Title

Application of mobile measured high-resolution air pollution data in urban planning, health exposure, and economic impact study

Author

Tang, Minmeng

Publication Date

2021

Peer reviewed|Thesis/dissertation

Application of mobile measured high-resolution air pollution data in urban planning, health exposure, and economic impact study

By

MINMENG TANG
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Atmospheric Science

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Deb A. Niemeier, Chair

Christopher D. Cappa

Elena Craft

Committee in Charge

2021

DEDICATION

I dedicate this dissertation to my dear my family. A special feeling of gratitude to my loving parents and wife who support and encourage me with all the difficulties and challenges.

Abstract

Air pollution is a major global risk causing a large number of illnesses and deaths every year. Many literatures have shown the robust causal relations between health and outdoor exposure to various air pollutants. The complex urban environment causes the uneven distribution of air pollution concentrations, which can change sharply within a short distance. Therefore, understanding the within city air pollution gradients is crucial for various studies including exposure assessment, urban planning, air pollution monitoring, and environmental equity.

Mobile-based air pollution monitoring has been proposed to tackle this challenge since it can typically achieve higher spatial resolution measurement of air pollution concentrations than other methods. However, the inherently different nature of measurement from mobile monitors makes it difficult to apply methods designed for stationary sensors. This dissertation focuses on developing methods that are suitable for mobile sensor data to take advantage of the high spatial resolution nature for exposure assessment, air pollution monitoring, and socioeconomic impact studies.

For the exposure assessment study, we calculate exposure concentrations of traffic-related air pollutants with three different travel modes in the complex urban environment. We simulate bicycle, transit, and vehicle trips within Oakland CA. based on the local road network. With highly resolved mobile sensor data, we calculate the average concentration and the cumulative exposure of nitric oxide (NO), nitrogen dioxide (NO_2), and black carbon (BC) for each bicycle, transit, and vehicle trip that we simulate. The results show that cumulative exposure may be a better metric than the more typical average ambient concentration when evaluating the air

pollution exposure with different travel modes. The average concentrations of each trip are not significantly different among bicycle, transit, and vehicle. However, the cumulative exposure varies dramatically because it takes trip duration, route variations for different travel modes, and inhalation rates into consideration. Vehicle passengers tend to experience the lowest cumulative exposure, as well as have the lowest average per meter and per minute exposure. Because of the increased inhalation rates for bicyclists and longer trip duration for public transit users, they tend to experience higher cumulative exposure. Our study also compares the importance of trip duration and trip distance influencing exposure, which turns out that total trip duration is more influential than the total trip distance in terms of cumulative exposure. Our work finds better metrics to assess travel air pollution exposure by using big data and modern simulation techniques.

In another study, we combine the land use model with different regression methods to estimate black carbon (BC) concentrations in Oakland, CA. The regression methods used in this study include linear regression, Random Forest (RF), Support Vector Regression (SVR), and Neural Network (NN). The least absolute shrinkage and selection operator (LASSO), principle component analysis (PCA), conditional independence feature ordering (FOCI), and genetic algorithm (GA) are used for feature selection and dimension reduction of the SVR method to reduce overfitting and improve prediction accuracy. The tuning of RF and SVR are automatically conducted with the Bayesian Optimization method, while we manually tune the NN method. Among all these regression methods, RF performs the best with the highest prediction accuracy and robustness. Even though SVR shows much better prediction accuracy than linear and NN methods, the complex feature selection and dimension reduction processes make it less efficient than RF. NN

has the highest prediction accuracy on the train set, but the lowest accuracy on the independent validation set, which suggests an overfitting issue. With the one-factor-at-a-time (OAT) sensitivity analysis and localized hotspots identifications, our study shows that the LURs with a common approach are not efficient at identifying localized hotspots. However, LUR coupling RF can achieve higher air pollution prediction accuracy and robustness using mobile sensor measurements. This approach can be used in air pollution exposure assessment to more accurately identify vulnerable population groups or communities and better highlight environmental justice issues.

For the socioeconomic impacts of air pollution, we study the effects of air pollution on housing price in Oakland CA. We evaluate the ambient air quality on a parcel by parcel basis with the high-resolution mobile-based air pollution measurements of NO, NO_2 and BC. In this study, a hedonic price model is constructed with a spatial lag model and instrumental variable method to cover both spatial autocorrelation and endogeneity effects between air pollution concentrations and housing price. The results indicate the air pollution influences housing price positively, which is surprising. The results could be explained in two ways: people are not sensitive to air pollution when the overall ambient air quality is good; the low variability of air pollution concentrations leads to false positive results. The explanations could be verified with the high-resolution mobile-based air pollution measurements covering more diversified regions.

Table of Contents

List of Figures

List of Tables

Chapter 1. Introduction

There is little question that air pollution is a major global risk resulting in high incidences of illness and deaths.[1,2] A large number of epidemiological studies have established robust causal relations between health and outdoor exposure to ultrafine particulate matter (PM), black carbon (BC), carbon monoxide (CO), oxides of nitrogen, including nitric oxide (NO), and nitrogen dioxide (NO_2).[2–5]

Urban air pollutant concentrations are unevenly distributed and can vary dramatically even within short distances.[6–8] The spatial variations in air pollutant concentrations can be as large as the contrast between cities.[9] Epidemiological studies clearly show that within-city PM exposure is larger than the between-city effect.[10] One of the critical gaps in our understanding is how to characterize within city air pollutant concentration gradients, which is crucial for exposure assessment[11], urban planning[12,13], air pollution monitoring[14], and environmental equity[15].

Federal law authorizes the Environmental Protection Agency (EPA) to establish National Ambient Air Quality Standards (NAAQS) for six air pollutants, including CO, lead (Pb), NO_2, sulfur dioxide (SO_2), ozone (O_3), and PM. Each state is required to develop State Implementation Plans (SIPs), which demonstrate how the state will meet the NAAQS by reducing mobile, industrial and area pollutant sources. California's air quality standards are generally more stringent than the national standards and include four additional pollutants: hydrogen sulfide (H_2S), sulfate, vinyl chloride, and visibility reducing particles.[16] An area that routinely exceeds the state or federal air quality standards is referred as nonattainment area, and an attainment plan is required as part of the SIP to be reviewed and approved by California Air Resource Board (CARB) and EPA. The size of

the designated area is typically either at air basin or county level, both of which clearly do not capture air pollutants variations on the order of several hundred meters or even at the urban scale.

The limitation in spatial resolution constrains within area air pollution exposure assessment. The most common approaches for assessing intra-city air pollution related health exposure use geo-statistical methods, including among others inverse distance weighting (IDW) methods and Kriging,[17–19] Land Use Regression (LUR),[20–22] chemical transport models,[23,24] and remote sensing.[25] Each of these approaches has distinct advantages and limitations (Table 1).

Geo-statistical and LUR methods require a large number of stationary monitoring sites to ensure model estimation or prediction accuracy. The chemical transport models require accurate measurement of meteorological parameters, topography information, and significant computational resources to achieve high spatial resolution. Dispersion models, like AERMOD, can achieve higher spatial resolution than the regional models, but these models can only be applied for near field (<50 km) assessment and they are not able to represent complex photochemical reactions[26]. It is difficult to characterize fine-scale ground level gradients, which relate closely with health exposure, with remote sensing data and the cloud cover issue can also reduce the accuracy of remote sensing techniques.

Table 1. Typical spatial resolution of different types of techniques.

Technique type	Multiscale regional Models (CMAQ[a], WRF[b])	Dispersion models (AERMOD[c])	Remote sensing methods	Google car measurement
Spatial resolution	1km - 4 km[27,28]	> 30 meters[29]	250 meters – 1 km[30]	~ 30 meters[31]

Note: [a]CMAQ: Community Multi-scale Air Quality model; [b]WRF: Weather Research and Forecasting model; [c]AERMOD: American Meteorology Society/Environmental Protection Agency Regulatory Model.

Air pollution is monitored and regulated by the stationary monitoring sites operated by federal and state authorities; these fixed monitors provide high accuracy and reliable data. However, these stationary sites not only occupy a large area but also have high construction and maintenance costs,[32] which limits the number of stationary sites present in urban areas. According to Apte et al. estimation, the mean number of monitoring stations per million people and 1000 km^2 is between two and five stations for 60% of the U.S. census urban areas. Hence, the current stationary sites are insufficient for capturing intra-urban air pollution gradients accurately.

With the development of high accuracy portable pollution sensing instruments and Global Positioning System (GPS) technology, the use of vehicles for mobile air pollution monitoring has been proposed to tackle some of the challenges of stationary monitoring sites. These mobile sensors can typically achieve high spatial resolution air pollution measurement, but can rarely be deployed for extended periods of time for a variety of reasons.[33,34] Limited exploration of mobile sensing systems has been conducted in urban areas for air pollution exposure assessment and health effects studies.[35] There is a trade-off between the high spatial resolution that these devices can achieve and the cost of temporal resolution. Most researchers apply methods designed for stationary sensor networks to process the mobile data, even though these methods are not always suited because of the inherently different nature of spatiotemporal observations from mobile monitors.[36–38] **One of the critical needs is the development of methods that transition and harmonize stationary sensor networks with the kinds of mobile sensor data that**

can now be collected.[32,39] The transition between these two inherently different data producers brings up issues associated with the study of transferability between two sensor types (stationary and mobile) and how data from each of these sensors can and should be used for long-term pollutant estimation (e.g., for transportation).

Chapter 2. Research Objectives

In this dissertation, we focus on the potential application of the new mobile sensor data to better identify health exposure assessment, spatial resolution refinement for the existing air pollution monitoring network, and socioeconomic impacts from air quality degradation. Specifically, we are interested in,

Using big data techniques to better understand high-Resolution Cumulative Exposure Assessment of Traffic-Related Air Pollution (Chapter 3);

In this chapter, we calculate the cumulative exposed mass of traffic-related air pollutants with three different traffic modes in the urban environment, and compare the result with the conventional exposure assessment method. With the vehicle equipped mobile sensors, we are able to obtain the high-resolution air pollution concentrations in the complex urban environment. With detailed information about air pollution distributions, it is possible to develop more generalizable methods, which can assess traffic exposure more comprehensively and accurately than conventional methods. This work also builds the framework to achieve real-time personal exposure assessment for future studies. The work of this chapter has been published in *ES&T Engineering*, volume 1, No. 3, P 436-446.

Comparing Machine Learning, Deep Learning, and Land Use Regression Outcomes Using Google Street Mobile Source High-Resolution Black Carbon Concentrations in Oakland, California (Chapter 4);

In this chapter, we couple land use model (LUR) with machine learning and deep learning models to estimate air pollution concentrations. With thorough and complicate feature selection and

model tuning algorithms, we build Random Forest (RF), Support Vector Regression (SVR), and Neural Network (NN) models to predict air pollution concentrations at unmeasured locations. With this work, we are able to thoroughly explore the predictive ability of the LUR model over a varied urban landscape. Meanwhile, the modeling performance in hyper-localized air pollution prediction of the advanced methods are evaluated comprehensively and systematically. The result of this work is suggestive and helpful for air pollution epidemiology modeling, health exposure assessment, and other studies that require air pollution concentration measurement or estimation. The work of this chapter has been submitted to *Environmental Science and Technology*.

How Air Pollution Influences Housing Price in Bay Area? (Chapter 5);

In this chapter, we study the effects of localized air pollution concentrations on the housing price market in Oakland, CA. we construct the hedonic price model, which combines the spatial lag model with the instrumental variable model to cover both the spatial autocorrelation and endogeneity effects between air pollution concentration and housing price. This work fill the gap in literatures that considering both spatial autocorrelation and endogeneity effects simultaneously in the study of relation between housing price and air pollution concentration. Furthermore, our work is also the first to introduce the high-resolution air pollution mapping information into the housing valuation studies, which uses the real measurement of air pollution concentrations of every parcel instead of estimations from limited number of stationary monitors. The work of this chapter has been submitted to *International Journal of Environmental Research and Public Health*.

Published in *ES&T Engineering*, volume 1, No. 3, P 436-446.

Chapter 3. Using big data techniques to better understand high-Resolution Cumulative Exposure Assessment of Traffic-Related Air Pollution

*AUTHOR NAMES: Minmeng Tang[1], Deb A. Niemeier[2]**

AUTHOR ADDRESS: [1]Department of Land, Air, and Water Resources, University of California, Davis, One Shields Ave. Davis, CA, 95616, USA

[2]Department of Civil and Environmental Engineering, University of Maryland, 1173 Glenn Martin Hall, College Park, MD. 20742, USA

KEYWORDS: trips, travel modes, exposure, air pollution, route selection

Abstract

In this paper, we calculate exposure concentrations of traffic-related air pollutants for different travel modes in the urban environment. Using recent high-resolution mobile sensor measured air pollution concentration data, we simulate bicycle, transit and vehicle trips within Oakland CA and calculate exposure concentrations for nitric oxide (NO), nitrogen dioxide (NO_2), and black carbon (BC). We draw on highly resolved sensor data (on the order of seconds) collected by Aclima and Google, which was then aggregated to the annual median for every 30-meter road segment in the study area by Apte et al. (2017). For each bicycle, transit and vehicle trip that we simulate, we calculate the average concentration and the cumulative exposure for all three pollutants. The cumulative exposure is calculated as the total mass of pollutants inhaled during a trip. Our results show that cumulative exposure, rather than the more typical average ambient

concentration, may be a better metric for assessing travel pollutant exposure. For all three travel modes, the average concentrations of each trip are not significantly different. When we account for trip duration and route variations for different travel modes and inhalation rates, the cumulative exposure varies dramatically. Cumulative exposure for vehicle passengers tends to be the lowest, as well as having the lowest average per meter and per minute exposure. Bicyclists and public transit users tend to experience higher cumulative exposure due to increased inhalation rates for bicyclists and longer trip duration for public transit users. Last but not least, our study shows that total trip duration is more influential than total trip distance when estimating air pollution exposure. Our use of big data and modern simulation techniques point toward better metrics for assessing air pollutant exposure.

1. Introduction

Traffic emissions are a complex mixture of particles and gaseous compounds including ultrafine particulate matter (PM), black carbon (BC), carbon monoxide (CO), nitric oxide (NO), and nitrogen dioxide (NO_2). The causal relationship between health and outdoor air pollutant exposure is well established [2–5]. Proximity to traffic emissions contributes disproportionally to overall air pollution exposure, and exacerbates health effects [40]. Vehicle commuters tend to experience higher levels of air pollutant exposure as a result of traffic proximity, while active commuters (i.e., walking or cycling) inhale larger doses of air pollutants due to increased inhalation rates and longer commuting time [41].

Urban areas tend to experience higher traffic related air pollution [42,43] and the spatial variations can be very significant, even within short distances [8]. The differences in air pollutant

concentrations within a region can be as large as the contrast between cities [9]. Epidemiological studies also clearly show that within-city PM exposure is larger than the between-city effect [10] and that different traffic conditions will exacerbate exposure levels [44]. Using California bay area as an example, two monitors (Laney College, Oakland and Knox Ave, San Jose) in the Bay Area Air Quality Management District (BAAQMD) are used to compare NO, NO_2, and BC concentration variations between Oakland and San Jose in year 2015. For NO and NO_2, the annual averages are the same for monitors in Oakland and San Jose; however the monthly maximum concentrations differ from 841 ppb in Oakland to 219 ppb in San Jose for NO and from 72 ppb in Oakland to 61 ppb in San Jose for NO_2 [45]. The variations of BC concentration are even larger: the annual averages are 1.4 $\mu g/m^3$ and 1 $\mu g/m^3$ for monitors in Oakland and San Jose; while the monthly maximum values are 41.1 $\mu g/m^3$ and 20.8 $\mu g/m^3$ for the corresponding locations [45].

The research is mixed, however, as to which travel modes and under what conditions, pollutant concentration exposure pathways will be elevated. During a 2013-2014 Lisbon, Portugal air pollution campaign, car drivers and bus passengers experienced higher pollutants concentrations (PM, volatile organic compound, carbon dioxide, CO, and ozone) than bicyclists [46]. In contrast, a study in Colorado, United States found that bicyclists were exposed to higher BC and fine PM, but lower CO concentrations than driving [12]. The other two studies, however, found there are no significant differences in average concentrations among different travel modes [47,48]. One study in New Zealand compares students' exposure while walking from home to school along two route options including: the shortest distance route and an alternative lowest exposure route; they conclude that taking the alternative lowest exposure route is important to reduce exposure, especially when the shortest distance route overlaps or runs parallel to an arterial road [49].

The literature is mixed partially because of the variations in location, route selection, and exposure assessment methods used in the various studies. Differences in local road systems and urban planning strategies between different cities can also lead to contradictory findings in daily travel mode exposure studies. Most studies also rely on limited number of trips with a fixed route to measure air pollution exposure under different travel modes.[48,50,51] Measuring exposure by different travel modes on the same (often single) route obviously does not reflect the kind of exposure that might be experienced under actual daily travel activity. Finally, most studies also often rely on average regional or monitor level air pollution concentrations to represent traffic exposure when the local trip duration and breathe rate should be factored into inhalation concentrations [41,48].

With vehicles equipped with mobile sensors we are now able to measure high-resolution air pollution concentration data. With these data, it is possible to develop more generalizable approaches to more comprehensively assess traffic exposure with greater accuracy. In this paper, we use continuous sensor data and simulated trips to assess exposure under different traffic conditions and modes. Furthermore, we evaluate the performance of using average concentrations and cumulative exposed mass as an exposure assessment metric. Based on our work, cumulative exposed mass, rather than average concentrations, is more appropriate to assess exposure among different travel modes. Bicyclists and public transit users tend to experience higher cumulative exposures because of increased inhalation rates and the longer trip duration, respectively.

2. Materials and Methods

2.1 Study Area

We include three major areas within Oakland, CA: West Oakland (WO), Downtown Oakland (DO), and East Oakland (EO) (Figure 1). The WO area is about 10 km^2 with residential and industrial blocks, surrounded by three major interstate highways (I-880, I-980, and I-580), and the 9th-largest container port in the U.S. (Port of Oakland). The DO area is a mixture of residential and commercial blocks that encompasses about 5 km^2. The EO area is separated from WO and DO, and EO is connected with WO and DO by interstate highway I-880 and I-580. EO has a mix of industrial and residential areas of about a 15-km^2 area. These three areas are transected by a highly concentrated system of roads, connecting San Francisco and the east bay shoreline, which are among the most densely populated areas in California [52]. One way commutes in the Bay Area are approximately 32 minutes, on average; this is the 5th-longest commute time for US metro regions [53].

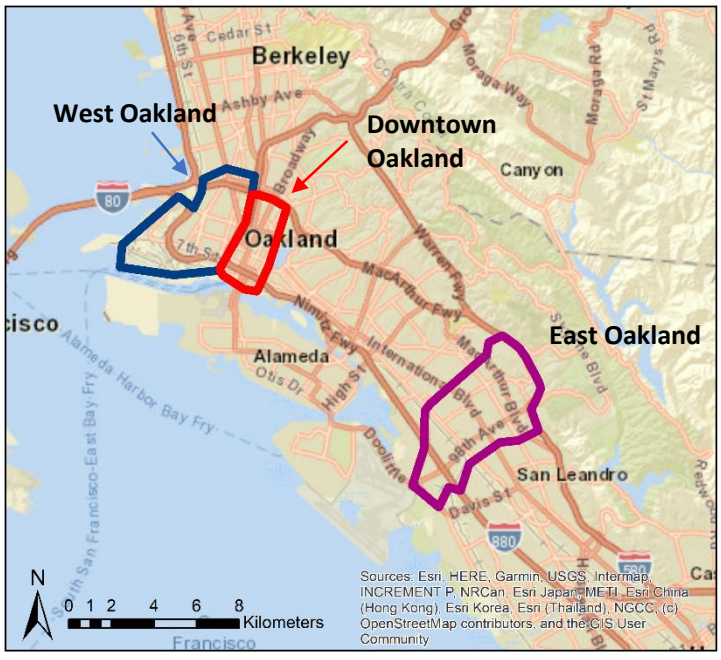

Figure 1. Study domain.

2.2 Pollutant concentration data

Between June 2015 and May 2016, two Google street view mapping vehicles were equipped with Aclima environmental intelligence sensors and a data integration platform, which provides fast-response and laboratory-grade air pollution measurements (Aclima, Inc., San Francisco). Two inlets were collocated and positioned in forward-facing direction, which were used to sample aerosols and gases, respectively. The inlets were installed several inches above the roof line at the rear edge of the front window to minimize the self-emission influences. The post-installation tests have proven that the self-sampling did not occur in most circumstances. More details about the sampling instrument layout and post-installation tests design are available in Apte et al's supporting information [6]. The vehicles repeatedly measured weekday daytime concentrations of BC, NO, and NO_2 in Oakland, CA. [54]. Apte et al. applied a data reduction and aggregation algorithm to convert these data of about 3 million instantaneous observations into estimates of median annual weekday concentrations for about 16,500 different 30-m roadway segments within the study domain [6]. We use the high-resolution concentration data from Apte et al's supporting information without any pre-process procedures as the air pollution concentration measurements for our study.

2.3 Trip and route generation

To create trips we randomly sample two points from our 30m roadway segment data and designate one point an origin (O) and one point a destination (D). We connect 30m roadway segments between the ODs and estimate exposure using the simulated trips. We constrain the trip ends to be within the study areas of WO, DO, and EO. To identify the path (or route) between

endpoints, we calculate a great circle distance using the haversine formula [55] and the 'geosphere' package in R [56,57]. The great circle distance produces the shortest path between two points. To be more conservative, we regard trips with great circle distance of less than 1 km as walking trips. Based on the 2009 National Household Travel Survey, the mean and median walking distance is 1.1 km and 0.8 km, respectively [58]. We only use those trips with great circle distance longer than 1 km for exposure analysis under auto, bicycle, and transit modes.

We create great circle distance simulations of 1,000 to 50,000 trips (Figure 2). We use a Chi-square distance to compare the similarity between histograms with different numbers of simulated trips; the equation to calculate two normalized histograms is:

$$\chi^2(H_1, H_2) = \sum_i \frac{(H_{1i} - H_{2i})^2}{H_{1i} + H_{2i}}$$

where H_{1i} and H_{2i} are normalized density values of the pairwise bins in two histograms. The Chi-square distance between histograms using 1,000 and 5,000 trips is 2.3×10^{-4}; between 5,000 and 10,000 trips is 1.8×10^{-5}; and between 10,000 and 50,000 trips is 2.2×10^{-5}. Based on the Chi-square distances, increasing our simulations from 1,000 to 5,000 leads to the largest change in the histogram distribution, but any further increases in simulations do not significantly improve the histogram representation. That is, histogram distributions with 5,000, 10,000, and 50,000 trips are relatively stable, so we use a simulation of 5,000 trips to represent the spatial distance pattern within our study domain. The bimodality in the distributions arises from the spatial proximity of our study areas. The WO and DO areas are co-located, while EO area is separated from WO and DO. For our exposure estimations, we regard WO and DO together as one area, while considering EO as separate area. The intra-area trips contribute to the first peak (less than

5 km) in distance distribution, while inter-area trips contribute to the second peak (between 10 km and 15 km). Our bimodal distance distribution is very useful for estimating cumulative exposure of different travel modes (e.g., short trips versus longer trips).

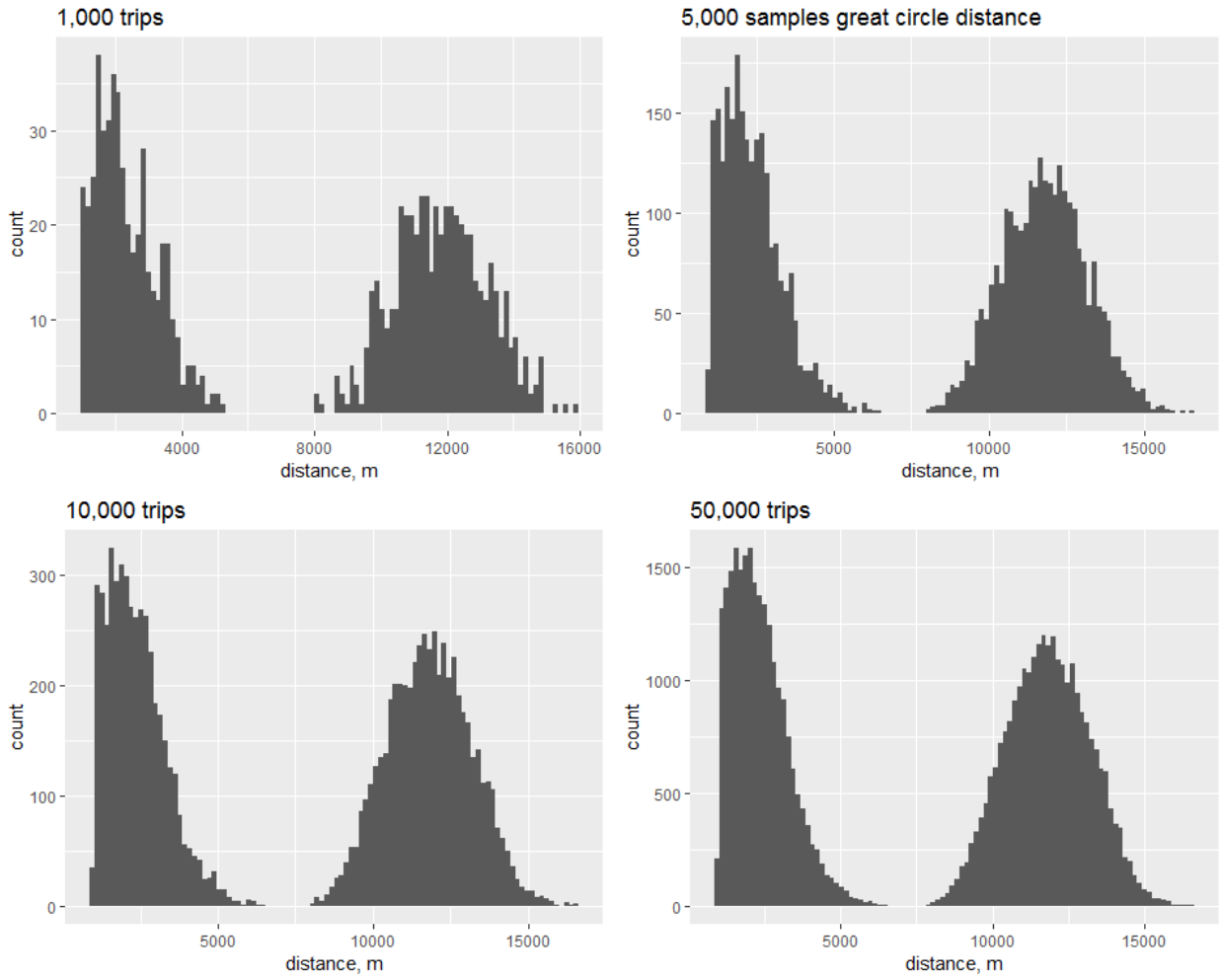

Figure 2. Great circle distance distribution with different number of trips generated.

