Environmental Science & Technology LETTERS

pubs.acs.org/journal/estlcu

Letter

A Random Forest Model for Daily $PM_{2.5}$ Personal Exposure Assessment for a Chinese Cohort

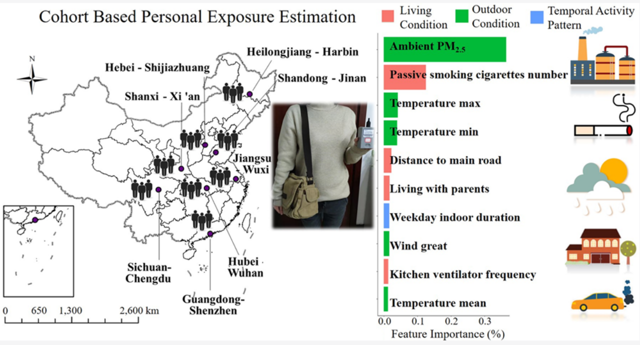
Yanwen Wang, Yanjun Du, Jianlong Fang, Xiaoyan Dong, Qiong Wang, Jie Ban, Qinghua Sun, Runmei Ma, Wenjing Zhang, Mike Z. He, Cong Liu, Yue Niu, Renjie Chen, Haidong Kan, and Tiantian Li*

Cite This: *Environ. Sci. Technol. Lett.* 2022, 9, 466−472

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Errors in air pollution exposure assessment are often considered as a major limitation in epidemiological studies. However, it is difficult to obtain accurate personal level exposure on cohort populations due to the often prohibitive expense. Personal exposure estimation models are used in lieu of direct personal exposure measures but still suffer from issues of availability and accuracy. We aim to establish a personal $PM_{2.5}$ exposure assessment model for a cohort population and assess its performance by applying our model on cohort subjects. We analyzed data from representative sites selected from the subclinical outcomes of polluted air in China (SCOPA-China) cohort study and established a random forest model for estimating daily $PM_{2.5}$ personal exposure. We also applied the model among subjects recruited in the project mentioned above within the same area and study period to estimate the reliability of the model. The established model showed a good fit with an R^2 of 0.81. The model application results showed similar patterns with empirically measured data. Our pilot study provided a validated and feasible modeling approach for assessing daily personal $PM_{2.5}$ exposure for large cohort populations. The promising model framework can improve $PM_{2.5}$ exposure assessment accuracy for future environmental health studies of large populations.



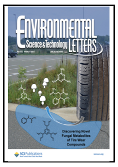
KEYWORDS: *personal exposure model, $PM_{2.5}$, random forest model, cohort*

INTRODUCTION

Exposure measurement error in fine particulate matter [particles with an aerodynamic diameter of ≤2.5 μm ($PM_{2.5}$)] has long been considered as a limitation of air pollution epidemiological studies.[1,2] Early epidemiological studies tended to use particle concentrations from the nearest fixed monitoring stations to represent exposure levels of study subjects,[3,4] and the ambient concentration is important for the evaluation of population health effects.[5] However, the monitored ambient concentrations could show large differences compared to personal $PM_{2.5}$ exposure measurements,[6,7] indicating that $PM_{2.5}$ concentrations from fixed monitoring stations may serve as a poor proxy for personal exposure. Furthermore, using fixed site monitoring data as a substitution for direct personal exposure measures may introduce errors into the effect estimates of epidemiological studies,[8,9] indicating that inaccurate exposure assessment can bias effect estimates toward the null[10,11] or even lead to invalid inferences.[12] Therefore, it is worth studying and has important supporting significance for establishing the relationship between personal exposure and ambient exposure, and accurately assessing personal $PM_{2.5}$ exposure is critical for minimizing exposure measurement error for epidemiological studies.

Cohorts are important epidemiological study designs with the benefits of often having sufficient power for health effect analysis, as well as reflecting the heterogeneity of a target population.[13] However, due to the prohibitive costs and intense resource demands, air pollution prediction models are often used as a practical solution for exposure assessment. Previous cohort studies established estimation models as a proxy for exposure,[14] and some models can obtain $PM_{2.5}$ concentrations at the individual level.[15] However, these modeling methods need to measure indoor pollutants and observe housing characteristics,[16] which are difficult to apply to large populations. Most previous personal exposure model

Received: December 1, 2021
Revised: March 28, 2022
Accepted: April 1, 2022
Published: April 7, 2022



https://doi.org/10.1021/acs.estlett.1c00970

assessment studies used sophisticated model input parameters,[17] and the coefficients were difficult to generalize in large study populations such as cohorts.

Recent studies have built personal exposure estimation models in small scale populations, which provide a preliminary foundation for more advanced modeling methods.[18−20] The studies attempted to facilitate the obtaining of model monitoring parameters, yet the model still has various shortcomings such as large estimation errors that lead to deficiencies in accuracy.[18,19] Some studies included activity pattern information and used linear regression to model the relationship between $PM_{2.5}$ personal exposure concentration and multiple variables.[18,19] However, the studies either lacked model validation or had low predictive accuracy, where the R^2 of testing data is unsatisfactory. Moreover, the predictors included in the models, such as meteorological variables, were found to have a nonlinear relationship with $PM_{2.5}$ concentration,[20] indicating that linear regression may not be the best approach. Considering these shortcomings, it is important to employ improved assessment methods to address the problems described above. Random forest models have been recently applied to estimate the spatiotemporal distribution of pollutants,[21−23] provided reliable performance,[24] and also have the ability to capture nonlinear relationships of variables. The successful applications of random forest models in other settings make it a promising method for establishing personal exposure estimation for large population studies.

