New Opportunities in Crowd-Sourced Monitoring and Non-government Data Mining for Developing Urban Air Quality Models in the US

Tianjun Lu

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
In
Planning, Governance, and Globalization

Steve Hankey, Chair
Wenwen Zhang
Linsey Marr
Peter Sforza

May 11, 2020

Blacksburg, Virginia

Keywords: Hazardous air pollutants; volunteer-based monitoring; local emissions; exposure assessment; crowdsourcing; low-cost monitoring; LUR validation; hybrid models; open data; urban morphology; enhanced models

New Opportunities in Crowd-Sourced Monitoring and Non-government Data Mining for Developing Urban Air Quality Models in the US

Tianjun Lu

ACADEMIC ABSTRACT

Ambient air pollution is among the top 10 health risk factors in the US. With increasing concerns about adverse health effects of ambient air pollution among stakeholders including environmental scientists, health professionals, urban planners and community residents, improving air quality is a crucial goal for developing healthy communities. The US Environmental Protection Agency (EPA) aims to reduce air pollution by regulating emissions and continuously monitoring air pollution levels. Local communities also benefit from crowd-sourced monitoring to measure air pollution, particularly with the help of rapidly developed low-cost sampling technologies. The shift from relying only on government-based regulatory monitoring to crowd-sourced effort has provided new opportunities for air quality data. In addition, the fast-growing data sciences (e.g., data mining) allow for leveraging open data from different sources to improve air pollution exposure assessment. My dissertation investigates how new data sources of air quality (e.g., community-based monitoring, low-cost sensor platform) and model predictor variables (e.g., non-government open data) based on emerging modeling approaches (e.g., machine learning [ML]) could be used to improve air quality models (i.e., land use regression [LUR]) at local, regional, and national levels for refined exposure assessment.

LUR models are commonly used for predicting air pollution concentrations at locations without monitoring data based on neighboring land use and geographic variables. I explore the use of

crowd-sourced low-cost monitoring data, new/open dataset from government and non-government sponsored platforms, and emerging modeling techniques to develop LUR models in the US. I focus on testing whether: (1) air quality data from community-based monitoring is feasible for developing LUR models, (2) air quality data from non-government crowd-sourced low-cost sensor platforms could supplement regulatory monitors for LUR development, and (3) new/open data extracted from non-government sponsored platforms could serve as alternative datasets to traditional predictor variable sources (e.g., land use and geographic features) in LUR models.

In Chapter 3, I developed LUR models using community-based sampling (n = 50) for 60 volatile organic compounds (VOC) in the city of Minneapolis, US. I assessed whether adding area source-related features improves LUR model performance and compared model performance using variables featuring area sources from government vs. non-government sponsored platforms. I developed three sets of models: (1) base-case models with land use and transportation variables, (2) base-case models adding area source variables from local business permit data (government sponsored platform), and (3) base-case models adding Google point of interest (POI) data for area sources. Models with Google POI data performed the best; for example, the total VOC (TVOC) model had better goodness-of-fit (adj-R^2: 0.56; Root Mean Square Error [RMSE]: 0.32 μg/m^3) as compared to the permit data model (0.42; 0.37) and the base-case model (0.26; 0.41). This work suggests that VOC LUR models can be developed using community-based samples and adding Google POI could improve model performance as compared to using local business permit data.

In Chapter 4, I evaluated a national LUR model using annual average PM$_{2.5}$ concentrations from low-cost sensors (i.e., PurpleAir platform) in 6 US urban areas (n = 149) and tested the feasibility of using low-cost sensor data for developing LUR models. I compared LUR models using only the PurpleAir sensors vs. hybrid LUR models (combining both the EPA regulatory monitors and the PurpleAir sensors). I found that the low-cost sensor network could serve as a promising alternative to fill the gaps of existing regulatory networks. For example, the national regulatory monitor-based LUR (i.e., CACES LUR developed as part of the Center for Air, Climate, and Energy Solutions) may fail to capture locations with high $PM_{2.5}$ concentrations and the within-city spatial variability. Developing LUR models using the PurpleAir sensors was reasonable (PurpleAir sensors only: 10-fold CV $R^2 = 0.66$, MAE = 2.01 $\mu g/m^3$; PurpleAir and regulatory monitors: $R^2 = 0.85$, MAE = 1.02 $\mu g/m^3$). I also observed that incorporating PurpleAir sensor data into LUR models could help capture within-city variability and merit further investigation on areas of disagreement with the regulatory monitors. This work suggests that the use of crowd-sourced low-cost sensor networks for LUR models could potentially help exposure assessment and inform environmental and health policies, particularly for places (e.g., developing countries) where regulatory monitoring network is limited.

In Chapter 5, I developed national LUR models to predict annual average concentrations of 6 criteria pollutants (NO_2, $PM_{2.5}$, O_3, CO, SO_2 and PM_{10}) in the US to compare models using new data (Google POI, Google Street View [GSV] and Local Climate Zone [LCZ]) vs. traditional geographic variables (e.g., road lengths, area of built land) based on different modeling approaches (partial least square [PLS], stepwise regression and machine learning [ML] with and without Kriging effect). Model performance was similar for both variable scenarios (e.g., random 10-fold CV R^2 of ML-kriging models for NO_2, new vs. traditional: 0.89 vs. 0.91); whereas adding the new variables to the traditional LUR models didn’t necessarily improve model performance. Models with kriging effect outperformed those without (e.g., CV R^2 for $PM_{2.5}$

using the new variables, ML-kriging vs. ML: 0.83 vs. 0.67). The importance of the new variables to LUR models highlights the potential of substituting traditional variables, thus enabling LUR models for areas with limited or no data (e.g., developing countries) and across cities.

The dissertation presents the integration of new/open data from non-government sponsored platform and crowd-sourced low-cost sensor networks in LUR models based on different modeling approaches for predicting ambient air pollution. The analyses provide evidence that using new data sources of both air quality and predictor variables could serve as promising strategies to improve LUR models for tracking exposures more accurately. The results could inform environment scientists, health policy makers, as well as urban planners interested in promoting healthy communities.

New Opportunities in Crowd-Sourced Monitoring and Non-government Data Mining for Developing Urban Air Quality Models in the US

Tianjun Lu

GENERAL AUDIENCE ABSTRACT

According to the US Centers for Disease Control and Prevention (CDC), a healthy community aims at preventing disease, reducing health gaps, and creating more accessible options for a wider population. Outdoor air pollution has been evidenced to cause a wide range of diseases (e.g., cardiovascular diseases, respiratory diseases, diabetes and adverse birth outcome), ranking as the top 10 health risks in the US. Thus, improving understanding of ambient air quality is one of the common goals among environmental scientists, urban planners, health professionals, and local residents to achieving healthy communities.

To understand air pollution exposures in different areas, US Environmental Protection Agency (EPA) has regulatory monitors for outdoor air pollution measurements across the country. For locations without these regulatory monitors, land use regression (LUR) models (one type of air quality models) are commonly employed to make a prediction. Usually, information including number of people, location of bus stops, and type of roads are shared online from government websites. These datasets are often used as significant predictor variables for developing LUR models. Questions remain on whether new air quality data and alternative land use data from non-government sources could improve air quality modeling. In recent years, local communities have been actively involving in air pollution monitoring using rapidly developed low-cost sensors and sampling campaigns with the help of local residents. In the meantime, advances in

data sciences make open data much easier to acquire and use, particularly from non-government sponsored platforms. My dissertation aims to explore the use of new data sources including community-based low-cost monitoring data and open dataset from non-government websites in LUR modes based on emerging modeling techniques (e.g. machine learning) to predict air pollution levels in the US.

I first built LUR models for volatile organic compounds (VOC: organic chemicals with a high vapor pressure at room temperature [e.g., Benzene]) based on community-based sampling data in the City of Minneapolis, US. I added information on number of neighboring gas stations, dry cleaners, paint booths, and auto shops from both the local government and Google website into the model and compared the model performance for both data sources (Chapter 3). Then, I used $PM_{2.5}$ data from a non-government website (PurpleAir low-cost sensors) for 6 US cities evaluating an existing air quality model that used air quality data from government websites. I further developed LUR models using the PurpleAir $PM_{2.5}$ data to see whether this non-government source of low-cost sensor data could be as reasonable as the government data for LUR model development. I finally extracted new/open data from non-government sponsored platforms (e.g., Google products and local climate zone [LCZ: a map that describes the development patterns of land, such as high-rise vs. low-rise or trees vs. sands]) in the US to investigate if these data sources can be used to alternate the land use and geographic data often used in national LUR model development.

I found that: (1) adding information (e.g., number of neighboring gas stations) from non-government sponsored sources (e.g., Google) could improve the air quality model performance for VOCs, (2) integrating non-government low-cost $PM_{2.5}$ sensor data into government regulatory monitoring data to develop LUR models could improve model performance and offer more insights on the air pollution exposure, (3) new/open data from non-government sponsored platforms could be used to replace the land use and geographic data previous obtained from government websites for air quality models. These findings mean that air quality data and street-level land use characteristics could serve as alternative data sources and are capable of developing better air quality models for promoting healthy communities.

ACKNOWLEDGEMENT

Thanks to all who have helped me over the years. First, I would like to express my sincere thanks to Dr. Steve Hankey (as my mentor and committee chair). It was Dr. Hankey who introduced me to the academic world in active transportation and air quality – two totally new fields for me when I first arrived in the US. Steve, you have encouraged and inspired me through all the difficulties I have during my graduate education. You have always been a good model to me for conducting research, teaching/mentoring, community involvement, and work-life balance. I do appreciate your supports and advices in both my academic research as well as my personal life. Your calm, organized manner and thoughtful character have guided me along the way. Having the opportunities to work with you is one of the most valuable experiences I have in my life. Things I've learned from you make me a better researcher, quick learner, and fun person. I look forward to continuing learning more from you.

I would like to sincerely thank my doctoral committee members, Dr. Linsey Marr. Dr. Wenwen Zhang, and Dr. Peter Sforza. Thanks for all your help and support to my dissertation work and research. I do appreciate Dr. Linsey Marr for your support on the air quality research. I still remember that your course of CEE 5154: Air Pollution Transport Chemistry helped me gain much knowledge of air pollution. I also learned a lot from our bi-weekly AIM seminars organized by you. Thanks to Dr. Wenwen Zhang for helping me with the data analytics. I've learned so much from you on machine learning and data mining techniques. I enjoyed the chat along our way to the coffee shops, which has inspired me a lot on research ideas and life experience. I'd like to thank Dr. Peter Sforza for your help on the remote sensing and geospatial techniques. I enjoyed the time brainstorming the fantastic remote sensing data sources with you.

I appreciate all of your insightful comments. Your instructions helped me think more thoroughly about my research questions and motivated me to do better work.

I also thank my collaborators at University of Washington (Dr. Julian Marshall, Dr. Matthew Bechle, and Dr. Sun-young Kim), Carnegie Mellon University (Dr. Albert Presto), Oregon State University (Dr. Perry Hystad, Andy Larkin), and the City of Minneapolis (Jennifer Lansing). I have been inspired by your impactful research and valuable comments. Thanks for sharing your data, methods, and experience with me. Your supports mean a lot.

I am lucky to work with all UAP faculty, staff and peers. It is pleasant to work with you in our department and I appreciate your helpful suggestions along the way. In particular, I'd like to thank Dr. Ralph Buehler, Dr. Yang Zhang, Dr. Theo Lim, Dr. Todd Schenk, Dr. Ralph Hall, Dr. Tom Sanchez, and Dr. Anne Khademian. I also thank Leigh, Krystal, Leslie, Han, Omchand, Seungbee, Hye-jeong, Kaidi, Raj, Qiong, Ruixiang, Meng, and other UAP members. You've always been good friends and co-workers of me.

I owe a big thank you to my girlfriend-Yisi, who is a PhD candidate in the School of Public Health at University of Washington. Thanks to the annual conference of International Society for Environmental Epidemiology (ISEE) for bringing us together in such a fantastic way. Yisi, thank you so much for giving me valuable comments on my research and dissertation. I am so lucky to have you during this difficult time of the coronavirus pandemic. Your love and help really matter to me. Mum and dad, thank you for your support to my study here and respect to every choice I've made. Thank you for your love, your sacrifice, and your encouragement in the past years. Lastly, I would like to acknowledge all my soccer teammates (Shaoyang, Rick, Yuntian, Laoyu, TR, and other family members) and Poker friends in Chinese Soccer Club. Playing soccer and poker with you brings a lot of fun to my life. I cherish the days we play, talk and eat together.

Table of Contents

List of Figures

List of Tables

Chapter 1: INTRODUCTION

1.1. Overview

Exposure to ambient air pollution is among the top 10 health risk factors in the US (WHO, 2009). In the past decades, US agencies at national and local levels have launched various programs and grant opportunities to promote healthy communities (CDC, 2012; EPA, 2019; HUD, 2014). Improving air quality in communities is one of the common goals for developing healthy cities shared by urban planners, environmental scientists, and local governments (ISGlobal, 2018). Strategies to gain better understanding of human exposure to air pollution to achieve human health objectives is an evolving process (Gulia et al., 2015). For example, the US Environmental Protection Agency (EPA) has engaged each state to establish the national air quality monitoring network as required by the Clean Air Act to evaluate ambient air quality for major pollutants (i.e., “criteria” pollutants: Ozone [O_3], Nitrogen Dioxide [NO_2], Particulate Matter [PM], Sulfur Dioxide [SO_2], and Carbon Monoxide [CO]; EPA, 1990). Local governments have also implemented measures that aim to reduce ambient air pollution levels through land uses (e.g., siting and planning through permit control; CARB, 2020), transportation modes (e.g., encouraging active transportation; APHA, 2010), and urban vegetation (e.g., planning trees and building parks to alter urban atmosphere; USDA, 2010). However, information on air pollution levels and its spatial variation is still limited due to insufficient air monitors from the national regulatory monitoring network (EPA, 2020). This limitation in spatial density and coverage of monitors further hinders modeling effort to efficiently assess human exposure. One area which warrants further research is how to leverage the new trend in community-based monitoring, crowd-sourced low-cost monitoring, open data, and modeling techniques to improve air pollution exposure assessment. This dissertation aims to enhance existing air quality models by incorporating new data of air pollution and predictor variables.

Existing costly regulatory air quality monitors are deployed to support standards compliance, at limited number of locations ranging from background (e.g., away from urban areas and emission sources) to population centers (e.g., near roads or hospitals; EPA, 2020b). One strategy to improve monitoring network is through community-based effort, which mainly refers to monitoring activities organized or partnered with local communities (NERL, 2015). Another ongoing effort to supplement existing regulatory monitoring network is to leverage the crowd-sourced low-cost sensor networks in areas of interest. The use of low-cost sensors allows for a relatively cheap and dense network of air quality monitors (e.g., neighborhoods; Kimbrough et al., 2019). While development in both community-based monitoring campaigns and the low-cost sensor network platforms is rapid, an understudied topic is how to integrate the new air quality monitoring network into modeling process for improving human exposure assessment.

Another major trend is the fast-growing data sciences that allow for mining open data from different platforms (e.g., Kitchin, 2014; Monino & Sedkaoui, 2016). These platforms provide a wide variety of data that could potentially serve as predictor variables to improve understanding of air quality through population dynamics, land use, transportation, and atmospheric environment (Bechle et al., 2017; Engel-Cox et al., 2004; Hankey et al., 2019). For decades, one of the most important open data platforms is the US Census, which stores reliable demographic and economic facts of communities across the country (US Census Bureau, 2020). The National Land Cover Database (NLCD) generated by a group of federal agencies (e.g., USGS [US Geological Survey]) allows for tracking national land use and land cover information consistently over years (MRLC, 2016). Other widely used open data platforms have provided publicly available atmospheric science datasets characterizing aerosols, clouds, and tropospheric composition based on remote sensing (e.g., satellite and aerial imagery; NASA, 2020; NOAA, 2020). However, these open data are mainly retrieved from government-sponsored initiatives. Emerging practices and data sciences (e.g., crowdsourcing and deep learning) has witnessed significant advances in non-government sponsored platforms. For example, Google Cloud provides easy access to Google Maps Platform, including rich map (e.g., Google street view [GSV]) and place (e.g., point of interest [POI]) details covering over 200 countries and territories (Google, 2011). The OpenStreetMap is a popular crowd-sourced map platform offering street network data across the world (OpenStreetMap, 2020). Similarly, Microsoft has released 125 million deep learning-derived building footprints across US (Microsoft, 2018). While non-government sponsored open datasets continue to serve the public, questions remain on whether some of these datasets with better consistency, coverage, and accessibility could be incorporated into air quality modeling to track air pollution exposures more accurately. This topic necessitates studies on exploring novel and open datasets for improving urban air quality models.

1.2. Summary of Dissertation Objectives

This dissertation aims to improve air quality models through community-based monitoring, crowd-sourced low-cost sensor network, new/open datasets, and emerging modeling techniques. The primary goal of my work includes assessing the use of these strategies in commonly used urban air quality models (i.e., land use regression [LUR]) at local, regional, and national scales. Specifically, this dissertation centers on three sets of analyses:

1. Comparing the impacts of different types of open datasets (government vs. non-government) on LUR models developed using a community-based sampling for 60 volatile organic compounds (VOC) in the city of Minneapolis, US (Chapter 3).

2. Evaluating the contribution of a crowd-sourced low-cost sensor network to improving LUR models using data from 6 urban areas in the US (Chapter 4).

3. Exploring the potential of innovative open datasets of predictor variables based on emerging modeling techniques (e.g., machine learning [ML]) for the development of national LUR models in the US (Chapter 5).

In the second chapter I summarize the motivation of assessing air pollution, the trend of community-based air quality monitoring and crowd-sourced low-cost sensing, and the development and opportunities of LUR models. I then present the three dissertation key analyses (Chapter 3 – Chapter 5). Finally, I conclude with major findings, limitations and implications of these LUR modeling effort.

Chapter 2: LITERATURE REVIEW

2.1. Assessing Human Exposure to Ambient Air Pollution

2.1.1. Health Effects of Air Pollution

An increasing number of studies have reported adverse health impacts of ambient air pollution (Jerrett et al., 2009; Laden et al., 2006; Pedersen et al., 2013), including cardiovascular diseases (Hoek et al., 2001; Hvidtfeldt et al., 2019; Laden et al., 2006b), respiratory diseases (X. Chen et al., 2017; Hvidtfeldt et al., 2019; Jerrett et al., 2009), birth outcomes (Ballester et al., 2010; Pedersen et al., 2013; Stieb et al., 2016), and cancer (Cakmak et al., 2018; Villeneuve et al., 2013). These impacts involve major air pollutants including NO_2 (Ballester et al., 2010; Cesaroni et al., 2012; Hoek et al., 2001), O_3 (Jerrett et al., 2009; Jung et al., 2013; Lin et al., 2008), $PM_{2.5}$ (particle diameter less than 2.5 micron; Cakmak et al., 2018; Laden et al., 2006b; Stieb et al., 2016; Thurston et al., 2016), PM_{10} (particle diameter less than 10 micron; Hvidtfeldt et al., 2019; Pedersen et al., 2013), SO_2 (X. Chen et al., 2017; Liang et al., 2019), and volatile organic compounds (VOCs; Villeneuve et al., 2013; Zhang et al., 2019). In general, different groups of population could be impacted by the exposure to air pollution including children (Ballester et al., 2010; S. Lin et al., 2008) and elders (Hvidtfeldt et al., 2019; H.-W. Zhang et al., 2019). Even for residents living in regions with relatively low air pollution levels, health risks of air pollution exposure still exist (Hoek et al., 2001; Stieb et al., 2016). Developing efficient tools and models for assessing air pollution exposure to cover a wide range of population and areas is necessary. Table 2.1 lists a sample of studies on long-term exposure to different air pollutants in multiple countries.

Table 2.1 Sample of long-term health effects studies of major air pollutants

NO$_2$ (per 10 g/m^3 increase)			
Study	**Study details**	**Health endpoint**	**Estimates and 95% CI**[a]
Cesaroni et al., 2012	The Rome Longitudinal Study (n = 45,006)	total mortality	HR = 1.04 (95% CI:1.03, 1.05)
Hoek et al., 2002	NLCS on diet and cancer cohort study (n = 4,973)	cardiopulmonary diseases mortality	HR = 1.27 (95% CI: 1.00, 1.78)
Ballester et al., 2010	INMA cohort in Valencia (n = 785), exposures in second trimester	small for gestational age	OR = 1.37 (95% CI: 1.01, 1.85)
O$_3$ (per 10 ppb increase)			
Jerrett et al., 2009	American Cancer Society Cancer Prevention Study II (CPS II) cohort (n = 448,850)	respiratory diseases mortality	HR = 1.04 (95% CI: 1.01, 1.07)
Lin et al., 2008	New York State birth cohort (n = 1,204,396)	asthma hospital admissions	OR = 11.6 (95% CI: 11.5, 11.7)
Jung et al., 2013	cohort study in Taiwan (n = 49,833)	autism spectrum disorder incidence	HR = 1.59 (95% CI: 1.42, 1.79)
PM$_{2.5}$ (per 10 g/m^3 increase)			
Laden et al., 2006	Harvard Six Cities Cohort (n = 8,096)	total mortality	RR = 1.16 (95% CI: 1.07, 1.26)
		cardiovascular diseases mortality	RR = 1.28 (95% CI: 1.13, 1.44)
Stieb et al., 2016	Singleton live births between 1999 and 2008 in Canada (n = 2,969,380)	small for gestational age	OR = 1.04 (95% CI: 1.01, 1.07)
		reduced term birth weight	OR = –20.5 g (95% CI: –24.7, –16.4)
Thurston et al., 2016	NIH-AARP Study never smoked (n = 19,785)	respiratory mortality	HR = 1.27 (95% CI: 1.03, 1.56)
Cakmak et al., 2017	CanCHEC (n = 3.6 million)	lung cancer mortality	HR = 1.54 (95% CI: 1.27, 1.87)
		ischemic heart disease mortality	HR = 1.13 (95% CI: 1.08, 1.19)
PM$_{10}$ (per 10 g/m^3 increase)			
Hvidtfeldt et al., 2018	The Danish Diet, Cancer and Health cohort of 50–64 yr adults (n = 49,564)	total mortality	HR = 1.12 (95% CI: 1.03, 1.22)
		cardiovascular diseases mortality	HR = 1.30 (95% CI: 1.11, 1.53)

		respiratory diseases mortality	HR = 1.04 (95% CI:0.87, 1.24)
Pedersen et al., 2013	Singleton delivery of ESCAPE (n = 74,178)	low birthweight at term	OR = 1.16 (95% CI 1.00, 1.35)
SO$_2$ (per 10 g/m^3 increase)			
Liang et al., 2019	retrospective birth cohort in seven Chinese cities in Pearl River Delta (n = 320,238)	preterm birth	HR = 1.48 (95% CI: 1.40, 1.57)
Chen et al., 2017	retrospective cohort in four cities in northern China (n = 39,054)	respiratory diseases mortality	HR = 1.11 (95% CI: 1.02,1.20)
		COPD mortality	HR = 1.15 (95% CI: 1.05,1.25)
VOC			
Villeneuve et al., 2013	Ontario Tax Cohort study (n = 58,750)	cancer mortality	benzene (per IQR = 0.13 μg/m^3) HR = 1.06 (95% CI: 1.02, 1.11)
		total mortality	n-Hexane (per IQR = 1.20 μg/m^3) HR = 1.02 (95% CI: 1.01, 1.05)
		cancer mortality	total hydrocarbons (per IQR = 9.02 μg/m^3) HR = 1.06 (95% CI: 1.02, 1.09)
Zhang et al., 2019	cohort study in Taiwan (n = 283,666), subject ≥ 40yr	Ischemic stroke mortality	total hydrocarbons (per 0.16 ppm) HR = 2.69 (95% CI: 2.64, 2.74)
			nonmethane hydrocarbons (per 0.11 ppm) HR = 1.62 (95% CI: 1.59, 1.66)

[a]HR (hazard risk), RR (relative risk) and OR (odds ratio) are different measures of risk. HR is estimated using survival curves and gives instantaneous estimates; RR is estimated by averaging events over a specific time period; OR is the ratio of the odds of an outcome in the presence of a risk and the odds of an outcome in the absence of a risk.

2.1.2. Regulatory Air Quality Monitoring Network in the US.

In the US, studies addressing long-term health impacts of many air pollutants (Jerrett et al., 2009; Laden et al., 2006b; Abbey et al., 1993; Brook et al., 2004; Daniels et al., 2000) have guided the EPA to regulate allowable ambient concentrations (i.e., National Ambient Air Quality Standards [NAAQS] for criteria pollutants) and facilitated the establishment of the air quality management framework (Demerjian, 2000). In particular, the EPA has initiated a national regulatory-grade monitoring network to track ambient air pollution concentrations over time. Table 2.2 shows the number of regulatory EPA monitoring sites of 6 criteria pollutants in 2000 and 2010 in the US. The primary purpose of the EPA network is to capture air quality trends for specific urban areas including background and population centers (EPA, 2020b). In this case, long-term health studies may be constrained by assigning residents exposure values from limited number of monitors for one city or even larger areas (Cohen et al., 2009; Laden et al., 2000; Pope et al., 2004). While this practice may be enough for between-city health impact assessment of long-term air pollution exposures, the within-city spatial variability of air pollution is often not fully captured particularly for cities without air quality monitors.

Table 2.2 Number of regulatory EPA monitoring sites of 6 criteria pollutants of 2000 and 2010 in the continental US

Pollutant	Year	Number of monitors	Annual mean concentrations	NAAQS[a]
NO_2 (ppb)	2000	345	15.6	53
	2010	327	9.6	
O_3 (ppb)	2000	768	49.4	80
	2010	850	45.8	
$PM_{2.5}$ (µg/m3)	2000	950	12.5	15
	2010	934	9.0	
CO (ppm)	2000	293	0.6	9
	2010	218	0.4	
PM_{10} (µg/m3)	2000	1021	23.8	50
	2010	829	18.6	
SO_2 (ppb)	2000	496	4.7	30
	2010	370	2.2	

[a]NAAQS: National Ambient Air Quality System. Averaging time: NO_2 (1 year); O_3 (8 hours); $PM_{2.5}$ (1 year); CO (8 hours); PM_{10} (1 year); and SO_2 (1 year).

2.1.3. Air Pollution Exposure Models

Different models are developed to predict air pollution concentrations at unmonitored locations. Dispersion models are used to predict pollution concentrations at downwind locations based on emission and meteorological data (Bellander et al., 2001; Nafstad et al., 2003). Spatial interpolation creates a surface based on monitoring data and assigns values to locations without measurements. For example, inverse-distance weighting (IDW) uses the inverse of the distance to locations with monitoring data (Hystad et al., 2012; Marshall et al., 2008) while kriging weighted the surrounding measurements to develop continuous surfaces (Jerrett et al., 2005; Mercer et al., 2011). Comparatively, LUR model has been developed for two decades to provide air pollution predictions at finer spatial resolutions (Hoek et al., 2008). Generally, LUR accounts for geospatial features (e.g., traffic, land use) surrounding the monitors and predicts air pollution concentrations at unmonitored locations typically using a regression approach. Up-to-date, LUR has been applied in countries across the globe, including North America (Hystad et al., 2011; Novotny et al., 2011), South America (Habermann & Gouveia, 2012), Europe (Eeftens et al., 2012a; Fernández-Somoano et al., 2011), Australia (Dirgawati et al., 2015; Knibbs et al., 2014), Africa (Dionisio et al., 2010), and Asia (Hassan Amini et al., 2014; H. Xu et al., 2019). These LUR models have been used to assess different air pollutants (e.g., criteria pollutants; VOC; Hoek et al., 2008; Ryan & Lemasters, 2007; Jerrett et al., 2005). More recently, researchers have developed LUR models using hybrid (Kim et al., 2020; Su et al., 2008) or machine learning approaches (Di et al., 2016; Meng et al., 2018). According to the commonly cited review of LUR models, measurements with at least 40-80 locations would enable a LUR model in urban area (Hoek et al., 2008); however, LUR models developed for large geographies may not fully capture within-city spatial variability of the pollution levels in the study area. As such, methods to increase the density and coverage of monitors for LUR development may be helpful for improving exposure assessment.

2.2. Community-based Monitoring and Crowd-Sourced Low-cost Sensing of Air Pollution

2.2.1. Benefits of Community-based Monitoring and Low-cost Sensors

Opportunities to fill the gaps of existing monitoring network may exist in ongoing community-based monitoring effort (involving local communities to monitoring) and crowd-sourced low-cost sensor networks. Emerging community-based and crowd-sourced monitoring is not tied to siting policies and could potentially track air pollution hot spots, which may supplement existing regulatory monitors towards reducing exposure (Kaufman et al., 2017; Muller et al., 2015; Rada et al., 2016). In addition, regulatory air quality monitoring is often costly, complex, and less portable while low-cost monitoring could be much cheaper, user-friendly, and more compact

(Commodore et al., 2017; Castell et al., 2017). For example, low-cost air quality sensors are often referred to devices costing less than $2,500 (EPA, 2014) for the entire monitoring system (e.g., measuring component, battery, and data storage; Borrego et al., 2016). More importantly, the use of low-cost sensors is capable of covering more locations of interest (Ahangar et al., 2019). Particularly, the monitoring network could target locations near traffic segments, residential areas, industrial facilities, and rural communities (Barzyk et al., 2016; EPA, 2020a; Hauser et al., 2015; Kinney et al., 2000; Williams et al., 2009). Another advantage of the low-cost monitoring is that it could track air pollutants that are not regularly monitored (e.g., volatile organic compounds [VOC; Williams et al., 2009], elemental carbon [EC; Kinney et al., 2000], black carbon [EPA, 2016], and ultrafine particles [Minkler et al., 2010; Truax et al., 2013]). Although performance of different low-cost sensors varied widely by pollutant (especially for gaseous pollutants), studies have shown successful practices in applying low-cost air quality sensors with careful sensor calibration and data quality assurance/quality control (QA/QC; Malings et al., et al., 2019; Y. Wang et al., 2015; Williams, 2019). The EPA has advanced the development of low-cost sensors by providing Air Sensor Toolbox with operation procedures, performance evaluation, data interpretation and data communication (EPA, 2014). With increased engagement among concerned residents, environmental and social groups, and local institution collaborators (Whitelaw et al., 2003), crowd-sourced low-cost monitoring can play significant roles in achieving clean and healthy communities.

2.2.2. Practices and Platforms of Crowd-Sourced Low-cost Monitoring

The rapid development in low-cost sensing has witnessed revolutionary advances in air quality monitoring from relying only on government-operated networks to including complementary crowd-sourced low-cost sensors (Morawska et al., 2018; Snyder et al., 2013). The most recent reviews have summarized existing development and applications of low-cost sensor networks, highlighting the potential of broader participation in air quality monitoring and network expansion (Morawska et al., 2018; Borghi et al., 2017; Clements et al., 2017; Jova et al., 2015; Rai et al., 2017; Thompson, 2016). The number of sampling locations of existing community-based monitoring could range from 4 to 400 depending on the study purpose and monitoring period (Barzyk et al., 2016; EPA, 2020a; Hauser et al., 2015; Kinney et al., 2000; Williams et al., 2009; English et al., 2017). In recent years, EPA has also provided grants to initiate multiple projects to explore the application of low-cost sensors in local communities (EPA, 2020a). These local practices have facilitated crowd-sourced low-cost monitoring to search effective strategies to monitor and analyze air pollution. The trend in communication technology (e.g., WIFI) and open data (e.g., application programming interface [API]) has also witnessed open data platforms of crowd-sourced low-cost sensor network in the US (Air Quality Egg, 2020; AirVisual, 2020; PurpleAir, 2020; AirCasting, 2020). The rise of such platforms may ease monitoring air quality within communities and sharing low-cost sensor data online, thus improving air pollution exposure assessment at the national or global level. Integrating crowd-sourced monitoring with low-cost sensors into existing regulatory monitoring network is promising. Table 2.3 lists sample studies of crowd-sourced air quality monitoring with low-cost sensors in the US.

Table 2.3 Sample studies of crowd-sourced air quality monitoring with low-cost sensors in the US

Study description	Study area	Year	Number of sites	Monitoring type	Approximate cost	Study purpose	Pollutant	Reference
Air monitoring on Harlem sidewalks	Harlem, NY	1996	4	Gravimetric monitoring (pumps and Teflon filters)	NA	Assessing street-level air pollution and relationship with diesel sources	$PM_{2.5}$ and elemental carbon (EC)	(Kinney et al., 2000)
Detroit Exposure and Aerosol Research Study	Wayne County, MI	2004-2007	140	Multiple monitoring types (e.g., gravimetric monitoring [PEM], Ogawa passive monitor) for personal, indoor, outdoor and community:	e.g., PEM: 2,000$; Ogawa: 200$	Investigating residential area near air pollution sources	VOC, $PM_{2.5}$, PM_{10}, EC and organic carbon (OC)	(Williams et al., 2009)
Community-based monitoring for education and communication	Multiple counties in NC	2011	7	Ogawa passive monitor (NOx, O_3)	200$/device	Monitoring counties without air monitors and comparing results to counties with regulatory monitors	NOx and O_3	(Hauser et al., 2015)
Air monitoring in the Ironbound community	Newark, NJ	2015	21	$PM_{2.5}$: Nephelometer; NO_2: electrochemical sensors (CairClip NO_2)	NA	Community-based monitoring example on borders with highways, waterways, railroads, and airport.	$PM_{2.5}$ and NO_2	(Barzyk et al., 2016)
The Hawai'i Island Volcanic Smog Sensor Network (HI-Vog)	Island of Hawai'i	2016-2021	40	$PM_{2.5}$: Optical Particle Counter; SO_2: Electrochemical sensors (e.g., Alphasense)	SO_2: 300$/device	Tracking volcanic smog with high spatial and temporal resolution	$PM_{2.5}$ and SO_2	(EPA, 2020)
Engage, Educate, and Empower California Communities on the Use and Applications of "Low-cost" Air Monitoring Sensors	Southern CA	2016-2021	400	Multiple monitoring types (e.g., Nephelometer, gravimetric)	NA	Informing selection, use, maintenance, and interpretation of low-cost sensors	$PM_{2.5}$ and PM_{10}	(EPA, 2020)

Imperial County Community Air Monitoring Network	Imperial County, CA	2017	40	Optical Particle Counter (e.g., Dylos DC1700)	200-300	Filling more detailed data on PM using low-cost sensors with collaboration from community, academic, nongovernmental, and governmental partners.	$PM_{2.5}$ and PM_{10}	(English et al., 2017)
Real-time Affordable Multi-Pollutant (RAMP)	Pittsburgh, PA	2016-2020	> 70	Electrochemical sensors (e.g., NO_2: Alphasense ID: NO2-B43F)	NA	Supplementing sparse regulatory-grade monitoring network through low-cost sensors to track exposures from restaurants, truck traffic, and environment justice communities	PM, CO, SO_2, NO_2, O_3, and VOC	(EPA, 2020)
Shared Air/Shared Action (SA2): Community Empowerment through Low-cost Air Pollution Monitoring	Chicago, IL	2016-2020	40	$PM_{2.5}$, PM_{10}: Optical Particle Counter (e.g., PurpleAir, AirBeam, and MET One); NO_2, O_3: Electrochemical sensors (e.g., Terrier, Aeroqual 500)	PM: 200-400; NO_2, O_3: 1,500$-2,000$	Developing effective strategies to monitor and analyze air pollution; monitoring in four communities using low-cost portable sensors	$PM_{2.5}$, PM_{10}, O_3, and $NO_{2,}$	(EPA, 2020)
Monitoring the Air in Our Community: Engaging Citizens in Research	Denver, CO	2016-2020	17	$PM_{2.5}$: gravimetric monitoring (PEM); NO_2: electrochemical sensors (CairClip NO_2)	PEM: 2,000$;	Identifying the needs and interpretation of community-based air pollution data among stakeholders; approaches to modify behaviors of reducing air pollution exposure	$PM_{2.5}$ and NO_2	(EPA, 2020)
Putting Next Generation Sensors and Scientists in Practice to Reduce Wood Smoke in a Highly Impacted, Multicultural Rural Setting (NextGenSS)	Yakama, WA	2016-2020	8	$PM_{2.5}$: Optical Particle Counter; Black carbon: Micro-aethalometer (Aethlabs MA200)	NA	Deploying next-generation low cost sensors on heavy wood impacts on rural community	PM and black carbon	(EPA, 2020)

2.2.3. Integrating Community-based and Crowd-Sourced Low-cost Monitoring into LUR Models

One research question is how low-cost sensors could be integrated to improve air quality modeling. A few recent studies have developed LUR models using monitoring data from low-cost sensors (Bi et al., 2020a; Carvlin et al., 2019; Huang et al., 2019; Lim et al., 2019; Masiol et al., 2018; 2019; Miskell et al., 2018; Weissert et al., 2018). In general, model performance in these studies varied by pollutant, study area, and modeling approach, with R^2/adj-R^2 ranging from 0.47 to 0.83. While most models were developed using only the low-cost sensors, few studies have used data from both low-cost sensors and regulatory monitors (Bi et al., 2020b; Huang et al., 2019). For example, Bi et al. (2020) found that the hybrid machine learning-based LUR model using both the regulatory and the low-cost sensors monitoring data outperformed that with only the regulatory monitors (CV R^2: 0.73 vs. 0.53). However, in another study using the same LUR approach, the model with hybrid datasets performed worse than the regulatory monitor-only models (CV R^2: 0.73 vs. 0.85; Huang et al., 2019). Further studies are warrented to explore the contribution of the low-cost sensors to LUR models. Existing low-cost sensor-based LUR models often focused on specific areas or single cities, use of low-cost sensor data for multi-city or national-level LUR modeling is questionable. Table 2.4 lists a sample of recent studies on LUR models using low-cost sensors.

Table 2.4 Recent sample studies of LUR models using low-cost sensors

Pollutant	**Study area**	**Number of sites**	**Model type**	**Model R^2**	**Major predictors**	**Reference**
NO_2	Auckland, New Zealand	40	LUR (stepwise regression)	Adjusted R^2 = 0.66	Distance to major road, number of bus stops, street width, and awnings	(Weissert et al., 2018)
NO_2	Vancouver, BC, Canada	mobile monitoring	LUR (stepwise regression)	Adjusted R^2 = 0.53	high-rise urban land use, sum of bus stops, speed limit	(Miskell et al., 2018)
$PM_{2.5}$	Imperial County, Southern CA, US	39	LUR (random forest)	CV R^2 =0.74 (low-cost only); 0.73 (hybrid model)	Land use, $PM_{2.5}$-ancillary variables, AOD, and meteorological inputs	(Bi et al., 2020)
$PM_{2.5}$	New York City, NY, US	63	LUR (random forest)	CV R^2 = 0.73 (hybrid model); 0.85 (regulatory model)	Land use, impervious surface, AOD, and meteorological inputs	(Huang et al., 2019)
$PM_{2.5}$	Seoul, South Korea	mobile monitoring	LUR (machine learning)	CV R^2 = 0.63-0.80	Land cover and land use	(Lim et al., 2019)
$PM_{2.5}$ and PM_{coarse}	Imperial County, Southern CA, US	35	LUR (Bayesian additive regression trees, lasso, partial least squares)	CV R^2 = 0.47-0.54 ($PM_{2.5}$); 0.55-0.65 (PM_{coarse})	Geolocation, proximity to the border, land cover, road, satellite measurement, and meteorological inputs	(Carvlin et al., 2019)
PM	Monroe County, NY, US	23	LUR (D/S/A algorithm)	Adjusted R^2 = 0.70	Land use, elevation, housing, population density, and roadways	(Masiol et al., 2018)
O_3	Monroe County, NY, US	10	LUR (D/S/A algorithm)	Adjusted R^2 = 0.83	Land use, elevation, housing, population density, and roadways	(Masiol et al., 2019)

2.3. LUR Model Improvement: Alternative/New Variables and Modeling Approaches

2.3.1. Traditional land use and geographic variables used in national LUR models

While increasing the monitoring density and coverage using community-based low-cost sensor networks may improve LUR models, other strategies may be used to seek for alternative/new variables, particularly for developing national models. Existing national-level LUR studies have used significant predictors including (1) traffic intensity/road features, (2) land use/land cover, (3) population dynamics, and (4) geographical characteristics (Beelen et al., 2013; de Hoogh et al., et al., 2016; Di et al., 2016; Hoek et al., 2015; Kerckhoffs et al., 2015; Kryza et al., 2011; Stedman et al., 1997). Some studies have found that meteorological, temporal information, and emission estimates were also significantly associated with air pollution concentrations (de Hoogh et al., 2016; Di et al., 2016; Kim et al., 2020; Knibbs et al., 2014; Z. Zhang et al., 2018). Assembling and calculating hundreds of variables may be time-consuming and computational-intensive for LUR model development. Another limitation is that the data used for national LUR models may be retrieved from various jurisdictions, making it difficult to generalize models across regions. Countries (e.g., developing countries) with limited and no data of these kinds would also suffer from lack of variables for developing air pollution LUR models. Further studies are needed to explore new/alternative variables that potentially improve LUR models in terms of data consistency, variable-reduction, and model generalizability. Table 2.5 shows existing major national LUR models.

Table 2.5 Existing major national LUR models

Pollutant	**Study country**	**Number of sites**	**Model type**	**Major predictors**	**Reference**
$PM_{2.5}$, PM_{10}, NO_2, CO, O_3, SO_2	US	~ 1,000	LUR (PLS-UK)	Traffic, population, land use/land cover, elevation, and satellite measurements	(Kim et al., 2020)
$PM_{2.5}$	US	1,928	LUR (convolutional neural network)	Satellite AOD, chemical transport model estimates, meteorological data, aerosol index data, and land use	(Di et al., 2016)
NO_2	US	423	LUR (stepwise regression)	Satellite measurements, impervious, tree canopy, roads, and elevation	(Novotny et al., 2011)
NO_2, $PM_{2.5}$, VOC	Canada	53-177	LUR (stepwise regression)	Satellite measurement, emissions, industrial area, and road length	(Hystad et al., 2011)
NO_2, NO_x, $PM_{2.5}$, PM_{10}, PMcoarse, etc.	More than 30 European countries	40-80 per country	LUR (stepwise regression)	Land use, road, population density, altitude, and local data	(Beelen et al., 2013)
NO_2, $PM_{2.5}$	Western Europe	1426 for NO_2; 436 for $PM_{2.5}$	LUR (stepwise regression)	Satellite measurements, chemical transport model estimates, roads, land cover, altitude, and north-south trend	(de Hoogh et al., 2016)
O_3	The Netherlands	90	LUR (stepwise regression)	Traffic intensity, low density residential land, road length, and urban green space	(Kerskhoffs et al., 2015)
NO_2	The Netherlands	144	LUR (stepwise regression)	Satellite measurements, industry, population, port, region indicators, traffic, and road	(Hoek et al., 2015)
NO_2, NO_x	UK	37	LUR (stepwise regression)	Land cover, NO_x emissions	(Stedman et al., 1997)
NO_x	Poland	58-104	LUR (stepwise regression)	Road length, traffic intensity, urban fabric, population density, and elevation	(Kryza et al., 2011)
NO_2	Australia	68	LUR (stepwise regression)	Satellite measurements, land use, road, population density, and elevation	(Knibbs et al., 2014)
$PM_{2.5}$, PM_{10}, NO_2	China	1,382	LUR (generalized additive mixed models)	Satellite AOD, meteorological data, and geographic data (e.g., roads, population density, proximity to emissions).	(Z.Zhang et al., 2018)
$PM_{2.5}$	China	1,452	LUR (stepwise regression, Bayesian maximum entropy)	Industrial area, road length, population density, and wind speed	(Chen et al., 2018)
$PM_{2.5}$, NO_2	China	~900	LUR (PLS-UK)	Road, land cover, POI, fire count data, elevation, meteorology, and satellite measurements	(H. Xu et al., 2019)

2.3.2. Satellite Data for LUR Models

Increasing studies have used satellite-derived measurements/estimates in LUR model development. The advanced image processing and sensor technologies of satellites have provided avenues to estimate ground-level concentrations and column abundance (i.e., the amount of trace gas in vertical atmosphere) for a wide range of species including aerosols, O_3, NO_2, CO, HCHO, and SO_2 (Martin, 2008a). The key advantage of satellite-derived air pollution estimates is the data consistency and wide spatial coverage across cities, regions, and countries beyond political boundary, thus being used to improve LUR models (Tulloch & Li, 2004; Lamsal et al., 2008). One of the commonly used satellite products is the aerosol optical depth (AOD), which estimates the amount of aerosol in the atmosphere by measuring the light extinction in the vertical direction (Kloog et al., 2011). Multiple studies have used various AOD products (e.g., moderate resolution imaging spectroradiometer [MODIS]) to predict ground-level concentrations of $PM_{2.5}$ based on their good agreement (Donkelaar et al., 2010; Donkelaar et al., 2006; C. Lin et al., 2015). Likewise, satellite-derived column abundance (e.g., ozone monitoring instrument [OMI]) of NO_2 is found to be significantly correlated with ground-level observations (Bechle et al., 2013; Lamsal et al., 2008) and has been successfully used in LUR models to predict NO_2 concentrations (Kim et al., 2020; Novotny et al., 2011). In summary, these satellite-based tropospheric column and ground-level measurements/estimates along with traditional land use and geographic variables could be used in LUR models (de Hoogh et al., 2016; Yang et al., 2017; Vienneau et al., 2013; H. Xu et al., 2019; Z. Zhang et al., 2018; Hoek et al., 2015). An understudied topic is how the satellite-derived estimates contribute to LUR models with different variable inputs and modeling approaches for multiple air pollutants.

2.3.3. Potential New Variables for LUR Models

2.3.3.1. Google Point of Interest (POI)

Emerging data sciences (e.g., data mining) allow for extracting information from new data sources sponsored by non-government bodies to develop LUR models. For example, as one of an important place-based products, Google POI represents a geolocation of a point with attributes (e.g., coordinates, ratings) that falls into different categories (e.g., restaurants, bus stations); these categories could characterize potential emission sources of air pollutants (e.g., land use, transportation). Traditionally, land use and traffic variables used in LUR models have several limitations. First, data inconsistency and incompleteness are often barriers in cross-region LUR studies. Another limitation is that existing national land use datasets (e.g., national land cover database [NLCD]) generally delineate area of interest rather than point of interest, which may fail to capture some local emission sources (e.g., restaurants, gas stations). Comparatively, Google POI data may serve as alternative or supplemental variables in LUR models. To date, only few studies have explored the contribution of categorized POI data to air quality modeling (Wu et al., 2017; H. Xu et al., 2019; Zheng et al., 2013). Google POI could provide more consistent and detailed point-based local indicators as compared to traditional dataset, allowing for developing LUR models to assess multi-city and intra-urban spatial variability of air pollution exposure (French et al., 2015; Madaio et al., 2016). Further studies are needed to explore whether the Google POI data can be used to replace or supplement traditional land use data sources in LUR models.

2.3.3.2. Google Street View (GSV)

As another popular non-government sponsored open dataset, GSV data is also a potential new variable to be applied for LUR models, which consists of street-level georeferenced panorama images regarding natural environment (e.g., tree, water) and built environment (e.g., sidewalk, building; Rzotkiewicz et al., 2018). GSV imagery has been used in identifying greenness (Li et al., 2016), assessing walkability (Yin set al., 2015), tracking traffic crashes (Hanson et al., 2013), evaluating aesthetics (Lafontaine et al., 2017), and informing other physical and mental health-related research (Rzotkiewicz et al., 2018). Other applications include land use mapping (Mitraka et al., 2015), housing price estimation (Law et al., 2018), and even political election prediction (Gebru et al., 2017). To my knowledge, for air quality application, no study has been found to explore how GSV imagery data could be used in air quality modeling. Improvement in image processing algorithms (e.g., machine learning, deep learning) has made it possible to extract rich contextual and microenvironment features from GSV imagery for characterizing air pollution exposure (e.g., vehicle and pedestrian traffic, street greenness; Li et al., 2017; Larkin & Hystad, 2017; 2019), questions remain on how these GSV-derived variables could be used in LUR models and whether this consistent and large dataset could alternate traditional land use and geographic variables.

2.3.3.3. Local Climate Zones (LCZ)

A limitation of existing LUR studies is that input variables are often collected from the US Census (e.g., population density, mixing of jobs) and land use surfaces (e.g., NLCD), but do not characterize urban morphology (e.g., building height, form). Lack of accounting for this information may underestimate its impact on air quality at street level (e.g., urban heat island, street canyon effect; McCarty & Kaza, 2015; Stone, 2005; Yuan et al., 2017). One recently developed urban form metric called LCZ classifies built and natural environment based on climate-relevant surface properties (Stewart & Oke, 2012a). Compared to traditional dataset, LCZ provides consistent method to measure urban form and functions including vertical characteristics of buildings (Bechtel et al., 2015; Mills et al., 2015). Due to its detailed and climate-based classification of urban areas, LCZ has gained popularity in temperature and climate studies (Alexander et al., 2015; Z. Lin et al., 2016; Middel et al., 2014; Petralli et al., 2014; Quan et al., 2017; C. Wang et al., 2018; Y. Xu et al., 2017). Only a few studies have explored LCZ in urban air quality models for specific study area (Steeneveld et al., 2016; Ching, 2013). To date, no study has tested the use of LCZ in empirical air quality models across regions. Questions remain on how LCZ data could serve as alternative and supplemental variables for LUR models at the national level.

2.3.3. Emerging Modeling Approaches for LUR Models

Aside from the increasing new data sources and types that could be integrated to improve LUR models, emerging modeling approaches may be applied to more efficiently handle these data. Traditional LUR models typically use the stepwise regression approach (Table 2.5; Beelen et al., 2013; de Hoogh et al., et al., 2016; Hoek et al., 2015; Kerckhoffs et al., 2015; Kryza et al., 2011; Stedman et al., 1997; Knibbs et al., 2014; Hystad et al., 2011). Other recent studies have developed hybrid models including combining universal kriging (UK), stepwise regression, and partial least squares (PLS; H. Xu et al., 2019; Kim et al., 2020), as well as chemical transport modeling (CTM) and stepwise regression (M. Wang et al., 2017). Particularly, some studies have successfully developed machine learning (ML)-based LUR models to predict air pollution concentrations with the goal of treating data-intensive process and non-traditional dataset, such as convolutional neural networks (Di et al., 2016), random forest (Bi, Stowell, et al., 2020b; Zhan et al., 2018), and gradient boosting (Reid et al., 2015). Best practices for predicting urban air pollution using emerging modeling approachs and new, large, and open datasets have yet to be developed.

2.4. Organization of Dissertation

In summary, for air quality data, existing LUR models may be constrained by the limited monitoring density and coverage of regulatory air pollution monitors. As a promising strategy to improve the monitoring network, community-based monitoring has a potential to track both ambient criteria pollutants and pollutants that are less frequently monitored (e.g., VOC). In addition, the rising non-government open platforms of low-cost sensor data (e.g., PurpleAir) could benefit local communities and supplement regulatory monitoring network to improve LUR models for air quality. This improvement could be represented by providing independent air quality data for model evaluation and serving as dependent data to increase number of monitors used in LUR models. For new predictor variables, traditional land use and geographic data sources may fail to capture the street-level features in a uniform way, which hinders the effort to generalize LUR models across cities. Seeking for new/alternative and non-government sponsored data (e.g., Google products) could serve as effective avenues to develop refined local and national LUR models. Emerging modeling approaches (e.g., ML) may also serve as efficient measures to deal with such types of big and open datasets for LUR models.

I explore these issues in Chapter 3 to Chapter 5:

Chapter 3 entitled "**Land Use Regression Models for 60 Volatile Organic Compounds: Comparing Google Point of Interest (POI) and City Permit Data**" presents the development of LUR models for VOC using different sets of predictor variables based on a community-based effort in a single city (i.e., City of Minneapolis) in the US.

Research questions:

1. Whether community-based air quality monitoring could be used to develop LUR models?

2. Should area sources be incorporated into models to characterize local emission sources for VOCs.

3. Whether non-government data sources could serve as feasible predictor variables for LUR models?

Chapter 4 entitled “**Using A Crowd-Sourced Low-cost Sensor Network in National Land Use Regression Models for $PM_{2.5}$**” explores the contribution of a crowd-sourced low-cost air quality sensor network to the evaluation and development of LUR models based on the data from 6 urban areas in the US.

Research questions:

1. Whether air quality data from non-government crowd-sourced platforms could provide a promising data source for LUR model development?

2. Whether adding the low-cost sensor data could improve existing LUR models that are developed based only on regulatory monitors?

3. Whether the open data of air quality could offer more insights to capture within-city spatial variability for exposure assessment?

Chapter 5 entitled “**Exploring New Predictor Data Sources to Develop National Land Use Regression Models for Criteria Air Pollutants**” examines the use of new predictor variables for improving national LUR models for criteria pollutants in the US.

Research questions:

1. Whether predictor data from non-government open platforms could be used to alternate traditional land use and geographic variables for LUR models?

2. Whether adding the new variables into existing traditional variables could improve national LUR model performance?

3. How LUR model performance varies by pollutant and modeling approach?

4. Are there some new insights that the new variables could offer for characterizing different air pollutants and for improving future LUR modeling?

Chapter 3. LAND USE REGRESSION MODELS FOR 60 VOLATILE ORGANIC COMPOUNDS: COMPARING GOOGLE POINT OF INTEREST (POI) AND CITY PERMIT DATA

ABSTRACT

Land Use Regression (LUR) models of Volatile Organic Compounds (VOC) normally focus on land use (e.g., industrial area) or transportation facilities (e.g., roadway); here, I incorporate area sources (e.g., gas stations) from city permitting data and Google Point of Interest (POI) data to compare model performance. I used measurements from 50 community-based sampling locations (2013-2015) in Minneapolis, MN, USA to develop LUR models for 60 VOCs. I used three sets of independent variables: (1) base-case models with land use and transportation variables, (2) models that add area source variables from local business permit data, and (3) models that use Google POI data for area sources. The models with Google POI data performed best; for example, the total VOC (TVOC) model has better goodness-of-fit (adj-R^2: 0.56; Root Mean Square Error [RMSE]: 0.32 μg/m^3) as compared to the permit data model (0.42; 0.37) and the base-case model (0.26; 0.41). Area source variables were selected in over two thirds of models among the 60 VOCs at small-scale buffer sizes (e.g., 25m-500m). My work suggests that VOC LUR models can be developed using community-based sampling and that models improve by including area sources as measured by business permit and Google POI data.

Keywords:

Hazardous air pollutants; volunteer-based monitoring; local emissions; exposure assessment

3.1. Introduction

Land use regression (LUR) is commonly used to model air pollutants using regulatory monitoring networks (e.g., NO_2, particulate matter) with the goal of estimating pollutant concentrations at unmonitored locations (Brauer et al., 2003; Jerrett et al., 2005; Marshall et al., 2008; Ross et al., 2007). Volatile organic compounds (VOCs) are precursors to ozone formation (WHO, 2000) and may pose long-term health risks (e.g., lung cancer, blood disorders) even at low concentrations (Glass et al., 2003; M. Lin et al., 2004; Villeneuve et al., 2014) suggesting a need to also properly characterize spatial patterns of VOCs using LUR models. However, ambient VOCs are less frequently monitored, thus reducing the possibility to model VOCs using LUR for exposure assessment studies (Guerreiro et al., 2014; Pankow et al., 2003). A variety of factors have limited the ability of previous studies to monitor and model VOCs including: (1) fewer than the commonly recommended 40-80 sampling locations for model development (Hoek et al., 2008), (2) complex emission sources for many VOC species (Brown et al., 2007; Kim et al., 2001; Piccot et al., 1992), and (3) limited and inconsistent data availability for small-scale local emission sources (e.g., area sources; Hochadel et al., 2006; Madsen et al., 2007).

