TYPE Original Research
PUBLISHED 10 August 2023
DOI 10.3389/fpubh.2023.1213453

Check for updates

OPEN ACCESS

EDITED BY
Perng-Jy Tsai,
National Cheng Kung University, Taiwan

REVIEWED BY
Mostafa Yuness Abdelfatah Mostafa,
Minia University, Egypt
Wanzhou Wang,
Peking University, China

*CORRESPONDENCE
Chunyang Dong
✉ Dongchunyang@scdc.sh.cn
Shaofeng Sui
✉ Suishaofeng@scdc.sh.cn

[†]These authors share first authorship

RECEIVED 16 May 2023
ACCEPTED 28 July 2023
PUBLISHED 10 August 2023

CITATION
Shi Y, Du Z, Zhang J, Han F, Chen F, Wang D, Liu M, Zhang H, Dong C and Sui S (2023) Construction and evaluation of hourly average indoor $PM_{2.5}$ concentration prediction models based on multiple types of places.
Front. Public Health 11:1213453.
doi: 10.3389/fpubh.2023.1213453

Construction and evaluation of hourly average indoor $PM_{2.5}$ concentration prediction models based on multiple types of places

Yewen Shi[1†], Zhiyuan Du[2†], Jianghua Zhang[1], Fengchan Han[1], Feier Chen[1], Duo Wang[1], Mengshuang Liu[1], Hao Zhang[2], Chunyang Dong[1*] and Shaofeng Sui[1*]

[1]Shanghai Municipal Center for Disease Control and Prevention, Shanghai, China, [2]Department of Environmental Health, Key Laboratory of the Public Health Safety, Ministry of Education, School of Public Health, Fudan University, Shanghai, China

Background: People usually spend most of their time indoors, so indoor fine particulate matter ($PM_{2.5}$) concentrations are crucial for refining individual $PM_{2.5}$ exposure evaluation. The development of indoor $PM_{2.5}$ concentration prediction models is essential for the health risk assessment of $PM_{2.5}$ in epidemiological studies involving large populations.

Methods: In this study, based on the monitoring data of multiple types of places, the classical multiple linear regression (MLR) method and random forest regression (RFR) algorithm of machine learning were used to develop hourly average indoor $PM_{2.5}$ concentration prediction models. Indoor $PM_{2.5}$ concentration data, which included 11,712 records from five types of places, were obtained by on-site monitoring. Moreover, the potential predictor variable data were derived from outdoor monitoring stations and meteorological databases. A ten-fold cross-validation was conducted to examine the performance of all proposed models.

Results: The final predictor variables incorporated in the MLR model were outdoor $PM_{2.5}$ concentration, type of place, season, wind direction, surface wind speed, hour, precipitation, air pressure, and relative humidity. The ten-fold cross-validation results indicated that both models constructed had good predictive performance, with the determination coefficients (R^2) of RFR and MLR were 72.20 and 60.35%, respectively. Generally, the RFR model had better predictive performance than the MLR model (RFR model developed using the same predictor variables as the MLR model, $R^2 = 71.86\%$). In terms of predictors, the importance results of predictor variables for both types of models suggested that outdoor $PM_{2.5}$ concentration, type of place, season, hour, wind direction, and surface wind speed were the most important predictor variables.

Conclusion: In this research, hourly average indoor $PM_{2.5}$ concentration prediction models based on multiple types of places were developed for the first time. Both the MLR and RFR models based on easily accessible indicators displayed promising predictive performance, in which the machine learning domain RFR model outperformed the classical MLR model, and this result suggests the potential application of RFR algorithms for indoor air pollutant concentration prediction.

KEYWORDS

indoor air quality, $PM_{2.5}$, prediction models, machine learning, random forest

1. Introduction

$PM_{2.5}$ refers to particulate matter with an aerodynamic diameter of 2.5 μm or less, which is one of the environmental pollutants with the greatest impact on public health (1–3). Numerous epidemiological studies have shown that both long-term and short-term exposure to $PM_{2.5}$ increases the risk of death from respiratory and cardiovascular diseases in the population (4–6). Studies have shown that for every 10 g/m^3 increase in the average concentration of $PM_{2.5}$ in ambient air, there is a 3.1% increase in hospital admissions and a 2.5% increase in mortality from chronic obstructive pulmonary disease (7). Furthermore, there is a 3% increase in emergency department visits for bronchial asthma (8), a 16% increase in the risk of death from ischemic heart disease, and a 14% increase in mortality from stroke (4, 9).

Currently, most relevant studies use ambient $PM_{2.5}$ concentrations as a surrogate for human $PM_{2.5}$ exposure without taking into account the difference between indoor and outdoor $PM_{2.5}$ concentrations as well as the contribution of indoor $PM_{2.5}$ exposure to actual human exposure, which limits the interpretation of their results. As most people spend at least 80% of their day indoors, and for some specific populations such as the older adults and children, this percentage is even higher (10–12). Therefore, indoor $PM_{2.5}$ concentration is crucial for accurate $PM_{2.5}$ exposure assessment and health risk assessment. Direct measurement of indoor $PM_{2.5}$ concentration can provide the most accurate data; however, such practice is not easy to achieve, as it requires a lot of manpower and material resources as well as the compliance of the research participants, especially for large-scale population and/or long-term studies. When direct measurement is difficult to achieve, it is important to construct appropriate predictive models.

At present, many studies have been conducted to establish prediction models for indoor $PM_{2.5}$ concentration (12–18), mainly involving multiple linear regression (MLR) models and random forest regression (RFR) models, which have their own advantages and disadvantages. For indoor $PM_{2.5}$ concentration, there is still controversy about which model has a better predictive effect. In addition, the models in these studies have mostly predicted the average indoor $PM_{2.5}$ concentration on one or more days, and do not adequately account for the fluctuation of indoor $PM_{2.5}$ concentration during the day (or longer) and the variability of individual behaviors over time (19–21). Obviously, the establishment of indoor $PM_{2.5}$ concentration prediction models with higher temporal resolution is of more practical significance to improve individual $PM_{2.5}$ exposure assessment. The existing models were constructed using indoor $PM_{2.5}$ concentration monitoring data from a single type of place, which is not universal enough and inevitably limits the practical application to different types of places. No study has yet established prediction models for hourly average indoor $PM_{2.5}$ concentration based on data from multiple types of places.

In this study, monitored data on indoor $PM_{2.5}$ concentrations from five types of typical sites (offices, primary and secondary schools, kindergartens, shopping malls, and restaurants) in Shanghai were collected during different seasons. The data were used to develop and evaluate predictive MLR and RFR models for indoor $PM_{2.5}$ temporal average concentrations based on multiple types of places. The aim of the study was to provide a feasible way to improve individual $PM_{2.5}$ exposure assessment.

2. Materials and methods

2.1. Data collection

Five types of typical locations – offices, middle and primary schools, kindergartens, shopping malls, and restaurants – were selected for indoor $PM_{2.5}$ concentration field monitoring in 16 districts of Shanghai. A TSI DustTrak 8,530 benchtop aerosol monitor (TSI Incorporated, Shoreview, MN, United States) was used for the monitoring. One floor was selected as the monitoring site for the high, middle, and low areas of office buildings, shopping malls, and restaurants. Two, four, and six monitoring points were set for indoor areas of 200–1,000 m^2, 1,001–5,000 m^2, and over 5,000 m^2, respectively. Two classrooms from each floor were used as monitoring sites in high, middle, and low areas of kindergartens, middle, and primary schools. One, three, and five monitoring points were set for indoor areas of less than 50 m^2, 50–100 m^2, and more than 100 m^2, respectively. All of the above points were distributed evenly on the diagonal of the room or in a plum style, and the height of each point was set at the level of a human respiratory belt (0.8–1.2 m). The actual measurement time was in January, April, July, and October of 2018 (the 4 months represented the four seasons of the year: January for winter, April for spring, July for summer, and October for autumn). Indoor $PM_{2.5}$ concentrations in each location were monitored for 1 week during these 4 months, with each instrument monitoring the concentrations every 15 min, which covered all times of the day (00,00–23,00 h) to ensure full coverage of people's activities in various places as much as possible.

