

## Urban Air Pollution Mapping Using Fleet Vehicles as Mobile Monitors and Machine Learning

Bu Zhao, Long Yu, Chunyan Wang, Chenyang Shuai, Ji Zhu, Shen Qu, Morteza Taiebat, and Ming Xu\*



Cite This: *Environ. Sci. Technol.* 2021, 55, 5579–5588



Read Online

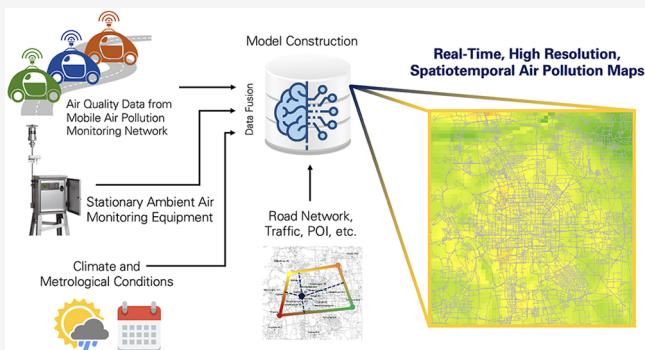
ACCESS |

Metrics & More

Article Recommendations

Supporting Information

**ABSTRACT:** Spatially explicit urban air quality information is important for developing effective air quality control measures. Traditionally, urban air quality is measured by networks of stationary monitors that are not universally available and sparsely sited. Mobile air quality monitoring using equipped vehicles is a promising alternative but has focused on vehicle-level experiments and lacks fleet-level demonstration. Here, we equipped 260 electric vehicles in a ride-hailing fleet in Beijing, China with low-cost sensors to collect real-time, spatial-resolved data on fine particulate matter ( $PM_{2.5}$ ) concentrations. Using this data, we developed a decision tree model to infer the distribution of  $PM_{2.5}$  concentrations in Beijing at 1 km by 1 km and 1 h resolution. Our results are able to show both short- and long-term variations of urban  $PM_{2.5}$  concentrations and identify local air pollution hotspots. Compared with a benchmark model that only uses data from stationary monitoring sites, our model has shown significant improvement with the coefficient of determination increased from 0.56 to 0.80 and root mean square error decreased from 12.6 to  $8.1 \mu\text{g}/\text{m}^3$ . To the best of our knowledge, this study collects the largest mobile sensor data for urban air quality monitoring, which are augmented by state-of-the-art machine learning techniques to derive high-quality urban air pollution mapping. Our results demonstrate the potential and necessity of using fleet vehicles as routine mobile sensors combined with advanced data science methods to provide high-resolution urban air quality monitoring.



Ambient air pollution has become a major threat to public health in both developed and developing countries.<sup>1,2</sup> Today, 91% of the world's population lives in areas where air pollution exceeds the limits of the World Health Organization (WHO) guidelines.<sup>3</sup> Notably, the majority of air pollution-related health impact occurs in urban areas where more than half of the world's population lives.<sup>4</sup>

Improving urban air quality requires air pollution measurements and monitoring, but site-based urban air pollution monitoring is limited worldwide. Robust air pollution observations are largely absent in many developing countries.<sup>5</sup> In more affluent regions, fixed-site ambient air pollution monitors are also sparse, primarily due to high cost and space limitation.<sup>6–9</sup> For example, there are 17 continuous regulatory monitors in the New York metropolitan area, with one per 1.2 million people and  $526 \text{ km}^2$ .<sup>10</sup> In Beijing, China, there are currently 35 continuous regulatory monitors with approximately two per million people and  $1000 \text{ km}^2$ . On the other hand, the spatial and temporal distribution of air pollutants can vary greatly over short distances and periods due to uneven distribution of emission sources and complex physical-chemical transformations, especially in populous urban areas.<sup>11,12</sup> Such heterogeneous spatiotemporal distribution is not well captured by the traditional fixed-site stationary monitoring system with sparse observations. The lack of

understanding of the spatiotemporal distribution of air pollutants can significantly affect air quality control, exposure assessment, and environmental justice.<sup>11,13,14</sup>

A variety of methods have been used to improve the spatiotemporal coverage and resolution of air quality monitoring, but still with various limitations. Air quality data estimated from satellite remote sensing is usually spatially coarse (>1–10 km resolution)<sup>15</sup> and frequently unavailable (e.g., due to cloud covering).<sup>16–18</sup> Physical and chemical models such as dispersion models and chemical transport models can offer high fidelity, but tend to be computationally expensive and highly rely on the quality of underlying emission inventories.<sup>19–23</sup> Geostatistical methods such as land use regression can provide estimations with high spatial resolution, but lack sufficient temporal resolution and relies on the availability of updated local land-use data.<sup>24–28</sup>

Received: December 1, 2020

Revised: March 13, 2021

Accepted: March 15, 2021

Published: March 24, 2021



Mobile sensing using equipped vehicles has been considered as a cost effective and operationally viable alternative to provide real-time, high-resolution air quality measures within cities and neighborhoods.<sup>6,7,29,30</sup> Traditionally, mobile air quality monitoring has been done using lab-grade equipment (i.e., mobile labs) operated by trained staff. Due to the high operating and maintenance cost, these equipped vehicles are sparsely deployed (e.g., one or two for a city) on fixed routes for a limited period. Thus, the spatial coverage and resolution are still not sufficient for high-resolution, real-time air pollution monitoring (i.e., only one or two data points at each timestamp across the entire city). As a result, large-scale, continuous application of mobile air quality monitoring is still limited. Alternatively, low-cost sensors can be used to equip a larger number of vehicles for larger spatial coverage and higher resolution. However, such studies are rare in the literature and data obtained from low-cost sensors are not necessarily accurate, thus cannot be directly used to infer air quality distribution.<sup>31</sup>

Here, we address these challenges of mobile air quality monitoring by exploring the potential of combining a large network of low-cost mobile sensors and state-of-the-art machine learning techniques. Specifically, we equipped a total of 260 electric vehicles in a ride-hailing fleet in Beijing with low-cost sensors to collect real-time, spatial-resolved data on fine particulate matter ( $\text{PM}_{2.5}$ ) concentrations in two 1 month experiments with one in winter and the other in summer. After filtering out incomplete and erroneous data, the resulting dataset has almost 28 million data points each of which includes a record of the  $\text{PM}_{2.5}$  concentration, the location (latitude and longitude), and time of the record. Using these data and the publicly available hourly average  $\text{PM}_{2.5}$  concentration data from regulatory monitoring stations, we develop a machine learning model to infer  $\text{PM}_{2.5}$  concentrations for the core urban area of Beijing (within the Sixth Ring Road) at 1 km by 1 km and 1 h resolution. We demonstrate that our model shows superior performance in inferring urban air quality distribution when compared with a baseline model only using stationary monitoring data and other machine learning techniques. To the best of our knowledge, our study collects the largest amount of mobile sensing data for urban air pollution monitoring. Our results demonstrate the potential and necessity of mobile monitoring using low-cost air quality sensors and provide evidence for developing effective urban air pollution control strategies.