Existing studies (see Table A3.1) have used LUR to model VOCs based on monitors at traffic segments (Carr et al., 2002), schools (Chang et al., 2006; Mukerjee et al., 2009; Smith et al., 2006), fire stations (Smith et al., 2011), airports (Gaeta et al., 2016) and across cities in North America (Atari & Luginaah, 2009; Johnson et al., 2010; Kheirbek et al., 2012; Oiamo et al., 2015; Poirier et al., 2015; Su et al., 2010; Wheeler et al., 2008), Europe (Aguilera et al., 2008; Carr et al., 2002; Fernández-Somoano et al., 2011; Gaeta et al., 2016) and Asia (Amini et al., 2017); only one existing LUR model (in Canada) has successfully assessed national VOC concentrations (Hystad et al., 2011). Previous VOC LUR models are often limited by the (1) number of monitors (n < 40; Atari & Luginaah, 2009; Mukerjee et al., 2009; Smith et al., 2006, 2011), (2) sampling period (a few weeks or seasons; Fernández-Somoano et al., 2011; Gaeta et al., 2016; Mukerjee et al., 2009; Oiamo et al., 2015; Su et al., 2010) and (3) number of VOC species monitored (n < 10; Aguilera et al., 2008; Amini, et al., 2017; Atari & Luginaah, 2009; Carr et al., 2002; Fernández-Somoano et al., 2011; Gaeta et al., 2016; Hystad et al., 2011; Johnson et al., 2010; Kheirbek et al., 2009, 2012; Oiamo et al., 2015; Poirier et al., 2015; Su et al., 2010; Wheeler et al., 2008). These limitations often hinder development of robust models to characterize a wide range of VOC species for long-term concentrations (e.g., annual averages). Community-based sampling offers an opportunity to gather sampling data for many VOC species and many time periods that would otherwise be difficult to collect (Conrad & Hilchey, 2011; Smith et al., 2007). Collaborative sampling efforts among local agencies and communities may facilitate more effective LUR modeling for pollutants that are otherwise less commonly monitored (e.g., VOCs) by approximating traditional fixed-site sampling for LUR.

To account for VOC emission sources, existing VOC LURs mainly include variables for transportation facilities (e.g., proximity to roads; Kheirbek et al., 2012) and land uses (e.g., industrial; Wheeler et al., 2008). Studies including source apportionment (Baldasano et al., 1998; Brown et al., 2007; Watson et al., 2001) and targeted measurements at specific locations (Kwon et al., 2006) found that ambient VOCs may be linked to area sources (e.g., dry cleaners, gas stations) that are often neglected due to their low (yet collectively high) individual emissions. A recent VOC review calls for critical evaluation of VOC-specific local characteristics and sources, which may be significant contributors to the spatial distribution of VOCs (Amini et al., 2017). Few VOC LUR studies attempt to account for the emissions from area sources (Amini et al., 2017; Hystad et al., 2011), partly due to a lack of such data that are often difficult to acquire

(Aguilera et al., 2008). Often, LUR models may be limited by inconsistent land use, emission source, or transportation data across political boundaries, making it difficult to generalize concentration estimates across jurisdictions (Amini et al., 2017). Google Point of Interest (POI) data offers rich information on land use patterns that may provide an alternative to traditional data sources (French et al., 2015; Madaio et al., 2016). A potential advantage of using POI data in LUR models is the ability to consistently assess the contribution of area sources to VOCs across different regions or countries which traditional city-level land use data cannot (details regarding the Google POI data is below).

In this paper, I developed LUR models for concentrations of 60 VOC species using community-based sampling data collected at 186 total locations (50 locations had sufficient data for model building) from November 2013 to August 2015 in Minneapolis, MN (Lansing et al., 2016). Specifically, my goals were to: (1) assess the feasibility of LUR modeling using data from community-based sampling, (2) explore whether information on area sources improves LUR models, and (3) investigate whether Google POI data could serve as an alternative data input in LUR modeling. I developed LUR models with and without information on area sources to compare how different data inputs could improve LUR model performance for a wide range of VOC species and explore their various spatial patterns.

3.2. Materials and Methods

3.2.1. Community-based Sampling Campaign for LUR Development

I developed LUR models using data collected as part of a community-based VOC sampling effort (Lansing et al., 2016). The sampling campaign was implemented by Minneapolis health department employees and volunteers (e.g., local residents) trained by local agencies. The City of Minneapolis was divided into 34 grid cells and sampling locations were selected so that at least two locations were in each grid cell. The campaign resulted in 186 sampling locations across Minneapolis including residential locations (56%), participating businesses that may emit VOCs (20%), Minneapolis Park and Recreation Board (MPRB) properties (17%), Minnesota Pollution Control Agency (MPCA) monitoring locations (2%), and others (e.g., formaldehyde collocated samples and residents who sponsored canisters: 5%; Lansing et al., 2016). The campaign measured 61 VOC species (e.g., benzene, toluene, and naphthalene) using a performance-based air sampling method (TO-15) and passivated stainless steel (Summa) passive sampling canisters. Specifically, TO-15 is a method developed by the US EPA for monitoring the 97 VOCs included in the 189 hazardous air pollutants (HAPs). The sampling campaign used 1-liter Summa canisters (a spherical container interiorly rendered inactive to most organic compounds) to collect air samples over a 72-hour period. All samples were sent to the lab (Pace Analytical Services) to analyze 60 VOCs (out of original 61; formaldehyde was not modeled due to the use of a different sampling strategy). The VOC measurements were collected across eight sampling events during November, February, May, and August of each year; the campaign started in November 2013 and ended in August 2015. Detailed information regarding how measurements were collected, analyzed, and processed as well as QA/QC methods can be found in the City of Minneapolis report (Lansing et al., 2016). I compared the community-based sampling measurements at the four MPCA monitoring stations (2%) to the MPCA data. Generally, there was a slight mismatch between the two different sampling campaigns; for example, the average normalized measurement gap was 23% for priority VOCs (BTEX: 21%, naphthalene: 26%; see Figure A3.1).

3.2.2. Dependent Variables for LUR

3.2.2.1. VOC species

Previous LUR studies model a limited number of VOC species (e.g., aromatic alkylbenzenes [mainly derivatives of benzene]) and fail to capture concentrations of other species (Amini et al., 2017). I developed LUR models of annual-average concentrations for the 60 VOC species sampled in the community-based sampling campaign. In the main text of this article I describe four VOC species that were of interest to the City of Minneapolis (hereafter referred to as priority VOCs), partly due to the fact that these VOC species exceeded the chronic health benchmarks in the initial study by the Minneapolis Health Department (MHD; Lansing et al., 2016); all LUR results for other VOC species are in the Appendix. The four priority VOCs include BTEX (benzene, toluene, ethylbenzene, m&p-xylene and o-xylene), naphthalene, tetrachloroethene, and TVOC (total VOCs-sum of all VOC species monitored as an overall measure of VOCs; Hodgson, 1995; Singh et al., 2016). I replaced all non-detects with half of the method detection limit of the sampling approach following U.S. EPA guidance (EPA, 1991).

3.2.2.2. Sampling periods for modeling. The community-based sampling campaign resulted in VOC measurements at 186 locations; however, only a small number of locations (n=24) were sampled during all eight events. Also, many locations did not have four consecutive sampling events to estimate annual averages for a single year. The average number of sampling events per site was 3.8. Using the second year of the sampling campaign (seasons 5-8 [November 2014 to August 2015] of the 8-season campaign) yielded the largest number of locations to model annual averages (n=50); thus, I report LUR model results of annual-average VOC concentrations for the second year as my core model scenario. I also developed a number of alternative modeling scenarios as described in the sensitivity analysis 2.5.1. Table 3.1 shows the summary statistics for my core modeling scenario for the four priority VOCs. A map of sampling locations and summary statistics for all 60 VOCs are shown in Figure A3.2 and Table A3.2.

Table 3.1 Summary statistics of the priority VOC measurements

VOC	Arithmetic Mean[a]	Geometric Mean[a]	Median[a]	Max[a]	Min[a]	IQR[b]
BTEX	5.02	4.12	3.64	27.61	1.97	2.88-5.01
Naphthalene	0.73	0.56	0.54	6.12	0.19	0.35-0.84
Tetrachloroethene	5.19	0.62	0.36	184.14	0.16	0.19-1.23
TVOC	61.76	53.81	45.91	228.82	30.68	36.37-69.29

[a] All units are in $\mu g/m^3$.
[b] Interquartile range; number of locations is 50.

3.2.2.3. Correlation matrix for all VOCs

I developed a correlation matrix using the measurements among VOC species. I identified comparatively high correlation among some of the VOC species (e.g., 1,3-Butadiene, 1,2-Dichlorobenzene); however, many of the 60 VOC species presented very low correlation. I decided to model the 60 VOCs separately in this paper; however, future work might assess when it is appropriate to model species together. I aggregated BTEX species for modeling due to the known high correlation (Pankow et al., 2003) and to compare to other LUR studies (Aguilera et al., 2008; Amini et al., 2017; Atari & Luginaah, 2009; Kheirbek et al., 2012; Mukerjee et al., 2012). The correlation among the priority VOCs was below 0.32. Figure A3.3 shows the correlation matrix for all VOCs.

3.2.3. Independent Variables for LUR

I assembled four subsets of candidate independent variables: (1) area sources as measured by city business permit data retrieved from the MHD, (2) area sources as measured by web-scraped Google POI data, (3) transportation variables, and (4) land use variables. Specifically, the city permit data includes four types of business licensing facilities (dry cleaners, paint booths, auto shops, and gas stations) that may emit VOCs and are of interest to the MHD; the data showed sources as of November 2013. To explore an alternative data source for area sources, I retrieved 90 categories of POI data from the Google Places Application Programming Interface (API) and identified four categories that most closely matched the city permit data (i.e., laundry, painter, car repair, gas station). The Google Places API is a service that returns information about POIs based on a search query (e.g., all restaurants within a specified distance from a central point). In this study, I used a Python script (shown in the Appendix) to automatically retrieve POI data in May 2018 to cover the study area. Google POI data is based, in part, on crowd-sourced information and may have errors. Importantly, Google POI data may be a promising set of variables to assess impacts of localized sources since the data can be tabulated across political boundaries, potentially allowing for modeling the relationship between VOCs and area sources across multiple jurisdictions. Variables were tabulated as point, proximity, or buffer variables as appropriate; I used 16 buffer lengths (25m, 50m, 75m, 100m, 150m, 200m, 250m, 300m, 400m, 500m, 750m, 1000m, 1500m, 2000m, 3000m, 5000m) following a previous LUR study in Minneapolis (Hankey & Marshall, 2015). This process resulted in a total of 228 (i.e., 16 × 14 buffer variables plus 4 point/proximity variables) candidate variables for selection (Table 3.2). Transportation and land use variables were offered for all models; area source variables offered varied depending on the model (see model building description below).

Table 3.2 Candidate independent variables in the LUR models

Category	Variable Name	Variable Type	Unit	Data Source
Area Sources: City Permit Data	Dry Cleaners	Count in buffer[a]	Count total	Minneapolis Health Department
	Gas Stations	Count in buffer	Count total	Minneapolis Health Department
	Paint Booths	Count in buffer	Count total	Minneapolis Health Department
	Auto Shops	Count in buffer	Count total	Minneapolis Health Department
Area Sources: Google POI	Laundry	Count in buffer	Count total	Google POI
	Gas Stations	Count in buffer	Count total	Google POI
	Painter	Count in buffer	Count total	Google POI
	Car Repair	Count in buffer	Count total	Google POI
Transportation	Principal Arterials	Length in buffer	Meters	N'compass
	Arterials	Length in buffer	Meters	N'compass
	Collectors	Length in buffer	Meters	N'compass
	Local Roads	Length in buffer	Meters	N'compass
	Dis. to Freeway	Length	Meters	Calculated
	Dis. to Major Road	Length	Meters	Calculated
	Traffic Intensity	Point	AADT m^{-2}	Minnesota Pollution Control Agency
	Transit Stops	Count in buffer	Count total	Minnesota Geospatial Commons
Land Use	Elevation	Elevation	Meters	Minnesota Geospatial Commons
	Industrial Area	Area in buffer	Square meters	Minnesota Geospatial Commons
	Open Space	Area in buffer	Square meters	Minnesota Geospatial Commons
	Retail Area	Area in buffer	Square meters	Minnesota Geospatial Commons
	Wtd. Household Income	Area-weighted average	Dollars	US Census Bureau
	Wtd. Housing Dens.	Area-weighted average	Unit km^{-2}	US Census Bureau

[a] Buffers in meters: 25, 50, 75, 100, 150, 200, 250, 300, 400, 500, 750, 1000, 1500, 2000, 3000, 5000

3.2.4. LUR Model Building

My LUR modeling approach was based on a commonly used forward stepwise regression technique (Su et al., 2009). The approach includes two steps: (1) add the independent variable most correlated with the VOC concentration (tested for normality and log-transformed for LUR modeling) and (2) sequentially add the independent variables most correlated with model residuals until the last variable is not significant ($p > 0.05$) or the Variance Inflation Factor (VIF; multicollinearity indicator) is larger than 5. I allowed for only one buffer size to be selected for each variable to further avoid collinearity (Wilton et al., 2010). I report model performance based on adjusted R^2, Root Mean Square Error (RMSE), and 10-fold cross validated (10-fold CV) R^2. I developed three types of LUR models with the four subsets of candidate independent variables using MATLAB R2014b to assess the impact of including area source information in LUR models of VOCs:

Base-case: No Area Sources. To replicate the majority of previous LUR models for VOCs (Aguilera et al., 2008; Carr et al., 2002; Kheirbek et al., 2012; Mukerjee et al., 2012; Smith et al., 2011), I developed LUR models with only transportation and land use variables as covariates.

Area Sources: City Business Permit Data. To explore whether area sources contribute to model performance, I added area sources from city business permit data in addition to the transportation and land use variables.

Area Sources: Google POI. To explore how an alternative data source for area sources – Google POI – impacts model performance, I replaced area source city permit data with Google POI data (while still including the transportation and land use variables).

I mapped model estimates of VOC concentrations (100m × 100m grid) for all three types of LUR models using ArcGIS 10.6 to assess spatial patterns among models. I tabulated the independent variables (variables that were significant in my LUR models) at the centroid of each grid, and used corresponding model results to estimate the VOC concentrations for all grid cells. To compare the impact of variables among VOC species and models, I fully normalized the model coefficients by multiplying each coefficient by the 95th-5th percentile difference of the independent variable divided by the 95th-5th percentile difference of the dependent variable.

3.2.5. Sensitivity Analysis

I performed sensitivity analyses to explore (1) LUR model results using different scenarios for aggregating to annual averages among sampling periods and (2) seasonal LUR models to assess whether seasonal trends exist among VOC species.

3.2.5.1. Scenarios to estimate annual-average concentrations among sampling periods

I aggregated VOC concentrations for LUR modeling based on multiple scenarios: (1) four consecutive sampling events (one event per season) during the first year (November 2013 to August 2014; n=40), second year (November 2014 to August 2015; n=50), or year-2014 calendar year (February 2014 to November 2014; n=45); (2) measurements during all 8 sampling events (n=24); and (3) non-consecutive coverage of 4 seasons among the 8 sampling events (n=79). Summary statistics for all 5 VOC annual-average scenarios of different sampling periods for the priority VOCs are shown in Table A3.3.

3.2.5.2. Seasonal LURs

In addition to the annual-average models, I also developed seasonal models with all available sampling data for each season (i.e., spring, summer, fall, and winter). I report model performance for each season as compared to the annual-average concentration models for the priority VOCs.

3.2.6. Model Validation

I examined Cook's distance to identify potential outliers that may influence my model results. I checked for spatial autocorrelation of model residuals using Moran's I, and further explored where spatial autocorrelation arose (if any) using LISA (Local Indicators of Spatial Association) for the priority VOCs (Anselin, 1995).

3.3. Results and Discussion

I developed three types of LUR models for 60 VOCs to explore the impact of different independent variables including different measures of area sources on model performance. I report detailed findings for the priority VOC species (BTEX, naphthalene, tetrachloroethene, and TVOC); detailed analyses for all 60 VOCs are in the Appendix.

3.3.1. LUR Model Results for Priority VOCs

I developed core LUR models using three sets of candidate independent variables: (1) transportation and land use variables (base-case models), (2) the base-case variables plus area sources measured from city permit data, and (3) the base-case variables plus area sources measured by Google POI data. I compare model results using performance indicators (e.g., adj-R^2; RMSE; 10-fold CV), variable selection (e.g., coefficient direction; buffer sizes) and by mapping concentration estimates for visual inspection. Table 3.3 shows model results for the priority VOCs. Table A3.4 shows model results for all 60 VOCs.

Table 3.3 LUR model coefficients for the priority VOCs

		Base-case: No Area Sources				Area Sources: City Permit Data				Area Sources: Google POI			
Category	Variable	BTEX	Naphthalene	Tetrachlo-roethene	TVOC	BTEX	Naphthalene	Tetrachlo-roethene	TVOC	BTEX	Naphthalene	Tetrachlo-roethene	TVOC
Area Sources: City Permit Data	Dry Cleaners	▨	▨	▨	▨			0.39 (25)	0.21 (25)	▨	▨	▨	▨
	Gas Stations	▨	▨	▨	▨		0.37 (200)			▨	▨	▨	▨
	Paint Booths	▨	▨	▨	▨		0.29 (500)			▨	▨	▨	▨
	Auto Shops	▨	▨	▨	▨	0.05 (50)			0.02 (75)	▨	▨	▨	▨
Area Sources: Google POI	Laundry	▨	▨	▨	▨	▨	▨	▨	▨			0.41 (25)	0.55 (1,000)
	Gas Stations	▨	▨	▨	▨	▨	▨	▨	▨			0.41 (200)	
	Painter	▨	▨	▨	▨	▨	▨	▨	▨	0.31 (400)	0.76 (400)		0.28 (400)
	Car Repair	▨	▨	▨	▨	▨	▨	▨	▨	0.16 (50)			0.12 (150)
Transportation	Principal Arterials		-0.45 (500)		0.47 (2,000)	0.28 (5,000)	-0.38 (750)		0.53 (2,000)	0.30 (5,000)	-0.73 (500)	0.34 (300)	0.63 (3,000)
	Arterials					0.33 (250)		0.30 (200)					
	Collectors		0.34 (1,500)		0.40 (5,000)	0.21 (75)				0.15 (100)	0.50 (1,000)		
	Transit Stops	0.55 (300)[a]			-0.29 (150)								
Land Use	Industrial Area								0.15 (200)				
	Open Space								0.35 (2,000)				0.29 (2,000)
	Retail Area			0.68 (150)									0.18 (25)
	Wtd. Housing Dens.							0.26 (25)				0.31 (25)	-0.41 (150)
Intercept		1.36	0.25	0.41	1.36	0.80	0.42	0.24	3.32	0.78	0.23	0.26	3.09
Adj-R^2		0.15	0.20	0.31	0.26	0.37	0.40	0.64	0.42	0.47	0.50	0.75	0.56
RMSE[b]		0.43	0.27	0.76	0.41	0.37	0.23	0.55	0.37	0.34	0.21	0.46	0.32
10-fold CV-R^2		0.14	0.17	0.26	0.21	0.32	0.32	0.40	0.36	0.41	0.47	0.56	0.48

[a] Model coefficients are normalized coefficients with buffers in parentheses. All variables are at $p < 0.05$. Number of locations used for modeling is 50.

[b] All units are in μg/m^3. Grey shading indicates variables that were not offered for the three LUR models during model building.

3.3.1.1. Goodness of fit

In general, adding city permit data to the LUR models improved model performance and outperformed the base-case models; this finding suggests that area sources are an important factor in explaining the variability of VOC concentrations. I also found that models with Google POI data outperformed models with city permit data for both the priority VOCs and among all 60 VOCs. For example, the BTEX models performed much better when including Google POI data (adj-R^2: 0.47; RMSE: 0.34 µg/m^3) as compared to city permit data (0.37; 0.37) and the base-case model (0.15; 0.43). These results are consistent with the reported R^2 of five previous LUR models for total BTEX ranging from 0.40 (moderate) in Detroit, USA to 0.81 (good) in Sarnia, Canada (Aguilera et al., 2008; Amini et al., 2017; Atari & Luginaah, 2009; Kheirbek et al., 2012; Mukerjee et al., 2012). For tetrachloroethene (commonly used at dry cleaners), the model performance improved from the base-case model (adj-R^2: 0.31; RMSE: 0.76 µg/m^3) with the addition of city permit data model (0.64; 0.55) and Google POI model (0.75; 0.46). I also aggregated all VOC species (TVOC) to compare to other measurement and modeling campaigns (Chen et al., 2016; Mečiarová et al. , 2017; Singh et al., 2016). Similar to the individual VOC species, the Google POI model (adj-R^2: 0.56; RMSE: 0.32 µg/m^3) outperformed the city permit data model (0.42; 0.37) and the base-case model (0.26; 0.41). These results indicate that area sources are important for explaining spatial patterns of VOC concentrations and that Google POI data may serve as a useful data source for LUR modeling. Figure 3.1 shows a summary of model performance among all 60 VOCs.

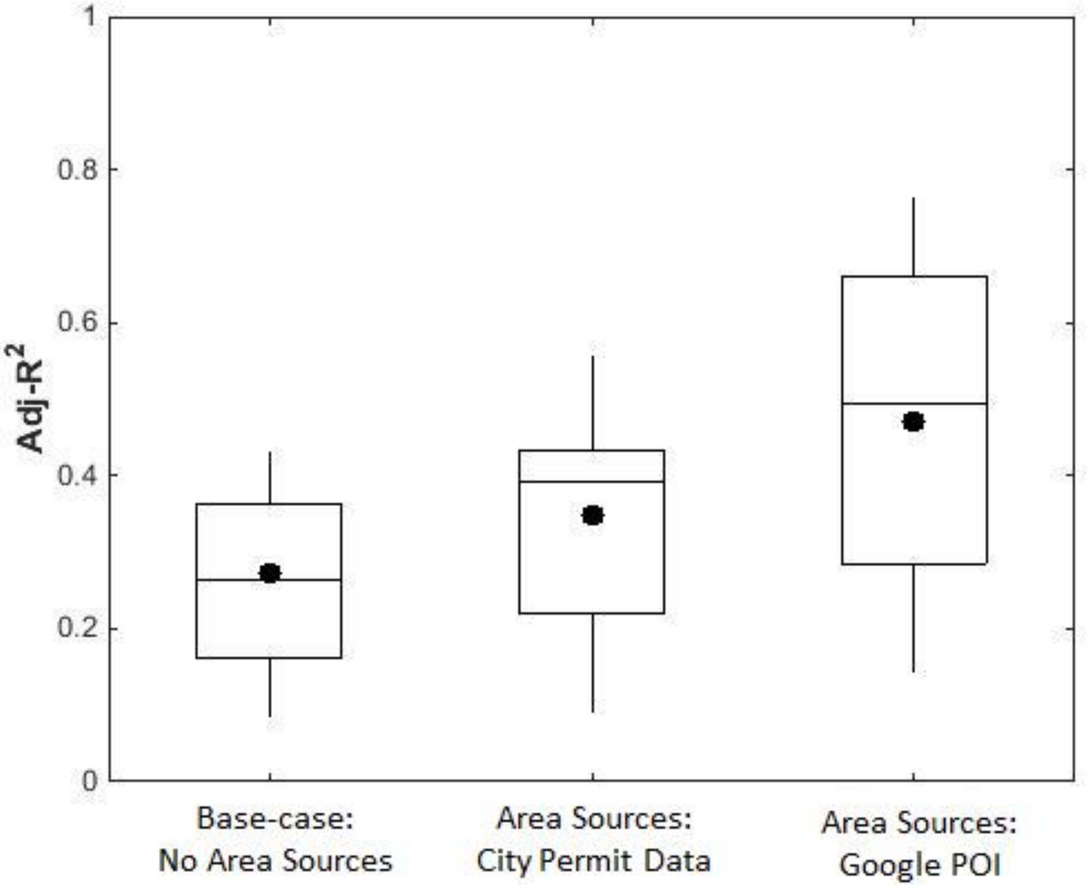


Figure 3.1. LUR model performance among 60 VOC species. The three input datasets represent the addition of area source information as candidate variables.

3.3.1.2. City permit data vs. Google POI

I developed LUR models by including four categories of area sources from two data sources (city business permit and Google POI). I noticed that LUR models using Google POI data outperformed those using the city permit data. There are several differences between the city permit and Google POI data that may explain this result. First, the Google POI area source categories were not perfectly matched with the city permit data and I had to choose the closest Google category; thus, the two data sources may capture slightly different sets of locations due to this choice. Second, there was a temporal mismatch among the Google POI data (year-2018), the city permit data (year-2013), and the VOC sampling data (2013-2015). Since business locations change over time, there are differences among locations in each category that could be due to this temporal mismatch. Third, the Google POI data is crowd-sourced (i.e., Google includes information available from businesses on the internet) which may lead to additional locations (e.g., capture of businesses without formal permitting) or missing locations (e.g., businesses that do not have an online presence) as compared to the city permit data. In general, the Google POI data captured more area source locations (e.g., average number of gas stations within 500m buffer: 0.73 Google POI vs. 0.68 city permit locations) and had a larger coefficient of variation as compared to that of the city permit data (e.g., gas stations: 1.40 vs. 1.29). Table A3.5 shows the average number of area source locations and coefficient of variation for the city permit data and the Google POI data.

3.3.1.3. Significant variable selection

The base-case models selected comparatively fewer variables ($n \leq 3$) for the priority VOCs – most of which were transportation variables (e.g., transit stops, principal arterials; Table 3.3). This choice of spatial predictors is similar to other studies that assess these VOCs (Amini et al., 2017; Atari & Luginaah, 2009; Kheirbek et al., 2012). When adding area sources as candidate variables, either from city permit or Google POI data, models consistently selected area sources (e.g., dry cleaners, gas stations) suggesting that traditional models (base-case model) may neglect the impact of important area sources. For example, BTEX, which is predominately from auto-related emissions, was associated with auto shops in the city permit data models; in the Google POI model, a similar area source (e.g., car repairs) was selected reinforcing the importance of including these data in the LUR models. One study in Tehran, Iran also found that the proximity to gas stations was associated with toluene and BTEX (Amini et al., 2017) while one national LUR study in Canada also incorporated this variable but failed to capture the variability of local-scale benzene concentrations (Hystad et al., 2011). A unique aspect of my models is the capability to compare the area sources that may be important to explain the variations in VOC concentrations. To support my findings, I also modeled all 60 VOCs (in addition to the priority VOCs) to explore how many VOC species may be linked to these small-scale sources. This exercise resulted in 45 out of 60 VOCs selecting area sources for the city permit data models and 52 out of 60 VOCs for the Google POI models, which further points to the importance of the area sources (see Figure A3.4).

3.3.1.4. Model coefficients

The normalized model coefficients allow for comparing variables across VOC species and indicate that area sources were as important predictors as commonly recognized transportation and land use variables. For the best performing Google POI model (tetrachloroethene), the area source coefficients had a slightly larger magnitude of association (0.41 for both laundry and gas stations) as compared to coefficients for transportation variables (0.34 for principal arterials) and land use variables (0.31 for housing density). These findings may help highlight the importance of specific area sources to inform policy choices (e.g., elimination of tetrachloroethene from all dry-cleaners in Minneapolis). Coefficients among models mostly followed a priori assumptions; however, results for certain variables and VOC species were counterintuitive, which was also the case in other LUR studies of VOCs (Fernández-Somoano et al., 2011; Mukerjee et al., 2012; Smith et al., 2011). For example, principal arterials had a negative association with naphthalene among all three models. To my knowledge, no study has explored naphthalene in LUR models; however, one review of emission sources of naphthalene pointed out that vehicle emissions were important sources (Jia & Batterman, 2010). This conflicting result indicates that area sources (e.g., paint booths, gas stations) may be correlated with other traditional predictor variables (e.g., road classification) and produce confounding results in some cases.

3.3.1.5. Buffer sizes of significant variables

Almost all area sources were selected at small buffer sizes (e.g., 25m-500m) suggesting that area sources are associated with VOC concentrations at small spatial resolutions reinforcing findings from previous studies (Baldasano et al., 1998; Brown et al., 2007; Watson et al., 2001; Mukund et al., 1996; Sun et al., 2016). For example, the Tehran LUR study found that being near a gas station was associated with higher VOC concentrations (Amini et al., 2017). Buffer sizes of transportation variables differed among VOC species; for example, traffic-related VOC species (e.g., BTEX) were associated with principal arterials at a buffer length of 5,000m while TVOC selected road classifications at a buffer length at 3,000m. The choice of these large buffer sizes is consistent with a suggestion to include traffic-related variables at buffers up to 5,000m from a recent VOC review (Amini et al., 2017). BTEX was also associated with lower road hierarchy (e.g., collectors) at smaller buffer lengths (e.g., 100m), which is similar to other LUR studies (Smith et al., 2006; Su et al., 2010). These findings imply that modeling individual or grouped VOC species may help identify specific variables of importance at different spatial resolutions.

3.3.1.6. Mapping concentration estimates

I mapped VOC concentrations for the entire city of Minneapolis on a 100 × 100 meter grid. For the purposes of mapping concentrations, locations with predicator data values that were outside of the variable space in the model building data were truncated to the highest (or lowest) value at my sampling sites as suggested by previous LUR studies (Beelen et al., 2010; 2009). In general, spatial patterns differed among VOC species and model types underscoring that (1) it is necessary to model VOC species separately to assess VOC-specific predictors (Amini et al., 2017), (2) different methods of obtaining area source data appear to provide different information, and (3) incorporating information on area sources from Google POI (or from local permitting data) offers an opportunity to improve model performance. For example, apart from the higher concentrations along transportation segments, the BTEX maps also showed vast hot spots partly due to area sources (e.g., car repair). Concentration hotspots in these maps visualize the spatial patterns of naphthalene and tetrachloroethene resulting from the significant association with area sources as compared to transportation variables. Potentially, these maps could be used for selecting additional sampling locations and to identify differences between the city permit and Google POI maps (to find potential emitters not captured with the city permit data, or possibly to identify errors or improve classification of the Google POI data). Figure 3.2 shows model estimates for the priority VOCs using all three types of models in Minneapolis, MN. Figure 3.3 shows scatterplots of the predicted vs. observed values for the priority VOCs.

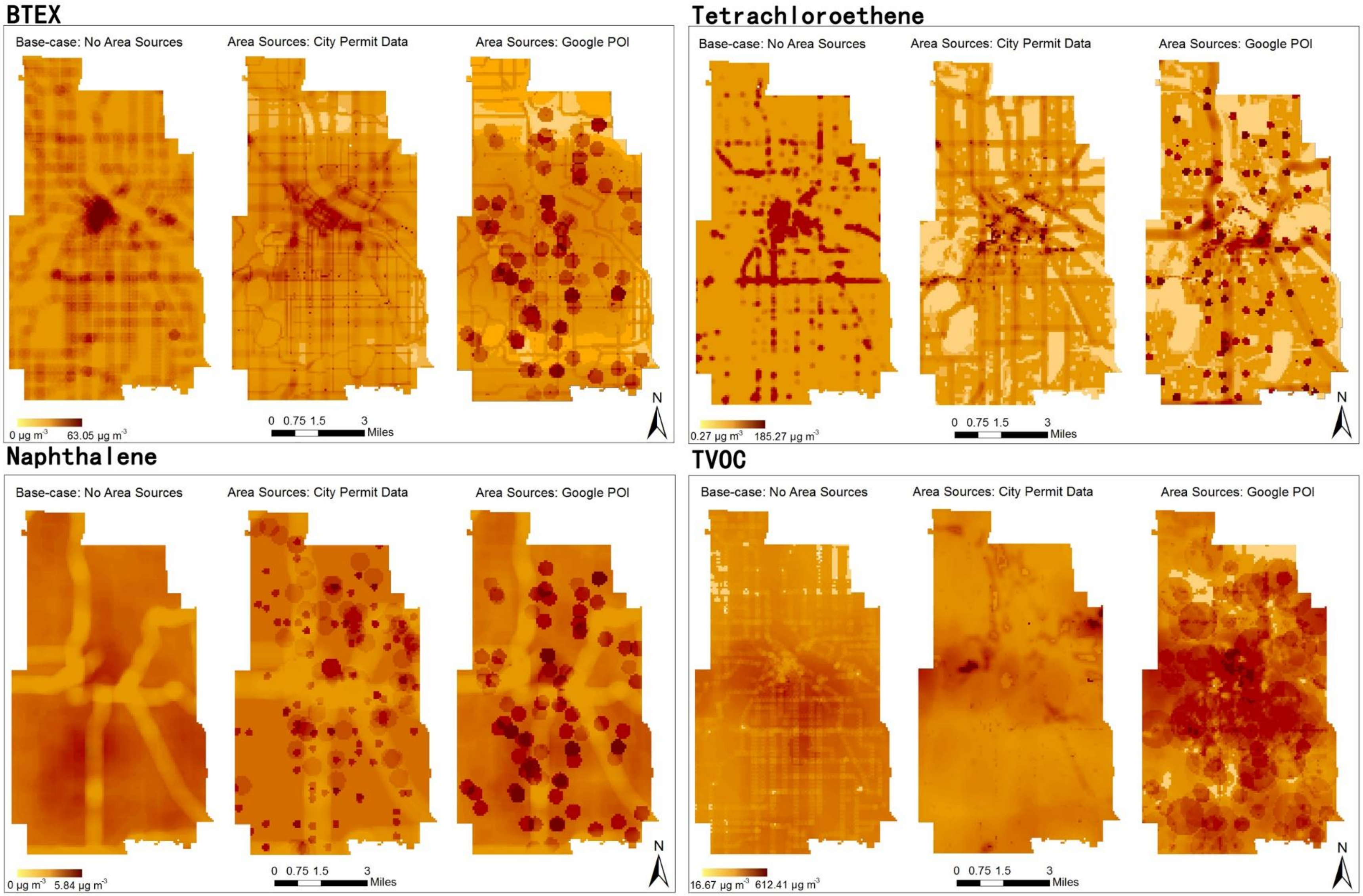


Figure 3.2. LUR model estimates for the priority VOCs among model types in Minneapolis, MN.

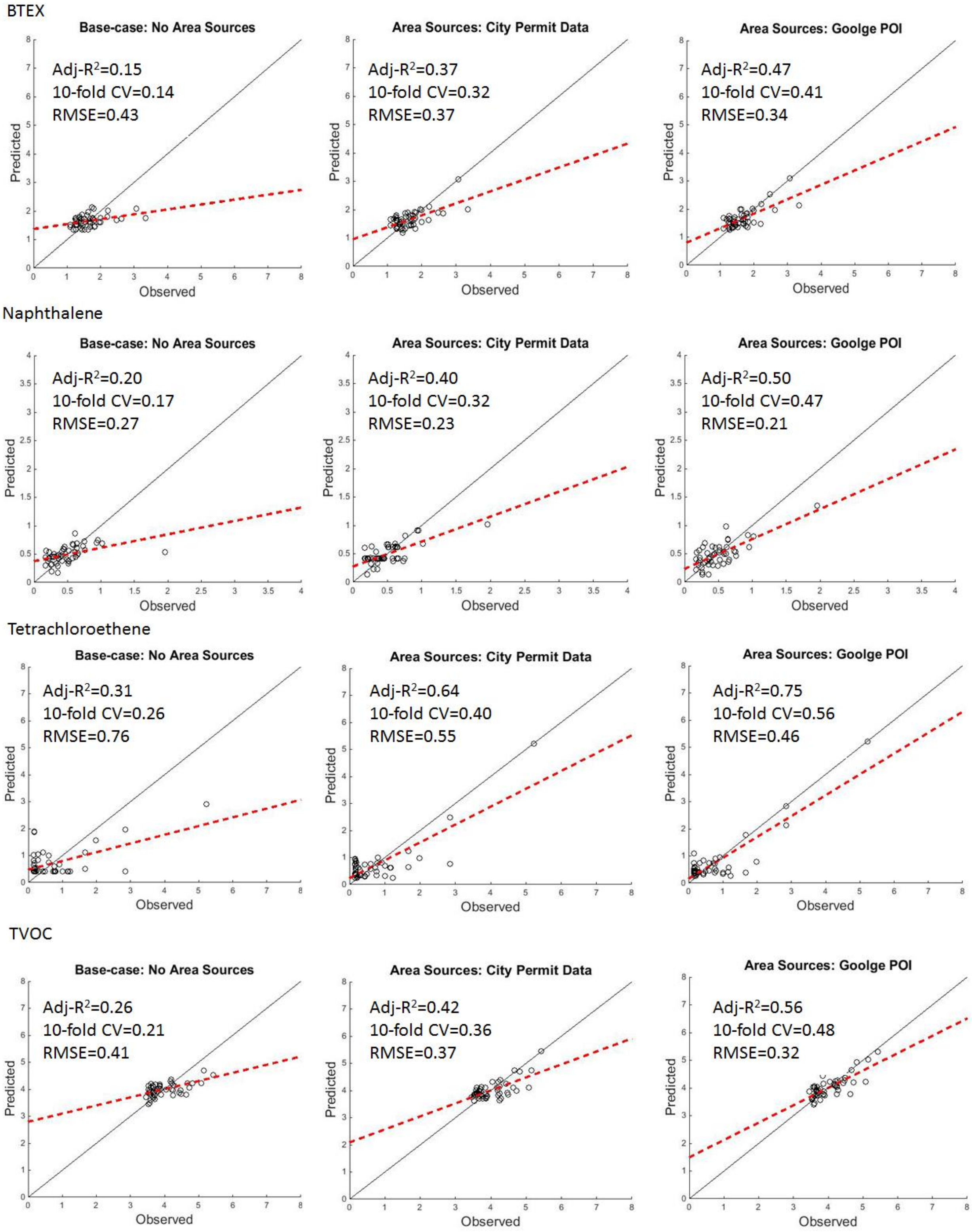


Figure 3.3. Scatterplots of predicted vs. observed values for the priority VOCs. Solid black lines represent the 1:1 line; dashed red-lines represent the best fit line.

3.3.2. Sensitivity Analysis

3.3.2.1. Scenarios to estimate annual-average concentrations among sampling periods

I developed LUR models using five scenarios for estimating annual-average VOC concentrations. Generally, the model scenario using only locations with all eight sampling events had the highest adj-R^2; however, this scenario included only 24 locations and demonstrated issues with overfitting (e.g., too many significant variables selected). All other scenarios had similar performance to my core model scenario (second year; n = 50 sampling locations; Figure A3.5 and Figure A3.6). My core scenario is consistent with the adequate number of sampling location (n = ~40-80) recommended for LUR modeling over small geographic areas (Hoek et al., 2008).

3.3.2.2. Seasonal LUR models

I developed seasonal LUR models and compared performance to the annual-average models for the priority VOCs (see Figure A3.7). In general, model performance varied by VOC and season with no obvious pattern of seasonal performance among the priority VOCs. On average, model fit was better for the annual-average models (mean adj-R^2 with Google POI: 0.57) as compared to the seasonal-average models (mean adj-R^2 with Google POI: 0.28). Seasonal fluctuations were not consistent among VOCs and annual-average concentrations are likely a stronger rationale for policy decisions which aim to reduce overall exposure.

3.3.3. Model Validation

3.3.3.1. 10-fold CV

In general, 10-fold CV R^2 values (Table 3.3) were slightly lower than the Adj-R^2 for the full models, with generally consistent patterns between Adj-R^2 and CV R^2. The drop in R^2 was largest for the Google POI models (e.g., tetrachloroethene: 0.75 to 0.56) suggesting that the Google POI models may be the most likely to encounter overfitting issues. However, my sample size was small (n=50) and this result should be tested in studies with more sampling data available for VOC-specific modeling.

3.3.3.2. Cook’s distance and spatial autocorrelation

Examination of Cook’s distance for my priority VOCs confirmed that no significant outliers influenced my model results (see Table A3.6). Based on the Moran’s I test, no significant spatial residual correlation was found in the LUR models for the priority VOCs except for some instances in the BTEX models (Table A3.7). For example, BTEX showed significant autocorrelation in two models (Moran’s I index with $p < 0.05$ for the base-case models [models with Google POI data]: 0.26 [-0.19]) but not the third model (city permit data). I further explored this issue using the LISA procedure and found that among the priority VOCs, only BTEX was flagged at a cluster of locations near a principal arterial indicating that spatial autocorrelation existed at this heavy traffic corridor for BTEX (see Figure A3.7). Future research should explore how sampling locations can be specifically designed for the purpose of spatial modeling to reduce spatial autocorrelation, or how spatial autocorrelation can be included within the modeling framework.

3.3.4. Implications for Developing VOC LUR Models

3.3.4.1. Implications for modeling 60 VOCs

To my knowledge, this is the first LUR study that measures and models 60 VOC species; existing LUR studies explored a limited number of species (n<10; see Table A3.1). I was able to model 60 VOCs to explore how VOCs show various spatial variability. My correlation matrix indicates that certain VOC species may be highly correlated and share similar spatial characteristics. For example, BTEX and some aromatic compounds (e.g., 1,3-Dichlorobenzene, 1,4-Dichlorobenzene) were generally correlated with each other (Pankow et al., 2003). This finding suggests that future LUR models could be refined by grouping certain VOCs (e.g., factor analysis, principal component analysis). However, many of the VOC species showed little correlation and warranted individual LUR models for those VOC species. The different toxicity and complex health risks of each VOC also necessitates my targeted modeling strategy for individual VOCs in certain cases (Lansing et al., 2016).

3.3.4.2. Implications for community-based sampling campaigns

Few studies have developed LUR models for air pollutants using community-based sampling data. Community- and volunteer-based sampling campaigns offer the potential to monitor at spatial and temporal scales that would otherwise be difficult for some pollutants. To ensure measurement quality, this approach requires training sessions for volunteers and adequate collection devices for rotation. My work shows that community-based efforts can provide useful data for modeling and estimating VOC concentrations. Learning from previously established best practices for LUR models (Larson et al., 2007; Su et al., 2013), future community-based campaigns could be designed to ensure that annual-average concentrations are available at a large number of locations for modeling.

A limitation of my study is that out of 186 total locations, I was only able to use 50 locations for modeling due to a lack of sampling data during specific sampling events among locations; my community-based sampling may also be limited by the fact that its original purpose was not for LUR. My LUR models were based on locations that had four consecutive events of data available to capture annual-average concentrations of VOCs. This issue may be important for future sampling campaigns to capture the spatiotemporal nature of VOC emissions. The seasonal models didn't demonstrate obvious patterns across seasons among the priority VOCs; however, the annual model fit outperformed the seasonal models indicating that it is necessary to capture concentration patterns during all four seasonal events to evaluate the annual-average concentrations. Previous studies typically use a limited number of sampling events (e.g., 1-2 weeks) to build LUR models, which may misrepresent spatial patterns or annual-average values of VOC concentrations (Amini et al., 2017; Atari & Luginaah, 2009; Kheirbek et al., 2012; Su et al., 2010; Wheeler et al., 2008).

3.3.4.3. Implications for comparing area sources in VOC LUR

Most LUR studies are developed for criteria pollutants and rely on traditional transportation and land use variables and do not include information on small-scale air pollution sources (i.e., area sources; Kheirbek et al., 2012; Wheeler et al., 2008). However, VOCs embody comparatively different characteristics and emission sources as compared to criteria pollutants (e.g., NO_2). Existing LUR VOC studies have only analyzed a few VOC species (e.g., BTEX) that are mainly from traffic and industrial emissions (Amini et al., 2017). Comparatively, my study analyzed 60 VOCs and presented a more comprehensive assessment of different VOC species in my LUR models.

A contribution of my models is that adding area sources helps to assess whether these small-scale sources are correlated with VOC concentrations. I was able to explore this relationship by comparing our base-case models that exclude area sources to models with information on area sources. By normalizing the model coefficients, I found that area sources may be as important (or even more important for some VOC species) as traditional transportation and land use variables. More work is needed to add area sources into LUR modeling for other jurisdictions and pollutants to further assess utility of these data sources. For example, in one LUR study in Iran, proximity to gas stations were flagged as significant variables for toluene and BTEX (Amini et al., 2017); the inclusion of this variable suggests that it is necessary to consider area sources when modeling in both developed countries and developing countries (often with higher air pollution levels; Amini et al., 2017).

3.3.4.4. Implications for using non-traditional data such as Google POI

An open question is how best to measure small-scale emission sources for air quality modeling. I used two measures of area source data (Google POI and city permit data) to explore the potential benefits of each dataset. I found that inclusion of both datasets improved model performance but that the Google POI models demonstrated the best model performance among all models and the 60 VOC species. My LUR models show that online mapping data could provide a useful input for LUR modeling. I only included variables from the Google POI database that were closely matched to the categories of area sources I had from the city business permit database; future work could replicate and expand my approach by including all data available in the Google POI database.

My modeling approach of combining an open dataset (Google POI) on area source emissions with a community-based sampling campaign offers promising potential for creating community driven modeling efforts to better characterize the spatial patterns of VOCs. To date, only one national LUR model for limited VOC species is available (Hystad et al., 2011). My work suggests that it may be possible to develop generalizable LUR models for VOCs across different regions or countries when using open access variables to pool datasets among study locations. However, such data sources may introduce biases, particularly from user generated and user verified content (Crutcher & Zook, 2009; Stephens, 2013). For example, businesses without an online presence, which are more likely in low socioeconomic regions, are less likely to add themselves to the dataset (e.g., Google Maps), and lower internet/smartphone usage in these regions may exacerbate that divide. Future work could refine this modeling approach and allow for expanding the geographic scope of these models towards developing models capable of providing generalizable information for siting and planning efforts.

3.4. Conclusion

I developed LUR models for 60 ambient VOC species using measurements at 50 sampling locations (out of 186 locations) from a community-based sampling campaign during November 2013 to August 2015 in Minneapolis, MN. I was able to assemble three sets of independent variables to develop my core LUR models: (1) land use and transportation variables, (2) area source variables from local business permit data, and (3) Google POI data for area sources. I found that models with the Google POI area source data performed better as compared to the base-case model and the permit data model. I found that area sources had a similar or bigger magnitude of correlation with VOCs than traditional land use and transportation variables. Among the 60 VOCs, over two-thirds of the LUR models indicated that area sources were significantly correlated with the VOC concentrations at small spatial scales. My work suggests that community-based sampling could be used as a valuable input for LUR models to estimate VOC concentrations. My study explores the spatial patterns of a wide breath of VOCs (a novel aspect is the number of VOCs studied) and identifies differences among data inputs for important area sources. The use of Google POI data also offers a more generalizable data source for national VOC LUR models in the future. My work could be used to inform planning policies to reduce emissions from area sources.

Chapter 4. USING A CROWD-SOURCED LOW-COST SENSOR NETWORK IN NATIONAL LAND USE REGRESSION MODELS FOR $PM_{2.5}$

ABSTRACT

Exposure assessment from existing national scale land use regression (LUR) models are typically based on sparse regulatory monitoring networks. Emerging low-cost sensor networks that are located by a variety of users (i.e., crowd sourcing of locations) offer the opportunity to supplement LUR models with improved measurement density and coverage (e.g., where the regulatory monitors are unavailable). I evaluated an existing national LUR model using annual average $PM_{2.5}$ concentrations from PurpleAir sensors in 6 US urban areas (n = 149). I developed LUR models (using only the PurpleAir sensors) and hybrid LUR models (using both the regulatory and low-cost monitors). I found that the low-cost sensor network may offer a promising alternative to regulatory networks where there are gaps in regulatory network. For example, developing LUR models using the PurpleAir sensors yielded promising results (6-city PurpleAir sensors: 10-fold CV $R^2 = 0.66$, MAE = 2.01 µg/m^3; PurpleAir and national regulatory monitors: $R^2 = 0.85$, MAE = 1.02 µg/m^3). I observed that integrating PurpleAir into LUR may be helpful to capture within-city variability and identify areas of disagreement (e.g., near-industrial neighborhoods). Integrating crowd-sourced low-cost sensor network in LUR models could help track exposures more accurately and inform environmental and health policies.

Keywords:
Crowdsourcing; low-cost monitoring; LUR validation; hybrid models

4.1. Introduction

Exposure to ambient air pollution (e.g., $PM_{2.5}$: particulate matter less than 2.5 micrometers in diameter) is a significant global burden of disease resulting in various health effects (e.g., cardiovascular, metabolic, and respiratory diseases; Dabass et al., 2016; F. Liu et al., 2019; Pun et al., 2017). Land Use Regression (LUR) is frequently used to predict and assess ambient exposure to $PM_{2.5}$ at unmonitored locations (Hankey & Marshall, 2015; Hoek et al., 2008). LUR models are often trained with limited and expensive ground-level regulatory monitors (e.g., Environmental Protection Agency Air Quality System [EPA AQS]; Clark et al., 2011; Di et al., 2016; Ross et al., 2007). The relatively sparse regulatory monitoring network is designed for compliance with the national ambient air quality standards and may fail to reliably capture within-city spatial variability (Hall et al., 2014). Dense air quality monitoring networks are needed to improve exposure assessment (Vlaanderen et al., 2019).

Growing interest in public environmental data collection and the innovation of inexpensive instruments have shifted the paradigm of air quality monitoring from government agencies toward crowd-sourced efforts (e.g., non-governmental organizations, citizen scientists; Jiao et al., 2016; Snyder et al., 2013b; Thompson, 2016b). For example, Google Project Air View initiative has been expanded across the world to measure street-level air quality data through sensors equipped on the Google Street View vehicles (Google Earth Outreach, 2020). Others focus on involving citizens to monitor and share monitoring data through internet-enabled platforms (Kumar et al., 2015). Emerging crowd-sourced monitoring is not tied to siting policies that ignores hot spots in favor of capturing ambient concentrations and provides increased spatial resolutions (e.g., neighborhood) with potential large inventories of data as compared to regulatory fixed monitors (Muller et al., 2015; Rada et al., 2016). This crowd-sourced trend allows for public participation in air quality data collection and holds promise for supplementing the regulatory data towards reducing exposure and informing health policies (English et al., 2017; Kaufman et al., 2017).

In addition, the inexpensive, portable, and operation-friendly low-cost sensors have revolutionized the crowd-sourced effort (Morawska et al., 2018). The non-regulatory grade low-cost sensors do not hinder their popularity and ubiquitous monitoring networks (Morawska et al., 2018). For example, low-cost sensors have known performance issues related to inter-device consistency and sensitivity to relative humidity, temperature, and particle coincidence (e.g., two particles get counted at once; Y. Wang et al., 2015; R. Xu, 2015). However, data from laboratory and ambient calibrations are becoming increasingly available to correct for these sampling artifacts; these corrections combined with improving performance from low-cost sensing provide opportunities to refine air quality monitoring (Broday et al., 2017; Holstius et al., 2014; Kelly et al., 2017; Malings et al., 2019). While such low-cost sensor networks have been deployed and readily accessible, the retrieving data haven't been used in concert with the regulatory monitors for exposure estimation.

A few studies have integrated low-cost sensing into LUR modeling for a single city (Dijkema et al., 2011; Eilenberg et al., 2020; Jain et al., 2020; Masiol et al., 2018; Weissert et al., 2019); yet limited work examines their contribution to LUR models (e.g., $PM_{2.5}$) at multi-city and national scale (Bi et al., 2020). Emerging low-cost sensor networks may be useful for improving LURs due to the increased network density within many urban areas and coverage in places not covered by EPA. For example, PurpleAir (hereafter PPA) is a crowd-sourced low-cost sensor network that offers publicly available real-time and historic data (e.g., $PM_{2.5}$) worldwide (PurpleAir, 2020). Prior work has shown that the PPA data is well correlated with reference measurements in US and Canada (R^2 = ~ 0.90; Brauer & Lee, 2018; South Coast Air Quality Management District, 2018). An understudied topic is whether the crowd-sourced low-cost sensor data from multiple cities is of sufficient quality to improve LUR models and to capture spatial variability that is not well characterized by regulatory monitoring alone. Properly characterizing the spatial patterns of air quality is important for assigning exposures in epidemiological studies and estimating the resulting health impacts.

In this study, I use a crowd-sourced low-cost sensor network (i.e., PPA) in 6 US urban areas to explore the contribution of multi-city low-cost sensor data to regulatory monitor-based LUR models. Specifically, I firstly use calibrated concentration estimates from the 6-city PPA sensors (n = 149) to evaluate an existing $PM_{2.5}$ LUR model and identify potential opportunities for the PPA data. I then develop LUR models using the PPA data (along with the same geographic variables from the existing model) and evaluate this model using the EPA regulatory monitors in the same 6 urban areas (n = 68). I also develop hybrid models using a combination of the PPA and EPA monitors. Finally, I focus on comparing how long-term concentration estimates derived from different LUR modeling approaches impact exposure assessment.

4.2. Materials and Methods

4.2.1. PPA Data Preparation

4.2.1.1. National model data

In this study, I used an existing national LUR model for year-2015 annual average PM2.5 developed as part of the Center for Air, Climate, and Energy Solutions (CACES; hereafter CACES LUR; Kim et al., 2020). Briefly, the model is developed using a partial least squares-universal kriging (PLS-UK) approach; the universal kriging framework partitions annual average concentrations into (1) a variance component that accounts for spatial and non-spatial variability and (2) a mean component based on a small number of reduced dimension variables from partial least squares of a larger number of independent variables Kim et al., 2020). Independent variables included 11 categories of geographic variables (e.g., traffic, population, land use, and satellite air pollution measurements) around the regulatory $PM_{2.5}$ monitoring sites (n = 757) in 2015. Many variables were calculated as point data or at various buffer sizes (50m – 15 km) resulting in 339 independent variables; a list of independent variables is in Table A4.1. Generally, the CACES LUR had a reasonable performance using internal data (e.g., random 10-fold cross validation $PM_{2.5}$ in 2010: $R^2 = 0.85$; standardized RMSE = 0.13). Details of model development and internal evaluation can be found here (Kim et al., 2020).

4.2.1.2. PPA data assembly

I retrieved $PM_{2.5}$ measurements from a crowd-sourced low-cost sensor network (i.e., PPA). The PPA sensors measure real-time $PM_{1.0}$, $PM_{2.5}$, and PM_{10} concentrations that are calculated based on light scattering and conversion factors from the manufacturer (https://www2.purpleair.com/). Since our analyses aimed to assess performance from multi-city low-cost sensors, I selected core-based statistical areas (CBSA; hereafter cities; n = 6) that have at least 7 EPA regulatory monitors and PPA sensors each. Namely, Los Angeles-Long Beach-Santa Ana, CA (LA), New York-Northern New Jersey-Long Island, NY-NJ-PA (New York), Phoenix-Mesa-Glendale, AZ (Phoenix), Pittsburgh, PA (Pittsburgh), Riverside-San Bernardino-Ontario, CA (Riverside), and Washington-Arlington-Alexandria, DC-VA-MD-WV (DC). I retrieved and assembled ambient (with “outdoor” label on the website) hourly $PM_{2.5}$ data (2015 – 2018) at all available PPA sensor sites in the 6 cities using Python 3.6 (Python Software Foundation) and RStudio 3.5.2 from the ThingSpeak and PPA Application Interface Programming.

4.2.1.3. Quality assurance/quality control (QA/QC) and the PPA data correction

I selected PPA sensor sites using the same criteria as the CACES national LUR model. Specifically, I retained sites with at least 18 hours/day, 244 days/year, and no more than 45 consecutive days without measurements with the goal of calculating annual averages (S. Kim et al., 2020). The PPA monitors have two sensors ("channels") in each monitor. I primarily used measurements from channel A and used channel B data only when data from channel A was not available. I also used the two channels to exclude spurious data by removing hours when the absolute difference between the two channels were larger than a predefined threshold (i.e., 3 μg/m^3 or 20% of the maximum channel readings, whichever is greater; Malings et al., 2019). For sites where it was possible to calculate annual averages for more than one year, I chose the year that was closest to 2015 to most closely align with the CACES LUR model estimates. Due to sensitivity of the PPA sensors to relative humidity and temperature (Y. Wang et al., 2015; R. Xu, 2015), I removed obvious sensor errors (e.g., $PM_{2.5}$ concentrations > 1,000 μg/m^3, temperature > 120 °F, and relative humidity > 100%). These procedures resulted in 149 valid PPA sites in the 6 cities. A summary of obtaining valid PPA measurements is in Table A4.2.

