Hindawi
Indoor Air
Volume 2024, Article ID 5589891, 11 pages
https://doi.org/10.1155/2024/5589891

Research Article

Predicting Personal Exposure to $PM_{2.5}$ Using Different Determinants and Machine Learning Algorithms in Two Megacities, China

Na Li,[1] **Yunpu Li**,[1] **Dongqun Xu**,[1] **Zhe Liu**,[1] **Ning Li**,[2] **Ryan Chartier**,[3] **Junrui Chang**,[1] **Qin Wang**,[1] **and Chunyu Xu**[1]

[1]*China CDC Key Laboratory of Environment and Population Health, National Institute of Environmental Health, Chinese Center for Disease Control and Prevention, Beijing 100021, China*
[2]*Nanjing Jiangning Center for Disease Control and Prevention, Nanjing 211100, China*
[3]*RTI International, Research Triangle Park, NC 27709, USA*

Correspondence should be addressed to Chunyu Xu; xuchunyu@nieh.chinacdc.cn

Received 11 January 2023; Revised 30 January 2024; Accepted 23 February 2024; Published 8 March 2024

Academic Editor: Giovanni Pernigotto

The primary aim of this study is to explore the utility of machine learning algorithms for predicting personal $PM_{2.5}$ exposures of elderly participants and to evaluate the effect of individual variables on model performance. Personal $PM_{2.5}$ was measured on five consecutive days across seasons in 66 retired adults in Beijing (BJ) and Nanjing (NJ), China. The potential predictors were extracted from routine monitoring data (ambient $PM_{2.5}$ concentrations and meteorological factors), basic questionnaires (personal and household characteristics), and time-activity diary (TAD). Prediction models were developed based on either traditional multiple linear regression (MLR) or five advanced machine learning methods. Our results revealed that personal $PM_{2.5}$ exposures were well predicted by both MLR and machine learning models with predictors extracted from routine monitoring data, which was indicated by the high nested cross-validation (CV) R^2 ranging from 0.76 to 0.88. The addition of predictors from either the questionnaire or TAD did not improve predictive accuracy for all algorithms. The ambient $PM_{2.5}$ concentrations were the most important predictor. Overall, the random forest, support vector machine, and extreme gradient boosting algorithms outperformed the reference MLR method. Compared with the traditional MLR approach, the CV R^2 of the RF model increased up to 7% (from 0.82 ± 0.13 to 0.88 ± 0.10), while the RMSE reduced up to 18% (from 19.8 ± 5.4 to 16.3 ± 4.5) in BJ.

1. Introduction

Accurate assessment of personal exposures to fine particulate matter ($PM_{2.5}$) is essential to study its health effects and provide risk assessments. Direct measurement of personal exposure to $PM_{2.5}$ via wearable monitors is currently regarded the most accurate exposure assessment method [1, 2]. However, the collection of personal exposure data is too logistically complicated and expensive for most budget-constrained large-scale population. Instead, the outdoor concentrations from nearby fixed-site monitors are used as a proxy for exposure in many epidemiological studies [3–5]. This approximation method leads to exposure misclassification as people usually spend greater than 80% of their time indoors [6, 7], and indoor air quality can vary substantially from outdoor environments. This variation is often driven by building ventilation rates and proximity to indoor sources of pollution such as cooking, heating, cleaning activities, tobacco smoking, and other domestic combustion sources [8, 9].

To overcome this significant limitation, investigators have tried to develop personal exposure models accounting for potential influential factors. Personal exposure surveys have shown that measured $PM_{2.5}$ concentrations can be correlated with influencing factors using statistical models that can subsequently be applied to estimate personal exposures

of new subjects [10]. The statistical algorithm used in model development is one of the crucial factors influencing the overall predictive power of the model. Multiple linear regression (MLR) has been the most commonly used method for model development because of its lower computational cost and ease of interpretability of the results [11, 12]. However, MLR models also have disadvantages such as the inability to capture complex and nonlinear interactions. Increased computing power has enabled the development of advanced machine learning algorithms to overcome some of the shortcomings of MLR models. To date, there have been hundreds of machine learning algorithms described in the literature, such as tree-based algorithms, artificial neural network (ANN) algorithms, kernel-based algorithms, and Bayesian method [13]. Recently, machine learning algorithms have been used to accurately predict the concentrations of atmospheric pollutants, and the performance of these algorithms was generally better than the MLR method [14–19]. However, to the best of our knowledge, the application of machine learning algorithms to estimate personal exposure is still in the early stages [11, 20–23]. The application of this approach in urban areas with a higher burden of ambient $PM_{2.5}$ pollution remains understudied [20].

Significant predictors of personal $PM_{2.5}$ exposures have been reported to be outdoor and indoor environmental concentrations, meteorological factors, personal and household characteristics, and human activities such as cooking, heating, smoking, and air conditioner and air purifier use [12, 23–27]. However, the relative importance of these predictors varied across investigations of different population groups, regions (rural vs. urban), and atmospheric air pollution conditions. In addition to selection of modeling algorithms, feature selection is another key process that can significantly influence model prediction performance. Exclusion of the effective determinants of personal exposure will reduce predictive accuracy, while inclusion of redundant and irrelevant variables may lead to overfitting and decrease the generalizability of the model [28–30]. In addition, removing noisy features will decrease the effort associated with collecting information for these variables when the model is applied. Several methods of feature selection are available for MLR algorithms, such as best subset selection and backward and forward stepwise selection. Statisticians have also developed feature selection methods suitable for machine learning algorithms, such as recursive feature elimination (RFE), genetic algorithms, and simulated annealing [31, 32]. However, these methods have not been used to develop models for estimating personal $PM_{2.5}$ exposures [11, 20–23].

The elderly is one of the most susceptible groups to air pollution exposure, due to generally weaker immune systems, or undiagnosed respiratory or cardiovascular health conditions [33–35]. However, most exposure studies conducted with elderly participants have been carried out in developed countries with relatively low ambient pollution levels. Unfortunately, the results of these studies cannot be directly extrapolated to elderly populations that suffer from exposures to high levels of $PM_{2.5}$ pollution in Chinese cities. To better characterize the exposure characteristics of this population, we conducted a repeated measurement study of outdoor-indoor-personal exposure in Beijing (BJ) and Nanjing (NJ) during 2015 and 2016. Our previous analyses showed that measured personal exposure concentrations were significantly lower than concentrations measured outdoors, confirming that using nearby outdoor $PM_{2.5}$ measurements as a direct proxy for personal exposure would inaccurately represent true exposures [12]. Therefore, a validated personal exposure prediction model should be developed, tested, and used to further investigate exposure-health effect relationships in at-risk populations. The primary aims of this analysis include the following two aspects: (1) to explore whether the use of machine learning algorithms can improve the accuracy of exposure prediction models and (2) to identify the key variables needed for accurate $PM_{2.5}$ prediction of elderly exposures in urban areas with high background pollution levels.

