Supplementary Material

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], 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, United State

Corresponding Author:

*Telephone: + 86 010 50930157, E-mail: xuchunyu@nieh.chinacdc.cn

TABLE S1 Hyperparameters for tuning machine learning model

NO	Machine learning Algorithm	Hyperparameter
1	Random forest	n_{tree} = 500; m_{try} = 1:(number of candidate variables)

2	Artificial neural network	Decay = seq(0.01, 0.1, by=0.02);
		size = 1:20

3	Support vector machine	sigma = c(0.01, 0.015, 0.1, 0.2, 0.4, 0.5, 0.8, 1, 2, 4); *C* = c(0.01, 0.05, 0.1, 0.2, 0.25, 0.5, 1, 2, 4, 8, 16)

4	Extreme gradient boosting	nrounds = seq(100, 600, by=100);
		max_depth = 3:7;
		gamma = c(0.01, 0.02);
		eta= c(0.05, 0.1, 1);
		colsample_bytree = 0.75;
		subsample = 0.5;
		min_child_weight = 0

5	Gradient boosting machine	interaction.depth = c(1:10);
		n.trees = c(25, 50, 100, 150, 200, 250, 300);
		shrinkage = c(0.05, 0.1, 0.2);
		n.minobsinnode = c(2, 5, 10, 20, 30, 40)

TABLE S2 The list of candidate predictors for 24-h average personal $PM_{2.5}$

NO	Variable	Type	Value
Routine monitoring			
1	Ambient $PM_{2.5}$	Continuous variable	μg/m^3
2	Outdoor temperature	Continuous variable	℃
3	Outdoor relative humidity	Continuous variable	%
4	Wind speed	Continuous variable	m/s
5	Air pressure	Continuous variable	hPa
Basic questionnaire			
6	Gender	Categorical variable	0=Female; 1=Male
7	Age	Continuous variable	year
8	Education degree	Categorical variable	1 = Primary School and below 2 = Junior High School 3 = High School or Junior College 4 = College and above

9	Household income	Categorical variable	1 = ≤50,000 RMB/year;
			2 = 50,001~100,000 RMB /year;
			3 = 100,001 ~ 150,000 RMB /year;
			4 = 150,001 ~ 200,000 RMB /year;
			5 = 200,001 ~ 250,000 RMB /year;
			6 = > 250,000 RMB /year

10	Number of children	Continuous variable	n
11	Number of family members	Continuous variable	n
12	Number of pets	Continuous variable	n
13	Floors	Continuous variable	n
14	Building age	Continuous variable	year
15	Years since last housing renovation	Continuous variable	year
16	Distance to the nearest main road	Continuous variable	m

17	Room volume	Continuous variable	m^3

18	Cooking frequency	Categorical variable	1 = 3 times/day;
			2 = 2 times/day;
			3 = 1 times/day;
			4 = none

19	Duration of each cooking session	Categorical variable	1 = <20 min; 2 = 20 ~ 40 min 3 = 40 ~ 60 min 4 = >60 min

20	Room cleaning frequency	Categorical variable	1 = every day;
			2 = every 2~3 days;
			3 = every 4~5 days;
			4 = weekly

21	Windows opening number	Continuous variable	n
22	Window opening width	Categorical variable	1 = <10%; 2 = 11% ~ 20%; 3 = 21% ~ 50%; 4 = 51% ~ 80%; 5 = >80%
23	Window opening time	Continuous variable	h
24	Air conditioner use	Continuous variable	h
Time-activity diary			
25	Time in transit	Continuous variable	%
26	Time at home	Continuous variable	%
27	Time in indoor public place	Continuous variable	%
28	Time outdoors	Continuous variable	%
29	Exposure to ETS	Continuous variable	%
30	Cooking time	Continuous variable	%
31	Time percent of cleaning	Continuous variable	%