I mainly used a physics-based Hygroscopic Growth (HG) correction method from co-location studies to adjust the raw hourly $PM_{2.5}$ measurements of the PPA sensors (Malings et al., 2019). In general, I calibrated using (1) the HG factor that accounts for temperature and relative humidity by referencing at conditions of reported regulatory data (22 °C and 35% relative humidity) and (2) hygroscopicity of bulk aerosol that considers particle composition changes during seasons. I used the PPA measurements with relative humidity below 95%. Finally, I applied additional linear corrections to calibrate factory-to-ambient differences using co-located measurements with the federal equivalent Beta Attenuation Monitors (BAM) in Pittsburgh (Malings et al., 2019) and

Riverside (South Coast Air Quality Management District, 2018). I applied the best available city-specific chemical composition, thus hygroscopicity to the HG method for each city. Preferably, for cities (or nearby cities) where I had co-located regulatory-grade monitors, I calculated hygroscopicity based on the co-located monitors (i.e., Pittsburgh, LA, and Riverside); for cities where this was not possible, I used the Interagency Monitoring of Protected Visual Environments (IMPROVE) samples for the monitors closest to each city (i.e., New York, Phoenix, and DC; Figure A4.1; IMPROVE, 2019). Since city-specific composition may not be readily available if more cities are included, I also tested another Empirical Correction (EC)-based method for the 6 cities to compare calibration performance (Figure A4.2; Malings et al., 2019). The PPA data assembly framework is presented in Figure A4.3.

4.2.2. LUR Development Using the PPA Data

4.2.2.1. External evaluation of the CACES LUR

I used the adjusted hourly PPA $PM_{2.5}$ measurements to calculate the annual average concentrations of all PPA sites for our analyses. I drew a map to compare the PPA data and the CACES LUR predictions in LA area and to identify the disagreement. I then performed external evaluation by comparing the PPA data to the CACES LUR predictions for all 6 cities. I explored how the external evaluation varied by the proximity of a PPA site to potential sources, including (1) sites near major road segments (< 200m) based on the 2015 TIGER/Line Shapefiles, (2) sites with more than 5 restaurants within 500m from Google Point of Interest data (Chapter 3), and (3) sites with major facility emissions (annual emissions > 45 short tons per year) within 5,000m according to the 2017 National Emissions Inventory data. I investigated the PPA sites near major emissions vs. all other sites (i.e., background) for the three types of sources.

4.2.2.2. Development of the PPA LUR

To test the potential of using a low-cost sensor network for multiple cities, I used the PPA measurements in the 6 cities to develop LUR models using the same PLS-UK approach as the CACES LUR. Our dependent variable was $PM_{2.5}$ annual average concentrations at the 149 PPA sites. Our independent variables included the same 339 geographic variables as the CACES LUR models and were assigned based on values from the nearest Census Block, the smallest spatial resolution for the existing CACES LUR.

I conducted internal evaluation of the PPA LUR using the PPA data of the 6 cities (n= 149). Briefly, I used 10-fold cross-validation (CV): randomly divided all PPA sites into 10 groups, held out one group, developed models using the remaining groups, and predicted the hold-out group. Similar to the CACES LUR external evaluation, I performed external evaluation using the EPA monitors of the 6 cities (n = 68). Then, I compared the internal and external evaluation of the PPA LUR model.

4.2.3. LUR Development Using Both the PPA and EPA Data

I developed a hybrid LUR model (hereafter Hybrid LUR) by incorporating $PM_{2.5}$ measurements of (1) the 6-city EPA monitors (n = 68) and the 6-city PPA sensors in the 6 cities and (2) the national EPA monitors (n = 757) and the 6-city PPA sensors (n = 149). Similar to the process of internal evaluation of the PPA LUR, I conducted internal evaluation for the Hybrid LUR.

4.2.4. Exposure Assessment Based on Different LUR Modeling

I used population-weighted $PM_{2.5}$ annual average predictions from the three different LUR models based on combinations of the national EPA monitors and 6-city PPA monitors (CACES LUR, PPA LUR, and Hybrid LUR) to assess human exposures that could be used in epidemiological studies. To further investigate how the EPA data from national monitors vs. the 6 cities impacts the exposure assessment based on the CACES LUR and the Hybrid LUR, I created 5 scenarios of exposure maps: CACES LUR (National EPA data), CACES LUR (6-city EPA data), PPA LUR (6-city PPA data), Hybrid LUR (6-city PPA + National EPA data), and Hybrid LUR (6-city PPA + 6-city EPA data). Predictions were developed for all locations within the 6 cities except areas with zero population based on the US Census, at the Block Group (n = 31,911).

4.3. Results and Discussion

4.3.1. External Evaluation Performance of the CACES LUR

4.3.1.1. Descriptive statistics of measurements and CACES LUR

After conducting QA/QC for the PPA measurements, I drew the boxplot of the EPA monitors (n = 68), PPA sensors (n = 149), and CACES LUR predictions at Census Blocks (n = 478,756) in Figure 4.1. Generally, the PPA measurements reported higher concentrations than the EPA monitors. This result may be attributable to a bias towards high concentration locations (e.g., near traffic or industrial area) from PPA users. For example, LA had many more PPA than EPA sites (103 vs. 12), thus the PPA sites potentially had better chance to capture more localized emission sources and spatial variability. Comparatively, Phoenix had equal number of the EPA and the PPA monitoring sites (n = 7) reporting similar annual average concentrations. In addition, interquartile range (IQR) of the concentrations varied by city suggesting that different sampling locations between the regulatory and crowd-sourced sensors may impact the variability of measurements. Developed using the EPA monitors, the CACES LUR predicted similar-to-smaller IQR than the EPA measurements, which may be explained by the fact that (1) the Census

Block predictions derived from the national CACES LUR includes many more locations than the EPA monitoring network (478,756 vs. 68) including many likely low-concentration areas, and (2) owing to the model structure, the CACES LUR may not capture all of the variability demonstrated within the EPA monitoring network.

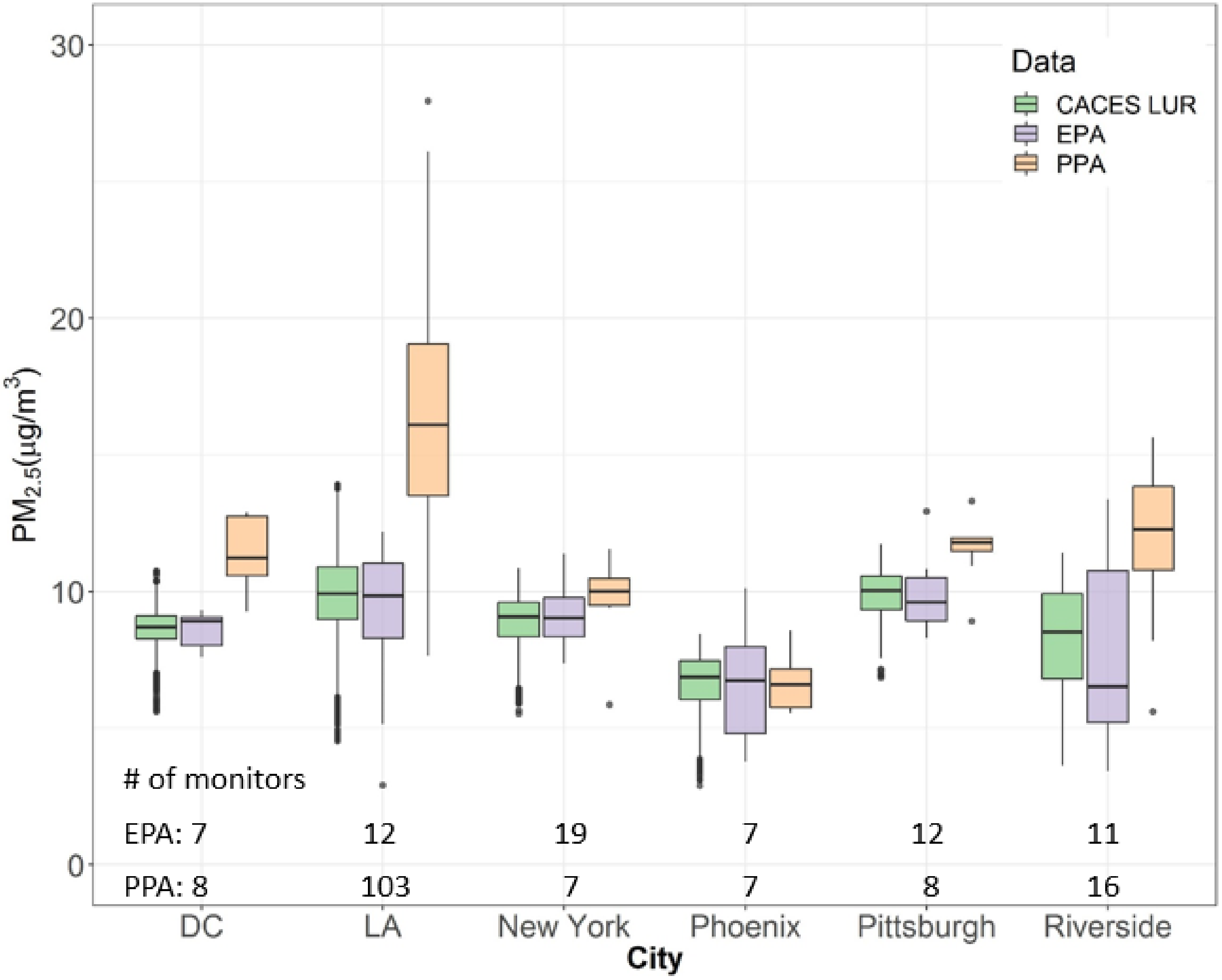


Figure 4.1. Boxplot of $PM_{2.5}$ concentrations by city and data source: EPA monitors (n = 68), PPA sensors (n = 149), and CACES LUR predictions at Census Blocks (n = 478,756).

4.3.1.2. External evaluation of the CACES LUR

The comparison of the CACES LUR predictions and the PPA measurements of LA area identified that concentration disagreement seemed to be larger at residential areas near industrial facilities or highway segments (Figure 4.2). This finding suggests that the PPA sensors may help pick up “hotspots” that were not be captured by the CACES LUR model. Table 4.1 shows a summary of the EPA and PPA data for the 6 cities as well as external evaluation results of the CACES LUR. PPA mean concentrations were higher than the EPA monitors (15 vs. 8.7 µg/m^3); the disagreement was most prominent in LA and Riverside. The performance of the external evaluation of the CACES LUR using all 6-city PPA measurements (pooled evaluation) showed modest performance (R^2 = 0.41, MAE = 5.5 µg/m^3) indicating that the CACES LUR may fail to capture the PPA sites where high concentrations were reported (Figure A4.4). Comparatively, the single-city evaluation performed worse than the pooled evaluation for most cities (Table 4.1, Figure A4.5). Pittsburgh (R^2 = 0.72, MAE = 1.7 µg/m^3) and Riverside (R^2 = 0.64, MAE = 4.0 µg/m^3), where co-located PPA and regulatory monitors were available for the PPA data correction, performed better than the pooled evaluation. This finding indicates that co-locating the PPA sensors with the EPA monitors improves the PPA calibration. The disagreement between the CACES LUR and the PPA measurements implies the potential use of the PPA data in LUR models. The external evaluation using the HG corrected PPA data outperformed that using the EC method for the 6 cities in our study Figure A4.6 – A4.11, which suggests that our use of the HG correction method is more appropriate in this study.

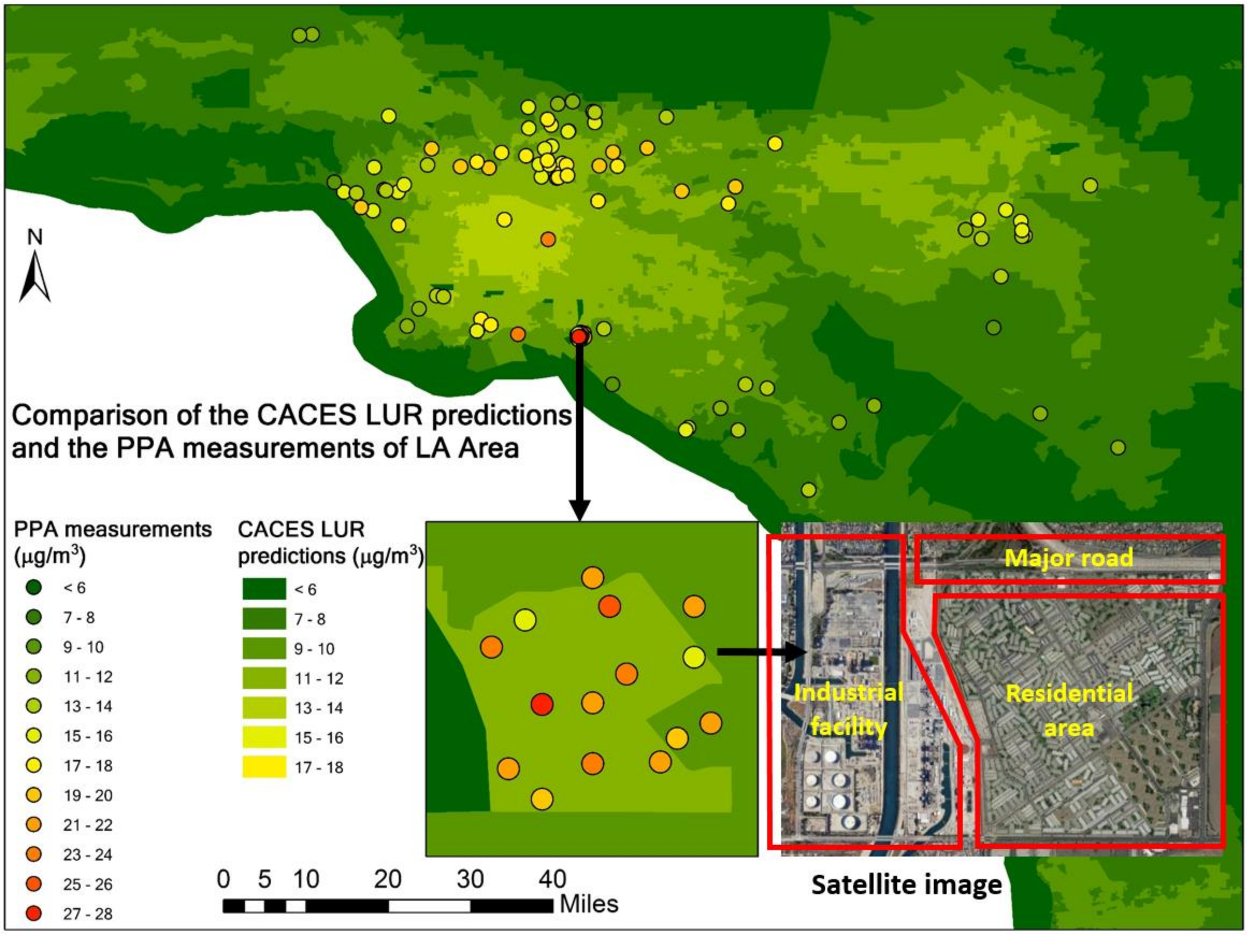


Figure 4.2. Comparison of the CACES LUR predictions and the PPA measurements of LA area.

Table 4.1 Summary Statistics of the EPA Monitors and PPA Sensors and External Evaluation of the CACES LUR

Core-based Statistical Area (CBSA)	# Monitors		Mean Concentrations (μg/m^3)		External Evaluation	
	EPA	PPA	EPA	PPA	R^2	MAE
Los Angeles-Long Beach-Santa Ana, CA	12	103	9.1	16	0.28	6.8
New York-Northern New Jersey-Long Island, NY-NJ-PA	19	7	9.1	9.6	0.31	2.1
Phoenix-Mesa-Glendale, AZ	7	7	6.6	6.6	0.28	0.71
Pittsburgh, PA	12	8	9.9	12	0.72	1.7
Riverside-San Bernardino-Ontario, CA	11	16	7.7	12	0.64	4.0
Washington-Arlington-Alexandria, DC-VA-MD-WV	7	8	8.6	11	0.020	2.5
All 6 cities	68	149	8.7	15	0.41	5.5

Note: MAE: mean absolute error (μg/m^3). Concentrations and R^2 are at two significant figures.

To further investigate whether the disagreement is a general trend for potential emission sources. I conducted external evaluation using the PPA sites that were near major emission sources (i.e., traffic, restaurants, and NEI facilities) vs. all other sites. Generally, when accounting for all the three major sources, no general trend was found at sites in proximity (R^2: 0.21, MAE: 6.9 μg/m^3) vs. background (R^2: 0.41, MAE: 5.5 μg/m^3; Figure A4.12). Similar results were found in the city-specific evaluations (Figure A4.13 – A4.18). While looking at each source separately, no obvious pattern was found for all 6 cities except for the NEI facilities (Figure A4.19 – A4.21). For example, for PPA sites near high NEI emissions (within 5,000m), the external evaluation performed (R^2: 0, MAE: 8.0 μg/m^3) worse than the background (R^2: 0.51, MAE: 4.7 μg/m^3) indicating that such high emissions may not be fully captured in the CACES LUR models (given limited number of EPA monitors surrounding major industrial facilities). This finding implies that the PPA sensors may be integrated into LUR models to help capture the near-source variability.

4.3.2. PPA LUR Models

4.3.2.1. Internal and external evaluation of the PPA LUR

I used the PPA data (n = 149) instead of the EPA data to develop LUR models (i.e., PPA LUR); I conducted external evaluation using the EPA monitors (n = 68). The internal evaluation showed reasonable performance (R^2: 0.66; MAE: 2.01 μg/m^3) indicating that the PPA data may be feasible to develop LUR models. The external evaluation performance dropped (R^2: 0.39; MAE: 4.1 μg/m^3) as compared to the internal evaluation; but the slope was consistent around the 1:1 line (Figure 4.3). Notably, the PPA LUR seemed to consistently over-predict concentrations at the EPA sites (Figure 4.3, right panel); however, there were no EPA concentrations higher than 20 μg/m^3 available to further explore if such pattern remains at high-concentration sites. These findings suggest that the PPA network could potentially be used successfully in places where large regulatory monitoring network don't exist.

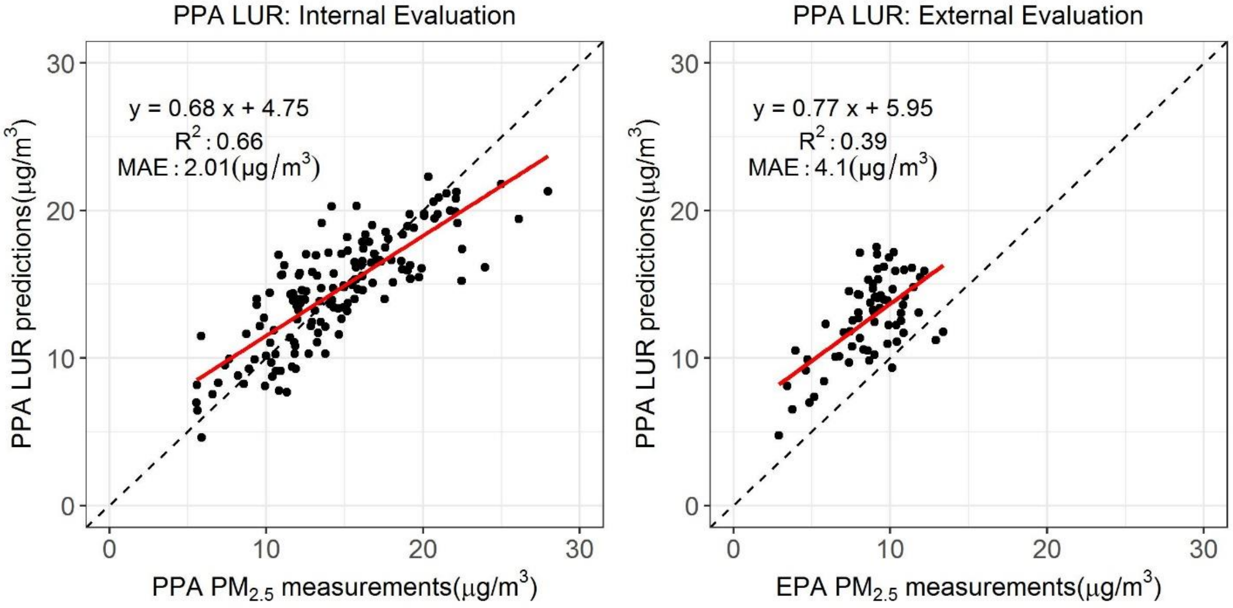


Figure 4.3. Internal evaluation vs. external evaluation of the PPA LUR models. Internal evaluation: 149 PPA measurements; external evaluation: 68 EPA measurements.

4.3.3 Hybrid LUR Models

I developed the Hybrid LUR models using a combination of the EPA and PPA measurements (Figure 4.5). Generally, the Hybrid LUR performed better than the separate models (i.e., CACES LUR and PPA LUR). Specifically, when using data from the national EPA monitors, the Hybrid LUR performed similarly (R^2: 0.85; MAE: 1.02 µg/m^3) as compared to the CACES LUR (R^2: 0.83; MAE: 0.72 µg/m^3; Figure 4.5: panel D vs. panel A). When using the EPA and PPA data from only the 6 cities, the PPA LUR (R^2: 0.66; MAE: 2.01 µg/m^3) showed similar performance as compared to the CACES LUR (R^2: 0.67; MAE: 0.99 µg/m^3; Figure 4.5: panel C vs. panel B); however, combining both data improved model performance: Hybrid LUR (R^2: 0.77; MAE: 1.62

μg/m^3; Figure 4.5: panel E). These findings indicate that integrating the PPA data into LUR models could improve model performance.

4.3.4. Exposure Assessment Using Population-weighted Concentrations

4.3.4.1. Boxplot of the three exposure assessment scenarios

Figure 4.4 shows the boxplot of $PM_{2.5}$ Predictions of the CACES LUR, PPA LUR, and Hybrid LUR at Block Group level for the 6 cities. In general, the PPA LUR had the highest median concentrations and prediction variability. A likely reason for the inflated PPA LUR results is that the full sample was dominated by measurements in LA (~2/3 of the measurements), where concentrations from the PPA sensors were higher than the EPA monitors (Figure 4.1), and the consistently higher concentrations in LA was probably transferred to other cities via LUR modeling. This transferability may also be found when the Hybrid LUR predictions in Phoenix and Riverside were slightly higher than the CACES LUR; however, the CACES LUR and Hybrid LUR shared similar central tendencies of predictions for most cities.

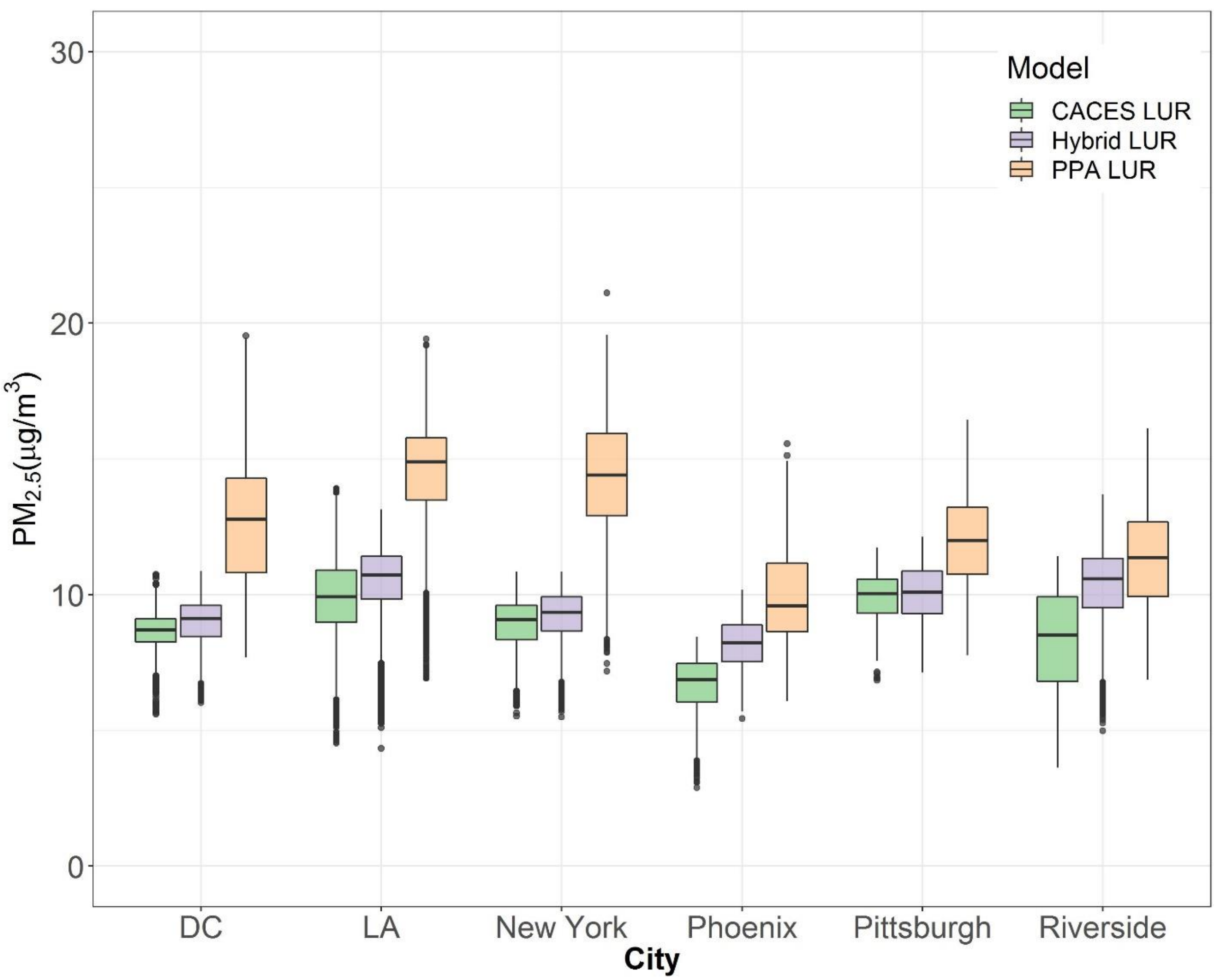


Figure 4.4. $PM_{2.5}$ Predictions of CACES LUR, PPA LUR, and Hybrid LUR at the Block Group level (n = 31,911).

4.3.4.2. Comparison of the CACES LUR, PPA LUR and Hybrid LUR

I compared the three exposure assignment models (5 scenarios) of population-weighted $PM_{2.5}$ concentrations for LA at the BG level in the main text (Figure 4.5) since the PPA data in LA had (1) the highest variability and (2) most number of sensors (n = 103) among the 6 cities (comparisons for other cities are in Figure A4.22 – A4.26). I found that the PPA LUR predicted higher concentrations as compared to the CACES LUR and the Hybrid LUR. To investigate whether the PPA LUR predictions were mostly due to elevated PPA measurements, I normalized the predictions to the model mean. I found that the PPA LUR had more sparse areas of high concentrations as compared to the CACES LUR (developed using the national EPA data or the 6-city data; Figure A4.27 – A4.32). This result indicates that the PPA LUR could potentially capture more within-city variability due to higher density of sensors deployed around the city as compared to the CACES LUR. This finding could also be reflected by the sparsely distributed higher concentrations in the Hybrid models as compared to the CACES LUR (Figure 4.5: panel D vs. panel A and panel E vs. panel B). The difference between the models using the national EPA data and the 6-city EPA data was not large (Figure 4.5: panel A vs. panel B and panel D vs. panel E). To include as many as the EPA monitors, I focused on analyzing the model scenarios with the national EPA data: CACES LUR (National EPA data) and Hybrid LUR (6-city PPA + National EPA data).

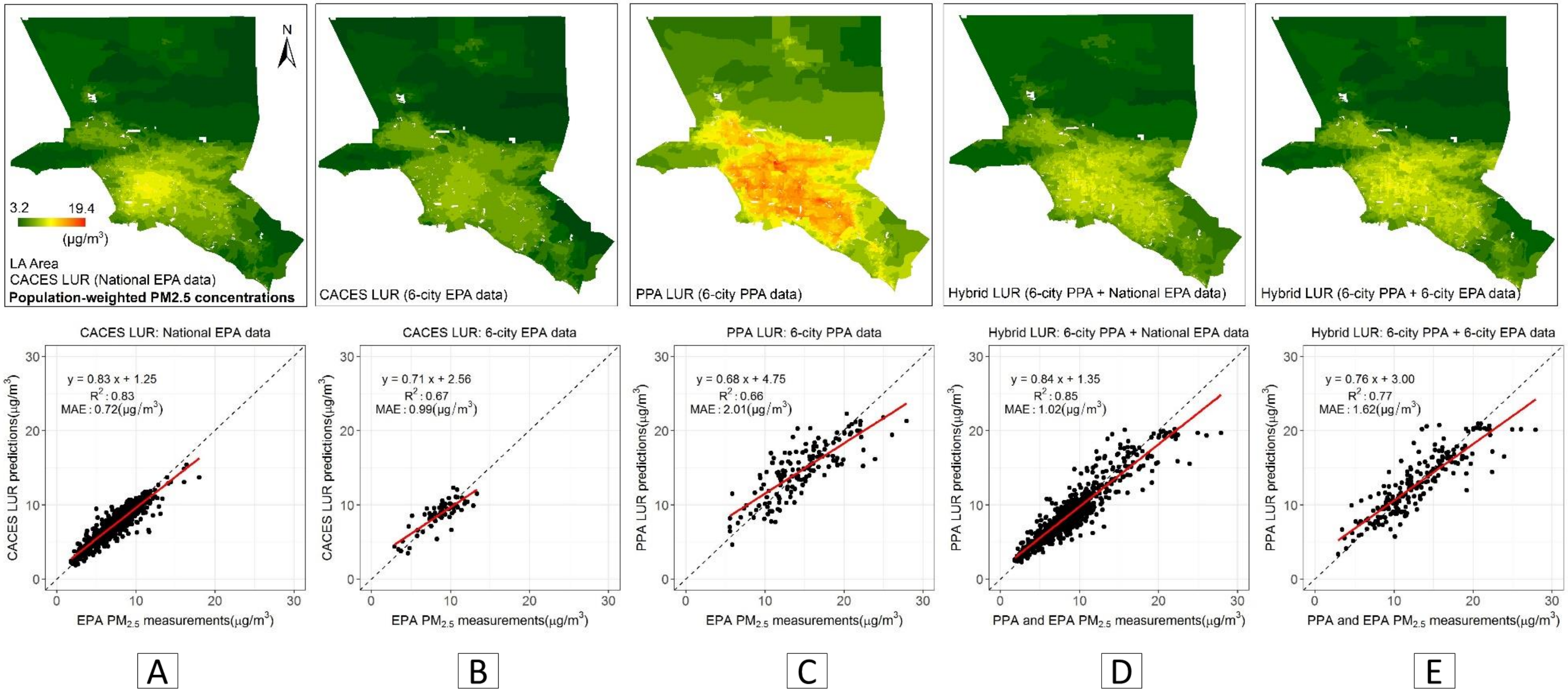


Figure 4.5. Population-weighted PM$_{2.5}$ concentration maps of the CACES LUR, PPA LUR, and Hybrid LUR in LA (top 5 panels) and internal evaluations of LUR model comparisons (bottom 5 panels). Panel A is the CACES LUR developed using the national EPA data; panel B is the CACES LUR using the 6-city EPA data; panel C is the PPA LUR using the 6-city PPA data; panel D is the Hybrid LUR using the national EPA data and the 6-city PPA data; panel E is the Hybrid LUR using the 6-city EPA data and the 6-city PPA data.

Figure 4.6 shows scatterplots of the three exposure scenarios of LA and the 6 cities (scatterplots of other cities are presented in Figure A4.33 – A4.37). I found that the PPA LUR and the CACES LUR predictions were spatially correlated while the PPA LUR predicted higher $PM_{2.5}$ concentrations overall. The association between the PPA LUR and the CACES LUR predictions was the strongest in LA among all 6 cities (R^2: 0.62, MAE: 4.52 µg/m^3 with a near 1:1 slope). This alignment was also consistent for the CACES LUR including all 6-city data, though with a drop in R^2 (R^2: 0.46, MAE: 4.50 µg/m^3 with a near 1:1 slope). Both scatterplots (top two panels of Figure 4.6) reveal that the PPA LUR predictions were consistently higher than those from the CACES LUR. Similarly, the Hybrid LUR predictions were also correlated with the CACES LUR for the 6 cities (strongest for LA); the correlation dropped when including data of all 6 cities (R^2: 0.82 vs. 0.64). These findings indicate that including the PPA data might introduce some areas of higher concentration that may not be predicted by the CACES LUR; however, we need a bit more care of (or confidence in) the absolute concentrations reported by the PPA sensors to verify whether these are "real" areas of high concentrations.

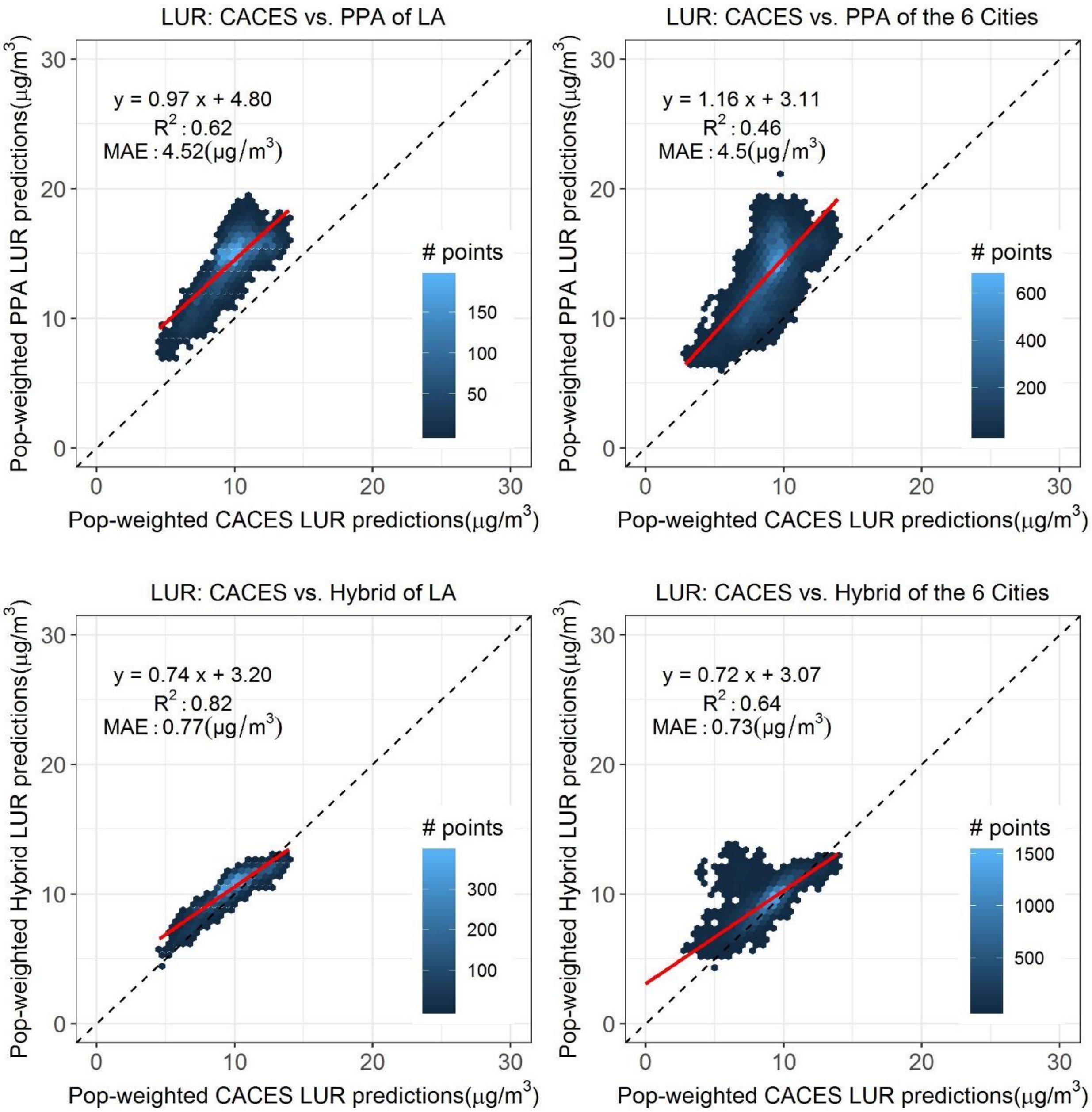


Figure 4.6. Scatterplots of the PPA LUR and CACES LUR comparisons (top two panels) and the Hybrid LUR and CACES LUR comparisons in LA and the 6 cities (bottom two panels).

4.3.5. Implications for LUR Evaluation and Development Using Low-cost Sensors

I used a crowd-sourced low-cost sensor network using data from multiple US cities rather than single cities (Dijkema et al., 2011; Eilenberg et al., 2020; Jain et al., 2020; Masiol et al., 2018; Weissert et al., 2019) to explore the contribution of low-cost sensing to LUR models. I developed LUR models using the PPA data from 6 US cities and obtained promising performance (i.e., PPA LUR: R^2: 0.66), which indicates that the crowd-sourced multi-city low-cost sensor data may be plausible for LUR development. In addition, integrating the 6-city low-cost sensor data into existing regulatory-based CACES LUR achieved improved performance. For example, when using the EPA data nationwide, adding the 6-city PPA data into the LUR model resulted in similar performance (R^2: Hybrid PPA [0.85] vs. CACES LUR [0.83]). While combining both the EPA data and the PPA data from the same 6 cities, the Hybrid model outperformed the CACES LUR (R^2: Hybrid PPA [0.77] vs. CACES LUR [0.67]). These findings suggest that the PPA data is a promising dataset to improve existing LUR. Further, the contribution of adding only the PPA data from 6 cities (n = 149) to the national EPA data network (n = 757) may not be as obvious as Hybrid models developed using the 6-city PPA data and the 6-city EPA data (n = 68). As such, the low-cost sensor networks could potentially be as useful as regulatory networks for model building when high-grade measurements aren't available. Our study reinforced the potential contribution of the low-cost sensor network to LUR modeling similar to other studies (Bi et al., 2020b; Huang et al., 2019; Schneider et al., 2017; Weissert et al., 2019) Particularly for $PM_{2.5}$, a recent study in Southern California found that combining the low-cost sensors with the regulatory monitors using a satellite-based random forest hourly LUR models could effectively improve $PM_{2.5}$ predictions (CV R^2 increased by ~0.2; Bi et al., 2020).

Developing LUR models with refined spatial resolution and coverage is an important goal for exposure assessment. LUR models developed from only regulatory monitors may mischaracterize within-city variability especially for sparsely monitored locations (Pope et al., 2019). Our approach of using the crowd-sourced low-cost sensor network to develop LUR models may offer insight to environmental and health scientists. For example, the well-correlated scatterplots among the population-weighted PPA LUR, Hybrid LUR, and the CACES LUR are positive signs for using the low-cost sensors. For example, they could serve as a substitute to some places where monitoring is sparse and be integrated into existing regulatory monitor-based LUR models for improving exposure assessment. In addition, the exposure assessment maps of the PPA LUR and the Hybrid LUR models had identified more within-city variability as compared to the CACES LUR predictions. The different spatial patterns indicate that the low-cost sensor network may be useful for near-source areas. Similar findings have been found in another study using an ongoing low-cost sensor network in the New York City Community Air Survey, which identified more “hotspots” in traffic and populous areas (Huang et al., 2019). While our study could not confirm whether the disagreement came from the underestimation of the CACES LUR models or the overestimation of the PPA LUR, our results are encouraging for the potential of the multi-city low-cost sensors to be used in LUR models to merit further investigation on these areas of discrepancy.

The usefulness of the low-cost sensor network for LUR models may also be reflected by some disagreement between the PPA data and the regulatory monitor-based CACES LUR (Figure 4.2 and Table 4.1). The PPA users tend to monitor high-concentration locations (e.g., industrial areas or traffic segments), which may help pick up such areas not predicted by the CACES LUR. I found that the pooled CACES LUR model evaluation outperformed the city-specific models, similar to a study in Canada that found the national models could generally capture pollutant variability at regional-scale instead of local-scale (Hystad et al., 2011b). This is likely due to the fact that the CACES LUR models (1) were based on national patterns and potentially mischaracterized within-urban variability and (2) may be sensitive to individual cities with different siting of regulatory monitors. In this case, models incorporating the PPA data might help capture local variability and resolve the complex spatial heterogeneity of within-city air pollution levels.

One key potential of using the low-cost PPA sensors is the increased and promising network. The PPA sensors has by far the largest global network (n = ~ 7000 nodes) and the expansion of such network would be more significant for places where regulatory-grade monitoring networks are limited (e.g., developing countries) or don’t exist (e.g., rural areas; Reis et al., 2015). For example, Africa lacks sufficient monitors to track air quality leaving nearly a billion people have no pollution exposure information at all (IQAir, 2019). The available PPA sensor data as well as other similar types of low-cost sensor network could provide opportunities for developing LUR models to predict $PM_{2.5}$ exposure. Another strength of the PPA network is that it has many sensors tracking real-time $PM_{2.5}$ for several months of a year rather than a few weeks, contributing to more representative annual averages. The PPA data I used to calculate annual averages in this study included at least 2/3 days of the entire year. While the regulatory EPA data is known to be the “gold standard”, the crowd-sourced PPA data could potentially become a valuable dataset with careful quality control and side-by-side calibration from the regulatory-grade monitors.

I followed the EPA criteria to filter and calibrate the PPA data with the goal of obtaining reliable annual averages of $PM_{2.5}$. Previous studies typically used measurements based on one sensor per site or duplicated measurements at limited sites at best (Eeftens et al., 2012b; English et al., 2017; Huang et al., 2019). I was able to censor the two channels of the PPA sensor per site to reduce the bias of measurements. Although the low-cost sensors may not be as reliable as regulatory monitors, a good correlation between the low-cost sensors and the regulatory reference sites has been identified (Brauer & Lee, 2018; Mead et al., 2013; South Coast Air Quality Management District, 2018). To account for the sensitivity of the low-cost sensors to humidity and temperature and to approach to regulatory-grade data, I have applied the HG correction approach (Malings et al., 2019) based on available reference monitors in Pittsburgh (n = 9) and Riverside (n = 3). Our data calibration suggests that more co-located sensors with regulatory-grade monitors from a variety of locations could reduce the uncertainty of the PPA measurements and mitigate the concerns for developing LUR models.

4.3.7. Limitations and Future Research

Our work could be improved in several ways in the future. Our 6-city PPA data was dominated by the LA sample (103 out of 149) and could introduce bias to predictions of other cities for capturing within-city variability. Developing LUR models with a larger sample of PPA data (e.g., integrating PPA sensors nationwide) could allow for improving exposure estimates for multiple cities. I web-scraped crowd-sourced PPA data, but had limited information on monitoring purpose, sampling campaign, and sensor ownership among different cities. For example, the user-oriented sensor network could lead to a bias of sampling at locations near high emission sources (e.g., industrial facility). As such, LUR models developed using only sensors based on such preference are more likely to over-predict other areas. A useful test for investigating this issue is to assess locations where the low-cost sensor and regulatory-based models disagree. Another limitation is that our QA/QC procedure was not able to identify whether the PPA sensors had a systematic bias, which warrants caution for its sole use of model development. The improved performance of the Hybrid LUR models suggests that combining data from both the low-cost sensors and regulatory monitors could help reduce the uncertainty (Williams, 2019), thus allowing for spatially-refined exposure estimates to investigate the air quality impacts on human health. The confidence of using the low-cost sensor data could also be gained through testing more co-located PPA sensors with the regulatory monitors. For example, prior work has showed that the PPA data was reliable based on investigations of two cities (Brauer & Lee, 2018; South Coast Air Quality Management District, 2018), future work could aim to reduce uncertainty by expanding the evaluation of the PPA sensors for various locations and multiple cities. Lastly, the PPA measurements I used to calculate annual averages were mostly from 2017 and 2018 due to data availability while the latest CACES LUR predictions were for 2015. For example, the California wildfire in December 2017 may be one incident to elevate the air pollution level; however, a filtered boxplot (excluding the PPA data of December 2017 in LA and Riverside; Figure A4.38) indicated a minor influence on the annual averages. To better inform application of the PPA data, I remained using the full year of data. Future studies can be refined by using same years of PPA sensor measurements and CACES model predictions once newer versions of LUR models become available.

4.4. Conclusion

The novelty of this study is to use an emerging crowd-sourced low-cost sensor network (i.e., PPA) to develop LUR models for multiple cities. Our work suggests that the low-cost sensors may offer a promising alternative to fill the gaps of existing regulatory monitoring network (e.g., sparse or no monitors). Most importantly, I demonstrate how the approach of using crowd-sourced sensor network for LUR models could provide information for places where regulatory monitors are not available and serve as a promising strategy to improve exposure assessment of spatial resolution and coverage. Further, our approach of using data from the 6-city PPA sensors for LUR models could be expanded to include the PPA data nationwide or other low-cost sensor networks available. This idea could become more significant for rural areas and developing countries where regulatory monitor networks are not rich, thus tracking exposure patterns of $PM_{2.5}$ more accurately and informing environmental and health policies.

Chapter 5: EXPLORING NEW PREDICTOR DATA SOURCES TO DEVELOP NATIONAL LAND USE REGRESSION MODELS FOR CRITERIA POLLUTANTS

ABSTRACT

Most empirical land use regression (LUR) models are developed using hundreds of geographic variables (e.g., road lengths, area of built land; hereafter “traditional” variables). An understudied topic is whether “new” data sources capable of capturing street-level features uniformly (e.g., point of interest [POI], Google street view [GSV], and local climate zones [LCZ]) could alternate traditional variables and improve LUR models. I developed national LUR models for predicting annual average concentrations of six criteria pollutants (e.g., NO_2, $PM_{2.5}$) in the US. I compared models with combinations of new and traditional variables based on different modeling approaches (e.g., machine learning [ML]). Model performance was similar for both variable scenarios (e.g., random 10-fold CV R^2 of ML-kriging models for NO_2, new vs. traditional: 0.89 vs. 0.91); whereas adding the new variables to the traditional LUR models didn’t necessarily improve model performance. Models with kriging effect outperformed those without (e.g., CV R^2 for $PM_{2.5}$ using the new variables, ML-kriging vs. ML: 0.83 vs. 0.67). The contribution of the new variables to LUR models highlights the potential of substituting traditional variables, thus enabling LUR models in areas with limited or no data (e.g., developing countries) and across political boundaries (e.g., cities).

Keywords:
Open data; urban morphology; local emissions; enhanced models

5.1. Introduction

Ambient air pollution contains a complex mixture of particles and gases that have shown adverse effects to human health (Jerrett et al., 2009; Laden et al., 2006). Knowledge of air quality primarily relies on limited regulatory air monitors, yet fail to capture air pollution concentrations at unmonitored locations (Demerjian, 2000; EPA, 2020). Land use regression (LUR) has been identified as an effective tool to predict air pollution concentrations at these locations (Adam-Poupart et al., 2014; Keller et al., 2015; Marshall et al., 2008; Van Donkelaar et al., 2016). Estimating air pollution levels with improved efficiency and spatial resolution/coverage necessitates better LUR models aiming at intra-city and multi-city air pollution assessment for health studies (Beelen et al., 2014; Di et al., 2017), environmental justice (Clark et al., 2014; Su et al., 2010), and urban planning (Hankey et al., 2017; Shi et al., 2016).

“Traditional” variables used for developing LUR models often include hundreds of geographic features (e.g., traffic, land use/land cover, and population dynamics; Hoek et al., 2008; Jerrett et al., 2005). These variables are often extracted and calculated based on various government-sponsored sources that may not be timely updated (e.g., US Census, national land cover database [NLCD]) using geospatial information system (GIS) and remote sensing techniques (Beckerman et al., 2013; Hankey & Marshall, 2015; H. Xu et al., 2019). The variable collection and processing demand large effort from data sources at different levels, thus difficult for generalizing LUR models across regions. For example, land use data is often limited at the national level since local jurisdictions classify and archive this information in different manners (Theobald, 2014). Another limitation is that traditional variables do not include urban morphology (e.g., street configuration, building height), which may be important to capture street-level air pollution variability (Edussuriya et al., 2014; Yuan et al., 2014; Tang et al., 2013).

New variables that require less different data sources and characterize more air quality-related features may be helpful for improving LUR models.

More recently, the column abundance and ground-level estimates of air pollution from satellite products (e.g., aerosol optical depth [AOD]) have been identified as significant variables for developing LUR models (Bechle et al., 2013; Martin, 2008), particularly at national or global levels (Knibbs et al., 2014; Novotny et al., 2011; Vienneau et al., 2013). The rapid development in data sciences (e.g., data mining) has enabled the access to multiple non-government sponsored data platforms (Bellinger et al., 2017; Lary et al., 2016; Sheng & Tang, 2011; Zheng et al., 2013). For example, Google point of interest (POI) provides place-based attributes (e.g., restaurants, gas stations) related to air quality (Hassan Amini et al., 2014; Wu et al., 2017b). In addition, Google street view (GSV) imagery-derived features could characterize street-level built environment (e.g., greenness, infrastructure; Larkin & Hystad, 2019; Rzotkiewicz et al., 2018). Another recently developed database is the local climate zones (LCZ), which classifies built and natural environment based on urban morphology and climate-related properties (Bechtel et al., 2015; Demuzere et al., 2020; Stewart & Oke, 2012a). The LCZ data has been widely used in temperature and climate studies, which may offer some insights for air quality modeling (Brousse et al., 2016; Steeneveld et al., 2016; Y. Xu et al., 2017). While these new sets of variables may offer a uniform way of capturing local features across large geographies, an understudied topic is whether they could shed some light on LUR model development and generalizability.

Conventional LUR models apply a stepwise regression approach that allows for multiple significant variables with varying buffers to be selected (Eeftens et al., 2012; Su et al., 2009). Recent progress has been made to develop parsimonious and hybrid models (e.g., partial least squares in a kriging framework [PLS-UK]) that input much less number of predictor variables (Kim et al., 2020). Machine learning (ML) models (e.g., random forest, neural network) allow for processing big data input with high predictive power for air quality (Bi et al., 2020a; Di et al., 2016). Questions remain on how these LUR approaches impact model performance using different sets of variables.

In this study, I develop national LUR models of six criteria pollutants (i.e., NO_2, O_3, $PM_{2.5}$, CO, PM_{10}, and SO_2) for the contiguous US that predict annual average concentrations based on regulatory monitors in 2015. My work aims to compare LUR model performance using different scenarios of predictor variables (i.e., “traditional”, “new”, and “all”) and check their performance consistency across different modeling approaches (e.g., stepwise regression, PLS-UK, and ML). That is, the “traditional” scenario includes geographic and satellite categories. The “new” scenario includes satellite, POI, GSV, and LCZ data. Lastly, the “all” scenario contains all of the candidate independent variables. I test the feasibility of using new variables and assess merging new variables with traditional variables in LUR models. I also evaluate the contribution of adding kriging effect to LUR models. I focus on how choice of variable input across different modeling approaches impacts LUR model performance.

5.2 Materials and Methods

5.2.1. Dependent Variables

The dependent variables of the models included criteria pollutant (i.e., pollutants that have been regularly monitored due to their adverse health effects; NO_2, O_3, $PM_{2.5}$, CO, PM_{10}, and SO_2) concentrations retrieved from the EPA Air Quality System (AQS) monitoring locations in 2015 (Kim et al., 2020). I developed models for all criteria pollutants, but focused on reporting only criteria pollutants (i.e., NO_2, O_3, and $PM_{2.5}$) in the manuscript. All the concentrations were annualized (except O_3) mainly based on the following criteria: computing locations with (1) at least 18 hours valid measurements per day and 244 days per year and (2) at most 45 consecutive missing days of measurements. The concentrations of O_3 was calculated using the daily maximum of the 8-hour moving average at locations with at least 18 hours per day during the summer season with predominant photochemical reactions (i.e., May to September). All pollution concentrations were square rooted to meet the normal distribution assumption. Detailed description of data preparation can be found in another study (Kim et al., 2020).

5.2.2. Independent Variables

I assembled multiple categories of candidate independent variables for developing LUR models in the goal of exploring how different variable input impacts model performance. Generally, I used three scenarios of variable inputs: traditional (i.e., geographic and satellite) vs. new (i.e., satellite, POI, GSV, and LCZ) vs. all (all candidate variables). Table 5.1 shows the full list of the candidate independent variables.

5.2.2.1. Geographic variables

To compare to a recent LUR study (Kim et al., 2020), I used the same sets of geographic variables including eight major categories (e.g., traffic, population, land use/land cover, and vegetation). Variables were calculated and tabulated as count, length, and area within appropriate buffers according to different data types; in general, buffer sizes ranged from 0.05 to 30 kilometers. This process resulted in ~ 360 variables of geographic category for the LUR models (Table 5.1). Detailed data processing and variable calculation process is also available in Kim et al., 2020.

Table 5.1 Candidate independent variables in the LUR models

Variable scenario			Variable category	Variable name	Variable type	Description	Data source
All	New	Traditional					
X		X	Geographic[a]	Traffic	Length in buffer (km)	Any road, A1, truck route, intersections, etc. (0.05-15 km)	TeleAtlas (http://www.teleatlas.com/OurProducts/MapData/Dynamap/index.htm)
X		X		Population	Count in buffer (person)	Population in block groups (0.5-3 km)	US Census (http://arcdata.esri.com/data/tiger2000/tiger_download.cfm)
X		X		Land use/land cover	Area in buffer (%)	Built land, open space, agricultural land, etc. (0.05-15 km)	US Geological Survey (http://water.usgs.gov/GIS/dsdl/ds240/index.html); MRLC (http://www.mrlc.gov/index.php)
X		X		Sources	Length in buffer (m)	Distance to the nearest source (e.g., railroad, airport)	National Emission Inventory (http://www.epa.gov/ttn/chief/net/2002inventory.html)
X		X		Emissions	Point in buffer (lb/ton)	Sum of site-specific facility emissions (3-30 km)	National Emission Inventory (http://www.epa.gov/ttn/chief/net/2002inventory.html)
X		X		Vegetation	Area in buffer (quantile)	Normalized difference vegetation index (0.5-10 km)	Satellite (http://glcf. umd.edu/data/ndvi/)
X		X		Impervious	Area in buffer (%)	Impervious surface value (0.05-5 km)	National Land Cover Database (http://www.mrlc.gov/index.php)
X		X		Elevation	Counts	Elevation above sea levels (1-5 km)	Calculated from (http://nationalmap.gov/elevation.htm)
X	X	X	Satellite[b]	Air pollution estimates	Column abundance or surface ($\mu g/m^3$ or ppb)	Satellite-based estimates (NO_2, SO_2, CO, HCHO, $PM_{2.5}$)	Multiple sources[b]
X	X		POI[b]	Point of interest	Count in buffer	90 categories of POI (e.g., gas station, restaurant)	Google Places API (https://developers.google.com/places/web-service/intro)
X	X		GSV[b]	Google street view	Object pixel (%)	57 categories of GSV-related features (e.g., tree, grass, person, building)	Calculated (https://developers.google.com/maps/documentation/streetview/intro)
X	X		LCZ[b]	Local climate zones	Counts in buffer	17 LCZs (e.g., compact high-rise, dense trees)	Calculated from (Demuzere et al., 2020)

[a] Detailed description can be found in the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air) Database (Kim et al., 2020).
[b] Detailed description is shown in the Appendix.

5.2.2.2. Satellite measurements/estimates

The annual average estimates of satellite-based air pollution concentrations (i.e., column abundance [atmospheric trace gas in the vertical column] or surface [ground-level estimates]) for NO_2, $PM_{2.5}$, SO_2, CO, and HCHO (formaldehyde) were obtained from different satellite products and datasets (e.g., aerosol optical depth [AOD], ozone monitoring instrument [OMI]; Boersma et al., 2011; Chance, 2007; Deeter et al., 2017; OMI Science Team, 2012; Stavrakou et al., 2008). The resolution of these gridded satellite-based estimates varied (NO_2, $PM_{2.5}$, HCHO: 0.1˚ * 0.1˚; SO_2, CO: 0.25˚ * 0.25˚; Table A5.1). The satellite estimates was assigned to the grid where EPA monitor was located.

5.2.2.3. Google POI

To explore alternative datasets for traditional land use and geographic data, I web-scraped Google POI from Google Places application programming interface (API) that returns all point-based location of interest within a specific buffer around the target location. Particularly, I used Python code to retrieve the POI data in 2018. This process resulted in number of counts for 90 POI categories (Table A5.2) at all EPA monitors of the criteria pollutants with buffer sizes of 100m, 250m, 500m, 750m and 1,000m (n = 450 variables). In general, the candidate categories could pick up details in land use that were not included in other datasets, including direct emissions from local business (e.g., gas stations, restaurants, auto shops, and dry cleaners; Table 5.1). More importantly, the POI data may serve as a uniform and localized land use proxy for assessing air quality impacts across political boundary (e.g., cities, regions), allowing for more generalizable air quality model in large geographies.