We use the Google map API to calculate the route between trip ends [59,60]. We specify three different travel modes for each API trip request: auto, bike, and transit, where transit refers to the public transportation systems including bus and light rail. For each trip, we deploy two Google route options: shortest distance and shortest duration. In total, we generate 30,000 routes for

the 5,000 random trips for three different travel modes and two different route features. To assign cycling trips, we use a maximum range of 10 km, which is between an easy 5-mile bike commute [61] and an approximate maximum bike commute of 10 miles [62]. After eliminating trips longer than 10 km, we have a total of 2,332 bike trips. There are 72 trips that are not available with public transit, which lead to a total of 4,928 transit trips.

Figure 3 shows the distributions of different route options by travel mode. The black in Figure 3a reflects auto, red reflects transit, and blue reflects bicycle routes. Figure 3b and 3c show similar bimodal distributions as Figure 3a. Most trips have distance less than 25 km for both auto and transit options. Figure 3d shows the distribution of the shortest duration trips; this clearly shows that driving and bicycling trips generally exhibit shorter duration and transit trips are generally of longer duration.

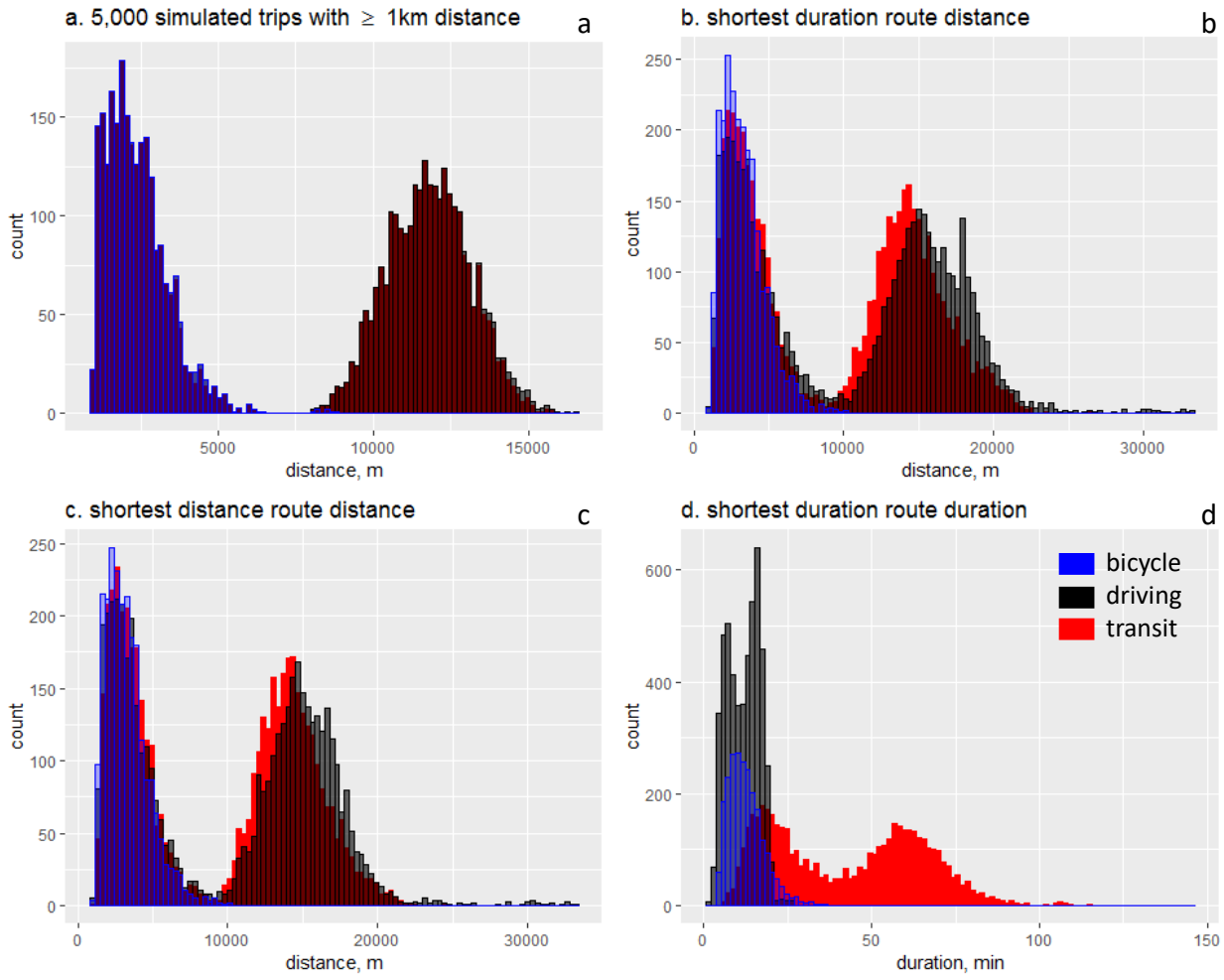

Figure 3. Trip distance and duration distributions by different travel modes. (Note: in part a, transit and driving overlap since they share very similar distributions of simulated trips)

2.4 Exposure assessment

To calculate the cumulative exposure for each trip by different travel modes, we use,

$$cumulative\ exposure\ of\ trip\ j\ with\ traffic\ mode\ k = \frac{\sum_{i=1}^{N_{j,k}} C_{i,j,k} \cdot V_k \cdot T_{j,k}}{N_{j,k}}$$

where j represents a simulated trip and k indicates travel mode, including bicycle, auto, and public transit. $C_{i,j,k}$ is the air pollutant concentration measurement for each 30-meter road

segment within trip j by travel mode k; V_k is the ventilation rate for travel mode k; $T_{j,k}$ is the total duration of trip j for travel mode k; and $N_{j,k}$ is the total number of 30-meter road segments in trip j with travel mode k. We employ estimated ventilation rates based on heart rates. On average, cyclists have a ventilation rate of 23.5 L/min, vehicle users have a rate of 11.8 L/min, and transit users have a rate of 12.7 L/min [63]. Our cumulative exposure calculation estimates the total mass of air pollutants inhaled by a traveler on each route by different travel modes, and considers the critical factors of breathe rate variations and route length and duration.

Finally, we note that some of our trips are inter-area trips, with routes passing through roadway segments linking DO and EO for which we have no concentration observations. To estimate air pollution concentrations on these roadway segments, we bootstrap our resampling method to increase estimation accuracy. We classify both measured and unmeasured roadway segments into four categories: highway, major arterial, residential, and residential within a 400-m distance from the highway. Meta-analyses have clearly shown that the spatial extent of mobile sources is on the order of 100-400 meters for PM and 200-500 meters for NO_2 [8,14]; therefore, we use a 400-m threshold to identify whether residential roadways are influenced by highway generated pollutants. For the unmeasured inter-area roadway segments in each travel mode category, we randomly assign the air pollutant concentrations from the corresponding category of measured road segments. We repeat this random assignment 1,000 times and use the average concentrations as the estimate of the unmeasured road segments.

3. Results and discussion

3.1 Average exposure concentration

We calculate the average air pollutant exposure concentrations for each route (Figure 4). The top three panels show the average exposure concentrations of NO, NO_2, and BC for the shortest distance routes; the lower panels show the average exposure concentrations of NO, NO_2, and BC for the shortest duration routes. Since NO, NO_2, and BC are all common pollutants in traffic related emissions, we expect them to be highly correlated, and the NO, NO_2, and BC plots look very similar. Figure 4 indicates that the average exposure concentrations of the three different travel modes are within similar range and overlap with each other for route distance less than 10 km. For mid-range distance (10 km – 35 km), transit users tend to experience higher concentrations than auto drivers. Using NO concentrations as an example, the mean concentration difference between transit and driving in mid-range distance is 6.53 ppb ± 0.17 ppb for shortest distance routes with 90% confidence, and 6.63 ppb ± 0.17 ppb for shortest duration routes with 90% confidence. However, it is worth noting that the average exposure concentrations for transit and driving still overlap at some distances/durations.

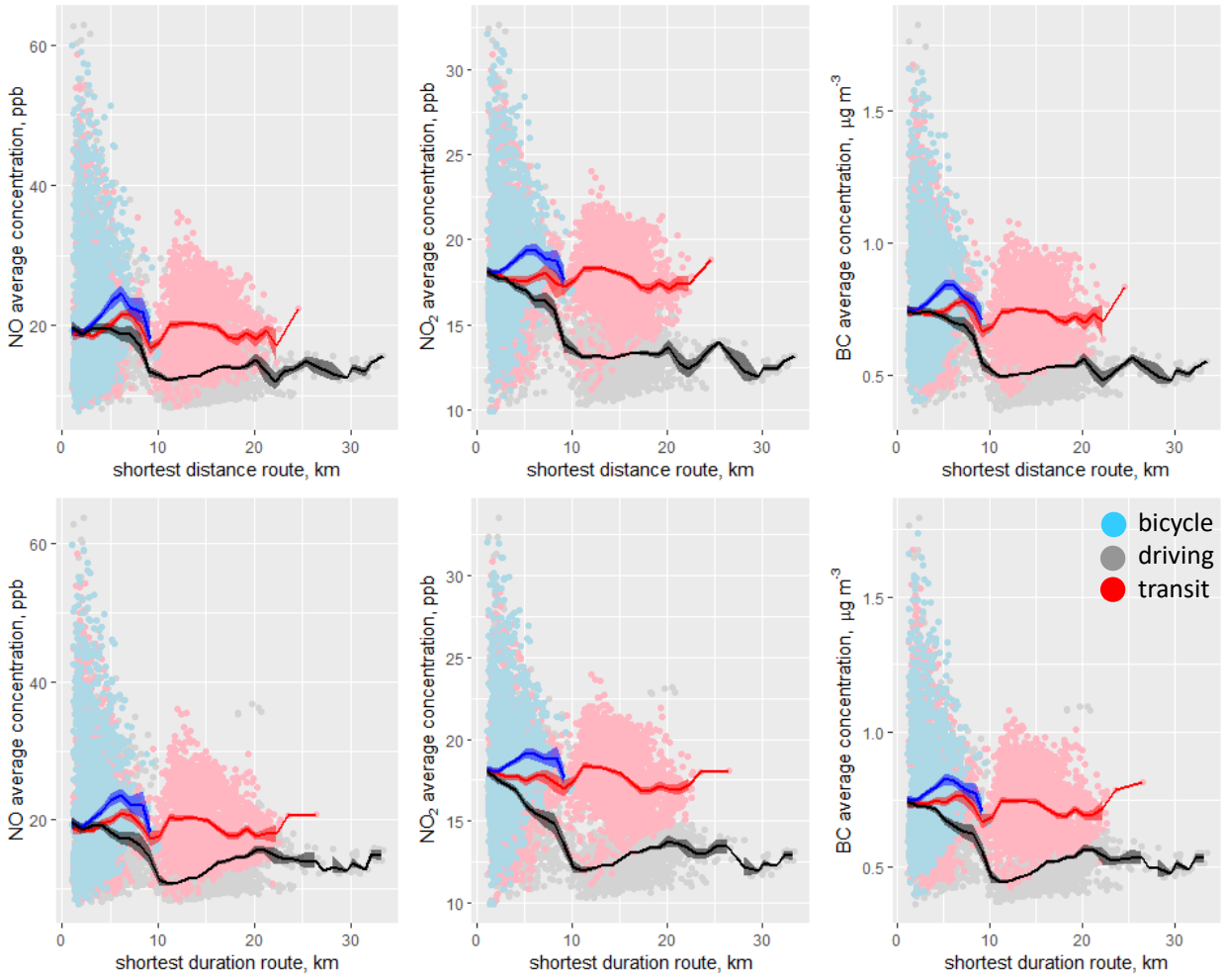

Figure 4. Average exposure concentrations of NO, NO2, and BC change with route distance (lines and shaded areas are the mean and 90% confidence interval for every 1 km distance interval).

When we look at duration (Figure 5), the longest duration routes for driving and bicycle require less than 30 minutes and 40 minutes, respectively. The average concentrations of the three travel modes clearly overlap between each other within similar duration ranges. This suggests that exposure does not significantly differ between travel modes.

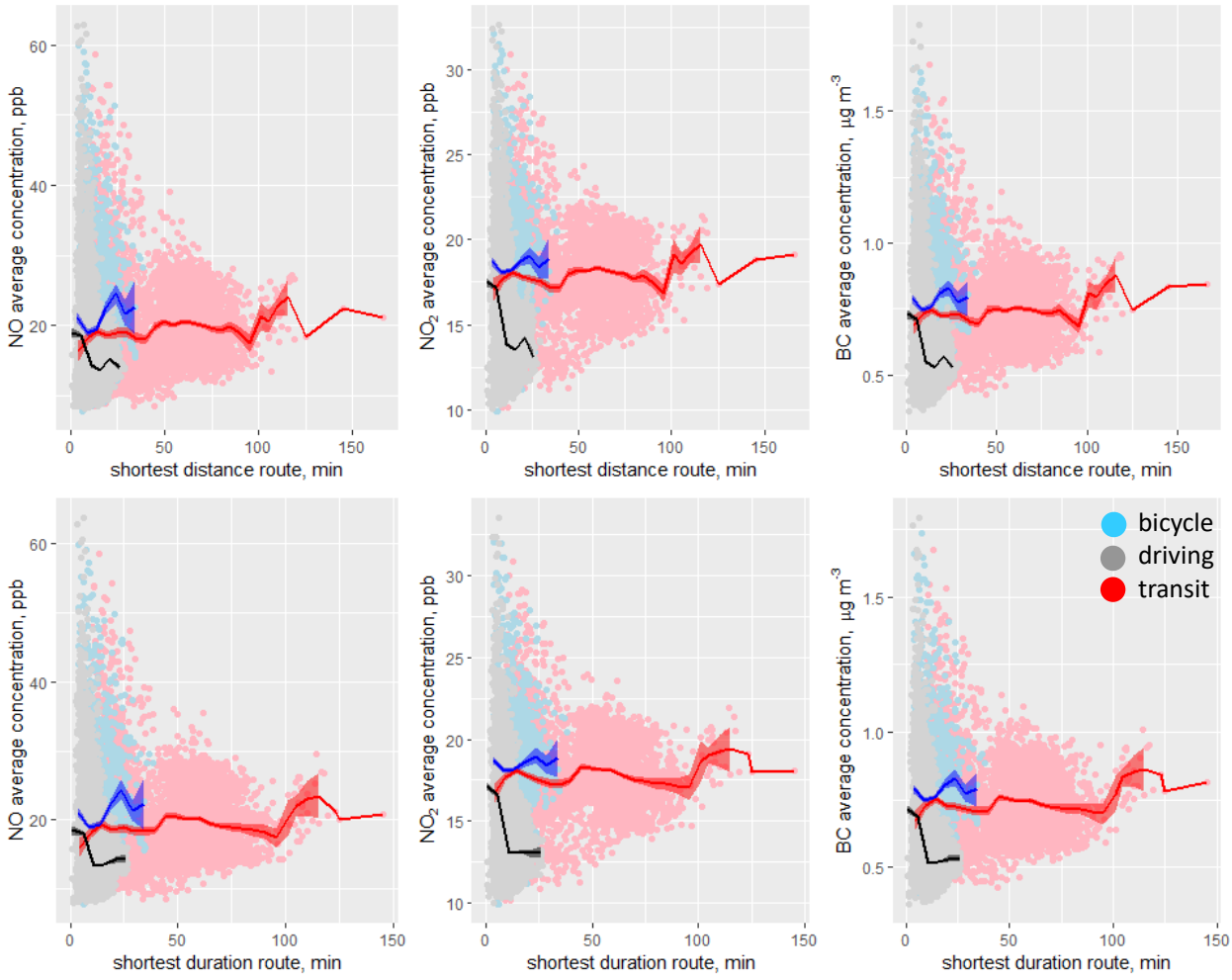

Figure 5. Average exposure concentrations of NO, NO2, and BC change with route duration (lines and shaded areas are the mean and 90% confidence interval for every 5-minute duration interval).

To compare the average exposure concentrations in more detail, we look at the scatterplot of the average NO exposure concentrations for the three travel modes (Figure 6). There are no obvious differences in NO average concentrations among different travel modes. The trends are consistent with Figure 4 and 5, which also shed new insight on the conflict in the literature on travel exposure assessment that we discussed in the introduction. Since most of the literature uses a limited number of trips to assess exposure concentrations, it may be the specificity of

those trips used in the literature that leads to higher exposure concentrations for one travel mode versus others.

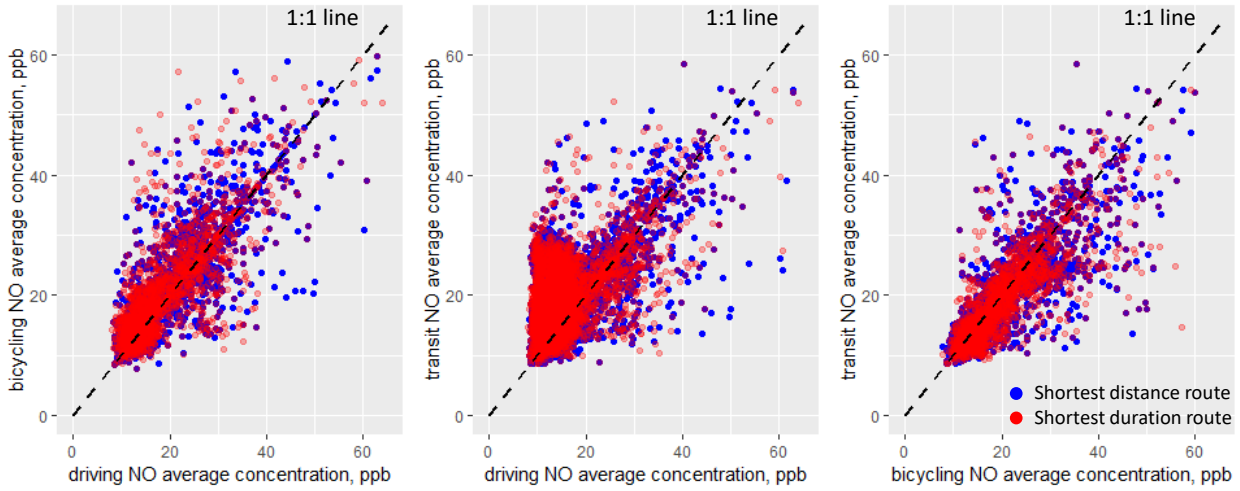

Figure 6. Comparison of average NO exposure concentrations among three travel modes.

3.2 Cumulative exposure

The cumulative exposure against route distance and duration are shown in Figures 7 and 8, respectively. Bicycling and driving have a not unexpected linear relationship to route distance and duration. Transit is clustered with respect to route distance, which is likely caused by the availability of bus and light rail. For example, transferring between buses will produce different patterns of trip duration and distance compared with direct route buses. Most buses also travel on arterial and residential roads (versus highways), which means the cumulative exposure of the transit mode tends to be within driving and bicycling exposure ranges. Driving has the lowest per meter or per minute cumulative exposure, while bicycling has the highest per meter or per minute cumulative exposure. Transit is much closer to bicycling for per meter cumulative exposure, both of which are much larger than driving. However, for per minute cumulative exposure, transit is much closer to driving, and both of them are much lower than bicycling.

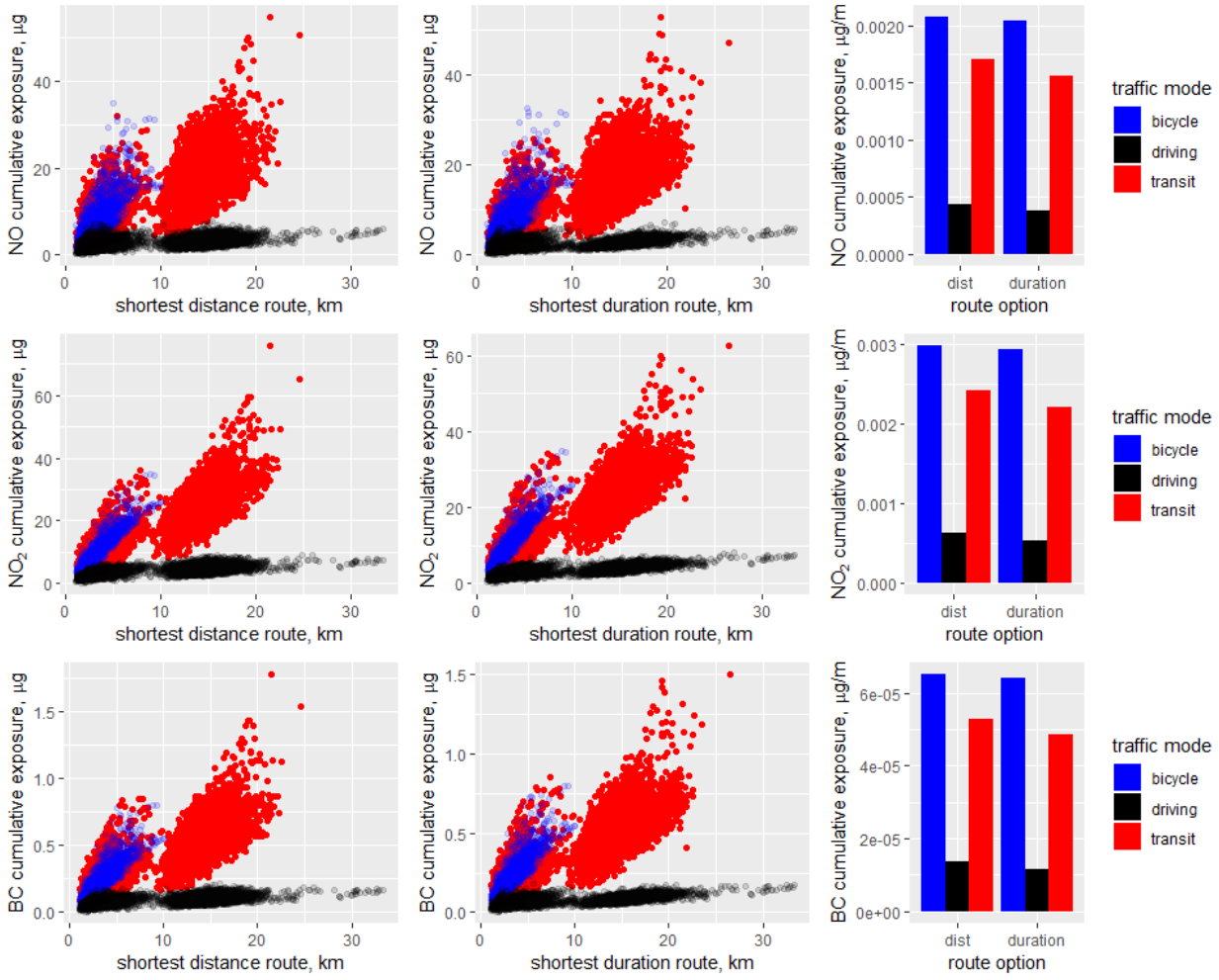

Figure 7. Cumulative exposure of NO, NO2, and BC changes with route distance.

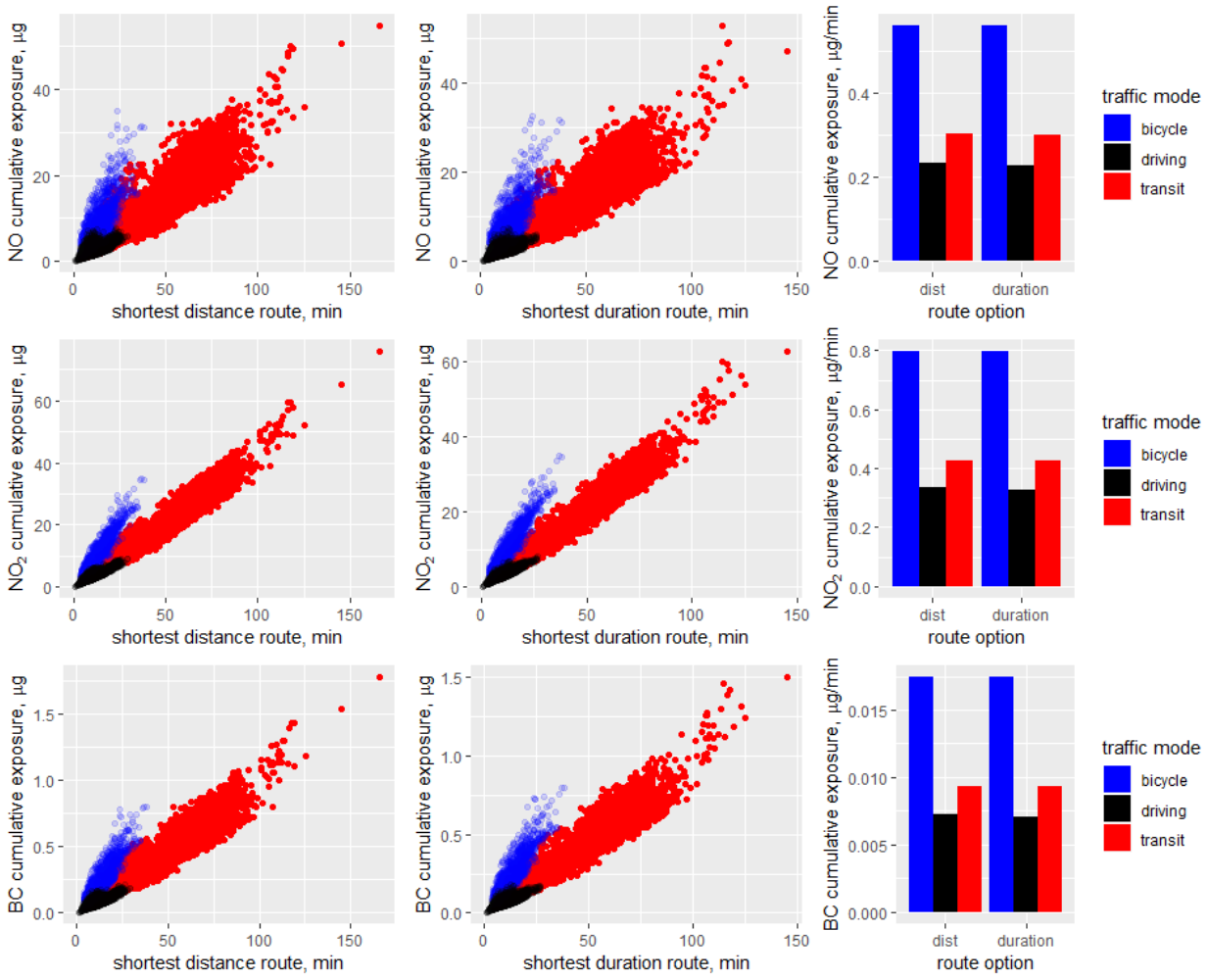

Figure 8. Cumulative exposure of NO, NO2, and BC changes with route duration.