This study aims to establish a daily personal $PM_{2.5}$ exposure assessment model that can be applied for a large cohort population and evaluate model performance by applying the established model to a representative air pollution cohort in China.

■ MATERIALS AND METHODS

Study Subjects and Collection of Data. Our study was based on the subclinical outcomes of polluted air in China (SCOPA-China) cohort.[25] The location of the study sites is shown in Figure S1. The cohort selected a study period of nonheating seasons from September to December 2017 and 2018 and selected nine communities of urban areas in eight provinces, with 214 subjects included. The selection of the study sites included different latitudes, climate types, and geographic regions. The study subjects were 40−89 years of age with a sex ratio of approximately 1:1 and were healthy with no smoking habits. The cohort obtained approvals (2017016 and 201820) from the ethics committee of the National Institute of Environmental Health, Chinese Center for Disease Control and Prevention (NIEH, China CDC).

Subjects in our study were instructed to carry a personal $PM_{2.5}$ exposure monitor (MicroPEM) for 3 days. The MicroPEM monitors were calibrated with fixed site monitors before each visit. During the 3 day study period, subjects were instructed to document their daily activities. We collected baseline information about subjects from community surveys and questionnaires. Ambient $PM_{2.5}$ concentrations, as well as meteorological parameters, were documented from the nearest monitoring stations of the residential communities of the subjects. The monitoring and meteorological data were matched and processed as daily average values, and data of each monitoring day of each subject were treated as an independent data record, with a total of 579 entries included in our study.

Model Establishment. Random forest models consist of regression trees for obtaining average prediction results of multiple subsample sets through several rounds of training.[21,26] The method was applied in our study to predict daily $PM_{2.5}$ personal exposure concentrations by parameters such as basic information about sites and populations, living conditions, outdoor conditions, and temporal activity patterns. Model parameters were prescreened using the Boruta algorithm before model establishment (https://CRAN.R-project.org/package=Boruta). The filtering algorithm can reduce the misleading impact of random fluctuations and correlations by adding randomness to the system and collecting results from the ensemble of randomized samples.[27] Detailed information about the selected variables is presented in Table S1. The study data were separated randomly into two data sets, with 90% of the data served as training data and 10% as testing data. Model sensitivity analyses were conducted through each stage of cross validation on the basic parameters of the random forest model, which included the number of variables tried at each split (mtry) and the number of trees (ntree) to ensure the reliability of the model, which followed the principle of gridsearch.[28,29] The accuracy of our model was verified by 10-fold cross-validation. The contributions of selected parameters to prediction results were assessed by impurity-based feature importance, which serves as an indicator of variable importance.

Model Validation and Application. We used the Bland−Altman method to assess the validation of the established model.[30] The number of paired measurements that were beyond the 95% limit of agreement reflected the performance of model prediction. Additionally, the testing data set was applied in the established model, and we used linear regression to evaluate the fit of observed and predicted concentrations.

We also conducted additional sensitivity analysis. As each of the study sites represents typical geographical regions, to maintain the unique nature of the geographical region, we randomly chose 90% of data from each of the study areas to form a training data set. The rest of the data formed a testing data set. From this analysis, we aimed to spatially validate the accuracy of the model. To further verify the accuracy of personal exposure prediction model results compared to outdoor monitoring and modeling concentrations, we compared the differences between personal monitoring concentrations and outdoor monitoring concentrations, outdoor modeling concentrations, and personal modeling prediction concentrations, respectively. The outdoor modeling concentrations were from the full coverage ground level $PM_{2.5}$ 0.01 degree data of the CHAP data product (ChinaHighPM$_{2.5}$).[32,33] We also establish the prediction model by using the top 20 important predictors and tested the model performance as further sensitive analysis.

The established model was applied among the subjects recruited in the project mentioned above from the same study areas during the same study period. We used the established model on the cohort data set and obtained the ratio between monitored $PM_{2.5}$ personal exposure concentration and monitored ambient concentration from the data set that underwent personal exposure assessment and the ratio between predicted personal exposure concentration and monitored ambient concentration from the application data set. We also analyzed the distribution of the ratios to investigate the difference between personal exposure and ambient concentrations.

■ RESULTS AND DISCUSSION

In this study, 51 variables were selected in the model by the filtering of the Boruta algorithm. Detailed descriptions are listed in Table 1. During the study period, the average daily ambient $PM_{2.5}$ concentration was 49.5 $\mu g/m^3$. The feature importance of the top 20 model parameters of the established random forest model is shown in Figure 1.