2. Methods

2.1. Study Design and Subjects. A detailed description of this $PM_{2.5}$ exposure longitudinal panel study of the elderly has been reported previously [12]. Briefly, this study was conducted in urban districts of BJ and NJ during both the heating season (HS; Nov.–Mar.) and the nonheating season (NHS; Jun.–Sep.) in 2015–2016. BJ is located in the northern region of China, while NJ is in the southern region, leading to distinct climate types (BJ: temperature monsoon climate, NJ: subtropical monsoon climate). These climate differences result in the use of different heating methods in winter (BJ: centralized heating, NJ: no centralized heating) and behavioral patterns, including window opening behavior and air conditioning usage, all of which may influence personal exposure. Outdoor-indoor-personal $PM_{2.5}$ levels were measured simultaneously for five consecutive days in each season. The sampling periods covered both weekdays and weekends as the participants generally exhibited distinct activity patterns during these days [36, 37]. Previous studies have also used this sampling strategy of monitoring exposure for 3–7 consecutive days [38–42]. Household characteristics and personal activity factors affecting exposure levels were also collected during this time period. In each city, thirty-three healthy, nonsmoking retired adults were recruited through leaflets placed in residential communities. In BJ, 31 and 30 participants were monitored during the HS and the NHS, respectively, with 85% (28/33) of the participants completing the monitoring in both seasons. Similarly, 31 participants in NJ were monitored during each season, with 88% (29/33) taking part in both seasons. The study was approved by the Human Investigation Committee of National Institute of Environmental Health, China CDC, and all participants signed informed consent.

2.2. Measurement of $PM_{2.5}$. The personal-indoor-outdoor exposure to $PM_{2.5}$ was simultaneously measured with RTI MicroPEM (v3.2, RTI International, NC, USA) for five consecutive days including weekends and weekdays during both heating and nonheating season. The MicroPEMs allow for gravimetric (filter-based) sampling while simultaneously logging real-time data via nephelometry. The MicroPEMs

were operated at a nominal flow rate of 0.5 L/min and were programmed to sample using a 25% duty cycle (1 min on and 3 min off for every 4 min cycle) to prolong battery life and prevent filter overloading. The MicroPEMs measuring personal exposure were worn in a shoulder bag, and the sampling inlet of the MicroPEM was extended into the breathing zone with a 0.3 m length of conductive silicone tubing. Participants were instructed to carry the shoulder bag with them at all times with the exception of sleeping, dressing, bathing, or performing other activities that did not allow for the bag to be carried. During these time periods, they were asked to place the bag nearby (<2 m). Indoor monitors were located in the household area in which the participant reported spending most of their waking hours. The outdoor MicroPEM was placed near a window in the residence, and the sample inlet was extended approximately 0.5 m out of the window using conductive silicone tubing. To minimize the influence of indoor air flow on the measurement of outdoor $PM_{2.5}$, any openings around the window used for outdoor monitoring were sealed with adhesive tape. All monitors were installed in the participant's residence by trained technicians, and monitoring typically started between 8 and 10 a.m. and ended at approximately the same time following the 5-day sample period.

Teflon sample filters were equilibrated in a chamber (Binder, Germany) with constant environmental conditions ($25 \pm 1^{\circ}C$, $50 \pm 5\%$ RH) for a minimum of 24 hours (CN HJ 656-2013) and then weighed using a microbalance with 1 μg precision (XP6, Mettler Toledo International Inc., Switzerland) before and after sampling. Each filter (25 mm, 3.0 μm porosity polytetrafluoroethylene with support ring, Pall Corporation, Mexico) was sampled for five days, and the five-day integrated $PM_{2.5}$ mass concentration was calculated by dividing the $PM_{2.5}$ mass collected on the filter (μg) by the corresponding air sample volume (m^3). These filter concentrations were then used to post-correct and calibrate the corresponding real-time concentrations for each individual sample using the following equation.

$$C = C_0 \times \frac{C_{\text{gav}}}{C_{\text{nep}}}, \tag{1}$$

where C is the corrected real-time $PM_{2.5}$ concentration, C_0 is the raw real-time concentration from the nephelometer, C_{gav} is the five-day weighted mass concentrations measured by the gravimetric method, and C_{nep} is the concurrent five-day mean concentration calculated using the raw real-time nephelometer data. The 24 h time-weighted $PM_{2.5}$ concentrations were calculated using these calibrated real-time data.

2.3. Ambient Air Quality and Meteorological Data. Ambient $PM_{2.5}$ data were retrieved from the China National Environmental Monitoring Center Network, which provides hourly $PM_{2.5}$ concentrations from local air quality monitoring stations (AQMS). The straight-line distance between participant's address and local AQMS was calculated. Data from the closest AQMS to each participant's address was used to produce 24 h time-weighted $PM_{2.5}$ concentrations corresponding to the sampling periods for personal exposure. In addition, meteorological data (temperature, relative humidity, atmospheric pressure, and wind speed) was also obtained from government-run monitoring sites in BJ and NJ.

2.4. Questionnaire and Time-Activity Diary (TAD). Prior to deployment of the sampling equipment, a standardized questionnaire was used to gather subjects' demographics (e.g., gender, age, and household income), home description (e.g., floors, room volume, building age, number of inhabitants, pet ownership, and primary cooking fuel), and lifestyle (e.g., window opening, cooking and cleaning frequency, and air conditioner and air purifier use), which potentially affect personal $PM_{2.5}$ exposures. The participants were also instructed to complete a daily TAD during sampling periods. Time-location information, as well as certain activities of pollutant-generating (i.e., cooking, cleaning, and environmental tobacco smoke (ETS) exposure), was recorded on the standardized time-based diaries.

A global position system (GPS) data logger (model BT-Q1000XT, Qstarz International, Taiwan, China) was carried by each participant to collect timestamped data on position (latitude, longitude) every 10 s. The recorded GPS track was displayed in Google Maps to verify the trips manually recorded in the TADs. When any inconsistencies between TAD recordings and GPS data were identified, the individual participants were contacted immediately for information confirmation. If the inconsistencies could not be clarified with the participant, the more objective GPS data were used for microenvironment identification. Finally, potential predictors of exposure levels and patterns were extracted from the manually inspected pooled GPS-TAD data.

2.5. Quality Assurance and Quality Control. The nephelometer baseline and nominal flow rate of MicroPEMs were calibrated before sampling and measured again at the conclusion of sampling. Filters were weighed in duplicate, and the values were averaged to obtain the final weight. The duplicate weights are needed to be within 4.0 μg of each other; otherwise, the filter was reweighed. Field blanks were collected at a rate of 10% of the samples. The method detection limit (MDL) for gravimetric method was estimated as three times the standard deviation (SD) of the field blanks divided by the nominal sample volume, and all the masses of samples greatly exceeded the MDL (4.3 $\mu g/m^3$). Field duplicate samples were collected for 6% of the samples. The difference between the time-weighted average $PM_{2.5}$ concentrations of duplicate samples was within 10% or 5 $\mu g/m^3$ in all cases. During HS, some real-time personal exposure data was lost due to an unknown source of instrument failure likely due to large temperature swings and the potential for condensation within the MicroPEM. This was more frequently an issue with BJ, which has colder outdoor temperatures during HS (BJ: -8.5°C to -7.7°C, NJ: 4.0°C to 10.4°C). Additionally, some samples were stopped early on request from the participant and scheduling considerations. Therefore, the calculated daily exposure from the calibrated real-time measurements was considered valid only if the

sample contained more than 22 h of valid data within a 24 h period. In total, 89% (271/305) and 96% (297/310) of the daily data were included in this analysis for BJ and NJ, respectively.