For the construction of prediction models, we used the findings of relevant publications (17, 21–24) to identify 11 easily accessible indicators that may have significant effects on indoor $PM_{2.5}$ concentrations. The relevant information of the indicators could be found in Supplementary Table S1. The outdoor $PM_{2.5}$ and PM_{10} concentration data were obtained from the monitoring stations of 16 municipal control points in Shanghai. By calculating the distance between all government-controlled monitoring stations and the indoor places we monitored, the data from the closest station was selected as outdoor $PM_{2.5}$ and PM_{10} concentration data for indoor places. Meteorological data for the same period were obtained from the European Center for Medium and Long-Range Weather Forecasts, which included outdoor temperature, relative humidity, air pressure, precipitation, surface wind speed, and wind direction.

2.2. Data analysis

The data analysis in this study was based on the arithmetic mean of time, that is, the indoor and outdoor $PM_{2.5}$ concentrations, outdoor PM_{10} concentration, as well as related meteorological parameters were processed as hourly mean values for use. For example, the indoor $PM_{2.5}$ concentration at 09:00 h was actually the mean value of 08:00 h to 09:00 h. Following a series of data washing, the final database consisted of 11,712 records, 11 potential predictor variables, and natural log-transformed indoor $PM_{2.5}$ concentrations (approximately normally distributed) as response variables for MLR and RFR model construction. Data analysis and model construction in this study were performed with R software (version 4.1.0), and statistical significance levels were set at p values of <0.01 and < 0.05 (both sides).

2.3. MLR model construction steps

A sensitivity analysis was conducted for the effects of different variable screening methods on the predictive efficacy of MLR models. The three adopted types of variable screening were as follows: 1) manually supervised forward linear regression commonly used in reference to classical land-use regression modeling (25, 26), 2) stepwise regression (backward, variables with regression coefficient $p < 0.05$ were retained), and 3) least absolute shrinkage and selection operator (Lasso). The manually supervised forward linear regression method was used to build a basal multiple regression model in three steps: 1) After testing the premise assumptions of the regression model, all potential predictor variables expected to be included in the model were first univariately regressed against the response variable (natural log-transformed hourly average $PM_{2.5}$ concentration), and predictor variables with significant ($p < 0.05$) regression coefficients were retained for the next step, 2) Correlations between prediction variables were tested. Among the prediction variables that were highly correlated with other prediction variables (Spearman $r > 0.50$, $p < 0.05$), only the prediction variable with the highest coefficient of determination (R^2) was retained for further analysis, 3) The predictor variables that remained after the previous two steps were sorted according to R^2 (from highest to lowest), and then each predictor was entered into the regression model in order. Finally, only those predictor variables with significant partial regression coefficients ($p < 0.05$), which boosted the R^2 of the model by more than 1% and whose coefficients were consistent with the priori hypothesis (such as a positive coefficient of outdoor $PM_{2.5}$), were retained.

In the process of MLR model diagnosis, variance inflation factors of the predictive variables were tested to evaluate multicollinearity. Additionally, considering that season may modify the effects of other potential predictor variables on indoor $PM_{2.5}$ concentration, we stratified the data by winter–spring (January, April) and summer-autumn (July, October) seasons and developed season-specific prediction models.

2.4. RFR model construction steps

Random forest model is a machine learning model that realizes the classification and/or prediction for unknown samples through the integrated learning with a large number of decision trees, which is now widely used in the processing of big data due to its fast computing speed, high prediction accuracy, and strong anti-interference (27–29). This model possesses two significant characteristics, namely sample randomization and variable randomization. Bagging algorithm is the basis of the random forest model, which is also known as bootstrap sampling algorithm, in short, there is put back to the random collection of samples to form a different set of data to train the base learner, so as to realize the mutual independence of individual learners. The Random Forest algorithm extends and expands the Bagging algorithm. In addition to random sampling of samples, the Random Forest algorithm also incorporates random selection of variables at each attribute node of the classification tree, which further enhances the diversity of each decision tree, reduces the risk of model overfitting, and can effectively improve the generalization performance of the final ensemble model (27, 29). The prediction accuracy and generalization of a Random Forest model are closely related to two important hyperparameters, which are ntree (the number of trees used) and mtry (the number of variables used for binary trees in the specified nodes). The randomForest package of R software (version 4.1.0) was used to construct the RFR model. In our analysis, different values were set for these two parameters as sensitivity analysis in order to obtain maximum model prediction effectiveness. The increase in mean squared error (%IncMSE) of the predicted value was taken as an indicator to measure the importance of a variable, in other words, a random value was assigned to each prediction variable. If the prediction variable is important, the prediction error of the model will increase after its value is randomly replaced, so the larger the value, the more important the variable is.

In order to evaluate and compare the prediction efficiency of the MLR model and the RFR model for indoor hourly average $PM_{2.5}$ concentration in various types of places, we developed two RFR models. The first RFR model was called the Full variables-RFR model (Full-RFR). Since the RFR model does not need to consider preconditions such as the independence of predictive variables that are faced by general MLR models, all 11 potential predictive variables were included in the model. The second RFR model was called the Conjoint-RFR model (Conjoint-RFR). In order to compare the MLR and RFR models, this Conjoint-RFR model was established using the same predictor variables as the MLR model with the best prediction performance identified in the previous steps.

2.5. Evaluation of models

The R^2 and root mean squared error (RMSE) calculated based on the predicted and measured values of the model were used as the model performance evaluation indexes. In addition, the generalization performance of the model was evaluated by a ten-fold cross-validation (CV) method. In short, the entire dataset was randomly and equally divided into ten subsets, nine of which were selected as the training set and the remaining one was used as the test set to test the prediction performance of the model. This process was repeated 10 times until each subset was used for one verification (30).

3. Results

3.1. Indoor $PM_{2.5}$ pollution in various places

The summary of hourly average indoor $PM_{2.5}$ concentration statistics for each site was shown in Table 1. In general, the median hourly average indoor $PM_{2.5}$ concentration was 34.9 μg/m^3 and the interquartile range was 24.5 μg/m^3, with a few readings on the high side and a maximum value of 288 μg/m^3. The result of Welch analysis of variance (Welch ANOVA) (31) showed significant differences ($p < 0.01$) in the hourly average indoor $PM_{2.5}$ concentrations in different types of places. The highest hourly average indoor $PM_{2.5}$ concentrations were found in restaurants (44.4 μg/m^3), probably because of frequent cooking in restaurants that produces a large amount of grease smoke and causes indoor $PM_{2.5}$ concentrations to increase (32). The Ambient Air Quality Standards (GB 3095–2012) of China and the Environmental Protection Agency of the United States have set the daily average ambient $PM_{2.5}$ concentration limit at 35 μg/m^3. No clearly established indoor $PM_{2.5}$ concentration standard exists in China; therefore, the daily average ambient $PM_{2.5}$ concentration standard and the classification method of the China Environmental Monitoring

TABLE 1 Hourly average indoor $PM_{2.5}$ concentrations in each place (μg/m^3).