Table 1. Descriptive Statistics of Outdoor and Personal Exposure Variables

	mean	SD[a]	min	max	medium
precipitation 20−20 h (0.1 mm)	0.6	2.4	0.0	20.0	0.0
RH mean[b] (%)	62.7	17.6	27.0	93.0	65.0
RH min[c] (%)	41.4	17.3	15.0	74.0	42.0
temperature mean (°C)	14.1	4.5	5.5	24.4	13.9
temperature max (°C)	19.3	4.2	9.7	28.2	19.2
temperature min (°C)	10.1	5.4	0.1	22.9	9.8
wind speed mean (m/s)	1.6	0.5	0.4	2.7	1.7
wind max[d] (m/s)	3.7	1.0	2.2	7.6	3.7
wind great[e] (m/s)	6.9	1.7	3.7	14.2	6.6
ambient $PM_{2.5}$ ($\mu g/m^3$)	49.5	37.4	8.0	195.0	40.8
personal $PM_{2.5}$ ($\mu g/m^3$)	36.8	29.1	0.5	314.6	28.5

[a] Standard deviation. [b] Mean relative humidity. [c] Minimum relative humidity. [d] Maximum value of the 10 min average wind speed. [e] Maximum value of the instantaneous wind speed in a given period.

Outdoor conditions proved to make a large contribution to personal exposure in our study, and ambient $PM_{2.5}$ concentrations contributed the most in the established model. Systematic review studies reported that outdoor sources can contribute 33−55% of total personal $PM_{2.5}$ exposure, indicating a large proportion of the contribution to personal exposure.[34] Our study found that the feature importance by ambient $PM_{2.5}$ ranks first, which indicates that ambient $PM_{2.5}$ contributed the most to personal exposure. The high rank of the ambient concentration indicated that ambient particle pollution contributes greatly to personal particle exposure. Temperature variables ranked within the top 20 of feature importance, which may be due to the effects of temperature on indoor and outdoor air exchange rates.[18,35] Other meteorological variables such as daily maximum wind speed also ranked high in feature importance, which may be related to the influence of the wash-off effects of particle air pollution,[36] which indicated wind might reduce the concentration of pollutants.

The number of cigarettes from indoor passive smoking ranked second in the feature importance, which agrees with previous studies of indoor particulate pollution.[37,38] Indoor concentrations of $PM_{2.5}$ with household smokers could reach approximately 2.5 times the concentrations of $PM_{2.5}$ when no smoking was reported,[39] indicating the importance of passive smoking as a particle source. The window opening range conditions of parlors and bedrooms during weekdays and weekend and kitchen ventilator use contributed greatly to feature importance. This finding together with previous indoor air quality studies[40] shows that the ventilation of indoor air could drastically reduce particle concentrations.

Daily activities, including commuting duration, sleeping duration, and duration being indoors, also ranked highly in the feature importance, which supports the influence of the time spent in different microenvironments on personal particle exposure. Our study suggests that it is important to take the

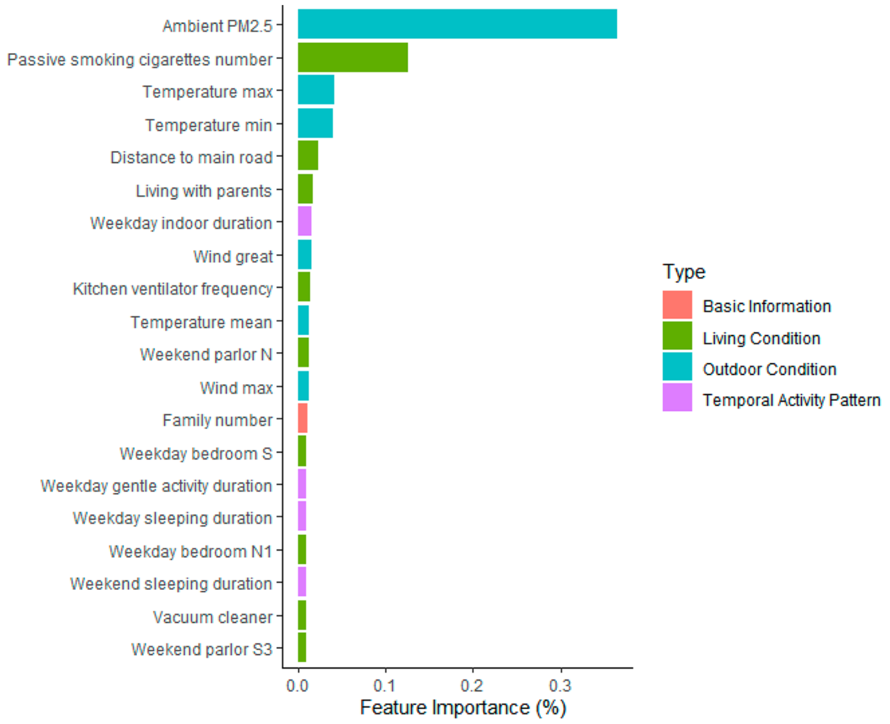


Figure 1. Importance ranking of the top 20 selected variables. Abbreviations: Weekend parlor N, duration of windows being closed in the parlor during the weekend; Weekday bedroom N1, duration of windows being closed in the bedroom during weekdays; Weekend parlor S3, duration of windows being more than two-thirds open in the parlor during the weekend; Wind max, maximum value of the 10 min average wind speed; Wind great, maximum value of the instantaneous wind speed in a given period.