## METHODS AND MATERIALS

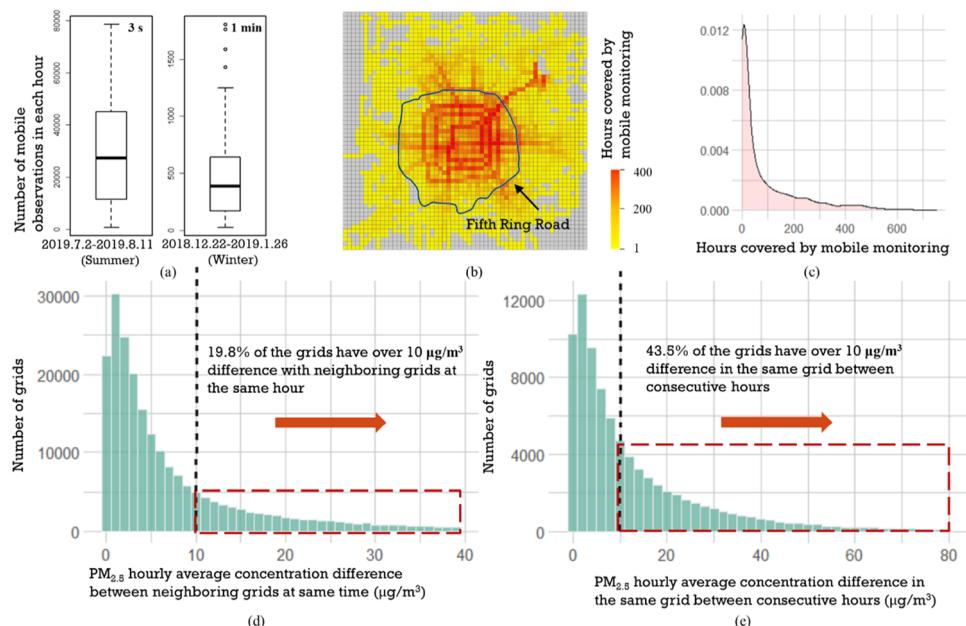
**Study Area.** We focus on the area within the Sixth Ring Road in Beijing (Figure S4), which covers about 3363  $\text{km}^2$  and is the core administrative, business, and residential area in Beijing. A total of 22 regulatory air quality monitoring stations and 35 meteorological stations are located in this area (Figure S1). The study area has a population of over 16 million<sup>32</sup> with high population density ( $\sim 5000/\text{km}^2$ ), complex road structure ( $\sim 10,826 \text{ km}$  total road length), and diverse land use types (35 types). This area regularly suffers from severe air pollution. The largest source of locally generated  $\text{PM}_{2.5}$  emissions in this area is vehicle emissions followed by road and construction dust.<sup>33</sup> We conducted two mobile monitoring campaigns from December 22, 2018 to January 26, 2019 (winter) and July 2, 2019 to August 11, 2019 (summer) in Beijing.

**Mobile Monitoring Campaign.** A two-phase mobile monitoring campaign for  $\text{PM}_{2.5}$  pollution was conducted in the

study area in winter and summer. During the first-phase (winter) monitoring, 100 fleet vehicles (electric vehicles to minimize pipeline emissions) were equipped. BR-SMART sensors from BRAMC Medical & Technology Co. Ltd. (Beijing, China) were used as mobile monitoring probes. This mobile sensor is capable of continuously collecting on-road  $\text{PM}_{2.5}$  concentrations at 1 min frequency for 6–8 h (powered by battery). The first phase was conducted from December 22, 2018 to January 26, 2019 and yielded a total of 392,295 min of data. A total of 256,782 min (65.5%) of data remained after removing GPS and sampling errors. The relative low data recovery rate is mainly due to the misoperation by the drivers (e.g., forgetting to turn off the device during charging, out of battery). For the second phase, 160 fleet vehicles (also electric vehicles) were equipped. XD-YL-1 sensors from Elklin Technology Co. Ltd. (Xi'an, China) were used. This customized mobile sensor is capable of continuously collecting on-road  $\text{PM}_{2.5}$  concentrations at 3 s frequency when the vehicle is operating using on-board charging. The second phase was conducted from July 2 to August 11, 2019 and collected 34,012,344 observations, equivalent to 1,700,617 min of data. After data preprocessing, 27,621,091 (81.2%) observations remained. In both phases, the sensors were mounted on the top end of the vehicles, which were operating as normal to provide ride-hailing services. The sensors were mounted in a way that the air inlet faced backward to avoid the effect of headwind and the exhaust from the vehicle in front of the equipped vehicle. Both sensors were calibrated during production and adjusted again before installation. The accuracy of the sensors was also validated by comparing with  $\text{PM}_{2.5}$  concentration data from collocated stationary monitors. More details about the sensors parameters, installation, and calibration can be found in the Supporting Information (Figures S2–S4).

**Data Collection.** The study area is divided into 57 by 59 (3363 in total) 1 km by 1 km grids based on the sampling frequency of the sensors and the number of grids that can be covered in each hour by the mobile monitoring campaign (Figure S5). A variety of data are collected and allocated to each grid. Hourly  $\text{PM}_{2.5}$  concentration data during the study periods are collected from the 22 stations located in the study area. These stations are administered by the National Meteorological Information Center (NMIC) and provided hourly concentrations of particulate matters ( $\text{PM}_{2.5}$ ,  $\text{PM}_{10}$ ) and inorganic gaseous pollutants ( $\text{CO}_2$ ,  $\text{CO}$ ,  $\text{SO}_2$ , and  $\text{NO}_x$ ). We consider the hourly  $\text{PM}_{2.5}$  concentration data from the stationary monitoring as the ground truth. Road network data are collected from OpenStreetMap (OSM, <https://www.openstreetmap.org>) with extracted features including road type (primary road, secondary road, tertiary road, motorway, and trunk), length, and the number of intersections. Meteorological data are from the 35 meteorological stations administrated by the China Meteorological Administration (CMA) and Beijing Meteorology Bureau, including hourly air pressure, temperature, relative humidity, wind direction, and speed. Land use data include the area of land in each of the 35 types (e.g., building, factory, and commercial) in each grid. Traffic volume is approximated by the average speed (km/h) for the equipped vehicles in real time. Additional details for these data can be found in Table S1.

The collected mobile monitoring data include real-time  $\text{PM}_{2.5}$  concentrations, timestamp, and location (longitude and latitude). First, we remove outliers including abnormal extreme



**Figure 1.** Spatial and temporal characteristics of the mobile monitoring data. (a) Number of mobile monitoring samples in each hour. The difference between the two monitoring phases is due to different sampling frequencies (1 min in winter and 3 s in summer); (b) spatial coverage (2621 grids) is 78% of the entire study area. Due to range limitation of the electric vehicles, most of the samples are located within the Fifth Ring Road; (c) probability density plot for the grids covered by different number of hours of mobile monitoring campaign; (d) distribution of the differences of hourly average PM<sub>2.5</sub> concentrations of two adjacent grids with the same timestamp; and (e) distribution of the differences of hourly average PM<sub>2.5</sub> concentrations of the same grid in two consecutive hours.

