaohua, Qi hang, Yao Ying, Guo Miao, Guo Jingfeng, Zhang Yunlong. JSUE-EE 1. 10(2022).

[2] Huang Chao, Li Qiaoping, Xie Yijun, Peng Jiadong.Journal of atmospheric science 2,12 (2022).

[3] Li Kexiu, sun Chanjuan, Zhang Jialing, Zou Zhijun, Huang Chen.Environmental monitoring management and technology 1. 4 (2022).

[4] Cai Jiao.Shanghai University of Technology.

[5] Platts-Mills, T. , Erwin, E. A. , Heymann, P. W. , Woodfolk, J. A. .Am. J. Respir. Crit. Care Med. 2. 109 (2009).

[6] Stephenson, J., P. .Publ. Health. 3. 95 (1991) .

[7] Wei, L. A. , Jiao, C. B. , Chen, H. C. , Zz, C. , Cs, C. , Bl, B. .Sustainable Cities and Society. 54(2019).

[8] Pei, J. , Gong, J. , Wang, Z. Building and Environment. 183(2020).

[9] Cs, A. , Ping, W. A. , Xin, H. A. , Kl, A. , Sh, B. , Zz, A. , . Building and Environment. 206(2021).

[10] Sun, C. , Li, K. , Zhang, J. , Huang, C. .Indoor and Built Environment.1. 230 (2022).

[11] Cai, J. , Huang, C. , Liu, W. , Hu, Y. , Zou, Z. . Procedia Engineering. 121. 1948(2015).

[12] Huang, C. , Cai, J. , Liu, W. , Wang, X. , Zou, Z. , Sun, C. Building and Environment. 151. 198 (2019) .

[13] Ostrom, Nancy, K. Pediatrics. 2. 300(1992).

[14] Ya-Nan F U. Journal of Environment and Health, (2013).

[15] Jiang, W. , Dong, J. I. , Gui, X. Chinese Archives of Otolaryngology-Head and Neck Surgery. (2016).

[16] Couper, Ponsonby, Dwyer. Clinical & Experimental Allergy, 6. 715(2010).

[17] K., C. , Dannemiller, J. , F. , Gent, . Indoor Air, (2015).

[18] Friedman, J. H. .Computational Statistics & Data Analysis. 38(2002).

[19] Han Qidi, Zhang Xiaotong, Shen Wei. Bulletin of mineral and rock geochemistry. 6. 8(2018).

[20] Zheng Kaiwen, Yang Chao. Guizhou power technology,2. 4 (2017).

[21] Y Zhang, J Chen. . Highway, 01. 221(2022).

[22] J Cheng, X Chen. Journal of Southeast University : English version, 3. 6(2019).