5.2.2.4. Google GSV

Another promising data source capable of capturing local features is the publicly available Google GSV, which provides georeferenced images containing street-level information along major road network in developed and underdeveloped countries (Rzotkiewicz et al., 2018a). In general, the GSV image collection and processing of this study involve two steps. First, I web-scrapped GSV images (n = 133,252) around the coordinates of the EPA monitors for criteria pollutants (1979-2015). At least five random locations within 100m-buffer of each monitoring location were sampled with a threshold of 20m for distance to the monitoring locations. Four panorama images covering four directions (0˚, 90˚, 180˚, 270˚) were extracted per location. This step resulted in at least 20 GSV images per monitoring location. To match the EPA monitors chosen in this study, 5,470 images were assembled in total for the image processing step. Second, each GSV image was processed using a deep learning algorithm (i.e., pyramid scene parsing network) to classify each pixel in the image (Zhao et al., 2017). Then, a python script was used to summarize each image to give a percentage of 150 feature categories (Andrew Larkin & Hystad, 2019). For the purpose of ambient air quality modeling, I tabulated only 57 categories of outdoor-related GSV features (e.g., tree, grass, and building; Table A5.3). The pixel results of the sampling locations around each monitoring location were averaged to obtain the final GSV-based variables for all air quality monitors.

5.2.2.5. LCZ

LCZ data classifies urban form into a total of 17 categories based on climate-relevant properties (Stewart & Oke, 2012). Most importantly, it enables characterizing urban areas in a consistent manner worldwide and captures urban morphology (e.g., building heights) that is often missing in existing urban form measures. Recently, the first continental US-wide LCZ map based on deep learning, remote sensing, and crowd-sourced data is available (Demuzere et al., 2020). Figure A5.1 shows the 17 LCZ classifications: LCZ 1-10 features built environment while LCZ A-G characterizes land cover. The LCZ surface for contiguous US includes 15 categories (excluding LCZ 7 and LCZ 9). I calculated the counts of the 15 LCZ categories within 5 buffer sizes (i.e., 500m, 1,000m, 1,500m, 2,000, and 2,500m) of the EPA monitors based on the LCZ surface. LCZ count calculation was conducted in ESRI ArcGIS (version: 10.6).

5.2.3. Modeling Approach

5.2.3.1. Stepwise Regression

I compared multiple modeling approaches using different scenarios of independent variables (traditional vs. new vs. all). First, I used a forward stepwise linear regression, which has been used commonly in LUR models for air pollution (Su et al., 2009), to select significant variables from candidate independent variables. Generally, it included two steps: (1) selecting the most correlated variable with the dependent variable and (2) adding the variable that was mostly correlated with the model residuals among the remaining variables. This process stopped when either a variable was not significant ($p > 0.05$) or the multi-collinearity indicator (variance inflation factor [VIF]) was greater than 5. I allowed variables to be selected with multiple buffers since the main purpose of this study is to make predictions.

I also followed a similar approach in a model comparison study (Mercer et al., 2011b) and added a kriging process (based on a minimum mean squared error interpolation) after the stepwise regression. That is, I incorporated spatial smoothing by kriging the residuals as a second stage after first estimating a trend in the stepwise regression (hereafter stepwise-kriging). The assumption is that this approach may improve model performance as compared to the traditional stepwise-based LUR models.

5.2.3.2. Partial least squares-universal kriging (PLS-UK)

I used PLS-UK modeling approach same as a recent LUR study (Kim et al., 2020) for the three independent variable scenarios. The modeling process involved two major components: variance and mean. Specifically, the universal kriging (using exponential covariance function for variogram) accounts for the variance component and PLS accounts for the mean component by reducing the dimensions of independent variables that served in a linear regression process. All kriging covariance parameters and PLS summary variables were based on a maximum likelihood approach. According to the study (Kim et al., 2020), models using 3-30 variables performed best depending on pollutant but mostly that was marginal compared to those using full dataset. Therefore, I used the 30-variable parsimonious model to develop LUR models for each pollutant. That is, only top 30 variables were selected by forward selection to enter PLS reduction and regression modeling. Details were described in Kim et al., 2020. I conducted the PLS-UK modeling in R (version: 3.5.2).

5.2.3.3. Machine learning (ML)

Finally, to explore how emerging ML algorithms could impact LUR model performance, I developed LUR models using ML algorithms for all three scenarios of variables (traditional vs. new vs. all). I compared nine common algorithms (e.g., random forest, gradient boosting) integrated in Python scikit-learn packages (Python version: 3.6.10) to develop LUR models. For example, random forest has a set of decision trees (constructed by the best splits randomly chosen through subset predictors) averaging for regression results in the final prediction. Gradient boosting optimizes model prediction in an iterative fashion by fitting on the negative gradients. I fine-tuned the parameters and selected the ML algorithm with the best performance (potentially highest predictive power and lowest error) among the nine ML algorithms. Then, I selected the best algorithm to represent ML approach to compare to the stepwise regression and PLS-UK approaches.

Similar to the stepwise regression, I also added a second-stage kriging step after estimating a trend in the ML process. I used kriging for the residuals from the ML algorithm with the best performance (using exponential covariance function for variogram; hereafter ML-kriging).

5.2.4. Modeling Evaluation

I conducted two types of 10-fold cross-validation (CV) to evaluate the LUR models. In general, random CV divided the monitoring locations into 10 groups randomly for training and testing while spatial CV divided the 10 groups using k-means clustering (Young et al., 2016). Each CV separately involves 10 times of following processes: (1) selecting one group out of the 10 groups as the hold-out group, (2) developing models using the remaining nine groups to predict concentrations at the hold-out group. In general, random CV accounts for model performance at random air monitoring locations while spatial CV reflects locations distant from a monitor.

I used standardized root mean square error (RMSE) and mean square error (MSE)-based R^2 to evaluate the CV performance. Briefly, the standardized RMSE (i.e., RMSE/mean concentrations of all monitors; hereafter RMSE) allows for comparison across the criteria pollutants. The MSE-R^2 (i.e., one minus the sum of squared error between the observations and the predictions divided by the sum of squared error between the observations and the mean of the observations; hereafter R^2) assesses agreement between observations and predictions on the 1:1 line instead of the regression line (Keller et al., 2015).

5.2.5. Modeling Comparison

To investigate how the LUR models performed among different scenarios, I focused on several comparisons below. First, assessing the model performance using the three scenarios of variable input: traditional vs. new vs. all. The goal was to evaluate whether (1) the new sets of variables could be alternative choices for LUR models when traditional data sources are not available and (2) adding the new variables to the traditional variables improves model performance. Second, I compared across the three modeling approaches with two extra approaches adding a second-step kriging smoothing (i.e., stepwise, stepwise-kriging, PLS-UK, ML, and ML-kriging) to assess how different modeling approaches impact model performance. Third, I compared across the pollutants to see whether the LUR models are sensitive to pollutant species. Then, I demonstrated the results of the two types of CV to indicate how spatial relationships among monitors would impact LUR predictions. Lastly, I investigated how different variable categories were chosen by each model to reveal their contribution to air pollution prediction. Specifically, variable importance (i.e., relative importance) in ML models was characterized by number of times each variable was selected (ML and ML-kriging shared the same variable selection process). For stepwise regression (stepwise and stepwise-kriging shared the same variable selection process), variable importance was represented by the normalized coefficients in the regression process during variable selection. That is, to multiple the variable coefficient by a factor that equals the difference of 95th and 5th percentile of the independent variable divided by the difference of 95th and 5th percentile of the dependent variable. I used variable importance in projection (VIP) score to measure the contribution of candidate variables in PLS-UK approach. Variables with VIP score close to or greater than one are recognized as important features for the models. The larger value of the variable importance indicators (i.e., relative importance, normalized coefficients, and VIP), the greater contribution of the variable to the LUR models. The 20 variables with greatest importance were reported.

5.3. Results and Discussion

5.3.1. Summary of Monitored Air Pollution Concentrations

The number of valid monitors in 2015 based on the selection criteria differed among pollutants, ranging from 196 (CO) to 821 (O_3). I summarized descriptive statistics of major criteria pollutants: NO_2, O_3, and $PM_{2.5}$ (Table 2). The annual mean (median) concentrations of each pollutant is NO_2: 7.8 (7.1) ppb, O_3: 44.2 (44.5) ppb, and $PM_{2.5}$: 7.7 (8.0) μg/m^3. The results of other criteria pollutants are in Table A5.4.

Table 5.2 Summary statistics of major criteria pollutants in 2015

Statistics	**NO_2[a]**	**O_3[a]**	**$PM_{2.5}$[b]**
Number of monitors	320	821	757
Mean concentrations	7.8	44.2	7.7
Std	4.9	5.2	2.4
Min	0.4	25.4	1.8
Q1	3.9	41.1	6.2
Median	7.1	44.5	8.0
Q3	10.9	47.5	9.2
Max	22.3	57.3	18.0

[a]Concentration unit is ppb;
[b]Concentration unit is μg/m^3.

5.3.2. LUR Model Performance by Variable Input

In general, LUR models using the new scenario of variables performed similarly as compared to those using the traditional scenario in both random and spatial CV. For example, for NO_2 in random CV among different modeling approaches, the gap of R^2 between the traditional scenario models and the new scenario models ranged from 0.02 to 0.09. However, when applying stepwise and stepwise-kriging, this pattern of performance varied: for O_3, models with the traditional variables performed much better than the new scenario (e.g., in random CV, stepwise R^2 [RMSE]: 0.56 [0.08] vs. 0.01 [0.12]; stepwise-kriging R^2 [RMSE]: 0.72 [0.06] vs. 0.36 [0.09]; Figure 5.1). This finding suggests that the use of new variable input may be sensitive to modeling approaches. Overall, using the alternative new variables in LUR models is feasible for predicting concentrations of criteria pollutants. Figure A5.2 shows the random and spatial CV results of CO, PM_{10}, and SO_2.

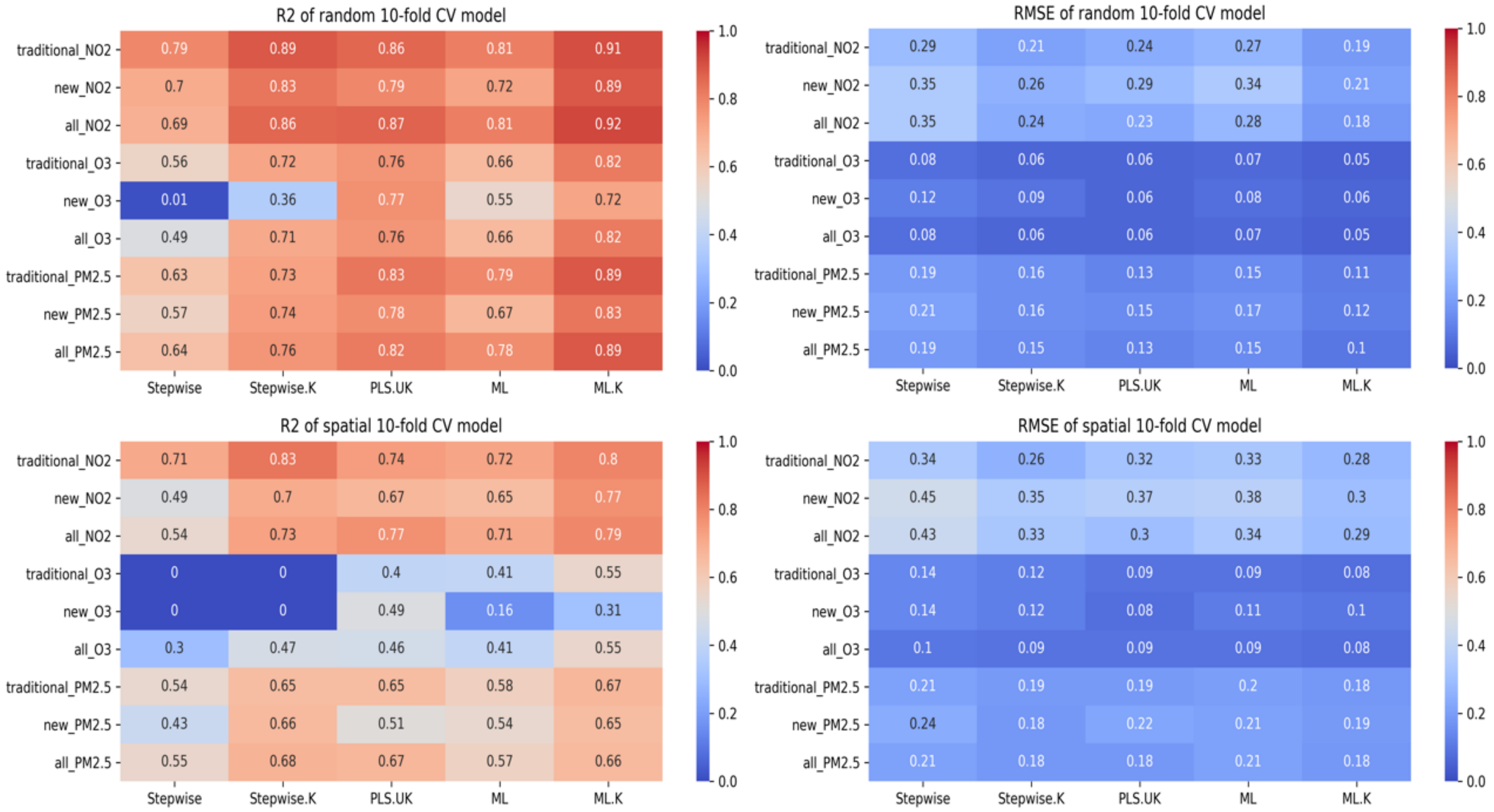


Figure 5.1. Random and spatial 10-fold CV results of major criteria pollutants (NO_2, O_3, and $PM_{2.5}$).

The LUR models using all variables showed similar performance as those with the traditional scenario in both random and spatial CV. For example, for most criteria pollutants (e.g., NO_2, $PM_{2.5}$, and SO_2) in random CV among different modeling approaches, the gap of R^2 between the traditional scenario models and the all scenario models was less than 0.10. Exceptions applied when the gap of R^2 became bigger in stepwise and stepwise-kriging models for O_3 and PM_{10}. For example, PM_{10} using all variables outperformed the traditional-only models in both random and spatial CV (e.g., in random CV, stepwise R^2 [RMSE]: 0.45 [0.37] vs. 0.26 [0.43]; stepwise-kriging R^2 [RMSE]: 0.71 [0.27] vs. 0.49 [0.35]; Figure A5.2). This finding suggests that in most cases, adding the new variables to the traditional variables in LUR models doesn't necessarily improve model performance; while in certain cases, incorporating the new variables may be helpful to capture spatial variability.

5.3.3. LUR Model Performance by Modeling Approach

Among the nine ML algorithms, gradient boosting and random forest generally showed the best performance in both random and spatial CV (Figure A5.3-A5.4). For example, for NO_2, gradient boosting performed slightly better than random forest using the three scenarios of variables (e.g., in random CV, gradient boosting vs. random forest: traditional R^2 [RMSE]: 0.79 [0.29] vs. 0.74 [0.32]; new R^2 [RMSE]: 0.78 [0.29] vs. 0.75 [0.31]; all R^2 [RMSE]: 0.80 [0.28] vs. 0.75 [0.31]; Figure A5.3), while other ML algorithms had worse performance (e.g., in random CV, traditional R^2 [RMSE]: 0.44-0.69 [0.18-0.24]; new R^2 [RMSE]: 0-0.70 [0.17-0.54]; all R^2 [RMSE]: 0.31-0.74 [0.15-0.25]). Since gradient boosting performed slightly better for the major criteria pollutants focused here (i.e., NO_2, O_3, and $PM_{2.5}$), I used gradient boosting algorithm to represent ML approach to compare to other LUR approaches.

Comparing among the five modeling approaches in LUR (i.e., stepwise, stepwise-kriging, PLS-UK, ML, and ML-kriging), models with kriging (i.e., stepwise-kriging, PLS-UK, and ML-kriging) generally performed well for all criteria pollutants. Specifically, ML-kriging showed a marginal improvement for both random and spatial CV (e.g., in random CV of NO_2 using all variables, ML-kriging R^2 [RMSE]: 0.92 [0.18]; PLS-UK R^2 [RMSE]: 0.87[0.23]; stepwise-kriging R^2 [RMSE]: 0.86 [0.24]). For models without kriging effect, ML generally outperformed the stepwise regression approach slightly in random and spatial CV for most variable scenarios; in some ways, ML really helped with specific pollutants (e.g., O_3, PM_{10}; Figure 5.1 and Figure A5.2). Another finding is that CO, PM_{10}, and SO_2 were more sensitive to modeling approaches. For example, for CO, the random CV R^2 could range from 0 to 0.80 depending on modeling approaches while for $PM_{2.5}$, the random CV R^2 showed smaller variability (0.57-0.89).

5.3.4. LUR Model Performance by Pollutant

The LUR models performed differently by pollutant. Generally, models for major criteria pollutants (i.e., NO_2, O_3, and $PM_{2.5}$) outperformed others (i.e., CO, PM_{10}, and SO_2). For NO_2, O_3, and $PM_{2.5}$, all models performed well except for O_3 using stepwise regression. For example, in random CV for O_3, the R^2 (RMSE) of stepwise regression with the new variables was 0.01 (0.12; Figure 5.1). Particularly, LUR models for NO_2 performed the best (e.g., random CV R^2: 0.69-0.92) among all pollutants whereas those for SO_2 had the worst performance (e.g., random CV R^2: 0-0.57).

5.3.5. LUR Model Performance by CV

In terms of CV evaluation, random CV consistently performed better than spatial CV indicating improved model performance when EPA monitors used for model development were in proximity. For example, when comparing the best-performing ML-kriging, the difference of R^2

(RMSE) in each CV for NO_2 was 0.11(0.09); while for O_3, the gap could be as large as R^2 (RMSE): 0.41(0.04). This finding suggests that the magnitude of impact of model evaluation method may vary by pollutant and model type.

5.3.6. Variable Importance

The traditional categories were identified as important variables contributing to the LUR models, even when adding the new variables. Among all three variable scenarios (i.e., traditional, new, and all), satellite data was mostly selected as a top 5 variable. The importance of satellite data was more prominent in models with the new variable scenario. This finding indicates that satellite variables were important predictors for air pollution levels.

For major criteria pollutants (i.e., NO_2, O_3, and $PM_{2.5}$), variables contributed differently among pollutants and modeling approaches in the new variable scenario. For example, of the top 20 variables selected in ML models, POI was an important contributor for NO_2 while LCZ contributed the most apart from the satellite data for O_3. Satellite, POI, and LCZ were all among the top 20 variables for the $PM_{2.5}$ model. All the three categories of new variables could also potentially contribute to LUR models. The top 20 variables selected by the stepwise regression and PLS-UK were similar, but slightly different from those in the ML models. For example, for O_3 of the new variable scenario, 10 out of 20 important variables were LCZ categories using ML approach. In contrast, 9 (13) out of 20 important variables were POI categories using stepwise (PLS-UK) approaches; GSV variables were also among the top 5 important predictors. Even in the all variable scenario, new variables were also identified as important contributors highlighting the potential to substitute some traditional variables.

While for the other three pollutants (i.e., CO, PM_{10}, and SO_2), GSV, POI and LCZ were important variable categories for ML, PLS-UK and stepwise regression respectively. For example for CO, 10 out of 20 important variables were GSV category using ML and 11 out of 20 variables were POI using PLS-UK. For PM_{10}, 6 out 12 variables were LCZ using stepwise regression approach. This finding suggests that the contribution of the new variables may be sensitive to modeling approaches. Figure 5.2 shows top 20 most important features of the ML models for NO_2, O_3, and $PM_{2.5}$ (other pollutants are shown in Figure A5.5). Variable importance of the stepwise regression and PLS-UK models are shown in Figure A5.6-A5.9.

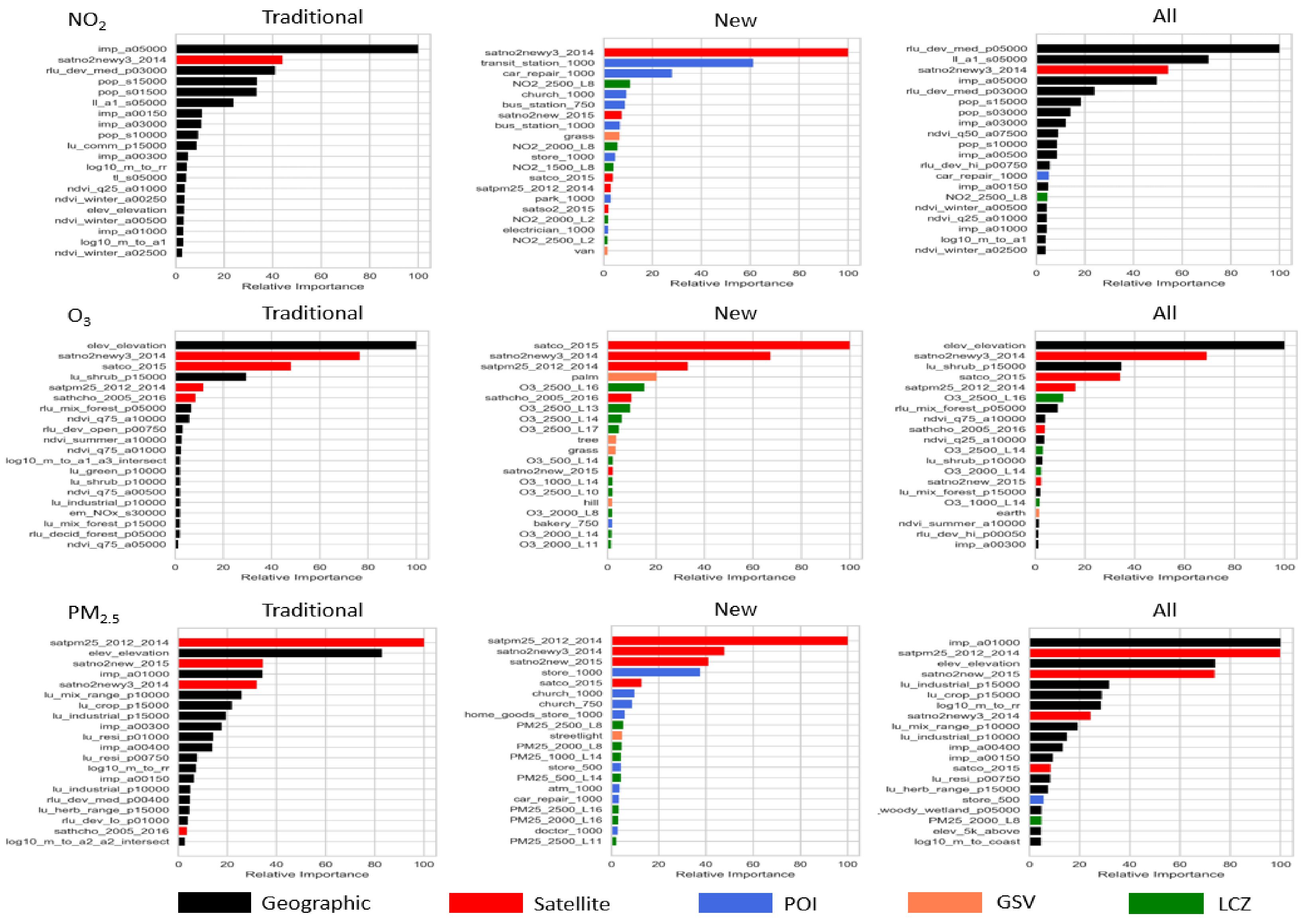


Figure 5.2. Top 20 most important features of the ML models for NO_2, O_3, and $PM_{2.5}$.

5.3.7. Implications for Developing LUR Models

5.3.7.1. Implications for using new variables for LUR models

I developed LUR models for six criteria pollutants in 2015 using five different modeling approaches based on the US EPA monitors. I investigated on how model performance varied using three different sets of variable scenarios: traditional vs. new vs. all. I used both random and spatial CV to evaluate the model performance.

One motivation of this study is to explore whether some new variables could replace or supplement traditional datasets. Current LUR models (particularly for large geographies) were often developed using variables that mainly from government-sponsored sources (e.g., US Census, NLCD) that may not updated in a timely fashion; some datasets were only available before 2010. The tabulation of hundreds of variables that involves data sources at different levels and across multiple jurisdictions may hinder the effort to develop LUR models at relatively high spatial resolutions (e.g., Census blocks) or beyond political boundaries. This limitation further impacts the subsequent research in environmental justice, exposure assessment, urban planning, and epidemiology. While there is ongoing effort in developing air quality models at national or global scales, the challenge of assembling and analyzing the huge magnitude of data deserves attention. Additionally, underdeveloped countries may suffer from data sources with limited spatial coverage and public accessibility to develop air quality models. Seeking for alternative new variables for developing LUR models may be helpful to resolve these issues.

I found that LUR models using the new scenario of variables generally demonstrated similar performance as compared to those with only the traditional variables (e.g., land use and geographic variables). This finding indicates that the new variables could be alternative choices for LUR models when preparation and processing of the traditional variables are cumbersome. The use of the new variables can also benefit regions with limited and no traditional data sources (e.g., developing countries). While replacing existing traditional variables with the new variables could be feasible for LUR development, adding both variables would not necessarily yield improved performance. The minimal change in model performance suggests that models developed with more variables may perform as well as those with adequate ones, which matches a recent study exploring the sensitivity of variable input (Kim et al., 2020). Another explanation may be that the new variable category consists of features that resemble the traditional variables (e.g., POI vs. land use), reassuring the potential application of using the alternative variables for LUR models.

Satellite-derived air pollution estimates could mostly be selected as top 5 significant variables among different pollutants regardless of modeling approaches indicating the importance of maintaining satellite variables for developing LUR models, particularly for models using only the new variable scenario. Such significant contribution has also been identified in other studies (Bechle et al., 2013; Knibbs et al., 2014; Novotny et al., 2011; Vienneau et al., 2013; Di et al., 2016; Kim et al., 2020). For example, Kim et al. (2020) found that satellite-derived estimates were almost prioritized for model contribution based on PLS-UK approach. In another ML-based (i.e., neural network) LUR study, satellite products (e.g., AOD) served as key components of the hybrid model for predicting $PM_{2.5}$ exposure (Di et al., 2016).

5.3.7.2. Implications for exploring variable importance of new variables

Aside from the traditional variables, variable importance of the new variables for LUR models varied by pollutant and modeling approach. For example, POI data was the second largest contributor (frequently selected for modeling) for NO_2 prediction in the ML models, including transit station and car repair. This finding is reasonable since NO_2 is mainly associated with traffic-related emissions (Bechle et al., 2015; de Hoogh et al., 2016; Hochadel et al., 2006). For O_3, LCZ was commonly important contributors for ML whereas for PLS-UK and stepwise regression, POI stood out among the top 20 important variables. This result may be explained by that O_3 is a secondary pollutant and highly correlated with urban heat and the emissions of NO_x and VOCs. LCZ could be an indicator to characterize urban heat (Petralli et al., 2014) and POI may feature impacts of local emission sources (Hassan Amini et al., 2014; Wu et al., 2017b). Though LCZ variables were not commonly identified as important variables in LUR models by each pollutant, it has the capacity to account for urban morphology and track land use and land cover over time. It may also contribute to develop LUR models for areas suffered from urban heat and street canyon effect. The widely recognized important new categories reveal great potential for alternating traditional land use and geographic dataset to uncover local factors in a uniform way.

Likewise, GSV variables commonly contributed to CO, PM_{10}, and SO_2 models using ML approach highlighting the potential application for predicting these pollutants. Some subset categories of GSV (e.g., building, sidewalk) may indicate traffic and infrastructure conditions and others (e.g., grass, palm) may characterize natural components, which has been successfully used to identify infrastructure and green space that may be associated with air quality (Larkin & Hystad, 2019; Rzotkiewicz et al., 2018). Some studies have already identified the importance such microscale predictors, including presence of awnings (Miskell et al., 2018; Weissert et al., 2018), tree canopy (Knibbs et al., 2014; Novotny et al., 2011a), and building (Madsen et al., 2011; Shi et al., 2016b; Su et al., 2008). These features could be characterized by the GSV variables. This finding further implies that future studies could use GSV-related variables to develop LUR models for areas where the traditional data sources are limited and pollutant species that are difficult to model. While new insights may be offered through these important new variables, the sensitivity of these new variables to different modeling approaches and pollutants cautions the use and interpretation for developing LUR models.

5.3.7.3. Implications for developing LUR models based on different modeling approaches

The difference in model performance of the various modeling approaches reflected that models integrated with kriging performed better than those without. This result is consistent with other similar studies that integrated models with kriging (Kim et al., 2020; H. Xu et al., 2019; Young et al., 2016). In addition, this improvement from kriging is greater than a study in US (Young et al., 2016), but similar to another study in China (H. Xu et al., 2019). Such benefit could be greater for models using the new variables (e.g., in random CV R^2 increase of $PM_{2.5}$, stepwise to stepwise-kriging: 0.17 [new] vs. 0.10 [traditional]; ML to ML-kriging: 0.16 [new] vs. 0.10 [traditional]). Among these models, ML-kriging performed best for most pollutants indicating that developing hybrid models may be one effective strategy to improve LUR models. For models without kriging effect, ML models offered modest improvement compared to stepwise regression approach in most cases; ML approach helped improve models of O_3 and PM_{10}. This finding is similar to another study comparing linear regression and ML approach within the LUR framework (Weichenthal et al., 2016). Another noticeable pattern was that using ML models could be more adaptable to the use of the new variables (as compared to stepwise regression) suggesting that when non-traditional variables are readily available, ML approach may be more appropriate for developing LUR models. More studies on the sensitivity of ML model performance to different variable input should be investigated.

As noted, model performance varied by pollutant. In general, models for major criteria pollutants (i.e., NO_2, O_3, and $PM_{2.5}$) were better than those for CO, SO_2, and PM_{10}, confirming the finding of a recent LUR model study using PLS-UK (Kim et al., 2020). In addition, some criteria pollutants were more sensitive to modeling approaches (e.g., O_3, CO, SO_2, and PM_{10}). This could be attributable to different factors including chemical features and physics, emission sources, spatial patterns, satellite data availability, and data quality. This finding further supports incorporating kriging into existing modeling effort. The model evaluation results indicate that random CV outperformed spatial CV for all pollutants. This finding means that regardless of modeling approach, poor model performance may be expected when there are few monitors in the vicinity. One should caution the use to predict air pollution concentrations where sparse network of monitors are available.

5.3.7.4. Limitations and future research

This study has several limitations. One limitation is that some of the new variables may be inaccurate. For example, I explored the use of Google products (i.e., POI, GSV) in LUR in the goal of alternating the traditional variables; however, the time of the extracted POI and GSV data may not well aligned with the local environment of 2015 (when my LUR models are developed for). While most of these new variables may be more updated (e.g., LCZ for 2016 surface; POI was retrieved in 2018), the development of more updated version of LUR models may help explore the impacts of such temporal mismatch. While inaccuracies of these new data sources may exist, some of these variables may be better than the traditional variables in capturing street-level features and urban morphology. Thus, the contribution of these alternative variables should not be neglected. To compare to a recent published study using PLS-UK (Kim et al., 2020), I used the same sets of traditional variables to develop LUR models. While some of the traditional variables may be limited in temporal mismatch (e.g., population data in 2000), future studies could use updated datasets of traditional variables for comparison. Another limitation relates to the modeling approach. I followed the literature (Mercer et al., 2011b) to conduct a two-step modeling approach to include kriging process (i.e., ML-kriging and stepwise-kriging) instead of developing a kriging framework similar to PLS-UK. Future model improvement could be achieved by integrating the ML and stepwise regression into a kriging framework, thus testing the impacts on model performance. The different variable categories selected by various modeling approaches for the same pollutant may reveal intertwining relationships between variables. Future effort can explore modeling approaches that better account for multicollinearity and complexity (e.g., neural network; Di et al., 2016).

There are some implications for future research. Developing empirical models of large geographies is an important goal for tracking air pollution exposure for large populations. However, most large-scale air quality models may be limited in spatial resolution and variable consistency. My study of using new variables capable of capturing street-level features and providing relatively uniform method may offer some insights for environmental scientists and urban planners. I used Google POI and GSV to explore the feasibility of developing LUR models. The reasonable model performance of using these alternative variables implicates that future LUR models could use variables of these types when traditional land use and geographic variables are not readily available. Although Google POI and GSV are not available in some countries (e.g., China), similar products may be used (e.g., Baidu POI, Gaode Map) for LUR development. Such alternative variables further increase the possibility of developing multi-country LUR models – an understudied topic. I only used satellite, Google POI, GSV, and LCZ to serve as new variables; other potential new variables could also be incorporated. For example, Yelp open dataset provides institutional, retail, and entertainment place-based data that may capture street-level features for air quality. One important application of these new variables is that they enable characterizing new information on street-level features that may be associated with air quality. Further studies on the use of such variables may inform urban planning strategies and landscape designs for healthy cities. In addition, traditional cross-year LUR models often used land use data without accounting for temporal changes. One advantage of using LCZ dataset is that it has the ability to track land use and land cover information over decades, which could be used to develop LUR model over years. Lastly, existing national LUR models were developed based on the EPA monitoring network, which was originally launched for regulation compliance covering mostly urban areas. Future national models could be developed integrating air quality monitors with improved coverage and density (e.g., low-cost sensors).

5.4. Conclusion

In summary, this study reveals important findings on feasibility of using alternative new variables with improved consistency, availability, and generalizability for LUR model development. Results indicate that models using the alternative variables demonstrated similar performance as compared to those with the traditional variables. Additionally, models adding kriging effect could improve model performance and ML-based approach may be more applicable for non-traditional data sources. My approach suggests that LUR models for predicting air pollution concentrations developed using emerging modeling approaches and new variables may be promising for areas or countries where traditional land use and geographic information are limited and unavailable.

Chapter 6. CONCLUSIONS

The dissertation describes empirical evidence on improving LUR models for predicting air pollution concentrations using new air quality data, alternative predictor variables, and emerging modeling approaches. In this Chapter I summarize the core conclusions of this dissertation. My conclusions include key findings, limitations, and implications for future research.

6.1. Key Findings

With the rapid development in crowdsourcing, sensor technologies, data sciences, and modeling approaches, there are opportunities to improve existing LUR models for refined exposure assessment. This dissertation aims at exploring the impacts of new data sources and emerging modeling methods on LUR model performance. I tested the feasibility of these strategies for cities/regions at different levels (i.e., a single city, multiple cities, and the entire country). I found that LUR models could be improved through integrating new data sources of air quality, community-based low-cost monitoring, and non-government sponsored crowd-sourced platforms into the modeling process. Below I discuss the key findings of these studies.

I found that the area source-related features was an important factor for predicting VOC concentrations. For example, in the BTEX model, models with area sources (e.g., dry cleaners, gas stations; city permit data: adj-R^2: 0.37; RMSE: 0.37 μg/m^3) outperformed the base-case model that didn't include these sources (adj-R^2: 0.15; RMSE: 0.43 μg/m^3). Further, in the TVOC models (aggregating all 60 VOC species), this pattern was similar to the BTEX models: the city permit model (adj-R^2: 0.42; RMSE: 0.37 μg/m^3) performed better than the base-case model (adj-R^2: 0.26; RMSE: 0.41 μg/m^3). In addition, I found that area sources may be as important as traditional transportation and land use variables in LUR models. These findings suggest that predictions of VOCs with comparatively different emission sources from criteria pollutants (e.g., NO_2) could benefit from accounting for area sources in LUR models.

For models with variables of area sources, LUR models using Google POI data outperformed those with city permit data among all 60 VOCs. For example, the BTEX models with Google POI data (adj-R^2: 0.47; RMSE: 0.34 μg/m^3) showed better performance as compared to models with city permit data (0.37; 0.37). Likewise, for TVOC models, the model performance improved when including Google POI data: Google POI (adj-R^2: 0.56; RMSE: 0.32 μg/m^3) vs. city permit data (adj-R^2: 0.42; RMSE: 0.37 μg/m^3). This comparison between models with the city permit data and the Google POI data indicates that non-government data sources may be helpful for improving performance of LUR models.

For models adding area sources, variables including dry cleaners and gas stations were consistently identified as important predictors for LUR models of VOCs. For example, more than 45 out 60 VOC species models selected these small-scale emission sources. One unique aspect of my work is that I was able to compare variable importance across VOC species. For tetrachloroethene using the Google POI model, the magnitude of association was slightly larger for both laundry and gas stations (0.41) as compared to traditional (0.34) and land use variables (0.31). These results highlight the importance of area sources and necessitate of their integration into LUR models for VOCs.

Additionally, my work indicates that air pollution data form community-based sampling could be feasible to develop LUR models for pollutants that were not commonly monitored (i.e., VOC). I was able to model 60 VOC species to identify various spatial variability of each VOC (group) using data from this community-based effort in a single city (i.e., City of Minneapolis) in the US. This practice implies that community-based monitoring could be useful data sources for developing air quality models. My approach of combining the community-driven effort with new open data sources may enhance LUR models (as emphasized in Chapter 3).

In Chapter 4, I first tested the contribution of a non-government air quality data source to LUR models using a crowd-sourced low-cost sensor platform (i.e., PurpleAir) from multiple US cities rather than single cities. I found that the PPA data could be reasonably used to develop LUR models. For example, the internal evaluation result (R^2: 0.66) indicates that LUR models based on the PPA data may be capable of capturing spatial variability of $PM_{2.5}$. In addition, combining both the PPA data and EPA data could improve LUR performance. For example, if including the national EPA data, the Hybrid PPA model performed as well as the CACES LUR (R^2: 0.85 vs. 0.83); if using only the 6-city EPA data, the Hybrid PPA model performed better than the CACES LUR (R^2: 0.77 vs. 0.67). Air quality data from the crowd-sourced low-cost sensor network could be as useful as the regulatory networks for LUR models, particularly at locations where regulatory monitors are not available.

Improving the spatial resolution and coverage for LUR models is significant for exposure assessment. Inclusion of the low-cost PPA measurements could potentially account for locations with high $PM_{2.5}$ concentrations, thus improving the capability of capturing within-city spatial variability. For example, the population-weighted PPA LUR, CACES LUR, and Hybrid LUR were well correlated indicating the contribution of the low-cost sensor data. Such contribution could be expanded when filling the gaps of sparse regulatory monitors and integrating into regulatory-based LUR models. Another benefit of the low-cost sensors is that they may be capable of capturing more within-city variability, including near-source areas (e.g., industrial facilities, road segments). My work could be helpful to identify such places with prediction disagreement.

I found that the CACES LUR that was developed based on regulatory EPA monitors may fail to capture locations with higher concentrations. The PPA data may help fill these gaps. For example, I found that the evaluating the CACES LUR using the PPA data showed modest performance (e.g., $R^2 = 0.41$, MAE = 5.5 μg/m^3). This finding suggests that the PPA users may tend to deploy sensors at areas typically reporting high concentrations, which may pick up locations not predicted by the CACES LUR, thus identifying the complex spatial heterogeneity of within-city air pollution. Finally, the crow-sourced low-cost sensor network could benefit from careful quality control and more side-by-side regulatory-grade calibration to warrant better application in LUR models.

Aside from improving LUR models via integrating data from low-cost sensors, searching for enhanced predictor variables based on emerging modeling approaches could be another strategy (Chapter 5). I compared national LUR models for criteria pollutants (e.g., NO_2) using different sets of variables in the goal of exploring alternative variables from non-government sponsored sources that could improve traditional land use and geographic variables in terms of data availability, spatial resolution, and model generalizability.

I found that the use of the new variables (i.e., satellite data, Google POI, GSV, and LCZ) for LUR models was as reasonable as using the traditional variables. For example, for NO_2 in random CV, the difference of R^2 between the models with the two type of variable scenarios (new variable vs. traditional variable scenario) ranged only from 0.02 to 0.09. This result indicates that the new variable sources may be alternative choices when the traditional variables were not easily obtained. This strategy could become more valuable for underdeveloped regions where traditional variables were not readily available.

I also found that combining both the traditional and new variables (the “all” variable scenario) may not yield improved LUR performance. For example, for most criteria pollutants (e.g., NO_2, $PM_{2.5}$, and SO_2) in random CV, such improvement of R^2 could be minimal (e.g., 0.10 in maximum). This result further highlights that using new variables for LUR models may also serve as parsimonious yet reliable option.

While the new variables are promising, each of the variable categories may offer insights to urban planners and environmental scientists. For example, satellite-derived estimates were frequently included in LUR models for most criteria pollutants, which suggests that maintaining satellite data may help enhance LUR performance. This finding was particularly useful for models that only use the new variables. The contribution of other new data categories revealed the typical characteristics for some pollutants, including (1) the importance of traffic-related POI data for NO_2 concentrations, (2) the significance of LCZ that characterize urban heat and street canyon effect for O_3 concentrations, and (3) the contribution of GSV variables to feature street-level conditions (e.g., sidewalk, awning) for pollutants (e.g., CO, PM_{10}, and SO_2) that were less frequently modeled.

In terms of modeling approaches used for LUR models, I found that models integrated with kriging could perform better than those without. Such improvement could be more obvious for models using the new variables. For models without kriging, ML may be more plausible for places where the new variables were more accessible. Due to different characteristics of each pollutant (e.g., emission sources, data quality), some pollutants may be more sensitive to modeling approaches. As such, adding kriging into existing modeling effort may be an effective strategy to improve LUR models.

6.2. Limitations and Potential Future Research

My dissertation highlights the potential of using community-based effort, new data sources, and emerging modeling approaches for improving LUR models. However, there are some limitations that could be addressed in future research. I briefly discuss some of the major limitations and implications for future research.

One limitation of my work is the community-based air quality monitoring effort. For example, in Chapter 3 for a local practice, I used data from a community-based sampling with consecutive 8 seasons of 60 VOC measurements. However, the annual average concentrations were calculated based on the 72-hour data of each season, which was not consecutive measurements for the entire year. While in Chapter 4 for a national study, I extracted $PM_{2.5}$ concentrations at low-cost sensor sites (i.e., PPA) with at least 244 days out of a year to calculate the annual averages, but the detailed information on monitoring purpose, location characteristics, and deployment strategies were unknown. These limitations should caution the use and interpretation of LUR model results for long-term estimates. One should be aware of the trade-off in several aspects: data availability vs. data accuracy, site coverage vs. targeted sampling, and monitoring duration vs. temporal representativeness. Future work could be improved in these community-based effort including (1) merging the goal of air quality modeling with community-based monitoring and (2) using data from other available low-cost sensor platform or refining the models once more details regarding to sensor locations are available.

Another limitation is from the predictor variables. For example, in Chapter 4 and 5 for model evaluation and comparison, I used the same set of traditional variables served in the CACES LUR models (i.e., PLS-UK). These variables were not well aligned with the target time of the LUR models (i.e., 2015). For example, some of the land use and geographic variables were before 2010 (e.g., population data). Such temporal mismatch issue also appeared in non-government data sources. For example, in Chapter 3 and 5, Google POI data was retrieved mainly in 2018 while the LUR models were developed for 2015. Similar issue also exists in the GSV (2018) and LCZ (2016) data. Developing newer version of LUR models with more updated data sources could further explore the impact on LUR performance. In addition, these data sources may be biased from the user generated and verified perspective. Some businesses without online presence and some rural areas without GSV imagery may suffer from data unavailability.

A novel aspect of my dissertation is the alternative data sources from non-government sponsored open platforms. The new data I used for these studies mainly include two parts: air quality data (served as dependent variables in LUR models) and predictor variables (served as independent variables in LUR models). In terms of air quality data, many monitoring campaigns of air quality are ongoing across communities in the US. Future LUR models could be developed using these community-based data, particularly when regulatory air monitors are not available or pollutant of interest is not regularly monitored at all. One implication of my work (Chapter 3) is to develop first-of-its-kind regional or national LUR models using available VOC data collected from different cities across the US would be interesting. Similarly based on effort in Chapter 4, developing the first national PPA LUR models for $PM_{2.5}$ using all the PurpleAir sensor data in the US may be another direction to explore the contribution of low-cost sensor data from non-government open platforms. These air quality monitoring and modeling products may help inform monitoring gaps and readjust existing regulatory monitoring network, thus targeting pollutants and areas with more challenges and concerns.

For the future of non-traditional predictor datasets, they have the advantage of capturing street-level features that may be important for predicting air pollution. In Chapter 3, I first identified the contribution of Google POI data to LUR models using only four categories (i.e., laundry, painter, car repair, gas station). Further in Chapter 5, I used all 90 categories to explore the application for criteria pollutants and found they were feasible variables to alternate traditional predictors. Future LUR models could be developed to generalize across regions or countries, particularly for underdeveloped areas where traditional variables are not rich and accessible. Developing LUR models using these types of uniform data could allow for multi-city and multi-country comparisons. For example, other alternative data sources may be also tested and explored (e.g., Baidu POI, Yelp open dataset) for LUR models in different countries. The use of these new variables could further inform urban planning and landscape design strategies for developing healthy cities. One example would be using LCZ dataset as a proxy of urban form to explore how urban form impacts air quality and identify sustainable and resilient urban development patterns for clean and healthy communities. Finally, my dissertation could be valuable for those interested in using crowdsourcing, big/open data, data analytics, and emerging modeling approaches to develop generalizable air quality models in large scales.

REFERENCE

Abbey, D. E., Petersen, F., Mills, P. K., & Beeson, W. L. (1993). Long-term ambient concentrations of total suspended particulates, ozone, and sulfur dioxide and respiratory symptoms in a nonsmoking population. *Archives of Environmental Health: An International Journal*, *48*(1), 33–46.

Adam-Poupart, A., Brand, A., Fournier, M., Jerrett, M., & Smargiassi, A. (2014). Spatiotemporal Modeling of Ozone Levels in Quebec (Canada): A Comparison of Kriging , Land-Use Regression (LUR), and Combined. *Environmental Health Perspectives*, *970*(9), 970–976.

Aguilera, I., Sunyer, J., Fernandez-Patier, R., Aguirre-Alfaro, A., Meliefste, K., Bomboi-Mingarro, M. T., … Brunekreef, B. (2008). Estimation of outdoor NOx, NO2, and BTEX exposure in a cohort of pregnant women using land use regression modeling. *Environmental Science & Technology*, *42*(3), 815–21. http://doi.org/10.1021/es0715492

Ahangar, F. E., Freedman, F. R., & Venkatram, A. (2019). Using low-cost air quality sensor networks to improve the spatial and temporal resolution of concentration maps. *International Journal of Environmental Research and Public Health*, *16*(7), 1252. http://doi.org/10.3390/ijerph16071252

Air Quality Egg. (2020). The Egg. Retrieved from https://airqualityegg.com/egg [18 March 2020]

AirCasting. (2020). AirCasting is an open-source environmental data visualization platform that consists of an Android app and online mapping system. Retrieved from https://www.habitatmap.org/aircasting [18 March 2020]

AirVisual. (2020). Explore the air quality anywhere in the world.

Alexander, P. J., Mills, G., & Fealy, R. (2015). Using LCZ data to run an urban energy balance model. *Urban Climate*, *13*, 14–37. http://doi.org/10.1016/j.uclim.2015.05.001

Amini, H., Hosseini, V., Schindler, C., Hassankhany, H., Yunesian, M., Henderson, S. B., & Künzli, N. (2017). Spatiotemporal description of BTEX volatile organic compounds in a Middle Eastern megacity: Tehran Study of Exposure Prediction for Environmental Health Research (Tehran SEPEHR) *. *Environmental Pollution*, *226*, 219–229. http://doi.org/10.1016/j.envpol.2017.04.027

Amini, H., Schindler, C., Hosseini, V., & Yunesian, M. (2017). Land Use Regression Models for Alkylbenzenes in a Middle Eastern Megacity: Tehran Study of Exposure Prediction for Environmental Health Research (Tehran SEPEHR). *Environmental Science & Technology*, *51*, 8481–8490. http://doi.org/10.1021/acs.est.7b02238

Amini, H., Taghavi-shahri, S. M., Henderson, S. B., Nadda, K., Nabizadeh, R., & Yunesian, M. (2014). Land use regression models to estimate the annual and seasonal spatial variability of sulfur dioxide and particulate matter in Tehran, Iran. *Science of the Total Environment*, *489*, 343–353. http://doi.org/10.1016/j.scitotenv.2014.04.106

Amini, H., Yunesian, M., Hosseini, V., Schindler, C., Sarah, B., & Künzli, N. (2017). A systematic review of land use regression models for volatile organic compounds.

Atmospheric Environment, *171*, 1–16. http://doi.org/10.1016/j.atmosenv.2017.10.010

Anselin, L. (1995). Local indicators of spatial association — LISA. *Geographical Analysis*, *27*(2), 93–115. http://doi.org/10.1111/j.1538-4632.1995.tb00338.x

APHA (American Public Health Association). (2010). *Promoting active transportation: An opportunity for public health*. Retrieved from https://www.albany.edu/ihi/files/APHA_Promoting_Active_Transportation_Report.pdf [7 March 2020]

Atari, D. O., & Luginaah, I. N. (2009). Assessing the distribution of volatile organic compounds using land use regression in Sarnia, "Chemical Valley", Ontario, Canada. *Environmental Health*, *14*, 1–14. http://doi.org/10.1186/1476-069X-8-16

Baldasano, J. M., Delgado, R., & Calbo, J. (1998). Applying Receptor Models To Analyze Urban / Suburban VOCs Air Quality in Martorell (Spain). *Environmental Science & Technology*, *32*(3), 405–412. http://doi.org/10.1021/es970008h

Ballester, F., Estarlich, M., Iñiguez, C., Llop, S., Ramón, R., Esplugues, A., … Rebagliato, M. (2010). Air pollution exposure during pregnancy and reduced birth size: a prospective birth cohort study in Valencia, Spain. *Environmental Health*, *9*(1), 6.

Barzyk, T., Williams, R., Kaufman, A., & Greenberg, M. (2016). *Citizen science air monitoring in the Ironbound community*. Washington DC. Retrieved from https://cfpub.epa.gov/si/si_public_record_report.cfm?Lab=NERL&dirEntryId=315154 [18 March 2020]

Bechle, M. J., Millet, D. B., & Marshall, J. D. (2013). Remote sensing of exposure to NO2: Satellite versus ground-based measurement in a large urban area. *Atmospheric Environment*, *69*(2), 345–353. http://doi.org/10.1016/j.atmosenv.2012.11.046

Bechle, M. J., Millet, D. B., & Marshall, J. D. (2015). National Spatiotemporal Exposure Surface for NO2: Monthly Scaling of a Satellite-Derived Land-Use Regression, 2000-2010. *Environmental Science and Technology*, *49*(20), 12297–12305. http://doi.org/10.1021/acs.est.5b02882

Bechle, M. J., Millet, D. B., & Marshall, J. D. (2017). Does urban form affect urban NO2? Satellite-based evidence for more than 1200 cities. *Environmental Science and Technology*, *51*(21), 12707–12716. http://doi.org/10.1021/acs.est.7b01194

Bechtel, B., Alexander, P. J., Böhner, J., Ching, J., & Conrad, O. (2015a). Mapping Local Climate Zones for a worldwide database of the form and function of cities. *ISPRS International Journal of Geo-Information*, *4*, 199–219. http://doi.org/10.3390/ijgi4010199

Bechtel, B., Alexander, P. J., Böhner, J., Ching, J., & Conrad, O. (2015b). Mapping Local Climate Zones for a Worldwide Database of the Form and Function of Cities. *ISPRS International Journal of Geo-Information*, *4*, 199–219. http://doi.org/10.3390/ijgi4010199

Beckerman, B. S., Jerrett, M., Martin, R. V, van Donkelaar, A., Ross, Z., & Burnett, R. T. (2013). Application of the deletion/substitution/addition algorithm to selecting land use regression models for interpolating air pollution measurements in California. *Atmospheric Environment*, *77*, 172–177. http://doi.org/10.1016/j.atmosenv.2013.04.024

Beelen, R., Hoek, G., Pebesma, E., Vienneau, D., de Hoogh, K., & Briggs, D. J. (2009). Mapping of background air pollution at a fine spatial scale across the European Union. *Science of the Total Environment*, *407*(6), 1852–1867. http://doi.org/10.1016/j.scitotenv.2008.11.048

Beelen, R., Hoek, G., Vienneau, D., Eeftens, M., Dimakopoulou, K., Pedeli, X., … Hoogh, K. De. (2013). Development of NO2 and NOx land use regression models for estimating air pollution exposure in 36 study areas in Europe-The ESCAPE project. *Atmospheric Environment*, *72*, 10–23. http://doi.org/10.1016/j.atmosenv.2013.02.037

Beelen, R., Raaschou-Nielsen, O., Stafoggia, M., Andersen, Z. J., Weinmayr, G., Hoffmann, B., … others. (2014). Effects of long-term exposure to air pollution on natural-cause mortality: an analysis of 22 European cohorts within the multicentre ESCAPE project. *The Lancet*, *383*(9919), 785–795.

Beelen, R., Voogt, M., Duyzer, J., Zandveld, P., & Hoek, G. (2010). Comparison of the performances of land use regression modelling and dispersion modelling in estimating small-scale variations in long-term air pollution concentrations in a Dutch urban area. *Atmospheric Environment*, *44*(36), 4614–4621. http://doi.org/10.1016/j.atmosenv.2010.08.005

Bellander, T., Berglind, N., Gustavsson, P., Jonson, T., Nyberg, F., Pershagen, G., & Järup, L. (2001). Exposure to Air Pollution from Traffic and House Heating in Stockholm. *Environmental Health Perspectives*, *109*(6), 633–639.

Bellinger, C., Jabbar, M. S. M., Zaiane, O., & Osornio-Vargas, A. (2017). A systematic review of data mining and machine learning for air pollution epidemiology. *BMC Public Health*, *17*(1), 907.

Bi, J., Stowell, J., Seto, E. Y. W., English, P. B., Al-hamdan, M. Z., Kinney, P. L., … Liu, Y. (2020a). Contribution of low-cost sensor measurements to the prediction of PM2.5 levels: A case study in Imperial County, California, USA. *Environmental Research*, *180*(August 2019), 108810. http://doi.org/10.1016/j.envres.2019.108810

Bi, J., Stowell, J., Seto, E. Y. W., English, P. B., Al-hamdan, M. Z., Kinney, P. L., … Liu, Y. (2020b). Contribution of low-cost sensor measurements to the prediction of PM2.5 levels: A case study in Imperial County, California, USA. *Environmental Research*, *180*, 108810. http://doi.org/10.1016/j.envres.2019.108810

Bi, J., Wildani, A., Chang, H. H., & Liu, Y. (2020). Incorporating low-cost sensor measurements into high-resolution PM2.5 modeling at a large spatial scale. *Environmental Science & Technology*, *54*, 2152–2162. http://doi.org/10.1021/acs.est.9b06046

Boersma, K. F., Eskes, H. J., Dirksen, R. J., Veefkind, J. P., Stammes, P., & Huijnen, V. (2011). An improved tropospheric NO 2 column retrieval algorithm for the Ozone Monitoring Instrument. *Atmospheric Chemistry and Physics*, *4*(2), 1905–1928. http://doi.org/10.5194/amt-4-1905-2011

Borghi, F., Spinazz, A., Rovelli, S., Campagnolo, D., Buono, L. Del, Cattaneo, A., & Cavallo, D. M. (2017). Miniaturized monitors for assessment of exposure to air pollutants: A review. *International Journal of Environmental Research and Public Health*, *14*, 909.

http://doi.org/10.3390/ijerph14080909

Borrego, C., Costa, A. M., Ginja, J., Amorim, M., Coutinho, M., Karatzas, K., … others. (2016). Assessment of air quality microsensors versus reference methods: The EuNetAir joint exercise. *Atmospheric Environment*, *147*, 246–263.

Brauer, M., Hoek, G., Vliet, P. Van, Meliefste, K., Fischer, P., Gehring, U., … Brunekreef, B. (2003). Estimating Long-Term Average Particulate Air Pollution Concentrations: Application of Traffic Indicators and Geographic Information Systems. *Epidemiology*, *14*(2), 228–239.