2.6. Statistical Analysis. Five state-of-the-art machine learning algorithms were tested to identify the most effective algorithm at predicting personal $PM_{2.5}$ exposure. The selected algorithms included commonly used algorithms with different underlying principles that have been shown to have good predictive ability for estimating outdoor or indoor air quality [10, 13, 14]. These algorithms included ANN with a single hidden layer, random forest (RF), support vector machine with Gaussian kernels (SVM), extreme gradient boosting (XGBoost), and gradient boosting machine (GBM). The MLR algorithm served as a reference method for comparison of the results. To meet the normality requirements of MLR, all 24 h $PM_{2.5}$ concentration data were natural log-transformed. Grid search optimization was used to tune the hyperparameters for each of the machine learning algorithms. To this end, we defined a wide range of variance for each of the hyperparameters (Table S1). The model performance for each combination of hyperparameters was evaluated using a cross-validation (CV) method, and the one with the best performance was selected for the final model.

All candidate predictors are listed in Table S2 and were divided into three categories according to the data source and difficulty of information acquisition: routine monitoring (including ambient concentrations and meteorological factors), basic questionnaire (including personal and household characteristics), and TAD (including time-location information and certain activities). Dummy coding, using the *dummyVars* function, was applied to handle the categorical variables as the machine learning algorithms are unable to process these variables. A series of prediction models were developed with different sets of potential predictors, beginning with those that are easiest to collect (routine monitoring) and followed by increasingly complex data (basic questionnaire and TAD). The improvement of model performance following the inclusion of additional more complex information was assessed by comparing between models.

The RFE method was applied for feature selection from each set of candidate predictors for the MLR and machine learning-based models. The RFE method is a search algorithm that treats the predictors as the inputs and uses model performance as the output to be optimized. Initially, the algorithm fits the model to all predictors. Each predictor is ranked using its importance to the model. Let S be a sequence of ordered numbers which are candidate values for the number of predictors to retain $(S_1 > S_2, \cdots)$. At each iteration of feature selection, the S_i top ranked predictors are retained, the model is refit, and the performance is assessed. The value of S_i with the best performance is determined, and the top S_i predictors are used to fit the final model [31]. The method was implemented by function RFE using the "caret" package in R software (version 3.5.1).

To better understand the relative influence of each predictor on model performance, variable importance (VI) scores and variable importance plots (VIPs) were constructed based on individual conditional expectation (ICE) curves [43–45]. This method identifies VI as the flatness of ICE curves in which the flatter curves represent the lower relative VI for the predictor of interest [44]. This analysis was performed by R software (version 3.5.1) with "vip" package.

A nested CV strategy was employed to evaluate the performance and generalization errors associated with the prediction models. This method overcomes the bias in performance evaluation caused by information leakage when the same data are used to tune model hyperparameters and evaluate model performance in non-nested CV [29]. The nested CV strategy contains an inner loop CV nested in an outer CV. The inner loop is responsible for hyperparameter tuning as mentioned above, while the outer loop is for error estimation [46]. For our analysis, 10% of samples were used for validation in the outer loop (10-fold CV), and 20% of samples were used for validation in the inner loop (5-fold CV). Measurements from the same participant were forced into the same group in each sampling procedure, and thus, artificial increases in the fitting degree related to repeated measurements of the same participant were eliminated. The coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE) between the measured and model predicted values were calculated and used for model comparison.

3. Results

3.1. Personal Characteristics. The median participant age was 62 and 59 in BJ and NJ, respectively. All participants were nonsmokers, but exposure to ETS was recorded for 12.9% (35/271) and 27.3% (81/297) person-days in BJ and NJ, respectively. All subjects lived in apartment, and natural ventilation was the only ventilation mode. Window opening was more prevalent in NJ than BJ due to differences in climate. Air purifiers were not frequently used and accounted for less than 3% (BJ: 8/271, NJ: 6/297) of monitoring person-days in both cities. Air conditioner usage time accounted for 23.2% (63/271) and 16.5% (49/297) in BJ and NJ, respectively.

According to time-activity data from pooled GPS-TADs, the participants spent more than 90% (median) of their time at home (BJ: 90.4%, NJ: 92.8%), followed by transportation (BJ: 3.1%, NJ: 1.9%), outdoors (BJ: 1.7%, NJ: 1.7%), and indoor public places (BJ: 0.9%, NJ: 1.1%). Other characteristics of subjects, their residences, and time-activity patterns that may influence personal exposure to $PM_{2.5}$ are shown in Table S3.

3.2. $PM_{2.5}$ Concentrations. Table 1 shows the summary statistics of ambient, outdoor, indoor, and personal $PM_{2.5}$ concentrations by city. Though large variations existed within each city, high levels of $PM_{2.5}$ pollution were observed in both cities. Overall, 95% of person-day measurements exceeded the World Health Organization (WHO) guideline

Table 1: The daily ambient, outdoor, indoor, and personal $PM_{2.5}$ measurement ($\mu g/m^3$) by city.

	N	Mean	Standard deviation	Min	P_{10}	P_{25}	P_{50}	P_{75}	P_{90}	Max
				Percentiles						
BJ										
Ambient $PM_{2.5}$	271	65.4	51.4	6.6	11.8	31.3	56.2	90.8	111.8	341.5
Outdoor $PM_{2.5}$	271	64.6	56.1	4.5	15.4	26.2	54.8	85.7	118.1	389.9
Indoor $PM_{2.5}$	271	49.8	36.3	3.6	14.4	23.5	40.6	65.3	93.1	256.0
Personal $PM_{2.5}$	271	53.5	39.5	4.2	15.1	26.2	45.2	68.5	99.7	285.0
NJ										
Ambient $PM_{2.5}$	297	59.7	31.8	16.4	28.3	34.2	52.3	74.2	105.8	156.6
Outdoor $PM_{2.5}$	297	71.9	44.3	16.0	27.7	37.3	58.1	95.5	137.2	210.4
Indoor $PM_{2.5}$	292[a]	62.7	37.9	16.3	27.4	35.8	52.7	79.3	108.3	232.0
Personal $PM_{2.5}$	297	65.7	40.2	16.4	27.8	36.0	51.8	82.0	123.2	218.9

Notes: N: the number of person-day; Min: the minimum value; Max: the maximum value; P_{10}, P_{25}, P_{50}, P_{75}, and P_{90}: the 10th, 25th, 50th, 75th, and 90th percentiles, respectively. [a]Five days of indoor $PM_{2.5}$ concentrations from a single subject was missing due to instrument failure in nonheating season.