Since the high correlations among three pollutants, in Figure 9 we only show the cumulative exposure of NO in order to compare cumulative exposure for the same trip with different travel modes. Other pollutants share similar trends with NO. From Figure 9, it is very clear that driving has the lowest cumulative exposure among all three travel modes, and there are no clear differences between bicycling and transit. Although they have different per minute and per meter cumulative exposure, bicycling trips tend to have shorter distance and duration than transit, which offset these differences and lead to similar cumulative exposure for each trip. This result is consistent with the scant literature that does use cumulative exposure to compare

among different travel modes: active commuters inhale higher doses of pollutants due to increased inhalation rates and commute time [41]. From our result, not only is the cumulative exposure of bicycling higher than driving, we have shown that per minute and per meter cumulative exposure of bicycling is also higher (Figure 7 and 8). The per minute cumulative exposure of bicycling is about 2.4 times higher than driving, while the ventilation rate of bicycling used in cumulative exposure calculations is only 1.99 times higher than driving. Therefore, the increased ventilation rate of bicycling does not by itself increase exposure, the bicycle route options also increase the cumulative exposure doses when compared to driving.

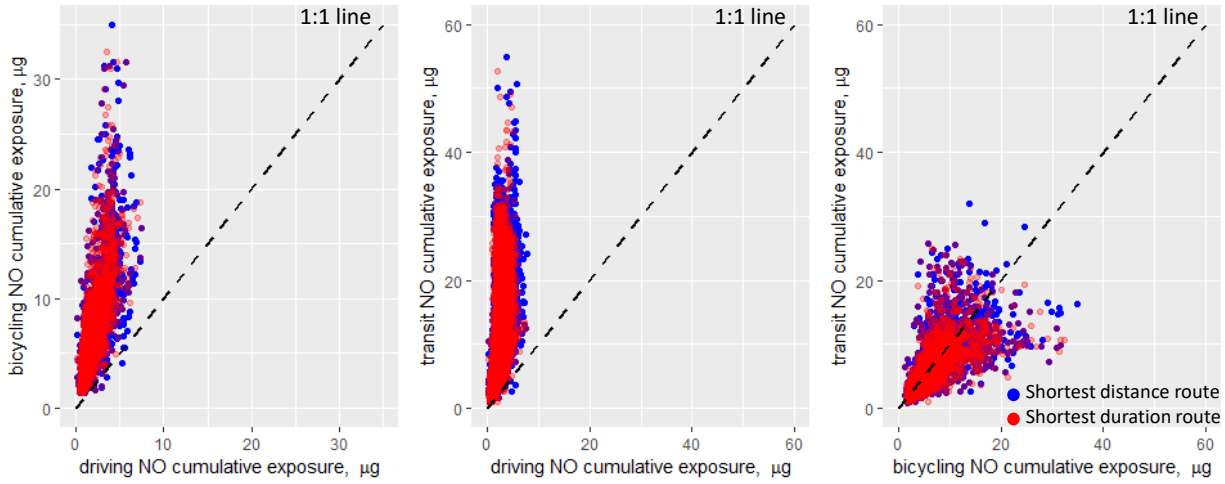

Figure 9. Comparison of NO cumulative exposure among different travel modes.

3.3 Shortest distance and shortest duration routes comparison

Here, we compare the importance of route distance and duration in influencing the cumulative exposure. Figure 10 shows the scatterplot of cumulative trip exposure between the shortest distance routes and the shortest duration routes. For all pollutants and all travel modes, the shortest duration routes tend to have lower cumulative exposure than the shortest distance

routes. This would suggest that trip duration is more important in influencing the cumulative exposure than distance (i.e., the shortest duration route trips will experience lower cumulative exposure in a majority of conditions). The variations on the shortest distance routes don't have an important effect on the cumulative exposure. Alternatively, when the inhalation dose is factored in, shorter duration routes are likely to produce lower cumulative exposure than shortest distance routes.

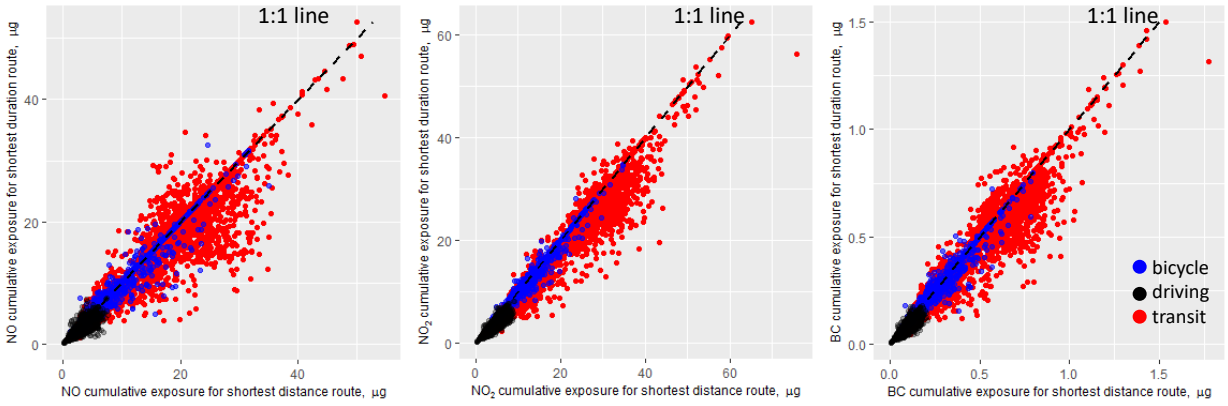

Figure 10. Cumulative exposure comparison between shortest distance and shortest duration routes.

4. Limitations and conclusion

To help contextualize our study, in this section we elaborate on the possible limitations and offer concluding comments. First, our air pollution measurement data reflects annual average estimates aggregated from Google Street View vehicle measurements. Thus, our estimates of cycling and transit exposure may be upwardly biased. Bicycle and transit routings can also include residential streets, separated pathways, and a 4-km underground subway tunnel; in these routings, air pollution concentrations may be less than the roads where we have Google observations. A second limitation is related to the ventilation rates for both auto and transit. Air

pollution concentrations in the micro-environment within car or on a bus may be different than the ambient concentrations; different vehicle characteristics, ventilation settings, driving conditions, and air exchange rates can all influence the in cabin air pollution concentrations. For cars, the traffic related air pollution concentrations tend to be lower inside than outside [64,65], while the particulate matter concentrations inside buses are higher than outside [66]. This would result in an overestimation of the exposure associated with driving and an underestimation with transit. However, these transport micro-environment studies are of very specific design (e.g. vehicle types, air conditioning settings). Our study focuses on the overall exposure assessment with wide mixture of fleet and large number of trips, where the specific micro-environment settings are less important. Therefore, we don't expect that either of these limitations would significantly affect our results.

The methodology discussed in this paper can also be applied to other areas with high-resolution air pollution data. Until 2019, the Google street view vehicle air quality mapping campaigns have been deployed in London, Copenhagen, Amsterdam, and Houston; and they are expanding the air quality mapping in more places around the globe [67]. Even for areas without high-resolution air pollution measurement, the use of land use regression model, geo-statistical model, chemical transport model, and remote sensing technology are also able to estimate the high-resolution air pollution distributions based on the regulatory air pollution monitors and the portable air pollution sensors. With the estimated air pollution mapping, these areas are capable of applying the method discussed in this paper to quantify the cumulative exposure with different traffic modes and route options.

In summary, this paper uniformly simulates 5,000 trips in Oakland CA. and calculates cumulative exposure of NO, NO_2, and BC for all trips using three different travel modes: driving, bicycling, and transit. In Oakland, CA. transit tends to travel the longest duration, while bicycle routes tend to have the shortest distance/duration among our travel modes. We show that cumulative exposure is more useful to assess travel exposure than average concentration, which is more typical in literature. For all three travel modes, the average concentrations are not significantly different. But due to the travel duration and distance variations of different travel modes, the cumulative exposure varies by mode. Traveling by auto has the lowest cumulative exposure, and also has the lowest average per meter and per minute exposure. Finally, we show that trip duration is more important than trip distance for air pollution exposure; there is less exposure for shortest duration trips than there is for shortest distance routes.

Submitted to *Environmental Science & Technology*

Chapter 4. Comparing Machine Learning, Deep Learning, and Land Use Regression Outcomes Using Google Street Mobile Source High-Resolution Black Carbon Concentrations in Oakland, California

*AUTHOR NAMES: Minmeng Tang[1], Deb A. Niemeier[2]**

AUTHOR ADDRESS: [1]Department of Land, Air, and Water Resources, University of California, Davis, One Shields Ave. Davis, CA, 95616, USA

[2]Department of Civil and Environmental Engineering, University of Maryland, 1173 Glenn Martin Hall, College Park, MD. 20742, USA

KEYWORDS: air pollution prediction, land use regression, machine learning, random forest, support vector machine, neural network

Abstract:

In this paper, we couple land use model with linear regression and three different machine learning and deep learning models including Random Forest (RF), Support Vector Regression (SVR), and Neural Network (NN) to estimate black carbon (BC) concentrations in Oakland, CA. Since SVR is sensitive to input features, we apply least absolute shrinkage and selection operator (LASSO), principle component analysis (PCA), conditional independence feature ordering (FOCI), and genetic algorithm (GA) for feature selections and dimension reduction from the output of the land use model; while no feature selection and dimension reduction methods are used for RF and NN. We use Bayesian Optimization method to automatically tune RF and SVR; while NN is

manually tuned because of the complexity of the model's structure. RF performs the best among all algorithms with regression coefficient (R^2) values of 0.701 on train set with 5-fold cross-validation and 0.694 on the independent validation set. SVR shows lower prediction accuracy than RF with R^2 values at 0.693 and 0.667 on train and validation sets respectively. The LASSO is used in the LR model to select features, which results to R^2 values of 0.596 and 0.594 on train and validation sets respectively. NN has the highest R^2 value on the train set at 0.723, but the lowest R^2 on the validation set at 0.466, which is likely due to overfitting issue. The one-factor-at-a-time (OAT) sensitivity analysis suggests that RF is the most robust model among all four models. The features that are most sensitive in predicting BC concentrations are vehicle speed and the total length of local road systems (highways, arterials, and residential roads) within different buffer sizes. Highways and truck routes are both significant sources linked to local hotspots. The most common approach using LURs is not particularly efficient at identifying localized hotspots. However, higher accuracy and robustness for predicting air pollution using LUR can be achieved by coupling the RF model using high-resolution mobile measurements. Together, this modeling approach can improve air quality exposure assessment for vulnerable population groups or communities and help to address environmental justice issues.

Synopsis:

Our results indicate that machine learning approaches are more robust than other land-use regression approaches. Continued appraisal and development of these new approaches can significantly enhance our understanding of the spatial variation of air pollutants.

1. Introduction

Urban areas are the hotspots of global air pollution problems. The surface topography, emission source variation, and population distribution all lend themselves to highly variable air pollutant concentrations in urban areas; concentrations that can vary dramatically even within short distances [6–8]. The spatial variations in air pollutant concentrations can be as large as the contrast between cities [9]. Epidemiological studies clearly show that within-city PM exposure is larger than the between-city effect [68]. One of the critical gaps in our understanding is how to characterize *within*-city air pollutant concentration gradients, which is crucial for exposure assessment [11], urban planning [12,13], air pollution monitoring [14], and environmental equity [15].

With the development of high accuracy portable pollution sensing instruments and Global Positioning System (GPS) technology, the use of vehicles for mobile air pollution monitoring is increasingly tackling some of the challenges of estimating pollutants based on stationary monitoring sites. These mobile sensors can typically achieve high spatial resolution for air pollutants measurement, but there is a trade-off. The high spatial resolutions come at the cost of low temporal resolution for any given location. In this paper, we bridge the gap between the high spatial resolution-low temporal resolution offered by the mobile sensors and the low spatial resolution-high temporal resolution offered by the stationary source monitors to estimate pollutants at unmeasured locations. To do this, we take advantage of modern computing techniques.

One approach frequently used to estimate localized air pollutant concentrations at unmeasured locations is Land Use Regression (LUR). LUR is frequently used in intra-city air pollution prediction

and health exposure studies due to its simplicity, and interpretability. Traditional LUR models rely on stationary monitoring observations. There have been studies developed using mobile observations [69–81]. However, to take advantage of the high spatial resolution of mobile measurement, the corresponding land use variables at a similar spatial scale are necessary to maximize model prediction accuracy.

Hankey & Marshall (2015) constructed a LUR model using bicycle-based mobile source measurements; this research suggests a framework for constructing a LUR model from the mobile source measurements. The model R^2's vary with pollutant, with the highest adjusted R^2 (0.5) associated with particle number and an adjusted R^2 for BC prediction of only 0.35. It is not uncommon for LUR models to produce lower prediction accuracies using pollutant source data generated at the microscale [72,73,75,76,78]. The few studies using the LUR approach achieving higher R^2s (between 0.6 and 0.8) [70,71,80] are usually constructed with measurements at carefully selected locations for a short period of time. [75] On average, most LURs use around 40 pre-selected locations, which increases the modeling variability.[75] For campaigns with short-term measurement periods and a limited number of locations, the LURs may not be the optimal approach for capturing the spatial variation of long term average concentrations [81].

This study addresses two gaps in the literature. First, we explore the predictive ability of the LUR for longer term measurement over a varied urban landscape. We use the google street view mobile source measurements in West Oakland, California, which covers slightly more than one year and includes every street in West Oakland. Our second contribution is to couple the LUR with modern computational methods. We specify different regression and prediction models (linear regression, Random Forest (RF), Support Vector Regression (SVR), and Neural Network

(NN) and couple these with a land use model to predict air pollution concentrations. In this way, we are able to evaluate both the advanced methods and the modeling performance together.

2. Materials and methods

2.1. Study domain and air pollution data

Our study domain covers the West Oakland (WO) area in Oakland, CA (Figure 11), which about 10 km^2 with a mix of residential and industrial blocks, surrounded by three major interstate highways (I-880, I-980, and I-580); the area also host the 9th-largest container port in the U.S. (Port of Oakland). Two Google street view mapping vehicles, equipped with Aclima environmental intelligence sensors and a data integration platform, were deployed in the study area between June 2015 and May 2016. The vehicles repeatedly measured weekday daytime concentrations of black carbon (BC), nitric oxide (NO), and nitrogen dioxide (NO_2) in every road in Oakland, CA. [54]. Apte et al. applied a data reduction and aggregation algorithm to convert these data of instantaneous observations into estimates of median annual weekday concentrations for every 30-m road segment in the study domain [6]. We use the high-resolution black carbon (BC) concentrations from Apte et al's supporting information (Figure 11).

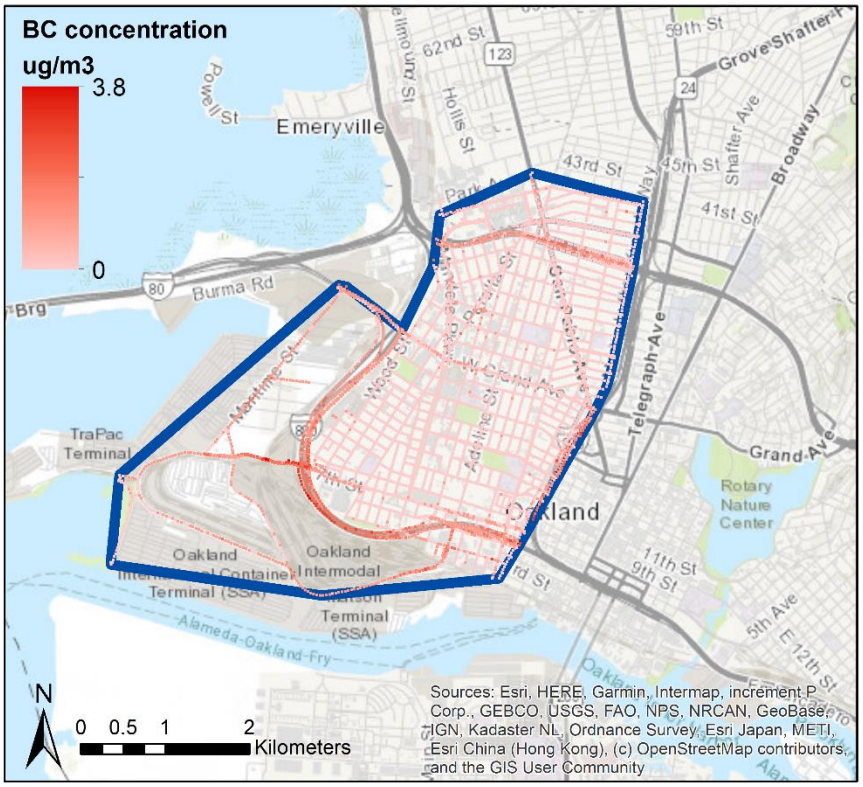

Figure 11. Study domain and high-resolution BC concentration map.

2.2. Land use model specification

As noted, our study problem is how to optimally use LUR models with large, spatially resolved data. In this case, we have more than 5,500 30m roadway segments, each with a BC annual average concentration. To take advantage of the spatial resolution offered by the concentrations, we also need land use variables that optimally vary at, or near the same spatial scale. To specify the LUR, we use the log of BC concentrations as the dependent variable. Using Messier et al. (2018) as a guide [82], we calculate independent variables using road length, road classifications, truck routes, local zoning classifications, normalized difference in vegetative index, land cover, population, point sources, and elevation, among others (see section S1 in supporting information). In total we assembled 108 number of variables for which we derived values using

six buffer sizes, 50 m, 100 m, 250m, 500 m, 1000 m, and 2500 m. We normalized numeric variables with zero-mean and unit-variance before conducting regression analysis.

2.3. Regression model specification

We construct four regression algorithms to couple with the land use features: 1) linear regression (LR); 2) random forest regression (RF); 3) support vector regression (SVR), and 4) neural network regression (NN). Our models are built in Python 3.7.6 [83] using scikit-learn 0.22 [84]. For the linear regression, we apply the least absolute shrinkage and selection operator (LASSO) method for feature selection. We regularize the coefficients of the independent variables with a shrinkage parameter to constrain the magnitude; this helps to avoid over-fitting and to select influential features. Random forest is a supervised learning algorithm, which uses resampling to estimate large numbers of regression trees. The individual trees act as an ensemble, with the important features of the final model emerging in the aggregation. SVR uses kernel functions to map the nonlinear data into high-dimensional feature space where complicated nonlinear relationships become linear. By optimizing the support vectors, the dependent variable can be regressed against independent variables. SVR is sensitive to the input features and requires careful feature selection. To optimize the SVR model, we apply three different feature selection methods: random forest feature importance ordering, feature ordering by conditional independence (FOCI), and the genetic algorithm (GA) and we use principal component analysis (PCA) for a dimension reduction method. Further details on the approach are discussed in section 2.4, model tuning. Our final approach, artificial neural network (NN), originating in neuroscience, is capable of simulating complex, nonlinear patterns. We construct an easy-to-tune feed-forward neural network model which we then couple with the land use model for BC predictions. To reduce the

NN training time , we use a graphics processing unit (GPU) from the Google Colaboratory cloud platform [85].

2.4 Model tuning

Tuning is the process of optimizing a ML model, which is accomplished by choosing the appropriate hyper-parameters to govern the learning process.

For our models, we have to specify a different means of specifying the hyper-parameters. We use a grid search in the LR model to select the (constant) LASSO shrinkage parameter. In the RF model, there are several hyper-parameters that must be tuned, which can be computationally expensive. We combine a RF with the Bayesian hyper-parameter optimization algorithm, which is a probabilistic model based approach. In general, this method builds a probability model of the objective function based on the previous iterations and optimizes the hyper-parameters of the true objective function according to this probability model. This approach is efficient because iterations are informed by past results. We implement the Bayesian hyper-parameter optimization process using Hyperopt [86]. We define the search space for the RF hyper-parameters (Table 2) and the optimization function in Hyperopt provides the optimized values of all hyper-parameters (Table 3). To be consistent, we set the maximum iteration numbers as 100 for all the Bayesian hyper-parameter optimization processes.

As noted earlier, to tune the SVR model, we try three different feature selection algorithms (FOCI, RF feature importance and GA) and one dimension reduction approach (PCA). FOCI is a forward stepwise feature selection algorithm, which uses the conditional dependence coefficient to select a subset of variables based on the predictive power [87]; the algorithm has been developed as the FOCI package in the R environment. Within the RF method, this approach provides not only

predictions of the dependent variable, it also evaluates the importance of input features using the regression trees. Based on the RF feature importance ordering, we can select different features as input for the SVR model. The GA based feature selection algorithm is a particular designed optimization algorithm, which selects subset of features and optimizes the hyper-parameters of the SVR model simultaneously [88].

The FOCI method resulted in selection of 13 features out of our 108 input features (Table 4). We use these 13 features as independent variables for the SVR model and use the Bayesian hyper-parameter optimization algorithm to tune the SVR model. For the RF feature importance and PCA methods, we select different feature groups as SVR model input, and apply Bayesian hyper-parameter optimization algorithm to automatically tune each SVR model.

Next, the genetic algorithm (GA) uses a fittest survives approach to produce next generation of offspring. In this algorithm, each solution has a "chromosome", which represents a set of parameters (in our case features and hyper-parameters). Each individual has a fitness value (R^2), which represents the quality of the solution. We first randomly initialize the algorithm by generating 100 individuals as the mating pool, where the two individuals with the highest fitness values are selected as parents. Every two parents will generate 8 offspring, and the two offspring with the highest fitness values are selected as the parents for the next generation. Mutation and crossover of the parents' chromosomes are introduced into the mating process, which generate different individuals from parents. Details about this GA method can be found in D. Zhang et al. [88] and our parameters setup is available in section S3.2 in supporting information.

Finally, NN models are difficult to tune because of their complicated structure and the lack of a robust automatic tuning algorithm. For this approach, we manually tune the NN model and apply

our expert knowledge to select appropriate structures with good prediction performance while avoiding overfitting. In the tuning process, we vary the number of layers, number of neurons in each layer and the activation function in each layer of the NN model and select the final model with best prediction accuracy.

2.5 Model validation

For model validation, we randomly split the data into an 80% training set and 20% validation set. We use only the training data in the model tuning process (section 2.4); the validation data is used to calculate R^2 for each optimized model, which is the criteria we use to evaluate each model's performance. To be consistent, we use the coefficient of regression (R^2) as the criteria to evaluate performance for all models, and 5-fold cross-validation is applied to calculate R^2 to avoid overfitting.

3. Results and Discussion

3.1 Model development

In the LASSO model, it gives 75 non-zero features and the regression coefficients are available in Table 7. Based on the training data, LASSO model has $R^2 = 0.596$, while the R^2 calculated based on the validation data equals 0.594. The very close R^2s between train and validation sets suggests that this LASSO model is accurate and robust in predicting BC concentrations coupling with land use model.

In the RF model, the optimized hyper-parameters are listed in Table 3, which provides $R^2 = 0.701$ based on the train data and $R^2 = 0.694$ based on the validation data. Similar with LASSO model,

RF provides very close R^2s between train and validation data sets, suggesting the robustness of this model.

For the SVR model, the tuning process is complicated as the feature selection or the dimension reduction are necessary. In order to achieve the best possible predicting performance of the SVR model, Figure 12 shows how the train set R^2 changes with different number of input features. The RF feature importance method provides the highest train set R^2 = 0.650 with 10 features selected, while the PCA method provides the highest train set R^2 = 0.627 with 100 features selected. FOCI method provides the train set R^2 = 0.604, and GA method provides the highest train set R^2 = 0.693 with 45 features. Among all these feature selection and dimension reduction methods, GA provides the highest possible train set R^2, which provides the validation set R^2 = 0.667. Table 5 and 6 list the detailed values of hyper-parameters and the selected features of the SVR model coupling with GA tuning method. The difference between the train and validation R^2s of SVR model is larger than LASSO and RF, suggesting that SVR does not generalize as well as the other two models.