Brauer, M., & Lee, M. (2018). *Evaluation of portable air quality sensors at the Vancouver (Clark Drive) near-road air quality monitoring site*. Retrieved from https://open.library.ubc.ca/cIRcle/collections/facultyresearchandpublications/52383/items/1.0380965

Broday, D. M., & the Citi-Sense Project Collaborators. (2017). Wireless distributed environmental sensor networks for air pollution measurement—The promise and the current reality. *Sensors*, *17*, 2263. http://doi.org/10.3390/s17102263

Brook, R. D., Franklin, B., Cascio, W., Hong, Y., Howard, G., Lipsett, M., … others. (2004). Air pollution and cardiovascular disease: a statement for healthcare professionals from the Expert Panel on Population and Prevention Science of the American Heart Association. *Circulation*, *109*(21), 2655–2671.

Brousse, O., Martilli, A., Foley, M., Mills, G., & Bechtel, B. (2016). Urban Climate WUDAPT, an efficient land use producing data tool for mesoscale models? Integration of urban LCZ in WRF over Madrid. *UCLIM*, *17*, 116–134. http://doi.org/10.1016/j.uclim.2016.04.001

Brown, S. G., Frankel, A., & Hafner, H. R. (2007). Source apportionment of VOCs in the Los Angeles area using positive matrix factorization. *Atmospheric Environment*, *41*(2), 227–237. http://doi.org/10.1016/j.atmosenv.2006.08.021

Cakmak, S., Hebbern, C., Pinault, L., Lavigne, E., Vanos, J., Crouse, D. L., & Tjepkema, M. (2018). Associations between long-term PM2.5 and ozone exposure and mortality in the Canadian Census Health and Environment Cohort (CANCHEC), by spatial synoptic classification zone. *Environment International*, *111*, 200–211.

CARB (California Air Resources Board). (2020). Community Air Protection Program Resource Center Land Use Resources. Retrieved from https://ww2.arb.ca.gov/our-work/programs/community-air-protection-program-resource-center/strategy-development/land-use [7 March 2020]

Carr, D., Ehrenstein, O. Von, Weiland, S., Wagner, C., Wellie, O., Nicolai, T., & Mutius, E. Von. (2002). Modeling annual benzene, toluene, NO2, and soot concentrations on the basis of road traffic characteristics. *Environmental Research Section A*, *118*(2), 111–118. http://doi.org/10.1006/enrs.2002.4393

Carvlin, G. N., Lugo, H., Olmedo, L., Bejarano, E., Wilkie, A., Meltzer, D., … Seto, E. (2019). Use of citizen science-derived data for spatial and temporal modeling of particulate matter near the US/Mexico Border. *Atmosphere*, *10*(9), 495.

CDC (Centers for Disease Control and Prevention). (2012). *CDC's healthy communities program: Program overview*. Retrieved from https://www.cdc.gov/nccdphp/dch/programs/healthycommunitiesprogram/index.htm [7 March 2020]

Cesaroni, G., Porta, D., Badaloni, C., Stafoggia, M., Eeftens, M., Meliefste, K., & Forastiere, F. (2012). Nitrogen dioxide levels estimated from land use regression models several years apart and association with mortality in a large cohort study. *Environmental Health*, *11*(1), 48.

Chance, K. (2007). OMI/Aura Formaldehyde (HCHO) Total Column 1-orbit L2 Swath 13x24 km V003, Greenbelt, MD, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC). Retrieved from 10.5067/Aura/OMI/DATA2015 [27 April 2017]

Chang, S. J., Chen, C. J., Lien, C. H., & Sung, F. C. (2006). Hearing loss in workers exposed to toluene and noise. *Environmental Health Perspectives*, *114*(8), 1283–1286. http://doi.org/10.1289/ehp.8959

Chen, W. H., Chen, Z. Bin, Yuan, C. S., Hung, C. H., & Ning, S. K. (2016). Investigating the differences between receptor and dispersion modeling for concentration prediction and health risk assessment of volatile organic compounds from petrochemical industrial complexes. *Journal of Environmental Management*, *166*, 440–449. http://doi.org/10.1016/j.jenvman.2015.10.050

Chen, X., Wang, X., Huang, J., Zhang, L., Song, F., Mao, H., … others. (2017). Nonmalignant respiratory mortality and long-term exposure to PM10 and SO2: A 12-year cohort study in northern China. *Environmental Pollution, 231*, 761e767.

Ching, J. K. S. (2013). A perspective on urban canopy layer modeling for weather, climate and air quality applications. *Urban Climate, 3*, 13–39.

Clark, L. P., Millet, D. B., & Marshall, J. D. (2011). Air Quality and Urban Form in U.S. Urban Areas: Evidence from Regulatory Monitors. *Environmental Science and Technology*, *45*, 7028–7035. http://doi.org/10.1021/es2006786

Clark, L. P., Millet, D. B., & Marshall, J. D. (2014). National Patterns in Environmental Injustice and Inequality: Outdoor NO2 Air Pollution in the United States. *PLoS ONE*, *9*(4), e94431. http://doi.org/10.1371/journal.pone.0094431

Clements, A. L., Griswold, W. G., Rs, A., Johnston, J. E., Herting, M. M., Thorson, J., … Hannigan, M. (2017). Low-cost air quality monitoring tools: From research to practice (a workshop summary). *Sensors*, *17*(11), 2478. http://doi.org/10.3390/s17112478

Cohen, M. A., Adar, S. D., Allen, R. W., Avol, E., Curl, C. L., Gould, T., … others. (2009). Approach to estimating participant pollutant exposures in the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air). *Environmental Science & Technology*, *43*(13), 4687–4693.

Commodore, A., Wilson, S., Muhammad, O., Svendsen, E., & Pearce, J. (2017). Community-based participatory research for the study of air pollution: a review of motivations, approaches, and outcomes. *Environmental Monitoring Assessment*, *189*(8), 378.

http://doi.org/10.1007/s10661-017-6063-7

Conrad, C. C., & Hilchey, K. G. (2011). A review of citizen science and community-based environmental monitoring: Issues and opportunities. *Environmental Monitoring and Assessment*, *176*(1–4), 273–291. http://doi.org/10.1007/s10661-010-1582-5

Crutcher, M., & Zook, M. (2009). Geoforum Placemarks and waterlines: Racialized cyberscapes in post-Katrina Google Earth. *Geoforum*, *40*(4), 523–534. http://doi.org/10.1016/j.geoforum.2009.01.003

Daniels, M. J., Dominici, F., Samet, J. M., & Zeger, S. L. (2000). Estimating particulate matter-mortality dose-response curves and threshold levels: an analysis of daily time-series for the 20 largest US cities. *American Journal of Epidemiology*, *152*(5), 397–406.

de Hoogh, K., Gulliver, J., Donkelaar, A. van, Martin, R. V., Marshall, J. D., Bechle, M. J., … Hoek, G. (2016). Development of West-European PM2.5 and NO2 land use regression models incorporating satellite-derived and chemical transport modelling data. *Environmental Research*, *151*(2), 1–10. http://doi.org/10.1016/j.envres.2016.07.005

de Hoogh, K., Gulliver, J., van Donkelaar, A., Martin, R. V, Marshall, J. D., Bechle, M. J., … Hoek, G. (2016). Development of West-European PM2.5 and NO2 land use regression models incorporating satellite-derived and chemical transport modelling data. *Environmental Research*, *151*(2), 1–10. http://doi.org/10.1016/j.envres.2016.07.005

Deeter, M. N., Edwards, D. P., Francis, G. L., Gille, J. C., Martínez-alonso, S., & Worden, H. M. (2017). A climate-scale satellite record for carbon monoxide: the MOPITT Version 7 product. *Atmospheric Chemistry and Physics*, *10*, 2533–2555.

Demerjian, K. L. (2000). A review of national monitoring networks in North America. *Atmospheric Environment*, *34*, 1861–1884.

Demuzere, M., Hankey, S., Mills, G., Zhang, W., Lu, T., & Bechtel, B. (2020). Combining expert and crowd-sourced training data to map urban form and functions for the continental US. *In Review*.

Di, Q., Kloog, I., Koutrakis, P., Lyapustin, A., Wang, Y., & Schwartz, J. (2016). Assessing PM2.5 exposures with high spatiotemporal resolution across the continental United States. *Environmental Science & Technology*, *50*(9), 4712–4721. http://doi.org/10.1021/acs.est.5b06121

Di, Q., Koutrakis, P., & Schwartz, J. (2016). A hybrid prediction model for PM2.5 mass and components using a chemical transport model and land use regression. *Atmospheric Environment*, *131*, 390–399. http://doi.org/10.1016/j.atmosenv.2016.02.002

Di, Q., Wang, Y., Zanobetti, A., Wang, Y., Koutrakis, P., Choirat, C., … Schwartz, J. D. (2017). Air pollution and mortality in the Medicare population. *New England Journal of Medicine*, *376*(26), 2513–2522.

Dijkema, M. B., Gehring, U., Strien, R. T. Van, Zee, S. C. Van Der, Fischer, P., & Hoek, G. (2011). A comparison of different approaches to estimate small-scale spatial variation in outdoor NO2 concentrations. *Environmental Health Perspectives*, *670*(2), 670–675. http://doi.org/10.1289/ehp.0901818

Dionisio, K. L., Rooney, M. S., Arku, R. E., Friedman, A. B., & Hughes, A. F. (2010). Within-Neighborhood patterns and sources of particle pollution: Mobile monitoring and geographic information system analysis in four communities in Accra, Ghana. *Environmental Health Perspectives*, *118*(5), 607–613. http://doi.org/10.1289/ehp.0901365

Dirgawati, M., Barnes, R., Wheeler, A. J., Arnold, A., Mccaul, K. A., Stuart, A. L., … Heyworth, J. S. (2015). Development of Land Use Regression models for predicting exposure to NO2 and NOx in Metropolitan Perth, Western Australia. *Environmental Modelling and Software*, *74*, 258–267. http://doi.org/10.1016/j.envsoft.2015.07.008

Edussuriya, P., Chan, A., & Malvin, A. (2014). Urban morphology and air quality in dense residential environments: Correlations between morphological parameters and air pollution at street-level. *Journal of Engineering Science and Technology*, *9*(1), 64–80.

Eeftens, M., Beelen, R., De Hoogh, K., Bellander, T., Cesaroni, G., Cirach, M., … Hoek, G. (2012a). Development of land use regression models for PM2.5, PM 2.5 absorbance, PM10 and PMcoarse in 20 European study areas; Results of the ESCAPE project. *Environmental Science & Technology*, *46*(20), 11195–11205. http://doi.org/10.1021/es301948k

Eeftens, M., Beelen, R., De Hoogh, K., Bellander, T., Cesaroni, G., Cirach, M., … Hoek, G. (2012b). Development of land use regression models for PM2.5, PM 2.5 absorbance, PM10 and PMcoarse in 20 European study areas; Results of the ESCAPE project. *Environmental Science & Technology*, *46*(20), 11195–11205. http://doi.org/10.1021/es301948k

Eilenberg, R., Subramanian, R., Malings, C., Hauryliuk, A., Presto, A. A., & Robinson, A. L. (2020). Using a network of lower-cost monitors to identify the influence of modifiable factors driving spatial patterns in fine particulate matter concentrations in an urban environment. *In Review*.

Engel-Cox, J. A., Hoff, R. M., & Haymet, A. D. J. (2004). Recommendations on the Use of Satellite Remote-Sensing Data for Urban Air Quality. *Journal of the Air & Waste Management Association*, *54*(11), 1360–1371. http://doi.org/10.1080/10473289.2004.10471005

English, P. B., Olmedo, L., Bejarano, E., Lugo, H., Murillo, E., Seto, E., … Northcross, A. (2017). The Imperial County Community Air Monitoring Network: A Model for Community-based Environmental Monitoring for Public Health Action. *Environmental Health Perspectives*, *125*(7), 74501.

EPA (Environmental Protection Agency). (1990). Air pollution monitoring. Retrieved from https://www.epa.gov/amtic/monitoring-regulations [7 March 2020]

EPA (Environmental Protection Agency). (2014). Air Sensor Toolbox for citizen science. Retrieved from https://www.citizenscience.gov/air-sensor-toolbox/# [18 March 2020]

EPA (Environmental Protection Agency). (2015). *Community Air Sensor Network (CAIRSENSE) project: Lower cost, continuous ambient monitoring methods*. Retrieved from https://www.epa.gov/sites/production/files/2015-07/documents/cairsense.pdf [18 March 2020]

EPA (Environmental Protection Agency). (2016). Interpretation and communication of short-term air sensor data: A pilot project. Retrieved from https://www.epa.gov/sites/production/files/2016-05/documents/interpretation_and_communication_of_short-term_air_sensor_data_a_pilot_project.pdf [18 March 2020]

EPA (Environmental Protection Agency). (2019a). United States Environmental Protection Agency Region I, New England 2019 Healthy Communities Grant Program. Retrieved from https://www3.epa.gov/region1/eco/uep/hcgp.html [7 March 2020]

EPA (Environmental Protection Agency). (2019b). Village Green Project. Retrieved from https://www.epa.gov/air-research/village-green-project [18 March 2020]

EPA (Environmental Protection Agency). (2020a). Air Pollution Monitoring for Communities Grants. Retrieved from https://www.epa.gov/air-research/air-pollution-monitoring-communities-grants [18 March 2020]

EPA (Environmental Protection Agency). (2020b). Air Quality Systems (AQS). Retrieved from https://www.epa.gov/aqs [7 March 2020]

EPA (Environmental Protection Agency). (2020c). Managing air quality - Ambient air monitoring. Retrieved from https://www.epa.gov/air-quality-management-process/managing-air-quality-ambient-air-monitoring [7 March 2020]

Fernández-Somoano, A., Estarlich, M., Ballester, F., Fernández-Patier, R., Aguirre-Alfaro, A., Herce-Garraleta, M. D., & Tardón, A. (2011). Outdoor NO2and benzene exposure in the INMA (Environment and Childhood) Asturias cohort (Spain). *Atmospheric Environment*, *45*(29), 5240–5246. http://doi.org/10.1016/j.atmosenv.2011.02.010

French, S., Barchers, C., & Zhang, W. (2015). Moving beyond Operations: Leveraging Big Data for Urban Planning Decisions. In *56th Annual Conference of Association of College Schools of Planning (ACSP), Portland* (pp. 194-1-194–16).

Gaeta, A., Cattani, G., Di, A., Santis, A. De, Cesaroni, G., Badaloni, C., … Sacco, F. (2016). Development of nitrogen dioxide and volatile organic compounds land use regression models to estimate air pollution exposure near an Italian airport. *Atmospheric Environment*, *131*, 254–262. http://doi.org/10.1016/j.atmosenv.2016.01.052

Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Aiden, E. L., & Fei-Fei, L. (2017). Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States. *Proceedings of the National Academy of Sciences*, *114*(50), 13108–13113.

Glass, D. C., Gray, C. N., Jolley, D. J., Gibbons, C., Sim, M. R., Fritschi, L., … Manuell, R. (2003). Leukemia Risk Associated With Benzene Exposure. *Epidemiology*, *14*(5), 569–577. http://doi.org/10.1097/01.ede.0000082001.05563.e0

Google. (2011). Welcome to Google Maps Platform. Retrieved from https://cloud.google.com/maps-platform [3 April 2020]

Google Earth Outreach. (2020). Project Air View. Retrieved from https://www.google.com/earth/outreach/special-projects/air-quality/ [9 April 2020]

Guerreiro, C. B. B., Foltescu, V., & de Leeuw, F. (2014). Air quality status and trends in Europe. *Atmospheric Environment*, *98*, 376–384. http://doi.org/10.1016/j.atmosenv.2014.09.017

Gulia, S., Nagendra, S. M. S., Khare, M., & Khanna, I. (2015). Urban air quality management-A review. *Atmospheric Pollution Research*, *6*(2), 286–304. http://doi.org/10.5094/APR.2015.033

Habermann, M., & Gouveia, N. (2012). Application of land use regression to predict the concentration of inhalable particular matter in São Paulo city, Brazil. *Engenharia Sanit Ambient*, *17*, 155–162.

Hall, E. S., Kaushik, S. M., Vanderpool, R. W., & Duvall, R. M. (2014). Integrating Sensor Monitoring Technology into the Current Air Pollution Regulatory Support Paradigm: Practical Considerations. *American Journal of Environment Engineering*, *4*(6), 147–154. http://doi.org/10.5923/j.ajee.20140406.02

Hankey, S., Lindsey, G., & Marshall, J. D. (2017). Population-level exposure to particulate air pollution during active travel: Planning for low-exposure, health-promoting cities. *Environmental Health Perspectives*, *125*(4), 527–535. http://doi.org/10.1289/EHP442

Hankey, S., & Marshall, J. D. (2015). Land use regression models of on-road particulate air pollution (particle number, black carbon, pm2.5, particle size) using mobile monitoring. *Environmental Science & Technology*, *49*, 9194–9202. http://doi.org/10.1021/acs.est.5b01209

Hankey, S., Sforza, P., & Pierson, M. (2019). Using mobile monitoring to develop hourly empirical models of particulate air pollution in a rural Appalachian community. *Environmental Science & Technology*, *53*, 4305–4315. research-article. http://doi.org/10.1021/acs.est.8b05249

Hanson, C. S., Noland, R. B., & Brown, C. (2013). The severity of pedestrian crashes: an analysis using Google Street View imagery. *Journal of Transport Geography*, *33*, 42–53. http://doi.org/10.1016/j.jtrangeo.2013.09.002

Hauser, C. D., Buckley, A., & Porter, J. (2015). Passive samplers and community science in regional air quality measurement, education and communication. *Environmental Pollution, 203*, 243–249.

Hochadel, M., Heinrich, J., Gehring, U., Morgenstern, V., Kuhlbusch, T., Link, E., … Krämer, U. (2006a). Predicting long-term average concentrations of traffic-related air pollutants using GIS-based information. *Atmospheric Environment*, *40*(3), 542–553. http://doi.org/10.1016/j.atmosenv.2005.09.067

Hochadel, M., Heinrich, J., Gehring, U., Morgenstern, V., Kuhlbusch, T., Link, E., … Krämer, U. (2006b). Predicting long-term average concentrations of traffic-related air pollutants using GIS-based information. *Atmospheric Environment*, *40*(3), 542–553. http://doi.org/10.1016/j.atmosenv.2005.09.067

Hodgson, A. T. (1995). A Review and a Limited Comparison of Methods for Measuring Total Volatile Organic Compounds in Indoor Air. *Indoor Air*, *5*(4), 247–257. http://doi.org/10.1111/j.1600-0668.1995.00004.x

Hoek, G., Beelen, R., Hoogh, K. De, Vienneau, D., Gulliver, J., Fischer, P., & Briggs, D. (2008). A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric Environment*, *42*(33), 7561–7578. http://doi.org/10.1016/j.atmosenv.2008.05.057

Hoek, G., Eeftens, M., Beelen, R., Fischer, P., Brunekreef, B., Boersma, K. F., & Veefkind, P. (2015). Satellite NO2 data improve national land use regression models for ambient NO2 in a small densely populated country. *Atmospheric Environment*, *105*(2), 173–180. http://doi.org/10.1016/j.atmosenv.2015.01.053

Hoek, G., Fischer, P., Van Den Brandt, P., Goldbohm, S., & Brunekreef, B. (2001). Estimation of long-term average exposure to outdoor air pollution for a cohort study on mortality. *Journal of Exposure Science & Environmental Epidemiology*, *11*(6), 459–469.

Holstius, D. M., Pillarisetti, A., Smith, K. R., & Seto, E. (2014). Field calibrations of a low-cost aerosol sensor at a regulatory monitoring site in California. *Asmospheric Measurement Techniques*, *7*, 1121–1131. http://doi.org/10.5194/amt-7-1121-2014

Huang, K., Bi, J., Meng, X., Geng, G., Lyapustin, A., Lane, K. J., … Liu, Y. (2019). Estimating daily PM2.5 concentrations in New York City at the neighborhood-scale: Implications for integrating non-regulatory measurements. *Science of the Total Environment*, *697*, 134094. http://doi.org/10.1016/j.scitotenv.2019.134094

HUD (US Department of Housing and Urban Development). (2014). *HUD Healthy Communities Transformation Initiative*. Retrieved from https://healthyhousingsolutions.com/service/applied-field-research/hud-healthy-communities-transformation-initiative/ [7 March 2020]

Hvidtfeldt, U. A., Sørensen, M., Geels, C., Ketzel, M., Khan, J., Tjønneland, A., … Raaschou-Nielsen, O. (2019). Long-term residential exposure to PM2.5, PM10, black carbon, NO2, and ozone and mortality in a Danish cohort. *Environment International*, *123*, 265–272.

Hystad, P., Demers, P. A., Johnson, K. C., Brook, J., Donkelaar, A. Van, Lamsal, L., … Brauer, M. (2012). Spatiotemporal air pollution exposure assessment for a Canadian population-based lung cancer case-control study. *Environmental Health*, *11*(1), 1–13.

Hystad, P., Setton, E., Cervantes, A., & Poplawski, K. (2011). Creating national air pollution models for population exposure assessment in Canada. *Environmental Health Perspectives*, *119*(8), 1123–1129. Retrieved from https://open.library.ubc.ca/collections/facultyresearchandpublications/52383/items/1.0220728

Hystad, P., Setton, E., Cervantes, A., Poplawski, K., Deschenes, S., & Brauer, M. (2011a). Creating national air pollution models for population exposure assessment in Canada. *Environmental Health Perspectives*, *119*(8), 1123–1129. http://doi.org/10.1289/ehp.1002976

Hystad, P., Setton, E., Cervantes, A., Poplawski, K., Deschenes, S., & Brauer, M. (2011b). Creating National Air Pollution Models for Population Exposure Assessment in Canada. *Environmental Health Perspectives*, *119*(8), 1123–1129. http://doi.org/10.1289/ehp.1002976

Interagency Monitoring of Protected Visual Environments (IMPROVE). (2019). Federal Land Manager Environmental Database. Retrieved from http://views.cira.colostate.edu/fed/SiteBrowser/Default.aspx [28 April 2019]

IQAir. (2019). 2019 WORLD AIR QUALITY REPORT Region & City PM2.5 Ranking. Retrieved from file:///E:/Air Quality/Measurement Paper/Literature/2019-World-Air-Report-V8-20200318.pdf [21 April 2020]

ISGlobal (Barcelona Institute for Global Health). (2018). 5 Keys to healthier cities. Retrieved from https://www.isglobal.org/en/ciudadesquequeremos [7 March 2020]

Jain, S., Presto, A., & Zimmerman, N. (2020). Spatial modeling of daily PM2.5, NO2 and CO concentrations measured by a low-cost sensor network: Comparison of linear, machine learning, and hybrid land use models. *In Review*.

Jerrett, M., Arain, A., Kanaroglou, P., & Beckerman, B. (2005). A review and evaluation of intraurban air pollution exposure models. *Journal of Exposure Analysis and Environmental Epidemiology*, *15*, 185–204. http://doi.org/10.1038/sj.jea.7500388

Jerrett, M., Burnett, R. T., Pope, C. A., Ito, K., Thurston, G., Krewski, D., … Thun, M. (2009). Long-term ozone exposure and mortality. *New England Journal of Medicine*, *360*(11), 1085–1095. http://doi.org/10.1056/NEJMoa0803894

Jerrett, M., Burnett, R. T., Pope III, C. A., Ito, K., Thurston, G., Krewski, D., … Thun, M. (2009). Long-term ozone exposure and mortality. *New England Journal of Medicine*, *360*(11), 1085–1095.

Jia, C., & Batterman, S. (2010). A critical review of naphthalene sources and exposures relevant to indoor and outdoor air. *International Journal of Environmental Research and Public Health*, *7*(7), 2903–2939. http://doi.org/10.3390/ijerph7072903

Jiao, W., Hagler, G., Williams, R., Sharpe, R., Brown, R., Garver, D., … Buckley, K. (2016). Community Air Sensor Network (CAIRSENSE) project: evaluation of low-cost sensor performance in a suburban environment in the southeastern United States. *Atmospheric Measurement Techniques*, *9*, 5281–5292. http://doi.org/10.5194/amt-9-5281-2016

Johnson, M., Isakov, V., Touma, J. S., Mukerjee, S., & Özkaynak, H. (2010). Evaluation of land-use regression models used to predict air quality concentrations in an urban area. *Atmospheric Environment*, *44*(30), 3660–3668. http://doi.org/10.1016/j.atmosenv.2010.06.041

Jova, M., Lazovi, I., & Pokri, B. (2015). On the use of small and cheaper sensors and devices for indicative citizen-based monitoring of respirable particulate matter. *Environmental Pollution*, *206*, 696–704. http://doi.org/10.1016/j.envpol.2015.08.035

Jung, C.-R., Lin, Y.-T., & Hwang, B.-F. (2013). Air pollution and newly diagnostic autism spectrum disorders: a population-based cohort study in Taiwan. *PloS One*, *8*(9).

Kaufman, A., Williams, R., Barzyk, T., Greenberg, M., Shea, M. O., Sheridan, P., … Preuss, P. W. (2017). A Citizen Science and Government Collaboration: Developing Tools to Facilitate Community Air Monitoring. *Environmental Justice*, *10*(2), 2–6. http://doi.org/10.1089/env.2016.0044

Keller, J. P., Olives, C., Kim, S.-Y., Sheppard, L., Sampson, P. D., Szpiro, A. A., … Kaufman, J. D. (2015). A unified spatiotemporal modeling approach for predicting concentrations of multiple air pollutants in the multi-ethnic study of atherosclerosis and air pollution. *Environmental Health Perspectives*, *123*(4), 301–309.

Kelly, K. E., Whitaker, J., Petty, A., Widmer, C., Dybwad, A., Sleeth, D., … Butter, A. (2017). Ambient and laboratory evaluation of a low-cost particulate matter. *Environmental Pollution*, *221*, 491–500. http://doi.org/10.1016/j.envpol.2016.12.039

Kerckhoffs, J., Wang, M., Meliefste, K., Malmqvist, E., Fischer, P., Janssen, N. A. H., … Hoek, G. (2015). A national fine spatial scale land-use regression model for ozone. *Environmental Research*, *140*, 440–448. http://doi.org/10.1016/j.envres.2015.04.014

Kheirbek, I., Johnson, S., Ross, Z., Pezeshki, G., Ito, K., Eisl, H., & Matte, T. (2012). Spatial variability in levels of benzene, formaldehyde, and total benzene, toluene, ethylbenzene and xylenes in New York City: a land-use regression study. *Environmental Health*, *11*(1), 51.

Kim, S., Bechle, M., Hankey, S., Sheppard, L., Szpiro, A., & Marshall, J. D. (2020). Concentrations of criteria pollutants in the contiguous U.S., 1979 – 2015: Role of prediction model parsimony in integrated empirical geographic regression. *PLoS ONE*, *15*(2), e0228535. http://doi.org/https://doi.org/10.1371/journal.pone.0228535 February

Kim, Y. M., Harrad, S., & Harrison, R. M. (2001). Concentrations and sources of VOCs in urban domestic and public microenvironments. *Environmental Science & Technology*, *35*(6), 997–1004. http://doi.org/10.1021/es000192y

Kimbrough, S., Krabbe, S., Baldauf, R., Barzyk, T., Brown, M., Brown, S., … Shields, A. (2019). The Kansas City Transportation and Local-Scale Air Quality Study (KC-TRAQS): Integration of low-cost sensors and reference grade monitoring in a complex metropolitan area. Part 1: Overview of the project. *Chemosensors*, *7*(2), 26. http://doi.org/https://doi.org/10.3390/chemosensors7020026

Kinney, P. L., Aggarwal, M., Northridge, M. E., Janssen, N. A., & Shepard, P. (2000). Airborne concentrations of PM (2.5) and diesel exhaust particles on Harlem sidewalks: a community-based pilot study. *Environmental Health Perspectives*, *108*(3), 213–218.

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage.

Kloog, I., Koutrakis, P., Coull, B. A., Joo, H., & Schwartz, J. (2011). Assessing temporally and spatially resolved PM 2.5 exposures for epidemiological studies using satellite aerosol optical depth measurements. *Atmospheric Environment*, *45*(35), 6267–6275. http://doi.org/10.1016/j.atmosenv.2011.08.066

Knibbs, L. D., Hewson, M. G., Bechle, M. J., Marshall, J. D., & Barnett, A. G. (2014). A national satellite-based land-use regression model for air pollution exposure assessment in Australia. *Environmental Research*, *135*, 204–211. http://doi.org/10.1016/j.envres.2014.09.011

Kryza, M., Szymanowski, M., & Dore, A. J. (2011). Application of a land-use regression model for calculation of the spatial pattern of annual NO x air concentrations at national scale: a case study for Poland. *Procedia Environmental Sciences*, *7*, 98–103. http://doi.org/10.1016/j.proenv.2011.07.018

Kumar, P., Morawska, L., Martani, C., Biskos, G., Neophytou, M., Di, S., … Britter, R. (2015). The rise of low-cost sensing for managing air pollution in cities. *Environment International*, *75*, 199–205. http://doi.org/10.1016/j.envint.2014.11.019

Kwon, J., Weisel, C. P., Turpin, B. J., Zhang, J., Korn, L. R., Morandi, M. T., … Colome, S. (2006). Source proximity and outdoor-residential VOC concentrations: Results from the RIOPA study. *Environmental Science & Technology*, *40*(13), 4074–4082. http://doi.org/10.1021/es051828u

Laden, F., Neas, L. M., Dockery, D. W., & Schwartz, J. (2000). Association of fine particulate matter from different sources with daily mortality in six US cities. *Environmental Health Perspectives*, *108*(10), 941–947.

Laden, F., Schwartz, J., Speizer, F. E., & Dockery, D. W. (2006a). Reduction in fine particulate air pollution and mortality: extended follow-up of the Harvard Six Cities study. *American Journal of Respiratory and Critical Care Medicine*, *173*(6), 667–672.

Laden, F., Schwartz, J., Speizer, F. E., & Dockery, D. W. (2006b). Reduction in fine particulate air pollution and mortality: Extended follow-up of the Harvard Six Cities Study. *American Journal of Respiratory and Critical Care Medicine*, *173*(6), 667–672. http://doi.org/10.1164/rccm.200503-443OC

Laden, F., Schwartz, J., Speizer, F. E., & Dockery, D. W. (2006c). Reduction in fine particulate air pollution and mortality extended follow-up of the Harvard Six Cities Study. *American Journal of Respiratory and Critical Care Medicine*, *173*, 667–672. http://doi.org/10.1164/rccm.200503-443OC

Lafontaine, S. J. V, Sawada, M., & Kristjansson, E. (2017). A direct observation method for auditing large urban centers using stratified sampling, mobile GIS technology and virtual environments. *International Journal of Health Geographics*, *16*(1), 6.

Lamsal, L. N., Martin, R. V, Donkelaar, A. Van, Steinbacher, M., Celarier, E. A., Bucsela, E., … Pinto, J. P. (2008). Ground-level nitrogen dioxide concentrations inferred from the satellite-borne Ozone Monitoring Instrument. *Journal of Geophysical Research*, *113*(2), 1–15. http://doi.org/10.1029/2007JD009235

Lansing, J., Hanlon, P., & Doten, J. (2016). *Air Quality in Minneapolis: A Neighborhood Approach*. Retrieved from http://www.minneapolismn.gov/www/groups/public/@regservices/documents/webcontent/wcmsp-192216.pdf

Larkin, A., & Hystad, P. (2017). Towards personal exposures: How technology is changing air pollution and health research. *Current Environmental Health Report*, *4*, 463–471. http://doi.org/10.1007/s40572-017-0163-y

Larkin, A., & Hystad, P. (2019). Evaluating street view exposure measures of visible green space for health research. *Journal of Exposure Science & Environmental Epidemiology*, 447–456. http://doi.org/10.1038/s41370-018-0017-1

Larson, T., Su, J., Baribeau, A. M., Buzzelli, M., Setton, E., & Brauer, M. (2007). A spatial model of urban winter woodsmoke concentrations. *Environmental Science & Technology*, *41*(7), 2429–2436. http://doi.org/10.1021/es0614060

Lary, D. J., Alavi, A. H., Gandomi, A. H., & Walker, A. L. (2016). Machine learning in geosciences and remote sensing. *Geoscience Frontiers*, *7*(1), 3–10. http://doi.org/10.1016/j.gsf.2015.07.003

Law, S., Paige, B., & Russell, C. (2018). Take a Look Around: Using Street View and Satellite Images to Estimate House Prices. Retrieved from http://arxiv.org/abs/1807.07155

Li, X., Zhang, C., & Li, W. (2017). Building block level urban land-use information retrieval based on Google Street View images. *GIScience and Remote Sensing*, *54*(6), 819–835. http://doi.org/10.1080/15481603.2017.1338389

Li, X., Zhang, C., Li, W., & Kuzovkina, Y. A. (2016). Environmental inequities in terms of different types of urban greenery in Hartford, Connecticut. *Urban Forestry & Urban Greening*, *18*, 163–172. http://doi.org/10.1016/j.ufug.2016.06.002

Liang, Z., Yang, Y., Li, J., Zhu, X., Ruan, Z., Chen, S., … Zhao, Q. (2019). Migrant population is more vulnerable to the effect of air pollution on preterm birth: Results from a birth cohort study in seven Chinese cities. *International Journal of Hygiene and Environmental Health*, *222*(7), 1047–1053.

Lim, C. C., Kim, H., Vilcassim, M. J. R., Thurston, G. D., Gordon, T., Chen, L., … Kim, S. (2019). Mapping urban air quality using mobile sampling with low-cost sensors and machine learning in Seoul, South Korea. *Environment International*, *131*(June), 105022. http://doi.org/10.1016/j.envint.2019.105022

Lin, C., Li, Y., Yuan, Z., Lau, A. K. H., Li, C., & Fung, J. C. H. (2015). Using satellite remote sensing data to estimate the high-resolution distribution of ground-level PM2.5. *Remote Sensing of Environment*, *156*, 117–128. http://doi.org/10.1016/j.rse.2014.09.015

Lin, M., Chen, Y., Villeneuve, P. J., Burnett, R. T., Lemyre, L., Hertzman, C., … Krewski, D. (2004). Gaseous air pollutants and asthma hospitalization of children with low household income in Vancouver, British Columbia, Canada. *American Journal of Epidemiology*, *159*(3), 294–303. http://doi.org/10.1093/aje/kwh043

Lin, S., Liu, X., Le, L. H., & Hwang, S.-A. (2008). Chronic exposure to ambient ozone and asthma hospital admissions among children. *Environmental Health Perspectives*, *116*(12), 1725–1730.

Lin, Z., & Xu, H. (2016). A study of urban heat island intensity based on “local climate zones ”: A case study in Fuzhou , China. In *2016 Fourth International Workshop on Earth Observation and Remote Sensing Applications* (pp. 3–7).

Liu, Y., Arp, H. P. H., Song, X., & Song, Y. (2017). Research on the relationship between urban form and urban smog in China. *Environment and Planning B: Urban Analytics and City Science*, *44*(2), 328–342. http://doi.org/10.1177/0265813515624687

Madaio, M., Chen, S.-T., Haimson, O. L., Zhang, W., Cheng, X., Hinds-Aldrich, M., … Dilkina, B. (2016). Firebird: Predicting Fire Risk and Prioritizing Fire Inspections in Atlanta. In

Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 185–194). http://doi.org/10.475/123

Madsen, C., Carlsen, K. C. L., Hoek, G., Oftedal, B., Nafstad, P., Meliefste, K., … Brunekreef, B. (2007). Modeling the intra-urban variability of outdoor traffic pollution in Oslo, Norway-A GA2LEN project. *Atmospheric Environment*, *41*(35), 7500–7511. http://doi.org/10.1016/j.atmosenv.2007.05.039

Madsen, C., Gehring, U., Eldevik, S., Nafstad, P., Meliefste, K., Nystad, W., … Brunekreef, B. (2011). Comparison of land-use regression models for predicting spatial NOx contrasts over a three year period in Oslo , Norway. *Atmospheric Environment*, *45*(21), 3576–3583. http://doi.org/10.1016/j.atmosenv.2011.03.069

Malings, C., Tanzer, R., Hauryliuk, A., Kumar, S. P. N., Zimmerman, N., Kara, L. B., … Subramanian, R. (2019). Development of a general calibration model and long-term performance evaluation of low-cost sensors for air pollutant gas monitoring. *Atmospheric Measurement Techniques, 12*, 903–920.

Malings, C., Tanzer, R., Hauryliuk, A., Saha, P. K., Allen, L., Presto, A. A., … Subramanian, R. (2019). Fine particle mass monitoring with low-cost sensors: Corrections and long-term performance evaluation term performance evaluation. *Aerosol Science and Technology*, *0*(0), 1–15. http://doi.org/10.1080/02786826.2019.1623863

Marshall, J. D., Nethery, E., & Brauer, M. (2008). Within-urban variability in ambient air pollution: Comparison of estimation methods. *Atmospheric Environment*, *42*, 1359–1369. http://doi.org/10.1016/j.atmosenv.2007.08.012

Martin, R. V. (2008a). Satellite remote sensing of surface air quality. *Atmospheric Environment*, *42*(34), 7823–7843. http://doi.org/10.1016/j.atmosenv.2008.07.018

Martin, R. V. (2008b). Satellite remote sensing of surface air quality. *Atmospheric Environment*, *42*(34), 7823–7843. http://doi.org/10.1016/j.atmosenv.2008.07.018

Masiol, M., Chalupa, D. C., Rich, D. Q., Ferro, A. R., & Hopke, P. K. (2018). Hourly land-use regression models based on low-cost PM monitor data. *Environmental Research*, *167*(April), 7–14. http://doi.org/10.1016/j.envres.2018.06.052

Masiol, M., Squizzato, S., Chalupa, D., Rich, D. Q., & Hopke, P. K. (2019). Spatial-temporal variations of summertime ozone concentrations across a metropolitan area using a network of low-cost monitors to develop 24 hourly land-use regression models. *Science of the Total Environment*, *654*, 1167–1178. http://doi.org/10.1016/j.scitotenv.2018.11.111

McCarty, J., & Kaza, N. (2015). Urban form and air quality in the United States. *Landscape and Urban Planning*, *139*, 168–179. http://doi.org/10.1016/j.landurbplan.2015.03.008

Mead, M. I., Popoola, O. A. M., Stewart, G. B., Landshoff, P., Calleja, M., Hayes, M., … Jones, R. L. (2013). The use of electrochemical sensors for monitoring urban air quality in low-cost, high-density networks. *Atmospheric Environment*, *70*, 186–203. http://doi.org/10.1016/j.atmosenv.2012.11.060

Mečiarová, L., Vilčeková, S., Burdová, E. K., & Kiselák, J. (2017). Factors effecting the total volatile organic compound (TVOC) concentrations in slovak households. *International*

Journal of Environmental Research and Public Health, *14*(12). http://doi.org/10.3390/ijerph14121443

Meng, X., Hand, J. L., Schichtel, B. A., & Liu, Y. (2018). Space-time trends of PM2.5 constituents in the conterminous United States estimated by a machine learning approach, 2005–2015. *Environment International*, *121*, 1137–1147. http://doi.org/10.1016/j.envint.2018.10.029

Mercer, L. D., Szpiro, A. A., Sheppard, L., Lindström, J., Adar, S. D., Allen, R. W., … Kaufman, J. D. (2011a). Comparing universal kriging and land-use regression for predicting concentrations of gaseous oxides of nitrogen (NOx) for the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air). *Atmospheric Environment*, *45*(26), 4412–4420. http://doi.org/10.1016/j.atmosenv.2011.05.043

Mercer, L. D., Szpiro, A. A., Sheppard, L., Lindström, J., Adar, S. D., Allen, R. W., … Kaufman, J. D. (2011b). Comparing universal kriging and land-use regression for predicting concentrations of gaseous oxides of nitrogen (NOx) for the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air). *Atmospheric Environment*, *45*(26), 4412–4420. http://doi.org/10.1016/j.atmosenv.2011.05.043

Microsoft. (2018). Microsoft releases 125 million building footprints in the US as open data. Retrieved from https://blogs.bing.com/maps/2018-06/microsoft-releases-125-million-building-footprints-in-the-us-as-open-data [8 March 2020]

Middel, A., Häb, K., Brazel, A. J., Martin, C. A., & Guhathakurta, S. (2014). Impact of urban form and design on mid-afternoon microclimate in Phoenix Local Climate Zones. *Landscape and Urban Planning*, *122*, 16–28. http://doi.org/10.1016/j.landurbplan.2013.11.004

Mills, G., Bechtel, B., Ching, J., See, L., Feddema, J., Foley, M., … Connor, M. O. (2015). An Introduction to the WUDAPT project. In *ICUC9 - 9th International Conference on Urban Climate jointly with 12th Symposium on the Urban Environment*.

Minkler, M., Garcia, A. P., Williams, J., LoPresti, T., & Lilly, J. (2010). Si se puede: using participatory research to promote environmental justice in a Latino community in San Diego, California. *Journal of Urban Health*, *87*(5), 796–812.

Miskell, G., Salmond, J. A., & Williams, D. E. (2018). Use of a handheld low-cost sensor to explore the effect of urban design features on local-scale spatial and temporal air quality variability. *Science of the Total Environment*, *619–620*, 480–490. http://doi.org/10.1016/j.scitotenv.2017.11.024

Mitraka, Z., Del Frate, F., Chrysoulakis, N., & Gastellu-Etchegorry, J.-P. (2015). Exploiting earth observation data products for mapping local climate zones. In *2015 Joint Urban Remote Sensing Event (JURSE)* (pp. 1–4).

Monino, J.-L., & Sedkaoui, S. (2016). *Big data, open data and data development*. John Wiley & Sons.

Morawska, L., Thai, P. K., Liu, X., Asumadu-sakyi, A., Ayoko, G., Bartonova, A., … Williams, R. (2018). Applications of low-cost sensing technologies for air quality monitoring and exposure assessment: How far have they gone? *Environmental International*, *116*(April), 286–299. http://doi.org/10.1016/j.envint.2018.04.018

Mukerjee, S., Smith, L. A., Johnson, M. M., Neas, L. M., & Stallings, C. A. (2009). Spatial analysis and land use regression of VOCs and NO2 from school-based urban air monitoring in Detroit / Dearborn , USA. *Science of the Total Environment*, *407*(16), 4642–4651. http://doi.org/10.1016/j.scitotenv.2009.04.030

Mukerjee, S., Smith, L., Neas, L., & Norris, G. (2012). Evaluation of land use regression models for nitrogen dioxide and benzene in four US cities. *The Scientific World Journal*, *2012*. http://doi.org/10.1100/2012/865150

Mukund, R., Kelly, T. J., & Spicer, C. W. (1996). Source attribution of ambient air toxic and other VOCs in Columbus, Ohio. *Atmospheric Environment*, *30*(20), 3457–3470. http://doi.org/10.1016/1352-2310(95)00487-4

Muller, C. L., Chapman, L., Johnston, S., & Kidd, C. (2015). Crowdsourcing for climate and atmospheric sciences: current status and future potential. *International Journal of Climatology*, *3203*(January), 3185–3203. http://doi.org/10.1002/joc.4210

Multi-Resolution Land Characteristics (MRLC) Consortium. (2016). NLCD Land Cover. Retrieved from https://www.mrlc.gov/data [3 April 2020]

Nafstad, P., Ha, L. L., Oftedal, B., Gram, F., Holme, I., Hjermann, I., & Leren, P. (2003). Lung cancer and air pollution: a 27 year follow up of 16 209 Norwegian men. *Thorax*, *58*, 1071–1076.

NASA (National Aeronautics and Space Administration). (2020). Atmospheric Science Data Center. Retrieved from https://eosweb.larc.nasa.gov/ [8 March 2020]

NERL (National Exposure Research Laboratory). (2015). Air quality monitoring for citizen science. Retrieved from https://www.niehs.nih.gov/research/supported/translational/peph/podcasts/2015/may22_air-quality/index.cfm [7 March 2020]

NOAA (National Oceanic and Atmospheric Administration). (2020). Explore the world in real-time. Retrieved from https://www.nesdis.noaa.gov/content/imagery-and-data [7 March 2020]

Novotny, E. V., Bechle, M. J., Millet, D. B., & Marshall, J. D. (2011a). National satellite-based land-use regression: NO2 in the United States. *Environmental Science and Technology*, *45*(10), 4407–4414. http://doi.org/10.1021/es103578x

Novotny, E. V, Bechle, M. J., Millet, D. B., & Marshall, J. D. (2011b). National satellite-based land-use regression: NO2 in the United States. *Environmental Science and Technology*, *45*(10), 4407–4414. http://doi.org/10.1021/es103578x

Oiamo, T. H., Johnson, M., Tang, K., & Luginaah, I. N. (2015). Assessing traffi c and industrial contributions to ambient nitrogen dioxide and volatile organic compounds in a low pollution urban environment. *Science of the Total Environment, The*, *529*, 149–157. http://doi.org/10.1016/j.scitotenv.2015.05.032

OMI Science Team. (2012). OMI/Aura Level 2 Sulphur Dioxide (SO2) Trace Gas Column Data 1-Orbit subset Swath along CloudSat track 1-Orbit Swath 13x24 km, Edited by GES DISC, Greenbelt, MD, USA Goddard Earth Sciences Data and Information Services Center (GES DISC). Retrieved from https://disc.gsfc.nasa.gov/datacollection/OMSO2_CPR_003.html [1 April 2017]

OpenStreetMap. (2020). OpenStreetMap platform. Retrieved from https://www.openstreetmap.org/#map=4/38.01/-95.84 [7 March 2020]

Pankow, J. F., Luo, W., Bender, D. A., Isabelle, L. M., Hollingsworth, J. S., Chen, C., … Zogorski, J. S. (2003). Concentrations and co-occurrence correlations of 88 volatile organic compounds (VOCs) in the ambient air of 13 semi-rural to urban locations in the United States. *Atmospheric Environment*, *37*(36), 5023–5046. http://doi.org/10.1016/j.atmosenv.2003.08.006

Pedersen, M., Giorgis-Allemand, L., Bernard, C., Aguilera, I., Andersen, A.-M. N., Ballester, F., … others. (2013). Ambient air pollution and low birthweight: a European cohort study (ESCAPE). *The Lancet Respiratory Medicine*, *1*(9), 695–704.

Petralli, M., Massetti, L., & Orlandini, S. (2014). Urban planning indicators: useful tools to measure the effect of urbanization and vegetation on summer air temperatures. *International Journal of Climatology*, *34*, 1236–1244. http://doi.org/10.1002/joc.3760

Piccot, S. D., Watson, J. J., & Jones, J. W. (1992). A global inventory of volatile organic compound emissions from anthropogenic sources. *Journal of Geophysical Research: Atmospheres*, *97*(D9), 9897–9912. http://doi.org/10.1029/92JD00682

Poirier, A., Dodds, L., Dummer, T., Rainham, D., Maguire, B., & Johnson, M. (2015). Maternal exposure to air pollution and adverse birth outcomes in Halifax, Nova Scotia. *Journal of Occupational and Environmental Medicine*, *57*(12), 1291–1298. http://doi.org/10.1097/JOM.0000000000000604

Pope, C. A., Lefler, J. S., Ezzati, M., Higbee, J. D., Marshall, J. D., Kim, S., … Burnett, R. T. (2019). Mortality Risk and Fine Particulate Air Pollution in a Large, Representative Cohort. *Environmental Health Perspectives*, *127*(July), 1–9.

Pope III, C. A., Burnett, R. T., Thurston, G. D., Thun, M. J., Calle, E. E., Krewski, D., & Godleski, J. J. (2004). Cardiovascular mortality and long-term exposure to particulate air pollution: epidemiological evidence of general pathophysiological pathways of disease. *Circulation*, *109*(1), 71–77.

PurpleAir. (2020). PurpleAir: Real time air quality monitoring. Retrieved from https://www2.purpleair.com/?gclid=EAIaIQobChMIqeba6_Kk6AIVQIFaBR3MYwI4EAAYASAAEgL_yvD_BwE [18 March 2020]

Quan, S. J., Dutt, F., Woodworth, E., Yamagata, Y., & Yang, P. P. J. (2017). Local Climate Zone Mapping for Energy Resilience: A Fine-grained and 3D Approach. *Energy Procedia*, *105*, 3777–3783. http://doi.org/10.1016/j.egypro.2017.03.883

Rada, E. C., Ragazzi, M., Brini, M., Marmo, L., Zambelli, P., Chelodi, M., & Ciolli, M. (2016). Perspectives of Low-Cost Sensors Adoption for Air Quality Monitoring. In *Air Quality* (pp.

29–40). Apple Academic Press.

Rai, A. C., Kumar, P., Pilla, F., Skouloudis, A. N., Di, S., Ratti, C., … Rickerby, D. (2017). End-user perspective of low-cost sensors for outdoor air pollution monitoring. *Science of the Total Environment*, *607–608*, 691–705. http://doi.org/10.1016/j.scitotenv.2017.06.266

Reid, C. E., Jerrett, M., Petersen, M. L., Pfister, G. G., Morefield, P. E., Tager, I. B., … Balmes, J. R. (2015). Spatiotemporal prediction of fine particulate matter during the 2008 Northern California wildfires using machine learning. *Environmental Science and Technology*, *49*(6), 3887–3896. http://doi.org/10.1021/es505846r

Reis, S., Seto, E., Northcross, A., Quinn, N. W. T., Convertino, M., Jones, R. L., … Wimberly, M. C. (2015). Integrating modelling and smart sensors for environmental and human health. *Environmental Modelling and Software*, *74*, 238–246. http://doi.org/10.1016/j.envsoft.2015.06.003

Ross, Z., Jerrett, M., Ito, K., Tempalski, B., & Thurston, G. D. (2007). A land use regression for predicting fine particulate matter concentrations in the New York City region. *Atmospheric Environment*, *41*, 2255–2269. http://doi.org/10.1016/j.atmosenv.2006.11.012

Ryan, P. H., & Lemasters, G. K. (2007). A review of land-use regression models for characterizing intraurban air pollution exposure. *Inhalation Toxicology*, *19*(2), 127–133.

Rzotkiewicz, A., Pearson, A. L., Dougherty, B. V., Shortridge, A., & Wilson, N. (2018a). Systematic review of the use of Google Street View in health research: Major themes, strengths, weaknesses and possibilities for future research. *Health and Place*, *52*(July), 240–246. http://doi.org/10.1016/j.healthplace.2018.07.001

Rzotkiewicz, A., Pearson, A. L., Dougherty, B. V, Shortridge, A., & Wilson, N. (2018b). Systematic review of the use of Google Street View in health research: Major themes, strengths, weaknesses and possibilities for future research. *Health and Place*, *52*(July), 240–246. http://doi.org/10.1016/j.healthplace.2018.07.001

Schneider, P., Castell, N., Vogt, M., Dauge, F. R., Lahoz, W. A., & Bartonova, A. (2017). Mapping urban air quality in near real-time using observations from low- cost sensors and model information. *Environment International*, *106*(June), 234–247. http://doi.org/10.1016/j.envint.2017.05.005

Sheng, N., & Tang, U. W. (2011). A building-based data capture and data mining technique for air quality assessment. *Frontiers of Environmental Science & Engineering in China*, *5*(4), 543–551.

Shi, Y., Lau, K. K.-L., & Ng, E. (2016a). Developing street-level PM2. 5 and PM10 land use regression models in high-density Hong Kong with urban morphological factors. *Environmental Science & Technology*, *50*(15), 8178–8187.

Shi, Y., Lau, K. K., & Ng, E. (2016b). Developing street-level PM2.5 and PM10 land use regression models in high-density Hong Kong with urban morphological factors. *Environmental Science & Technology*, *50*, 8178–8187. http://doi.org/10.1021/acs.est.6b01807

Singh, D., Kumar, A., Kumar, K., Singh, B., Mina, U., Singh, B. B., & Jain, V. K. (2016).

Statistical modeling of O3, NOx, CO, PM2.5, VOCs and noise levels in commercial complex and associated health risk assessment in an academic institution. *Science of the Total Environment*, *572*(x), 586–594. http://doi.org/10.1016/j.scitotenv.2016.08.086

Singh, D., Kumar, A., Singh, B. P., Anandam, K., Singh, M., Mina, U., … Jain, V. K. (2016). Spatial and temporal variability of VOCs and its source estimation during rush/non-rush hours in ambient air of Delhi, India. *Air Quality, Atmosphere and Health*, *9*(5), 483–493. http://doi.org/10.1007/s11869-015-0354-3

Smith, L. A., Mukerjee, S., Chung, K. C., & Afghani, J. (2011). Spatial analysis and land use regression of VOCs and NO2 in Dallas, Texas during two seasons. *Journal of Environmental Monitoring*, *13*, 999–1007. http://doi.org/10.1039/c0em00724b

Smith, L. A., Stock, T. H., Chung, K. C., Mukerjee, S., Liao, X. L., Stallings, C., & Afshar, M. (2007). Spatial analysis of volatile organic compounds from a community-based air toxics monitoring network in Deer Park, Texas, USA. *Environmental Monitoring and Assessment*, *128*(1–3), 369–379. http://doi.org/10.1007/s10661-006-9320-8

Smith, L., Mukerjee, S., Gonzales, M., Stallings, C., Neas, L., & Norris, G. (2006). Use of GIS and ancillary variables to predict volatile organic compound and nitrogen dioxide levels at unmonitored locations. *Atmospheric Environment*, *40*, 3773–3787. http://doi.org/10.1016/j.atmosenv.2006.02.036

Snyder, E. G., Watkins, T. H., Solomon, P. A., Thoma, E. D., Williams, R. W., Hagler, G. S. W., … Preuss, P. W. (2013a). The changing paradigm of air pollution monitoring. *Environmental Science & Technology*, *47*, 11369–11377.

Snyder, E. G., Watkins, T. H., Solomon, P. A., Thoma, E. D., Williams, R. W., Hagler, G. S. W., … Preuss, P. W. (2013b). The Changing Paradigm of Air Pollution Monitoring. *Environmental Science & Technology*, *47*, 11369–11377. http://doi.org/10.1021/es4022602

South Coast Air Quality Management District. (2018). *Field Evaluation Purple Air PM Sensor Background*. Retrieved from http://www.aqmd.gov/docs/default-source/aq-spec/field-evaluations/purple-air-pa-ii---field-evaluation.pdf?sfvrsn=4

Stavrakou, T., Eskes, H., & Roozendael, M. Van. (2008). and Physics Twelve years of global observations of formaldehyde in the troposphere using GOME and SCIAMACHY sensors. *Atmospheric Chemistry and Physics*, *8*, 4947–4963.

Stedman, J. R., Vincent, K. J., Campbell, G. W., Goodwin, J. W. L., & Downing, C. E. H. (1997). New high resolution maps of estimated backgroup ambient NOx and NO2 concentrations in the U.K. *Atmospheric Environment*, *31*(21), 3591–3602.

Steeneveld, G., Klompmaker, J. O., Groen, R. J. A., & Holtslag, A. A. M. (2016a). An urban climate assessment and management tool for combined heat and air quality judgements at neighbourhood scales. *Resources, Conservation & Recycling*. http://doi.org/10.1016/j.resconrec.2016.12.002

Steeneveld, G., Klompmaker, J. O., Groen, R. J. A., & Holtslag, A. A. M. (2016b). An urban climate assessment and management tool for combined heat and air quality judgements at neighbourhood scales. *Resources, Conservation {&} Recycling*.

http://doi.org/10.1016/j.resconrec.2016.12.002

Stephens, M. (2013). Gender and the GeoWeb: divisions in the production of user-generated cartographic information. *GeoJournal*, *78,* 981–996. http://doi.org/10.1007/s10708-013-9492-z

Stewart, I. D., & Oke, T. R. (2012a). Local climate zones for urban temperature studies. *Bulletin of the American Meteorological Society*, *93*(12), 1879–1900. http://doi.org/10.1175/BAMS-D-11-00019.1

Stewart, I. D., & Oke, T. R. (2012b). Local climate zones for urban temperature studies. *Bulletin of the American Meteorological Society*, *93*(12), 1879–1900. http://doi.org/10.1175/BAMS-D-11-00019.1

Stieb, D. M., Chen, L., Beckerman, B. S., Jerrett, M., Crouse, D. L., Omariba, D. W. R., … others. (2016). Associations of pregnancy outcomes and PM2.5 in a national Canadian study. *Environmental Health Perspectives*, *124*(2), 243–249.