TABLE S3 Residence, demographic, and activity characteristics of study subjects

	BJ	**NJ**
Monitored participants, *n*	33	33
Age (years)	62 (53, 86)	59 (43, 78)
Gender, *n*, (%)		
Female	19 (57.6)	19 (57.6)
Male	14 (42.4)	14 (42.4)
Building age (years)	18 (5, 56)	11 (3, 32)
Years since the latest decoration (years)	12 (0, 20)	7 (3, 32)
Distance to the nearest major road (m)	45 (15, 354)	108 (11, 380)
Floor, *n*, (%)		
1st－3rd	10 (30.3)	10 (30.3)
4th－9th	11 (33.3)	13 (39.4)
≥10th	12 (36.4)	10 (30.3)
Total household income (Yuan), *n*, (%)		
≤50,000	4 (12.1)	3 (9.1)
50,001－100,000	11 (33.3)	10 (30.3)
100,001－150,000	12 (36.4)	15 (45.5)
150,001－200,000	3 (9.1)	1 (3.0)
200,001－250,000	2 (6.1)	4 (12.1)
＞250,000	1 (3.0)	0 (0)
Window opening width		
≤10%	16 (5.9)	24 (8.1)
11%－20%	49 (18.1)	66 (22.2)
21%－50%	50 (18.5)	40 (13.5)
51%－80%	15 (5.5)	30 (10.1)
>80%	141 (52.0)	137 (46.1)

Window opening time (min/d)	480 (0, 1440)	840 (0, 1440)
Have dog/cat, *n*, (%)	3 (9.1)	3 (9.1)
Use air conditioner (min/d)	0 (0, 992)	0 (0, 697)
Use air purifier (min/d)	0 (0, 1436)	0 (0, 609)
ETS exposure time (min/d)	0 (0, 127)	0 (0, 276)
Cooking time (min/d)	20 (0, 219)	40 (0, 225)
Cleaning time (min/d)	0 (0, 239)	32 (0, 311)
Meteorological factors		
Outdoor temperature (℃)	25.4 (-8.5, 29.7)	21.7 (4.0, 31.7)
Outdoor relative humidity (RH, %)	45.4 (11.9, 87.5)	62.8 (31.3, 83.2)
Wind speed (m/s)	2.1 (1.2, 5.2)	1.5 (0.7, 3.7)
Air pressure (kPa)	100.5 (99.2, 104.0)	101.7 (100.0, 103.4)
Time-activity data (%)		
Indoors, total	93.2 (55.0, 100.0)	95.0 (70.8, 100.0)
Residence	90.4 (53.9, 100.0)	92.8 (51.9, 100.0)
Public building	0.9 (0.0, 26.7)	1.1 (0.0, 31.1)
Transportation	3.1 (0.0, 18.6)	1.9 (0.0, 23.4)
Outdoors, not in traffic	1.7 (0.0, 35.6)	1.7 (0.0, 25.7)

Notes: Continuous variables are reported as median (min, max).

TABLE S4 Importance scores of variables included in the final prediction models

Data source	Variable	BJ						NJ					
		MLR	**RF**	**SVM**	**GBM**	**XGBoost**	**NNet**	**MLR**	**RF**	**SVM**	**GBM**	**XGBoost**	**NNet**
Routine monitoring	Ambient $PM_{2.5}$	0.82	39.15	28.49	60.08	47.28	63.75	0.62	25.56	35.55	49.92	31.89	53.38
	Air pressure		1.90	8.64		3.29	4.64		4.72	15.86		4.24	5.94
	Outdoor RH	0.08	8.65	12.26	5.93	8.46	9.04	0.12	5.34			5.32	
	Outdoor temperature		1.30			2.18			6.08	20.02		3.72	
	Wind speed		2.65	12.18		2.62		0.11	2.99		5.40	2.47	
Basic questionnaire	Age				2.83							2.80	
	Air conditioner use	0.16			4.27	1.95	6.35	0.08					
	Building age				3.01		3.88						
	Cleaning frequency							0.21					15.42
	Cooking frequency						2.23						
	Education degree						2.88						
	Floors							0.04					
	Household income						6.29						
	Window opening number						1.31	0.25				1.79	
	Window opening time		2.34		3.72	2.36							
	Window opening width						5.29	0.09					
Time-activity diary	Cooking time	0.21	3.41		4.56	4.81		0.10					
	Exposure to ETS	0.25	6.18		10.30	8.17		0.29	7.20		12.90	9.66	20.56

	Time at home			41.37
	Time in indoor public place	2.95		27.88
	Time in transit		1.95	31.97
	Time outdoors	3.25		32.14