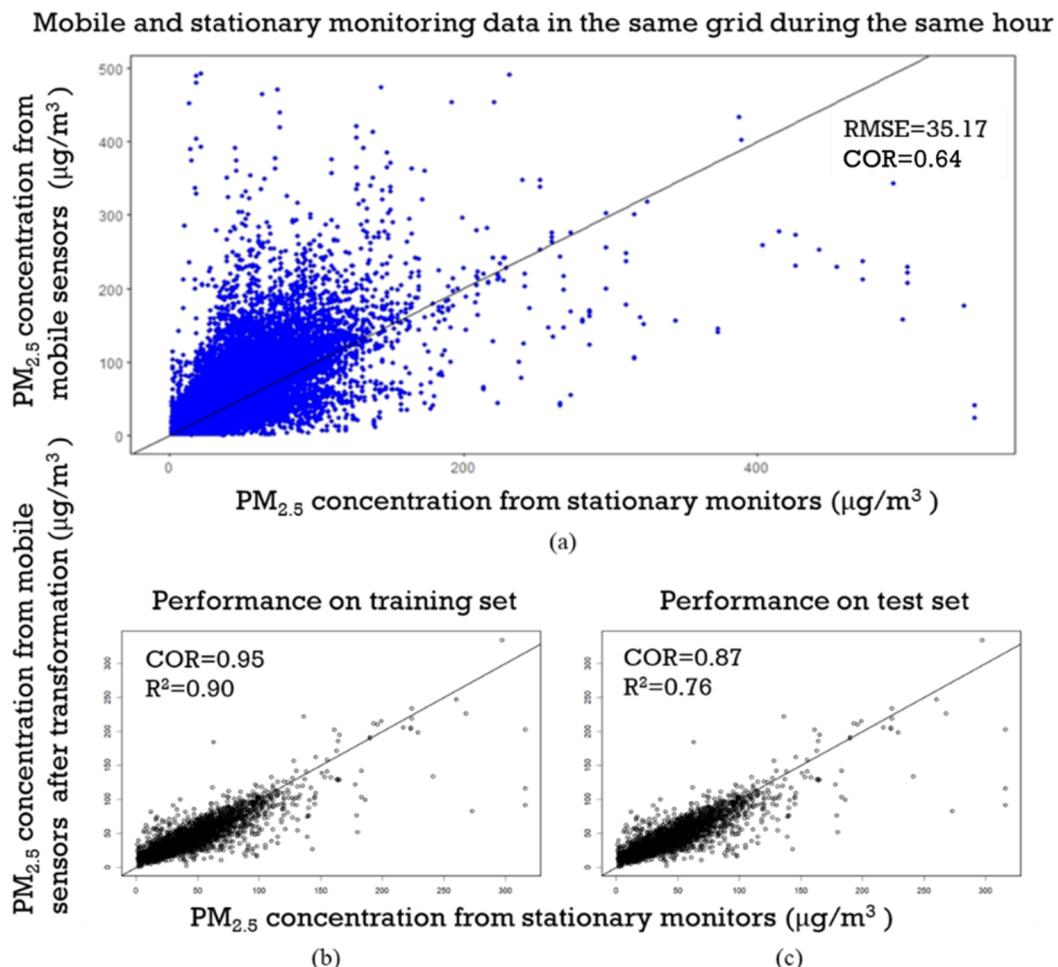
values such as those data with geolocation outside the study area.<sup>34</sup> Next, we allocate the remaining data into 57 by 59 (3363 in total) 1 km by 1 km grids. In theory, the vehicle could travel from one grid to another within one measurement cycle, but the number of cross-grid cases in our sample is small in both sampling phases (14.6% in the first phase and 0.7% in the second phase). Data in each grid are then aggregated into 1 h median and mean to match the sampling frequency of the stationary monitoring, which provides the ground truth. The aggregated data calculated by less than 2 observations with 1 min sampling frequency and 40 observations with 3 s sampling frequency (equivalent to 2 observations with 1 min sampling frequency) are removed to avoid extreme values due to few observations (Figure S8). This minimal number of observation needed in a single grid that also restricted the spatial resolution can be applied. Finally, we have the following data for each grid-hour: (1) location: latitude and longitude of the center of each grid; (2) road network: the length of five types of road and number of intersections in each grid; (3) meteorological data: hourly air pressure, temperature, relative humidity, wind direction, and speed from the nearest meteorological station; (4) land use: the size of each of the 35 land-use types in each grid; (5) stationary monitoring data: hourly PM<sub>2.5</sub> concentrations from the nearest monitoring station; and (6) mobile monitoring data: hourly median, mean, first and third quartiles, minimum, and maximum of PM<sub>2.5</sub> concentrations, the average speed of the equipped vehicle, and the number of mobile monitoring observations.

**Model Construction.** One common challenge faced by urban air quality mapping is the lack of continuous regulatory monitors to provide a spatially distributed benchmark as the ground truth. In our study area, there are only 22 continuous regulatory monitors located in 22 grids, leaving the other 3341 grids without ground truth measures to evaluate and validate

air quality mapping. Here, we address this challenge by transforming the mobile monitoring data to approximate the ground truth in grids without stationary monitoring data. Specifically, we first develop a transformation model to predict the hourly stationary monitoring results (ground truth) by using mobile monitoring data in the same grid during the same hour (grid-hour). We then use the tested model to estimate hourly PM<sub>2.5</sub> concentrations for grid-hours with mobile monitoring data available but without stationary monitoring data. The estimated values are used as proxies of the ground truth in these grid-hours for developing the PM<sub>2.5</sub> mapping model for the entire study area in the next step.

The mobile monitoring data and stationary monitoring data within the same grid-hour are used to construct the data transformation model. Our central hypothesis is that the mobile monitoring data represent the distribution of PM<sub>2.5</sub> concentration for a specific grid, which is also affected by the number of mobile observations, local characteristics, and meteorological conditions. Thus, we use the summary statistics of the mobile monitoring data (i.e., hourly median, mean, first, and third quartiles of PM<sub>2.5</sub> concentrations), meteorological data, and land use data as input variables to construct the transform model. The model is then tuned based on the average relative importance (Figure S9) and the results from 10-fold cross-validation to avoid overfitting. Model hyperparameters are optimized by choosing the set that minimizes RMSE in 10-fold cross-validation. We developed two separate data transformation models with different hyperparameters for the PM<sub>2.5</sub> concentration data collected from the two experiment phases (details in the Supporting Information). We used 70% of the data as the training set and the rest 30% as the test set.

For urban air quality inference, the PM<sub>2.5</sub> concentration of a given grid may be affected by local features such as land use



**Figure 2.** Comparison between stationary and mobile monitoring data before and after data transformation. (a) Before the transformation, the Pearson correlation coefficient between stationary and mobile monitoring data in the same grid during the same hour is 0.64 with a high RMSE (35.17). After the transformation, the correlation increases significantly and the RMSE drops to 8.78 and 14.33 on the training set (b) and test set (c), respectively.

characteristics, traffic, and road network, as well as external factors such as meteorological information, which can affect regional pollutants transport and local pollutant deposition. In addition, the PM<sub>2.5</sub> concentration of a given grid is also temporally and spatially correlated to its adjacent units in the space–time domain. A variety of methods are used to construct the inference model, and the needed predictors are selected through a backward elimination procedure (i.e., removing the variables with relative importance less than 5%). Finally, the PM<sub>2.5</sub> concentrations from the nearest 4 mobile observations, nearest 3 stationary observations during the same hour (i.e., spatial neighbors), and nearest mobile observation during the last hour (i.e., temporal neighbor), meteorological data, land use data, and road network data are selected as input variables. The model is then tuned based on their average relative importance (Figure S10) and the results from the 10-fold cross-validation to avoid overfitting. We also developed two separate models for air quality inference with different hyperparameters for the PM<sub>2.5</sub> concentration data collected from the two experiment phases. The results showed similar accuracy of PM<sub>2.5</sub> concentrations mapping with an R<sup>2</sup> of 0.786 and 0.804, indicating the validity of training a single model utilizing all data collected. In addition, a single model trained using data collected in different seasons (winter and summer) can help the general application of the model across

different levels of PM<sub>2.5</sub> concentrations. Model hyperparameters can be found in the Supporting Information. We used 70% of the data as the training set and the rest 30% as the test set.

## RESULTS

After removing erroneous data, we have a total of 27,877,873 data points—equivalent to 1,637,837 min of observation (Figure 1a). The observations are mainly concentrated between 10:00 a.m. and 23:00 p.m. (Figure S1). In the 1 km by 1 km gridded study area (3363 km<sup>2</sup>), our mobile monitoring data cover about 78% of the grids (Figure 1b). However, at any given hour, on average, only 152 grids (4.5%, Figure S1) are monitored by mobile sensors due to the limited number of equipped vehicles and the uneven temporal distribution of the monitoring data.

The PM<sub>2.5</sub> concentrations recorded by the mobile sensors vary greatly across space and time. Figure 1d shows the distribution of the PM<sub>2.5</sub> concentration differences between two adjacent grids with the same timestamp, with about 20% higher than 10 micrograms per cubic meter (μg/m<sup>3</sup>). The WHO recommends 10 μg/m<sup>3</sup> as the threshold for PM<sub>2.5</sub> emissions that pose health risks,<sup>3</sup> and several cohort studies have confirmed that a 10 μg/m<sup>3</sup> increase for both short-term

and long-term PM<sub>2.5</sub> can significantly increase the all-cause mortality as well as respiratory and cardiovascular disease hospitalizations.<sup>35–39</sup> As a comparison, the annual average PM<sub>2.5</sub> concentration in Beijing is 50 µg/m<sup>3</sup> in 2018,<sup>40</sup> which indicates that the variances observed in our mobile monitoring data are significant.