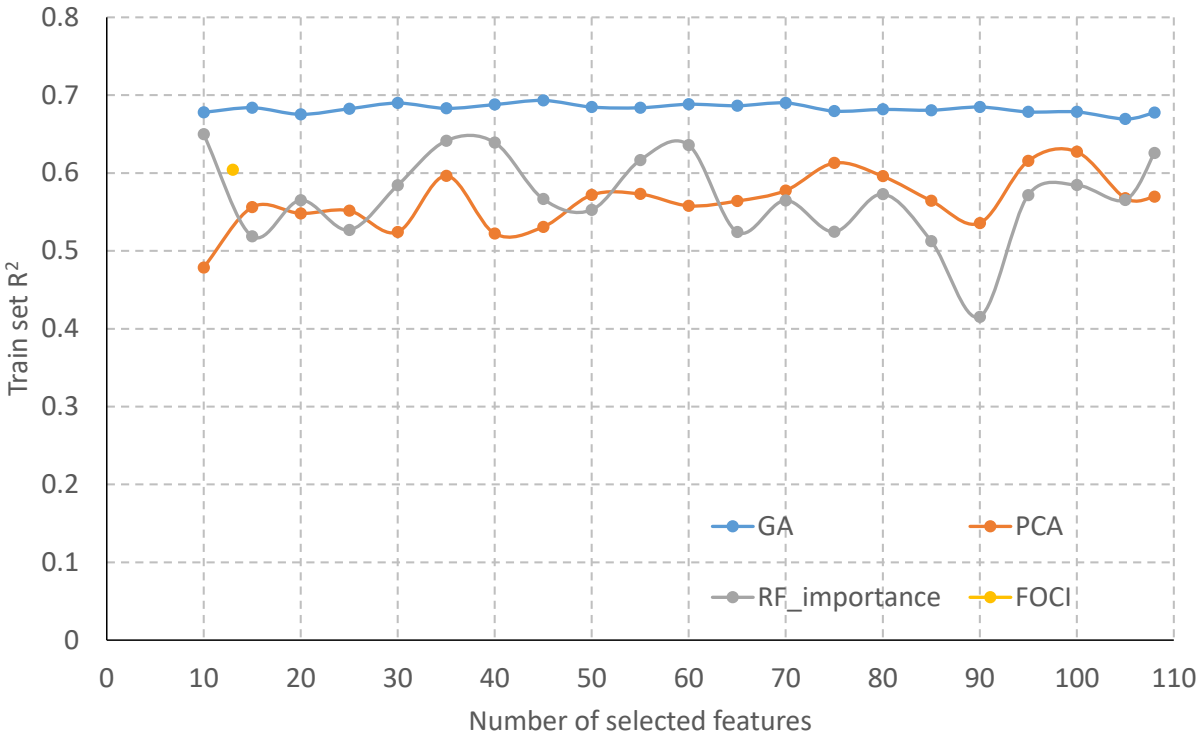

Figure 12. SVR train set R2 changes with varying number of input features selected by different feature selection and dimension reduction methods.

In the NN model, we construct the three-layer structure with sigmoid activation function and 50, 25, and 10 neurons in each layer, respectively. This model has estimation accuracy R^2 = 0.723 based on the train set, but the validation set R^2 is only 0.466. With better initialization of the NN model's parameters, the estimation accuracy of the train set can be even improved, however, this would lead to even lower accuracy for the validation set. The NN model has the highest train set R^2 but the lowest validation set R^2 among all the models in this paper. The large difference of R^2s between train and validation sets suggest that the NN model suffers severe overfitting issue.

3.2 Model performance evaluation

Among all the four models in this study, RF provides the highest validation R^2 with relative consistent prediction accuracy between train and validation sets. To better evaluate different models' performance, Figure 13 shows the scatter plot between predicted values and measured values of the validation set. There is no significant bias between predicted and measured values among all models. However, the NN model results in more outliers than other models, with the other three models sharing very similar patterns.

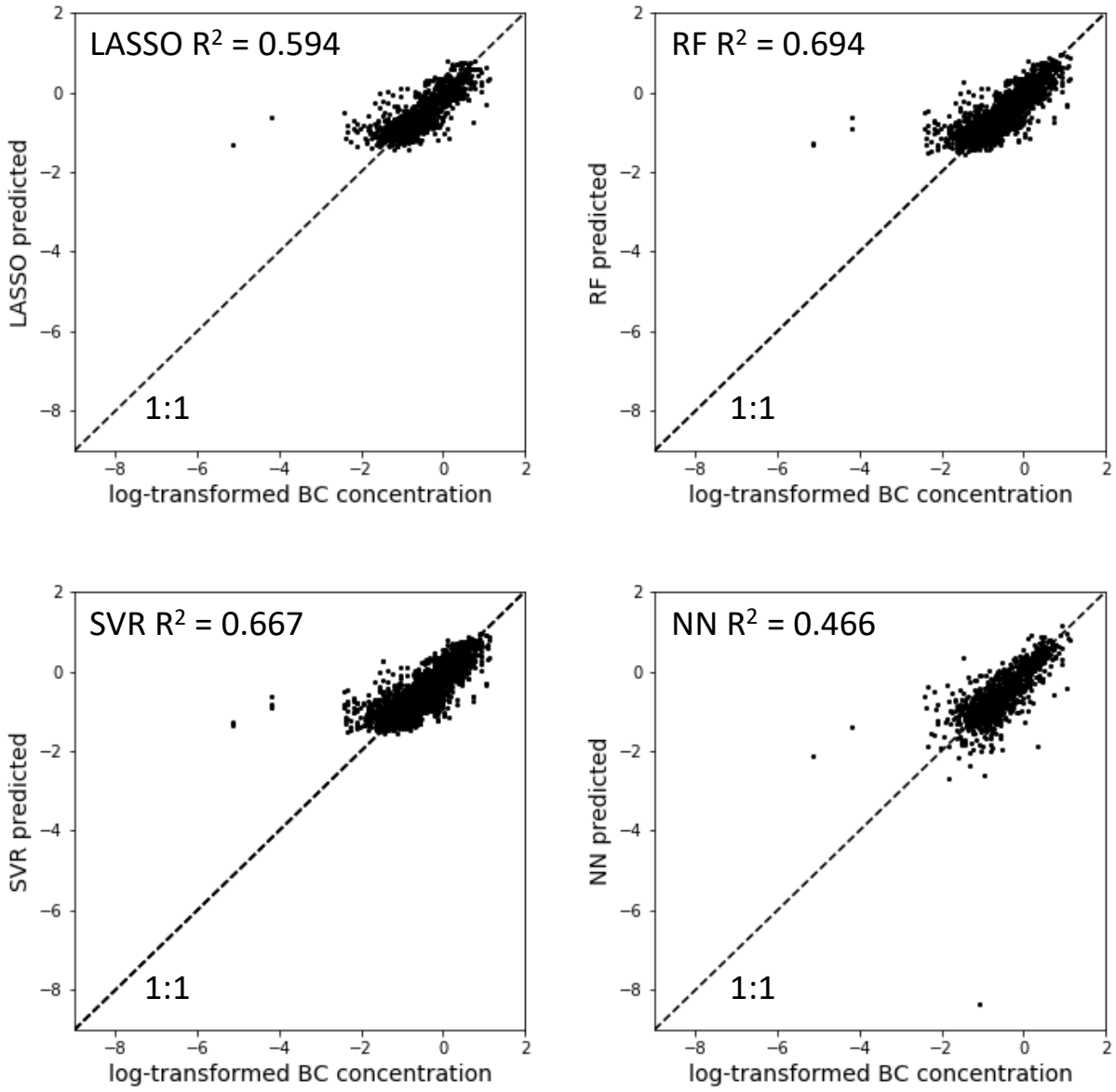

Figure 13. Predicted against measured values on validation set for all four models.

To better understand how the outliers spatially locate within the domain, we show the modeling between predicted values and true values in Figures 14, 15, and 16. The differences are normalized by the true (observed) values. Green dots designate points less than 10th percentiles

(i.e., the model underestimates BC concentrations); points greater than 90th percentiles are red (i.e., the model overestimates the observed BC concentrations).

In Figure 14, the normalized differences are plotted over the land use types. It's clear that most of the underestimated concentrations are largely located in three clusters, while the overestimated concentrations are more scattered. A large portion of the underestimations are located in the industrial and mixture of commercial and industrial regions, while the majority of overestimations are within the residential and commercial regions.

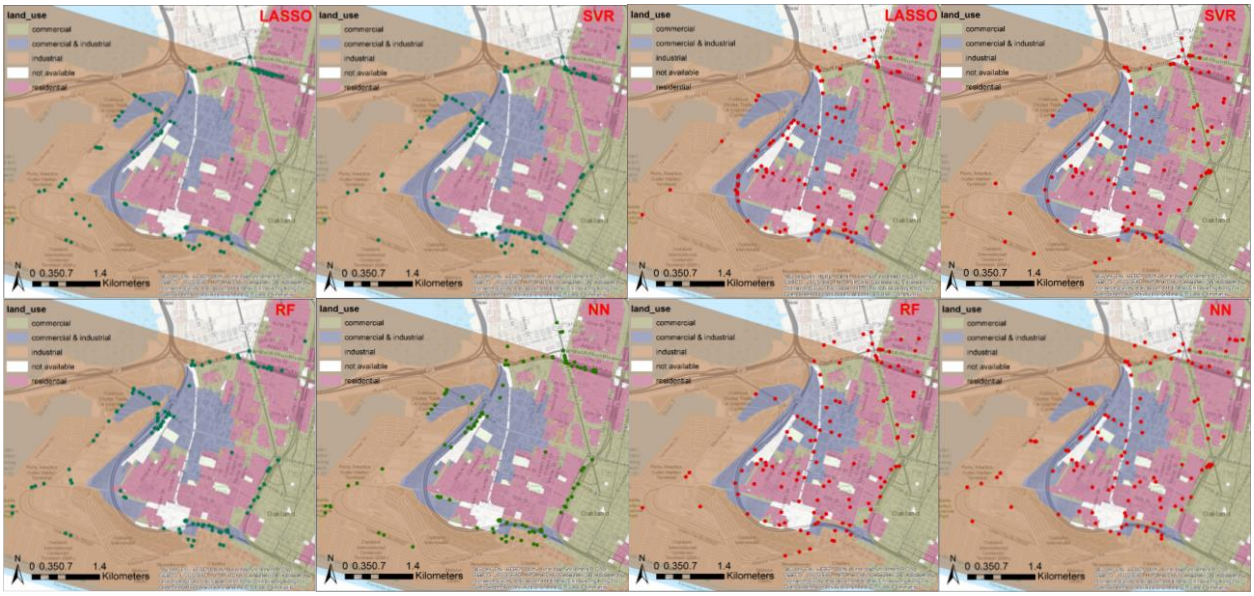

Figure 14. Less than 10th percentile (green dots) and greater than 90th percentile (red dots) of normalized differences of each model over the land use base map.

If we look at the relationship between the normalized difference points and the local highway systems, we find that most of the underestimations happen within a 100 meter distance from major highways, and nearly all of the underestimations are within a 500 meter distance from

major highways. We don't find a strong association between overestimations and local highway system.

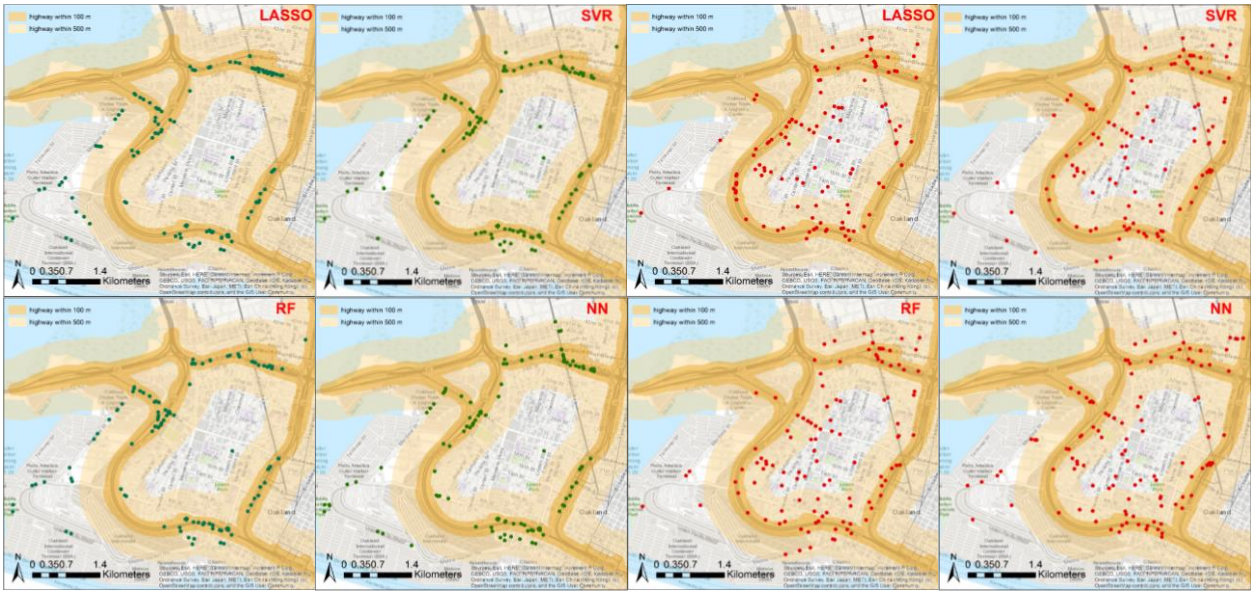

Figure 15. 10th percentile (green dots) and 90th percentile (red dots) of normalized differences of each model over the local highway system.

The routes that are designated for trucks have been noted in prior research as a parameter in BC predictions [69,70,76]. In Figure 16, we show the spatial relationship between designated truck routes and the normalized differences for each model. The majority of the underestimates generally locate within a 100 meter distance from truck routes and all of the underestimations are within a 500 meter distance from truck routes. The spatial variability of the over-estimates is much larger and scattered along both truck routes and truck prohibited routes.

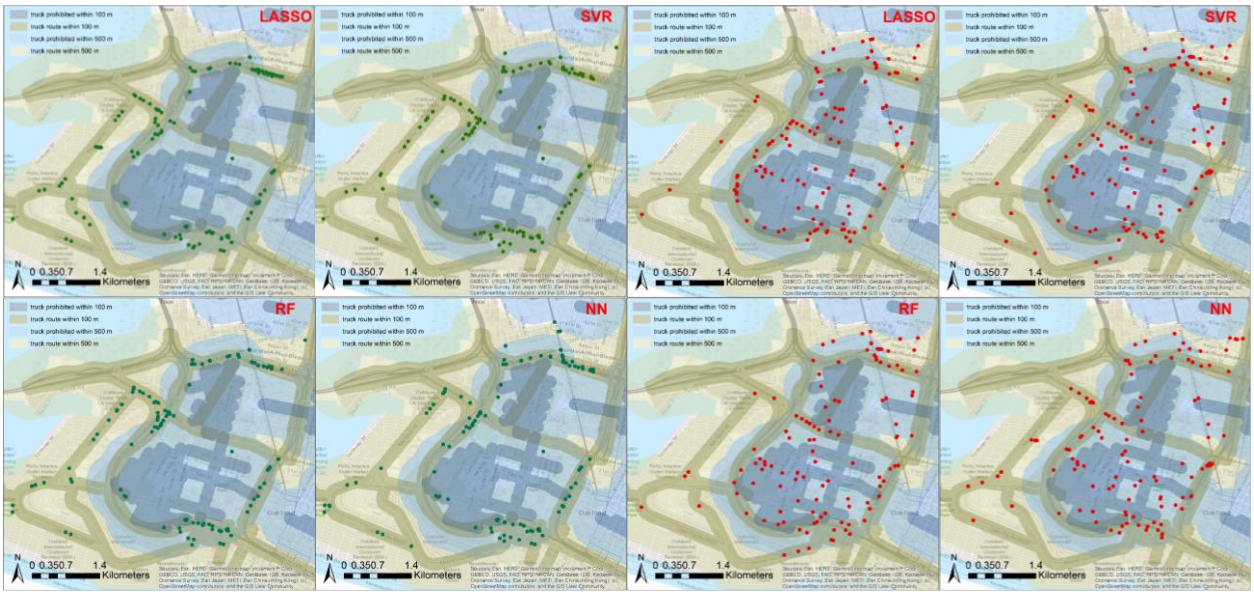

Figure 16. 10th percentile (green dots) and 90th percentile (red dots) of normalized differences of each model over the local truck routes and truck prohibited routes.

3.3 Sensitivity analysis

To understand the relative importance of the input features in predicting BC concentrations, we perform a sensitivity analysis using the one-factor-at-a-time (OAT) approach. In this method, we perturb one feature at a time from 0% to 200% with a 10% increment, keeping all other features unchanged. We then calculate by how much the model predicted BC concentrations vary from the true (observed) measurements. Figure 17 shows the features most sensitive to perturbations for each model. We also show how the predicted BC concentrations change as the input feature varies. Vehicle speed is highly sensitive to perturbations in both the SVR and RF modeling approaches, and is ranked high for sensitivity in the NN model. The features most sensitive to perturbations in the NN and LASSO models are the total length of highway within the 100 m buffer and the total length of residential road within the 2500 m buffer, respectively. The length of highways, arterials, and residential roads under different buffer sizes are all in the top five

features most sensitive to perturbations for all modeling approaches. Additional details about the features for each model are shown in Figure 18.

Across all four models, the RF model shows the most robustness; that is, the model performance in terms of the BC prediction is the least influenced by the variation of a single input feature. In contrast, the LASSO model is least stable and can be easily influenced by the variations of input features. SVR and NN show similar robustness, but both are less robust than the RF approach. Our sensitivity analysis suggests that vehicle speed, proximity and type of local road system are important in predicting BC concentrations across all approaches. This is a reasonable and expected finding. We also find that the RF model is the most robust and performs best among all four models when input features vary.

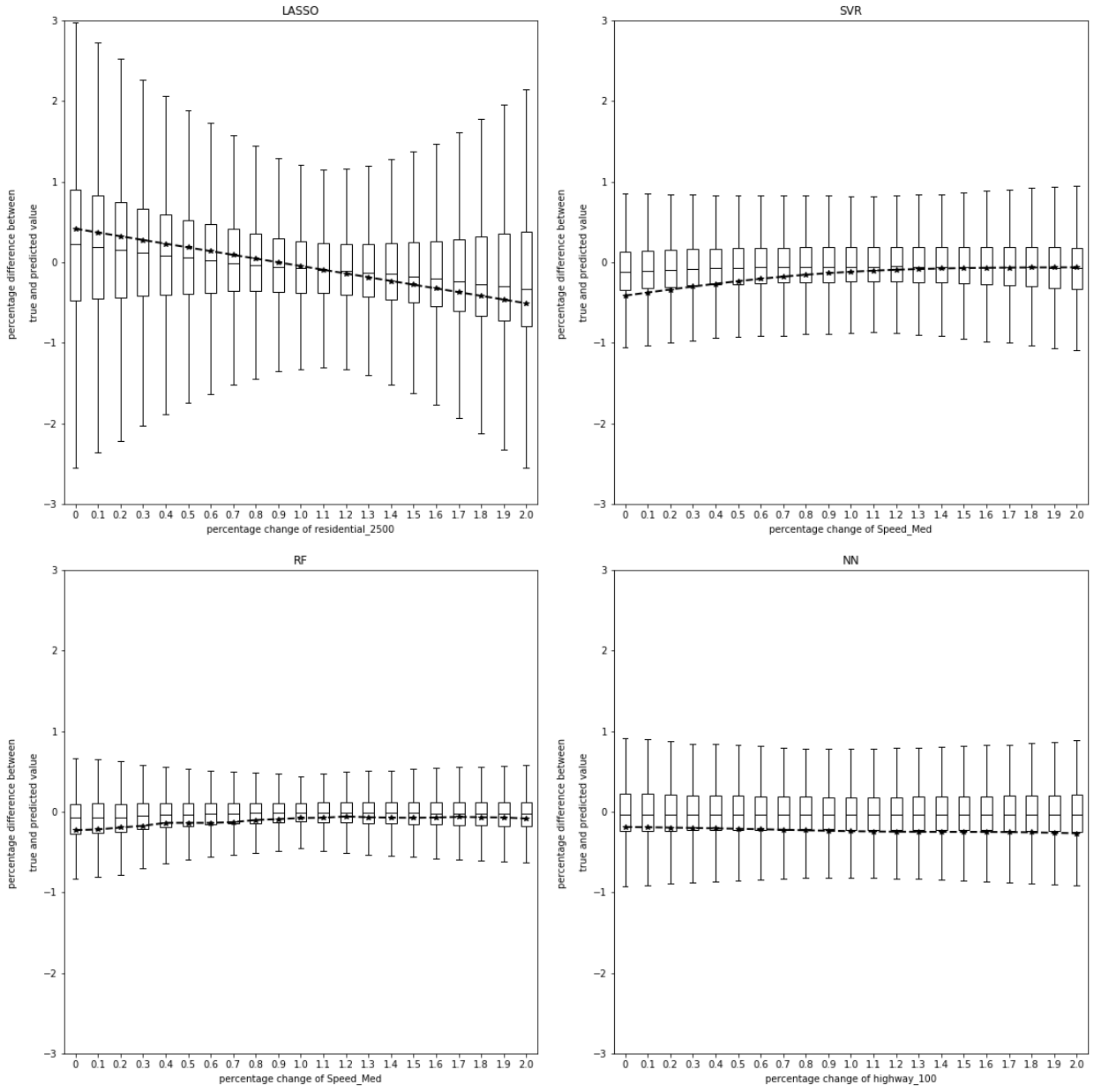

Figure 17. Most sensitive features for all four models and how their variations influence model performance in BC prediction (box shows 25th, 50th, and 75th percentiles; dot means mean value).

3.4 Discussion

In general, we found that there are some advantages to coupling modern computational methods with LUR models. However, we also found consistent patterns of under- and over-prediction that suggest these models need further development. All of the models consistently underestimated BC concentrations in three similar clusters, which had industrial and mixed commercial and industrial zoning. These clusters are also within 100 meter distance from major highways and truck routes. In contrast, we found that model over-prediction tended to occur within residential and commercial regions without any clear spatial patterns with respect to highways and truck routes. The locations of the normalized difference points do not show significant spatial patterns among the different models. Our underestimation and overestimation patterns suggest the LUR model, even with additional computational assistance and localized data, may be limited in identifying local hotspots near major highways and truck routes.

With respect to how the various models perform, LASSO is the simplest model and requires the shortest training time; however, its prediction accuracy is not as good as the SVR and RF models. In the SVR model, the feature selection is very important in optimizing the model's prediction accuracy, and the algorithm itself is computationally expensive. The GA method trains the SVR model more than 500 times to select 45 best feature combinations and optimize the hyper-parameters. Even though a carefully optimized GA method may increase the converging speed, the GA based feature selection algorithm still requires a large number of iterations and is computationally expensive. The RF approach requires a shorter training time because it is naturally suitable for parallel computing. The RF model without any feature selection or dimension reduction algorithm performs better than the SVR with carefully selected input

features. This suggest that the land use model of mobile measured air pollution data coupled with the RF model may be useful in predicting localized air pollution.

Although the NN model has the highest prediction accuracy based on the training data, it has the lowest accuracy on the validation data. Improving on feature selection or using a dimension reduction technique might improve the NN model's overfitting issue, but the training process of NN is much slower than the other three models, even after introducing the GPU. Because of the complexity of the NN model's structure, it is difficult to apply automated optimizing algorithms in the tuning process. Due to the computational resource limitations, we do not perform a comprehensive feature selection and dimension reduction on the NN model. We do believe that if a study is focused solely on achieving the highest possible prediction accuracy, it may be worth devoting the needed computational resources tuning the NN model with feature selection and dimension reduction algorithms. However, if predicting air pollution concentrations is only part of an analysis, the RF approach is a good first choice because of its high prediction accuracy, less training time, and robustness to overfitting issues. The sensitivity analysis also suggests that the RF model is the most robust among all four models.

Our study provides a reasonably good prediction accuracy compared to the literature. For example, Hove et al. conducted their BC mobile measurement campaign in Ghent, Belgium in December 2015 and constructed the land use linear regression model with cross-validation R^2 = 0.520 [73]. Messier et al. used the same air pollution data as our study but they also included the air pollution measurement in Downtown Oakland and East Oakland areas and constructed the land use kriging regression model with cross-validation R^2 = 0.43 [82]. These studies provide similar prediction accuracy as our work, but we offer a method that can be readily generalized. The

kriging regression tends to provide lower prediction accuracy than other models, while the linear model from Hove et al. shares a similar prediction accuracy as our LASSO model.

Lim et al. conducted mobile sampling campaigns with the low-cost sensors counting $PM_{2.5}$ particle number in Seoul, South Korea [89]. This study shares very similar prediction accuracy to our results and the stacked ensemble method provides higher prediction accuracy. The stacked ensemble method is an approach that combines the predictions of several other machine learning algorithms, which tends to achieve higher prediction accuracy compared to a single machine learning method. This method, although resulting in higher accuracy, is not easily generalizable and is computationally expensive. Our finding that the RF model is more robust is consistent with Ren et al. in which data from more than 1,000 stationary ozone monitors data in the U.S. were collected and used to populate spatial and spatiotemporal land use regression models with 13 linear regression and machine learning algorithms. The results find that RF and extreme gradient boosting are the best performing algorithms.[90]. We have generalized our results to hyper-localized pollutant data compared to Ren et al., which uses stationary monitoring data.

4. Conclusion

In this paper, we develop land use regression models based on high-resolution mobile observed BC concentrations in the West Oakland, CA.. Our study explores four linear regression and machine learning algorithms. The machine learning algorithms used in this study include RF, SVR, and NN. To comprehensively evaluate each model's performance in BC prediction, we carefully tune each model and compare their performance based on the regression coefficient from

validation set, which is independent from the data used for model tuning. Although the NN shows the highest prediction accuracy in training set, the model suffers from overfitting, which leads to low prediction accuracy across the validation data. Generally, RF performs the best among the four regression algorithms. We find it is most suited for hyper-localized air pollution concentration prediction, air pollution epidemiology modeling, and health exposure assessment studies, because of its high prediction accuracy, less training time, and robustness to overfitting issues.

5. Supporting Information

5.1 Land use variables

The land use variables are calculated using Messier et al. (2018)[82] as a guide and the detailed instructions are available in the Messier et al. (2018) supplementary material. In general, we construct 108 land use variables from the following datasets with six buffer sizes including 50 m, 100 m, 250 m, 500 m, 1000 m, and 2500 m:

we construct the binary road classification variable based on the OpenStreetMap dataset, which is an open source data that provides roads, trails, cafes, railway stations and so on all over the world[91]. In OpenStreetMap dataset, the road systems are classified into multiple categories based on the importance within the local road system as a whole. To simplify the road classification variable, we construct three categories from tens of categories in the OpenStreetMap dataset, which are highways, arterials, and residential roads. For the highways category, it contains motorway, motorway link, and trunk link; the arterials category contains

primary, primary link, secondary, secondary link, service, tertiary, tertiary link, and unclassified; the residential category contains living street and residential.

For each category (highways, arterials, and residential roads), we calculate the total road length within each buffer size based on the OpenStreetMap data[91].