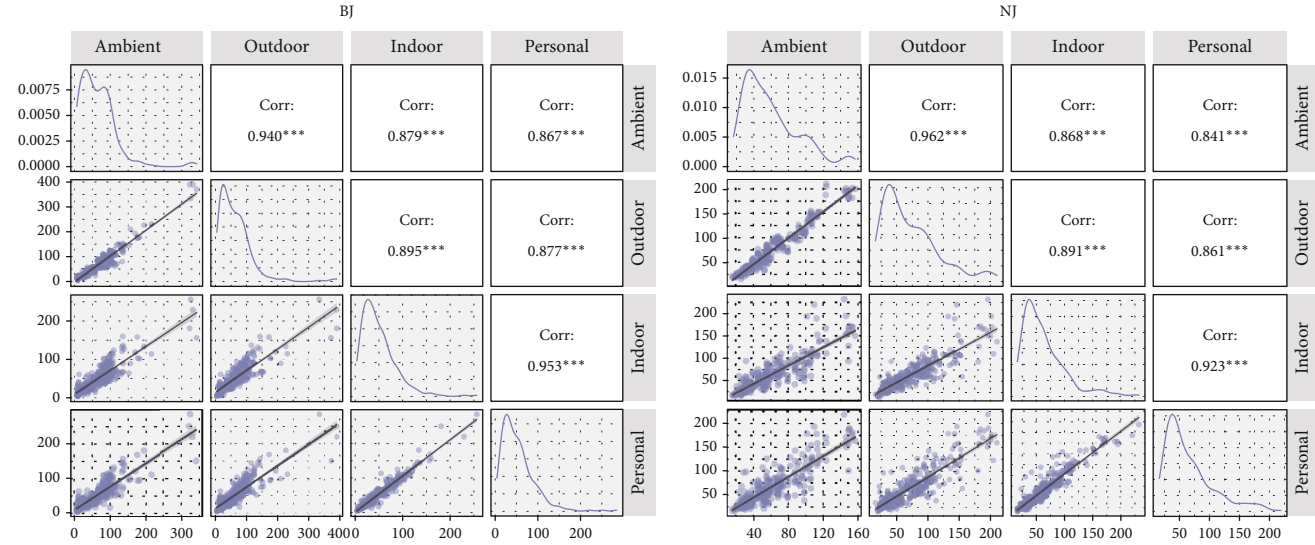

Figure 1: The correlation matrices of daily ambient, outdoor, indoor, and personal $PM_{2.5}$ measurements in BJ (Beijing) and NJ (Nanjing). The upper is the Spearman correlation coefficient, the lower is the scatter plot, and the diagonal is the density distribution plot. $^{***}p < 0.001$.

of $15\,\mu g/m^3$ (BJ: 90%, NJ: 100%). Regional differences in $PM_{2.5}$ exposures were found. The personal $PM_{2.5}$ concentrations in NJ were statistically significantly higher than BJ ($p < 0.001$), which was consistent with the indoor and outdoor $PM_{2.5}$ measurements.

Figure 1 illustrates the relationships among personal, indoor, outdoor, and ambient measurements. The residential outdoor $PM_{2.5}$ concentrations measured by MicroPEM were highly correlated with the ambient levels of the nearest AQMS, with the Spearman correlation coefficient of 0.94 and 0.96 in BJ and NJ, respectively. The personal $PM_{2.5}$ exposures were most related to indoor $PM_{2.5}$, followed by outdoor and ambient measurements.

3.3. Model Performance with Different Predictors and Algorithms. Table 2 shows the nested CV results for the prediction models based on different algorithms and candidate predictors. Overall, the prediction models performed better for the data collected in BJ than in NJ. Model 1 (including only ambient $PM_{2.5}$ and meteorological factors), based on either traditional MLR or machine learning algorithms, performed well with the CV R^2 ranging from 0.82 to 0.88 in BJ and from 0.76 to 0.80 in NJ. Model performance, including different candidate predictors, was then compared. However, the addition of variables from basic questionnaire (model 2) and TAD data (model 3) did not improve the model performance for all algorithms and in some instances slightly diminished model accuracy, possibly due to overfitting caused by redundant variables. For example, model 1 which is based on an RF algorithm, has a higher CV R^2 (0.88 ± 0.10) and lower RMSE ($16.3 \pm 4.5\,\mu g/m^3$) and MAE ($12.0 \pm 2.4\,\mu g/m^3$) than the corresponding model 3 (R^2: 0.85 ± 0.12, RMSE: $16.3 \pm 5.5\,\mu g/m^3$, MAE: $11.6 \pm 2.6\,\mu g/m^3$) in BJ.

Table 2: Nested CV results of prediction models with different algorithms and predictors.

City	Algorithm	Model 1			Model 2			Model 3		
		R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE
BJ	MLR	0.82 ± 0.13	19.8 ± 5.4	13.6 ± 2.6	0.80 ± 0.15	20.4 ± 6.0	13.9 ± 2.8	0.81 ± 0.10	19.7 ± 6.9	13.5 ± 3.3
	RF	0.88 ± 0.10[a]	16.3 ± 4.5[a]	12.0 ± 2.4	0.84 ± 0.11	18.3 ± 5.6	12.8 ± 3.8	0.85 ± 0.12[a]	16.3 ± 5.5	11.6 ± 2.6
	SVM	0.87 ± 0.11[a]	17.3 ± 5.8	12.4 ± 4.2	0.83 ± 0.16	19.3 ± 5.9	13.6 ± 4.3	0.83 ± 0.16	18.6 ± 5.9	13.4 ± 3.9
	XGBoost	0.84 ± 0.15	18.2 ± 5.5	12.8 ± 3.2	0.85 ± 0.12	18.0 ± 6.5	12.7 ± 4.8	0.84 ± 0.14	18.1 ± 6.3	13.0 ± 4.7
	GBM	0.82 ± 0.12	21.3 ± 8.1	14.4 ± 3.2	0.80 ± 0.18	23.1 ± 8.9	17.3 ± 6.3	0.85 ± 0.12[a]	29.0 ± 13.8	24.9 ± 12.2[a]
	ANN	0.84 ± 0.10	19.7 ± 7.5	14.5 ± 4.9	0.82 ± 0.13	21.6 ± 7.5	15.9 ± 4.7	0.84 ± 0.08	22.0 ± 11.8	15.2 ± 5.6
NJ	MLR	0.78 ± 0.17	21.1 ± 10.7	15.5 ± 8.4	0.78 ± 0.18	19.5 ± 5.1	13.5 ± 3.1	0.78 ± 0.18	20.1 ± 6.8	14.4 ± 4.9
	RF	0.79 ± 0.16	20.5 ± 9.7	14.4 ± 7.7	0.79 ± 0.17	21.1 ± 8.1	14.8 ± 6.3	0.79 ± 0.15	20.5 ± 8.4	15.0 ± 6.6
	SVM	0.80 ± 0.15	21.1 ± 10.1	14.6 ± 7.6	0.76 ± 0.20	21.8 ± 10.2	15.5 ± 7.8	0.78 ± 0.18	20.7 ± 10.8	14.9 ± 8.2
	XGBoost	0.79 ± 0.17	20.5 ± 9.0	14.7 ± 7.1	0.79 ± 0.16	22.4 ± 12.7	16.0 ± 9.7	0.78 ± 0.18	20.8 ± 8.9	15.5 ± 6.8
	GBM	0.76 ± 0.17[a]	23.0 ± 11.2	17.2 ± 9.4[a]	0.73 ± 0.22	25.9 ± 9.2[a]	20.6 ± 8.1[a]	0.73 ± 0.22	22.8 ± 6.7	17.0 ± 5.9
	ANN	0.78 ± 0.16	21.1 ± 9.8	15.9 ± 7.7	0.73 ± 0.18[a]	22.0 ± 4.8	16.5 ± 3.8[a]	0.71 ± 0.17[a]	25.4 ± 11.1[a]	18.2 ± 7.5[a]

Note: model 1 included variables for ambient $PM_{2.5}$ from the nearest AQMS and meteorological factors. Model 2 included both variables in model 1 and variables from basic questionnaire. Model 3 included both variables in model 2 and variables from TAD. The values were mean ± SD from nested cross-validation. [a]Statistically different from referenced MLR method.