Similarly, Figure 1e shows the distribution of PM<sub>2.5</sub> concentration differences for the same grid between two consecutive hours, with more than 40% over 10 µg/m<sup>3</sup>. These results confirm that local PM<sub>2.5</sub> concentrations in urban areas varies greatly in relatively small temporal and spatial scales, possibly driven by a combination of factors including the complex urban structure, nonuniform distribution of multiple emission sources, and spatially and temporally varying human activities.<sup>6,11,12</sup>

As shown in Figure 2a, there is a relatively high correlation (Pearson correlation coefficient: 0.64) between the hourly PM<sub>2.5</sub> concentration data from the stationary monitors and mobile monitoring data in the same grid-hour. In general, the PM<sub>2.5</sub> concentration measure from mobile sensors is higher than that from stationary monitors—root mean square error (RMSE) is 35.17, which is likely due to the pipeline emissions from nearby vehicles.

We developed and tested various machine learning models to predict the hourly stationary monitoring results by using mobile monitoring data in the same grid-hour. Specifically, we predict hourly PM<sub>2.5</sub> concentrations from stationary monitors using mobile monitoring data from the same grid-hour and corresponding meteorological and land use data as predictors (details in the Supporting Information). Based on the model performance (Table S1), the gradient boosting model using the extreme gradient boosting (XGBoost) algorithm is chosen to construct the transformation model. We randomly split the data into a training set (70%) and a test set (30%). The training set is used to train the model and the test set is only used to test the performance of the model. As shown in Figure 2b,c, the correlation between the predicted PM<sub>2.5</sub> concentration and the stationary monitoring data becomes much higher, with the Pearson correlation coefficient of 0.95 and 0.87 on the training set and test set, respectively. In addition, the difference of the predicted results from the stationary monitoring data becomes much smaller, with an RMSE of 14.33 and a coefficient of determination ( $R^2$ ) of 0.76 on the test set.

Using this model, we estimate hourly PM<sub>2.5</sub> concentrations for grid-hours with mobile monitoring data available but without stationary monitoring data as a proxy for the ground truth. As a result, we substantially expand the spatial coverage of our benchmarks with 78% of the grids in the study area. Next, we use the expanded dataset to predict PM<sub>2.5</sub> concentrations in grid-hours without any monitoring data.

**Model Performance.** We develop and test various machine learning models to predict PM<sub>2.5</sub> concentrations in grid-hours without either stationary or monitoring data. The model input variables include hourly PM<sub>2.5</sub> concentration data from the nearest 3 stationary monitors and 4 mobile monitors in the same hour, hourly concentration from the nearest mobile monitor in the last hour, and additional variables on land use, road networks, and meteorological data (details in the Supporting Information).

For each model, 10-fold cross-validation is applied with 70% of the data are used as the training set and the rest 30% as the test set in each fold. The performance of these models is

compared using mean absolute error (MAE), RMSE, and  $R^2$ . As shown in Table 1, decision tree models, XGBoost and

**Table 1. Model Performance Comparison<sup>a</sup>**

| methods                                  | MAE | RMSE | $R^2$ |
|--|-----|------|-------|
| extreme gradient boosting (XGBoost)      | 4.7 | 8.1  | 0.80  |
| XGBoost (without mobile monitoring data) | 6.7 | 12.6 | 0.56  |
| random forest (RF)                       | 5.1 | 8.3  | 0.79  |
| RF (without mobile monitoring data)      | 7.1 | 12.9 | 0.55  |
| spatial interpolation (SI)               | 7.1 | 12.8 | 0.55  |
| K-nearest neighbors (KNN)                | 7.8 | 13.6 | 0.53  |
| support vector regression (SVR)          | 6.0 | 12.4 | 0.59  |
| artificial neural network (ANN)          | 5.3 | 9.6  | 0.70  |

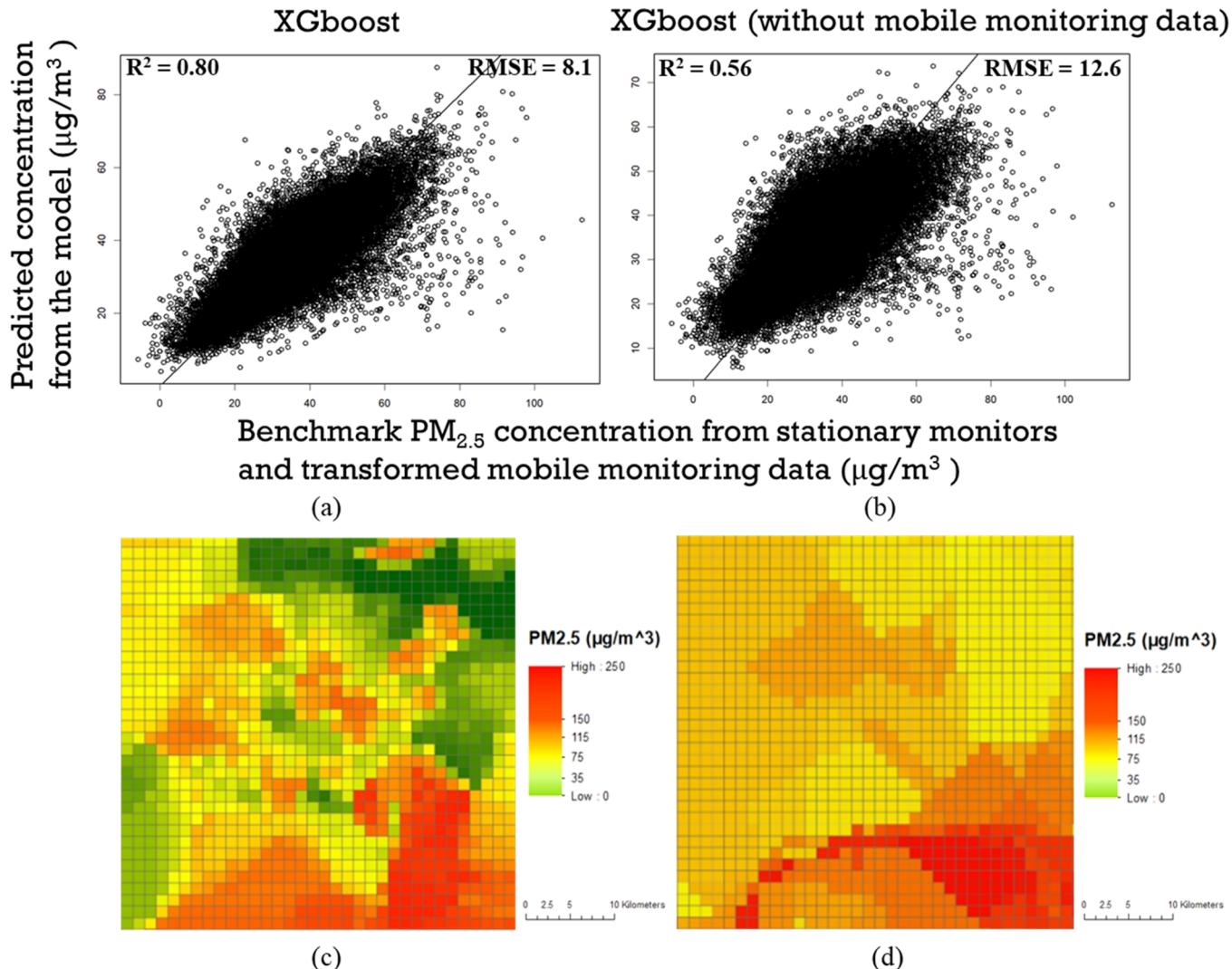
<sup>a</sup>Details of each model are in the Supporting Information.

random forest (RF), perform the best on all three metrics. Overall the XGBoost and RF models have similar performance, but XGBoost is less computational intensive due to the parallelization of tree construction.<sup>41</sup>

Using the mobile monitoring data can significantly improve the model performance. For example, the MAE and RMSE of the XGBoost model can be reduced from 6.7 to 4.7 and 12.6 to 8.1, respectively, and the  $R^2$  can be improved from 0.56 to 0.80. Note that the XGBoost model without using any mobile monitoring data can well infer the hourly PM<sub>2.5</sub> concentration on grids where stationary monitoring data are available ( $R^2 = 0.92$ , Figure S2). The high accuracy is mainly due to the high correlations between monitoring data at different stations (Figure S3). However, when the same model is applied to grids where stationary monitoring data are not available, the model performs much worse ( $R^2 = 0.56$ , Figure 3a). This is mainly due to the lack of variation in the input variables, as the input variables are the same for many grid-hours with the same nearest three stationary monitors. After adding the mobile monitoring data as an additional input to the model, the prediction accuracy is greatly improved ( $R^2$  from 0.56 to 0.80, Figure 3b) as anticipated.