Two binary variables indicating whether a road segment is on a designated heavy-duty truck route or on a road where heavy-duty trucks is prohibited are created based on the Oakland Truck Routes (2017) report created by the city of Oakland which is available online at https://www.arcgis.com/home/item.html?id=0fe7f165a9274b1182002ff9c0f4851d[92].

Three binary variables indicating commercial, industrial, and residential zonings are created based on the City of Oakland zoning classifications, which is available online at https://oakgis.maps.arcgis.com/apps/webappviewer/index.html?id=3676148ea4924fc7b75e7350903c7224[93]. These three land use zonings are generalized from the complex city zoning codes based on the explanations in Table 3 in the supplementary material in Messier et al. (2018)[82].

We also calculate the average Normalized Difference Vegetation Index (NDVI) within each buffer to represent the coverage of vegetation around each road segment. The NDVI is calculated based on the measurement from LANDSAT 8 in Google Earth Engine[94].

To include land cover information into the LUR model, we created 6 land cover types based on the USGS National Land Cover Database (NLCD) 2016[95,96] at 30 m resolution. These 6 land use types are Developed Open, Developed low, Developed medium, Developed high, evergreen forest and mixed forest. For each land cover type, six buffer sizes are used to calculate the corresponding variables, except evergreen forest and mixed forest, which only has 2500 m buffer

since the rest buffer sizes lead to variables with all zeros. For the land cover variables, the percentage of the land cover type within the buffer area is calculated by using the number of pixels representing the corresponding land cover type divided by the total number of pixels within the buffer area.

We use the 2010 census tract population data[97] to calculate the population density within each buffer. By assuming the population is evenly distributed within each census tract, we can compute the population density of each buffer based on buffer area and population density in each intersected census tract.

To calculate the mean elevation within each buffer area, we use the National Elevation Data (NED)[98] in Google Earth Engine[94], which has approximately 10 m resolution in US.

Some other point sources may also contribute to the air pollution concentrations. To include the point source contributions in the LUR model, we select four point source categories, which are ports, airports, National Priority Listing (NPL) sites, and Toxic Release Inventory (TRI) sites, and calculate the exponential decayed contributions from these sources by using Equation 1 in supplementary from Messier et al. (2018)[82]. We use 8 decay distances (λ_l) including 50 m, 100 m, 500 m, 1,000 m, 2,500 m, 5,000 m, 10,000 m, and 50,000 m for all the 4 point sources listed above.

Ports data is downloaded from Bureau of Transportation Statistics (https://data-usdot.opendata.arcgis.com/datasets/major-ports/data). Airports data is also available from Bureau of Transportation Statistics (http://osav-usdot.opendata.arcgis.com/datasets/831853ab8b714a81b6a3e21d0b164a4e_0). All ports and

airports within US are used in the calculation of point source contributions. NPL data is available from US Environmental Protection Agency (USEPA) (https://epa.maps.arcgis.com/home/item.html?id=c2b7cdff579c41bbba4898400aa38815#overview). We only use the NPL sites within Alameda County to prepare the point source contribution variables. TRI data is also available from USEPA (https://www.epa.gov/toxics-release-inventory-tri-program/tri-basic-data-files-calendar-years-1987-2018). We use the TRI sites within the surrounding 10 counties to prepare the corresponding variables.

Meanwhile we also calculate the inverse distance of the nearest point sources for each road segment with respect to all the 4 abovementioned point sources.

5.2 RF model tuning parameters

We are using Scikit-learn package in Python to train the Random Forest (RF) regression model[84]. Within this package, the RF regression model has multiple hyper-parameters and we are tuning five free parameters with the rest parameters using the default values, which are max_depth, max_features, n_estimators, min_samples_split, and min_samples_leaf. To apply the Hyperopt optimization algorithm, we need to first pre-define the search space, which limits the hyper-parameters' range. The pre-defined search space for the RF regression model is shown in Table 2.

Table 2. Search space for RF model hyper-parameters.

Hyper-parameters	**Values**		
max_features	'auto', 'sqrt', 'log2', 1, 0.5, 0.1, 0.05, 0.01		
	Minimum value	**Maximum value**	**Increment**
max_depth	10	50	1

n_estimators	450	1000	1
min_samples_split	2	10	1
min_samples_leaf	2	10	1

With the Hyperopt optimization algorithm, the optimized hyper-parameters for the RF regression model is listed in Table 3 below.

Table 3. Tuned values of RF model hyper-parameters.

Hyper-parameters	**Optimized value**
max_features	0.5
max_depth	24
n_estimators	781
min_samples_split	5
min_samples_leaf	2

5.3 SVR model feature selection, dimension reduction, and model tuning

5.3.1 FOCI feature selection

FOCI method selects 13 features out of the 108 input features and the 13 selected features are listed in Table 4.

Table 4. FOCI method selected 13 features for SVR model.

NDVI_1000	port_5000	residential_2500	residential_land_type	Developed_open_2500
Truck_prob	port_500	population_1000	Commercial_land_type	Industrial_land_type
Truck_route	airport_100	Mixed_forest_2500		

5.3.2 Genetic Algorithm (GA) parameter setup

In the GA model, the "chromosome" is set to have two sections. The first section contains three hyper-parameters of the SVR model, which are regularization parameter C, kernel coefficient gamma, and the loss function penalty parameter epsilon; the second section contains the selected features, which are the ID numbers of the corresponding features. Given the different nature of these two sections, the mating and mutation processes are calculated separately. The detailed mating and mutation processes are available in Zhang et al. (2015)[99]. In our GA model, the initial population size is set to be 100, mating cross rate 0.8, mutation rate 0.1, elite size 2, and offspring number 8. The GA optimized hyper-parameters of SVR model is listed in Table 5 and the corresponding features are given in Table 6.

Table 5. GA optimized hyper-parameters of SVR model.

Hyper-parameter	GA optimized value
C	2.35478
gamma	0.0831661
epsilon	0.129904

Table 6. GA selected 45 features for the SVR model.

Index_Hwy	NDVI_250	Developed_open_250	Developed_high_100	port_500
truck_route	NDVI_500	Developed_open_500	Developed_high_250	port_1000
Latitude	NDVI_1000	Developed_open_1000	Developed_high_1000	port_5000
Speed_Med	airport_1000	Developed_open_2500	Elevation_100	npl_50
highway_100	npl_10000	Developed_low_250	Population_250	npl_100

highway_1000	tri_50	Developed_low_1000	Population_500	npl_500
residential_50	tri_2500	Developed_low_2500	Population_1000	npl_1000
residential_500	tri_10000	Developed_medium_250	Population_2500	npl_2500
NDVI_100	port_50000	Developed_medium_2500	port_inverse_dist	npl_5000

5.4 LASSO model regression coefficients

The LASSO model gives 77 non-zero features. These features and the coefficients are listed in Table 7.

Table 7. LASSO selected features and the coefficients

feature	coefficient	feature	coefficient
intercept	-0.5849	Developed_low_50	0.0108
Road_Type	-0.0548	Developed_low_100	0.0130
Index_Hwy	0.1299	Developed_low_500	-0.0429
truck_route	0.1655	Developed_low_1000	-0.0067
truck_prob	-0.0281	Developed_low_2500	0.0897
commercial	0.0177	Developed_medium_50	0.0012
industrial	0.0659	Developed_medium_100	0.1765
Speed_Med	0.0993	Developed_medium_250	0.0694
highway_50	-0.0145	Developed_medium_500	0.0265
highway_100	0.1065	Developed_medium_1000	-0.0041
highway_250	-0.0182	Developed_high_50	0.0089
highway_500	-0.0114	Developed_high_100	0.2376
highway_1000	-0.0796	Developed_high_250	0.0941
highway_2500	-0.0593	Developed_high_1000	0.0695
arterial_50	0.0289	Developed_high_2500	-0.2894
arterial_100	-0.0316	Mixed_forest_2500	0.0041

arterial_250	-0.0199	Population_100	0.0269
arterial_500	-0.0247	Population_250	0.0356
arterial_1000	0.0225	Population_500	-0.0537
arterial_2500	0.0678	Population_1000	-0.0623
residential_50	-0.0507	Population_2500	-0.1372
residential_100	-0.0026	port_100	-0.0036
residential_250	-0.0168	port_500	-0.0484
residential_500	-0.0584	port_1000	0.1377
residential_2500	0.5946	airport_50	0.0018
NDVI_50	0.0034	airport_100	-0.0034
NDVI_100	0.0272	airport_500	-0.1449
NDVI_2500	0.0697	airport_1000	0.3026
Elevation_50	-0.0978	airport_10000	-0.2506
Elevation_100	-0.0919	npl_500	0.0004
Elevation_250	-0.0523	npl_1000	-0.1053
Developed_open_50	-0.0051	npl_5000	0.0220
Developed_open_250	0.0090	npl_10000	0.2574
Developed_open_500	0.0568	npl_50000	0.1345
Developed_open_1000	0.0069	tri_50	0.0571
Developed_open_2500	-0.1301	tri_100	-0.0266
tri_inverse_dist	-0.0370	tri_1000	0.0734
npl_inverse_dist	0.0095	tri_5000	-0.0984
tri_10000	-0.0013	tri_50000	-0.0530

5.5 Sensitivity analysis with the five most sensitive features

The top 5 most sensitive features for all four models based on the OAT sensitivity analysis are shown in Figure 18. For each column from top to bottom, it shows the five most sensitive features (most sensitive to least sensitive) for a specific model.

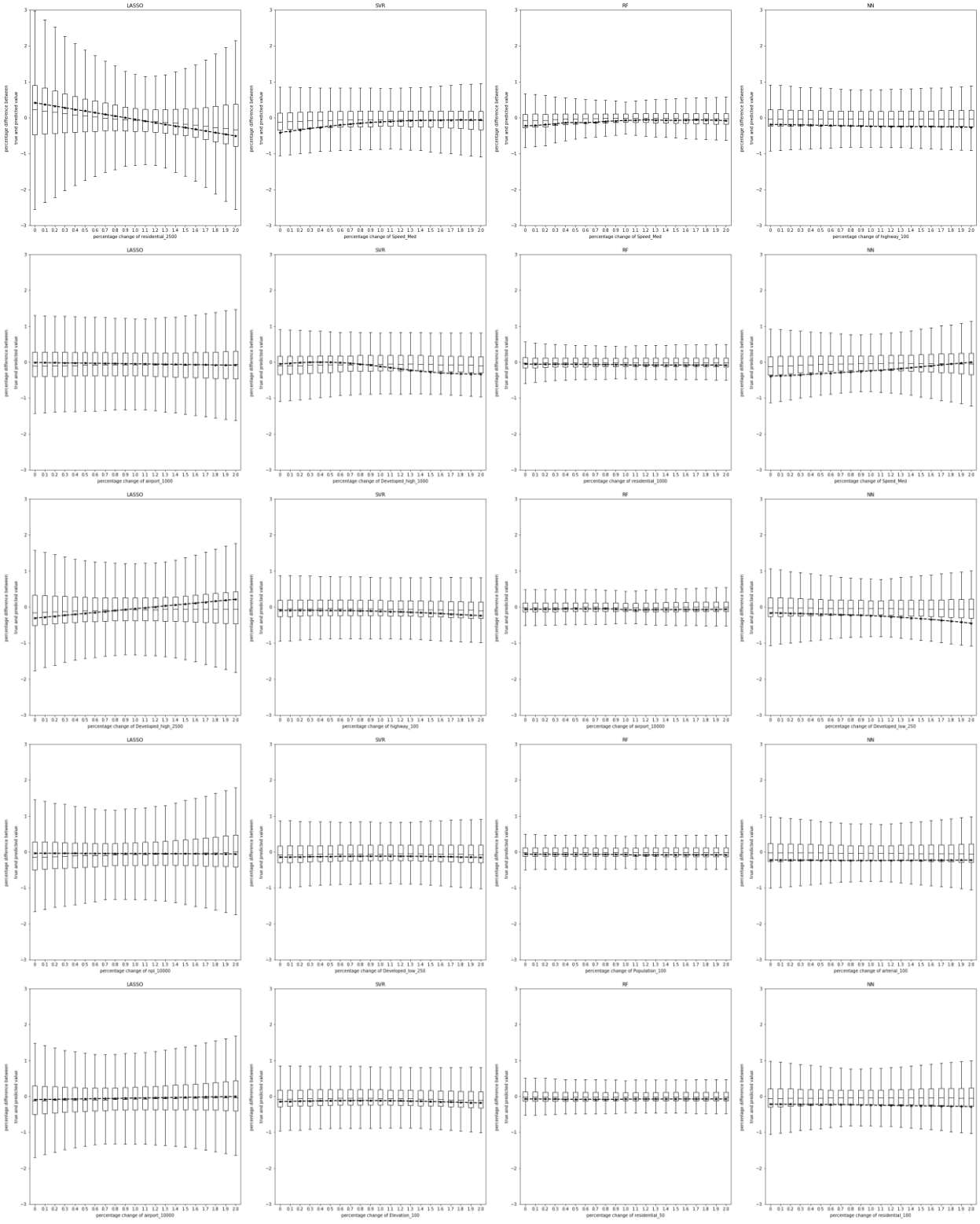
LASSO
SVR
RF
NN
percentage difference between true and predicted value
percentage change of residential_2500
percentage change of Speed_Med
percentage change of Speed_Med
percentage change of highway_100
percentage change of airport_1000
percentage change of Developed_high_1000
percentage change of residential_1000
percentage change of Speed_Med
percentage change of Developed_high_2500
percentage change of highway_100
percentage change of airport_10000
percentage change of Developed_low_250
percentage change of npl_10000
percentage change of Developed_low_250
percentage change of Population_100
percentage change of arterial_100
percentage change of airport_10000
percentage change of Elevation_100
percentage change of residential_50
percentage change of residential_100

Figure 18. Top 5 most sensitive features for each model and how their variations influence model performance in BC prediction (box shows 25th, 50th, and 75th percentiles; dot means mean value).

Submitted to *International Journal of Environmental Research and Public Health*

Chapter 5. How Air Pollution Influences Housing Price in Bay Area?

*AUTHOR NAMES: Minmeng Tang[1], Deb A. Niemeier[2]**

AUTHOR ADDRESS: [1]Department of Land, Air, and Water Resources, University of California, Davis, One Shields Ave. Davis, CA, 95616, USA

[2]Department of Civil and Environmental Engineering, University of Maryland, 1173 Glenn Martin Hall, College Park, MD. 20742, USA

KEYWORDS: air pollution, housing price, instrumental variable, spatial autocorrelation, spatial lag model.

Abstract

In this paper we examine the effects of localized air pollution measurements on the housing price in Oakland, CA. With high-resolution air pollution measurements for NO, NO_2, and BC, we can assess the ambient air quality on a parcel by parcel basis within the study domain. We combine a spatial lag model with an instrumental variable method to consider both the spatial autocorrelation and endogeneity effects between housing price and air pollution concentrations. We find a positive spatial autocorrelation with housing price using Moral's I (value of 0.276). Our results indicate air pollution positively influences housing price. It is somewhat surprising to find a positive relationship and we speculate that homeowners are insensitive to air pollution when the overall ambient air quality is good, or when there is low variability in pollutant concentrations.

Our result could be verified with more high-resolution air pollution measurements with a diversity of regions.

1. Introduction

Air pollution is not only a major global risk resulting in high incidences of illness and deaths,[1,100] but can also produce external damages to different economic sectors, including manufacturing, agriculture, transportation, and utilities.[101] In the U.S., air pollution costs are roughly equivalent to about 5% of the yearly gross domestic product (GDP) in 2014.[102] One sector we might expect to be highly sensitive to air quality is housing and there are a number of studies both nationally and globally focusing on the relationship between air quality and housing prices.

This literature mainly relies on the construction of the hedonic price models to evaluate the effect of air pollution on housing price. We can divide the body of research based on the approach. The first category uses an instrumental variable to address endogeneity effects and frequently uses a variable that is not related to housing price but directly related to air pollution as the instrumental variable to determine the exogeneous part of the variability from air pollution.[103,104] The second group uses spatial econometric models and hedonic price models to understand air pollution's influence on housing price accounting for spatial autocorrelation of housing price. The most common spatial hedonic models are Spatial Lag Model (SLM)[105,106], Spatial Error Model (SEM)[105,106], Spatial Durbin Model (SDM)[107], Geographically Weighted Regression (GWR)[108,109] and Quantile Regression Models (QRM)[109]. The results from this literature are inconclusive: some of the studies conclude that air pollution concentrations do not significantly influence housing

price[105,106,110], while others find that air pollution concentrations negatively and significantly influences housing price[104,111–114].

Previous studies have produced inconclusive findings in part, because there were limitations to the approaches. For example, nearly all the studies consider only spatial autocorrelation or endogeneity effects. Most studies rely on Moran's I to measure spatial autocorrelation[115] and results from cities in both China and U.S. suggests that there is positive and significant spatial autocorrelations in housing prices.[116,117] When air pollution is added to the mix, the endogeneity effect on housing price results in model estimation and causal inference biases.[103,109,111] We depart from previous studies by constructing a hedonic price model combining both spatial autocorrelation and endogeneity effects to examine the relationship between housing price and air pollution. To the best of our knowledge, this is the first study combining these two effects to comprehensively understand how air pollution influences housing price. We also introduce high-resolution air pollution mapping data into housing valuation studies. Prior research relied on air pollutant data from a limited number of stationary monitors to underpin estimation for a large region or a city. Our high-resolution mobile-based air pollution mapping data covers every street within the study domain, which allows us to draw on much more accurate ambient air quality measurements for each property.

2. Materials and Methods

2.1 Study area

Our study domain includes three major areas within Oakland, California: West Oakland (WO), Downtown Oakland (DO) and East Oakland (EO) (Figure 19). The WO and DO areas together cover

about 15 km^2 with residential, commercial, and industrial blocks and the EO area covers about 15 km^2 with a mix of industrial and residential blocks. The WO and DO areas have a total population of about 25,000 and EO area has a total population of about 58,000.[118]

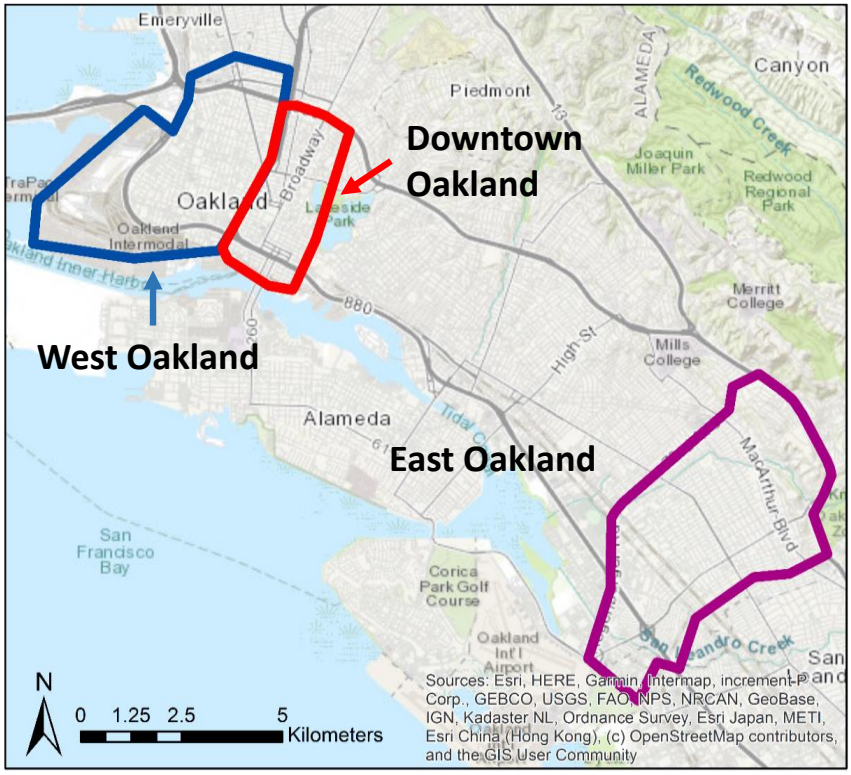

Figure 19. Study domain highlight.

2.2 Pollutant concentration and housing valuation data

Two Google street view mapping vehicles, carrying Aclima environmental intelligence sensors, were deployed in the study area between June 2015 and May 2016. The dataset covers the measurements of weekday daytime concentrations of black carbon (BC), nitric oxide (NO), and nitrogen dioxide (NO_2) with one second temporal resolution within the study area (Figure 19). A mobile-based data reduction and aggregation algorithm was developed by Apte et al. to average the instantaneous measurements into median annual weekday concentrations with 30-meter

resolution.[31,54] We use the high-resolution air pollution concentration product from Apte et al's supporting information as ambient air pollution measurement in our study.[31] Since meta-analyses have demonstrated that the spatial extent of mobile sources is on the order of 100-400 meters for particulate matter and 200-500 meters for NO_2,[8,14] we select 400 meters as the buffer size and calculate the mean air pollution concentrations within the buffer area of each property to represent the ambient air pollution concentrations. We also calculated air pollution concentrations with a 100-meter buffer and without any buffer. The results and conclusions were the same as produced with the 400-meter buffer. For the purposes of this paper, we use the 400-meter buffer air pollution concentrations calculated to ensure that we incorporate proximate roadway generated air pollution.

The housing valuation data (shown in Figure 20) is provided by Estated, Inc. (https://estated.com/) and includes land, improvement and total value for every property within our study domain. For each property, the detailed structure information includes year built, stories, room counts, parking type, construction type, and total area. Finally, social demographic variables at census tract level influencing housing price, including population density, income, and non-employment rate were assembled using the 2016 American Community Survey.

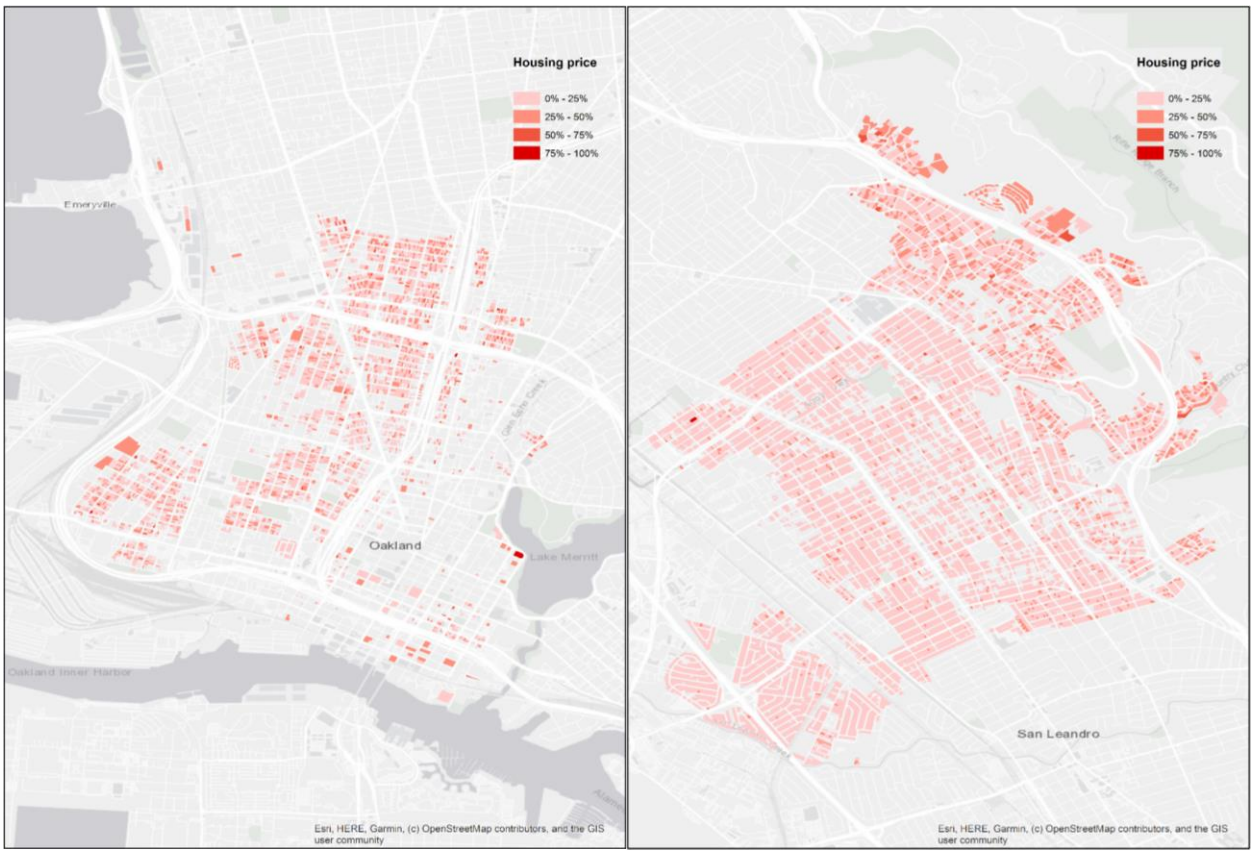

Figure 20. Housing price spatial distribution in the study domain.

2.3 Methods

Following Kim et al., in which the SLM model specification outperformed SEM on housing data in South Korea[110], we construct a spatial lag model (SLM) with an additional instrumental variable to include both the spatial autocorrelation and endogeneity effects (eq 1),

$$y = X\beta + \lambda Wy + \varepsilon, \varepsilon \sim N(0, \sigma^2) \quad (1)$$

where y is the logarithm of housing price, X's are independent variables including an instrumental variable, β are the estimated coefficients, W is the non-stochastic spatial weight matrix, Wy represents the spatial lag of the dependent variables, and ε is the error term. For the spatial weight matrix, there is no widely accepted spatial structures for housing price data, but some

studies use the queen contiguity weighting matrix since it is representative for contiguity-based weighting matrices[106]. We use the queen contiguity weighting matrix.

To address the endogeneity concern between housing price and air pollution concentrations, we combine the instrumental variable (IV) method together the SLM. We use the mean of the median vehicle speed within buffer area as the instrumental variable, which is positively related to air pollution concentrations but is not correlated with housing price.