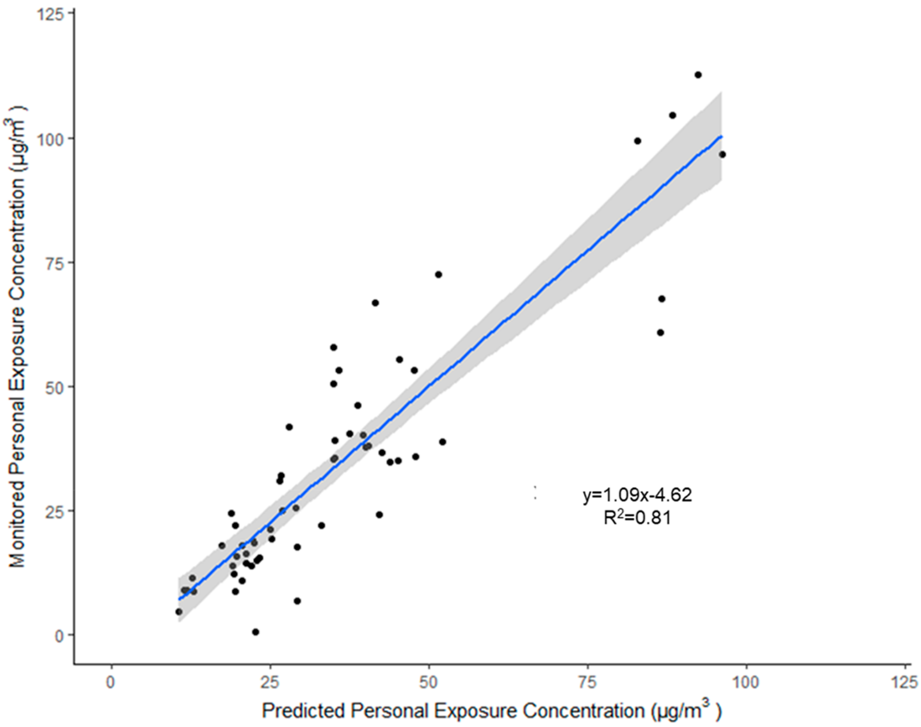


Figure 2. Linear fitting results of the predicted and observed concentrations using testing data. The gray shaded area indicates the 95% confidence interval for the fitted linear regression line.

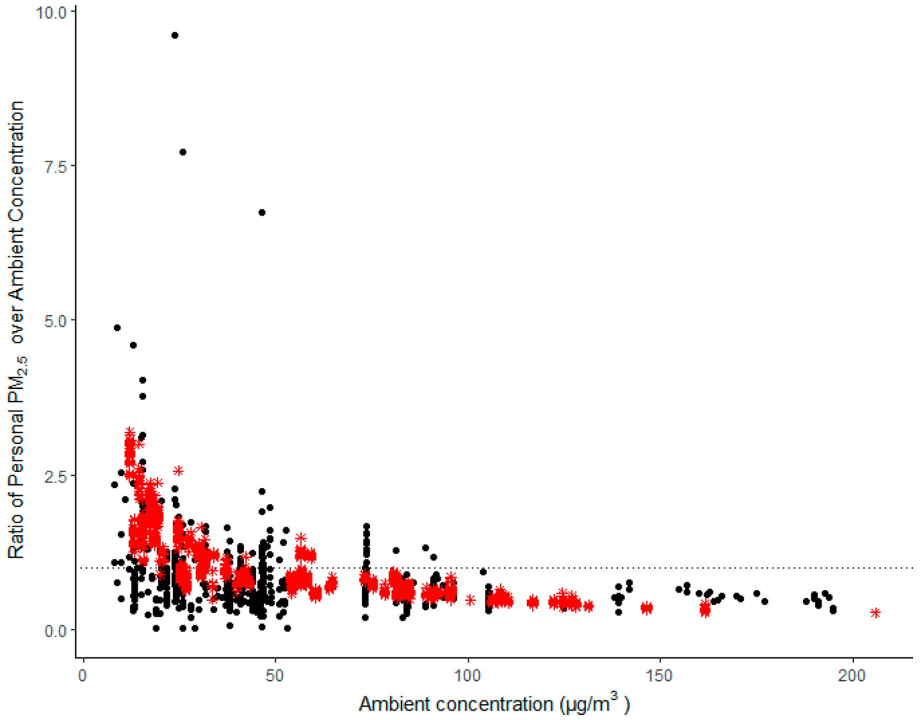


Figure 3. Comparison among ambient monitored, personal monitored, and personal predicted $PM_{2.5}$ concentrations. Black dots represent the ratios of monitored personal $PM_{2.5}$ to ambient concentration, while the red asterisks represent the ratio of personal predicted $PM_{2.5}$ concentration to ambient monitored concentrations.

time spent in different microenvironments into consideration when estimating the personal exposure level in epidemiological studies. The model establishment and sensitivity analysis results are shown in Figures S2−S8. Linear regression results of the predicted and observed $PM_{2.5}$ concentrations are shown in Figure 2. The linear regression analysis afforded an R^2 of 0.81 and a slope of 1.09.

Comparisons among ambient monitored, personal monitored, and personal predicted $PM_{2.5}$ exposure concentrations are shown in Figure 3. The ratios were below the line of 1:1 with ambient concentrations of >50 μg/m^3, and ratios increased at lower ambient concentrations. The pattern of the monitored ratio and predicted ratio was similar to that at increasing ambient concentrations. Linear regression analysis between outdoor particle concentration and personal exposure concentration is shown in Table S2. The slope of the linear regression of personal exposure and ambient concentration was 0.46.