To understand how the mobile monitoring data improves the model performance for different grids, we divided the grids into four categories: (1) the grids close to stationary monitors (<5 km) and with nearby mobile monitoring observations (<5 km), (2) the grids close to stationary monitors but without nearby mobile monitoring observations ( $>= 5$  km), (3) the grids away from stationary monitors but with nearby mobile monitoring observations ( $>= 5$  km), and (4) the grids away from stationary monitors and without nearby mobile monitoring observations. As shown in Table 2, overall, the use of mobile monitoring data can improve model performance. The largest improvement comes from the grids away from stationary monitors but with nearby mobile monitoring observations ( $R^2$  increased from 0.60 to 0.82), which is also the majority of the grids (90.3%). In addition, we also tested the robustness of our model by adding a random noise generated from the uniform distribution U(-1,1), U(-5,5), and U(-20,20), to the raw mobile monitoring data. The experiments were repeated 1000 times. After adding the noise, our model can still map the distribution of PM<sub>2.5</sub> concentrations with an  $R^2$  range from 0.786 to 0.808 (median: 0.800), 0.702 to 0.789 (median: 0.766), and 0.682 to 0.734 (median: 0.721), which demonstrates the robust of our model.

We then used the XGBoost model to map the PM<sub>2.5</sub> emissions in our study area at a specific time to further



**Figure 3.** Performance between XGBoost models with (a) and without (b) using mobile monitoring data as an input; and spatial distribution of inferred PM<sub>2.5</sub> concentrations of the central areas ( $34 \times 34 \text{ km}^2$ ) of Beijing at 12:00 p.m. on January 12, 2019, obtained from the model with (c) and without (d) mobile monitoring data.

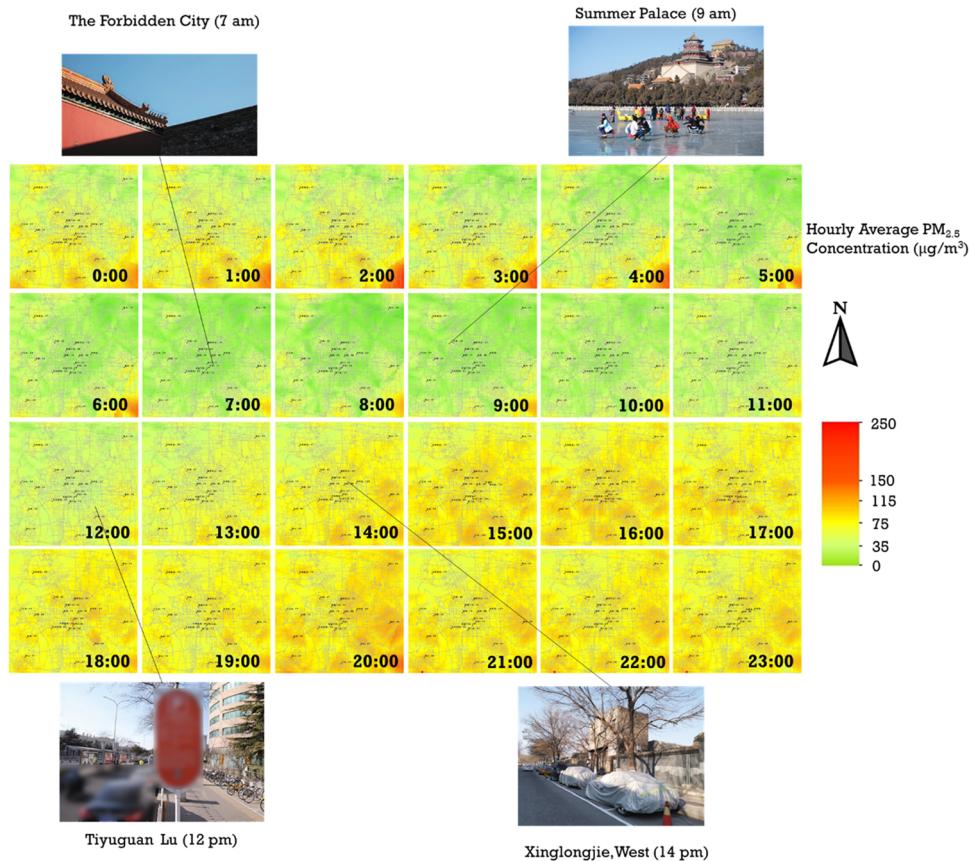
**Table 2. Model Performance on Different Grids**

| grids types based on distance/models                    | proportion of data in test set | XGBoost (with mobile monitoring data) |      |                | XGBoost (without mobile monitoring data) |      |                |
|---|--------------------------------|---------------------------------------|------|----------------|--|------|----------------|
|   |                                | MAE                                   | RMSE | R <sup>2</sup> | MAE                                      | RMSE | R <sup>2</sup> |
| stationary monitors (<5 km) and mobile monitors (<5 km) | 8.7%                           | 2.8                                   | 4.7  | 0.93           | 2.9                                      | 4.8  | 0.92           |
| stationary monitors (<5 km) and mobile monitors (>5 km) | 0.9%                           | 4.7                                   | 7.6  | 0.80           | 3.9                                      | 6.9  | 0.84           |
| stationary monitors (>5 km) and mobile monitors (<5 km) | 90.3%                          | 4.6                                   | 7.3  | 0.82           | 5.1                                      | 8.3  | 0.60           |
| stationary monitors (>5 km) and mobile monitors (>5 km) | 0.1%                           | 7.1                                   | 14.8 | 0.55           | 5.1                                      | 8.3  | 0.32           |