The spatial lag term in equation 1 is an endogenous variable and the instrumental variable is an additional endogenous variable, which can result in difficulty in estimating the model coefficients estimation due to the extra endogenous variable. We use a two-step Generalized Moments (GM) and Instrumental Variable (IV) method to estimate the coefficients in equation 1.[119–123] All the calculations are conducted in R[124] and the two-step GM/IV method is available in *sphet* package with function *spreg*.[125,126]

3. Results and discussion

3.1 Variable distribution

Housing price is not normally distributed (Figure 21a), so we apply the logarithm transformation of housing price (Figure 21b). The NO, NO_2, and BC concentrations in Figure c-e are the average of measurements within 400 m buffer of each parcel. Most parcels have NO concentrations less than 40 ppb, NO_2 concentrations less than 25 ppb, and BC concentrations less than 1.5 $\mu g/m^3$.

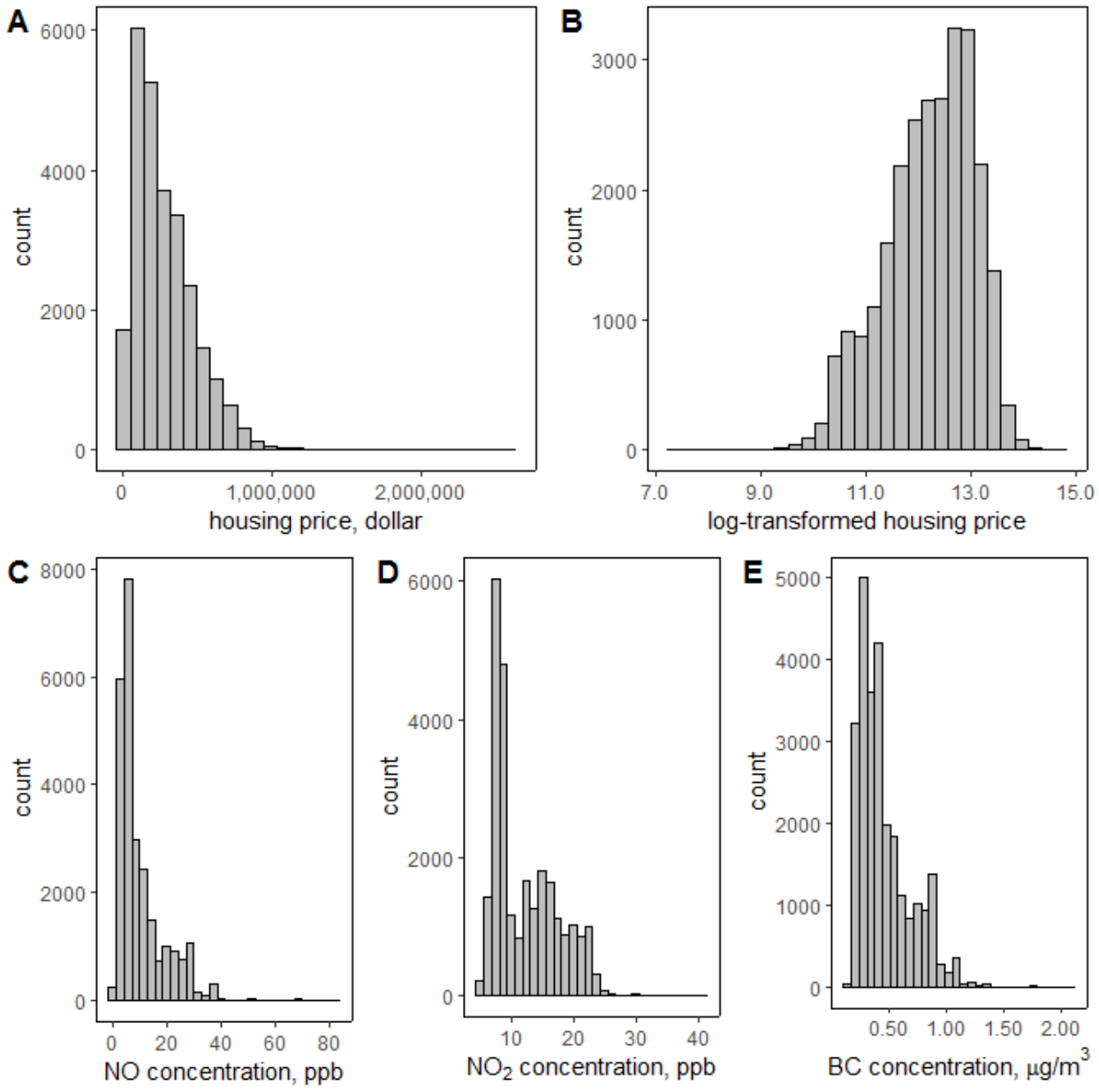

Figure 21. Distributions of housing price and concentrations of NO, NO2, and BC.

3.2 Spatial autocorrelation

We use the Moran scatter plot to examine the spatial autocorrelation of housing price and three air pollutants within the study domain (Figure 22). For Moran's I test, the test statistic is represented by the slope of the fitted line in the Moran scatter plot (Figure 22). We also use the

permutation based random Moran's I test, which uses the Monte-Carlo simulation method to randomly shuffle the data and calculate Moran's I statistic for each random shuffle and compare it with the actual Moran's I statistic. The results of both Moran's I tests for housing price and three pollutants are shown in Table 8. The housing price has Moran's I value equal to 0.276, suggesting a positive spatial autocorrelation. All of the pollutants have Moran's I values close to 0.99, suggesting highly positive autocorrelation.

Table 8. Moran's I test results for housing price and three pollutants.

	Housing price	**NO concentration**	**NO_2 concentration**	**BC concentration**
Moran's I test statistic	0.27643	0.98498	0.9927	0.99127
Analytical method p-value	<0.001	<0.001	<0.001	<0.001
Monte-Carlo based p-value	<0.001	<0.001	<0.001	<0.001

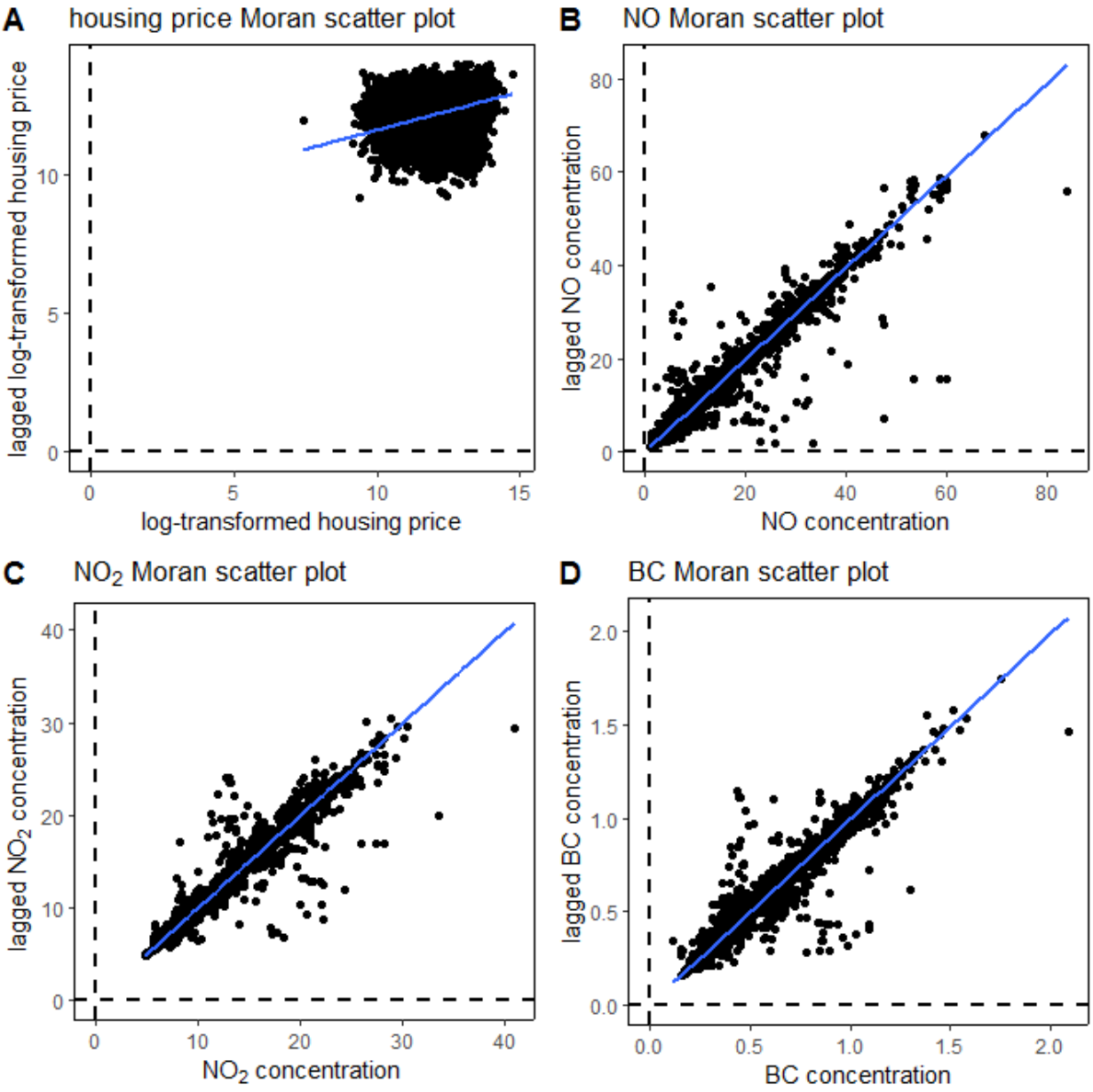

Figure 22. Moran's I scatter plots of housing price, NO, NO2, and BC concentrations (blue lines are the linear regression lines between variables and the lagged variables; the slopes of blue lines are the Moral's I statistic).

3.3 Spatial lag model results

The model results of all three pollutants are very similar (Table 9). As expected, the year the home was built negatively influences housing price and garage, bath number, total area, and median income positively influencing housing price. Air pollution concentrations positively and significantly influence housing price, which we will discuss in greater detail in section 3.4.

Table 9. Results of models with different pollutants[a].

Variables	NO concentration	NO_2 concentration	BC concentration
Intercept	2.9196*** (0.4688)	2.5027*** (0.45954)	2.8232*** (0.46502)
Year Built	-0.0070745*** (0.00033103)	-0.0068693*** (0.00032984)	-0.0070433 *** (0.0003305)
Effective Year Built	0.010137*** (0.0003401)	0.010268*** (0.00033955)	0.010166*** (0.00033986)
Construction type: concrete	-0.014669 (0.061875)	-0.0076211 (0.061662)	-0.0041038 (0.061783)
Construction type: frame	-0.3531*** (0.020988)	-0.32364*** (0.021187)	-0.34251*** (0.021)
Construction type: masonry	-0.36805*** (0.075532)	-0.32321*** (0.074869)	-0.35448*** (0.075232)
Other rooms: gym	-0.10416** (0.043856)	-0.080572* (0.04356)	-0.092731** (0.043693)
Other rooms: office	0.17428 (0.40627)	0.18374 (0.4051)	0. 18428 (0.40593)
Parking type: Carport	-0.027576 (0.029369)	-0.016681 (0.029342)	-0.020942 (0.029382)
Parking type: garage	0.051695*** (0.010082)	0.061715*** (0.010229)	0.055929*** (0.010152)
Parking type: Mixed	-0.0064772 (0.041319)	0.0046338 (0.041243)	-0.00011624 (0.04131)
Stories	0.020611*** (0.0021083)	0.017629*** (0.0021295)	0.020372*** (0.0021063)

Rooms	-0.0095808** (0.0040749)	-0.0096307** (0.0060424)	-0.094721** (0.0040706)
Beds	-0.010024 (0.0064244)	-0.0092185 (0.0064047)	-0.010209 (0.0064193)
Baths	0.084969*** (0.0086318)	0.082202*** (0.0086111)	0.08463*** (0.0086243)
Total area	0.00027102*** (0.000014127)	0.00026972*** (0.000014055)	0.0002711*** (0.000014107)
Population density	0.000014708*** (2.7835×10^{-6})	0.00017308*** (2.8254×10^{-6})	0.000018455*** (2.9621×10^{-6})
Median income	5.0184×10^{-6}*** (3.0363×10^{-7})	5.244×10^{-6}*** (2.9925×10^{-7})	5.3215×10^{-6}*** (2.9968×10^{-7})
Non-employment rate	0.07601 (0.050197)	0.031651 (0.050804)	0.081003 (0.04948)
NO concentration	0.0054361*** (0.00082701)	-	-
NO_2 concentration	-	0.013246*** (0.0016209)	-
BC concentration	-	-	0.22871*** (0.03212)
lambda	0.21710*** (0.019776)	0.18774*** (0.020526)	0.20761*** (0.020015)
R^2	0.3183	0.3175	0.3178

[a] *** significant at less than 0.1%, ** significant at less than 5%, * significant at 10%. (): standard error.

3.4 Discussion

Our models suggest that all three pollutants (NO, NO_2, and BC) have a positive and significant effect on housing price. This is unexpected and we have a few speculations as to why this occurs. First, the air pollution concentrations are low throughout the area. The average concentration of NO is 10.29 ppb, NO_2 is 12.12 ppb, and BC is 0.46 μg/m^3. We also include the air pollution

concentrations of BC and $PM_{2.5}$ from a stationary monitoring station (Oakland-west site) located in the center of West Oakland (https://ww3.arb.ca.gov/qaweb/iframe_site.php?s_arb_code=60349). For the stationary data, we calculate the mean values of the hourly measurements between June 2015 and May 2016, which covers the same time range (9 am to 5 pm) of the mobile air pollution measurement in our study. The mean concentrations of BC and $PM_{2.5}$ from the stationary monitor are 0.59 $\mu g/m^3$ and 8.36 $\mu g/m^3$, respectively. The BC concentrations are close between the stationary monitor measurement and the mobile measurement we use in this study, which gives a general estimate about the $PM_{2.5}$ concentrations across our study domain. Compare these to the National Ambient Air Quality Standards (NAAQS), the annual standard of NO_2 is at level of 53 ppb, and the annual standard of $PM_{2.5}$ is 12.0 $\mu g/m^3$ for primary source and 15.0 $\mu g/m^3$ for secondary source.[127] It is possible that when the ambient air quality is relatively clean, affordability dominates the need to pay a housing premium for even cleaner air. Of the 19 papers we found on housing and air quality, 10 papers were relevant to our research. Among these, the findings are mixed (Table 10). Three show insignificant effects of air pollution on housing price. In the remaining seven papers, the air pollution concentrations have significant and negative effect on house prices.

Table 10. Literature review summary.

Location	Air pollution concentrations									Method	air pollution impact on housing price
	CO, μg/m³	NO_2, μg/m³	O_3, μg/m³	$PM_{2.5}$, μg/m³	PM_{10}, μg/m³	SO_2, μg/m³	TSP, μg/m³	BC, μg/m³	NO, μg/m³		
Seoul, South Korea (Kim & Yoon, 2019)					45.611					SDM	insignificant
Seoul, South Korea (C. W. Kim, Phipps, & Anselin, 2003)		45.57 [a]								SLM, SEM	insignificant
						82.95					negative
18 districts in Warsaw, Poland (Ligus & Peternek, 2017)		—			—					Linear, Logarithm, SLM, SEM	insignificant [b]
Beijing, China (Mei, et al. 2020)	1399.1									Fixed-effect	negative
		60.34									negative
			53.66								positive
				88.24							negative
					111.27						negative
						20.5					negative
286 prefectural cities in China (Chen & Jin, 2019)				64.81						IV	negative

288 Chinese cities (Huang & Lanz, 2018)					77.44					IV & discontinuity regression	negative
3 largest cities in Mexico (Gonzalez, Leipnik, & Mazumder, 2013)					38.5, 51.7, 84					IV	negative
Metro areas US (Bayer et al., 2009)					42.21 (1990), 33.87 (2000)					IV	negative
all counties in US (Chay & Greenstone, 2005)							64.1 (1970), 56.3 (1980)			quasi-experimental discontinuity regression	negative
Lebanon (Marrouch & Sayour, 2021)					27.67					Fixed-effect	negative
Oakland, CA. USA		22.79						0.457	12.86	IV & SLM	positive

[a] paper reports NOx concentration in ppb and we convert it to µg/m3 with NO_2 molecular weight;

[b] insignificant in most districts, some districts are positive or negative.

All 10 papers we found use the hedonic price model to study the impact of air pollution on housing price. Their conclusions are derive from the model coefficients. If the regression coefficient of air pollution is statistically significantly less than zero, air pollution negatively influences housing price; if the coefficient is significantly greater than zero, air pollution positively influences housing price. If the coefficient is not significantly different from zero, air pollution is not significantly influencing housing price.

As we noted in the introduction even though all of the papers use the hedonic price model, the authors rely on different methods to emphasize different effects (e.g., instrumental variable (IV), spatial lag model (SLM), spatial error model (SEM), fixed effect, etc.).

In Table 10, among the three studies with insignificant results about air pollution influencing housing price, they all take the spatial autocorrelation effect into consideration when constructing the hedonic price model. In one study, the authors argue that the insignificant effect of NO_x concentrations on housing price is due to the fact that NO_x does not tend to exceed the standard; on the contrary, SO_2 shows significantly and negatively impact on housing price in the same study, because SO_2 has exceeded the official air quality standard over a long period of time .[110] While the other two studies believe that the insignificant results are caused by either insufficient degree of efficiency[105] or the change of air pollution concentration is more important than air pollution concentration itself[106].

In examining the literature, the results are suggestive that air pollution's effects tend to be insignificant when overall ambient air pollution concentrations are relatively low. In our study, the average air pollution concentrations across all of our sample observations are the lowest among these studies. It is possible that affordability is more important than a housing premium

for even cleaner air when the ambient air quality is already good. Therefore, the positive and significant coefficients of air pollution on housing price may be reasonable in area with good air quality.

A second possible reason why we find counter-intuitive results may be due to the very low variability in pollutants and housing prices. Within our buffer, standard deviations of NO, NO_2, and BC concentrations are 8.68 ppb, 5.07 ppb, and 0.23 $\mu g/m^3$, respectively. We compare the distribution of the three pollutants in our study with the stationary monitoring measurement located in the center of West Oakland (WO) in Figure 23. For the stationary data, we use the hourly measurements of NO, NO_2 and BC from the abovementioned Oakland-west site, covering the same date and time range of the mobile air pollution measurement in our study. Our data variability is close to the variation of air pollution concentrations at a single location. Low variability may lead to the positive and significant coefficients even if the results are not significant.

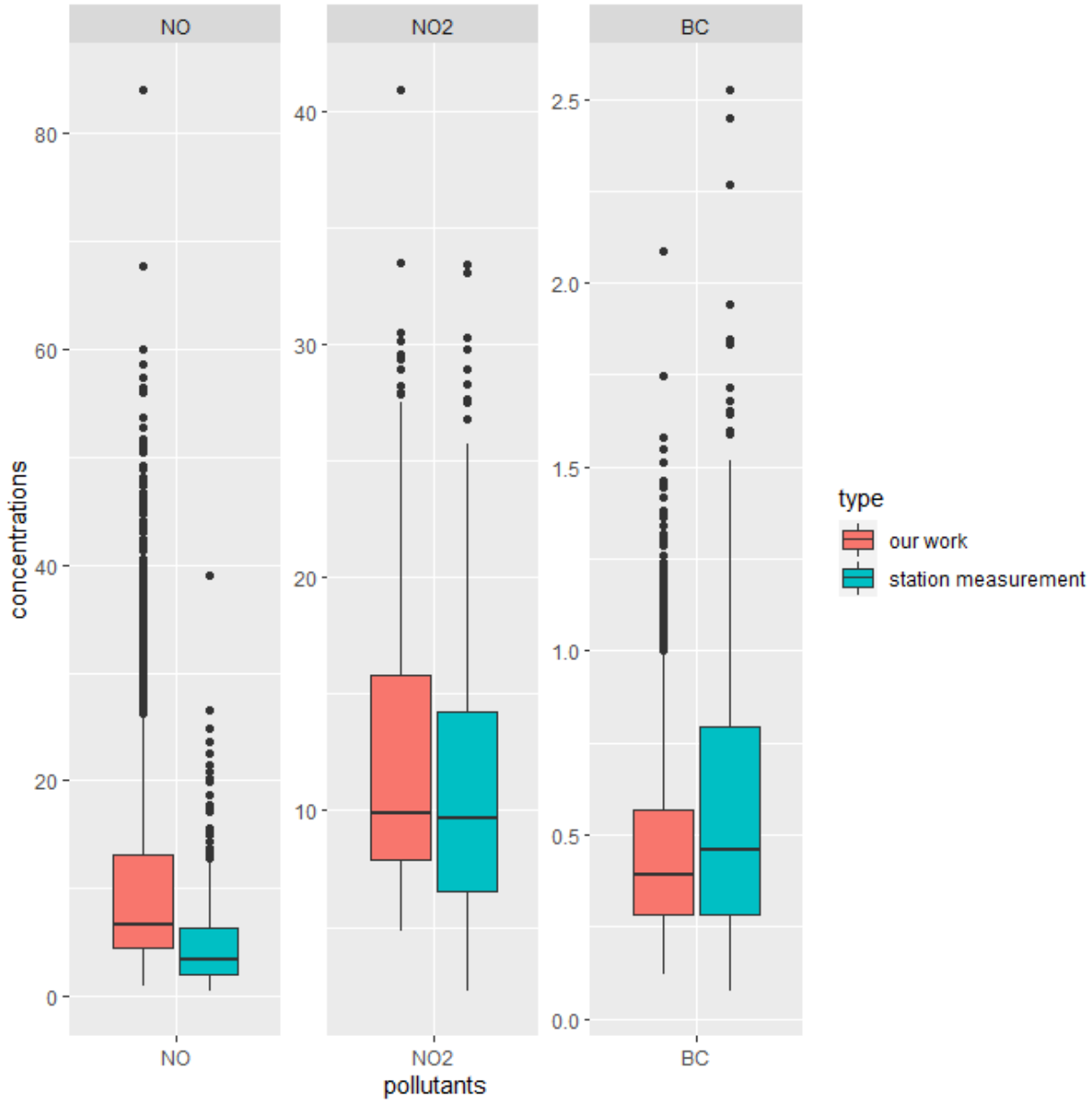

Figure 23. Pollutant distributions comparison between our work and one stationary monitor (NO and NO2 are in the unit of ppb, BC is in the unit of µg/m3).

4. Limitations and Conclusion

To the best of our knowledge, this is the first study that combines a spatial lag model with an instrumental variable method to capture both the spatial autocorrelation and endogeneity

effects of the relationship between housing price and air pollution. Our study demonstrates the use of high-resolution air pollution mapping data to quantify the localized ambient air quality at the parcel level.

Our results are counter-intuitive, suggesting that air pollution positively influences housing price in Oakland, CA. We believe this counter-intuitive result arises from two possible explanations. First, the results suggests that people may be insensitive to air quality if the overall ambient air quality is good, which is consistent with other literature we reviewed. Second, our study focuses on a relatively small study domain, where the variability of air pollution concentrations and housing price is low. The low variability of variables may lead to significant result even though the true influence is not significant.

Our results indicate that a larger, multi-regional study is probably the best way to determine the relationship between air pollution and housing price. Data from high-resolution air pollution measurement is expanding quickly. Google Earth Outreach team has conducted the high-resolution air pollution measurement in Houston, London, Copenhagen, and Amsterdam. These data may prove useful to better understand how air pollution affects housing price.

Chapter 6. Reference

(1) Apte, J. S.; Marshall, J. D.; Cohen, A. J.; Brauer, M. Addressing Global Mortality from Ambient PM2.5. *Environ. Sci. Technol.* **2015**, *49* (13), 8057–8066. https://doi.org/10.1021/acs.est.5b01236.

(2) Lim, S. S.; Vos, T.; Flaxman, A. D.; Danaei, G.; Shibuya, K.; Adair-Rohani, H.; Amann, M.; Anderson, H. R.; Andrews, K. G.; Aryee, M.; Atkinson, C.; Bacchus, L. J.; Bahalim, A. N.; Balakrishnan, K.; Balmes, J.; Barker-Collo, S.; Baxter, A.; Bell, M. L.; Blore, J. D.; Blyth, F.; Bonner, C.; Borges, G.; Bourne, R.; Boussinesq, M.; Brauer, M.; Brooks, P.; Bruce, N. G.; Brunekreef, B.; Bryan-Hancock, C.; Bucello, C.; Buchbinder, R.; Bull, F.; Burnett, R. T.; Byers, T. E.; Calabria, B.; Carapetis, J.; Carnahan, E.; Chafe, Z.; Charlson, F.; Chen, H.; Chen, J. S.; Cheng, A. T. A.; Child, J. C.; Cohen, A.; Colson, K. E.; Cowie, B. C.; Darby, S.; Darling, S.; Davis, A.; Degenhardt, L.; Dentener, F.; Des Jarlais, D. C.; Devries, K.; Dherani, M.; Ding, E. L.; Dorsey, E. R.; Driscoll, T.; Edmond, K.; Ali, S. E.; Engell, R. E.; Erwin, P. J.; Fahimi, S.; Falder, G.; Farzadfar, F.; Ferrari, A.; Finucane, M. M.; Flaxman, S.; Fowkes, F. G. R.; Freedman, G.; Freeman, M. K.; Gakidou, E.; Ghosh, S.; Giovannucci, E.; Gmel, G.; Graham, K.; Grainger, R.; Grant, B.; Gunnell, D.; Gutierrez, H. R.; Hall, W.; Hoek, H. W.; Hogan, A.; Hosgood, H. D.; Hoy, D.; Hu, H.; Hubbell, B. J.; Hutchings, S. J.; Ibeanusi, S. E.; Jacklyn, G. L.; Jasrasaria, R.; Jonas, J. B.; Kan, H.; Kanis, J. A.; Kassebaum, N.; Kawakami, N.; Khang, Y. H.; Khatibzadeh, S.; Khoo, J. P.; Kok, C.; Laden, F.; Lalloo, R.; Lan, Q.; Lathlean, T.; Leasher, J. L.; Leigh, J.; Li, Y.; Lin, J. K.; Lipshultz, S. E.; London, S.; Lozano, R.; Lu, Y.; Mak, J.; Malekzadeh, R.; Mallinger, L.; Marcenes, W.; March, L.; Marks, R.; Martin, R.; McGale, P.; McGrath, J.; Mehta, S.;

Mensah, G. A.; Merriman, T. R.; Micha, R.; Michaud, C.; Mishra, V.; Hanafiah, K. M.; Mokdad, A. A.; Morawska, L.; Mozaffarian, D.; Murphy, T.; Naghavi, M.; Neal, B.; Nelson, P. K.; Nolla, J. M.; Norman, R.; Olives, C.; Omer, S. B.; Orchard, J.; Osborne, R.; Ostro, B.; Page, A.; Pandey, K. D.; Parry, C. D. H.; Passmore, E.; Patra, J.; Pearce, N.; Pelizzari, P. M.; Petzold, M.; Phillips, M. R.; Pope, D.; Pope, C. A.; Powles, J.; Rao, M.; Razavi, H.; Rehfuess, E. A.; Rehm, J. T.; Ritz, B.; Rivara, F. P.; Roberts, T.; Robinson, C.; Rodriguez-Portales, J. A.; Romieu, I.; Room, R.; Rosenfeld, L. C.; Roy, A.; Rushton, L.; Salomon, J. A.; Sampson, U.; Sanchez-Riera, L.; Sanman, E.; Sapkota, A.; Seedat, S.; Shi, P.; Shield, K.; Shivakoti, R.; Singh, G. M.; Sleet, D. A.; Smith, E.; Smith, K. R.; Stapelberg, N. J. C.; Steenland, K.; Stöckl, H.; Stovner, L. J.; Straif, K.; Straney, L.; Thurston, G. D.; Tran, J. H.; Van Dingenen, R.; Van Donkelaar, A.; Veerman, J. L.; Vijayakumar, L.; Weintraub, R.; Weissman, M. M.; White, R. A.; Whiteford, H.; Wiersma, S. T.; Wilkinson, J. D.; Williams, H. C.; Williams, W.; Wilson, N.; Woolf, A. D.; Yip, P.; Zielinski, J. M.; Lopez, A. D.; Murray, C. J. L.; Ezzati, M. A Comparative Risk Assessment of Burden of Disease and Injury Attributable to 67 Risk Factors and Risk Factor Clusters in 21 Regions, 1990-2010: A Systematic Analysis for the Global Burden of Disease Study 2010. *Lancet* **2012**, *380* (9859), 2224–2260. https://doi.org/10.1016/S0140-6736(12)61766-8.