Compared with a traditional MLR algorithm, the machine learning-based models performed similarly or slightly better as indicated by a higher R^2 and lower RMSE and MAE. These results also demonstrated that RF and SVM were the most effective algorithms tested. As shown in Table 2, the CV R^2 of RF model increased by 7% (from 0.82 ± 0.13 to 0.88 ± 0.10), while RMSE decreased by 18% (from 19.8 ± 5.4 to 16.3 ± 4.5) compared to the traditional MLR approach in BJ. In addition, the lower SD of model performance metrics suggested that the performance of the RF and SVM algorithms was more stable.

3.4. Variable Importance. Figure 2 and Table S4 illustrate the relative variable importance in predicting personal $PM_{2.5}$ exposure based on different algorithms (model 3). Across all algorithms and cities, the ambient $PM_{2.5}$ was consistently the most import predictor and its contribution was much larger than any other factors. However, the other variables included in final models were quite different between cities and algorithms. For example, outdoor relative humidity (RH) was the only variable included in all models in BJ, while it was less important in NJ, where exposure to ETS played a more important role than other variables except ambient $PM_{2.5}$.

4. Discussion

MLR models were used for reference purposes during our development of machine learning algorithms for the prediction of personal $PM_{2.5}$ exposures. The nested CV results indicate that our MLR models yielded accurate 24 h exposure estimates. This MLR approach has been used extensively for $PM_{2.5}$ exposure prediction in previous studies, but the majority of these studies have been carried out in urban areas of developed countries with low air pollution levels, such as North America and Europe [47–49]. Recently, more research studies have been carried out in rural areas of developing countries (e.g., Kenya, India, Lao PDR, and China) [11, 21, 23, 27, 50, 51]. The predictive ability of the models included in these studies varied greatly with CV R^2 values ranging from 0.09 to 0.76. Compared with the studies mentioned above, our MLR model displayed stronger prediction ability as indicated by the higher nested CV R^2 values (BJ: 0.82, NJ: 0.78). This result was mainly due to the following two reasons. First, the personal exposure levels of our subjects covered a much broader range (BJ: 4.2-285.0 $\mu g/m^3$, NJ: 16.4-218.9 $\mu g/m^3$) than that studied in the developed countries. Second, ambient $PM_{2.5}$ was the dominant exposure source for our subjects, which has been accurately monitored and included in our MLR models. Contrary to our study, strong indoor sources (e.g., solid fuel combustion, cooking fumes, and ETS) and local outdoor source (e.g., vehicle emissions) also contributed a considerable proportion of exposure for participants in studies conducted in urban areas of developed countries [47–49, 52] or rural areas of developing countries [11, 21, 23], and the influence of these sources on personal exposure was difficult to accurately estimate.

A primary aim of this analysis was to explore whether the utility of machine learning algorithms could improve the accuracy of $PM_{2.5}$ exposure prediction compared to MLR methods. Our analysis found that all of the five machine learning algorithms we tested could provide accurate prediction with an R^2 ranging from 0.76 to 0.88 (model 1). The RF and SVM algorithms generally performed better than our MLR models with the same candidate explanatory variables, especially in BJ. To our knowledge, only a few studies have applied machine learning algorithms to predict personal $PM_{2.5}$ exposure [11, 20–23]. Among these studies, RF was the most commonly used algorithm. For example, in the

FIGURE 2: Bar plots of relative variable importance for personal $PM_{2.5}$ exposure prediction based on calculating flatness of partial dependence plot curves in BJ (Beijing) and NJ (Nanjing). MLR: multiple linear regression; RF: random forest; SVM: support vector machine; GBM: gradient boosting machine; XGBoost: extreme gradient boosting; ANN: artificial neural network.

Relationships of Indoor, Outdoor, and Personal Air (RIOPA) study, MLR and RF were used to predict chemical elements in 48 h personal $PM_{2.5}$ samples. Consistent with our findings, RF analysis performed better than MLR for most elements [22]. In rural Lao PDR, the mean 48 h $PM_{2.5}$ exposure concentrations for female cooks were estimated using machine learning models. These models produced an observed vs. CV predicted R^2 between 0.26 and 0.31, and the best candidate learner was RF, followed by cForest [21]. This, along with our findings, suggests that RF is a promising technology for personal exposure estimation for its ability to uncover and harness complex variable interrelationships to produce more accurate predictions [21]. However, inconsistent results were reported in a study conducted in rural area of Kenya. In this study, all five tested five machine learning algorithms (including RF, XGBoost, SVM, Rpart, and Glmnet) performed worse than MLR. The poorer machine learning model performance in this study may be partly explained by the relatively small sample size (~50) and failure to adopt appropriate variable selection methods [23]. Unlike the analysis presented here, a variable selection method specific to machine learning algorithms was not adopted in the Kenya study, but the same variables as MLR model were included, potentially limiting the predictive ability of the machine learning algorithms. Therefore, a suitable variable selection method is essential to improve the predictive power of the models based on machine learning algorithms. In a recent study conducted in Tianjin, a heavily polluted city in northern China, a total of 117 older adults over 60 years of age were recruited and their $PM_{2.5}$ exposures measured. Four modeling techniques, including time-integrated activity modeling, Monte Carlo simulation, ANN modeling, and combined use of principal component analysis (PCA) and ANN model, were used to evaluate their ability to predict $PM_{2.5}$ exposures in this study setting. The authors found that the combined use of PCA and ANN produced the most accurate results, yielding an R^2 of 0.99 and RMSE lower than 15 $\mu g/m^3$, while the traditional time-weighted activity modeling showed the lowest correlation with measured values with R^2 of less than 0.6. The high accuracy of the model used in this study may be very likely attributed to the inclusion of measured indoor $PM_{2.5}$ levels as predictors [20]. However, the indoor $PM_{2.5}$ measures were not used in our study, since only ambient measures can be accessed easily. In addition, contrary to the results in the Tianjin study, the prediction accuracy of our ANN model was slightly lower than MLR and the preprocess

method of PCA did not improve the model fit of ANN or any other machine learning based model.