Type of place	*n*	Arithmetic mean	SD	Percentiles				
				Min	P25	P50	P75	Max
Office	3,438	23.8	15.6	0.667	12.0	18.8	32.5	128
Middle and primary school	1,812	33.5	20.2	0.20	20.2	30.9	43.1	127
Kindergarten	1,441	37.6	19.8	5.19	24.2	34.3	45.9	154
Shopping mall	2,443	31.2	15.8	4.0	20.0	28.8	39.5	133
Restaurant	2,578	52.6	34.2	4.86	28.3	44.4	67.2	288
Overall	11,712	34.9	24.5	0.2	17.7	30.1	44.1	288

n, the number of samples; SD, standard deviation; Min, the minimum value; P25, P50, and P75: the 25th, 50th, and 75th percentiles, respectively; Max, the maximum value; $PM_{2.5}$ refers to particulate matter with an aerodynamic diameter of 2.5 μm or less.

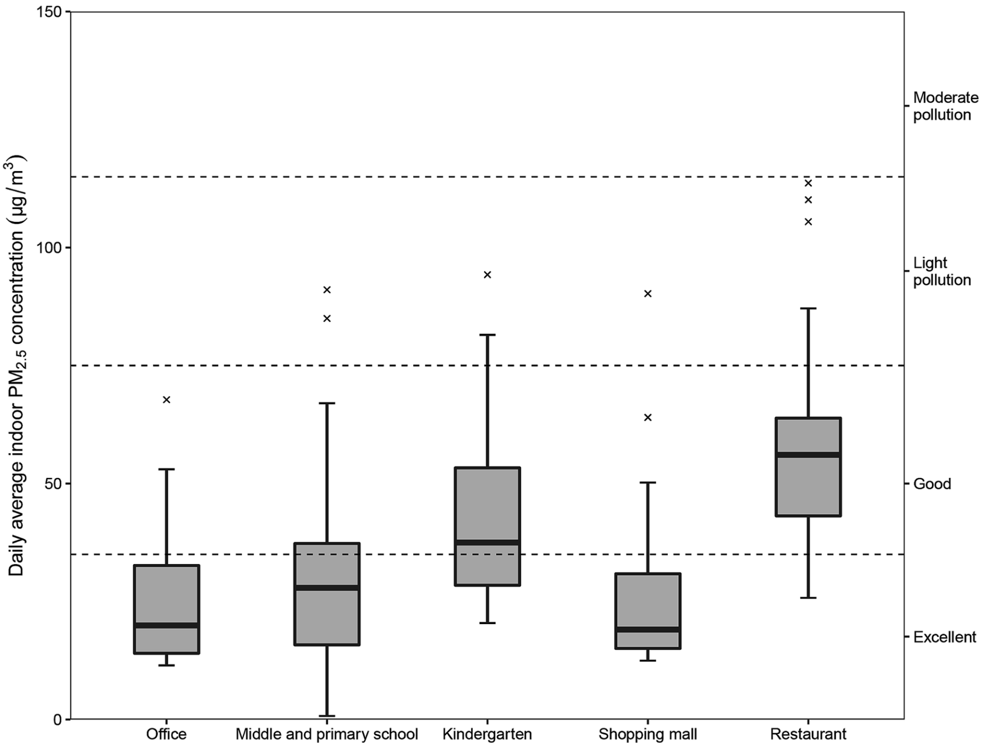

FIGURE 1
Daily average indoor $PM_{2.5}$ concentrations (μg/m^3) in each place. In reference to the classification method of the China Environmental Monitoring Station: 0–35 μg/m^3 is excellent; 35–75 μg/m^3 is good; 75–115 μg/m^3 is light pollution; and 115–150 μg/m^3 is moderate pollution. $PM_{2.5}$ refers to particulate matter with an aerodynamic diameter of 2.5 μm or less.

Station were used here to characterize the indoor $PM_{2.5}$ pollution in each location (Figure 1). In terms of 35 μg/m^3 as the standard, indoor $PM_{2.5}$ exceeded the standard in different degrees in all places and restaurants were the worst offender, followed by kindergartens. The monitoring results suggest that the indoor environmental quality of these two types of places needs to be improved.

The changes of hourly average indoor $PM_{2.5}$ concentration at different times are shown in Figure 2. Overall, there were significant differences ($p < 0.01$) in indoor $PM_{2.5}$ at different times of the day, and we also observed significant intraday fluctuations in the monitoring data for each type of place ($p < 0.05$). The variability of $PM_{2.5}$ concentration at different times of the day in multiple types of places is closely related to the nature of the place. For example, the fluctuation of $PM_{2.5}$ concentration in the restaurant was as expected ($p < 0.01$), with two peaks occurring after 11:00 and after 17:00, which are roughly the beginning of lunch and dinner. At these times, intensive cooking leads to higher indoor $PM_{2.5}$ concentrations, and similar patterns were observed in other places (Figure 2). These results demonstrate the intraday variability of indoor $PM_{2.5}$ concentration as well as the spatial variability across places.

3.2. MLR model results

Univariate regression model results for hourly average indoor $PM_{2.5}$ concentration were summarized in Supplementary Table S2. All 11 prediction variables were significantly associated with hourly average indoor $PM_{2.5}$ ($p < 0.05$). The R^2 of the 11 prediction variables

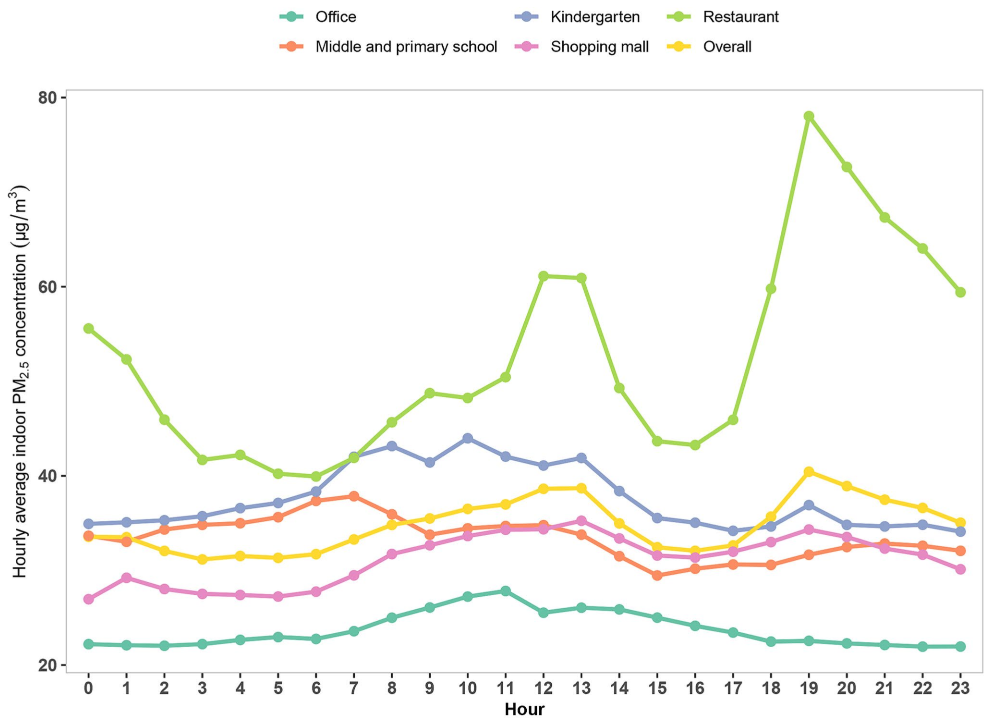

FIGURE 2
Variation of intraday hourly average indoor $PM_{2.5}$ concentration in each place (μg/m^3). $PM_{2.5}$ refers to particulate matter with an aerodynamic diameter of 2.5 μm or less.