Our work showed a higher model fit compared to those of previous studies. Previous $PM_{2.5}$ modeling studies used multiple statistical methods, including leave-one-out cross-validation (LOOCV) analysis and 10-fold CV analysis, to assess model fit, with R^2 values reaching 0.759 and 0.730,[18] respectively. The linear regression results from our random forest model showed an R^2 of 0.81 and a slope of 1.09, indicating a good model fit. Furthermore, previous studies suggested that meteorological variables such as temperature and relative humidity showed nonlinear relationships with $PM_{2.5}$ concentrations.[20] The random forest model was used to reveal the nonlinear relationship and was reported to show better model performance compared to linear regression models of the same study design.[31,41]

The validation and feasibility of the model application are the highlights of our study. The parameters in our study were obtained through environmental monitoring sites, meteorological stations, and questionnaires, which are typically accessible variables in cohort studies and increase the feasibility of using the model. Moreover, we conducted sensitivity analysis to estimate the predictive performance and the validity of the model. In our study, the study sites were chosen from typical geographical regions, and there was only one study site from each geographical region. The sensitive analysis ensured the model performance among the typical regions. The linear regression results showed that the model can be used among the geographical regions mentioned above.

Additionally, the performance of the established model was further discussed with respect to the application data set with the same variable information. The ratio values of predicted personal exposure and monitored personal exposure to monitored outdoor exposure concentrations were similar, which strongly supports the accuracy of the model. Personal exposure levels were generally higher than ambient levels when ambient particle concentrations were low, and the trend was consistent with the reviewed results reported in previous studies.[42]

There are several limitations of our study. First, the study period was relatively short compared to that of large cohort studies. Future studies with a longer study period should include the entire heating season. Second, as a classical problem of prediction analysis in both statistical models[21,43] and dynamical models,[44] our model found it difficult to capture extreme values, which may be due to the simplicity of the study subjects. Future studies should include more diverse subjects in a larger age range, subjects from rural areas, and active smokers to improve model applicability. Third, although our study contained data from eight provinces, the study sites were still relatively limited. Future studies should include sites at different economic levels and a variety of climatic zones to ensure regional representation and improved spatial validation. Fourth, because 70% of the study subjects were retired or unemployed in our study (Table S1) and the activity areas were around the residence areas, we used the monitoring site closest to their residential areas to represent ambient concentrations in the model. Future studies with more diverse subjects could include ambient concentrations around work places in the model for further analysis. Fifth, the feature importance used in our analysis can indicate a certain variable has strong effects on the target variable, yet to explain how the feature variables affect the target variables, there would need to be additional analysis such as Shapley value estimation to further determine the modes of the effect of the feature variables in future studies.

Our work is a pilot study aimed at establishing a daily $PM_{2.5}$ personal exposure assessment model in a cohort population of typical geographical regions using a random forest model. The established model can reduce research costs associated with field monitoring of individual level exposure measurements and provide a better understanding of the contributions of indoor and outdoor parameters to personal exposure concentrations. To the best of our knowledge, this is the first study using a random forest model approach to obtain personal $PM_{2.5}$ exposure estimation for a cohort population. The validation and feasibility of the model enable it to be easily replicated on cohort studies of large populations with questionnaire and station monitoring data. Our work provided a possible approach for estimating personal exposure levels and revealed a method for refined individual level exposure assessment of the public. Our model can be further applied in environmental health studies of cohort populations and reduce potential exposure assessment bias of epidemiological effect estimates.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.estlett.1c00970.

> Detailed information about the study subjects and the collection of data (Text S1), model establishment and sensitivity analysis results (Text S2), discussion of the limitations of spatial validation methods (Text S3), study site locations (Figure S1), descriptive statistics of selected variables (Table S1), Bland−Altman analysis (Figure S2), and model sensitivity analysis results (Figures S3−S8 and Table S2) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Tiantian Li − *China CDC Key Laboratory of Environment and Population Health, National Institute of Environmental Health, Chinese Center for Disease Control and Prevention, Beijing 100021, China; Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing, Jiangsu 211166, China;* orcid.org/0000-0003-2938-3917; Email: litiantian@nieh.chinacdc.cn

Authors

Yanwen Wang − *China CDC Key Laboratory of Environment and Population Health, National Institute of Environmental Health, Chinese Center for Disease Control and Prevention, Beijing 100021, China*

Yanjun Du − *China CDC Key Laboratory of Environment and Population Health, National Institute of Environmental Health, Chinese Center for Disease Control and Prevention, Beijing 100021, China*

Jianlong Fang − *China CDC Key Laboratory of Environment and Population Health, National Institute of Environmental*

Health, Chinese Center for Disease Control and Prevention, Beijing 100021, China; orcid.org/0000-0003-4716-390X

Xiaoyan Dong – *China CDC Key Laboratory of Environment and Population Health, National Institute of Environmental Health, Chinese Center for Disease Control and Prevention, Beijing 100021, China*

Qiong Wang – *China CDC Key Laboratory of Environment and Population Health, National Institute of Environmental Health, Chinese Center for Disease Control and Prevention, Beijing 100021, China*