Our comparison among models developed with different candidate predictors showed that the inclusion of variables from the basic questionnaire, and even the participant's TAD, could not improve prediction accuracy. The variable importance evaluation results also confirmed the rationality of this result. Our result may be of great practical significance as it shows that we can obtain the same prediction model performance for the elderly without the added burden needed to gather those data. However, extrapolating the current results to other age groups requires caution. In our study, the majority of participants were over 60 years old, and almost all of their time was spent at home (~90%), with only a small percentage spent during transportation (~3%) or in public places (~3%). It is noteworthy that their time-activity patterns significantly differ from other subgroups, such as office workers and school-age children. Thus, factors associated with time-activity patterns, such as commuting status and exposure to indoor pollution sources in public places, might assume greater significance. A study by Rojas-Bracho et al. found that personal $PM_{2.5}$ exposures increased by 2.5 $\mu g/m^3$ for each hour spent in a motor vehicle [48]. Our $PM_{2.5}$ real-time concentration data indicates that personal exposure levels are higher than environmental background levels during cycling or walking, with a personal/outdoor ratio of approximately 1.1 [53]. Moreover, our findings highlight that individuals frequenting restaurants were exposed to elevated levels of $PM_{2.5}$, as evidenced by considerably higher ratios of personal to outdoor $PM_{2.5}$ (BJ: 1.48, NJ: 1.37) [53]. This is consistent with previous studies conducted in Seoul [54, 55]. Taken together, it is important to consider that differences in time-activity patterns may significantly influence personal exposure models for populations other than the elderly.

In previous studies, exposure to ETS was found to be another important factor affecting overall $PM_{2.5}$ exposure [47, 48]. However, the ETS contribution to the prediction model is not evident in this analysis. It was reported that exposure to ETS for 1 h would increase the 24 h mean concentration of $PM_{2.5}$ exposure by about 4 $\mu g/m^3$ [47, 48, 56]. In our study, only 3.6% and 6.7% of participants in BJ and NJ were exposed to ETS for more than 1 h a day, which means its impact on $PM_{2.5}$ exposure levels was far less than ambient air and may be masked by the variation of ambient $PM_{2.5}$. Cooking behavior can lead to a sharp increase in indoor $PM_{2.5}$ level in a short period of time, which is also another important contributor of $PM_{2.5}$ exposure especially in rural areas in previous studies [21, 23, 48]. Chang et al. reported that cooking for 1 h increased 24 h personal exposures to $PM_{2.5}$ by about 4 $\mu g/m^3$ [47, 48, 57]. However, it should be noted that the magnitude of impact cooking can have on overall exposure is also strongly affected by the type of cooking, fuel type, who is cooking (participant or other), ventilation status, and building structures [57]. This suggests that a simple variable such as cooking duration could not accurately characterize its contribution to exposure. The TAD results from our study show that the median (P_{25}, P_{75}) daily cooking duration in subject's homes was 1.5 (1.0, 1.9) h and 1.5 (0.9, 2.1) h in BJ and NJ, respectively. Unfortunately, our questionnaires only included a cooking question related to fuel type. Natural gas was the dominant cooking fuel in both BJ and NJ. This uncertainty reduces the prediction ability of the family cooking time variable on individual exposure levels. Lack of detailed information on cooking behavior and high levels of background $PM_{2.5}$ pollution have reduced the role of cooking behavior in predicting personal exposure in our study, and future studies should attempt to collect more detailed information on cooking activities and patterns to better understand the potentially important relationship between household cooking and residential exposures.

Window opening was regarded as a predictor related to an increase in indoor and personal concentrations in previous reports [58–60], since window opening has a strong influence on air exchange rate, as well as increasing penetration by permitting ambient air to enter the indoor environment. However, we did not find that the inclusion of relevant variables of window opening behavior (window opening time and window opening width) had a significant impact on the accuracy of our models. A potential reason for this can be attributed to meteorological factors (e.g., temperature and wind speed), which can indirectly capture the opening windows status to a certain extent. In fact, our data indicated that more than 50% of the total variation of window opening time can be predicted by variables of temperature, humidity, and wind speed in BJ.

To our knowledge, this is the first study to develop prediction models for personal $PM_{2.5}$ exposure using multiple machine learning approaches in urban locations with high levels of ambient $PM_{2.5}$ pollution. This study was conducted in two Chinese megacities with uniform study design and measurement methods, and the consistent results between cities indicate that our findings are robust. However, we also note that the models in BJ and NJ did not include the same predictors, which suggests the need to develop city-specific assessment models. There were several limitations of this study. First, our study was only conducted with retired adults residing in urban areas, and as such, caution should be applied when extrapolating our results to other age groups with different time-activity patterns and people living in rural areas who are exposed to different $PM_{2.5}$ sources. Second, the sample size is relatively small, which is not conducive to developing machine learning models, especially for neural network models with complex structures. However, even with a relatively small number of training samples, the RF and SVM algorithms show advantages over the traditional MLR algorithm. Therefore, the machine learning approach shows promise for predicting personal air pollution exposures.

5. Conclusions

Our nested CV results showed that the models containing only predictors from routine air quality and meteorological monitoring data can accurately predict the personal $PM_{2.5}$ exposures of the elderly adults residing in urban areas with elevated levels of air pollution. The addition of individual

and household characteristics as well as time-activity information had a limited effect of predictive ability. The comparison statistics between MLR and machine learning models for the same data set indicated that the latter algorithms have advantages over the classic MLR method even at limited training sample sizes. Our results suggest that the machine learning approach could be a promising technology for predicting personal air pollution concentrations.

Abbreviations

$PM_{2.5}$: Fine particulate matter
MLR: Multiple linear regression
ANN: Artificial neural network
RFE: Recursive feature elimination
BJ: Beijing
NJ: Nanjing
HS: Heating season
NHS: Nonheating season
AQMS: Air quality monitoring stations
TAD: Time-activity diary
ETS: Environmental tobacco smoke
GPS: Global position system
MDL: Method detection limit
SD: Standard deviation
RF: Random forest
SVM: Support vector machine
XGBoost: Extreme gradient boosting
GBM: Gradient boosting machine
CV: Cross-validation
VI: Variable importance
VIPs: Variable importance plots
ICE: Individual conditional expectation
RMSE: Root mean square error
MAE: Mean absolute error
RH: Relative humidity
PCA: Principal component analysis.

Data Availability

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Additional Points

Practical Implications. Reliable and accurate models for estimating personal exposure are valuable tools for researchers. The statistical algorithm used in model development and predictors selected are the key factors influencing model prediction performance. Our results suggest that the machine learning approach could be a promising technology for predicting personal air pollution concentrations. Furthermore, our findings may be of great practical significance as it shows that we can obtain the same prediction model performance for the elderly without the added burden needed to gather information from basic questionnaire and the participant's TAD.

Ethical Approval

The study was approved by the Human Investigation Committee of National Institute of Environmental Health, China CDC.

Consent

All participants signed informed consent.

Conflicts of Interest

The authors have no conflicts of interest to declare.

Acknowledgments

The authors thank all the participants in this study. We also acknowledge Jiangsu Provincial and Nanjing Jiangning Center for Disease Control and Prevention as well as RTI International. The work was supported by the Public Welfare Research Program of National Health and Family Planning Commission of China (201402022) and National Natural Science Foundation of China (21677136).

Supplementary Materials

The supplementary material contains four tables. Table S1: hyperparameters for tuning machine learning model. Table S2: the list of candidate predictors for 24 h average personal $PM_{2.5}$. Table S3: residence, demographic, and activity characteristics of study subjects. Table S4: importance scores of variables included in the final prediction models. *(Supplementary Materials)*

References

[1] M.-A. Kioumourtzoglou, D. Spiegelman, A. A. Szpiro et al., "Exposure measurement error in $PM_{2.5}$ health effects studies: a pooled analysis of eight personal exposure validation studies," *Environmental Health*, vol. 13, no. 1, p. 2, 2014.