ranged from 0.1 to 30.54%, among which nine variables exceeded 2%, with the largest R^2 for outdoor $PM_{2.5}$ concentration (30.54%), followed by outdoor PM_{10} concentration ($R^2 = 28.76\%$), season ($R^2 = 24.05\%$), type of place ($R^2 = 17.11\%$), and wind direction ($R^2 = 8.64\%$). The final MLR model for log-transformed hourly average indoor $PM_{2.5}$ concentrations were shown in Table 2. The model which was developed based on the stepwise regression method had the best prediction performance (CV $R^2 = 60.48\%$) and the lowest prediction error (CV RMSE = 0.44) among the three MLR models (Table 3). In this paper, the relative importance of the predictor variables within MLR model was determined using the "Lindeman, Merenda and Gold (LMG)." LMG was evaluated as the most successful indicator of the relative importance of independent variables, which was implemented by using the "relaimpo" package of R software (33, 34) (Figure 3). Outdoor $PM_{2.5}$ concentration was the most important predictor variable, with an R^2 share of 33.91%, followed by type of place (27.62%), season (26.22%), wind direction (4.88%), and surface wind speed (2.80%). The two models developed after stratification by winter–spring and summer-autumn incorporated similar predictor variables, of which the R^2 and RMSE after cross-validation were also remarkably close (winter–spring model: $R^2 = 58.23\%$, RMSE = 0.38; summer-autumn model: $R^2 = 58.79\%$, RMSE = 0.49; Supplementary Tables S3–S5).

3.3. RFR model results

We compared and analyzed all RFR models with ntree of 200, 500, 1,000 and mtry of 1 ~ 11 (Supplementary Figure S1), and finally determined that ntree = 200 and mtry = 2 were the most suitable RFR parameters for this study after fully considering the model's prediction effectiveness, prediction error, and model efficiency. Results from the Conjoint-RFR model, which used the same predictor variables as the MLR model, showed that the RFR model explained a greater proportion of the variance of indoor $PM_{2.5}$ time-averaged concentrations with an R^2 (RMSE) of 89.65% (0.23), which decreased in predictive efficacy (CV $R^2 = 71.86\%$) and increased in prediction error (CV RMSE = 0.37) after ten-fold cross-validation. Nevertheless, the overall performance of the model was still better than that of the corresponding MLR model (CV $R^2 = 60.48$; CV RMSE = 0.44). The performance of the Full-RFR model incorporating all predictor variables was better than that of the Conjoint-RFR model, with a CV R^2 (RMSE) of 72.20% (0.36). The importance results of the predictor variables from the random forest algorithm (Figures 4A,B) indicated that the top five variables in the Conjoint-RFR model (Figure 4B) in order of importance were type of place, outdoor $PM_{2.5}$ concentration, season, hour, and surface wind speed. Comparison of the importance ranking results of the variables in the Conjoint-RFR model and the corresponding MLR model shows that the top three variables in both models are the same, namely, outdoor $PM_{2.5}$ concentration, type of place, and season, but with a different order. By contrast, the variable "hour" appears in the top five variables in the Conjoint-RFR model but wind direction is in the top five in the MLR model.

4. Discussion

Significant differences in indoor $PM_{2.5}$ concentrations between various types of places and at different times of day were found in our study. The variable of "type of place" ranked first and second in the importance assessment of the predictor variables of the RFR model and

TABLE 2 Multiple linear regression (MLR) model for log-transformed hourly average indoor $PM_{2.5}$.

Predictive variables	Coefficients	Standard error	*p*-value	Partial R^2 (%)
Intercept	17.60	1.73	**<0.01**	
Outdoor $PM_{2.5}$	0.014	0.013	**<0.01**	33.91
Type of place	—	—	—	27.62
Office (reference)	—	—	—	
Middle and primary school	0.035	0.014	**<0.01**	
Kindergarten	0.23	0.015	**<0.01**	
Shopping mall	0.16	0.012	**<0.01**	
Restaurant	0.76	0.011	**<0.01**	
Season	—	—	—	26.22
Winter (reference)	—	—	—	
Spring	0.015	0.027	0.58	
Summer	−0.73	0.037	**<0.01**	
Autumn	−0.03	0.022	0.14	
Wind direction	0.00024	0.000062	**<0.01**	4.88
Surface wind speed	−0.008	0.0024	**<0.01**	2.80
Hour	—	—	—	2.62
0 (reference)	—	—	—	
1	−0.024	0.028	0.39	
2	−0.06	0.028	**<0.05**	
3	−0.10	0.028	**<0.01**	
4	−0.12	0.029	**<0.01**	
5	−0.13	0.029	**<0.01**	
6	−0.12	0.029	**<0.01**	
7	−0.06	0.029	**<0.05**	
8	−0.018	0.029	0.51	
9	0.043	0.028	0.13	
10	0.069	0.028	**<0.05**	
11	0.10	0.028	**<0.01**	
12	0.16	0.028	**<0.01**	
13	0.17	0.028	**<0.01**	
14	0.16	0.028	**<0.01**	
15	0.095	0.028	**<0.01**	
16	0.092	0.028	**<0.01**	
17	0.10	0.028	**<0.01**	
18	0.16	0.028	**<0.01**	
19	0.21	0.028	**<0.01**	
20	0.16	0.028	**<0.01**	
21	0.10	0.028	**<0.01**	
22	0.076	0.028	**<0.01**	
23	0.016	0.028	0.09	
Precipitation	−0.19	0.012	**<0.01**	1.22
Air pressure	−0.015	0.0012	**<0.01**	0.91
Relative humidity	0.26	0.042	**<0.01**	0.83

Significant *p*-values are in bold. $PM_{2.5}$ refers to particulate matter with an aerodynamic diameter of 2.5 μm or less; R^2, coefficient of determination.

the MLR model in this study, respectively. This result emphasized the importance of place type in predicting indoor $PM_{2.5}$ concentration and suggested that it might be difficult to extrapolate the prediction model based on a single type of place for use in other types of places. In fact, it is not difficult to understand the conclusion that the different functional attributes of each place naturally create a unique indoor microenvironment, which consequently affects the occurrence, diffusion, deposition and other behaviors of $PM_{2.5}$ (35–38). For example, in an office, there is a high concentration of people, frequent use of office equipment (e.g., printers, photocopiers and computers), and air-conditioning equipment (air-conditioners, humidifiers, air filters), with low ventilation and a single source of indoor pollution, whereas in a shopping mall there is a higher flow of people, more frequent ventilation, and a more complex internal environment. In contrast, the frequent cooking activities in restaurants generate smoke and high temperatures, creating a different microenvironment than the places mentioned above (35, 39). However, the currently available prediction models for indoor $PM_{2.5}$ concentrations are all constructed based on monitoring data from a single type of place, such as residential buildings (16, 18, 40, 41), schools (19, 20), and offices (42), without considering the differences between various types of sites. This situation inevitably leads to limitations in the actual application for the assessment of indoor $PM_{2.5}$ exposure. At present, no study has attempted to construct an indoor $PM_{2.5}$ concentration prediction model based on monitoring data from multiple types of places, and our study has attempted to fill this gap. In addition, most existing studies have predicted indoor $PM_{2.5}$ concentration over a day or longer period (such as a week); however, many published studies have shown that indoor $PM_{2.5}$ concentrations have a large daily variability (19–21). According to a report by Che et al. (43), after conducting continuous monitoring of indoor air quality in 32 primary and secondary schools across Hong Kong, it was found that there were significant variations in $PM_{2.5}$ concentrations in classrooms at different times of the day. The $PM_{2.5}$ concentrations in classrooms during school hours were approximately 40% higher than non-school hours. Zhao et al. (44) reported that indoor $PM_{2.5}$ concentrations were 1.5 times higher at night than during the daytime in Beijing during winter. According to Xu et al. (13), indoor $PM_{2.5}$ concentrations at different moments of the day varied significantly, with the ratio of the highest to the lowest values even exceeding 15-fold. This temporal variability of indoor $PM_{2.5}$ may originate from outdoor sources, for example, factors such as changes in outdoor $PM_{2.5}$ concentrations, variations in wind direction, temperature, and atmospheric pressure throughout the day and night may contribute to the differences in indoor $PM_{2.5}$ concentrations (17, 44), or from indoor human activities, such as cooking, smoking, use of air purifiers, etc. (45, 46). No matter what causes this variability, establishing a higher temporal resolution in an indoor $PM_{2.5}$ concentration prediction model is more practical for refining individual $PM_{2.5}$ exposure assessment and health risk evaluation.