(3) Brunekreef, B.; Holgate, S. T. Air Pollution and Health. *Lancet* **2002**, *360* (9341), 1233–1242. https://doi.org/https://doi.org/10.1016/S0140-6736(02)11274-8.

(4) Pope, C. A.; Dockery, D. W. Health Effects of Fine Particulate Air Pollution: Lines That Connect. *J. Air Waste Manage. Assoc.* **2006**, *56* (6), 709–742. https://doi.org/10.1080/10473289.2006.10464485.

(5) Brook, R. D.; Rajagopalan, S.; Pope, C. A. 3rd; Brook, J. R.; Bhatnagar, A.; Diez-Roux, A. V; Holguin, F.; Hong, Y.; Luepker, R. V; Mittleman, M. A.; Peters, A.; Siscovick, D.; Smith, S. C. J.; Whitsel, L.; Kaufman, J. D. Particulate Matter Air Pollution and Cardiovascular Disease: An Update to the Scientific Statement from the American Heart Association. *Circulation* **2010**, *121* (21), 2331–2378. https://doi.org/10.1161/CIR.0b013e3181dbece1.

(6) 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* (12), 6999–7008. https://doi.org/10.1021/acs.est.7b00891.

(7) Brugge, D.; Durant, J. L.; Rioux, C. Near-Highway Pollutants in Motor Vehicle Exhaust: A Review of Epidemiologic Evidence of Cardiac and Pulmonary Health Risks. *Environ. Heal.* **2007**, *6*, 23. https://doi.org/10.1186/1476-069X-6-23.

(8) Karner, A. A.; Eisinger, D. S.; Niemeier, D. A. Near-Roadway Air Quality: Synthesizing the Findings from Real-World Data. *Environ. Sci. Technol.* **2010**, *44* (14), 5334–5344. https://doi.org/10.1021/es100008x.

(9) Zhu, Y.; Hinds, W. C.; Kim, S.; Shen, S.; Sioutas, C. Study of Ultrafine Particles near a Major Highway with Heavy-Duty Diesel Traffic. *Atmos. Environ.* **2002**, *36* (27), 4323–4335. https://doi.org/https://doi.org/10.1016/S1352-2310(02)00354-0.

(10) Miller, K. A.; Siscovick, D. S.; Sheppard, L.; Shepherd, K.; Sullivan, J. H.; Anderson, G. L.; Kaufman, J. D. Long-Term Exposure to Air Pollution and Incidence of Cardiovascular Events in Women. *N. Engl. J. Med.* **2007**, *356* (5), 447–458.

https://doi.org/10.1056/NEJMoa054409.

(11) Marshall, J. D.; Nethery, E.; Brauer, M. Within-Urban Variability in Ambient Air Pollution: Comparison of Estimation Methods. *Atmos. Environ.* **2008**, *42* (6), 1359–1369. https://doi.org/https://doi.org/10.1016/j.atmosenv.2007.08.012.

(12) Good, N.; Mölter, A.; Ackerson, C.; Bachand, A.; Carpenter, T.; Clark, M. L.; Fedak, K. M.; Kayne, A.; Koehler, K.; Moore, B.; L'orange, C.; Quinn, C.; Ugave, V.; Stuart, A. L.; Peel, J. L.; Volckens, J. The Fort Collins Commuter Study: Impact of Route Type and Transport Mode on Personal Exposure to Multiple Air Pollutants. *J. Expo. Sci. Environ. Epidemiol.* **2015**, *26*, 397–404. https://doi.org/10.1038/jes.2015.68.

(13) Farrell, W. J.; Weichenthal, S.; Goldberg, M.; Hatzopoulou, M. Evaluating Air Pollution Exposures across Cycling Infrastructure Types: Implications for Facility Design. *J. Transp. L. Use; Vol 8, No 3* **2015**.

(14) Zhou, Y.; Levy, J. I. Factors Influencing the Spatial Extent of Mobile Source Air Pollution Impacts: A Meta-Analysis. *BMC Public Health* **2007**, *7*, 89. https://doi.org/10.1186/1471-2458-7-89.

(15) 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*; 2001; Vol. 36. https://doi.org/10.1177/10780870122184993.

(16) CARB. California Ambient Air Quality Standards (CAAQS) https://www.arb.ca.gov/research/aaqs/caaqs/caaqs.htm.

(17) 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. https://doi.org/https://doi.org/10.1016/j.atmosenv.2008.05.057.

(18) Jerrett, M.; Arain A Fau - Kanaroglou, P.; Kanaroglou P Fau - Beckerman, B.; Beckerman B Fau - Potoglou, D.; Potoglou D Fau - Sahsuvaroglu, T.; Sahsuvaroglu T Fau - Morrison, J.; Morrison J Fau - Giovis, C.; Giovis, C.; Epidemiol, J. E. A. E.; Arain, A.; Kanaroglou, P.; Beckerman, B.; Potoglou, D.; Sahsuvaroglu, T.; Morrison, J.; Giovis, C.; Arain A Fau - Kanaroglou, P.; Kanaroglou P Fau - Beckerman, B.; Beckerman B Fau - Potoglou, D.; Potoglou D Fau - Sahsuvaroglu, T.; Sahsuvaroglu T Fau - Morrison, J.; Morrison J Fau - Giovis, C.; Giovis, C.; Epidemiol, J. E. A. E. A Review and Evaluation of Intraurban Air Pollution Exposure Models. *J. Expo. Anal. Environ. Epidemiol.* **2005**, *15* (1053-4245 (Print)), 185–204. https://doi.org/10.1038/sj.jea.7500388.

(19) Wong, D. W.; Yuan, L.; Perlin, S. A. Comparison of Spatial Interpolation Methods for the Estimation of Air Quality Data. *J. Expo. Anal. Environ. Epidemiol.* **2004**, *14* (5), 404–415. https://doi.org/10.1038/sj.jea.7500338.

(20) Ryan, P. H.; LeMasters, G. K. A Review of Land-Use Regression Models for Characterizing Intraurban Air Pollution Exposure. *Inhal. Toxicol.* **2007**, *19* (Suppl 1), 127–133. https://doi.org/10.1080/08958370701495998.

(21) Lee, J.-H.; Wu, C.-F.; Hoek, G.; de Hoogh, K.; Beelen, R.; Brunekreef, B.; Chan, C.-C. Land Use Regression Models for Estimating Individual NOx and NO(2) Exposures in a Metropolis with a High Density of Traffic Roads and Population. *Sci. Total Environ.* **2014**, *472*, 1163–

1171. https://doi.org/10.1016/j.scitotenv.2013.11.064.

(22) Wolf, K.; Cyrys, J.; Harciníková, T.; Gu, J.; Kusch, T.; Hampel, R.; Schneider, A.; Peters, A. Land Use Regression Modeling of Ultrafine Particles, Ozone, Nitrogen Oxides and Markers of Particulate Matter Pollution in Augsburg, Germany. *Sci. Total Environ.* **2017**, *579*, 1531–1540. https://doi.org/10.1016/J.SCITOTENV.2016.11.160.

(23) Tilmes, S.; Brandt, Jø.; FlatØy, F.; Bergström, R.; Flemming, J.; Langner, J.; Christensen, J. H.; Frohn, L. M.; Hov, Ø.; Jacobsen, I.; Reimer, E.; Stern, R.; Zimmermann, J. Comparison of Five Eulerian Air Pollution Forecasting Systems for the Summer of 1999 Using the German Ozone Monitoring Data. *J. Atmos. Chem.* **2002**, *42* (1), 91–121. https://doi.org/10.1023/A:1015753302760.

(24) Bellander, T.; Berglind, N.; Gustavsson, P.; Jonson, T.; Nyberg, F.; Pershagen, G.; Jarup, L. Using Geographic Information Systems to Assess Individual Historical Exposure to Air Pollution from Traffic and House Heating in Stockholm. *Environ. Health Perspect.* **2001**, *109* (6), 633–639.

(25) Ung, A.; Wald, L.; Ranchin, T.; Weber, C.; Hirsch, J.; Perron, G.; Kleinpeter, J. *Satellite Data for the Air Pollution Mapping over a City – The Use of Virtual Stations*; 2001.

(26) Rood, A. S. Performance Evaluation of AERMOD, CALPUFF, and Legacy Air Dispersion Models Using the Winter Validation Tracer Study Dataset. *Atmos. Environ.* **2014**, *89*, 707–720. https://doi.org/10.1016/j.atmosenv.2014.02.054.

(27) Rogers, R. E.; Deng, A.; Stauffer, D. R.; Gaudet, B. J.; Jia, Y.; Soong, S.-T.; Tanrikulu, S. Application of the Weather Research and Forecasting Model for Air Quality Modeling in

the San Francisco Bay Area. *J. Appl. Meteorol. Climatol.* **2013**, *52* (9), 1953–1973. https://doi.org/10.1175/JAMC-D-12-0280.1.

(28) Tanrikulu, S.; Soong, S.-T.; Tran, C.; Beaver, S. *Fine Particulate Matter Data Analysis and Modeling in the Bay Area*; 2009.

(29) USEPA. *AERMOD Implementation Guide*; 2018.

(30) 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* (30), 5880–5892. https://doi.org/https://doi.org/10.1016/j.atmosenv.2006.03.016.

(31) 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* (12), 6999–7008. https://doi.org/10.1021/acs.est.7b00891.

(32) Le, D. Van; Tham, C.-K. Machine Learning (ML)-Based Air Quality Monitoring Using Vehicular Sensor Networks. *2017 IEEE 23rd Int. Conf. Parallel Distrib. Syst.* **2017**. https://doi.org/10.1109/ICPADS.2017.00020.

(33) Kolb, C. E.; Herndon, S. C.; McManus, J. B.; Shorter, J. H.; Zahniser, M. S.; Nelson, D. D.; Jayne, J. T.; Canagaratna, M. R.; Worsnop, D. R. Mobile Laboratory with Rapid Response Instruments for Real-Time Measurements of Urban and Regional Trace Gas and Particulate Distributions and Emission Source Characteristics. *Environ. Sci. Technol.* **2004**, *38* (21), 5694–5703. https://doi.org/10.1021/es030718p.

(34) Bukowiecki, N.; Dommen, J.; Prévôt, A. S. H.; Richter, R.; Weingartner, E.; Baltensperger, U. A Mobile Pollutant Measurement Laboratory—Measuring Gas Phase and Aerosol Ambient Concentrations with High Spatial and Temporal Resolution. *Atmos. Environ.* **2002**, *36* (36), 5569–5579. https://doi.org/https://doi.org/10.1016/S1352-2310(02)00694-5.

(35) Xie, X.; Semanjski, I.; Gautama, S.; Tsiligianni, E.; Deligiannis, N.; Rajan, R.; Pasveer, F.; Philips, W. A Review of Urban Air Pollution Monitoring and Exposure Assessment Methods. *ISPRS Int. J. Geo-Information* **2017**, *6* (12), 389. https://doi.org/10.3390/ijgi6120389.

(36) Shi, Y.; Lau, K. K.-L.; Ng, E. Developing Street-Level PM2.5 and PM10 Land Use Regression Models in High-Density Hong Kong with Urban Morphological Factors. *Environ. Sci. Technol.* **2016**, *50* (15), 8178–8187. https://doi.org/10.1021/acs.est.6b01807.

(37) Hatzopoulou, M.; Valois, M. F.; Levy, I.; Mihele, C.; Lu, G.; Bagg, S.; Minet, L.; Brook, J. Robustness of Land-Use Regression Models Developed from Mobile Air Pollutant Measurements. *Environ. Sci. Technol.* **2017**, *51* (7), 3938–3947. https://doi.org/10.1021/acs.est.7b00366.

(38) Wang, A.; Fallah-Shorshani, M.; Xu, J.; Hatzopoulou, M. Characterizing Near-Road Air Pollution Using Local-Scale Emission and Dispersion Models and Validation against in-Situ Measurements. *Atmos. Environ.* **2016**, *142*, 452–464. https://doi.org/https://doi.org/10.1016/j.atmosenv.2016.08.020.

(39) 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.

https://doi.org/10.1016/j.jenvman.2015.12.012.

(40) von Schneidemesser, E.; Steinmar, K.; Weatherhead, E. C.; Bonn, B.; Gerwig, H.; Quedenau, J. Air Pollution at Human Scales in an Urban Environment: Impact of Local Environment and Vehicles on Particle Number Concentrations. *Sci. Total Environ.* **2019**, *688*, 691–700. https://doi.org/10.1016/j.scitotenv.2019.06.309.

(41) Cepeda, M.; Schoufour, J.; Freak-Poli, R.; Koolhaas, C. M.; Dhana, K.; Bramer, W. M.; Franco, O. H. Levels of Ambient Air Pollution According to Mode of Transport: A Systematic Review. *Lancet Public Heal.* **2017**, *2* (1), e23–e34. https://doi.org/10.1016/S2468-2667(16)30021-4.

(42) Wang, J.; Wu, Q.; Liu, J.; Yang, H.; Yin, M.; Chen, S.; Guo, P.; Ren, J.; Luo, X.; Linghu, W.; Huang, Q. Vehicle Emission and Atmospheric Pollution in China: Problems, Progress, and Prospects. *PeerJ* **2019**, *7*, e6932. https://doi.org/10.7717/peerj.6932.

(43) Lelieveld, J.; Evans, J. S.; Fnais, M.; Giannadaki, D.; Pozzer, A. The Contribution of Outdoor Air Pollution Sources to Premature Mortality on a Global Scale. *Nature* **2015**, *525* (7569), 367–371. https://doi.org/10.1038/nature15371.

(44) Timoshek, A.; Eisinger, D.; Bai, S.; Niemeier, D. Mobile Source Air Toxic Emissions: Sensitivity to Traffic Volume, Fleet Composition, and Average Speed. *Transp. Res. Rec.* **2010**, *2158* (1), 77–85. https://doi.org/10.3141/2158-10.

(45) BAAQMD. Bay Area Air Quality Management District - Air pollution https://www.baaqmd.gov/about-air-quality/current-air-quality/air-monitoring-data/#/airp?id=546&style=chart&zone=-1&date=2015-06-01&view=monthly (accessed

Nov 4, 2020).

(46) Ramos, C. A.; Wolterbeek, H. T.; Almeida, S. M. Air Pollutant Exposure and Inhaled Dose during Urban Commuting: A Comparison between Cycling and Motorized Modes. *Air Qual. Atmos. Heal.* **2016**, *9* (8), 867–879. https://doi.org/10.1007/s11869-015-0389-5.

(47) Adams, H. S.; Nieuwenhuijsen, M. J.; Colvile, R. N. Determinants of Fine Particle (PM 2.5) Personal Exposure Levels in Transport Microenvironments, London, UK. *Atmos. Environ.* **2001**, *35* (27), 4557–4566. https://doi.org/10.1016/S1352-2310(01)00194-7.

(48) Int Panis, L.; de Geus, B.; Vandenbulcke, G.; Willems, H.; Degraeuwe, B.; Bleux, N.; Mishra, V.; Thomas, I.; Meeusen, R. Exposure to Particulate Matter in Traffic: A Comparison of Cyclists and Car Passengers. *Atmos. Environ.* **2010**. https://doi.org/10.1016/j.atmosenv.2010.04.028.

(49) Ma, X.; Longley, I.; GAO, J.; Salmond, J. Evaluating the Effect of Ambient Concentrations, Route Choices and Environmental (In)Justice on Students' Dose of Ambient NO2 Whilst Walking to School at Population Scales. *Environ. Sci. Technol.* **2020**, No. 2. https://doi.org/10.1021/acs.est.0c05241.

(50) Briggs, D. J.; de Hoogh, K.; Morris, C.; Gulliver, J. Effects of Travel Mode on Exposures to Particulate Air Pollution. *Environ. Int.* **2008**, *34* (1), 12–22. https://doi.org/10.1016/j.envint.2007.06.011.

(51) de Nazelle, A.; Fruin, S.; Westerdahl, D.; Martinez, D.; Ripoll, A.; Kubesch, N.; Nieuwenhuijsen, M. A Travel Mode Comparison of Commuters' Exposures to Air Pollutants in Barcelona. *Atmos. Environ.* **2012**, *59*, 151–159.

https://doi.org/10.1016/j.atmosenv.2012.05.013.

(52) King, J. California cities most densely populated in U.S. https://www.sfgate.com/bayarea/place/article/California-cities-most-densely-populated-in-U-S-3436611.php#:~:text=Los Angeles is the nation's,6%2C266 people per square mile. (accessed Jan 11, 2020).

(53) VitalSigns. How long is it taking us to travel to work? http://www.vitalsigns.mtc.ca.gov/commute-time (accessed Nov 1, 2020).

(54) Google. Oakland_201506-201605_GoogleAclimaAQ www.google.com (accessed Nov 1, 2020).

(55) Sinnott, R. W. Virtues of the Haversine. *Sky Telescope* **1984**, *68* (2), 158.

(56) R Core Team. *R: A Language and Environment for Statistical Computing*; Vienna, Austria, 2018.

(57) Hijmans, R. J.; Williams, E.; Vennes, C. *Package "Geosphere";* 2019.

(58) Yong, Y.; Diez-Roux, A. V. Walking Distance by Trip Purpose and Population Subgroups Yong. *Am J Prev Med* **2013**, *43* (1), 11–19. https://doi.org/10.1016/j.amepre.2012.03.015.Walking.

(59) Google. Google maps direction http://maps.google.com (accessed Nov 1, 2020).

(60) Cooley, D.; Barcelos, P.; Rstudio. *Package "Googleway": Accesses Google Maps APIs to Retrieve Data and Plot Maps*; 2018.

(61) PinchFlat. How Far Would You Bike To Your Job? Here's What I Did http://www.pinch-

flat.com/how-far-is-too-far-to-bike-to-work/ (accessed Nov 1, 2020).

(62) Oswald, F. Bicycle Commuter's Guide http://www.bicyclinglife.com/PracticalCycling/commuteguide.htm (accessed Nov 1, 2010).

(63) Zuurbier, M.; Hoek, G.; Hazel, P. Van Den; Brunekreef, B. Minute Ventilation of Cyclists, Car and Bus Passengers: An Experimental Study. **2009**, *10*, 1–10. https://doi.org/10.1186/1476-069X-8-48.

(64) Hudda, N.; Eckel, S. P.; Knibbs, L. D.; Sioutas, C.; Delfino, R. J.; Fruin, S. A. Linking In-Vehicle Ultrafine Particle Exposures to on-Road Concentrations. *Atmos. Environ.* **2012**, *59*, 578–586. https://doi.org/10.1016/j.atmosenv.2012.05.021.

(65) Onat, B.; Stakeeva, B. Personal Exposure of Commuters in Public Transport to PM2.5 and Fine Particle Counts. *Atmos. Pollut. Res.* **2013**, *4* (3), 329–335. https://doi.org/10.5094/APR.2013.037.

(66) Song, W. W.; Ashmore, M. R.; Terry, A. C. The Influence of Passenger Activities on Exposure to Particles inside Buses. *Atmos. Environ.* **2009**, *43* (39), 6271–6278. https://doi.org/10.1016/j.atmosenv.2009.05.004.

(67) Google. Google Earth Outreach -- Air Quality https://www.google.com/earth/outreach/special-projects/air-quality/ (accessed Nov 4, 2020).

(68) Miller, K. A.; Siscovick, D. S.; Sheppard, L.; Shepherd, K.; Sullivan, J. H.; Anderson, G. L.; Kaufman, J. D.; Siscovick Ds Fau - Sheppard, L.; Sheppard L Fau - Shepherd, K.; Shepherd K

Fau - Sullivan, J. H.; Sullivan Jh Fau - Anderson, G. L.; Anderson Gl Fau - Kaufman, J. D.; Kaufman, J. D.; Med, N. E. J. Long-Term Exposure to Air Pollution and Incidence of Cardiovascular Events in Women. *N. Engl. J. Med.* **2007**, *356* (1533-4406 (Electronic)), 447–458. https://doi.org/10.1056/NEJMoa054409.

(69) Westerdahl, D.; Fruin, S.; Sax, T.; Fine, P. M.; Sioutas, C. Mobile Platform Measurements of Ultrafine Particles and Associated Pollutant Concentrations on Freeways and Residential Streets in Los Angeles. *Atmos. Environ.* **2005**, *39* (20), 3597–3610. https://doi.org/https://doi.org/10.1016/j.atmosenv.2005.02.034.

(70) Larson, T.; Henderson, S. B.; Brauer, M. Mobile Monitoring of Particle Light Absorption Coefficient in an Urban Area as a Basis for Land Use Regression. *Environ. Sci. Technol.* **2009**, *43* (13), 4672–4678. https://doi.org/10.1021/es803068e.

(71) Sabaliauskas, K.; Jeong, C. H.; Yao, X.; Reali, C.; Sun, T.; Evans, G. J. Development of a Land-Use Regression Model for Ultrafine Particles in Toronto, Canada. *Atmos. Environ.* **2015**, *110*, 84–92. https://doi.org/10.1016/j.atmosenv.2015.02.018.

(72) Rivera, M.; Basagaña, X.; Aguilera, I.; Agis, D.; Bouso, L.; Foraster, M.; Medina-Ramón, M.; Pey, J.; Künzli, N.; Hoek, G. Spatial Distribution of Ultrafine Particles in Urban Settings: A Land Use Regression Model. *Atmos. Environ.* **2012**, *54*, 657–666. https://doi.org/10.1016/j.atmosenv.2012.01.058.

(73) Van den Hove, A.; Verwaeren, J.; Van den Bossche, J.; Theunis, J.; De Baets, B. Development of a Land Use Regression Model for Black Carbon Using Mobile Monitoring Data and Its Application to Pollution-Avoiding Routing. *Environ. Res.* **2020**, *183*, 108619.

https://doi.org/10.1016/j.envres.2019.108619.

(74) Hagemann, R.; Corsmeier, U.; Kottmeier, C.; Rinke, R.; Wieser, A.; Vogel, B. Spatial Variability of Particle Number Concentrations and NOx in the Karlsruhe (Germany) Area Obtained with the Mobile Laboratory "AERO-TRAM." *Atmos. Environ.* **2014**, *94* (x), 341–352. https://doi.org/10.1016/j.atmosenv.2014.05.051.

(75) Hasenfratz, D.; Saukh, O.; Walser, C.; Hueglin, C.; Fierz, M.; Arn, T.; Beutel, J.; Thiele, L. Deriving High-Resolution Urban Air Pollution Maps Using Mobile Sensor Nodes. *Pervasive Mob. Comput.* **2015**, *16*, 268–285. https://doi.org/https://doi.org/10.1016/j.pmcj.2014.11.008.

(76) Abernethy, R. C.; Allen, R. W.; McKendry, I. G.; Brauer, M. A Land Use Regression Model for Ultrafine Particles in Vancouver, Canada. *Environ. Sci. Technol.* **2013**, *47* (10), 5217–5225. https://doi.org/10.1021/es304495s.

(77) Hankey, S.; Marshall, J. D. Land Use Regression Models of On-Road Particulate Air Pollution (Particle Number, Black Carbon, PM2.5, Particle Size) Using Mobile Monitoring. *Environ. Sci. Technol.* **2015**, *49* (15), 9194–9202. https://doi.org/10.1021/acs.est.5b01209.

(78) Kerckhoffs, J.; Hoek, G.; Messier, K. P.; Brunekreef, B.; Meliefste, K.; Klompmaker, J. O.; Vermeulen, R. Comparison of Ultrafine Particle and Black Carbon Concentration Predictions from a Mobile and Short-Term Stationary Land-Use Regression Model. *Environ. Sci. Technol.* **2016**, *50* (23), 12894–12902. https://doi.org/10.1021/acs.est.6b03476.

(79) Mueller, M. D.; Hasenfratz, D.; Saukh, O.; Fierz, M.; Hueglin, C. Statistical Modelling of Particle Number Concentration in Zurich at High Spatio-Temporal Resolution Utilizing Data

from a Mobile Sensor Network. *Atmos. Environ.* **2016**, *126*, 171–181. https://doi.org/10.1016/j.atmosenv.2015.11.033.