[2] C. L. Avery, K. T. Mills, R. Williams et al., "Estimating error in using ambient $PM_{2.5}$ concentrations as proxies for personal exposures: a review," *Epidemiology (Cambridge, Mass.)*, vol. 21, no. 2, pp. 215–223, 2010.

[3] D. W. Dockery, C. A. Pope 3rd, X. Xu et al., "An association between air pollution and mortality in six U.S. cities," *New England Journal of Medicine*, vol. 329, no. 24, pp. 1753–1759, 1993.

[4] F. Dominici, A. McDermott, S. L. Zeger, and J. M. Samet, "National maps of the effects of particulate matter on mortality: exploring geographical variation," *Environmental Health Perspectives*, vol. 111, no. 1, pp. 39–44, 2003.

[5] C. A. Pope 3rd and D. W. Dockery, "Health effects of fine particulate air pollution: lines that connect," *Journal of the Air & Waste Management Association*, vol. 56, no. 6, pp. 709–742, 2006.

[6] N. E. Klepeis, W. C. Nelson, W. R. Ott et al., "The National Human Activity Pattern Survey (NHAPS): a resource for assessing exposure to environmental pollutants," *Journal of Exposure Science & Environmental Epidemiology*, vol. 11, no. 3, pp. 231–252, 2001.

[7] X. Duan, *Highlights of the Chinese Exposure Factors Handbook*, Academic Press, 2015.

[8] É. C. Pagel, N. Costa Reis, C. E. de Alvarez et al., "Characterization of the indoor particles and their sources in an Antarctic research station," *Environmental Monitoring and Assessment*, vol. 188, no. 3, p. 167, 2016.

[9] E. Abt, H. H. Suh, G. Allen, and P. Koutrakis, "Characterization of indoor particle sources: a study conducted in the metropolitan Boston area," *Environmental Health Perspectives*, vol. 108, no. 1, pp. 35–44, 2000.

[10] W. Wei, O. Ramalho, L. Malingre, S. Sivanantham, J. C. Little, and C. Mandin, "Machine learning and statistical models for predicting indoor air quality," *Indoor Air*, vol. 29, no. 5, pp. 704–726, 2019.

[11] M. Sanchez, C. Milà, V. Sreekanth et al., "Personal exposure to particulate matter in peri-urban India: predictors and association with ambient concentration at residence," *Journal of Exposure Science & Environmental Epidemiology*, vol. 30, no. 4, pp. 596–605, 2020.

[12] N. Li, C. Xu, Z. Liu et al., "Determinants of personal exposure to fine particulate matter in the retired adults - results of a panel study in two megacities, China," *Environmental Pollution*, vol. 265, no. Part B, article 114989, 2020.

[13] S. Ray, "A quick review of machine learning algorithms," in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, pp. 35–39, Faridabad, India, 2019.

[14] M. Ashayeri, N. Abbasabadi, M. Heidarinejad, and B. Stephens, "Predicting intraurban $PM_{2.5}$ concentrations using enhanced machine learning approaches and incorporating human activity patterns," *Environmental Research*, vol. 196, p. 110423, 2021.

[15] S. Ausati and J. Amanollahi, "Assessing the accuracy of ANFIS, EEMD-GRNN, PCR, and MLR models in predicting $PM_{2.5}$," *Atmospheric Environment*, vol. 142, pp. 465–474, 2016.

[16] B. Choubin, M. Abdolshahnejad, E. Moradi et al., "Spatial hazard assessment of the PM_{10} using machine learning models in Barcelona, Spain," *Science of the Total Environment*, vol. 701, article 134474, 2020.

[17] H. Karimian, Q. Li, C. Wu et al., "Evaluation of different machine learning approaches to forecasting $PM_{2.5}$ mass concentrations," *Aerosol and Air Quality Research*, vol. 19, no. 6, pp. 1400–1410, 2019.

[18] M. Niu, Y. Wang, S. Sun, and Y. Li, "A novel hybrid decomposition-and-ensemble model based on CEEMD and GWO for short-term $PM_{2.5}$ concentration forecasting," *Atmospheric Environment*, vol. 134, pp. 168–180, 2016.

[19] W. Qiao, W. Tian, Y. Tian, Q. Yang, Y. Wang, and J. Zhang, "The forecasting of $PM_{2.5}$ using a hybrid model based on wavelet transform and an improved deep learning algorithm," *IEEE Access*, vol. 7, pp. 142814–142825, 2019.

[20] S. Gao, H. Zhao, Z. Bai et al., "Combined use of principal component analysis and artificial neural network approach to improve estimates of $PM_{2.5}$ personal exposure: a case study on older adults," *Science of the Total Environment*, vol. 726, p. 138533, 2020.

[21] L. Hill, A. Pillarisetti, S. Delapena et al., "Machine-learned modeling of $PM_{2.5}$ exposures in rural Lao PDR," *Science of the Total Environment*, vol. 676, pp. 811–822, 2019.

[22] P. H. Ryan, C. Brokamp, Z.-H. Fan, and M. B. Rao, "Analysis of personal and home characteristics associated with the elemental composition of $PM_{2.5}$ in indoor, outdoor, and personal air in the RIOPA study," *Research Report (Health Effects Institute)*, vol. 185, pp. 3–40, 2015.

[23] M. Johnson, R. Piedrahita, A. Pillarisetti et al., "Modeling approaches and performance for estimating personal exposure to household air pollution: a case study in Kenya," *Indoor Air*, vol. 31, no. 5, pp. 1441–1457, 2021.

[24] M. Lee, E. Carter, L. Yan et al., "Determinants of personal exposure to $PM_{2.5}$ and black carbon in Chinese adults: a repeated-measures study in villages using solid fuel energy," *Environment International*, vol. 146, p. 106297, 2021.

[25] X.-C. Chen, J. C. Chow, T. J. Ward et al., "Estimation of personal exposure to fine particles ($PM_{2.5}$) of ambient origin for healthy adults in Hong Kong," *Science of the Total Environment*, vol. 654, pp. 514–524, 2019.

[26] X.-C. Chen, H. J. Jahn, T. J. Ward et al., "Characteristics and determinants of personal exposure to $PM_{2.5}$ mass and components in adult subjects in the megacity of Guangzhou, China," *Atmospheric Environment*, vol. 224, article 117295, 2020.

[27] C. Chen, J. Cai, C. Wang et al., "Estimation of personal $PM_{2.5}$ and BC exposure by a modeling approach–results of a panel study in Shanghai, China," *Environment International*, vol. 118, pp. 194–202, 2018.

[28] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: a new perspective," *Neurocomputing*, vol. 300, pp. 70–79, 2018.

[29] G. C. Cawley and N. L. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *The Journal of Machine Learning Research*, vol. 11, pp. 2079–2107, 2010.

[30] S.-y. Jiang and L.-x. Wang, "Efficient feature selection based on correlation measure between continuous and discrete features," *Information Processing Letters*, vol. 116, no. 2, pp. 203–215, 2016.

[31] M. Kuhn, "Variable selection using the caret package," 2012, http://cran.cermin.lipi.go.id/web/packages/caret/vignettes/caretSelection.pdf.