MLR models are widely used for indoor air quality prediction because of the advantages of simple methodology, easy application, and strong interpretation of results (13, 17, 47). However, prerequisites exist for MLR application. First, a linear relationship must exist between the prediction variable and the response variable. Second, the response variable must obey a normal distribution when each predictor variable takes a certain definite value. Third, the response variable must satisfy the homogeneity of variance when each predictor variable takes different values. Fourth, the predictor variables are independent of each other and do not have a very close statistical

TABLE 3 Summary of model performance evaluation results.

Models	Model-based indicators		Ten-fold cross-validation indicators	
	Coefficient of determination (R^2, %)	Root mean square error (RMSE)	Coefficient of determination (R^2, %)	Root mean square error (RMSE)
Basal MLR model	59.51	0.44	59.38	0.45
MLR with lasso selection	60.54	0.43	60.35	0.45
MLR with stepwise selection	60.67	0.43	60.48	0.44
Conjoint-RFR model	89.65	0.23	71.86	0.37
Full-RFR model	91.20	0.21	72.20	0.36

The multiple linear regression (MLR) models were developed by three different variable selection methods. Two random forest regression (RFR) models were developed using 200 trees.

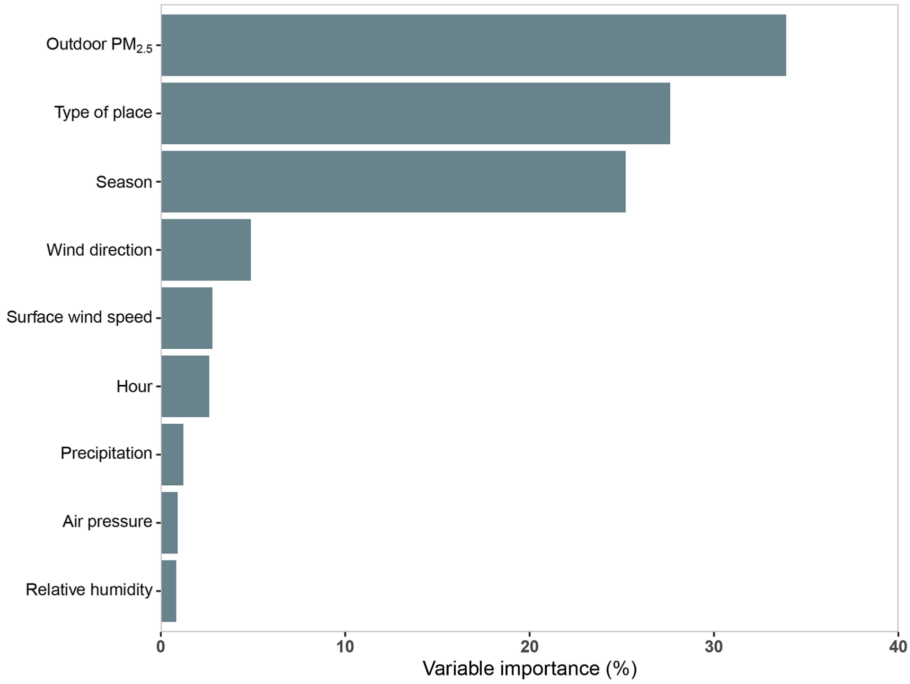

FIGURE 3
Relative importance of the multiple linear regression (MLR) model predictor variables. R^2, coefficient of determination; $PM_{2.5}$ refers to particulate matter with an aerodynamic diameter of 2.5 µm or less.

correlation. These prerequisites for MLR in practical applications are sometimes not easily satisfied.

With improvements in computing power and the advent of the era of big data, machine learning algorithms have been constantly enhanced and widely focused. The random forest algorithm is an integrated decision tree-based algorithm proposed by Breiman and Cutler in 2001, which can simultaneously construct a large number of decision trees in parallel and achieve significantly higher computational efficiency than other machine learning methods by integrating the learning of multiple decision trees (27, 29). Due to the inherent inclusion of interactions between variables in the random forest algorithm, there is no need to consider the issue of multicollinearity among variables in general models, and the algorithm performs robustly with mixed data types, missing data, non-equilibrium data, and extreme data, leading to a high prediction accuracy of the model (28). In addition, owing to the inclusion of sample perturbation and attribute perturbation in the algorithm, the random forest model can effectively limit overfitting and is regarded as one of the best algorithms today (48–50). Of course, random forest models also have certain drawbacks, such as poor interpretability of the model, which is usually considered as a black box model. Furthermore, categorical variables with more levels will have a greater impact on the model results than those with fewer levels, which may lead to a deviation in the prediction results (48, 51).

In our study, MLR and RFR prediction models were developed for hourly average indoor $PM_{2.5}$ concentrations based on monitoring data from multiple types of places. As a conventional and classical prediction model, the MLR model is widely used to predict indoor $PM_{2.5}$ concentration. Our MLR model (CV $R^2 = 60.48\%$) had a relatively high predictive performance compared with published MLR prediction models of indoor $PM_{2.5}$ concentration based on 1 day or longer (such as 1 week) whose R^2 values ranged from 33 to 87% (13, 16, 18, 19, 52–54). To the best of our knowledge, only one study by Xu et al. (13) has developed an MLR prediction model for hourly average indoor $PM_{2.5}$ concentration. In this study, two MLR models were developed for two regions with CV R^2 values of 71 and 75%.