Jie Ban – *China CDC Key Laboratory of Environment and Population Health, National Institute of Environmental Health, Chinese Center for Disease Control and Prevention, Beijing 100021, China*

Qinghua Sun – *China CDC Key Laboratory of Environment and Population Health, National Institute of Environmental Health, Chinese Center for Disease Control and Prevention, Beijing 100021, China*

Runmei Ma – *China CDC Key Laboratory of Environment and Population Health, National Institute of Environmental Health, Chinese Center for Disease Control and Prevention, Beijing 100021, China*

Wenjing Zhang – *China CDC Key Laboratory of Environment and Population Health, National Institute of Environmental Health, Chinese Center for Disease Control and Prevention, Beijing 100021, China*

Mike Z. He – *Department of Environmental Medicine and Public Health, Icahn School of Medicine at Mount Sinai, New York, New York 10029, United States;* orcid.org/0000-0003-2357-3883

Cong Liu – *School of Public Health, Key Lab of Public Health Safety of the Ministry of Education and NHC Key Lab of Health Technology Assessment, Fudan University, Shanghai 200032, China*

Yue Niu – *School of Public Health, Key Lab of Public Health Safety of the Ministry of Education and NHC Key Lab of Health Technology Assessment, Fudan University, Shanghai 200032, China*

Renjie Chen – *School of Public Health, Key Lab of Public Health Safety of the Ministry of Education and NHC Key Lab of Health Technology Assessment, Fudan University, Shanghai 200032, China*

Haidong Kan – *School of Public Health, Key Lab of Public Health Safety of the Ministry of Education and NHC Key Lab of Health Technology Assessment, Fudan University, Shanghai 200032, China;* orcid.org/0000-0002-1871-8999

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.estlett.1c00970

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was funded by the National Key Research and Development Program of China (2017YFC0211706 and 2017YFC0211701), the National Natural Science Foundation of China (91543111, 41907368, 41907367, 21906156, 81573247, 91643205, and 91743111), and the National Research Program for Key Issues in Air Pollution Control (Premier Funding DQGG0401).

REFERENCES

(1) Armstrong, B. Study design for exposure assessment in epidemiological studies. *Sci. Total Environ.* **1995**, *168* (2), 187–94.

(2) Bateson, T. F.; Wright, J. M. Regression calibration for classical exposure measurement error in environmental epidemiology studies using multiple local surrogate exposures. *Am. J. Epidemiol.* **2010**, *172* (3), 344–52.

(3) Lanki, T.; Ahokas, A.; Alm, S.; Janssen, N. A.; Hoek, G.; De Hartog, J. J.; Brunekreef, B.; Pekkanen, J. Determinants of personal and indoor PM2.5 and absorbance among elderly subjects with coronary heart disease. *J. Expo Sci. Environ. Epidemiol.* **2007**, *17* (2), 124–33.

(4) Meng, Q. Y.; Spector, D.; Colome, S.; Turpin, B. Determinants of Indoor and Personal Exposure to PM(2.5) of Indoor and Outdoor Origin during the RIOPA Study. *Atmos Environ. (1994)* **2009**, *43* (36), 5750–5758.

(5) Weisskopf, M. G.; Webster, T. F. Trade-offs of Personal Versus More Proxy Exposure Measures in Environmental Epidemiology. *Epidemiology.* **2017**, *28* (5), 635–643.

(6) Adgate, J. L.; Ramachandran, G.; Pratt, G. C.; Waller, L. A.; Sexton, K. Longitudinal variability in outdoor, indoor, and personal PM2.5 exposure in healthy non-smoking adults. *Atmos. Environ.* **2003**, *37* (7), 993–1002.

(7) Janssen, N. A.; Hoek, G.; Harssema, H.; Brunekreef, B. Personal exposure to fine particles in children correlates closely with ambient fine particles. *Arch. Environ. Health* **1999**, *54* (2), 95–101.

(8) Kioumourtzoglou, M. A.; Spiegelman, D.; Szpiro, A. A.; Sheppard, L.; Kaufman, J. D.; Yanosky, J. D.; Williams, R.; Laden, F.; Hong, B.; Suh, H. Exposure measurement error in PM2.5 health effects studies: a pooled analysis of eight personal exposure validation studies. *Environ. Health* **2014**, *13* (1), 2.

(9) Ni, Y.; Wu, S.; Ji, J.; Chen, Y.; Zhao, B.; Shi, S.; Tu, X.; Li, H.; Pan, L.; Deng, F.; Guo, X. The exposure metric choices have significant impact on the association between short-term exposure to outdoor particulate matter and changes in lung function: Findings from a panel study in chronic obstructive pulmonary disease patients. *Sci. Total Environ.* **2016**, *542*, 264–270.

(10) Alexeeff, S. E.; Schwartz, J.; Kloog, I.; Chudnovsky, A.; Koutrakis, P.; Coull, B. A. Consequences of kriging and land use regression for PM2.5 predictions in epidemiologic analyses: insights into spatial variability using high-resolution satellite data. *J. Expo Sci. Environ. Epidemiol.* **2015**, *25* (2), 138–44.

(11) Zeger, S. L.; Thomas, D.; Dominici, F.; Samet, J. M.; Schwartz, J.; Dockery, D.; Cohen, A. Exposure measurement error in time-series studies of air pollution: concepts and consequences. *Environ. Health Perspect.* **2000**, *108* (5), 419–26.