Stone, B. (2005). Urban heat and air pollution: An emerging role for planners in the climate change debate. *Journal of the American Planning Association*, *71*(1), 13–25. http://doi.org/10.1080/01944360508976402

Su, J. G., Allen, G., Miller, P. J., & Brauer, M. (2013). Spatial modeling of residential woodsmoke across a non-urban upstate New York region. *Air Quality, Atmosphere and Health*, *6*(1), 85–94. http://doi.org/10.1007/s11869-011-0148-1

Su, J. G., Brauer, M., Ainslie, B., Steyn, D., Larson, T., & Buzzelli, M. (2008). An innovative land use regression model incorporating meteorology for exposure analysis. *Science of the Total Environment*, *390,* 520–529. http://doi.org/10.1016/j.scitotenv.2007.10.032

Su, J. G., Brauer, M., & Buzzelli, M. (2008). Estimating urban morphometry at the neighborhood scale for improvement in modeling long-term average air pollution concentrations. *Atmospheric Environment*, *42*(34), 7884–7893. http://doi.org/10.1016/j.atmosenv.2008.07.023

Su, J. G., Jerrett, M., & Beckerman, B. (2009). A distance-decay variable selection strategy for land use regression modeling of ambient air pollution exposures. *Science of the Total Environment*, *407*(12), 3890–3898. http://doi.org/10.1016/j.scitotenv.2009.01.061

Su, J. G., Jerrett, M., Beckerman, B., Verma, D., Arain, M. A., Kanaroglou, P., … Brook, J. (2010). A land use regression model for predicting ambient volatile organic compound concentrations in Toronto , Canada. *Atmospheric Environment*, *44*(29), 3529–3537. http://doi.org/10.1016/j.atmosenv.2010.06.015

Su, J. G., Larson, T., Gould, T., Cohen, M., & Buzzelli, M. (2010). Transboundary air pollution and environmental justice: Vancouver and Seattle compared. *GeoJournal*, *75*(6), 595–608.

Sun, J., Wu, F., Hu, B., Tang, G., Zhang, J., & Wang, Y. (2016). VOC characteristics, emissions and contributions to SOA formation during hazy episodes. *Atmospheric Environment*, *141*, 560–570. http://doi.org/10.1016/j.atmosenv.2016.06.060

Tang, R., Blangiardo, M., & Gulliver, J. (2013). Using building heights and street configuration to enhance intraurban PM10, NOX, and NO2 Land Use Regression Models. *Environmental Science & Technology*, *47*, 11643–11650. http://doi.org/10.1021/es402156g

Theobald, D. M. (2014). Development and applications of a comprehensive land use classification and map for the US. *PloS One*, *9*(4).

Thompson, J. E. (2016a). Crowd-sourced air quality studies: A review of the literature & portable sensors. *Biochemical Pharmacology*, *11*, 23–34. http://doi.org/10.1016/j.teac.2016.06.001

Thompson, J. E. (2016b). Crowd-sourced air quality studies: A review of the literature & portable sensors. *Trends in Environmental Analytical Chemistry*, *11*, 23–34. http://doi.org/10.1016/j.teac.2016.06.001

Thurston, G. D., Ahn, J., Cromar, K. R., Shao, Y., Reynolds, H. R., Jerrett, M., … Hayes, R. B. (2016). Ambient particulate matter air pollution exposure and mortality in the NIH-AARP diet and health cohort. *Environmental Health Perspectives*, *124*(4), 484–490.

Truax, C., Hricko, A., Gottlieb, R., Tovar, J., Betancourt, S., & Chien-Hale, M. (2013). *Neighborhood Assessment Teams*. Retrieved from https://envhealthcenters.usc.edu/wp-content/uploads/2016/11/Neighborhood-Assessment-Teams-on-traffic-pollution-2013.pdf [18 March 2020]

Tulloch, M., & Li, J. (2004). Applications of Satellite Remote Sensing to Urban Air-Quality Monitoring: Status and Potential Solutions to Canada. *Environmental Informatics Archives*, *2*, 846–854. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.725.9058&rep=rep1&type=pdf

U.S. Environmental Protection Agency (EPA). (1991). *Chemical Concentration Data near the Detection Limit*. Retrieved from https://www.navfac.navy.mil/niris/MID_ATLANTIC/OCEANA_NAS/BASEWIDE/ADMIN RECORD/N60191_000280.pdf

US Census Bureau. (2020). The 2020 Census is happening now. Retrieved from https://www.census.gov/ [3 March 2020]

USDA (United States Department of Agriculture. (2010). Air quality effects of urban trees and parks. Retrieved from https://www.fs.usda.gov/treesearch/pubs/52881 [3 April 2020]

Van Donkelaar, A., Martin, R. V, Brauer, M., Hsu, N. C., Kahn, R. A., Levy, R. C., … Winker, D. M. (2016). Global estimates of fine particulate matter using a combined geophysical-statistical method with information from satellites, models, and monitors. *Environmental Science & Technology*, *50*(7), 3762–3772.

Vienneau, D., Hoogh, K. De, Bechle, M. J., Beelen, R., Donkelaar, A. Van, Martin, R. V, … Marshall, J. D. (2013). Western European land use regression incorporating satellite- and ground-based measurements of NO2 and PM10. *Environmental Science and Technology*, *47*(2), 13555–13564.

Villeneuve, P. J., Jerrett, M., Brenner, D., Su, J., Chen, H., & Mclaughlin, J. R. (2014). Original contribution a case-control study of long-term exposure to ambient volatile organic compounds and lung cancer in Toronto , Ontario , Canada. *American Journal of*

Epidemiology, *179*(4), 443–451. http://doi.org/10.1093/aje/kwt289

Villeneuve, P. J., Jerrett, M., Su, J., Burnett, R. T., Chen, H., Brook, J., … Goldberg, M. S. (2013). A cohort study of intra-urban variations in volatile organic compounds and mortality, Toronto, Canada. *Environmental Pollution*, *183*, 30–39.

Vlaanderen, J., Portengen, L., Adam, M. C., Ulrike, S., Bert, G., Hoek, G., & Vermeulen, R. (2019). Error in air pollution exposure model determinants and bias in health estimates. *Journal of Exposure Science & Environmental Epidemiology*, 258–266. http://doi.org/10.1038/s41370-018-0045-x

Wang, C., Middel, A., Myint, S. W., Kaplan, S., Brazel, A. J., & Lukasczyk, J. (2018). Assessing local climate zones in arid cities: The case of Phoenix, Arizona and Las Vegas, Nevada. *ISPRS Journal of Photogrammetry and Remote Sensing*, *141*, 59–71. http://doi.org/10.1016/j.isprsjprs.2018.04.009

Wang, M., Sampson, P. D., Hu, J., Kleeman, M., Keller, J. P., Olives, C., … Kaufman, J. D. (2017). Combining Land-Use Regression and Chemical Transport Modeling in a Spatio-temporal Geostatistical Model for Ozone and PM2.5. *Environmental Science & Technology*, *50*(10), 5111–5118. http://doi.org/10.1021/acs.est.5b06001.Combining

Wang, Y., Li, J., Jing, H., Zhang, Q., Jiang, J., Biswas, P., … Biswas, P. (2015). Laboratory evaluation and Calibration of three low-Cost particle sensors for particulate matter measurement. *Aerosol Science and Technology*, *49*(11), 1063–1077. http://doi.org/10.1080/02786826.2015.1100710

Watson, J. G., Chow, J. C., & Fujita, E. M. (2001). Review of volatile organic compound source apportionment by chemical mass balance. *Atmospheric Environment*, *35*(9), 1567–1584. http://doi.org/10.1016/S1352-2310(00)00461-1

Weichenthal, S., Ryswyk, K. Van, Goldstein, A., Bagg, S., Shekkarizfard, M., & Hatzopoulou, M. (2016). A land use regression model for ambient ultrafine particles in Montreal, Canada: A comparison of linear regression and a machine learning approach. *Environmental Research*, *146*, 65–72. http://doi.org/10.1016/j.envres.2015.12.016

Weissert, L. F., Alberti, K., Miskell, G., Pattinson, W., Salmond, J. A., & Henshaw, G. (2019). Low-cost sensors and microscale land use regression: Data fusion to resolve air quality variations with high spatial and temporal resolution. *Atmospheric Environment*, *213*(May), 285–295. http://doi.org/10.1016/j.atmosenv.2019.06.019

Weissert, L. F., Salmond, J. A., Miskell, G., Alavi-shoshtari, M., & Williams, D. E. (2018). Development of a microscale land use regression model for predicting NO2 concentrations at a heavy trafficked suburban area in Auckland , NZ. *Science of the Total Environment*, *619–620*, 112–119. http://doi.org/10.1016/j.scitotenv.2017.11.028

Wheeler, A. J., Smith-Doiron, M., Xu, X., Gilbert, N. L., & Brook, J. R. (2008). Intra-urban variability of air pollution in Windsor, Ontario-Measurement and modeling for human exposure assessment. *Environmental Research*, *106*(1), 7–16. http://doi.org/10.1016/j.envres.2007.09.004

Whitelaw, G., Vaughan, H., Craig, B., & Atkinson, D. (2003). Establishing the Canadian community monitoring network. *Environmental Monitoring and Assessment*, *88*(1–3), 409–418.

WHO (World Health Organization). (2009). *Global health risks: Mortality and Burden of Disease attributable to selected major risks*. Retrieved from https://apps.who.int/iris/handle/10665/44203 [7 March 2020]

Williams, D. E. (2019). Low cost sensor networks: How do we know the data are reliable? *ACS Sensors*, *4*, 2558–2565. other. http://doi.org/10.1021/acssensors.9b01455

Williams, R. O. N., Rea, A., Vette, A., Croghan, C., Whitaker, D., Stevens, C., … others. (2009). The design and field implementation of the Detroit Exposure and Aerosol Research Study. *Journal of Exposure Science & Environmental Epidemiology*, *19*(7), 643–659.

Wilton, D., Szpiro, A., Gould, T., & Larson, T. (2010). Improving spatial concentration estimates for nitrogen oxides using a hybrid meteorological dispersion / land use regression model in Los Angeles , CA and Seattle , WA. *Science of the Total Environment, The*, *408*(5), 1120–1130. http://doi.org/10.1016/j.scitotenv.2009.11.033

World Health Organization (WHO). (2000). *Air quality guidelines for Europe, 2nd edition*. http://doi.org/10.1007/BF02986808

Wu, C. Da, Chen, Y. C., Pan, W. C., Zeng, Y. T., Chen, M. J., Guo, Y. L., & Lung, S. C. C. (2017a). Land-use regression with long-term satellite-based greenness index and culture-specific sources to model PM2.5 spatial-temporal variability. *Environmental Pollution*, *224*, 148–157. http://doi.org/10.1016/j.envpol.2017.01.074

Wu, C. Da, Chen, Y. C., Pan, W. C., Zeng, Y. T., Chen, M. J., Guo, Y. L., & Lung, S. C. C. (2017b). Land-use regression with long-term satellite-based greenness index and culture-specific sources to model PM2.5 spatial-temporal variability. *Environmental Pollution*, *224*, 148–157. http://doi.org/10.1016/j.envpol.2017.01.074

Xu, H., Bechle, M. J., Wang, M., Szpiro, A. A., Vedal, S., Bai, Y., & Marshall, J. D. (2019). National PM2.5 and NO2 exposure models for China based on land use regression, satellite measurements, and universal kriging. *Science of the Total Environment*, *655*(2), 423–433. http://doi.org/10.1016/j.scitotenv.2018.11.125

Xu, R. (2015). Particuology Light scattering: A review of particle characterization applications. *Particuology*, *18*, 11–21. http://doi.org/10.1016/j.partic.2014.05.002

Xu, Y., Ren, C., Ma, P., Ho, J., Wang, W., Lau, K. K., … Ng, E. (2017). Urban morphology detection and computation for urban climate research. *Landscape and Urban Planning*, *167*(July), 212–224. http://doi.org/10.1016/j.landurbplan.2017.06.018

Yang, X., Zheng, Y., Geng, G., Liu, H., Man, H., Lv, Z., … de Hoogh, K. (2017). Development of PM2.5and NO2 models in a LUR framework incorporating satellite remote sensing and air quality model data in Pearl River Delta region, China. *Environmental Pollution*, *226*(2), 143–153. http://doi.org/10.1016/j.envpol.2017.03.079

Yin, L., Cheng, Q., Wang, Z., & Shao, Z. (2015). “Big data” for pedestrian volume: Exploring the use of Google Street View images for pedestrian counts. *Applied Geography*, *63*, 337–345. http://doi.org/10.1016/j.apgeog.2015.07.010

Young, M. T., Bechle, M. J., Sampson, P. D., Szpiro, A. A., Marshall, J. D., Sheppard, L., & Kaufman, J. D. (2016). Satellite-based NO2 and model validation in a national prediction model based on universal kriging and land-use regression. *Environmental Science & Technology*, *50*(7), 3686–3694.

Yuan, C., Ng, E., & Norford, L. K. (2014). Improving air quality in high-density cities by understanding the relationship between air pollutant dispersion and urban morphologies. *Building and Environment*, *71*(2), 245–258. http://doi.org/10.1016/j.buildenv.2013.10.008

Yuan, M., Song, Y., Huang, Y., Hong, S., & Huang, L. (2017). Exploring the Association between Urban Form and Air Quality in China. *Journal of Planning Education and Research*, 0739456X1771151. http://doi.org/10.1177/0739456X17711516

Zhan, Y., Luo, Y., Deng, X., Grieneisen, M. L., Zhang, M., & Di, B. (2018). Spatiotemporal prediction of daily ambient ozone levels across China using random forest for human exposure assessment. *Environmental Pollution*, *233*, 464–473. http://doi.org/10.1016/j.envpol.2017.10.029

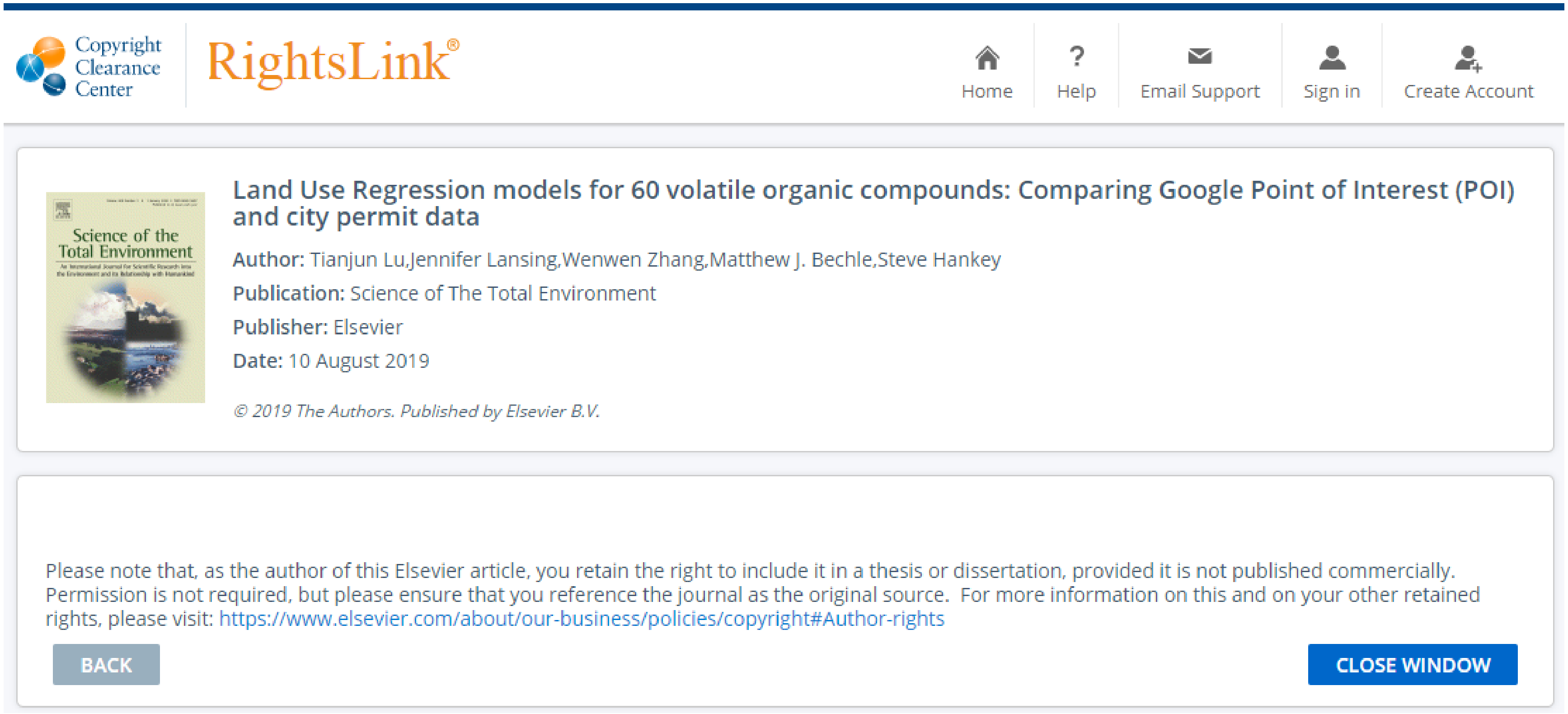
Zhang, H.-W., Kok, V. C., Chuang, S.-C., Tseng, C.-H., Lin, C.-T., Li, T.-C., … Hsu, C. Y. (2019). Long-term ambient hydrocarbons exposure and incidence of ischemic stroke. *PloS One*, *14*(12).

Zhang, Z., Wang, J., Hart, J. E., Laden, F., Zhao, C., Li, T., … Chen, K. (2018). National scale spatiotemporal land-use regression model for PM2.5 , PM10 and NO2 concentration in China. *Atmospheric Environment*, *192*, 48–54. http://doi.org/10.1016/j.atmosenv.2018.08.046

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., & Limited, S. G. (2017). Pyramid Scene Parsing Network. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zheng, Y., Liu, F., & Hsieh, H. (2013). U-Air: when urban air quality inference meets big data. *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’13*, 1436–1444. http://doi.org/10.1145/2487575.2488188

APPENDICES



Land Use Regression models for 60 volatile organic compounds: Comparing Google Point of Interest (POI) and city permit data

Author: Tianjun Lu,Jennifer Lansing,Wenwen Zhang,Matthew J. Bechle,Steve Hankey

Publication: Science of The Total Environment

Publisher: Elsevier

Date: 10 August 2019

© 2019 The Authors. Published by Elsevier B.V.

Please note that, as the author of this Elsevier article, you retain the right to include it in a thesis or dissertation, provided it is not published commercially. Permission is not required, but please ensure that you reference the journal as the original source. For more information on this and on your other retained rights, please visit: https://www.elsevier.com/about/our-business/policies/copyright#Author-rights

Table A3.1 Summary of existing VOC LUR model literature

Author	VOC specie	Number of sampling locations	City	Sampling period	Model R^2/Adj-R^2	Major predictors (direction of association)
Aguilera et al., 2008	BTEX	55	Sabadell, Spain	four 1-week; Apr 2005-March 2006	0.74/N/A	Altitude (-); Distance to road/parking lot (-)
Amini et al., 2017	Benzene, toluene, ethylbenzene, m/p-xylene and o-xylene, and BTEX	179	Tehran, Iran	three 2-week; Apr 2015-May 2016	0.64-0.70/0.62-0.68	Road (+); proximity to bus terminals (-); sensitive land use (-); distance to sewage treatment plants/gas filling stations (-)
Atari and Luginaah, 2009	Benzene, toluene, ethylbenzene, m/p-xylene and o-xylene, and BTEX	39	Sarnia, Canada	2 weeks; Oct 2005	0.81/0.79	Industry (+); highway (+); dwelling (+)
Carr et al., 2002	Benzene, toluene, and ethylbenzene	34: 18 traffic and 16 school sites	Munich, Germany	twelve 4-week; Dec 1996-Feb 1998	0.76-0.80/N/A	Traffic counts (N/A)
Fernandez-Somoano et al., 2011	Benzene	67	Asturias, Spain	two 1-week; Jun-Nov 2005	0.73/N/A	Altitude (-); continuous urban land cover (-); agriculture land use (-); proximity to road (+)
Gaeta et al., 2016	Benzene, toluene, acrolein, and formaldehyde	43: The Ciampino Airport	Rome, Italy	two 2-week; May-June 2011 and Jan 2012	0.29-0.57/0.24-0.53	The North latitude (+); product of traffic intensity of the nearest road and the inverse of distance to the nearest road (+); number of inhabitants (+)
Hystad et al., 2011	Benzene, ethylbenzene, and 1,3-butadiene	53	Entire Canada	Entire year of 2006	0.62-0.68/N/A	Major road length (+); population (+); highway (+); commercial land use (+); national pollutant release emissions (+)
Johnson et al., 2010	Benzene	25-285(pseudo measurements)	New Haven, USA	Jul-Aug 2001	N/A/0.67-0.89	Traffic intensity (N/A); proximity to roads/industrial sources (N/A)
Kheirbek et al., 2012	Benzene, BTEX, and formaldehyde	69	New York, USA	five 2-week; March-June 2011	0.65-0.83/N/A	Number of traffic signals (+); length of highway (+)

Mukerjee et al., 2009, 2012	Benzene, toluene, ethylbenzene, m/p-xylene and o-xylene, BTEX, styrene, and 1,3-butadiene	25: schools	Detroit, USA	one 5-week; Summer 2005	0.31-0.63/N/A	Proximity to nearest medium traffic road (-); traffic intensity (-); population density (+); distance to international border (+)
Oiamo et al., 2015	Benzene, toluene, and m/p-xylene	42	Ottawa, Canada	two 2-week; Oct 2008 and May 2009	0.75-0.79/N/A	Population (+); length of highway (+); distance to VOC facility (-); intersection count (+)
Poirier et al., 2015	Benzene and toluene	50	Halifax, Canada	two 2-week; Nov, Dec 1999	0.61-0.63/N/A	N/A
Smith et al., 2006	Benzene	22: schools	El Paso, USA	two 1-week; Nov, Dec 1999	0.93/N/A	Population density (+); proximity to international border crossing/petroleum facility (-)
Smith et al., 2011	Benzene, toluene, ethylbenzene, m/p-xylene and o-xylene, and 1,3-butadiene	24: fire stations	Dallas, USA	one 5-week; Aug-Sep 2006	0.41-0.72/N/A	Distance to nearest road (-/+); traffic intensity (+)
Su et al., 2010	Benzene, n-hexane, and total hydrocarbons	50	Toronto, Canada	one 2-week; Jul-Aug 2006	0.66-0.68/N/A	Expressway (+); major road (+); industrial and commercial land use (+); open land (-)
Wheeler et al., 2008	Benzene, toluene	54	Windsor, Canada	2 weeks; Feb, May, Aug, Oct 2005	0.46-0.73/N/A	Lengths of major roads/highways (+); VOC emission point sources (+)

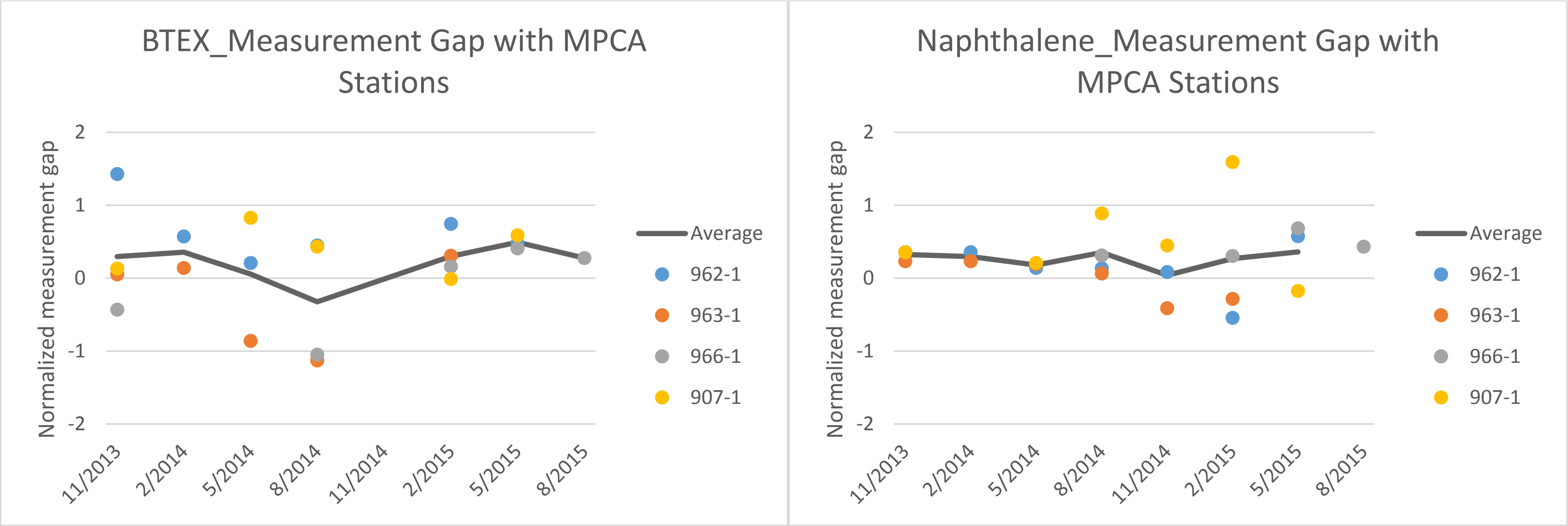


Figure A3.1. Measurement gap between the co-located community-based monitors and MPCA monitors. We attempted to normalize each sampling event by dividing by the average concentration of all available sampling events from November 2013 to August 2015 at each of the four MPCA monitoring stations (co-located by the Minneapolis samplers); however, the Minneapolis campaign had a 2-day sampling event gap with the MPCA campaign and was not available for all 8 events at these co-located stations. We then calculated the gap between the normalized values during each event of the two campaign. The normalized average measurement gap is 23% for priority VOCs (BTEX: 21%, naphthalene: 26%). 962-1, 963-1, 966-1 and 907-1 are the four co-located sampling/monitoring location IDs. Since tetrachloroethene measurements were not available (with invalid values) at MPCA monitors, we were not able to calculate the measurement gaps for tetrachloroethene and TVOC.

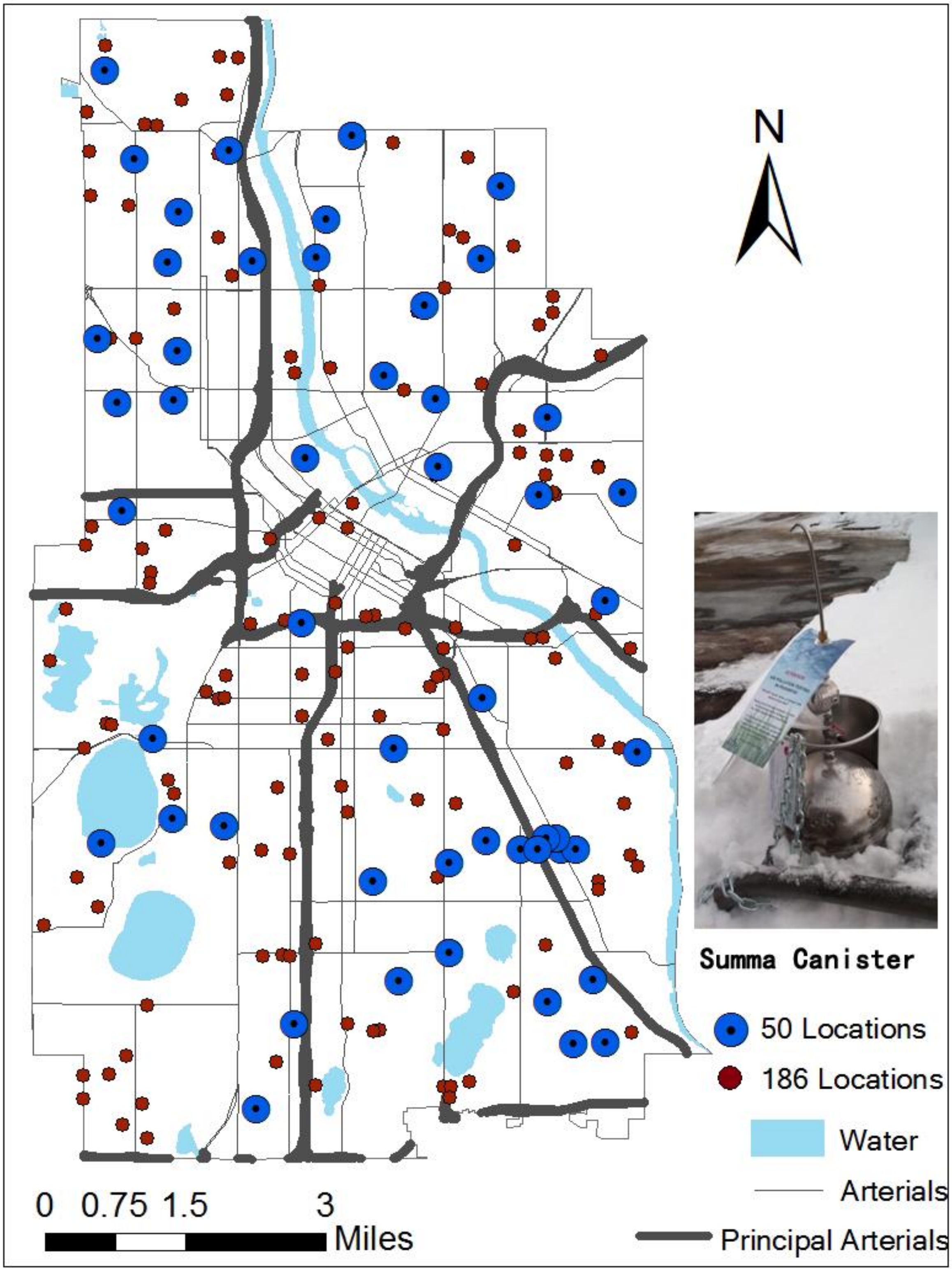


Figure A3.2. Sampling locations available during core modeling period (Second year: S5-S8 [n = 50]) and for all monitored locations (n=186) in Minneapolis, MN. Samples were collected using Summa Canister at each location.

Table A3.2 Summary statistics for all 60 VOCs during the core modeling period (Second year: S5-S8)[a]

VOC	Mean[b]	Median[b]	Max[b]	Min[b]	IQR[b]
1,1,1-Trichloroethane	0.18	0.16	0.47	0.13	0.14-0.18
1,1,2,2-Tetrachloroethane	0.26	0.23	0.75	0.18	0.20-0.26
1,1,2-Trichloroethane	0.23	0.20	0.76	0.16	0.18-0.23
1,1,2-Trichlorotrifluoroethane	0.22	0.19	0.59	0.15	0.17-0.23
1,1-Dichloroethane	0.14	0.12	0.44	0.10	0.11-0.14
1,1-Dichloroethene	0.16	0.14	0.36	0.11	0.13-0.16
1,2,4-Trichlorobenzene	0.69	0.53	5.19	0.42	0.48-0.62
1,2,4-Trimethylbenzene	0.54	0.28	9.17	0.08	0.13-0.52
1,2-Dibromoethane (EDB)	0.46	0.41	0.92	0.33	0.38-0.47
1,2-Dichlorobenzene	0.29	0.27	0.57	0.22	0.24-0.31
1,2-Dichloroethane	0.15	0.13	0.40	0.11	0.12-0.15
1,2-Dichloropropane	0.20	0.17	0.52	0.14	0.15-0.20
1,3,5-Trimethylbenzene	0.26	0.16	3.09	0.13	0.14-0.22
1,3-Butadiene	0.12	0.11	0.30	0.09	0.10-0.12
1,3-Dichlorobenzene	0.35	0.31	0.82	0.25	0.28-0.35
1,4-Dichlorobenzene	0.32	0.29	0.72	0.23	0.26-0.33
2-Butanone (MEK)	2.31	2.05	6.28	0.55	1.56-2.95
2-Hexanone	0.45	0.41	1.25	0.21	0.30-0.53
2-Propanol	3.97	0.84	119.43	0.11	0.42-1.31
4-Ethyltoluene	0.27	0.15	3.49	0.12	0.13-0.22
4-Methyl-2-pentanone (MIBK)	0.38	0.22	5.28	0.13	0.16-0.32
Acetone	11.63	9.53	42.16	6.08	8.59-11.53
Benzene	0.57	0.47	2.76	0.23	0.36-0.65
Benzyl chloride	0.35	0.28	1.51	0.22	0.24-0.33
Bromodichloromethane	0.18	0.16	0.57	0.12	0.14-0.18
Bromoform	0.59	0.50	2.74	0.41	0.45-0.58
Bromomethane	0.28	0.24	0.85	0.19	0.21-0.27
Carbon disulfide	0.50	0.09	10.41	0.06	0.06-0.24
Carbon tetrachloride	0.34	0.28	1.10	0.17	0.22-0.37
Chlorobenzene	0.30	0.15	1.67	0.08	0.09-0.34
Chloroethane	0.17	0.15	0.51	0.12	0.13-0.17
Chloroform	0.19	0.15	0.56	0.12	0.14-0.18
Chloromethane	0.68	0.67	1.17	0.31	0.58-0.78
cis-1,2-Dichloroethene	0.21	0.18	0.63	0.15	0.16-0.21
cis-1,3-Dichloropropene	0.23	0.20	0.50	0.17	0.19-0.24
Cyclohexane	0.64	0.24	14.90	0.15	0.18-0.33
Dibromochloromethane	0.82	0.71	2.69	0.55	0.62-0.80
Dichlorodifluoromethane	2.14	2.15	2.85	1.43	1.98-2.30

Dichlorotetrafluoroethane	0.26	0.23	0.80	0.18	0.20-0.26
Ethanol	9.26	7.03	44.25	2.59	5.71-8.92
Ethyl acetate	1.92	0.21	83.91	0.16	0.19-0.25
Ethylbenzene	0.38	0.28	2.58	0.20	0.23-0.38
Hexachloro-1,3-butadiene	0.54	0.43	3.17	0.35	0.38-0.51
m&p-Xylene	1.08	0.87	4.15	0.36	0.66-1.18
Methyl-tert-butyl ether	0.18	0.16	0.36	0.13	0.15-0.18
Naphthalene	0.73	0.55	6.12	0.19	0.35-0.84
n-Heptane	1.05	0.51	17.30	0.15	0.27-0.96
n-Hexane	2.79	1.37	33.21	0.17	0.78-2.97
o-Xylene	0.48	0.37	1.47	0.26	0.31-0.53
Propylene	0.60	0.54	2.05	0.09	0.38-0.76
Styrene	0.17	0.14	0.76	0.11	0.12-0.16
Tetrachloroethene	5.19	0.39	184.14	0.16	0.19-1.36
Tetrahydrofuran	0.15	0.11	1.08	0.08	0.09-0.12
Toluene	2.51	1.41	25.33	0.67	1.15-2.32
trans-1,2-Dichloroethene	0.25	0.22	0.59	0.18	0.20-0.26
trans-1,3-Dichloropropene	0.21	0.18	0.50	0.14	0.17-0.21
Trichloroethene	0.26	0.21	0.66	0.15	0.18-0.28
Trichlorofluoromethane	1.24	1.27	1.95	0.69	1.37-1.95
Vinyl acetate	0.59	0.45	1.69	0.22	0.75-1.69
Vinyl chloride	0.13	0.12	0.32	0.10	0.13-0.32
BTEX	5.02	3.64	27.61	1.97	2.88-5.01
TVOC	61.76	45.91	228.82	30.68	36.37-69.29

[a]Second year: S5-S8 (Nov 2014-Aug 2015); [b]All units are in $\mu g/m^3$.

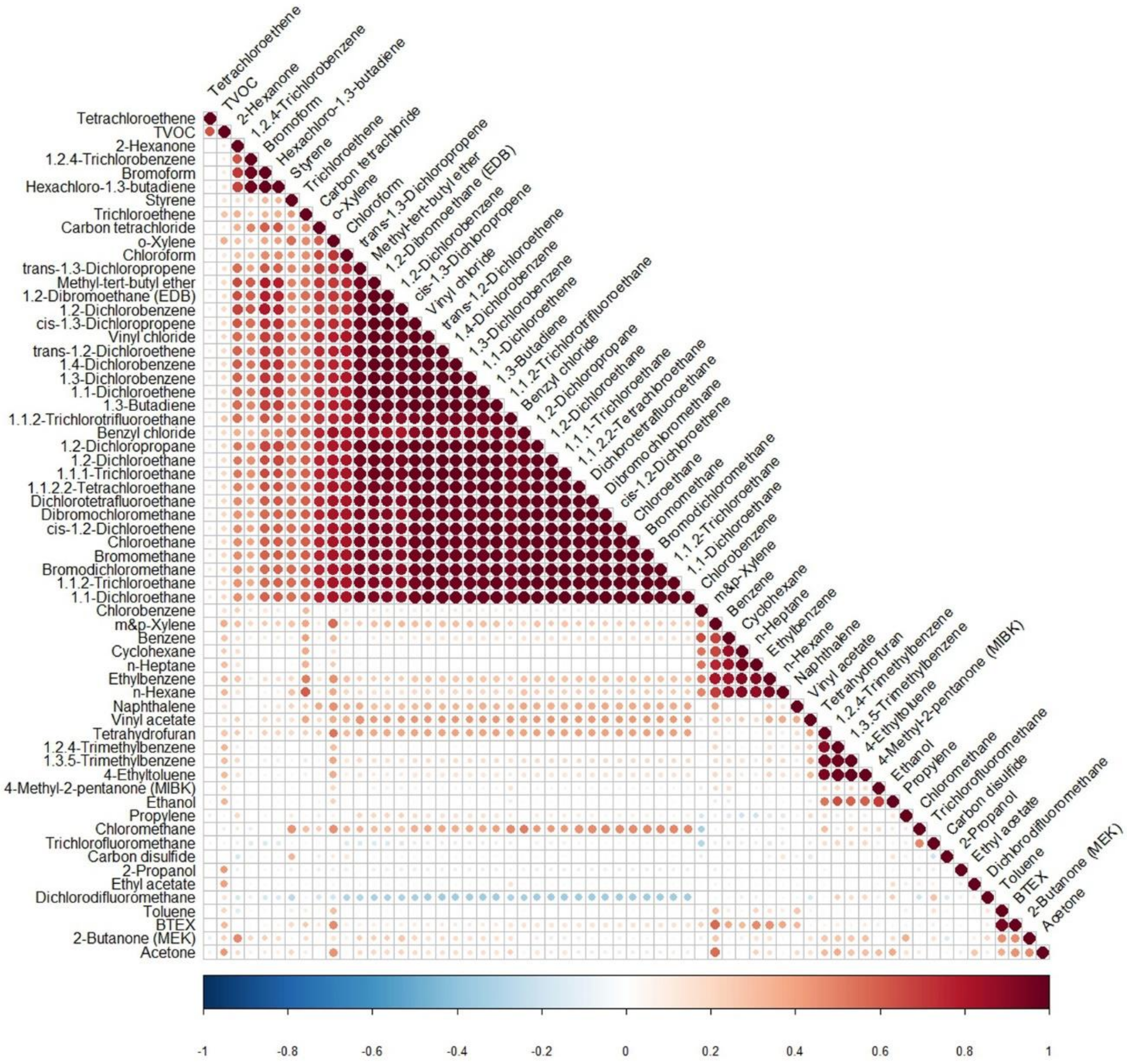


Figure A3.3. Correlation matrix for all VOCs during the core modeling period (Second year: S5-S8)

Table A3.3 Summary statistics of all 5 VOC annual-average scenarios of different sampling periods for the priority VOCs

VOC	Period	Number of Locations	Mean[f]	Median[f]	Max[f]	Min[f]	IQR[f]
BTEX	First year: S1-S4[a]	40	16.95	4.93	197.66	2.48	3.93-10.32
	Second year: S5-S8[b]	50	5.02	3.64	27.61	1.97	2.88-5.01
	Year-2014[c]	45	14.63	3.98	197.35	2.30	3.47-8.21
	8S[d]	24	13.57	5.57	99.96	2.58	3.32-7.90
	4S[e]	79	8.16	4.11	99.96	1.83	3.10-6.23
Naphthalene	First year: S1-S4	40	1.68	0.71	31.97	0.27	0.45-1.07
	Second year: S5-S8	50	0.73	0.54	6.12	0.19	0.35-0.84
	Year-2014	45	1.41	0.49	31.92	0.22	0.29-0.86
	8S	24	0.78	0.63	2.87	0.26	0.51-0.82
	4S	79	2.33	0.60	94.58	0.20	0.39-0.84
Tetrachloroethene	First year: S1-S4	40	5.04	0.27	181.13	0.17	0.19-0.58
	Second year: S5-S8	50	5.19	0.36	184.14	0.16	0.19-1.23
	Year-2014	45	5.51	0.61	176.99	0.15	0.20-1.48
	8S	24	9.00	0.83	182.63	0.17	0.28-1.33
	4S	79	3.33	0.55	182.63	0.16	0.23-1.17
TVOC	First year: S1-S4	40	131.83	62.05	1142.67	36.08	49.63-89.88
	Second year: S5-S8	50	61.76	45.91	228.82	30.68	36.37-69.29
	Year-2014	45	120.90	55.88	1140.84	32.46	46.31-102.20
	8S	24	100.52	74.96	375.66	38.35	46.95-112.14
	4S	79	85.48	52.44	1141.75	31.38	43.20-79.52

[a]First year: S1-S4 (Nov 2013-Aug 2014); [b]Second year: S5-S8 (Nov 2014-Aug 2015); [c]Year-2014: calendar year-2014 (Feb 2014-Nov 2014); [abc] are four consecutive sampling events (one event per season); [d]8S: measurements during all 8 sampling events; [e]4S: non-consecutive coverage of 4 seasons among the 8 sampling events; [f]All units are in µg/m^3.

Table A3.4 LUR model results for the 60 VOCs

Category	Variable	1,1,1-Trichloroethane			1,1,2,2-Tetrachloroethane			1,1,2-Trichloroethane			1,1,2-Trichlorotrifluoroethane			1,1-Dichloroethane		
		M1[a]	M2[b]	M3[c]	M1[a]	M2[b]	M3[c]	M1[a]	M2[b]	M3[c]	M1[a]	M2[b]	M3[c]	M1[a]	M2[b]	M3[c]
Area Sources: City Permit Data	Dry Cleaners															
	Gas Stations		0.63 (200)			0.67 (200)			0.73 (200)			0.71 (200)			0.74 (200)	
	Paint Booths															
	Auto Shops															
Area Sources: Google POI	Laundry															
	Gas Stations			0.35 (200)			0.36 (200)			0.40 (200)			0.33 (200)			0.40 (200)
	Painter			0.41 (400)			0.42 (400)			0.45 (400)			0.37 (400)			0.46 (400)
	Car Repair															
Transportation	Principal Arterials										0.64 (3,000)					
	Arterials					0.36 (200)			0.43 (200)		-1.53 (1,000)				0.43 (200)	
	Collectors															
	Local Roads										0.56 (25)					
	Dis. to Freeway															
	Dis. to Major Road										-0.32					
	Traffic Intensity															
	Transit Stops	-0.76 (2,000)		-0.32 (500)	-0.76 (2,000)		-0.34 (500)	-0.83 (2,000)		-0.37 (500)		-0.32 (100)		-0.84 (2,000)		-0.38 (500)
Land Use	Elevation															
	Industrial Area												0.17 (150)			
	Open Space		-0.33 (5,000)								-0.64 (5,000)	-0.46 (5,000)				
	Retail Area	0.45 (200)			0.44 (200)			0.47 (200)						0.47 (200)		
	Wtd. Household Income										-0.80 (2,000)					
	Wtd. Housing Dens.	0.56 (25)	0.52 (25)	0.45 (25)	0.57 (25)	0.43 (25)	0.46 (25)	0.65 (25)	0.48 (25)	0.53 (25)	0.70 (25)	0.57 (25)	0.47 (25)	0.64 (25)	0.48 (25)	0.53 (25)
Intercept		0.19	0.23	0.15	0.27	0.18	0.21	0.25	0.15	0.19	0.54	0.34	0.16	0.16	0.09	0.12
Adj-R^2		0.40	0.45	0.68	0.40	0.46	0.69	0.35	0.42	0.70	0.52	0.43	0.80	0.34	0.41	0.70
RMSE		0.04	0.04	0.03	0.06	0.06	0.05	0.07	0.07	0.05	0.05	0.05	0.03	0.05	0.04	0.03
10-fold CV-R^2		0.29	0.34	0.51	0.29	0.38	0.55	0.25	0.34	0.55	0.41	0.33	0.61	0.27	0.32	0.61

Category	Variable	1,1-Dichloroethene			1,2,4-Trichlorobenzene			1,2,4-Trimethylbenzene			1,2-Dibromoethane (EDB)			1,2-Dichlorobenzene		
		M1[a]	M2[b]	M3[c]	M1[a]	M2[b]	M3[c]	M1[a]	M2[b]	M3[c]	M1[a]	M2[b]	M3[c]	M1[a]	M2[b]	M3[c]
Area Sources: City Permit Data	Dry Cleaners															
	Gas Stations		0.53 (200)									0.39 (200)			0.34 (200)	
	Paint Booths															
	Auto Shops														0.05 (200)	
Area Sources: Google POI	Laundry															
	Gas Stations			0.29 (200)						0.80 (300)			0.21 (200)			0.20 (200)
	Painter			0.33 (500)									0.16 (500)			0.33 (1,000)
	Car Repair															
Transportation	Principal Arterials															
	Arterials															
	Collectors															
	Local Roads				0.47 (25)	0.47 (25)	0.47 (25)							0.25 (25)		0.23 (100)
	Dis. to Freeway															
	Dis. to Major Road															
	Traffic Intensity															
	Transit Stops															
Land Use	Elevation															
	Industrial Area															
	Open Space		-0.29 (5,000)													
	Retail Area															
	Wtd. Household Income															
	Wtd. Housing Dens.	0.36 (25)	0.43 (25)	0.33 (25)							0.26 (25)	0.27 (25)	0.24 (25)	0.25 (25)	0.25 (25)	
Intercept		0.13	0.20	0.12	0.39	0.39	0.39	0.33	0.33	0.28	0.35	0.34	0.33	0.18	0.23	0.20
Adj-R^2		0.34	0.51	0.61	0.09	0.09	0.09	0.00	0.00	0.17	0.21	0.52	0.45	0.22	0.55	0.44
RMSE		0.04	0.03	0.03	0.21	0.21	0.21	0.35	0.35	0.32	0.08	0.06	0.06	0.05	0.04	0.05
10-fold CV-R^2		0.28	0.41	0.50	0.06	0.06	0.07	0.03	0.01	0.08	0.16	0.41	0.38	0.18	0.46	0.39

Category	Variable	1,2-Dichloroethane			1,2-Dichloropropane			1,3,5-Trimethylbenzene			1,3-Butadiene			1,3-Dichlorobenzene		
		M1[a]	M2[b]	M3[c]	M1[a]	M2[b]	M3[c]	M1[a]	M2[b]	M3[c]	M1[a]	M2[b]	M3[c]	M1[a]	M2[b]	M3[c]
Area Sources: City Permit Data	Dry Cleaners															
	Gas Stations		0.65 (200)			0.61 (200)						0.60 (200)			0.52 (200)	
	Paint Booths															
	Auto Shops															
Area Sources: Google POI	Laundry															
	Gas Stations			0.36 (200)			0.36 (200)			0.69 (300)			0.33 (200)			0.29 (200)
	Painter			0.40 (400)			0.48 (400)						0.40 (400)			0.19 (500)
	Car Repair						-0.14 (300)									
Transportation	Principal Arterials															
	Arterials							0.74 (300)	0.74 (300)							
	Collectors															
	Local Roads															
	Dis. to Freeway															
	Dis. to Major Road															
	Traffic Intensity															
	Transit Stops	-0.77 (2,000)		-0.32 (500)	-0.74 (2,000)						-0.74 (2,000)		-0.31 (500)			
Land Use	Elevation															
	Industrial Area															
	Open Space		-0.33 (5,000)			-0.32 (5,000)						-0.31 (5,000)			-0.28 (5,000)	
	Retail Area	0.45 (200)			0.43 (200)						0.43 (200)					
	Wtd. Household Income															
	Wtd. Housing Dens.	0.57 (25)	0.54 (25)	0.46 (25)	0.54 (25)	0.51 (25)	0.41 (25)				0.53 (25)	0.49 (25)	0.42 (25)	0.36 (25)	0.42 (25)	0.33 (25)
Intercept		0.16	0.20	0.13	0.21	0.25	0.15	0.13	0.13	0.18	0.13	0.16	0.11	0.27	0.40	0.25
Adj-R^2		0.41	0.46	0.68	0.31	0.35	0.62	0.08	0.08	0.15	0.40	0.34	0.60	0.16	0.30	0.51
RMSE		0.04	0.04	0.03	0.05	0.05	0.04	0.19	0.19	0.18	0.03	0.03	0.02	0.08	0.07	0.06
10-fold CV-R^2		0.36	0.40	0.59	0.25	0.30	0.55	0.05	0.05	0.10	0.32	0.30	0.52	0.12	0.26	0.42

Category	Variable	4-Methyl-2-pentanone (MIBK)			Acetone			Benzene			Benzyl chloride			Bromodichloromethane		
		M1[a]	M2[b]	M3[c]	M1[a]	M2[b]	M3[c]	M1[a]	M2[b]	M3[c]	M1[a]	M2[b]	M3[c]	M1[a]	M2[b]	M3[c]
Area Sources: City Permit Data	Dry Cleaners															
	Gas Stations		-0.62 (5,000)									0.74 (200)			0.75 (200)	
	Paint Booths		0.28 (75)			0.19 (25)										
	Auto Shops															
Area Sources: Google POI	Laundry												0.29 (1,000)			
	Gas Stations			-0.86 (5,000)			0.40 (300)						0.43 (200)			0.40 (200)
	Painter												0.57 (400)			0.45 (400)
	Car Repair						0.13 (75)						-0.16 (300)			
Transportation	Principal Arterials				-0.51 (100)											
	Arterials				-0.30 (25)						0.45 (200)					
	Collectors															
	Local Roads				0.25 (2,000)	0.25 (1,000)										
	Dis. to Freeway															
	Dis. to Major Road															
	Traffic Intensity															
	Transit Stops	0.58 (25)	0.74 (25)	0.58 (25)				0.78 (300)	0.78 (300)	0.78 (300)				-0.84 (2,000)		-0.37 (500)
Land Use	Elevation															
	Industrial Area				0.43 (25)	0.23 (200)	0.18 (200)									
	Open Space												0.21 (200)			
	Retail Area		-0.57 (25)		0.60 (300)	0.28 (300)						0.49 (200)		0.48 (200)	0.44 (200)	
	Wtd. Household Income															
	Wtd. Housing Dens.										0.46 (25)	0.48 (25)	0.41 (25)	0.65 (25)	0.48 (25)	0.53 (25)
Intercept		0.23	0.45	0.59	1.86	1.96	2.33	0.31	0.31	0.31	0.20	0.18	0.16	0.19	0.11	0.15
Adj-R^2		0.10	0.54	0.24	0.43	0.41	0.56	0.16	0.16	0.16	0.31	0.47	0.77	0.35	0.42	0.70
RMSE		0.25	0.18	0.23	0.28	0.28	0.25	0.17	0.17	0.17	0.12	0.11	0.07	0.06	0.05	0.04
10-fold CV-R²		0.08	0.36	0.18	0.25	0.32	0.49	0.08	0.08	0.08	0.24	0.36	0.58	0.26	0.36	0.58

Category	Variable	Bromoform			Bromomethane			Carbon disulfide			Carbon tetrachloride			Chlorobenzene		
		M1[a]	M2[b]	M3[c]	M1[a]	M2[b]	M3[c]	M1[a]	M2[b]	M3[c]	M1[a]	M2[b]	M3[c]	M1[a]	M2[b]	M3[c]
Area Sources: City Permit Data	Dry Cleaners															
	Gas Stations					0.70 (200)										
	Paint Booths		0.43 (1,500)													
	Auto Shops											0.48 (3,000)				
Area Sources: Google POI	Laundry												0.14 (250)			
	Gas Stations			0.21 (200)			0.38 (200)						0.20 (200)			
	Painter			0.47 (750)			0.44 (400)						0.54 (1,500)			
	Car Repair															
Transportation	Principal Arterials										-0.35 (1,000)	-0.38 (1,000)	-0.42 (2,000)			
	Arterials					0.41 (200)							0.36 (400)			
	Collectors										0.31 (3,000)			-0.49 (1,000)	-0.49 (1,000)	-0.49 (1,000)
	Local Roads	0.42 (200)	0.48 (200)	0.33 (25)												
	Dis. to Freeway															
	Dis. to Major Road															
	Traffic Intensity															
	Transit Stops				-0.80 (2,000)		-0.35 (500)							0.39 (75)	0.39 (75)	0.39 (75)
Land Use	Elevation															
	Industrial Area															
	Open Space															
	Retail Area				0.46 (200)							-0.20 (25)	-0.23 (25)			
	Wtd. Household Income															
	Wtd. Housing Dens.				0.61 (25)	0.46 (25)	0.50 (25)				0.27 (25)	0.34 (25)	0.40 (2'000)			
Intercept		0.30	0.23	0.34	0.28	0.18	0.22	0.24	0.24	0.24	0.14	0.16	0.10	0.38	0.38	0.38
Adj-R^2		0.21	0.21	0.21	0.41	0.40	0.76	0.00	0.00	0.00	0.30	0.40	0.62	0.17	0.17	0.17
RMSE		0.14	0.14	0.14	0.07	0.07	0.04	0.43	0.43	0.43	0.11	0.10	0.08	0.20	0.20	0.20
10-fold CV-R^2		0.16	0.16	0.17	0.32	0.32	0.67	0.01	0.03	0.03	0.22	0.31	0.52	0.10	0.12	0.12

Category	Variable	Chloroethane			Chloroform			Chloromethane			cis-1,2-Dichloroethene			cis-1,3-Dichloropropene		
		M1[a]	M2[b]	M3[c]	M1[a]	M2[b]	M3[c]	M1[a]	M2[b]	M3[c]	M1[a]	M2[b]	M3[c]	M1[a]	M2[b]	M3[c]
Area Sources: City Permit Data	Dry Cleaners					-0.09 (200)										
	Gas Stations		0.72 (200)			0.91 (200)			-0.31 (5,000)			0.71 (200)			0.47 (200)	
	Paint Booths					-0.09 (50)										
	Auto Shops					0.39 (1,500)										
Area Sources: Google POI	Laundry															
	Gas Stations			0.39 (200)			0.22 (200)			0.83 (200)			0.38 (200)			0.26 (200)
	Painter			0.45 (400)									0.43 (400)			0.18 (500)
	Car Repair															
Transportation	Principal Arterials															
	Arterials		0.41 (200)			0.31 (200)		0.58 (200)	0.66 (200)	0.47 (200)		0.40 (200)				
	Collectors															
	Local Roads															
	Dis. to Freeway															
	Dis. to Major Road															
	Traffic Intensity															
	Transit Stops	-0.81 (2,000)		-0.36 (500)	-0.71 (2,000)	-0.27 (250)					-0.81 (2,000)		-0.36 (500)			
Land Use	Elevation															
	Industrial Area									0.15 (250)						
	Open Space							0.21 (150)		0.29 (150)					-0.26 (5,000)	
	Retail Area	0.46 (200)			0.66 (200)						0.46 (200)					
	Wtd. Household Income															
	Wtd. Housing Dens.	0.62 (25)	0.46 (25)	0.51 (25)	0.29 (25)	0.30 (25)	0.27 (25)				0.62 (25)	0.46 (25)	0.50 (25)	0.32 (25)	0.38 (25)	0.29 (25)
Intercept		0.18	0.11	0.14	0.21	0.10	0.14	0.46	0.55	0.44	0.22	0.14	0.17	0.19	0.28	0.17
Adj-R^2		0.42	0.41	0.68	0.35	0.67	0.40	0.26	0.26	0.37	0.39	0.40	0.77	0.14	0.28	0.48
RMSE		0.05	0.05	0.04	0.06	0.05	0.06	0.09	0.08	0.08	0.04	0.06	0.04	0.05	0.05	0.04
10-fold CV-R^2		0.32	0.35	0.60	0.31	0.59	0.33	0.20	0.20	0.30	0.31	0.31	0.65	0.10	0.21	0.42