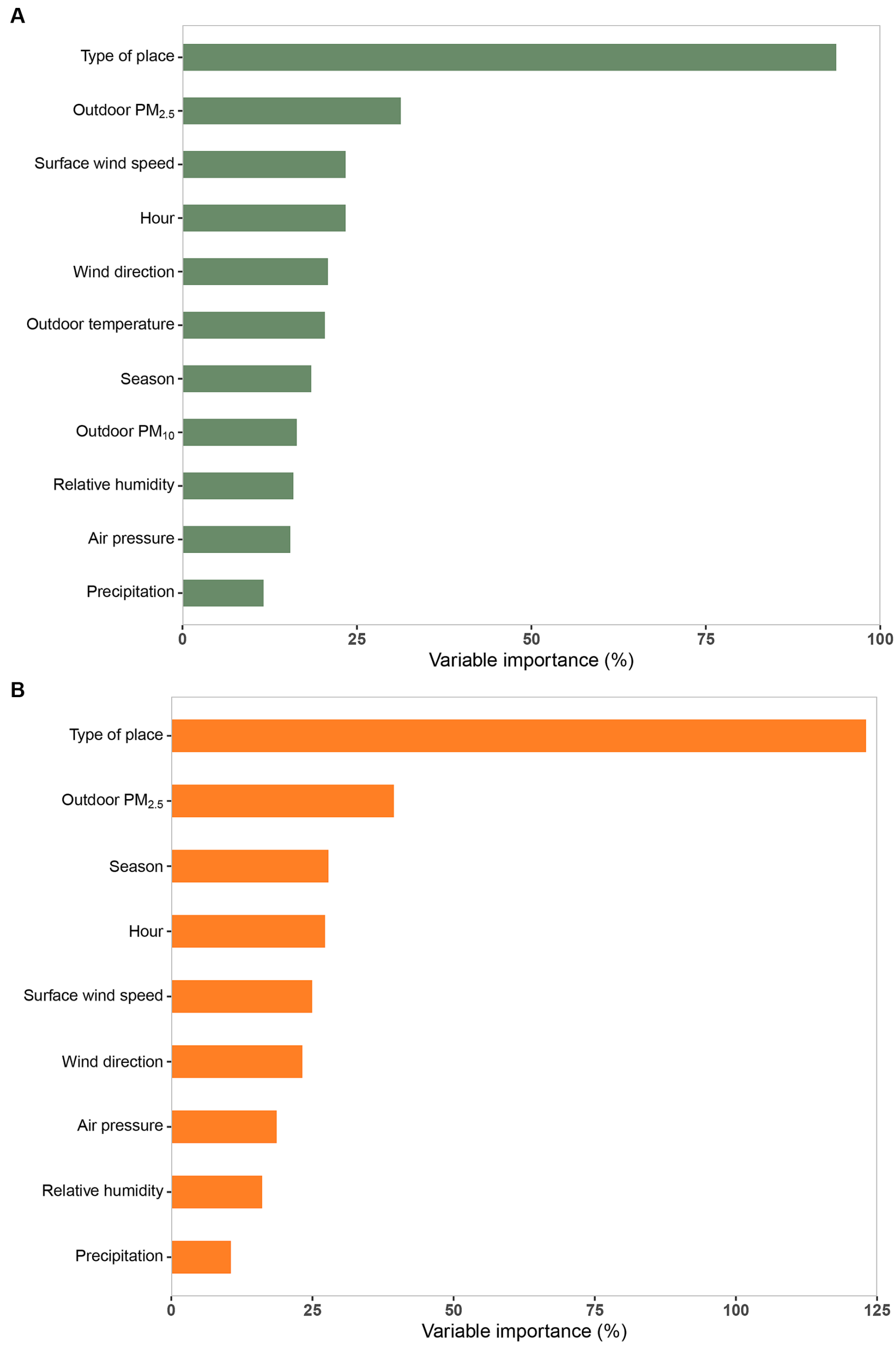

FIGURE 4
The importance of the predictor variables in random forest regression (RFR) models based on "%IncMSE." Full-RFR model **(A)**, Conjoint-RFR model **(B)**. $PM_{2.5}$ and PM_{10} refer to particulate matter with an aerodynamic diameter of 2.5 µm or less and of 10 µm.

The two CV R^2 values in the study by Xu et al. (13) indicated better model predictive performance than for our MLR model. This difference might be because the model development in our paper was based entirely on easily accessible temporal indicators and outdoor indicators. By contrast, the model construction in the study by Xu et al. (13) incorporated not only outdoor indicators (such as outdoor $PM_{2.5}$ concentration and outdoor relative humidity) but also indoor indicators (such as indoor smoking and cooking), with a wide range of indicator coverage. However, the model in that study also suffered from difficulties in the definition of relevant indicators, such as "whether or not to cook." In fact, cooking ingredients, cooking methods, cooking time, and the type of oil used have significant effects on indoor $PM_{2.5}$ concentration (55, 56). Moreover, these types of prediction indicators were not easy to obtain and the process was costly. Only several studies have developed RFR prediction models for indoor $PM_{2.5}$ concentration, and the CV R^2 values have ranged from 48.9 to 82% in these studies (13, 16, 18). The predictive efficacy of the Full-RFR model in this study (CV $R^2 = 72.20\%$) was also at a high level.

MLR and RFR models, as common indoor $PM_{2.5}$ concentration prediction models, are still controversial in terms of which approach can better predict indoor $PM_{2.5}$ concentrations. Previous studies have shown (16, 44) that using the same dataset, an RFR model usually outperforms an MLR model in terms of predictive efficacy owing to the strength of the algorithm itself, such as robustness to missing data and good characterization of interactions between different predictor variables. However, some studies have reached the opposite conclusion, as in the study by Yuchi et al. (18). In their study, two models had the same variables for the same dataset, and the MLR model (CV $R^2 = 50.2\%$) outperformed the RFR model (CV $R^2 = 48.9\%$) in terms of generalization performance. This issue was also explored in the current study, as the results of our sensitivity analysis for the modeling algorithm showed that the Full-RFR model, which used all predictor variables, and the Conjoint-RFR model, which used the same predictor variables as MLR, both performed better than the MLR model.

Compared with other studies, the current study had several strengths. First, the indoor $PM_{2.5}$ concentration monitoring data based on multiple types of places were used for modeling, which was more generalizable for predicting indoor $PM_{2.5}$ concentration than the models developed using data from a single type of place. Second, we developed modeling with high temporal resolution indoor $PM_{2.5}$ concentration data (hourly average data), which fully took into account the temporal variability of indoor $PM_{2.5}$. Third, the sample size used for modeling was sufficiently large ($n = 11{,}712$) to greatly exceed the number of predictor variables (11), so that the model was less prone to overfitting. Fourth, the model prediction cost was low, and the predictor variables in the model were all easy to obtain. For example, outdoor $PM_{2.5}$ concentration, wind direction, and surface wind can be found through the websites of relevant government departments. The model is suitable for epidemiological studies with large populations and/or long time periods.

Of course, there were some limitations in the study. First, the outdoor $PM_{2.5}$ concentration data of indoor places in the study were obtained from the nearest government-controlled monitoring sites. Although this approach has been used in many previous studies, it could introduce some errors in the model due to the spatial variability of outdoor $PM_{2.5}$ concentrations. Second, the absence of human indoor activity variables, such as smoking and cooking, might cause an increase in the prediction error of the model at certain time periods and contexts, for instance, during cooking and when air purifiers were used. Third, the model was developed and evaluated based on data from Shanghai, and there was a lack of equivalent data from other regions for further validation of model performance.

5. Conclusion

We found significant differences in indoor $PM_{2.5}$ concentration between types of places and time periods. This finding reflects the possible limitations of models based on indoor $PM_{2.5}$ concentration data from a single type of place as well as the necessity for a prediction model with a high temporal resolution in order to perfect individual $PM_{2.5}$ exposure assessment. Here, we aimed to develop MLR and RFR models for hourly average indoor $PM_{2.5}$ concentration over multiple types of places. Both statistical models were based on easy-to-access indicators and showed good predictive efficacy. They could, therefore, be used for quantitative estimation of indoor $PM_{2.5}$ exposure in large-scale population studies. In addition, the performance of the classical MLR model and machine learning RFR model were evaluated comparatively in predicting indoor $PM_{2.5}$ concentration, and the model performance metrics showed that the RFR model using the same dataset outperformed the MLR model. This finding suggests the potential of RFR models in predicting indoor air pollutant levels, and other machine learning algorithms may also be worthy of exploration.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

YS, ZD, CD, and SS designed the research. YS, FH, FC, DW, ML, and HZ performed data acquisition, organization and a part of analysis. ZD and YS analyzed the data and wrote the primary manuscript. YS, ZD, CD, SS, FH, and FC provided a contribution to the explanation of the findings and critically reviewed and edited the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by Shanghai Municipal Health Commission Science and Research Fund (No. 202040185).