(12) Butland, B. K.; Samoli, E.; Atkinson, R. W.; Barratt, B.; Katsouyanni, K. Measurement error in a multi-level analysis of air pollution and health: a simulation study. *Environ. Health* **2019**, *18* (1), 13.

(13) Wu, B. U.; Johannes, R. S.; Sun, X.; Tabak, Y.; Conwell, D. L.; Banks, P. A. The early prediction of mortality in acute pancreatitis: a large population-based study. *Gut.* **2008**, *57* (12), 1698–703.

(14) Gall, E. T.; Chen, A.; Chang, V. W.-C.; Nazaroff, W. W. Exposure to particulate matter and ozone of outdoor origin in Singapore. *Building and Environment.* **2015**, *93*, 3–13.

(15) Aaron, C. P.; Hoffman, E. A.; Kawut, S. M.; Austin, J. H. M.; Budoff, M.; Michos, E. D.; Hinckley Stukovsky, K.; Sack, C.; Szpiro, A. A.; Watson, K. D.; Kaufman, J. D.; Barr, R. G. Ambient air pollution and pulmonary vascular volume on computed tomography: the MESA Air Pollution and Lung cohort studies. *Eur. Respir. J.* **2019**, *53* (6), 1802116.

(16) Kaufman, J. D.; Adar, S. D.; Allen, R. W.; Barr, R. G.; Budoff, M. J.; Burke, G. L.; Casillas, A. M.; Cohen, M. A.; Curl, C. L.; Daviglus, M. L.; Diez Roux, A. V.; Jacobs, D. R., Jr.; Kronmal, R. A.; Larson, T. V.; Liu, S. L.; Lumley, T.; Navas-Acien, A.; O'Leary, D. H.; Rotter, J. I.; Sampson, P. D.; Sheppard, L.; Siscovick, D. S.; Stein, J. H.; Szpiro, A. A.; Tracy, R. P. Prospective study of particulate air pollution exposures, subclinical atherosclerosis, and clinical cardio-

vascular disease: The Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air). *Am. J. Epidemiol.* **2012**, *176* (9), 825−37.

(17) Burke, J. M.; Zufall, M. J.; Ozkaynak, H. A population exposure model for particulate matter: case study results for PM(2.5) in Philadelphia, PA. *J. Expo Anal Environ. Epidemiol.* **2001**, *11* (6), 470−89.

(18) Chen, C.; Cai, J.; Wang, C.; Shi, J.; Chen, R.; Yang, C.; Li, H.; Lin, Z.; Meng, X.; Zhao, A.; Liu, C.; Niu, Y.; Xia, Y.; Peng, L.; Zhao, Z.; Chillrud, S.; Yan, B.; Kan, H. Estimation of personal PM2.5 and BC exposure by a modeling approach - Results of a panel study in Shanghai, China. *Environ. Int.* **2018**, *118*, 194−202.

(19) Xu, M.; Jia, Y.; Li, G.; X, P. Evaluation of Personal Integrated Exposure to Fine Particle in a Community in Beijing. *J. Environ. Health* **2011**, *28* (11), 941−943.

(20) Wang, J.; Wang, Y.; Liu, H.; Yang, Y.; Zhang, X.; Li, Y.; Zhang, Y.; Deng, G. Diagnostic identification of the impact of meteorological conditions on PM2.5 concentrations in Beijing. *Atmos. Environ.* **2013**, *81*, 158−165.

(21) Ma, R.; Ban, J.; Wang, Q.; Li, T. Statistical spatial-temporal modeling of ambient ozone exposure for environmental epidemiology studies: A review. *Sci. Total Environ.* **2020**, *701*, 134463.

(22) Zhan, Y.; Luo, Y.; Deng, X.; Grieneisen, M. L.; Zhang, M.; Di, B. Spatiotemporal prediction of daily ambient ozone levels across China using random forest for human exposure assessment. *Environ. Pollut.* **2018**, *233*, 464−473.

(23) Hu, X.; Belle, J. H.; Meng, X.; Wildani, A.; Waller, L. A.; Strickland, M. J.; Liu, Y. Estimating PM2.5 Concentrations in the Conterminous United States Using the Random Forest Approach. *Environ. Sci. Technol.* **2017**, *51* (12), 6936−6944.

(24) Brokamp, C.; Jandarov, R.; Rao, M. B.; LeMasters, G.; Ryan, P. Exposure assessment models for elemental components of particulate matter in an urban environment: A comparison of regression and random forest approaches. *Atmos Environ. (1994)* **2017**, *151*, 1−11.

(25) Li, T.; Chen, R.; Zhang, Y.; Fang, J.; Zhao, F.; Chen, C.; Wang, J.; Du, P.; Wang, Q.; Shi, W.; Han, J.; Hu, X.; Kan, H.; Shi, X. Cohort profile: Sub-clinical outcomes of polluted air in China (SCOPA-China cohort). *Environ. Int.* **2020**, *134*, 105221.

(26) Zhao, C.; Wang, Q.; Ban, J.; Liu, Z.; Zhang, Y.; Ma, R.; Li, S.; Li, T. Estimating the daily PM2.5 concentration in the Beijing-Tianjin-Hebei region using a random forest model with a 0.01 degrees x0.01 degrees spatial resolution. *Environ. Int.* **2020**, *134*, 105297.