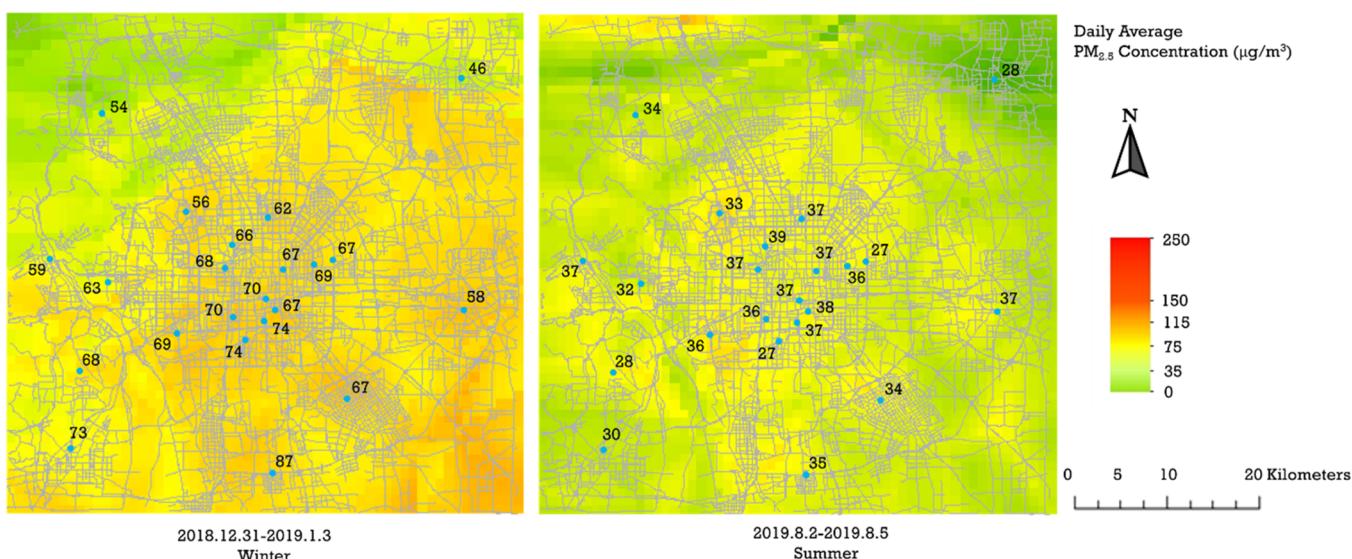
compare the results with and without mobile monitoring data. As depicted in Figure 3c, the inferred distribution of the PM<sub>2.5</sub> concentration from the model using mobile monitoring data shows great spatial heterogeneity that is consistent with the model performance. However, the mapping of PM<sub>2.5</sub> emissions without using mobile monitoring data consists of several large clusters without sufficient spatial heterogeneity (Figure 3d). This is mainly because without the mobile monitoring data, the model inputs for each grid are largely similar without significant variations (e.g., meteorological data, PM<sub>2.5</sub> concentration data from the nearest stationary monitors). The mobile monitoring data can provide the spatial variations as the model

input that is needed to infer the spatial heterogeneity of PM<sub>2.5</sub> emissions.

**Mapping PM<sub>2.5</sub> Emissions across Space and Time.** Figure 4 depicts the inferred spatial and temporal distribution of PM<sub>2.5</sub> emissions on a typical weekday in winter (December 31, 2018, Wednesday) in Beijing. We can easily identify a pollution hotspot in southeastern Beijing and the diffusion toward north and west throughout the day. This is consistent with the fact that there was a southeasterly wind of 1.864 mph after 12 p.m. according to the China Meteorological Administration (CMA).<sup>42</sup> In addition, a nearby monitoring station (Yongledian) that is just outside of our study area in



**Figure 4.** Hourly PM<sub>2.5</sub> concentration distribution on a typical weekday in winter (December 31, 2018) in Beijing inferred by our model, with real-time images from The Beijing News<sup>43</sup> for comparison.



**Figure 5.** Spatial-resolved daily average PM<sub>2.5</sub> concentration inferred in this study in winter (2018.12.31–2019.1.3) and summer (2019.8.2–2019.8.5) in Beijing, with daily average PM<sub>2.5</sub> concentration values from stationary monitors for the same periods highlighted as numbers.

the southeast also reported a high PM<sub>2.5</sub> concentration on the same day from 0:00 to 8:00, which is close to our model result for the observed hotspot (Figure S11 and S12).

Figure 5 shows the distribution of the long-term average PM<sub>2.5</sub> concentration in the study area during the two experiments inferred by our model. Clearly, there was a great spatial variation for the long-term PM<sub>2.5</sub> concentration.

Overall, the average PM<sub>2.5</sub> concentration was significantly higher in the winter experiment ( $72.8 \mu\text{g}/\text{m}^3$ ) than in the summer one ( $41.2 \mu\text{g}/\text{m}^3$ ). These results are validated by atmospheric studies for the same periods (winter:  $40\text{--}85 \mu\text{g}/\text{m}^3$ ; summer:  $23\text{--}57 \mu\text{g}/\text{m}^3$ ).<sup>40,44,45</sup> Such a seasonal pattern is also confirmed by atmospheric studies, mainly due to meteorological conditions and regional transport of air

pollutants.<sup>40,44,46</sup> For the winter experiment, the south and east areas were more polluted. This is consistent with long-term monitoring conducted in Beijing, which shows that administrative district Daxing (Southeast) had the highest average PM<sub>2.5</sub> concentration and Changping (Northwest) had a much lower concentration.<sup>47,48</sup> These results are also confirmed by multiple hypothesis tests ( $P < 0.0167$ )<sup>49</sup> and regression analyses (negative gradient from south to north).<sup>50</sup> During the summer, the average PM<sub>2.5</sub> concentration of different regions in Beijing was comparable with no significant differences, also confirmed by previous studies.<sup>51</sup> The central area had a slightly higher PM<sub>2.5</sub> concentration likely due to more traffic and heat island effects.<sup>52</sup> However, these previous atmospheric studies are based on point observations of stationary monitors, thus lacking the spatial resolution generated from our model. Our method can complement atmospheric studies to provide important spatiotemporal distribution of air pollutants at a resolution that cannot be differentiated in previous studies.

## DISCUSSION

The XGBoost model we developed in this study can infer the spatiotemporal distribution of urban air pollutants using large-scale mobile monitoring data obtained from equipped fleet vehicles. We apply the model in Beijing for PM<sub>2.5</sub> emissions in two 1 month experiments. Our results show that the model can infer the spatial distribution of PM<sub>2.5</sub> emissions at 1 km by 1 km resolution on an hourly average with an  $R^2$  value of 0.8 and an RMSE of 8.1  $\mu\text{g}/\text{m}^3$ . The resulting spatiotemporal distribution of PM<sub>2.5</sub> emissions clearly identifies emission hotspots and the transitions of hotspots throughout a day. We also show the superior performance of the model using the mobile monitoring data compared to without using it. Our results demonstrate the potential and necessity of mobile monitoring with low-cost air quality sensors for urban air quality monitoring with a high spatiotemporal resolution that is needed for effective air quality control measures.

Although similar studies exist on urban air quality mapping, we cannot make direct comparisons due to different local characteristics, sample size, and modeling methods. However, based on typical model performance metrics, our model is at least comparable, if not superior, to other models with similar purposes. Specifically, our model's  $R^2$  is 0.8, which is better than that of Zheng et al. (overall classification accuracy of 0.75),<sup>24</sup> Adam and Kanaroglou ( $R^2$  of 0.78, but without validation with stationary monitoring data or equivalent data),<sup>25</sup> and Lim et al. ( $R^2$  ranging from 0.63 to 0.80 but with a shorter period of mobile monitoring campaign of about 12 days),<sup>27</sup> similar to Song and Han ( $R^2$  ranging from 0.89 to 0.91 but with much less mobile monitoring observations collected from 15 vehicles).<sup>53</sup>

Our experience with the mobile monitoring campaign indicates the importance of reducing the need for drivers operating the sensors to improve data quality. The ideal case is to have fully automated sensors that do not require manual operations, are robust in extreme environmental conditions such as on rainy or windy days, and do not interfere with the regular operation of the vehicle.

The spatial and temporal coverage of the mobile monitoring data is the key to construct the machine learning model. Due to limited stationary monitors in our study, we extend the benchmarks to use transformed mobile monitoring data as a proxy for the ground truth. However, the spatial and temporal coverage is still not ideal given a limited number of equipped

vehicles and their relatively constrained operation time and areas due to range limits (electric vehicles). The majority of the study area only has a very limited number of observations, and about 20% of the area does not have any data. Additional ground truth information with accurate PM<sub>2.5</sub> monitoring data (either stationary or mobile) can greatly improve and further validate the model.

We recruited 260 electric vehicles in this study only based on budgetary and operational constraints. Theoretically, with more vehicles equipped with mobile sensors, our model can be further improved. However, it is not clear how many equipped vehicles are sufficient for a particular city. Future study should examine the relationship between the model performance and characteristics of the equipped vehicle fleet (e.g., number of vehicles, spatial and temporal coverage of the vehicle trajectories) to optimize the composition of the equipped vehicle fleet for the best model performance considering budgetary and operational constraints.

## ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.est.0c08034>.

Hyperparameters of the models, model evaluation metrics, summary of data used in this study, model performance for data transformation, study area and the location of fixed-site monitoring stations, device parameters, device installation and calibration, temporal characteristics for the mobile monitoring data, performance of the model only using stationary monitoring data, relative importance of the variables, and identified hotspots and other inference results ([PDF](#))

## AUTHOR INFORMATION

### Corresponding Author

Ming Xu – School for Environment and Sustainability and Department of Civil and Environmental Engineering, University of Michigan, Ann Arbor, Michigan 48109-1382, United States;  [orcid.org/0000-0002-7106-8390](https://orcid.org/0000-0002-7106-8390); Email: [mingxu@umich.edu](mailto:mingxu@umich.edu)

### Authors

Bu Zhao – School for Environment and Sustainability and Michigan Institute for Computational Discovery & Engineering, University of Michigan, Ann Arbor, Michigan 48109-1382, United States

Long Yu – Department of Statistics, Fudan University, Shanghai 200433, China

Chunyan Wang – School of Environment, Tsinghua University, Beijing 100084, China

Chenyang Shuai – School for Environment and Sustainability and Michigan Institute for Computational Discovery & Engineering, University of Michigan, Ann Arbor, Michigan 48109-1382, United States

Ji Zhu – Department of Statistics, University of Michigan, Ann Arbor, Michigan 48109-1107, United States

Shen Qu – School of Management and Economics and Center for Energy & Environmental Policy Research, Beijing Institute of Technology, Beijing 10081, China

Morteza Taiebat – School for Environment and Sustainability and Department of Civil and Environmental Engineering,

University of Michigan, Ann Arbor, Michigan 48109-1382,  
United States

Complete contact information is available at:  
<https://pubs.acs.org/10.1021/acs.est.0c08034>

## Notes

The authors declare no competing financial interest.

## ACKNOWLEDGMENTS

The authors acknowledge the financial support from the Michigan Institute for Data Science and invaluable feedback from the anonymous reviewers.

## REFERENCES

- (1) Li, X.; Jin, L.; Kan, H. *Air pollution: a global problem needs local fixes*; Nature Publishing Group: 2019.
- (2) Zhao, X.; Yu, X.; Wang, Y.; Fan, C. Economic evaluation of health losses from air pollution in Beijing, China. *Environ. Sci. Pollut. Res.* **2016**, *23*, 11716–11728.
- (3) World Health Organization *Ambient (outdoor) air pollution*; Available from: [https://www.who.int/news-room/fact-sheets/detail/ambient-\(outdoor\)-air-quality-and-health/](https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health/).
- (4) The Word Bank *Urban population*; Available from: <https://data.worldbank.org/indicator/SP.URB.TOTL>.
- (5) Carvalho, H. The air we breathe: differentials in global air quality monitoring. *Lancet Respir. Med.* **2016**, *4*, 603–605.
- (6) Kumar, P.; Morawska, L.; Martani, C.; Biskos, G.; Neophytou, M.; Di Sabatino, S.; Bell, M.; Norford, L.; Britter, R. The rise of low-cost sensing for managing air pollution in cities. *Environ. Int.* **2015**, *75*, 199–205.
- (7) Chong, C.-Y.; Kumar, S. P. Sensor networks: evolution, opportunities, and challenges. *Proc. IEEE* **2003**, *91*, 1247–1256.
- (8) Vardoulakis, S.; Gonzalez-Flesca, N.; Fisher, B. E. A.; Pericleous, K. Spatial variability of air pollution in the vicinity of a permanent monitoring station in central Paris. *Atmos. Environ.* **2005**, *39*, 2725–2736.
- (9) Hsieh, H.-P.; Lin, S.-D.; Zheng, Y. Inferring air quality for station location recommendation based on urban big data. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; IEEE: 2015.
- (10) New York State Department of Environmental Conservation *2020 ANNUAL MONITORING NETWORK PLAN-New York State Ambient Air Monitoring Program*; Elsevier: May 29th, 2020.
- (11) Boogaard, H.; Kos, G. P. A.; Weijers, E. P.; Janssen, N. A. H.; Fischer, P. H.; van der Zee, S. C.; de Hartog, J. J.; Hoek, G. Contrast in air pollution components between major streets and background locations: particulate matter mass, black carbon, elemental composition, nitrogen oxide and ultrafine particle number. *Atmos. Environ.* **2011**, *45*, 650–658.
- (12) Johnson, M.; Isakov, V.; Touma, J. S.; Mukerjee, S.; Özkaynak, H. Evaluation of land-use regression models used to predict air quality concentrations in an urban area. *Atmos. Environ.* **2010**, *44*, 3660–3668.
- (13) Marshall, J. D.; Nethery, E.; Brauer, M. Within-urban variability in ambient air pollution: comparison of estimation methods. *Atmos. Environ.* **2008**, *42*, 1359–1369.
- (14) Morello-Frosch, R.; Pastor, M.; Sadd, J. Environmental justice and Southern California's "riskscape": the distribution of air toxics exposures and health risks among diverse communities. *Urban Aff. Rev.* **2001**, *36*, 551–578.
- (15) Apte, J. S.; Messier, K. P.; Gani, S.; Brauer, M.; Kirchstetter, T. W.; Lunden, M. M.; Marshall, J. D.; Portier, C. J.; Vermeulen, R. C. H.; Hamburg, S. P. High-resolution air pollution mapping with Google street view cars: exploiting big data. *Environ. Sci. Technol.* **2017**, *51*, 6999–7008.
- (16) Gupta, P.; Christopher, S. A.; Wang, J.; Gehrig, R.; Lee, Y.; Kumar, N. Satellite remote sensing of particulate matter and air quality assessment over global cities. *Atmos. Environ.* **2006**, *40*, 5880–5892.
- (17) Martin, R. V. Satellite remote sensing of surface air quality. *Atmos. Environ.* **2008**, *42*, 7823–7843.
- (18) Christopher, S. A.; Gupta, P. Satellite remote sensing of particulate matter air quality: The cloud-cover problem. *J. Air Waste Manage. Assoc.* **2010**, *60*, 596–602.
- (19) Grell, G. A.; Peckham, S. E.; Schmitz, R.; McKeen, S. A.; Frost, G.; Skamarock, W. C.; Eder, B. Fully coupled "online" chemistry within the WRF model. *Atmos. Environ.* **2005**, *39*, 6957–6975.
- (20) Simpson, D.; Benedictow, A.; Berge, H.; Bergström, R.; Emberson, L. D.; Fagerli, H.; Flechard, C. R.; Hayman, G. D.; Gauss, M.; Jonson, J. E.; Jenkin, M. E.; Nyiri, A.; Richter, C.; Semeena, V. S.; Tsyro, S.; Tuovinen, J.-P.; Valdebenito, Á.; Wind, P. The EMEP MSC-W chemical transport model - technical description. *Atmos. Chem. Phys.* **2012**, *7825*.
- (21) Gibson, M. D.; Kundu, S.; Satish, M. Dispersion model evaluation of PM<sub>2.5</sub>, NO<sub>x</sub> and SO<sub>2</sub> from point and major line sources in Nova Scotia, Canada using AERMOD Gaussian plume air dispersion model. *Atmos. Pollut. Res.* **2013**, *4*, 157–167.
- (22) Zhang, L.; Yin, Y.; Chen, S. Robust signal timing optimization with environmental concerns. *Transp. Res. Part C: Emerging Technol.* **2013**, *29*, 55–71.
- (23) U.S. Environmental Protection Agency *CMAQ Models*; [cited 2020 July 19]; Available from: <https://www.epa.gov/cmaq/cmaq-models-0>.
- (24) Zheng, Y.; Liu, F.; Hsieh, H.-P. U-air: When urban air quality inference meets big data. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*; 2013.
- (25) Adams, M. D.; Kanaroglou, P. S. Mapping real-time air pollution health risk for environmental management: Combining mobile and stationary air pollution monitoring with neural network models. *J. Environ. Manage.* **2016**, *168*, 133–141.
- (26) Qi, Z.; Wang, T.; Song, G.; Hu, W.; Li, X.; Zhang, Z. Deep air learning: Interpolation, prediction, and feature analysis of fine-grained air quality. *IEEE Trans. Knowl. Data Eng.* **2018**, *30*, 2285–2297.
- (27) Lim, C. C.; Kim, H.; Vilcassim, M. J. R.; Thurston, G. D.; Gordon, T.; Chen, L.-C.; Lee, K.; Heimbinder, M.; Kim, S.-Y. Mapping urban air quality using mobile sampling with low-cost sensors and machine learning in Seoul, South Korea. *Environ. Int.* **2019**, *131*, 105022.
- (28) Schmitz, O.; Beelen, R.; Strak, M.; Hoek, G.; Soenario, I.; Brunekreef, B.; Vaartjes, I.; Dijst, M. J.; Grobbee, D. E.; Karssenberg, D. High resolution annual average air pollution concentration maps for the Netherlands. *Sci. Data* **2019**, *6*, 190035.
- (29) Hu, K.; Rahman, A.; Bhrugubanda, H.; Sivaraman, V. HazeEst: Machine learning based metropolitan air pollution estimation from fixed and mobile sensors. *IEEE Sens. J.* **2017**, *17*, 3517–3525.
- (30) Snyder, E. G.; Watkins, T. H.; Solomon, P. A.; Thoma, E. D.; Williams, R. W.; Hagler, G. S. W.; Shelow, D.; Hindin, D. A.; Kilaru, V. J.; Preuss, P. W. The changing paradigm of air pollution monitoring. *Environ. Sci. Technol.* **2013**, *11369*.
- (31) Williams, R.; Kilaru, V.; Snyder, E.; Kaufman, A.; Dye, T.; Rutter, A.; Russell, A.; Hafner, H. *Air sensor guidebook*; US Environmental Protection Agency: 2014.
- (32) Beijing Municipal Bureau of Statistics *Beijing Statistical Yearbook*, 2019; 2019.
- (33) China Daily *Beijing releases list of top air polluting sources*; 2018-05-15; Available from: <http://www.chinadaily.com.cn/a/201805/15/WSSafa22eca3103f6866ee8561.html>.
- (34) Brantley, H. L.; Hagler, G. S. W.; Kimbrough, E. S.; Williams, R. W.; Mukerjee, S.; Neas, L. M. Mobile air monitoring data-processing strategies and effects on spatial air pollution trends. *Atmos. Meas. Tech.* **2014**, *7*, 2169–2183.
- (35) Kheirbek, I.; Wheeler, K.; Walters, S.; Pezeshki, G.; Kass, D.; Matte, T. *Air pollution and the health of New Yorkers: The impact of fine particles and Ozone*; Environmental Protection: 2011.
- (36) Christidis, T.; Erickson, A. C.; Pappin, A. J.; Crouse, D. L.; Pinault, L. L.; Weichenthal, S. A.; Brook, J. R.; van Donkelaar, A.;