Acknowledgments

The authors would like to thank all the personnel at the monitoring sites for their cooperation.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2023.1213453/full#supplementary-material

References

1. Cohen AJ, Brauer M, Burnett R, Anderson HR, Frostad J, Estep K, et al. Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the global burden of diseases study 2015. *Lancet Lond Engl*. (2017) 389:1907–18. doi: 10.1016/S0140-6736(17)30505-6

2. Ritz B, Hoffmann B, Peters A. The effects of fine dust, ozone, and nitrogen dioxide on health. *Dtsch Ärztebl Int*. (2019) 51-52:881–6. doi: 10.3238/arztebl.2019.0881

3. Yang L, Li C, Tang X. The impact of PM2.5 on the host Defense of respiratory system. *Front Cell Dev Biol*. (2020) 8:91. doi: 10.3389/fcell.2020.00091

4. Yang H, Li S, Sun L, Zhang X, Cao Z, Xu C, et al. Smog and risk of overall and type-specific cardiovascular diseases: a pooled analysis of 53 cohort studies with 21.09 million participants. *Environ Res*. (2019) 172:375–83. doi: 10.1016/j.envres.2019.01.040

5. Kaufman JD, Adar SD, Barr RG, Budoff M, Burke GL, Curl CL, et al. Association between air pollution and coronary artery calcification within six metropolitan areas in the USA (the multi-ethnic study of atherosclerosis and air pollution): a longitudinal cohort study. *Lancet Lond Engl*. (2016) 388:696–704. doi: 10.1016/S0140-6736(16)00378-0

6. Pinault L, Tjepkema M, Crouse DL, Weichenthal S, van Donkelaar A, Martin RV, et al. Risk estimates of mortality attributed to low concentrations of ambient fine particulate matter in the Canadian community health survey cohort. *Environ Health Glob Access Sci Source*. (2016) 15:18. doi: 10.1186/s12940-016-0111-6

7. Li M-H, Fan L-C, Mao B, Yang J-W, Choi AMK, Cao W-J, et al. Short-term exposure to ambient fine particulate matter increases hospitalizations and mortality in COPD: a systematic review and meta-analysis. *Chest*. (2016) 149:447–58. doi: 10.1378/chest.15-0513

8. Fan J, Li S, Fan C, Bai Z, Yang K. The impact of PM2.5 on asthma emergency department visits: a systematic review and meta-analysis. *Environ Sci Pollut Res*. (2016) 23:843–50. doi: 10.1007/s11356-015-5321-x

9. Hayes RB, Chris L, Yilong Z, Kevin C, Yongzhao S, Reynolds HR, et al. PM2.5 air pollution and cause-specific cardiovascular disease mortality. *Int J Epidemiol*. (2020) 49:25–35. doi: 10.1093/ije/dyz114

10. Gauvin S, Reungoat P, Cassadou S, Déchenaux J, Momas I, Just J, et al. Contribution of indoor and outdoor environments to PM2.5 personal exposure of children—VESTA study. *Sci Total Environ*. (2002) 297:175–81. doi: 10.1016/S0048-9697(02)00136-5

11. Rivas I, Fussell JC, Kelly FJ, Querol X. Indoor sources of air pollutants. *Issues Environ Sci Technol*. (2019) 20:1–34. doi: 10.1039/9781788016179-00001

12. Xie W, Fan Y, Zhang X, Tian G, Si P. A mathematical model for predicting indoor PM2.5 concentration under different ventilation methods in residential buildings. *Build Serv Eng Res Technol*. (2020) 41:694–708. doi: 10.1177/0143624420905102

13. Xu C, Xu D, Liu Z, Li Y, Li N, Chartier R, et al. Estimating hourly average indoor PM2.5 using the random forest approach in two megacities, China. *Build Environ*. (2020) 180:107025. doi: 10.1016/j.buildenv.2020.107025

14. Tong X, Ho JMW, Li Z, Lui K-H, Kwok TCY, Tsoi KKF, et al. Prediction model for air particulate matter levels in the households of elderly individuals in Hong Kong. *Sci Total Environ*. (2020) 717:135323. doi: 10.1016/j.scitotenv.2019.135323

15. Elbayoumi M, Ramli NA, Md Yusof NFF, Yahaya ASB, Al Madhoun W, Ul-Saufie AZ. Multivariate methods for indoor PM10 and PM2.5 modelling in naturally ventilated schools buildings. *Atmos Environ*. (2014) 94:11–21. doi: 10.1016/j.atmosenv.2014.05.007

16. Li Z, Tong X, Ho JMW, Kwok TCY, Dong G, Ho K-F, et al. A practical framework for predicting residential indoor PM2.5 concentration using land-use regression and machine learning methods. *Chemosphere*. (2021) 265:129140. doi: 10.1016/j.chemosphere.2020.129140

17. Han Y, Qi M, Chen Y, Shen H, Liu J, Huang Y, et al. Influences of ambient air $PM_{2.5}$ concentration and meteorological condition on the indoor $PM_{2.5}$ concentrations in a residential apartment in Beijing using a new approach. *Environ Pollut*. (2015) 205:307–14. doi: 10.1016/j.envpol.2015.04.026

18. Yuchi W, Gombojav E, Boldbaatar B, Galsuren J, Enkhmaa S, Beejin B, et al. Evaluation of random forest regression and multiple linear regression for predicting indoor fine particulate matter concentrations in a highly polluted city. *Environ Pollut*. (2019) 245:746–53. doi: 10.1016/j.envpol.2018.11.034

19. Elbayoumi M, Ramli NA, Fitri Md Yusof NF. Development and comparison of regression models and feedforward backpropagation neural network models to predict seasonal indoor PM2.5–10 and PM2.5 concentrations in naturally ventilated schools. *Atmospheric Pollut Res*. (2015) 6:1013–23. doi: 10.1016/j.apr.2015.09.001

20. Che W, Frey HC, Li Z, Lao X, Lau AKH. Indoor exposure to ambient particles and its estimation using fixed site monitors. *Environ Sci Technol*. (2019) 53:808–19. doi: 10.1021/acs.est.8b04474

21. Streets DG, Fu JS, Jang CJ, Hao J, He K, Tang X, et al. Air quality during the 2008 Beijing Olympic games. *Atmos Environ*. (2007) 41:480–92. doi: 10.1016/j.atmosenv.2006.08.046

22. Zhou B, Shen H, Huang Y, Li W, Chen H, Zhang Y, et al. Daily variations of size-segregated ambient particulate matter in Beijing. *Environ Pollut*. (2015) 197:36–42. doi: 10.1016/j.envpol.2014.11.029

23. Huang L, Pu Z, Li M, Sundell J. Characterizing the indoor-outdoor relationship of fine particulate matter in non-heating season for urban residences in Beijing. *PLoS One*. (2015) 10:e0138559. doi: 10.1371/journal.pone.0138559

24. Guo H, Li W, Wu J. Ambient PM2.5 and annual lung cancer incidence: a Nationwide study in 295 Chinese counties. *Int J Environ Res Public Health*. (2020) 17:E1481. doi: 10.3390/ijerph17051481

25. Robinson ES, Shah RU, Messier K, Gu P, Li HZ, Apte JS, et al. Land-use regression Modeling of source-resolved fine particulate matter components from Mobile sampling. *Environ Sci Technol*. (2019) 53:8925–37. doi: 10.1021/acs.est.9b01897

26. Li Z, Ho KF, Chuang HC, Yim SHL. Development and intercity transferability of land-use regression models for predicting ambient PM_{10}, $PM_{2.5}$, NO_2 and O_3 concentrations in northern Taiwan. *Copernic GmbH*. (2021) 21:5063–78. doi: 10.5194/acp-21-5063-2021

27. Ebrahimy H, Mirbagheri B, Matkan AA, Azadbakht M. Per-pixel land cover accuracy prediction: a random forest-based method with limited reference sample data. *ISPRS J Photogramm Remote Sens*. (2021) 172:17–27. doi: 10.1016/j.isprsjprs.2020.11.024