Category	Variable	Cyclohexane			Dibromochloromethane			Dichlorodifluoromethane			Dichlorotetrafluoroethane			Ethanol		
		M1[a]	M2[b]	M3[c]	M1[a]	M2[b]	M3[c]	M1[a]	M2[b]	M3[c]	M1[a]	M2[b]	M3[c]	M1[a]	M2[b]	M3[c]
Area Sources: City Permit Data	Dry Cleaners															
	Gas Stations					0.66 (200)						0.69 (200)			-0.29 (1,500)	
	Paint Booths								-0.32 (1,500)							
	Auto Shops															
Area Sources: Google POI	Laundry									-0.32 (500)						
	Gas Stations						0.38 (200)						0.38 (200)			
	Painter			0.39 (200)			0.50 (400)						0.44 (400)			
	Car Repair						-0.15 (300)									
Transportation	Principal Arterials							0.40 (750)	0.44 (750)							
	Arterials					0.39 (200)			-0.29 (100)			0.39 (200)				
	Collectors															
	Local Roads															
	Dis. to Freeway															
	Dis. to Major Road							0.28								
	Traffic Intensity															
	Transit Stops									0.61 (750)	-0.79 (2,000)		-0.34 (500)			
Land Use	Elevation															
	Industrial Area							-0.41 (2,000)		-0.29 (2,000)						
	Open Space															
	Retail Area										0.46 (200)			0.57 (1,000)	0.69 (1,000)	0.57 (1,000)
	Wtd. Household Income															
	Wtd. Housing Dens.				0.47 (25)	0.42 (25)	0.44 (25)	-0.20 (25)		-0.27 (50)	0.60 (25)	0.45 (25)	0.49 (25)			
Intercept		0.33	0.33	0.26	0.52	0.46	0.49	1.15	1.16	1.10	0.27	0.17	0.21	1.89	2.10	1.89
Adj-R^2		0.00	0.00	0.11	0.39	0.39	0.76	0.30	0.22	0.37	0.41	0.39	0.75	0.16	0.22	0.16
RMSE		0.40	0.40	0.38	0.14	0.14	0.09	0.09	0.09	0.08	0.07	0.07	0.04	0.46	0.45	0.46
10-fold CV-R^2		0.00	0.01	0.08	0.31	0.32	0.58	0.22	0.16	0.33	0.35	0.36	0.62	0.08	0.18	0.13

Category	Variable	Ethyl acetate			Ethylbenzene			Hexachloro-1,3-butadiene			m&p-Xylene			Methyl-tert-butyl ether		
		M1^a	M2^b	M3^c	M1^a	M2^b	M3^c	M1^a	M2^b	M3^c	M1^a	M2^b	M3^c	M1^a	M2^b	M3^c
Area Sources: City Permit Data	Dry Cleaners															
	Gas Stations														0.39 (200)	
	Paint Booths											0.22 (250)				
	Auto Shops															
Area Sources: Google POI	Laundry															
	Gas Stations									0.22 (200)						0.22 (200)
	Painter						0.12 (300)			0.50 (1,000)			0.30 (250)			0.28 (500)
	Car Repair												0.05 (75)			
Transportation	Principal Arterials						0.62 (3,000)									
	Arterials															
	Collectors															
	Local Roads							0.40 (200)	0.40 (200)	0.33 (25)						
	Dis. to Freeway															
	Dis. to Major Road															
	Traffic Intensity															
	Transit Stops										0.33 (400)					
Land Use	Elevation															
	Industrial Area															
	Open Space															
	Retail Area															
	Wtd. Household Income										-0.37 (75)	-0.36 (75)	-0.30 (50)			
	Wtd. Housing Dens.													0.27 (25)	0.28 (25)	0.24 (25)
Intercept		0.30	0.30	0.30	0.31	0.31	0.16	0.24	0.24	0.26	0.72	0.84	0.78	0.15	0.15	0.14
Adj-R^2		0.00	0.00	0.00	0.00	0.00	0.31	0.06	0.06	0.25	0.23	0.29	0.44	0.20	0.56	0.47
RMSE		0.60	0.60	0.60	0.18	0.18	0.15	0.18	0.18	0.15	0.25	0.24	0.21	0.04	0.03	0.03
10-fold CV-R^2		0.01	0.01	0.02	0.02	0.06	0.25	0.05	0.05	0.20	0.18	0.22	0.34	0.06	0.38	0.38

Category	Variable	Naphthalene			n-Heptane			n-Hexane			o-Xylene			Propylene		
		M1^{a}	M2^{b}	M3^{c}	M1^{a}	M2^{b}	M3^{c}	M1^{a}	M2^{b}	M3^{c}	M1^{a}	M2^{b}	M3^{c}	M1^{a}	M2^{b}	M3^{c}
Area Sources: City Permit Data	Dry Cleaners															
	Gas Stations		0.66 (200)													
	Paint Booths		0.53 (500)			0.37 (25)										
	Auto Shops											0.17 (75)				
Area Sources: Google POI	Laundry															
	Gas Stations									0.71 (200)			0.93 (200)			
	Painter			0.76 (400)			0.37 (300)			0.33 (300)			0.19 (300)			
	Car Repair						0.18 (75)						0.06 (75)			
Transportation	Principal Arterials	-0.80 (500)	-0.68 (750)	-0.73 (500)			0.98 (3,000)									
	Arterials						-0.80 (1,500)				0.34 (200)	0.33 (250)	0.24 (250)		0.06 (200)	
	Collectors	0.60 (1,500)		0.50 (1,000)			-0.66 (1,000)									
	Local Roads															
	Dis. to Freeway															
	Dis. to Major Road													0.26	0.29	0.26
	Traffic Intensity															
	Transit Stops															
Land Use	Elevation															
	Industrial Area															
	Open Space															
	Retail Area															
	Wtd. Household Income													-0.30 (500)		-0.30 (500)
	Wtd. Housing Dens.										0.25 (25)	0.23 (25)	0.20 (25)			
Intercept		0.25	0.42	0.23	0.51	0.48	0.69	1.03	1.03	0.85	0.28	0.27	0.26	0.54	0.36	0.54
Adj-R^2		0.20	0.40	0.50	0.00	0.20	0.50	0.00	0.00	0.26	0.27	0.45	0.59	0.14	0.15	0.14
RMSE		0.27	0.23	0.21	0.48	0.43	0.34	0.66	0.66	0.57	0.14	0.12	0.10	0.19	0.19	0.19
10-fold CV-R^2		0.17	0.32	0.47	0.02	0.15	0.40	0.01	0.01	0.22	0.22	0.38	0.52	0.08	0.08	0.08

Category	Variable	Styrene			Tetrachloroethene			Tetrahydrofuran			Toluene			trans-1,2-Dichloroethene		
		M1[a]	M2[b]	M3[c]	M1[a]	M2[b]	M3[c]	M1[a]	M2[b]	M3[c]	M1[a]	M2[b]	M3[c]	M1[a]	M2[b]	M3[c]
Area Sources: City Permit Data	Dry Cleaners					0.39 (25)										
	Gas Stations								0.25 (250)						0.53 (200)	
	Paint Booths											0.35 (500)				
	Auto Shops											0.32 (50)				
Area Sources: Google POI	Laundry						0.41 (25)									
	Gas Stations			0.22 (200)			0.41 (200)			0.82 (300)						0.29 (200)
	Painter												0.55 (400)			0.32 (500)
	Car Repair			0.57 (1,000)									0.71 (1,000)			
Transportation	Principal Arterials						0.34 (300)									
	Arterials	0.68 (200)	0.68 (200)	0.60 (200)		0.30 (200)		0.65 (300)	0.62 (300)							
	Collectors											0.20 (75)	0.19 (75)			
	Local Roads			-0.24 (3,000)												
	Dis. to Freeway															
	Dis. to Major Road															
	Traffic Intensity															
	Transit Stops			-0.49 (250)									-0.38 (150)			
Land Use	Elevation															
	Industrial Area															
	Open Space														-0.28 (5,000)	
	Retail Area				0.68 (150)											
	Wtd. Household Income															
	Wtd. Housing Dens.					0.26 (25)	0.31 (25)							0.36 (25)	0.42 (25)	0.33 (25)
Intercept		0.12	0.12	0.23	0.41	0.24	0.26	0.09	0.08	0.11	1.05	0.82	0.62	0.20	0.30	0.18
Adj-R^2		0.26	0.26	0.66	0.31	0.64	0.75	0.08	0.14	0.29	0.00	0.40	0.44	0.34	0.31	0.52
RMSE		0.07	0.07	0.05	0.76	0.55	0.46	0.10	0.10	0.09	0.51	0.39	0.38	0.05	0.05	0.04
10-fold CV-R^2		0.18	0.21	0.51	0.26	0.40	0.56	0.05	0.10	0.21	0.02	0.31	0.35	0.25	0.25	0.44

Category	Variable	trans-1,3-Dichloropropene			Trichloroethene			Trichlorofluoromethane			Vinyl acetate			Vinyl chloride		
		M1[a]	M2[b]	M3[c]	M1[a]	M2[b]	M3[c]	M1[a]	M2[b]	M3[c]	M1[a]	M2[b]	M3[c]	M1[a]	M2[b]	M3[c]
Area Sources: City Permit Data	Dry Cleaners											0.71 (1,000)				
	Gas Stations		0.43 (200)						-0.41 (1,000)						0.58 (200)	
	Paint Booths		-0.10 (50)													
	Auto Shops		0.11 (400)													
Area Sources: Google POI	Laundry									-0.27 (500)						
	Gas Stations			0.23 (200)			0.44 (250)									0.32 (200)
	Painter			0.32 (500)												0.37 (400)
	Car Repair															
Transportation	Principal Arterials															
	Arterials		0.26 (200)							0.39 (300)						
	Collectors								-0.24 (200)							
	Local Roads	0.28 (300)										0.30 (25)				
	Dis. to Freeway															
	Dis. to Major Road															
	Traffic Intensity															
	Transit Stops		-0.33 (100)	-0.29 (400)												
Land Use	Elevation															
	Industrial Area															
	Open Space														-0.31 (5,000)	
	Retail Area						0.31 (5,000)									
	Wtd. Household Income															
	Wtd. Housing Dens.	0.31 (25)	0.31 (25)	0.31 (25)										0.41 (25)	0.47 (25)	0.37 (25)
Intercept		0.10	0.15	0.18	0.23	0.23	0.11	0.80	0.88	0.79	0.44	0.27	0.44	0.11	0.17	0.10
Adj-R^2		0.53	0.38	0.61	0.00	0.00	0.20	0.00	0.17	0.15	0.00	0.32	0.00	0.35	0.33	0.65
RMSE		0.04	0.05	0.04	0.10	0.10	0.09	0.10	0.10	0.10	0.23	0.19	0.23	0.03	0.03	0.02
10-fold CV-R^2		0.47	0.32	0.50	0.01	0.05	0.12	0.02	0.10	0.11	0.03	0.28	0.02	0.34	0.30	0.58

[a]M1: Base-case: No Area Sources; [b]M2: Area Sources: City Permit Data; [c]M3: Area Sources: Google POI; Model coefficients are normalized coefficients with buffers in parentheses. All variables are at $p < 0.05$. Number of locations used for modeling is 50.

Table A3.5 The average number of area source locations and coefficient of variation between the city permit data and Google POI data

Category	Area Sources: City Permit Data			Area Sources: Google POI		
	SD[a]	Average[b]	CV[c]	SD[a]	Average[b]	CV[c]
Dry Cleaners/Laundry	0.39	0.23	1.70	0.78	0.38	2.08
Gas Stations	0.87	0.68	1.29	1.02	0.73	1.40
Paint Booths/Painter	0.68	0.49	1.38	0.86	0.55	1.55
Auto Shops/Car Repair	3.99	2.54	1.57	5.43	3.29	1.65

[a]SD: standard deviation; [b]Average: average number of area source locations; [c]CV: coefficient of variation. The results were calculated for the 500m buffer scenario.

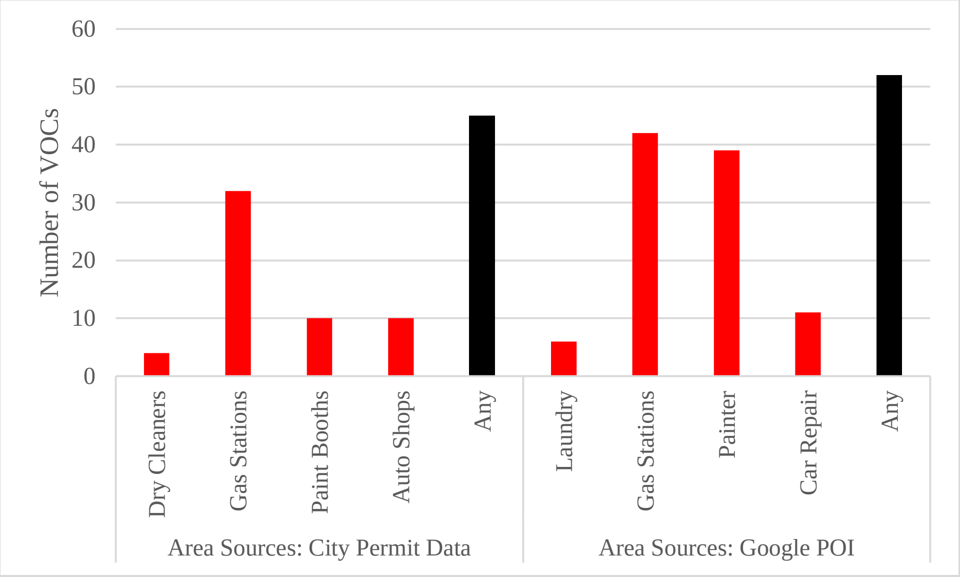


Figure A3.4. Number of VOCs that selected an area source in the core LUR models. The black bar "Any" presents number of VOCs that selected at least one type of area source.

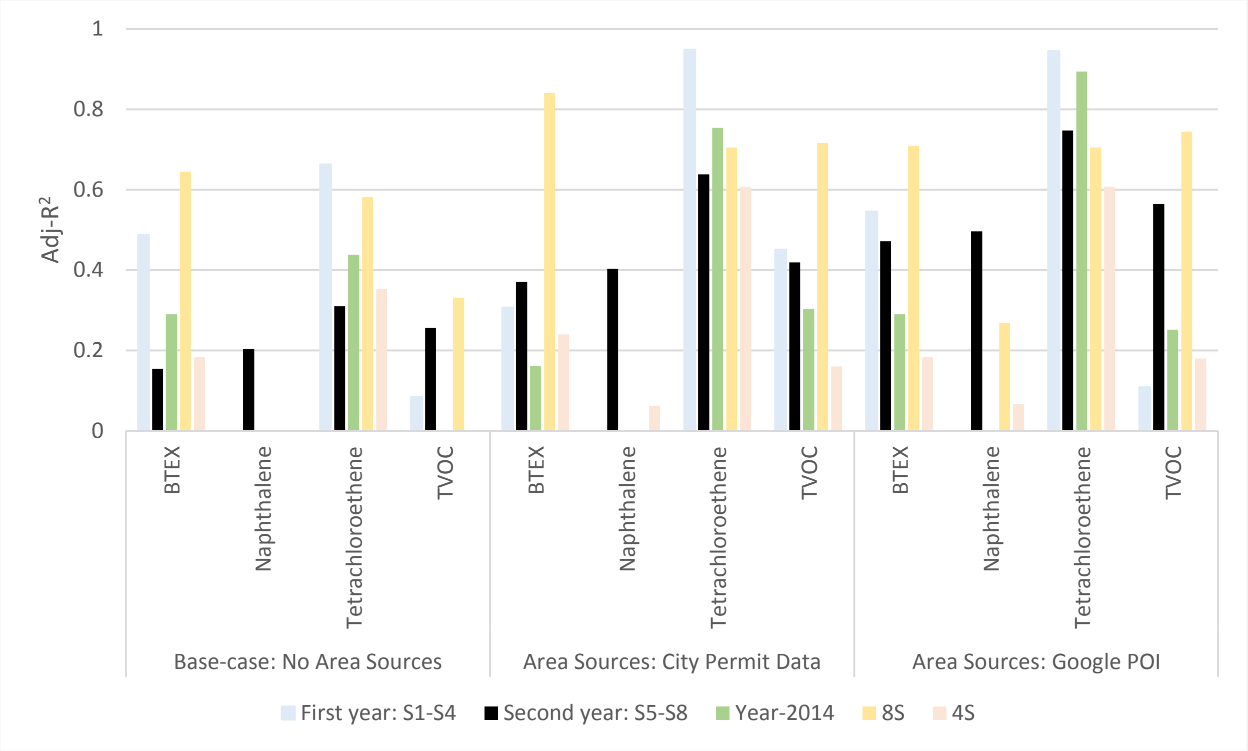


Figure A3.5. Model performance among the scenarios for estimating annual average VOC concentrations for the priority VOCs. The black bar (Second year: S5-S8) is our core model scenario.

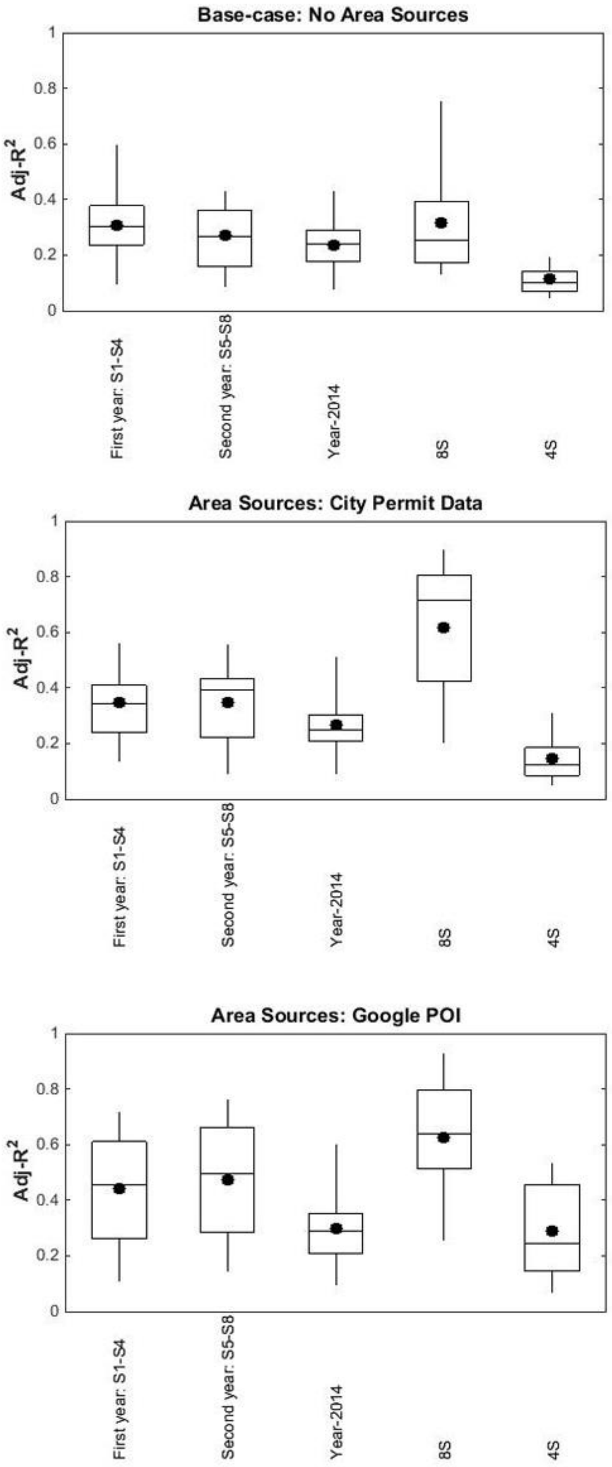


Figure A3.6. Model performance among the scenarios for estimating annual average VOC concentrations for all 60 VOCs. Black dots are means of model performance. First year: S1-S4 (Nov 2013-Aug 2014; n=40); Second year: S5-S8 (Nov 2014-Aug 2015; n=50); Year-2014: calendar year-2014 (Feb 2014-Nov 2014; n=24); 8S (measurements during all 8 sampling events; n=24); 4S (non-consecutive coverage of 4 seasons among the 8 sampling events; n=79).

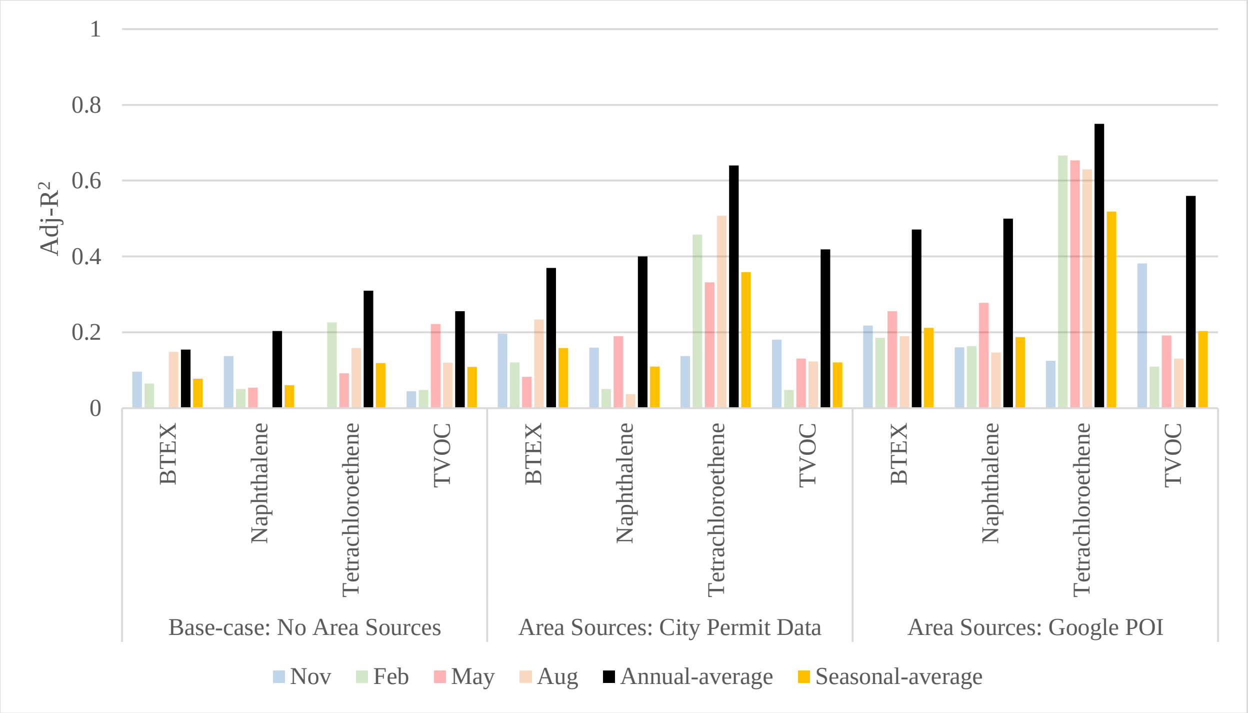


Figure A3.7. Adj-R^2 of seasonal models vs. annual models for the priority VOCs. VOC species without certain color-bars indicate no variables were selected in the seasonal models. Seasonal-average bar is the average value of the adj-R^2 of all LUR model using the VOC measurements of each season (if applicable); annual-average bar is the core LUR model performance using the VOC measurements of our second-year sampling events.

Table A3.6 Cook's distance of the Google POI model for the priority VOCs

VOC	Min	Median	Mean	Max
BTEX	0	0	0.03	0.81
Naphthalene	0	0.01	0.02	0.14
Tetrachloroethene	0	0	0.03	0.28
TVOC	0	0.01	0.02	0.10

Table A3.7 Moran's I results of the three models for the priority VOCs

Model Type	VOC	2500m	
		Moran's I	p-value
Base-case: No Area Sources	BTEX	0.26	0.01*
	Naphthalene	-0.03	0.90
	Tetrachloroethene	0.01	0.52
	TVOC	0.01	0.72
Area Sources: City Permit Data	BTEX	-0.17	0.07
	Naphthalene	-0.03	0.85
	Tetrachloroethene	-0.03	0.95
	TVOC	-0.05	0.76
Area Sources: Google POI	BTEX	-0.19	0.03*
	Naphthalene	-0.04	0.84
	Tetrachloroethene	-0.01	0.84
	TVOC	-0.07	0.60

* denotes p-value < 0.05. We set the threshold distance to 2500 meters to ensure a minimum number of neighbors to 1 based on inverse distance spatial relationships.

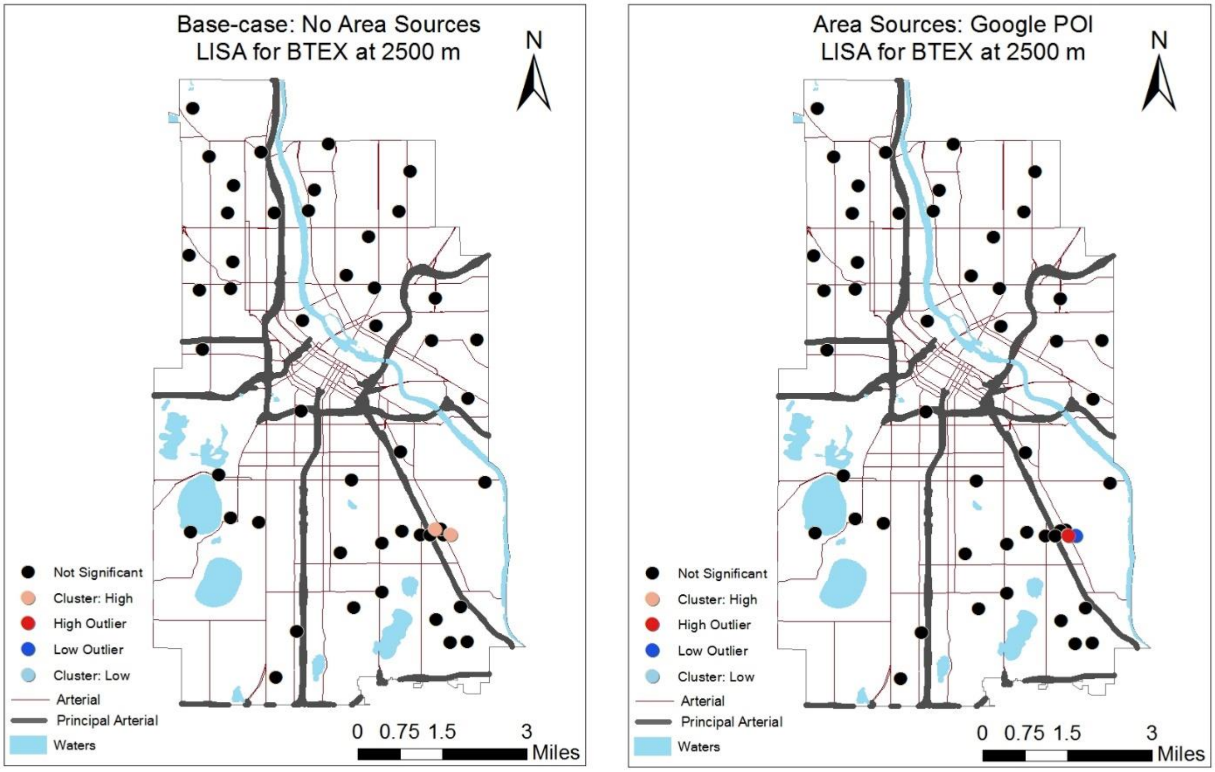


Figure A3.8. LISA results of BTEX for the models with spatial autocorrelation.

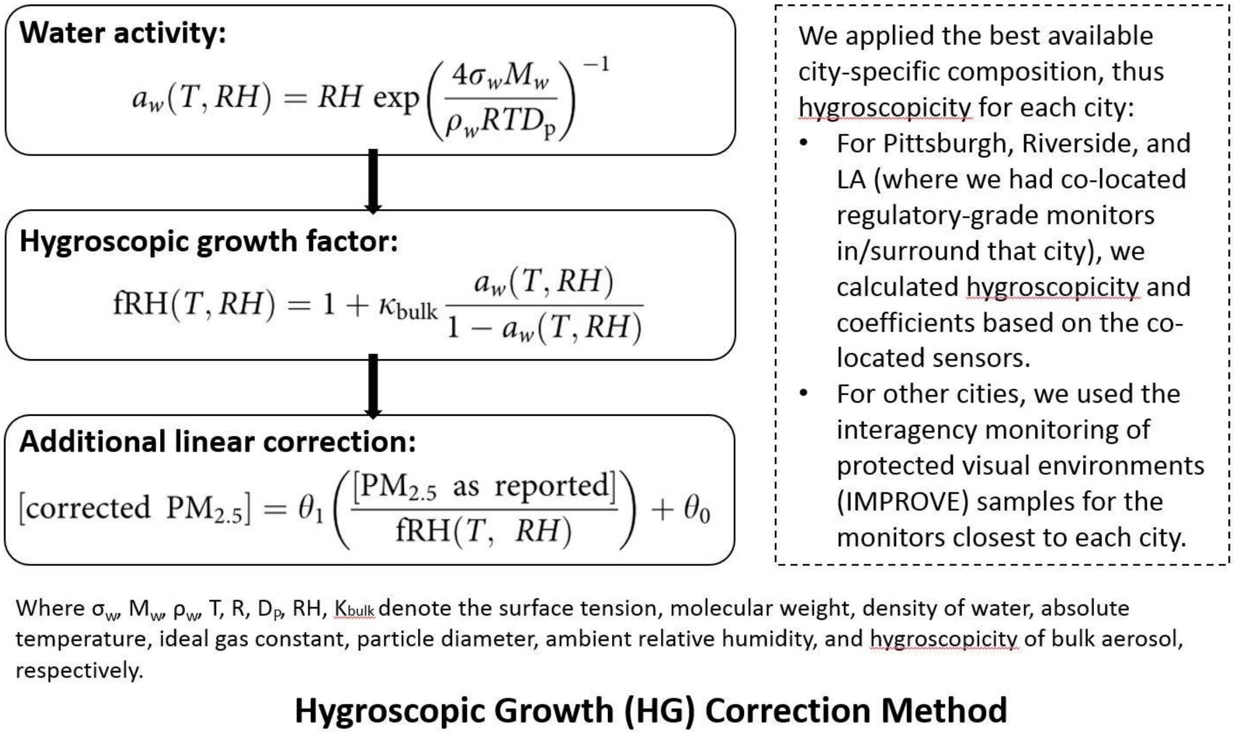
Table A4.1. Independent variables used in LUR models

Category	**Measure**	**Note**[a]
Traffic	Distance to the nearest road (0.05-15 km)	Any available road
Population	Sum (0.5-3 km)	Population in block groups
Land use (Urban)	Percent (0.05-15 km)	Urban or built-up land, etc.
Land use (Rural)	Percent (0.05-15 km)	Agriculture, forest, water, etc.
Source	Distance to the nearest source	Coastline, railroad, airport, etc.
Emission	Sum of cite-specific facility emissions (3-30 km)	$PM_{2.5}$
Vegetation	Quantiles (0.5-10 km)	Normalized Difference Vegetation Index
Imperviousness	Percent (0.05-5 km)	Impervious surface value
Elevation	Counts of points above/below a threshold (1-5 km)	Elevation value
Satellite estimate	Grid-level estimates	$PM_{2.5}$

[a] Detailed information can be found from the CACES LUR modeling study (Kim et al., 2020)

Table A4.2. Summary of obtaining valid PPA measurements

City	# of raw hours including both channels	# of valid hours of channel A data before filtering	# of valid hours of channel A data after filtering	# of valid sensors used
LA	8,663,963	2,579,804	2,360,159	103
New York	566,693	138,407	130,520	7
Phoenix	377,132	113,908	102,430	7
Pittsburgh	903,949	267,072	217,325	8
Riverside	3,548,497	1,066,588	697,999	16
DC	514,745	154,026	137,301	8



Water activity:

$$a_w(T, RH) = RH \exp\left(\frac{4\sigma_w M_w}{\rho_w RTD_\mathrm{p}}\right)^{-1}$$

Hygroscopic growth factor:

$$\mathrm{fRH}(T, RH) = 1 + \kappa_\mathrm{bulk} \frac{a_w(T, RH)}{1 - a_w(T, RH)}$$

Additional linear correction:

$$[\text{corrected PM}_{2.5}] = \theta_1 \left(\frac{[\text{PM}_{2.5} \text{ as reported}]}{\mathrm{fRH}(T,\ RH)}\right) + \theta_0$$

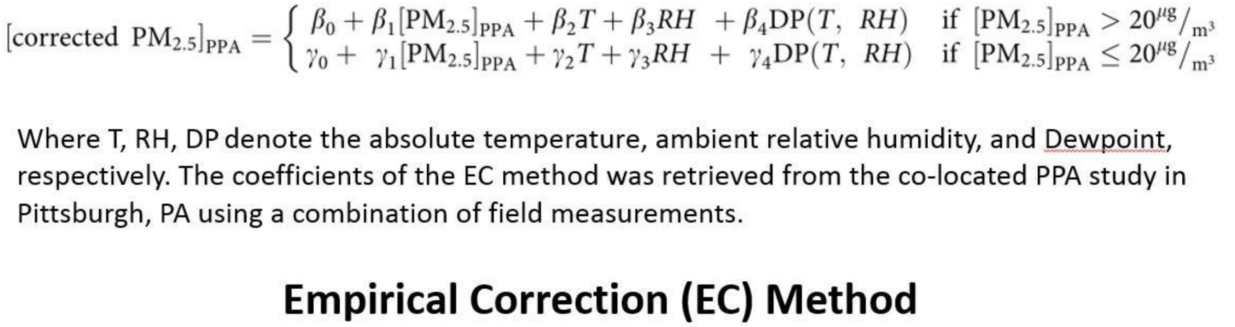
We applied the best available city-specific composition, thus hygroscopicity for each city:

- For Pittsburgh, Riverside, and LA (where we had co-located regulatory-grade monitors in/surround that city), we calculated hygroscopicity and coefficients based on the co-located sensors.
- For other cities, we used the interagency monitoring of protected visual environments (IMPROVE) samples for the monitors closest to each city.

Where σ_w, M_w, ρ_w, T, R, D_P, RH, K_{bulk} denote the surface tension, molecular weight, density of water, absolute temperature, ideal gas constant, particle diameter, ambient relative humidity, and hygroscopicity of bulk aerosol, respectively.

Hygroscopic Growth (HG) Correction Method

Figure A4.1. The overview of the Hygroscopic Growth (HG) Correction Method.



$$[\text{corrected PM}_{2.5}]_{\text{PPA}} = \begin{cases} \beta_0 + \beta_1[\text{PM}_{2.5}]_{\text{PPA}} + \beta_2 T + \beta_3 RH \ + \beta_4 \text{DP}(T,\ RH) & \text{if } [\text{PM}_{2.5}]_{\text{PPA}} > 20^{\mu\text{g}}/_{\text{m}^3} \\ \gamma_0 + \ \gamma_1[\text{PM}_{2.5}]_{\text{PPA}} + \gamma_2 T + \gamma_3 RH \ + \ \gamma_4 \text{DP}(T,\ RH) & \text{if } [\text{PM}_{2.5}]_{\text{PPA}} \leq 20^{\mu\text{g}}/_{\text{m}^3} \end{cases}$$

Where T, RH, DP denote the absolute temperature, ambient relative humidity, and Dewpoint, respectively. The coefficients of the EC method was retrieved from the co-located PPA study in Pittsburgh, PA using a combination of field measurements.

Empirical Correction (EC) Method

Figure A4.2. The overview of the Empirical Correction (EC) Correction Method.

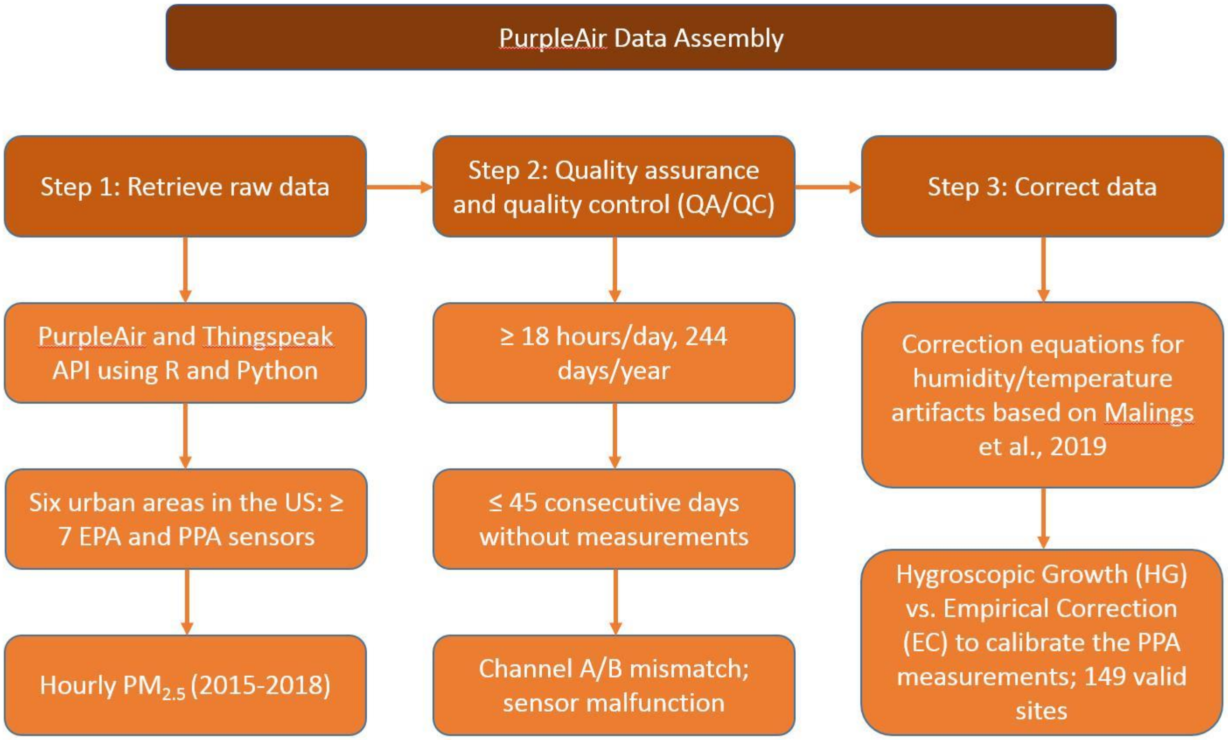

PurpleAir Data Assembly
Step 1: Retrieve raw data
Step 2: Quality assurance and quality control (QA/QC)
Step 3: Correct data
PurpleAir and Thingspeak API using R and Python
≥ 18 hours/day, 244 days/year
Correction equations for humidity/temperature artifacts based on Malings et al., 2019
Six urban areas in the US: ≥ 7 EPA and PPA sensors
≤ 45 consecutive days without measurements
Hygroscopic Growth (HG) vs. Empirical Correction (EC) to calibrate the PPA measurements; 149 valid sites
Hourly $PM_{2.5}$ (2015-2018)
Channel A/B mismatch; sensor malfunction

Figure A4.3. The PPA data assembly framework.

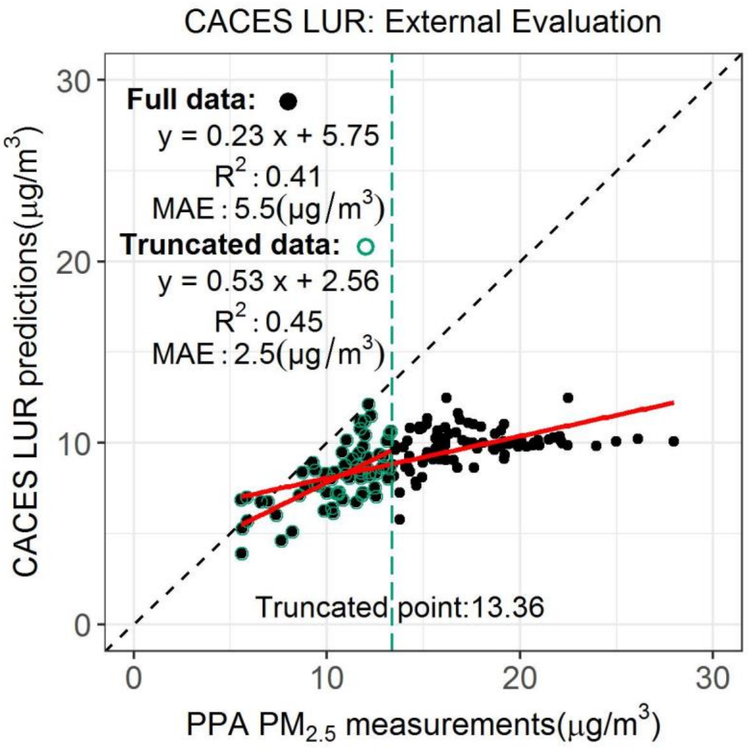


Figure A4.4. External evaluation of the CACES LUR using all 6-city PPA measurements (pooled evaluation). The truncated data was the PPA measurements below the maximum concentrations (13.36 μg/m^3) of the EPA monitors among the 6 cities.

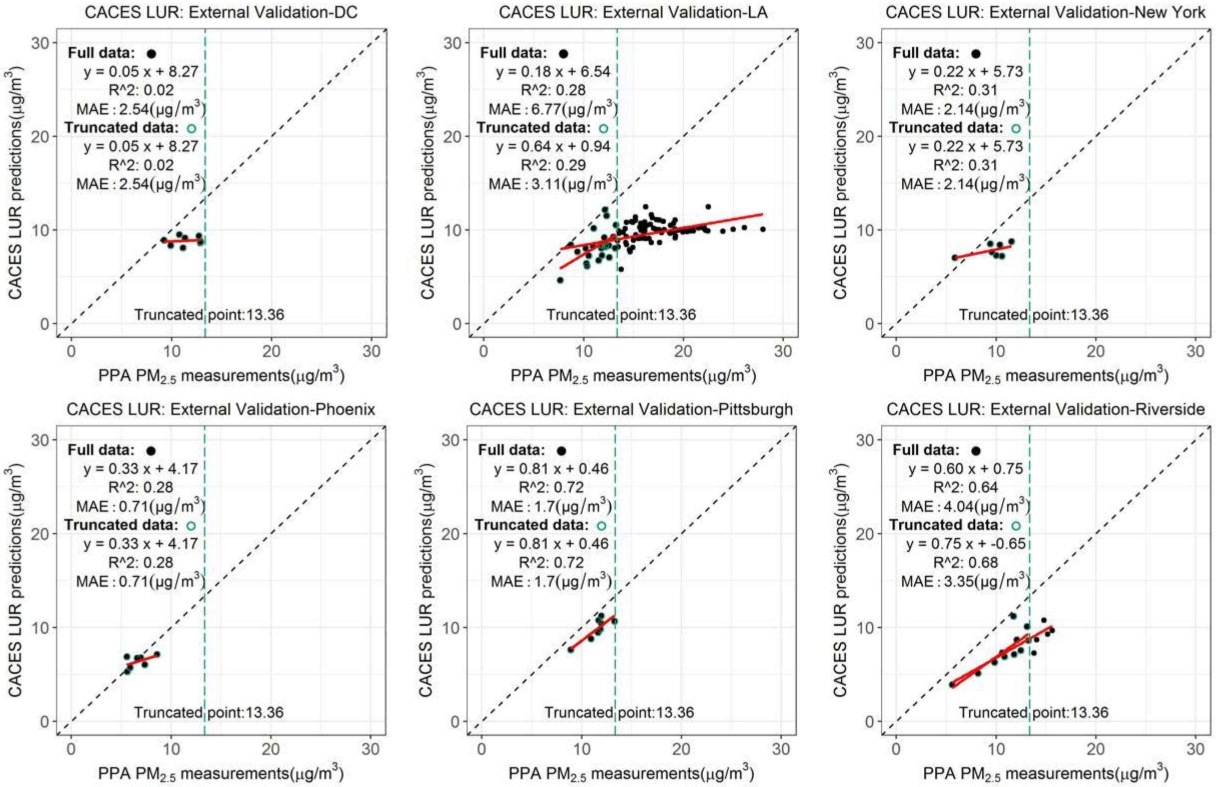


Figure A4.5. External evaluation of the CACES LUR by city (single-city evaluation). The truncated data was the PPA measurements below the maximum concentrations (13.36 µg/m^3) of the EPA monitors among the 6 cities.

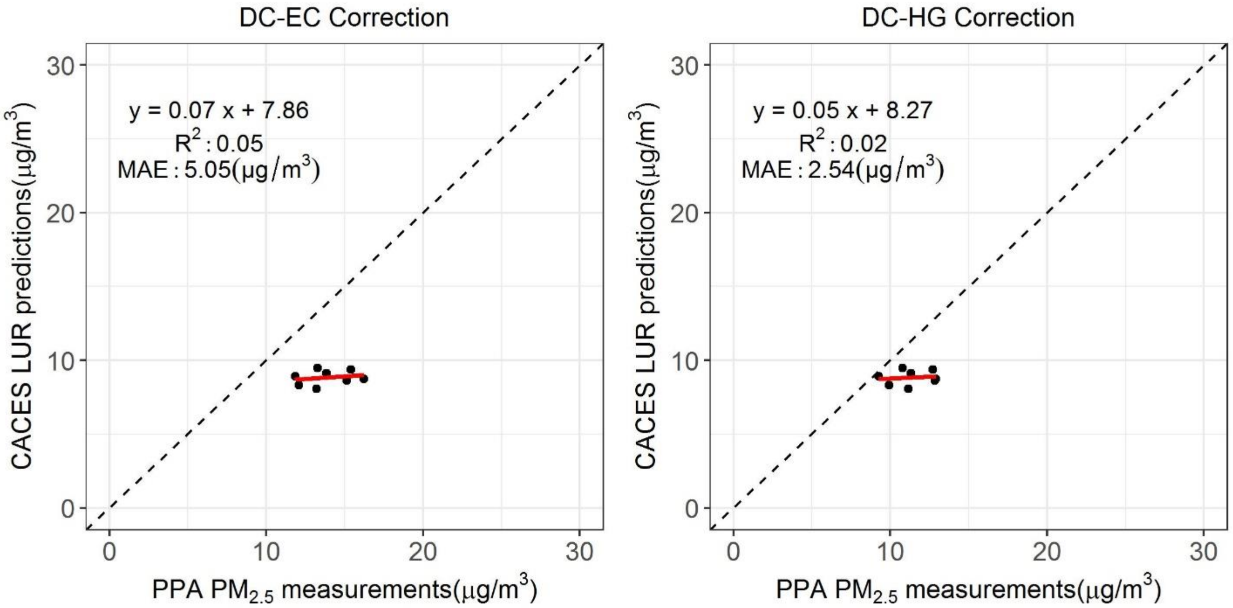


Figure A4.6. External evaluation of the CACES LUR using the HG vs. EC correction methods for DC.

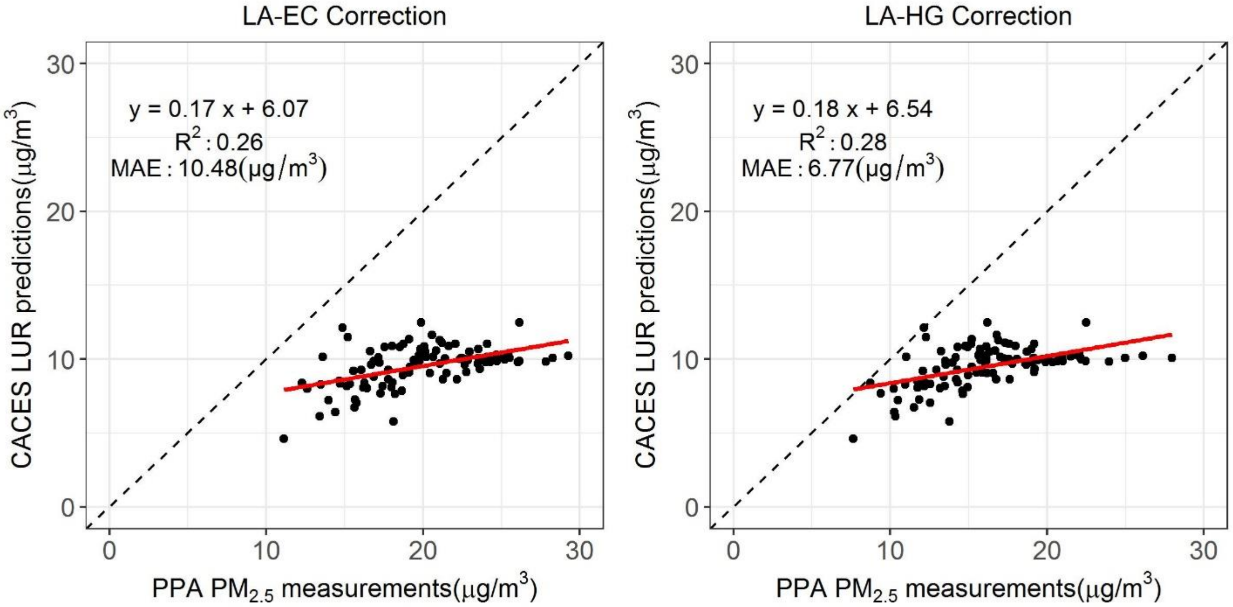


Figure A4.7. External evaluation of the CACES LUR using the HG vs. EC correction methods for LA.

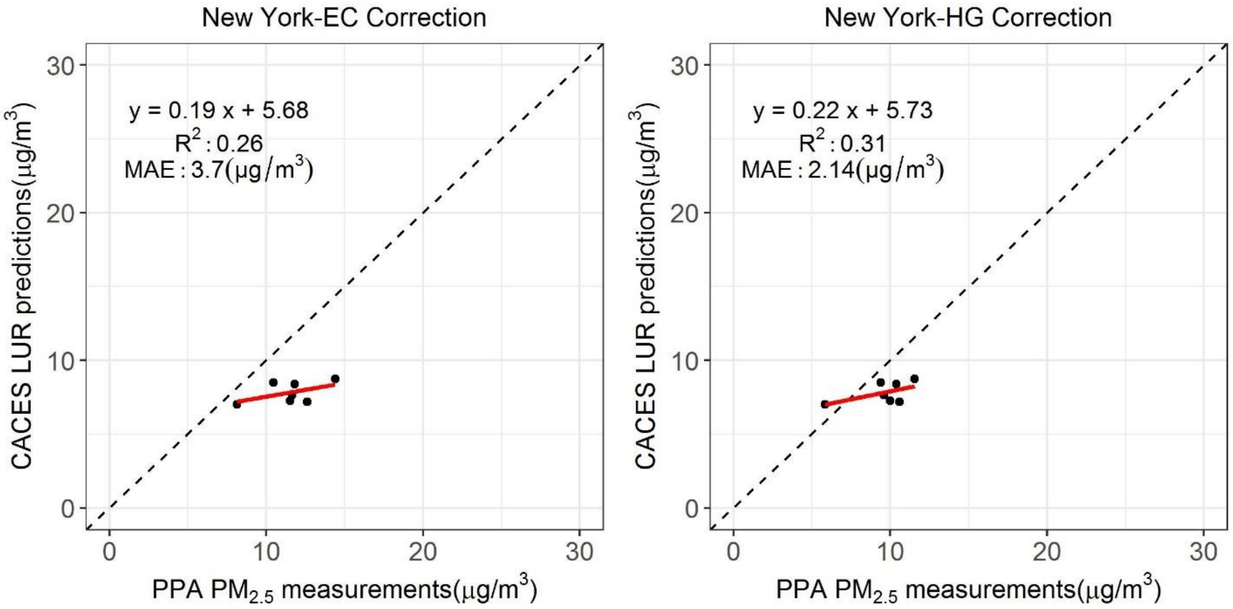


Figure A4.8. External evaluation of the CACES LUR using the HG vs. EC correction methods for New York.

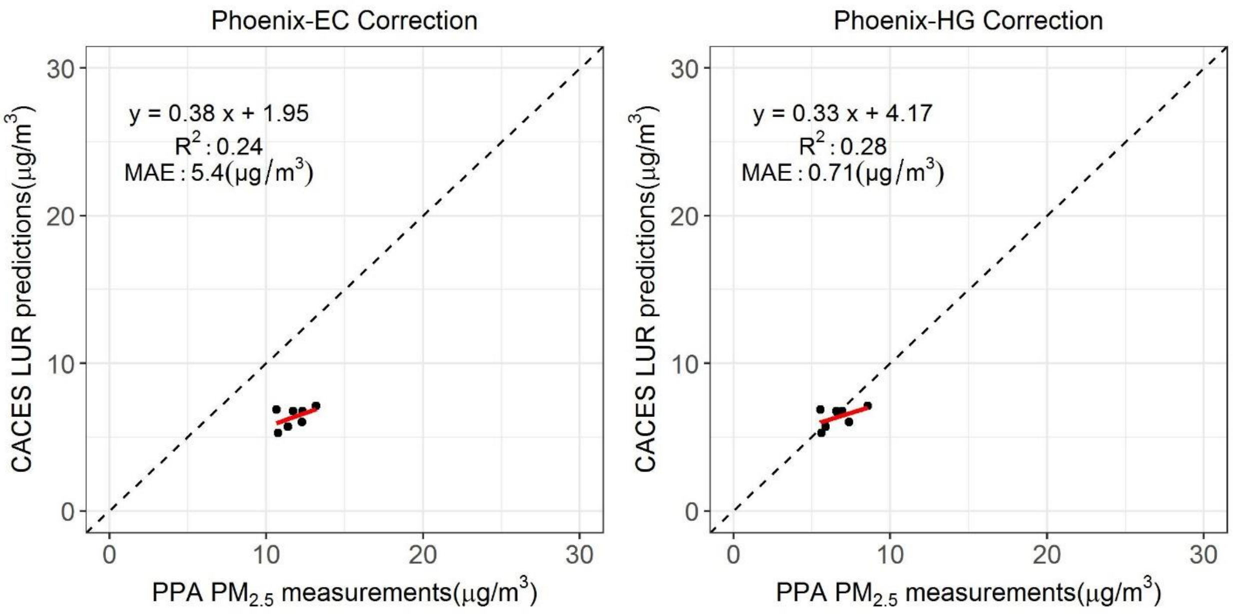


Figure A4.9. External evaluation of the CACES LUR using the HG vs. EC correction methods for Phoenix.

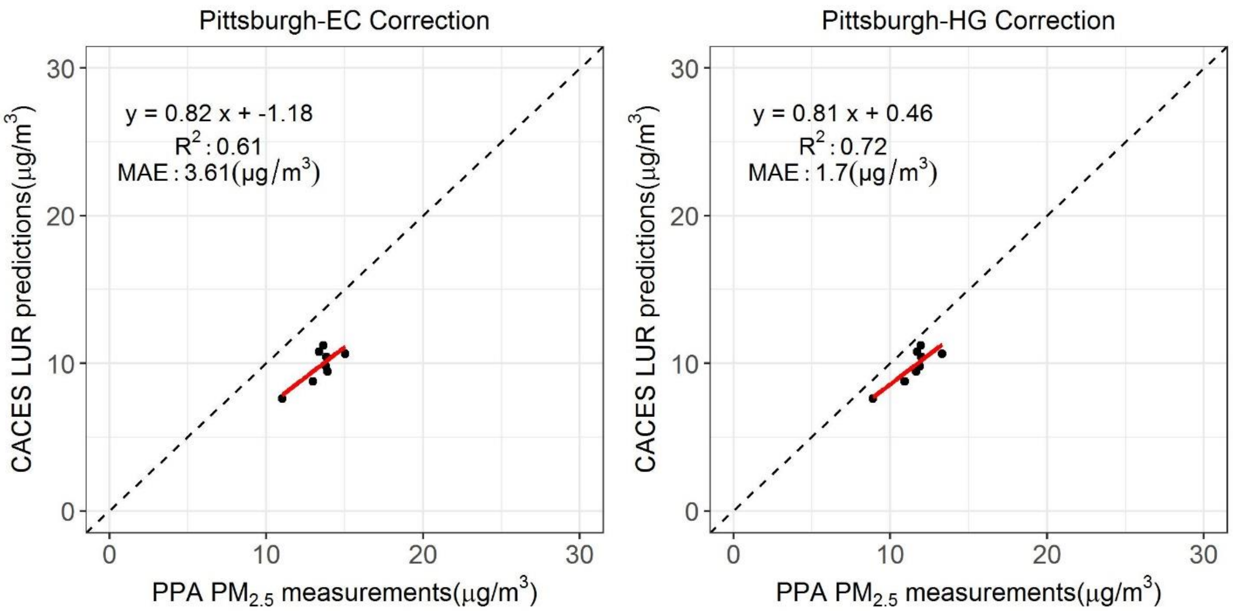


Figure A4.10. External evaluation of the CACES LUR using the HG vs. EC correction methods for Pittsburgh.