(27) Kursa, M. B.; Rudnicki, W. R. Feature Selection with the Boruta Package. *Journal of Statistical Software* **2010**, *36* (11), 1−13.

(28) Singh, H.; Singh, S.; Singla, D.; Agarwal, S. M.; Raghava, G. P. QSAR based model for discriminating EGFR inhibitors and non-inhibitors using Random forest. *Biol. Direct* **2015**, *10*, 10.

(29) Worachairungreung, M.; Ninsawat, S.; Witayangkurn, A.; Dailey, M. N. Identification of Road Traffic Injury Risk Prone Area Using Environmental Factors by Machine Learning Classification in Nonthaburi, Thailand. *Sustainability* **2021**, *13* (7), 3907.

(30) Bland, J. M.; Altman, D. G. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* **1986**, *1*, 307−310.

(31) Wang, Y.; Du, Y.; Wang, J.; Li, T. Calibration of a low-cost PM2.5 monitor using a random forest model. *Environ. Int.* **2019**, *133* (Part A), 105161.

(32) Wei, J.; Li, Z.; Lyapustin, A.; Sun, L.; Peng, Y.; Xue, W.; Su, T.; Cribb, M. Reconstructing 1-km-resolution high-quality PM2.5 data records from 2000 to 2018 in China: spatiotemporal variations and policy implications. *Remote Sensing of Environment* **2021**, *252*, 112136.

(33) Wei, J.; Li, Z.; Cribb, M.; Huang, W.; Xue, W.; Sun, L.; Guo, J.; Peng, Y.; Li, J.; Lyapustin, A.; Liu, L.; Wu, H.; Song, Y. Improved 1 km resolution PM2.5 estimates across China using enhanced space-time extremely randomized trees. *Atmos. Chem. Phys.* **2020**, *20*, 3273−3289.

(34) Evangelopoulos, D.; Katsouyanni, K.; Keogh, R. H.; Samoli, E.; Schwartz, J.; Barratt, B.; Zhang, H.; Walton, H. PM2.5 and NO2 exposure errors using proxy measures, including derived personal exposure from outdoor sources: A systematic review and meta-analysis. *Environ. Int.* **2020**, *137*, 105500.

(35) Zhou, X.; Cai, J.; Zhao, Y.; Chen, R.; Wang, C.; Zhao, A.; Yang, C.; Li, H.; Liu, S.; Cao, J.; Kan, H.; Xu, H. Estimation of residential fine particulate matter infiltration in Shanghai, China. *Environ. Pollut.* **2018**, *233*, 494−500.

(36) Xu, X.; Xia, J.; Gao, Y.; Zheng, W. Additional focus on particulate matter wash-off events from leaves is required: A review of studies of urban plants used to reduce airborne particulate matter pollution. *Urban Forestry & Urban Greening.* **2020**, *48*, 126559.

(37) Gemenetzis, P.; Moussas, P.; Arditsoglou, A.; Samara, C. Mass concentration and elemental composition of indoor PM2.5 and PM10 in University rooms in Thessaloniki, northern Greece. *Atmos. Environ.* **2006**, *40* (17), 3195−3206.

(38) Wallace, L. Indoor particles: a review. *J. Air Waste Manag Assoc.* **1996**, *46* (2), 98−126.

(39) Koistinen, K. J.; Hanninen, O.; Rotko, T.; Edwards, R. D.; Moschandreas, D.; Jantunen, M. J. Behavioral and environmental determinants of personal exposures to PM2.5 in EXPOLIS - Helsinki, Finland. *Atmos. Environ.* **2001**, *35* (14), 2473−2481.

(40) Chen, C.; Zhao, B. Review of relationship between indoor and outdoor particles: I/O ratio, infiltration factor and penetration factor. *Atmos. Environ.* **2011**, *45* (2), 275−288.

(41) Ryan, P. H.; Brokamp, C.; Fan, Z. H.; Rao, M. B. Analysis of Personal and Home Characteristics Associated with the Elemental Composition of PM2.5 in Indoor, Outdoor, and Personal Air in the RIOPA Study. *Res. Rep. - Health Eff. Inst.* **2015**, 3−40.

(42) Fan, Y.; Han, Y.; Liu, Y.; Wang, Y.; Chen, X.; Chen, W.; Liang, P.; Fang, Y.; Wang, J.; Xue, T.; Yao, Y.; Li, W.; Qiu, X.; Zhu, T. Biases Arising from the Use of Ambient Measurements to Represent Personal Exposure in Evaluating Inflammatory Responses to Fine Particulate Matter: Evidence from a Panel Study in Beijing, China. *Environmental Science & Technology Letters.* **2020**, *7* (10), 746−752.

(43) Hoek, G.; Beelen, R.; de Hoogh, K.; Vienneau, D.; Gulliver, J.; Fischer, P.; Briggs, D. A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmos. Environ.* **2008**, *42* (33), 7561−7578.

(44) Yang, M.; Fan, H.; Zhao, K. PM2.5 Prediction with a Novel Multi-Step-Ahead Forecasting Model Based on Dynamic Wind Field Distance. *Int. J. Environ. Res. Public Health* **2019**, *16* (22), 4482.