(80) Weichenthal, S.; Van Ryswyk, K.; Goldstein, A.; Shekarrizfard, M.; Hatzopoulou, M. Characterizing the Spatial Distribution of Ambient Ultrafine Particles in Toronto, Canada: A Land Use Regression Model. *Environ. Pollut.* **2016**, *208*, 241–248. https://doi.org/10.1016/j.envpol.2015.04.011.

(81) Klompmaker, J. O.; Montagne, D. R.; Meliefste, K.; Hoek, G.; Brunekreef, B. Spatial Variation of Ultrafine Particles and Black Carbon in Two Cities: Results from a Short-Term Measurement Campaign. *Sci. Total Environ.* **2015**, *508*, 266–275. https://doi.org/10.1016/j.scitotenv.2014.11.088.

(82) Messier, K. P.; Chambliss, S. E.; Gani, S.; Alvarez, R.; Brauer, M.; Choi, J. J.; Hamburg, S. P.; Kerckhoffs, J.; LaFranchi, B.; Lunden, M. M.; Marshall, J. D.; Portier, C. J.; Roy, A.; Szpiro, A. A.; Vermeulen, R. C. H.; Apte, J. S. Mapping Air Pollution with Google Street View Cars: Efficient Approaches with Mobile Monitoring and Land Use Regression. *Environ. Sci. Technol.* **2018**, *52* (21), 12563–12572. https://doi.org/10.1021/acs.est.8b03395.

(83) Van Rossum, G.; Drake Jr, F. L. *Python Tutorial*; Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands, 1995.

(84) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-Learn: Machine Learning in {P}ython. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(85) Bisong, E. Google Colaboratory BT - Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners; Bisong, E., Ed.; Apress: Berkeley, CA, 2019; pp 59–64. https://doi.org/10.1007/978-1-4842-4470-8_7.

(86) Bergstra, J.; Yamins, D.; Cox, D. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. In *International conference on machine learning*; 2013; pp 115–123.

(87) Azadkia, M.; Chatterjee, S. A Simple Measure of Conditional Dependence. **2019**, 1–35.

(88) Zhang, D.; Xiao, J.; Zhou, N.; Zheng, M.; Luo, X.; Jiang, H.; Chen, K. A Genetic Algorithm Based Support Vector Machine Model for Blood-Brain Barrier Penetration Prediction. *Biomed Res. Int.* **2015**, *2015*. https://doi.org/10.1155/2015/292683.

(89) 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* (March), 105022. https://doi.org/10.1016/j.envint.2019.105022.

(90) Ren, X.; Mi, Z.; Georgopoulos, P. G. Comparison of Machine Learning and Land Use Regression for Fine Scale Spatiotemporal Estimation of Ambient Air Pollution: Modeling Ozone Concentrations across the Contiguous United States. *Environ. Int.* **2020**, *142* (January), 105827. https://doi.org/10.1016/j.envint.2020.105827.

(91) Contributors, O. S. M. OpenStreetMap https://www.openstreetmap.org/ (accessed Apr 1, 2020).

(92) Davidlok. Oakland Truck Routes (2017)

https://services.arcgis.com/9tC74aDHuml0x5Yz/arcgis/rest/services/Oakland_Truck_Routes_(2017)/FeatureServer (accessed Mar 1, 2020).

(93) City of Oakland. Planning and Zoning Map https://oakgis.maps.arcgis.com/apps/webappviewer/index.html?id=3676148ea4924fc7b75e7350903c7224 (accessed Mar 1, 2020).

(94) Gorelick, Noel and Hancher, Matt and Dixon, Mike and Ilyushchenko, Simon and Thau, David and Moore, R. Google Earth Engine: Planetary-Scale Geospatial Analysis for Everyone. *Remote Sens. Environ.* **2017**. https://doi.org/10.1016/j.rse.2017.06.031.

(95) USGS. NLCD 2016 Land Cover (CONUS) https://www.mrlc.gov/data/nlcd-2016-land-cover-conus (accessed Mar 1, 2020).

(96) Yang, L.; Jin, S.; Danielson, P.; Homer, C.; Gass, L.; Bender, S. M.; Case, A.; Costello, C.; Dewitz, J.; Fry, J.; Funk, M.; Granneman, B.; Liknes, G. C.; Rigge, M.; Xian, G. A New Generation of the United States National Land Cover Database: Requirements, Research Priorities, Design, and Implementation Strategies. *ISPRS J. Photogramm. Remote Sens.* **2018**, *146*, 108–123. https://doi.org/https://doi.org/10.1016/j.isprsjprs.2018.09.006.

(97) U.S. Census Bureau. Population https://www.census.gov/en.html (accessed Mar 1, 2010).

(98) Gesch, D. B.; Evans, G. A.; Oimoen, M. J.; Arundel, S. The National Elevation Dataset; American Society for Photogrammetry and Remote Sensing, 2018; pp 83–110.

(99) Zhang, D.; Xiao, J.; Zhou, N.; Zheng, M.; Luo, X.; Jiang, H.; Chen, K. A Genetic Algorithm Based Support Vector Machine Model for Blood-Brain Barrier Penetration Prediction. **2015**. https://doi.org/10.1155/2015/292683.

(100) Lim, S. S.; Vos T Fau - Flaxman, A. D.; Flaxman Ad Fau - Danaei, G.; Danaei G Fau - Shibuya, K.; Shibuya K Fau - Adair-Rohani, H.; Adair-Rohani H Fau - Amann, M.; Amann M Fau - Anderson, H. R.; Anderson Hr Fau - Andrews, K. G.; Andrews Kg Fau - Aryee, M.; Aryee M Fau - Atkinson, C.; Atkinson C Fau - Bacchus, L. J.; Bacchus Lj Fau - Bahalim, A. N.; Bahalim An Fau - Balakrishnan, K.; Balakrishnan K Fau - Balmes, J.; Balmes J Fau - Barker-Collo, S.; Barker-Collo S Fau - Baxter, A.; Baxter A Fau - Bell, M. L.; Bell Ml Fau - Blore, J. D.; Blore Jd Fau - Blyth, F.; Blyth F Fau - Bonner, C.; Bonner C Fau - Borges, G.; Borges G Fau - Bourne, R.; Bourne R Fau - Boussinesq, M.; Boussinesq M Fau - Brauer, M.; Brauer M Fau - Brooks, P.; Brooks P Fau - Bruce, N. G.; Bruce Ng Fau - Brunekreef, B.; Brunekreef B Fau - Bryan-Hancock, C.; Bryan-Hancock C Fau - Bucello, C.; Bucello C Fau - Buchbinder, R.; Buchbinder R Fau - Bull, F.; Bull F Fau - Burnett, R. T.; Burnett Rt Fau - Byers, T. E.; Byers Te Fau - Calabria, B.; Calabria B Fau - Carapetis, J.; Carapetis J Fau - Carnahan, E.; Carnahan E Fau - Chafe, Z.; Chafe Z Fau - Charlson, F.; Charlson F Fau - Chen, H.; Chen H Fau - Chen, J. S.; Chen Js Fau - Cheng, A. T.-A.; Cheng At Fau - Child, J. C.; Child Jc Fau - Cohen, A.; Cohen A Fau - Colson, K. E.; Colson Ke Fau - Cowie, B. C.; Cowie Bc Fau - Darby, S.; Darby S Fau - Darling, S.; Darling S Fau - Davis, A.; Davis A Fau - Degenhardt, L.; Degenhardt L Fau - Dentener, F.; Dentener F Fau - Des Jarlais, D. C.; Des Jarlais Dc Fau - Devries, K.; Devries K Fau - Dherani, M.; Dherani M Fau - Ding, E. L.; Ding El Fau - Dorsey, E. R.; Dorsey Er Fau - Driscoll, T.; Driscoll T Fau - Edmond, K.; Edmond K Fau - Ali, S. E.; Ali Se Fau - Engell, R. E.; Engell Re Fau - Erwin, P. J.; Erwin Pj Fau - Fahimi, S.; Fahimi S Fau - Falder, G.; Falder G Fau - Farzadfar, F.; Farzadfar F Fau - Ferrari, A.; Ferrari A Fau - Finucane, M. M.; Finucane Mm Fau - Flaxman, S.; Flaxman S Fau - Fowkes, F. G. R.; Fowkes Fg Fau - Freedman, G.; Freedman G Fau - Freeman, M. K.;

Freeman Mk Fau - Gakidou, E.; Gakidou E Fau - Ghosh, S.; Ghosh S Fau - Giovannucci, E.; Giovannucci E Fau - Gmel, G.; Gmel G Fau - Graham, K.; Graham K Fau - Grainger, R.; Grainger R Fau - Grant, B.; Grant B Fau - Gunnell, D.; Gunnell D Fau - Gutierrez, H. R.; Gutierrez Hr Fau - Hall, W.; Hall W Fau - Hoek, H. W.; Hoek Hw Fau - Hogan, A.; Hogan A Fau - Hosgood 3rd, H. D.; Hosgood Hd 3rd Fau - Hoy, D.; Hoy D Fau - Hu, H.; Hu H Fau - Hubbell, B. J.; Hubbell Bj Fau - Hutchings, S. J.; Hutchings Sj Fau - Ibeanusi, S. E.; Ibeanusi Se Fau - Jacklyn, G. L.; Jacklyn Gl Fau - Jasrasaria, R.; Jasrasaria R Fau - Jonas, J. B.; Jonas Jb Fau - Kan, H.; Kan H Fau - Kanis, J. A.; Kanis Ja Fau - Kassebaum, N.; Kassebaum N Fau - Kawakami, N.; Kawakami N Fau - Khang, Y.-H.; Khang Yh Fau - Khatibzadeh, S.; Khatibzadeh S Fau - Khoo, J.-P.; Khoo Jp Fau - Kok, C.; Kok C Fau - Laden, F.; Laden F Fau - Lalloo, R.; Lalloo R Fau - Lan, Q.; Lan Q Fau - Lathlean, T.; Lathlean T Fau - Leasher, J. L.; Leasher Jl Fau - Leigh, J.; Leigh J Fau - Li, Y.; Li Y Fau - Lin, J. K.; Lin Jk Fau - Lipshultz, S. E.; Lipshultz Se Fau - London, S.; London S Fau - Lozano, R.; Lozano R Fau - Lu, Y.; Lu Y Fau - Mak, J.; Mak J Fau - Malekzadeh, R.; Malekzadeh R Fau - Mallinger, L.; Mallinger L Fau - Marcenes, W.; Marcenes W Fau - March, L.; March L Fau - Marks, R.; Marks R Fau - Martin, R.; Martin R Fau - McGale, P.; McGale P Fau - McGrath, J.; McGrath J Fau - Mehta, S.; Mehta S Fau - Mensah, G. A.; Mensah Ga Fau - Merriman, T. R.; Merriman Tr Fau - Micha, R.; Micha R Fau - Michaud, C.; Michaud C Fau - Mishra, V.; Mishra V Fau - Mohd Hanafiah, K.; Mohd Hanafiah K Fau - Mokdad, A. A.; Mokdad Aa Fau - Morawska, L.; Morawska L Fau - Mozaffarian, D.; Mozaffarian D Fau - Murphy, T.; Murphy T Fau - Naghavi, M.; Naghavi M Fau - Neal, B.; Neal B Fau - Nelson, P. K.; Nelson Pk Fau - Nolla, J. M.; Nolla Jm Fau - Norman, R.; Norman R Fau - Olives, C.; Olives C Fau - Omer, S. B.; Omer Sb Fau - Orchard, J.; Orchard

J Fau - Osborne, R.; Osborne R Fau - Ostro, B.; Ostro B Fau - Page, A.; Page A Fau - Pandey, K. D.; Pandey Kd Fau - Parry, C. D. H.; Parry Cd Fau - Passmore, E.; Passmore E Fau - Patra, J.; Patra J Fau - Pearce, N.; Pearce N Fau - Pelizzari, P. M.; Pelizzari Pm Fau - Petzold, M.; Petzold M Fau - Phillips, M. R.; Phillips Mr Fau - Pope, D.; Pope D Fau - Pope 3rd, C. A.; Pope Ca 3rd Fau - Powles, J.; Powles J Fau - Rao, M.; Rao M Fau - Razavi, H.; Razavi H Fau - Rehfuess, E. A.; Rehfuess Ea Fau - Rehm, J. T.; Rehm Jt Fau - Ritz, B.; Ritz B Fau - Rivara, F. P.; Rivara Fp Fau - Roberts, T.; Roberts T Fau - Robinson, C.; Robinson C Fau - Rodriguez-Portales, J. A.; Rodriguez-Portales Ja Fau - Romieu, I.; Romieu I Fau - Room, R.; Room R Fau - Rosenfeld, L. C.; Rosenfeld Lc Fau - Roy, A.; Roy A Fau - Rushton, L.; Rushton L Fau - Salomon, J. A.; Salomon Ja Fau - Sampson, U.; Sampson U Fau - Sanchez-Riera, L.; Sanchez-Riera L Fau - Sanman, E.; Sanman E Fau - Sapkota, A.; Sapkota A Fau - Seedat, S.; Seedat S Fau - Shi, P.; Shi P Fau - Shield, K.; Shield K Fau - Shivakoti, R.; Shivakoti R Fau - Singh, G. M.; Singh Gm Fau - Sleet, D. A.; Sleet Da Fau - Smith, E.; Smith E Fau - Smith, K. R.; Smith Kr Fau - Stapelberg, N. J. C.; Stapelberg Nj Fau - Steenland, K.; Steenland K Fau - Stockl, H.; Stockl H Fau - Stovner, L. J.; Stovner Lj Fau - Straif, K.; Straif K Fau - Straney, L.; Straney L Fau - Thurston, G. D.; Thurston Gd Fau - Tran, J. H.; Tran Jh Fau - Van Dingenen, R.; Van Dingenen R Fau - van Donkelaar, A.; van Donkelaar A Fau - Veerman, J. L.; Veerman Jl Fau - Vijayakumar, L.; Vijayakumar L Fau - Weintraub, R.; Weintraub R Fau - Weissman, M. M.; Weissman Mm Fau - White, R. A.; White Ra Fau - Whiteford, H.; Whiteford H Fau - Wiersma, S. T.; Wiersma St Fau - Wilkinson, J. D.; Wilkinson Jd Fau - Williams, H. C.; Williams Hc Fau - Williams, W.; Williams W Fau - Wilson, N.; Wilson N Fau - Woolf, A. D.; Woolf Ad Fau - Yip, P.; Yip P Fau - Zielinski, J. M.; Zielinski Jm Fau - Lopez, A. D.; Lopez Ad Fau - Murray, C.

J. L.; Murray Cj Fau - Ezzati, M.; Ezzati M Fau - AlMazroa, M. A.; AlMazroa Ma Fau - Memish, Z. A.; Memish, Z. A.; Lancet; Vos, T.; Flaxman, A. D.; Danaei, G.; Shibuya, K.; Adair-Rohani, H.; Amann, M.; Anderson, H. R.; Andrews, K. G.; Aryee, M.; Atkinson, C.; Bacchus, L. J.; Bahalim, A. N.; Balakrishnan, K.; Balmes, J.; Barker-Collo, S.; Baxter, A.; Bell, M. L.; Blore, J. D.; Blyth, F.; Bonner, C.; Borges, G.; Bourne, R.; Boussinesq, M.; Brauer, M.; Brooks, P.; Bruce, N. G.; Brunekreef, B.; Bryan-Hancock, C.; Bucello, C.; Buchbinder, R.; Bull, F.; Burnett, R. T.; Byers, T. E.; Calabria, B.; Carapetis, J.; Carnahan, E.; Chafe, Z.; Charlson, F.; Chen, H.; Chen, J. S.; Cheng, A. T. A.; Child, J. C.; Cohen, A.; Colson, K. E.; Cowie, B. C.; Darby, S.; Darling, S.; Davis, A.; Degenhardt, L.; Dentener, F.; Des Jarlais, D. C.; Devries, K.; Dherani, M.; Ding, E. L.; Dorsey, E. R.; Driscoll, T.; Edmond, K.; Ali, S. E.; Engell, R. E.; Erwin, P. J.; Fahimi, S.; Falder, G.; Farzadfar, F.; Ferrari, A.; Finucane, M. M.; Flaxman, S.; Fowkes, F. G. R.; Freedman, G.; Freeman, M. K.; Gakidou, E.; Ghosh, S.; Giovannucci, E.; Gmel, G.; Graham, K.; Grainger, R.; Grant, B.; Gunnell, D.; Gutierrez, H. R.; Hall, W.; Hoek, H. W.; Hogan, A.; Hosgood, H. D.; Hoy, D.; Hu, H.; Hubbell, B. J.; Hutchings, S. J.; Ibeanusi, S. E.; Jacklyn, G. L.; Jasrasaria, R.; Jonas, J. B.; Kan, H.; Kanis, J. A.; Kassebaum, N.; Kawakami, N.; Khang, Y. H.; Khatibzadeh, S.; Khoo, J. P.; Kok, C.; Laden, F.; Lalloo, R.; Lan, Q.; Lathlean, T.; Leasher, J. L.; Leigh, J.; Li, Y.; Lin, J. K.; Lipshultz, S. E.; London, S.; Lozano, R.; Lu, Y.; Mak, J.; Malekzadeh, R.; Mallinger, L.; Marcenes, W.; March, L.; Marks, R.; Martin, R.; McGale, P.; McGrath, J.; Mehta, S.; Mensah, G. A.; Merriman, T. R.; Micha, R.; Michaud, C.; Mishra, V.; Hanafiah, K. M.; Mokdad, A. A.; Morawska, L.; Mozaffarian, D.; Murphy, T.; Naghavi, M.; Neal, B.; Nelson, P. K.; Nolla, J. M.; Norman, R.; Olives, C.; Omer, S. B.; Orchard, J.; Osborne, R.; Ostro, B.; Page, A.; Pandey, K. D.; Parry, C. D. H.; Passmore, E.; Patra, J.; Pearce, N.;

Pelizzari, P. M.; Petzold, M.; Phillips, M. R.; Pope, D.; Pope, C. A.; Powles, J.; Rao, M.; Razavi, H.; Rehfuess, E. A.; Rehm, J. T.; Ritz, B.; Rivara, F. P.; Roberts, T.; Robinson, C.; Rodriguez-Portales, J. A.; Romieu, I.; Room, R.; Rosenfeld, L. C.; Roy, A.; Rushton, L.; Salomon, J. A.; Sampson, U.; Sanchez-Riera, L.; Sanman, E.; Sapkota, A.; Seedat, S.; Shi, P.; Shield, K.; Shivakoti, R.; Singh, G. M.; Sleet, D. A.; Smith, E.; Smith, K. R.; Stapelberg, N. J. C.; Steenland, K.; Stöckl, H.; Stovner, L. J.; Straif, K.; Straney, L.; Thurston, G. D.; Tran, J. H.; Van Dingenen, R.; Van Donkelaar, A.; Veerman, J. L.; Vijayakumar, L.; Weintraub, R.; Weissman, M. M.; White, R. A.; Whiteford, H.; Wiersma, S. T.; Wilkinson, J. D.; Williams, H. C.; Williams, W.; Wilson, N.; Woolf, A. D.; Yip, P.; Zielinski, J. M.; Lopez, A. D.; Murray, C. J. L.; Ezzati, M. A Comparative Risk Assessment of Burden of Disease and Injury Attributable to 67 Risk Factors and Risk Factor Clusters in 21 Regions, 1990-2010: A Systematic Analysis for the Global Burden of Disease Study 2010. *Lancet* **2012**, *380* (9859), 2224–2260. https://doi.org/10.1016/S0140-6736(12)61766-8.

(101) ELLIS ROBINSON. How much does air pollution cost the U.S.? https://earth.stanford.edu/news/how-much-does-air-pollution-cost-us#gs.6njexm.

(102) Tschofen, P.; Azevedo, I. L.; Muller, N. Z. Fine Particulate Matter Damages and Value Added in the US Economy. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116* (40), 19857–19862. https://doi.org/10.1073/pnas.1905030116.

(103) Betz, T.; Cook, S. J.; Hollenbach, F. M. Spatial Interdependence and Instrumental Variable Models. *Polit. Sci. Res. Methods* **2020**, *8* (4), 646–661. https://doi.org/10.1017/psrm.2018.61.

(104) Gonzalez, F.; Leipnik, M.; Mazumder, D. How Much Are Urban Residents in Mexico Willing to Pay for Cleaner Air? *Environ. Dev. Econ.* **2013**, *18* (3), 354–379. https://doi.org/10.1017/S1355770X13000077.

(105) Ligus, M.; Peternek, P. Impacts of Urban Environmental Attributes on Residential Housing Prices in Warsaw (Poland): Spatial Hedonic Analysis of City Districts BT - Contemporary Trends and Challenges in Finance; Jajuga, K., Orlowski, L. T., Staehr, K., Eds.; Springer International Publishing: Cham, 2017; pp 155–164.

(106) Kim, S. G.; Yoon, S. Measuring the Value of Airborne Particulate Matter Reduction in Seoul. *Air Qual. Atmos. Heal.* **2019**, *12* (5), 549–560. https://doi.org/10.1007/s11869-019-00668-x.

(107) Montero, J. M.; Mínguez, R.; Fernández-Avilés, G. Housing Price Prediction: Parametric versus Semi-Parametric Spatial Hedonic Models. *J. Geogr. Syst.* **2018**, *20* (1), 27–55. https://doi.org/10.1007/s10109-017-0257-y.

(108) De, U. K.; Vupru, V. Location and Neighbourhood Conditions for Housing Choice and Its Rental Value. *Int. J. Hous. Mark. Anal.* **2020**, *10* (4), 519–538. https://doi.org/10.1108/IJHMA-10-2016-0072.

(109) Sun, B.; Yang, S. Asymmetric and Spatial Non-Stationary Effects of Particulate Air Pollution on Urban Housing Prices in Chinese Cities. *Int. J. Environ. Res. Public Health* **2020**, *17* (20), 1–23. https://doi.org/10.3390/ijerph17207443.

(110) Won Kim, C.; Phipps, T. T.; Anselin, L. Measuring the Benefits of Air Quality Improvement: A Spatial Hedonic Approach. *J. Environ. Econ. Manage.* **2003**, *45* (1), 24–39.

https://doi.org/https://doi.org/10.1016/S0095-0696(02)00013-X.

(111) Chen, S.; Jin, H. Pricing for the Clean Air: Evidence from Chinese Housing Market. *J. Clean. Prod.* **2019**, *206*, 297–306. https://doi.org/10.1016/j.jclepro.2018.08.220.

(112) Bayer, P.; Keohane, N.; Timmins, C. Migration and Hedonic Valuation: The Case of Air Quality. *J. Environ. Econ. Manage.* **2009**, *58* (1), 1–14. https://doi.org/10.1016/j.jeem.2008.08.004.

(113) Mei, Y.; Gao, L.; Zhang, J.; Wang, J. Valuing Urban Air Quality: A Hedonic Price Analysis in Beijing, China. *Environ. Sci. Pollut. Res.* **2020**, *27* (2), 1373–1385. https://doi.org/10.1007/s11356-019-06874-5.

(114) Chay, K. Y.; Greenstone, M. Does Air Quality Matter? Evidence from the Housing Market. *J. Polit. Econ.* **2005**, *113* (2), 376–424. https://doi.org/10.1086/427462.

(115) MORAN, P. A. P. NOTES ON CONTINUOUS STOCHASTIC PHENOMENA. *Biometrika* **1950**, *37* (1–2), 17–23. https://doi.org/10.1093/biomet/37.1-2.17.

(116) Liu, R.; Yu, C.; Liu, C.; Jiang, J.; Xu, J. Impacts of Haze on Housing Prices: An Empirical Analysis Based on Data from Chengdu (China). *Int. J. Environ. Res. Public Health* **2018**, *15* (6). https://doi.org/10.3390/ijerph15061161.

(117) Bae, C. H. C.; Sandlin, G.; Bassok, A.; Kim, S. The Exposure of Disadvantaged Populations in Freeway Air-Pollution Sheds: A Case Study of the Seattle and Portland Regions. *Environ. Plan. B Plan. Des.* **2007**, *34* (1), 154–170. https://doi.org/10.1068/b32124.

(118) Poin2. Point2 https://www.point2homes.com/US/Neighborhood/CA/Oakland/Central-

East-Demographics.html (accessed Jul 25, 2020).

(119) Arraiz, I.; Drukker, D. M.; Kelejian, H. H.; Prucha, I. R. A Spatial Cliff-Ord-Type Model with Heteroskedastic Innovations: Small and Large Sample Results. *J. Reg. Sci.* **2010**, *50* (2), 592–614. https://doi.org/10.1111/j.1467-9787.2009.00618.x.

(120) Drukker, D. M.; Egger, P.; Prucha, I. R. On Two-Step Estimation of a Spatial Autoregressive Model with Autoregressive Disturbances and Endogenous Regressors. *Econom. Rev.* **2013**, *32* (5–6), 686–733. https://doi.org/10.1080/07474938.2013.741020.

(121) Kelejian, H. H.; Prucha, I. R. Specification and Estimation of Spatial Autoregressive Models with Autoregressive and Heteroskedastic Disturbances. *J. Econom.* **2010**, *157* (1), 53–67. https://doi.org/10.1016/j.jeconom.2009.10.025.

(122) Kelejian, H. H.; Prucha, I. R. A Generalized Moments Estimator for the Autoregressive Parameter in a Spatial Model. *Int. Econ. Rev. (Philadelphia).* **1999**, *40* (2), 509–533.

(123) Kelejian, H. H.; Prucha, I. R. A Generalized Spatial Two-Stage Least Squares Procedure for Estimating a Spatial Autoregressive Model with Autoregressive Disturbances. *J. Real Estate Financ. Econ.* **1998**, *17* (1), 99–121. https://doi.org/10.1023/A:1007707430416.

(124) R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria 2021.

(125) Piras, G. Sphet: Spatial Models with Heteroskedastic Innovations in R. *J. Stat. Software; Vol 1, Issue 1* **2010**.

(126) Bivand, R.; Piras, G. Comparing Implementations of Estimation Methods for Spatial

Econometrics. *J. Stat. Software; Vol 1, Issue 18* **2015**.

(127) EPA. NAAQS Table.