Hystad, P.; Martin, R. V.; Tjepkema, M.; Burnett, R. T.; Brauer, M. Low concentrations of fine particle air pollution and mortality in the Canadian Community Health Survey cohort. *Environ. Health* **2019**, *18*, 84.

(37) Chen, R.; Yin, P.; Meng, X.; Wang, L.; Liu, C.; Niu, Y.; Liu, Y.; Liu, J.; Qi, J.; You, J.; Kan, H.; Zhou, M. Associations between coarse particulate matter air pollution and cause-specific mortality: a nationwide analysis in 272 Chinese cities. *Environ. Health Perspect.* **2019**, *127*, No. 017008.

(38) Shi, L.; Zanobetti, A.; Kloog, I.; Coull, B. A.; Koutrakis, P.; Melly, S. J.; Schwartz, J. D. Low-concentration PM<sub>2.5</sub> and mortality: estimating acute and chronic effects in a population-based study. *Environ. Health Perspect.* **2016**, *124*, 46–52.

(39) Xing, Y.-F.; Xu, Y.-H.; Shi, M.-H.; Lian, Y.-X. The impact of PM<sub>2.5</sub> on the human respiratory system. *J. Thorac. Dis.* **2016**, *8*, E69.

(40) Xu, X.; Zhang, T. Spatial-temporal variability of PM<sub>2.5</sub> air quality in Beijing, China during 2013–2018. *J. Environ. Manage.* **2020**, *262*, 110263.

(41) Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*; 2016.

(42) December 2018 Weather in Beijing; Available from: <https://www.timeanddate.com/weather/china/beijing/historic?month=12&year=2018>.

(43) The Beijing News Press *The Beijing News*; . Available from: <http://www.bjnews.com.cn/>.

(44) Lv, D.; Chen, Y.; Zhu, T.; Li, T.; Shen, F.; Li, X.; Mehmood, T. The pollution characteristics of PM<sub>10</sub> and PM<sub>2.5</sub> during summer and winter in Beijing, Suning and Islamabad. *Atmos. Pollut. Res.* **2019**, *10*, 1159–1164.

(45) China National Environmental Monitoring Centre *National Urban Air Quality Report*; 2019.

(46) Zheng, G. J.; Duan, F. K.; Su, H.; Ma, Y. L.; Cheng, Y.; Zheng, B.; Zhang, Q.; Huang, T.; Kimoto, T.; Chang, D.; Pöschl, U.; Cheng, Y. F.; He, K. B. Exploring the severe winter haze in Beijing: the impact of synoptic weather, regional transport and heterogeneous reactions. *Atmos. Chem. Phys.* **2015**, *15*, 2969.

(47) Yan, D.; Lei, Y.; Shi, Y.; Zhu, Q.; Li, L.; Zhang, Z. Evolution of the spatiotemporal pattern of PM<sub>2.5</sub> concentrations in China—A case study from the Beijing-Tianjin-Hebei region. *Atmos. Environ.* **2018**, *183*, 225–233.

(48) Zhou, Y.; Cheng, S.; Chen, D.; Lang, J.; Wang, G.; Xu, T.; Wang, X.; Yao, S. Temporal and spatial characteristics of ambient air quality in Beijing, China. *Aerosol Air Qual. Res.* **2015**, *15*, 1868–1880.

(49) Huang, F.; Li, X.; Wang, C.; Xu, Q.; Wang, W.; Luo, Y.; Tao, L.; Gao, Q.; Guo, J.; Chen, S.; Cao, K.; Liu, L.; Gao, N.; Liu, X.; Yang, K.; Yan, A.; Guo, X. PM<sub>2.5</sub> spatiotemporal variations and the relationship with meteorological factors during 2013–2014 in Beijing, China. *PLoS One* **2015**, *10*, No. e0141642.

(50) Li, R.; Li, Z.; Gao, W.; Ding, W.; Xu, Q.; Song, X. Diurnal, seasonal, and spatial variation of PM<sub>2.5</sub> in Beijing. *Sci. Bull.* **2015**, *60*, 387–395.

(51) Yan, S.; Cao, H.; Chen, Y.; Wu, C.; Hong, T.; Fan, H. Spatial and temporal characteristics of air quality and air pollutants in 2013 in Beijing. *Environ. Sci. Pollut. Res.* **2016**, *23*, 13996–14007.

(52) Tian, G.; Liu, X.; Kong, L. Spatiotemporal Patterns and Cause Analysis of PM<sub>2.5</sub> Concentrations in Beijing, China. *Adv. Meteorol.* **2018**, 1–8.

(53) Song, J.; Han, K. Exploring Urban Air Quality with MAPS: Mobile Air Pollution Sensing. *arXiv* **2019**: arXiv: 1904.12303.