[32] Y. Nishihama, C.-R. Jung, S. F. Nakayama et al., "Indoor air quality of 5,000 households and its determinants. Part A: particulate matter ($PM_{2.5}$ and $PM_{10-2.5}$) concentrations in the Japan Environment and Children's Study," *Environmental Research*, vol. 198, article 111196, 2021.

[33] P. L. Ljungman and M. A. Mittleman, "Ambient air pollution and stroke," *Stroke*, vol. 45, no. 12, pp. 3734–3741, 2014.

[34] M. Brauer, G. Freedman, J. Frostad et al., "Ambient air pollution exposure estimation for the global burden of disease 2013," *Environmental Science & Technology*, vol. 50, no. 1, pp. 79–88, 2016.

[35] J. Song, Y. Gao, S. Hu et al., "Association of long-term exposure to $PM_{2.5}$ with hypertension prevalence and blood pressure in China: a cross-sectional study," *BMJ Open*, vol. 11, no. 12, article e050159, 2021.

[36] F. Shao, Y. Sui, X. Yu, and R. Sun, "Spatio-temporal travel patterns of elderly people – a comparative study based on buses usage in Qingdao, China," *Journal of Transport Geography*, vol. 76, pp. 178–190, 2019.

[37] Z. Shi, L. S. C. Pun-Cheng, X. Liu et al., "Analysis of the temporal characteristics of the elderly traveling by bus using smart card data," *ISPRS International Journal of Geo-Information*, vol. 9, no. 12, p. 751, 2020.

[38] C. Chen, J. Cai, C. Wang et al., "Estimation of personal $PM_{2.5}$ and BC exposure by a modeling approach - results of a panel study in Shanghai, China," *Environment International*, vol. 118, pp. 194–202, 2018.

[39] E. Dons, L. Int Panis, M. Van Poppel et al., "Impact of time–activity patterns on personal exposure to black carbon," *Atmospheric Environment*, vol. 45, no. 21, pp. 3594–3602, 2011.

[40] A. S. Geyh, J. Xue, H. Ozkaynak, and J. D. Spengler, "The Harvard Southern California chronic ozone exposure study: assessing ozone exposure of grade-school-age children in two Southern California communities," *Environmental Health Perspectives*, vol. 108, no. 3, pp. 265–270, 2000.

[41] L. Wallace and R. Williams, "Use of personal-indoor-outdoor sulfur concentrations to estimate the infiltration factor and outdoor exposure factor for individual homes and persons," *Environmental Science & Technology*, vol. 39, no. 6, pp. 1707–1714, 2005.

[42] L. Wallace, R. Williams, A. Rea, and C. Croghan, "Continuous weeklong measurements of personal exposures and indoor concentrations of fine particles for 37 health-impaired North Carolina residents for up to four seasons," *Atmospheric Environment*, vol. 40, no. 3, pp. 399–414, 2006.

[43] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, "Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation," *Journal of Computational and Graphical Statistics*, vol. 24, no. 1, pp. 44–65, 2015.

[44] B. M. Greenwell, B. C. Boehmke, and B. Gray, "Variable importance plots-an introduction to the vip package," *R J*, vol. 12, no. 1, p. 343, 2020.

[45] B. M. Greenwell, B. C. Boehmke, and A. J. McCarthy, "A simple and effective model-based variable importance measure," 2018, https://arxiv.org/abs/1805.04755.

[46] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, Springer, 2009.

[47] L.-T. Chang, P. Koutrakis, P. Catalano, and H. Suh, "Assessing the importance of different exposure metrics and time-activity data to predict 24-h personal $PM_{2.5}$ exposures," *Journal of Toxicology and Environmental Health, Part A*, vol. 66, no. 16-19, pp. 1825–1846, 2003.

[48] L. Rojas-Bracho, H. H. Suh, P. J. Catalano, and P. Koutrakis, "Personal exposures to particles and their relationships with personal activities for chronic obstructive pulmonary disease patients living in Boston," *Journal of the Air & Waste Management Association*, vol. 54, no. 2, pp. 207–217, 2004.

[49] B. J. Turpin, C. P. Weisel, M. Morandi et al., "Relationships of indoor, outdoor, and personal air (RIOPA): part II. Analyses of concentrations of particulate matter species," *Research report (Health Effects Institute)*, vol. 130, Part 2, pp. 1–77, 2007.

[50] Z. Qiu, W. Wang, J. Zheng, and H. Lv, "Exposure assessment of cyclists to UFP and PM on urban routes in Xi'an, China," *Environmental Pollution*, vol. 250, pp. 241–250, 2019.

[51] W. Ye, E. Saikawa, A. Avramov, S. H. Cho, and R. Chartier, "Household air pollution and personal exposure from burning firewood and yak dung in summer in the eastern *Tibetan Plateau*," *Environmental Pollution*, vol. 263, no. Part B, article 114531, 2020.

[52] Q. Y. Meng, D. Spector, S. Colome, and B. Turpin, "Determinants of indoor and personal exposure to $PM_{2.5}$ of indoor and outdoor origin during the RIOPA study," *Atmospheric Environment (1994)*, vol. 43, no. 36, pp. 5750–5758, 2009.

[53] N. Li, C. Xu, D. Xu et al., "Personal exposure to $PM_{2.5}$ in different microenvironments and activities for retired adults in two megacities, China," *Science of the Total Environment*, vol. 865, p. 161118, 2023.

[54] Y. Hwang and K. Lee, "Contribution of microenvironments to personal exposures to PM_{10} and $PM_{2.5}$ in summer and winter," *Atmospheric Environment*, vol. 175, pp. 192–198, 2018.

[55] S. Lim, J. Kim, T. Kim et al., "Personal exposures to $PM_{2.5}$ and their relationships with microenvironmental concentrations," *Atmospheric Environment*, vol. 47, pp. 407–412, 2012.

[56] K. J. Koistinen, O. Hänninen, T. Rotko, R. D. Edwards, D. Moschandreas, and M. J. Jantunen, "Behavioral and environmental determinants of personal exposures to $PM_{2.5}$ in EXPOLIS–Helsinki, Finland," *Atmospheric Environment*, vol. 35, no. 14, pp. 2473–2481, 2001.

[57] T. Lanki, A. Ahokas, S. Alm et al., "Determinants of personal and indoor $PM_{2.5}$ and absorbance among elderly subjects with coronary heart disease," *Journal of Exposure Science & Environmental Epidemiology*, vol. 17, no. 2, pp. 124–133, 2007.

[58] J. Kearney, L. Wallace, M. MacNeill, M.-E. Héroux, W. Kindzierski, and A. Wheeler, "Residential infiltration of fine and ultrafine particles in Edmonton," *Atmospheric Environment*, vol. 94, pp. 793–805, 2014.

[59] M. MacNeill, J. Kearney, L. Wallace et al., "Quantifying the contribution of ambient and indoor-generated fine particles to indoor air in residential environments," *Indoor Air*, vol. 24, no. 4, pp. 362–375, 2014.

[60] C. Xu, D. Xu, Z. Liu et al., "Estimating hourly average indoor $PM_{2.5}$ using the random forest approach in two megacities, China," *Building and Environment*, vol. 180, article 107025, 2020.