28. Li Y, Du Y, Deng Y, Fan R, Tao Y, Ma T, et al. Predicting the spatial distribution of phosphorus concentration in quaternary sedimentary aquifers using simple field parameters. *Appl Geochem*. (2022) 142:105349. doi: 10.1016/j.apgeochem.2022.105349

29. Leo B. Random forests. *Mach Learn*. (2001) 45:5–32. doi: 10.1023/A:1010933404324

30. Houwelingen H, Sauerbrei W. Cross-validation, shrinkage and variable selection in linear regression revisited. *Open J Stat*. (2013) 03:79–102. doi: 10.4236/ojs.2013.32011

31. Sebastian A, Madziarski M, Madej M, Proc K, Szymala-Pędzik M, Żórawska J, et al. The usefulness of the COVID-GRAM score in predicting the outcomes of study population with COVID-19. *Int J Environ Res Public Health*. (2022) 19:12537. doi: 10.3390/ijerph191912537

32. Li T, Cao S, Fan D, Zhang Y, Wang B, Zhao X, et al. Household concentrations and personal exposure of PM2.5 among urban residents using different cooking fuels. *Sci Total Environ*. (2016) 548-549:6–12. doi: 10.1016/j.scitotenv.2016.01.038

33. Grömping U. Relative importance for linear regression in *R*: the package relaimpo. *J Stat Softw*. (2006) 17:1–27. doi: 10.18637/jss.v017.i01

34. Kruskal W. Relative importance by averaging over orderings. *Am Stat*. (1987) 41:6–10. doi: 10.2307/2684310

35. Patel S, Sankhyan S, Boedicker EK, DeCarlo PF, Farmer DK, Goldstein AH, et al. Indoor particulate matter during HOMEChem: concentrations, size distributions, and exposures. *Environ Sci Technol*. (2020) 54:7107–16. doi: 10.1021/acs.est.0c00740

36. Bousiotis D, Alconcel L-NS, Beddows DCS, Harrison RM, Pope FD. Monitoring and apportioning sources of indoor air quality using low-cost particulate matter sensors. *Environ Int*. (2023) 174:107907. doi: 10.1016/j.envint.2023.107907

37. Szigeti T, Dunster C, Cattaneo A, Cavallo D, Spinazzè A, Saraga DE, et al. Oxidative potential and chemical composition of PM2.5 in office buildings across Europe – the OFFICAIR study. *Environ Int*. (2016) 92-93:324–33. doi: 10.1016/j.envint.2016.04.015

38. Matthaios VN, Kang C-M, Wolfson JM, Greco KF, Gaffin JM, Hauptman M, et al. Factors influencing classroom exposures to fine particles, black carbon, and nitrogen dioxide in Inner-City schools and their implications for indoor air quality. *Environ Health Perspect*. (2022) 130:47005. doi: 10.1289/EHP10007

39. Wallace L, Wang F, Howard-Reed C, Persily A. Contribution of gas and electric stoves to residential ultrafine particle concentrations between 2 and 64 nm: size distributions and emission and coagulation remission and coagulation rates. *Environ Sci Technol*. (2008) 42:8641–7. doi: 10.1021/es801402v

40. Dai X, Liu J, Li Y. A recurrent neural network using historical data to predict time series indoor PM2.5 concentrations for residential buildings. *Indoor Air*. (2021) 31:1228–37. doi: 10.1111/ina.12794

41. Yli-Tuomi T, Lanki T, Hoek G, Brunekreef B, Pekkanen J. Determination of the sources of indoor PM2.5 in Amsterdam and Helsinki. *Environ Sci Technol*. (2008) 42:4440–6. doi: 10.1021/es0716655

42. Carslaw N, Ashmore M, Terry AC, Carslaw DC. Crucial role for outdoor chemistry in ultrafine particle formation in modern office buildings. *Environ Sci Technol*. (2015) 49:11011–8. doi: 10.1021/acs.est.5b02241

43. Che W, Li ATY, Frey HC, Tang KTJ, Sun L, Wei P, et al. Factors affecting variability in gaseous and particle microenvironmental air pollutant concentrations in Hong Kong primary and secondary schools. *Indoor Air*. (2021) 31:170–87. doi: 10.1111/ina.12725

44. Zhao L, Chen C, Wang P, Chen Z, Cao S, Wang Q, et al. Influence of atmospheric fine particulate matter (PM2.5) pollution on indoor environment during winter in Beijing. *Build Environ*. (2015) 87:283–91. doi: 10.1016/j.buildenv.2015.02.008

45. Hadeed SJ, O'Rourke MK, Canales RA, Joshweseoma L, Sehongva G, Paukgana M, et al. Household and behavioral determinants of indoor PM2.5 in a rural solid fuel burning native American community. *Indoor Air*. (2021) 31:2008–19. doi: 10.1111/ina.12904

46. Omelekhina Y, Nordquist B, Alce G, Caltenco H, Wallenten P, Borell J, et al. Effect of energy renovation and occupants' activities on airborne particle concentrations in Swedish rental apartments. *Sci Total Environ*. (2022) 806:149995. doi: 10.1016/j.scitotenv.2021.149995

47. Zhou X, Cai J, Zhao Y, Chen R, Wang C, Zhao A, et al. Estimation of residential fine particulate matter infiltration in Shanghai, China. *Environ Pollut*. (2018) 233:494–500. doi: 10.1016/j.envpol.2017.10.054

48. Cutler DR, Edwards TC Jr, Beard KH, Cutler A, Hess KT, Gibson J, et al. Random forests for classification in ecology. *Ecology*. (2007) 88:2783–92. doi: 10.1890/07-0539.1

49. Genuer R, Poggi J-M, Tuleau-Malot C, Villa-Vialaneix N. Random Forests for Big Data. *Big Data Res*. (2017) 9:28–46. doi: 10.1016/j.bdr.2017.07.003

50. Iverson LR, Prasad AM, Matthews SN, Peters M. Estimating potential habitat for 134 eastern US tree species under six climate scenarios. *For Ecol Manag*. (2008) 254:390–406. doi: 10.1016/j.foreco.2007.07.023

51. Smith PF, Ganesh S, Liu P. A comparison of random forest regression and multiple linear regression for prediction in neuroscience. *J Neurosci Methods*. (2013) 220:85–91. doi: 10.1016/j.jneumeth.2013.08.024

52. Gaffin JM, Petty CR, Hauptman M, Kang C-M, Wolfson JM, Abu Awad Y, et al. Modeling indoor particulate exposures in inner-city school classrooms. *J Expo Sci Environ Epidemiol*. (2017) 27:451–7. doi: 10.1038/jes.2016.52

53. Raaschou-Nielsen O, Sørensen M, Hertel O, Chawes BLK, Vissing N, Bønnelykke K, et al. Predictors of indoor fine particulate matter in infants' bedrooms in Denmark. *Environ Res*. (2011) 111:87–93. doi: 10.1016/j.envres.2010.10.007

54. Jafta N, Barregard L, Jeena PM, Naidoo RN. Indoor air quality of low and middle income urban households in Durban. *South Africa Environ Res*. (2017) 156:47–56. doi: 10.1016/j.envres.2017.03.008

55. Gao X, Zhang M, Zou H, Zhou Z, Yuan W, Quan C, et al. Characteristics and risk assessment of occupational exposure to ultrafine particles generated from cooking in the Chinese restaurant. *Sci Rep*. (2021) 11:15586. doi: 10.1038/s41598-021-95038-y

56. Chen C, Zhao Y, Zhao B. Emission rates of multiple air pollutants generated from Chinese residential cooking. *Environ Sci Technol*. (2018) 52:1081–7. doi: 10.1021/acs.est.7b05600