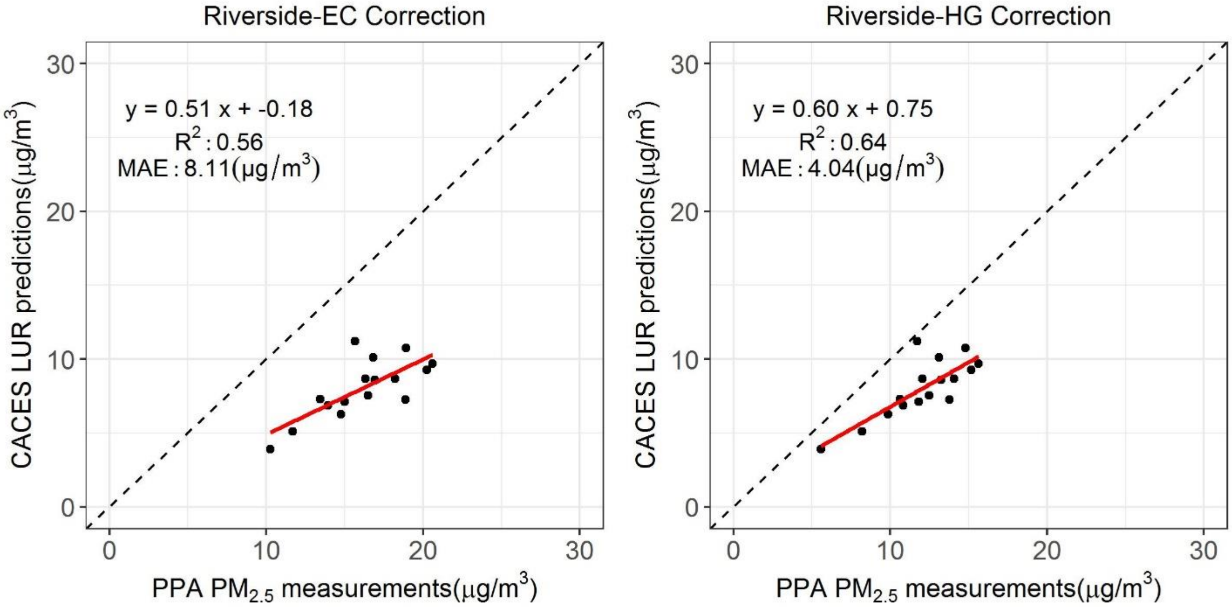


Figure A4.11. External evaluation of the CACES LUR using the HG vs. EC correction methods for Riverside.

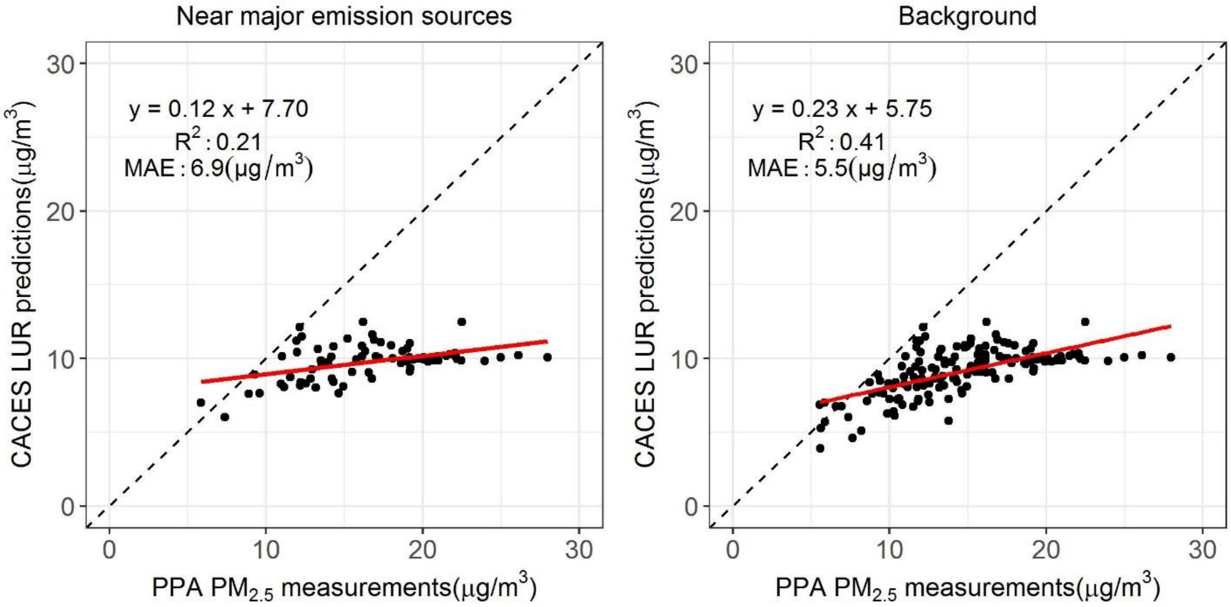


Figure A4.12. External evaluation using the PPA sites near major emission sources (i.e., traffic, restaurant, and NEI facilities) vs. background (i.e., all other sites) for all 6 cities.

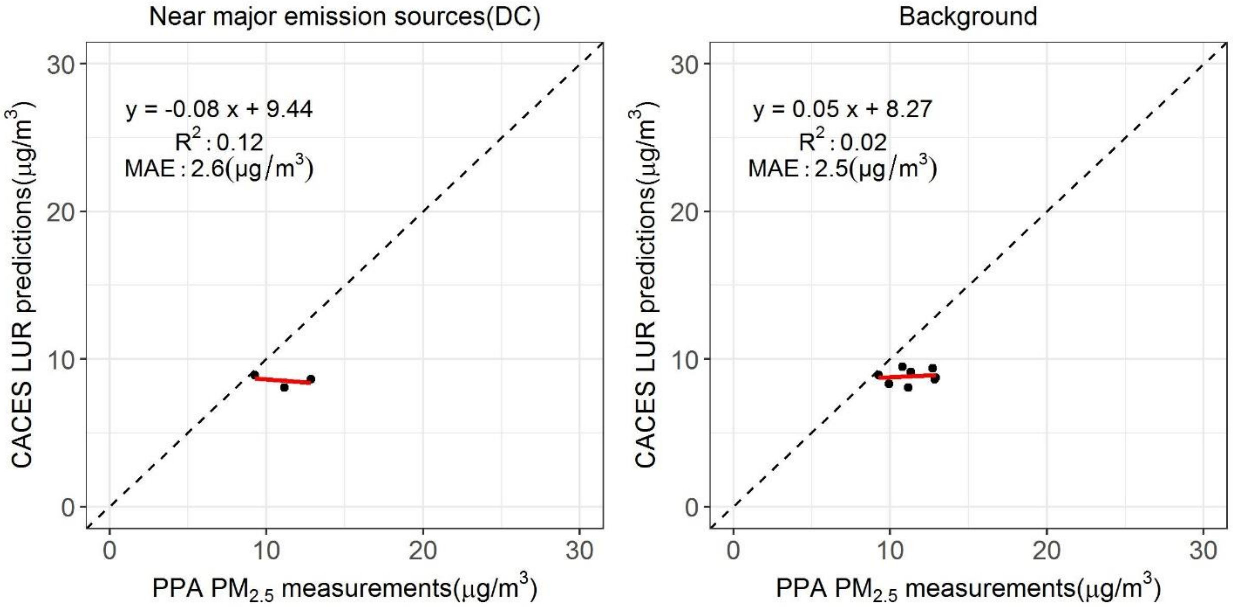


Figure A4.13. External evaluation using the PPA sites near major emission sources (i.e., traffic, restaurant, and NEI facilities) vs. background (i.e., all other sites) for DC.

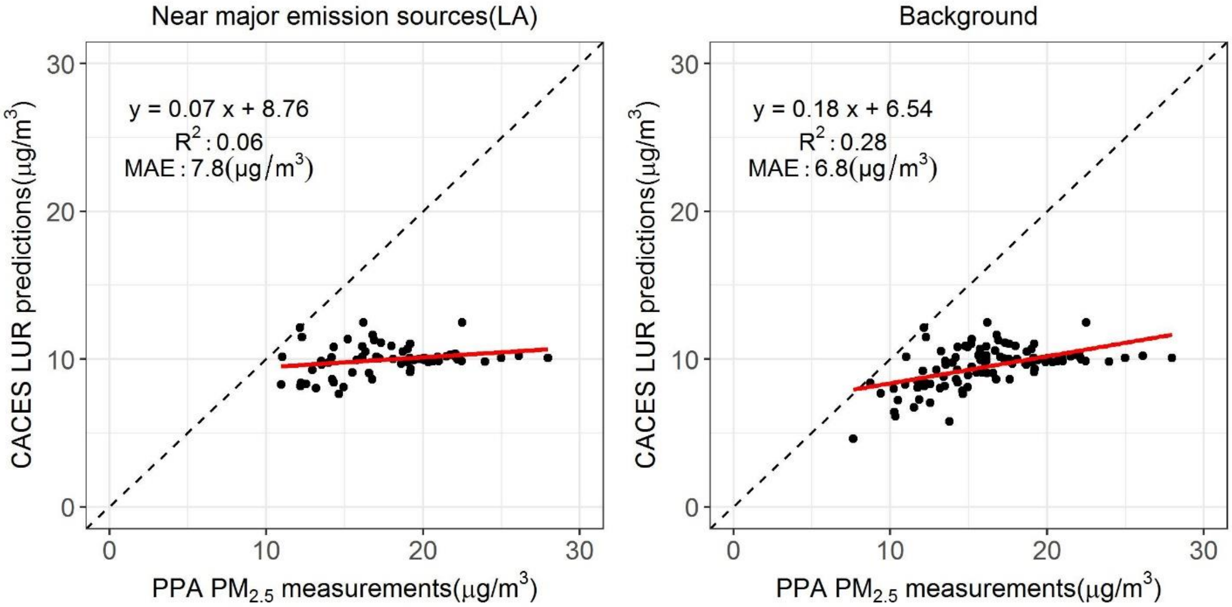


Figure A4.14. External evaluation using the PPA sites near major emission sources (i.e., traffic, restaurant, and NEI facilities) vs. background (i.e., all other sites) for LA.

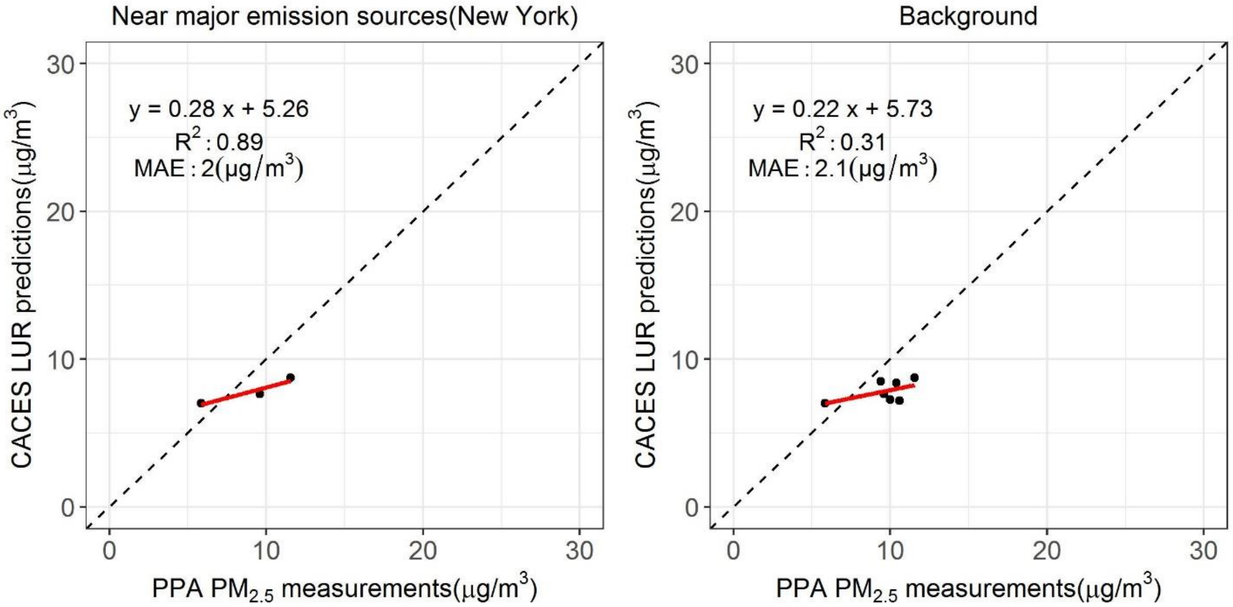


Figure A4.15. External evaluation using the PPA sites near major emission sources (i.e., traffic, restaurant, and NEI facilities) vs. background (i.e., all other sites) for New York.

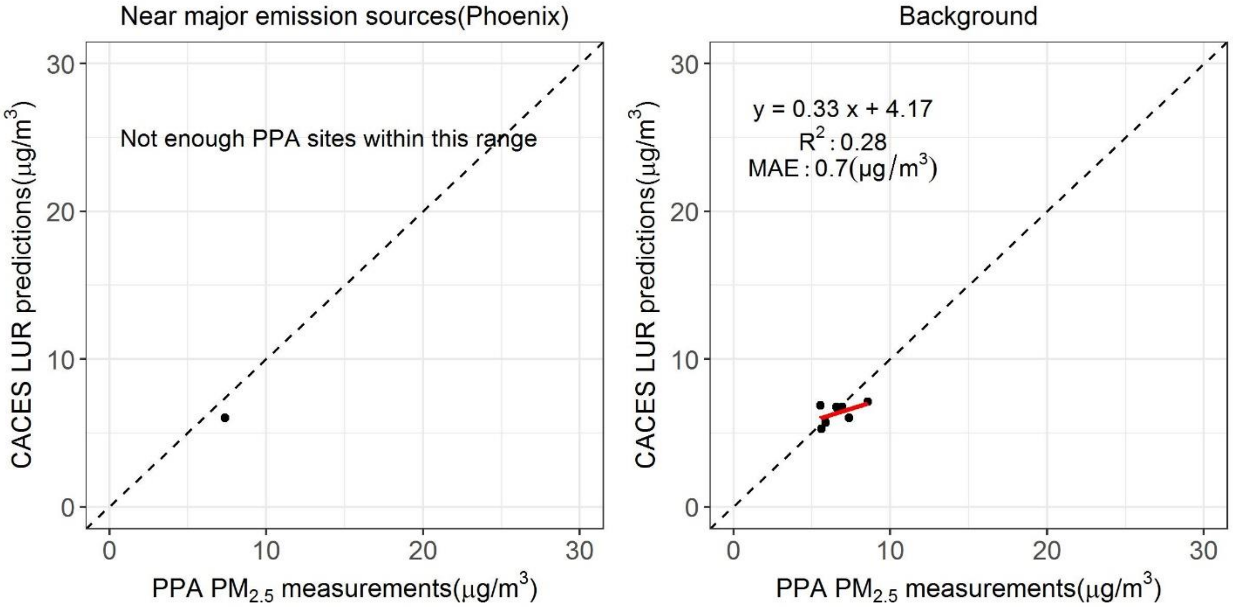


Figure A4.16. External evaluation using the PPA sites near major emission sources (i.e., traffic, restaurant, and NEI facilities) vs. background (i.e., all other sites) for Phoenix.

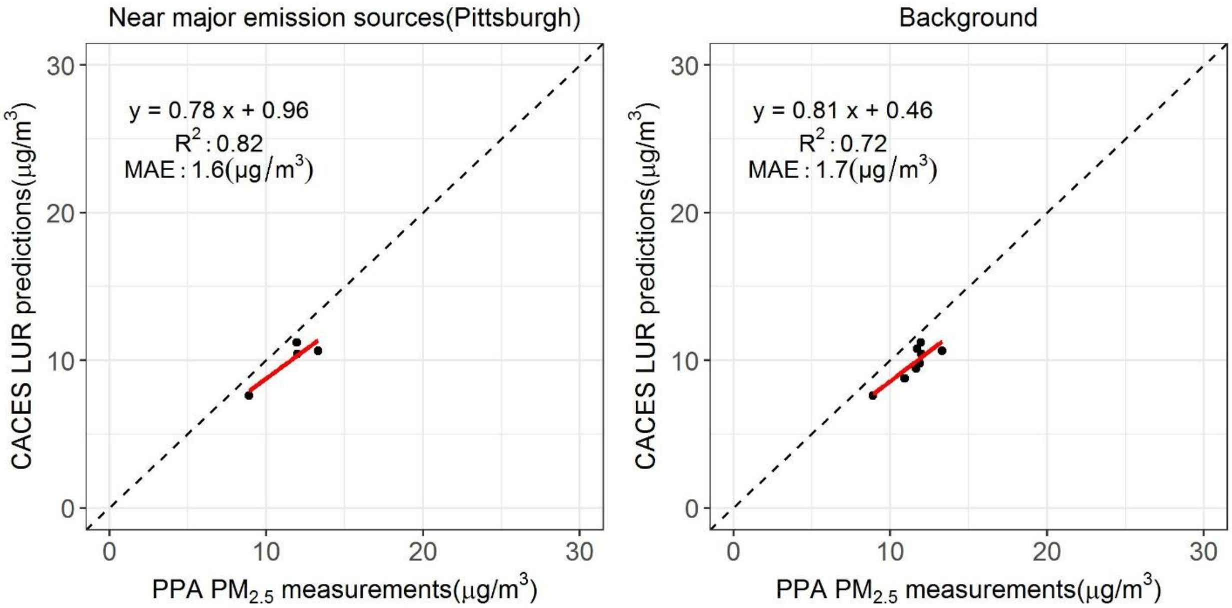


Figure A4.17. External evaluation using the PPA sites near major emission sources (i.e., traffic, restaurant, and NEI facilities) vs. background (i.e., all other sites) for Pittsburgh.

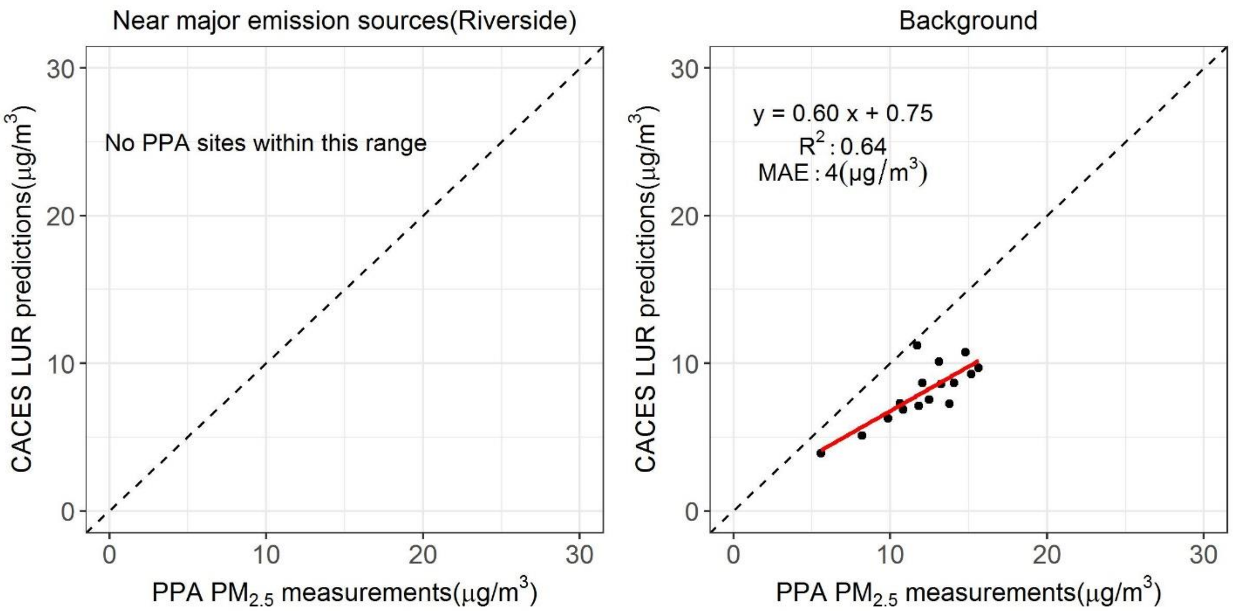


Figure A4.18. External evaluation using the PPA sites near major emission sources (i.e., traffic, restaurant, and NEI facilities) vs. background (i.e., all other sites) for Riverside.

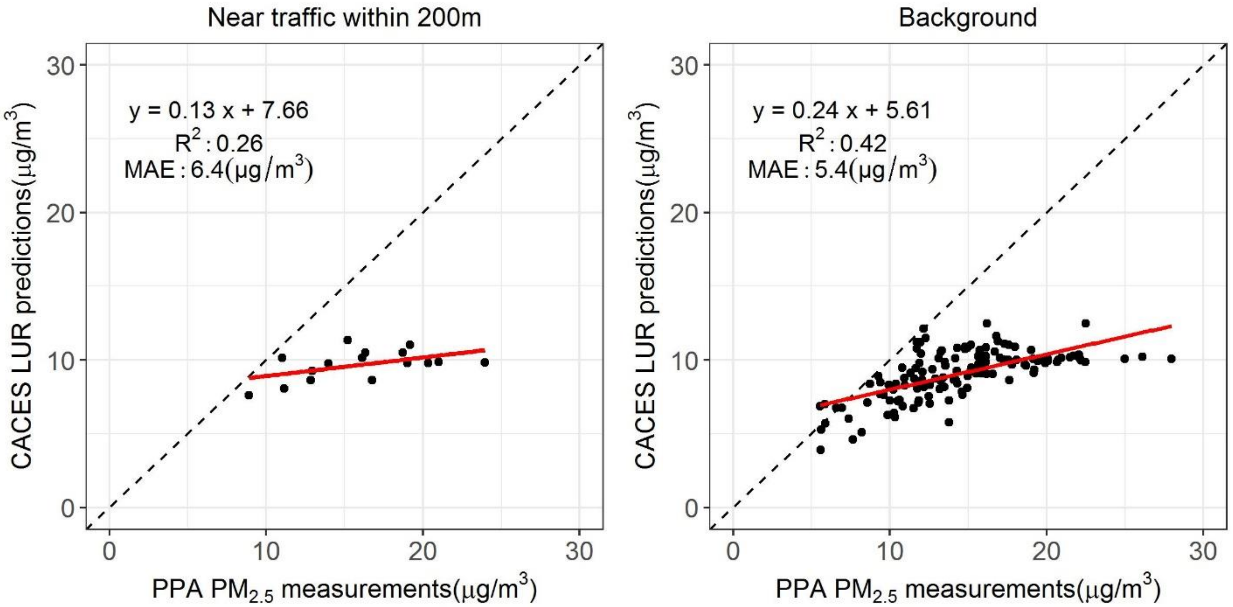


Figure A4.19. External evaluation using the PPA sites near traffic vs. background (i.e., all other sites) for all 6 cities.

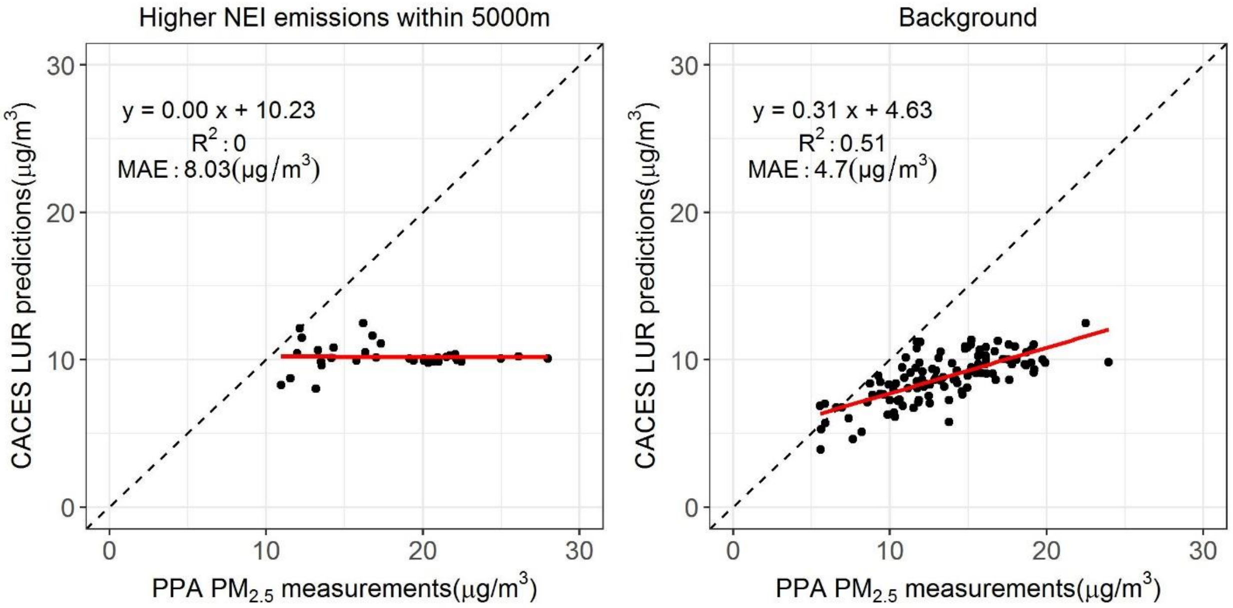


Figure A4.20. External evaluation using the PPA sites near higher NEI emissions vs. background (i.e., all other sites) for all 6 cities.

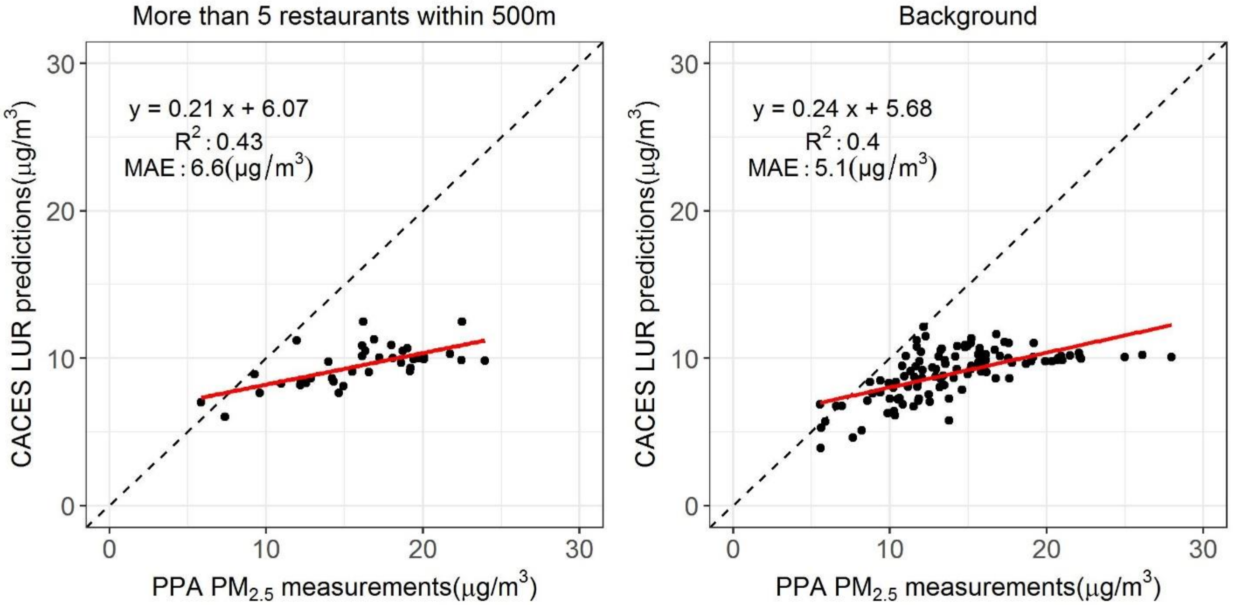


Figure A4.21. External evaluation of the PPA sites near restaurants vs. background (i.e., all other sites) for all 6 cities.

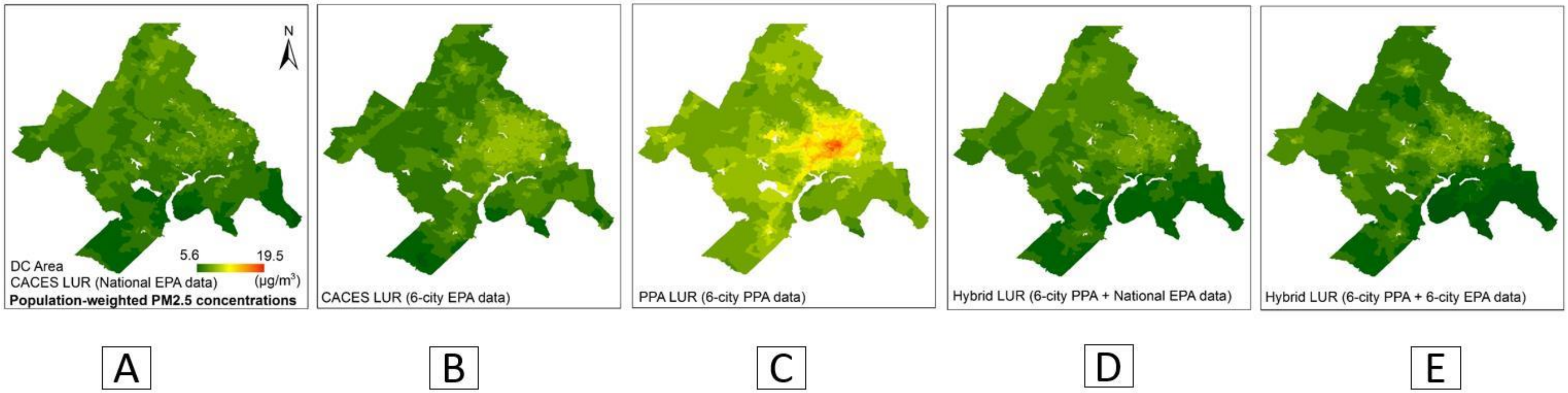

N
DC Area
5.6
19.5
CACES LUR (National EPA data)
(μg/m³)
Population-weighted PM2.5 concentrations
CACES LUR (6-city EPA data)
PPA LUR (6-city PPA data)
Hybrid LUR (6-city PPA + National EPA data)
Hybrid LUR (6-city PPA + 6-city EPA data)
A
B
C
D
E

Figure A4.22: Population-weighted $PM_{2.5}$ concentration maps of the CACES LUR, PPA LUR, and Hybrid LUR in DC. Panel A is the CACES LUR developed using the national EPA data; panel B is the CACES LUR using the 6-city EPA data; panel C is the PPA LUR using the 6-city PPA data; panel D is the Hybrid LUR using the national EPA data and the 6-city PPA data; panel E is the Hybrid LUR using the 6-city EPA data and the 6-city PPA data.

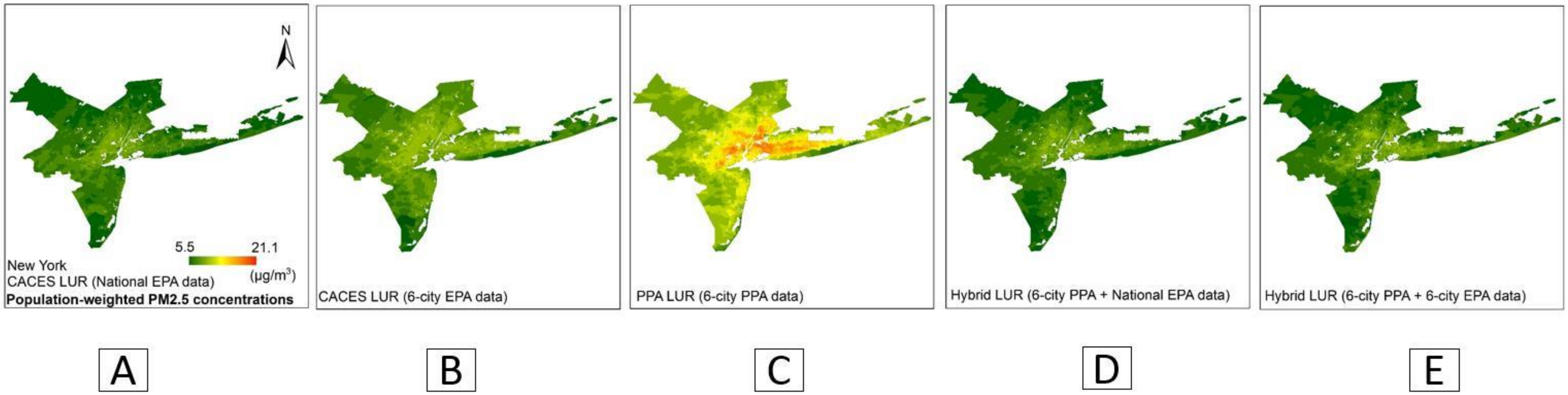

N
5.5
21.1
New York
CACES LUR (National EPA data)
(μg/m3)
Population-weighted PM2.5 concentrations
CACES LUR (6-city EPA data)
PPA LUR (6-city PPA data)
Hybrid LUR (6-city PPA + National EPA data)
Hybrid LUR (6-city PPA + 6-city EPA data)
A
B
C
D
E

Figure A4.23: Population-weighted $PM_{2.5}$ concentration maps of the CACES LUR, PPA LUR, and Hybrid LUR in New York. Panel A is the CACES LUR developed using the national EPA data; panel B is the CACES LUR using the 6-city EPA data; panel C is the PPA LUR using the 6-city PPA data; panel D is the Hybrid LUR using the national EPA data and the 6-city PPA data; panel E is the Hybrid LUR using the 6-city EPA data and the 6-city PPA data.

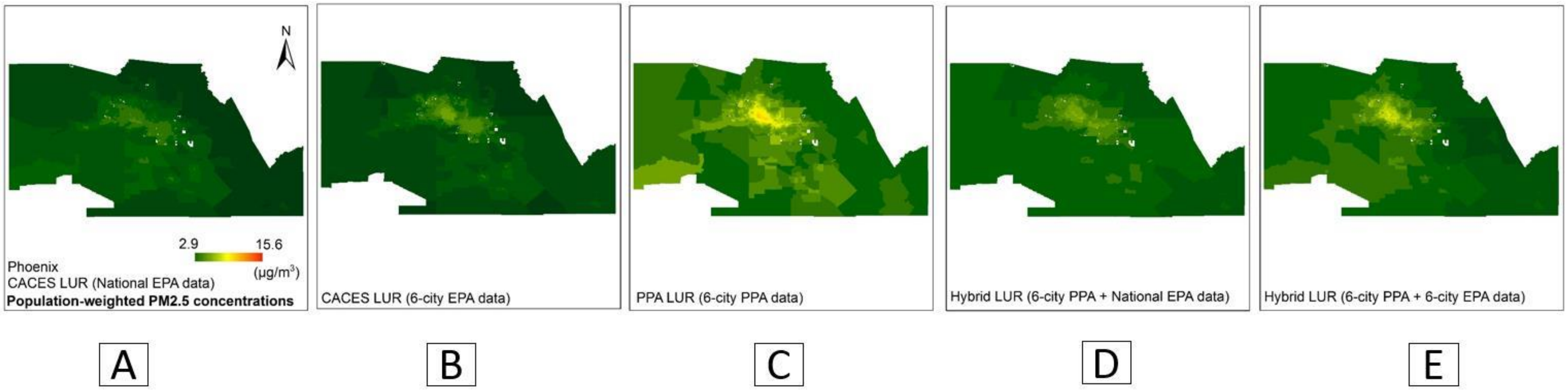

N
2.9
15.6
(μg/m³)
Phoenix
CACES LUR (National EPA data)
Population-weighted PM2.5 concentrations
CACES LUR (6-city EPA data)
PPA LUR (6-city PPA data)
Hybrid LUR (6-city PPA + National EPA data)
Hybrid LUR (6-city PPA + 6-city EPA data)
A
B
C
D
E

Figure A4.24: Population-weighted $PM_{2.5}$ concentration maps of the CACES LUR, PPA LUR, and Hybrid LUR in Phoenix. Panel A is the CACES LUR developed using the national EPA data; panel B is the CACES LUR using the 6-city EPA data; panel C is the PPA LUR using the 6-city PPA data; panel D is the Hybrid LUR using the national EPA data and the 6-city PPA data; panel E is the Hybrid LUR using the 6-city EPA data and the 6-city PPA data.

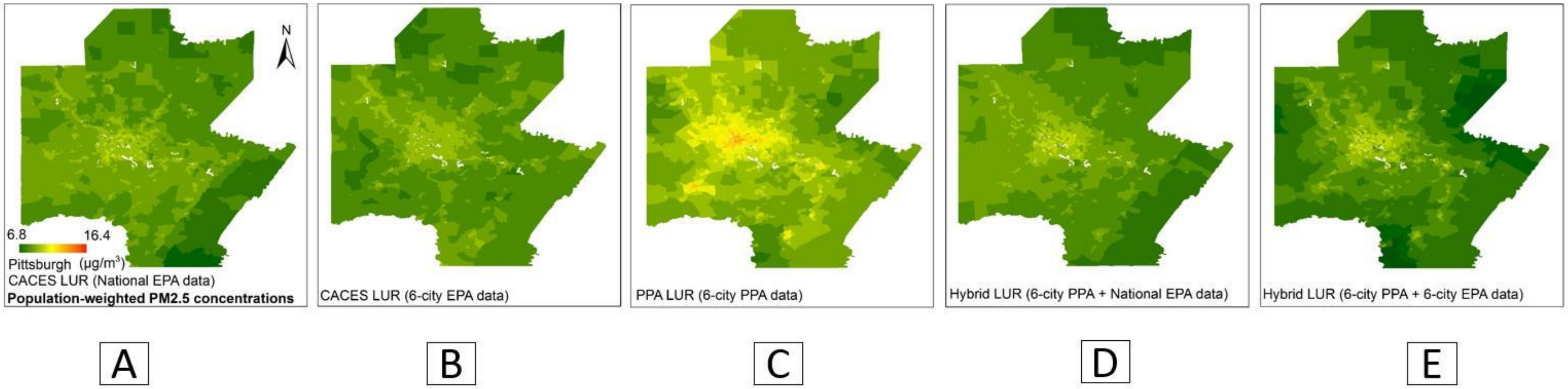

N
6.8
16.4
Pittsburgh (µg/m³)
CACES LUR (National EPA data)
Population-weighted PM2.5 concentrations
CACES LUR (6-city EPA data)
PPA LUR (6-city PPA data)
Hybrid LUR (6-city PPA + National EPA data)
Hybrid LUR (6-city PPA + 6-city EPA data)
A
B
C
D
E

Figure A4.25: Population-weighted $PM_{2.5}$ concentration maps of the CACES LUR, PPA LUR, and Hybrid LUR in Pittsburgh. Panel A is the CACES LUR developed using the national EPA data; panel B is the CACES LUR using the 6-city EPA data; panel C is the PPA LUR using the 6-city PPA data; panel D is the Hybrid LUR using the national EPA data and the 6-city PPA data; panel E is the Hybrid LUR using the 6-city EPA data and the 6-city PPA data.

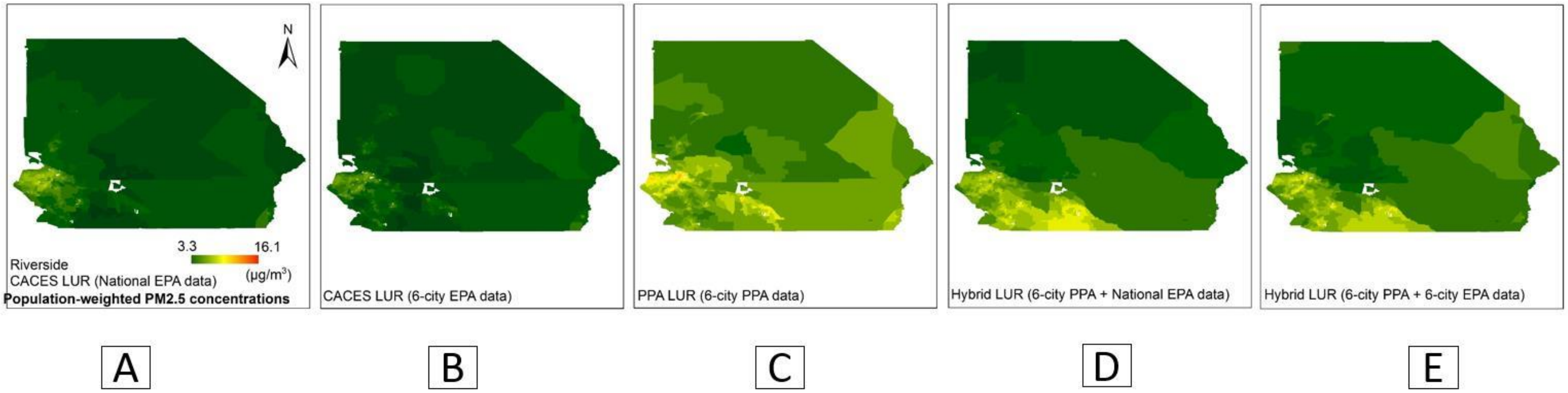

N
3.3
16.1
Riverside
CACES LUR (National EPA data)
(μg/m³)
Population-weighted PM2.5 concentrations
CACES LUR (6-city EPA data)
PPA LUR (6-city PPA data)
Hybrid LUR (6-city PPA + National EPA data)
Hybrid LUR (6-city PPA + 6-city EPA data)
A
B
C
D
E

Figure A4.26: Population-weighted $PM_{2.5}$ concentration maps of the CACES LUR, PPA LUR, and Hybrid LUR in Pittsburgh. Panel A is the CACES LUR developed using the national EPA data; panel B is the CACES LUR using the 6-city EPA data; panel C is the PPA LUR using the 6-city PPA data; panel D is the Hybrid LUR using the national EPA data and the 6-city PPA data; panel E is the Hybrid LUR using the 6-city EPA data and the 6-city PPA data.

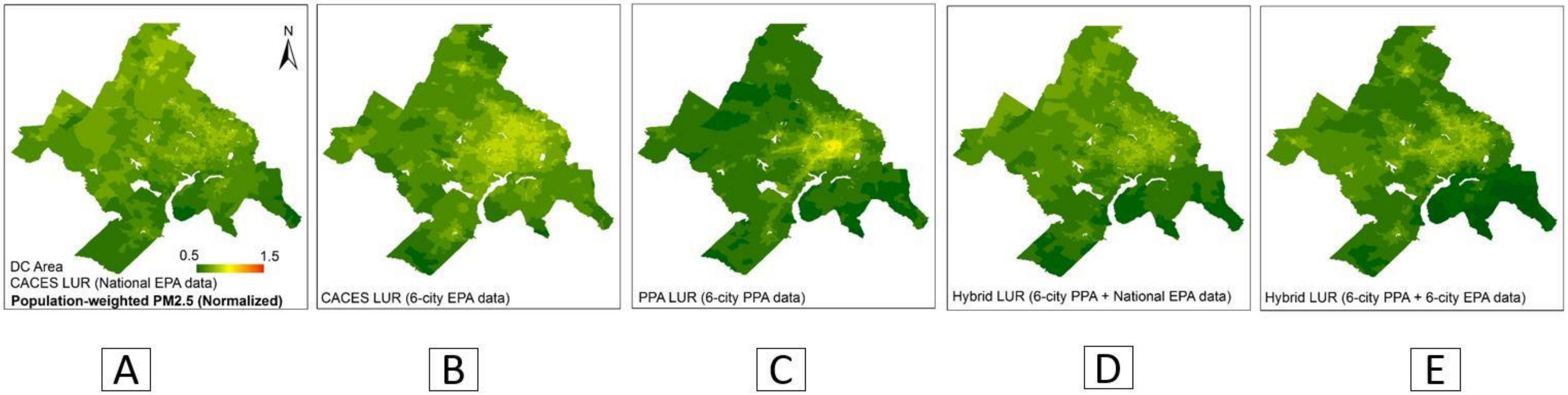

N
DC Area
CACES LUR (National EPA data)
Population-weighted PM2.5 (Normalized)
0.5
1.5
CACES LUR (6-city EPA data)
PPA LUR (6-city PPA data)
Hybrid LUR (6-city PPA + National EPA data)
Hybrid LUR (6-city PPA + 6-city EPA data)
A
B
C
D
E

Figure A4.27: Normalized population-weighted $PM_{2.5}$ concentration maps of the CACES LUR, PPA LUR, and Hybrid LUR in DC (Normalized to the mean concentrations of the 6 cities). Panel A is the CACES LUR developed using the national EPA data; panel B is the CACES LUR using the 6-city EPA data; panel C is the PPA LUR using the 6-city PPA data; panel D is the Hybrid LUR using the national EPA data and the 6-city PPA data; panel E is the Hybrid LUR using the 6-city EPA data and the 6-city PPA data.

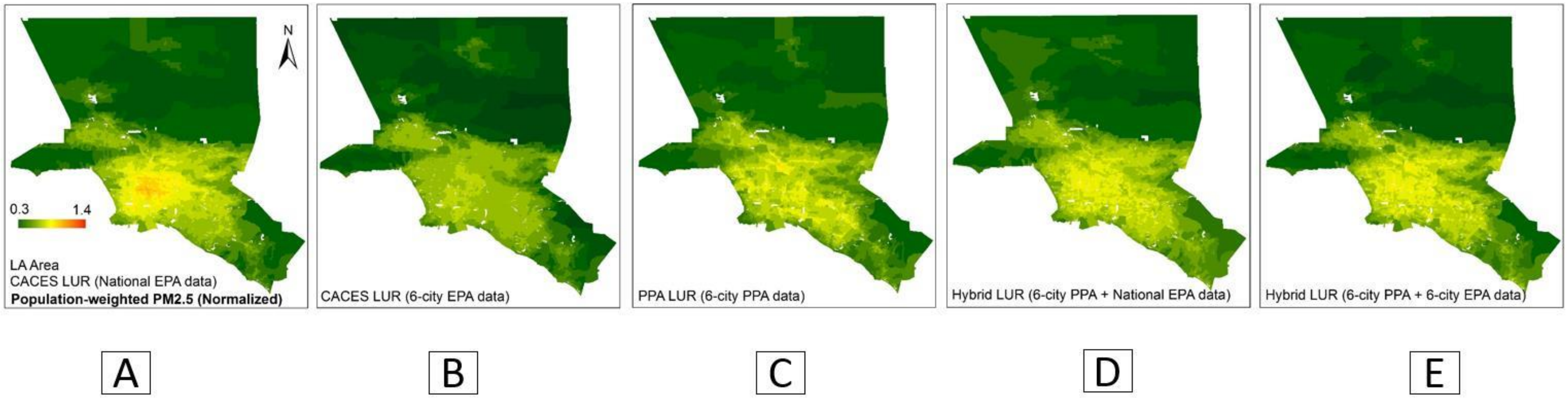

N
0.3
1.4
LA Area
CACES LUR (National EPA data)
Population-weighted PM2.5 (Normalized)
CACES LUR (6-city EPA data)
PPA LUR (6-city PPA data)
Hybrid LUR (6-city PPA + National EPA data)
Hybrid LUR (6-city PPA + 6-city EPA data)
A
B
C
D
E

Figure A4.28: Normalized population-weighted $PM_{2.5}$ concentration maps of the CACES LUR, PPA LUR, and Hybrid LUR in LA (Normalized to the mean concentrations of the 6 cities). Panel A is the CACES LUR developed using the national EPA data; panel B is the CACES LUR using the 6-city EPA data; panel C is the PPA LUR using the 6-city PPA data; panel D is the Hybrid LUR using the national EPA data and the 6-city PPA data; panel E is the Hybrid LUR using the 6-city EPA data and the 6-city PPA data.

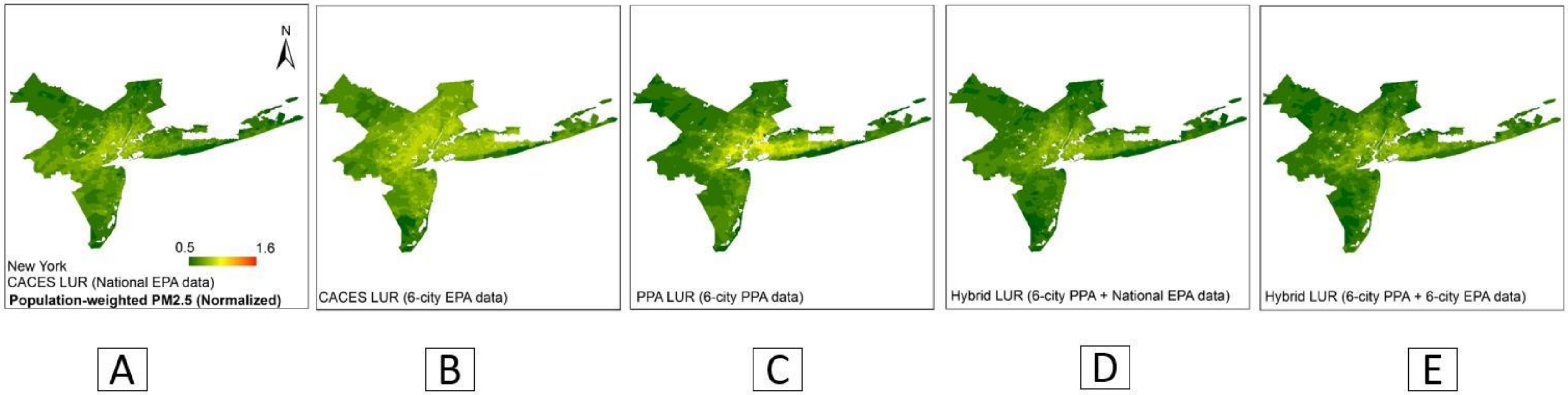

N
0.5
1.6
New York
CACES LUR (National EPA data)
Population-weighted PM2.5 (Normalized)
CACES LUR (6-city EPA data)
PPA LUR (6-city PPA data)
Hybrid LUR (6-city PPA + National EPA data)
Hybrid LUR (6-city PPA + 6-city EPA data)
A
B
C
D
E

Figure A4.29: Normalized population-weighted $PM_{2.5}$ concentration maps of the CACES LUR, PPA LUR, and Hybrid LUR in New York (Normalized to the mean concentrations of the 6 cities). Panel A is the CACES LUR developed using the national EPA data; panel B is the CACES LUR using the 6-city EPA data; panel C is the PPA LUR using the 6-city PPA data; panel D is the Hybrid LUR using the national EPA data and the 6-city PPA data; panel E is the Hybrid LUR using the 6-city EPA data and the 6-city PPA data.

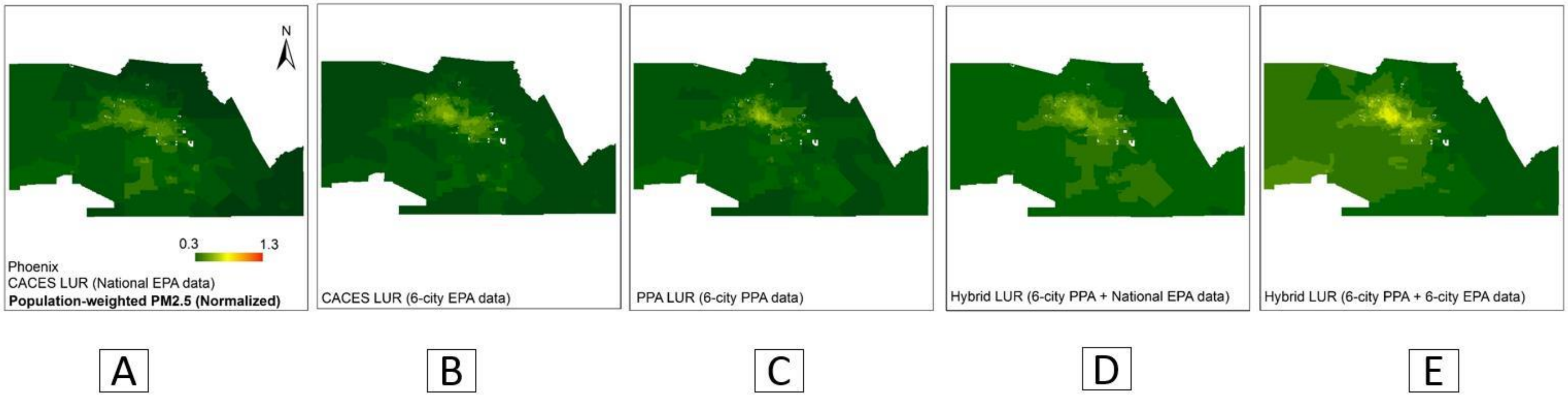

N
0.3
1.3
Phoenix
CACES LUR (National EPA data)
Population-weighted PM2.5 (Normalized)
CACES LUR (6-city EPA data)
PPA LUR (6-city PPA data)
Hybrid LUR (6-city PPA + National EPA data)
Hybrid LUR (6-city PPA + 6-city EPA data)
A
B
C
D
E

Figure A4.30: Normalized population-weighted $PM_{2.5}$ concentration maps of the CACES LUR, PPA LUR, and Hybrid LUR in Phoenix (Normalized to the mean concentrations of the 6 cities). Panel A is the CACES LUR developed using the national EPA data; panel B is the CACES LUR using the 6-city EPA data; panel C is the PPA LUR using the 6-city PPA data; panel D is the Hybrid LUR using the national EPA data and the 6-city PPA data; panel E is the Hybrid LUR using the 6-city EPA data and the 6-city PPA data.

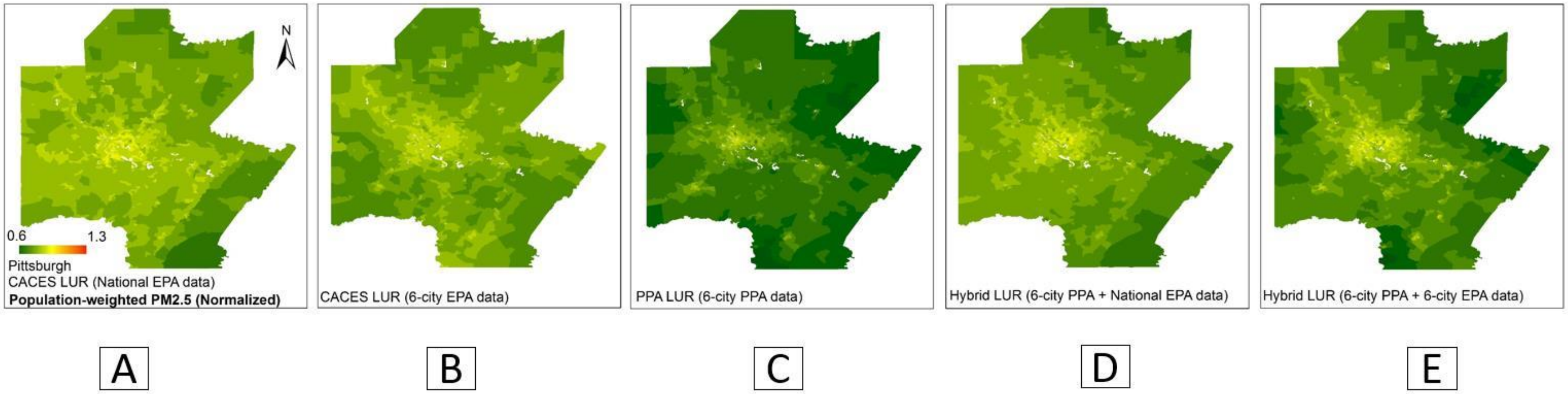

N
0.6
1.3
Pittsburgh
CACES LUR (National EPA data)
Population-weighted PM2.5 (Normalized)
CACES LUR (6-city EPA data)
PPA LUR (6-city PPA data)
Hybrid LUR (6-city PPA + National EPA data)
Hybrid LUR (6-city PPA + 6-city EPA data)
A
B
C
D
E

Figure A4.31: Normalized population-weighted $PM_{2.5}$ concentration maps of the CACES LUR, PPA LUR, and Hybrid LUR in Pittsburgh (Normalized to the mean concentrations of the 6 cities). Panel A is the CACES LUR developed using the national EPA data; panel B is the CACES LUR using the 6-city EPA data; panel C is the PPA LUR using the 6-city PPA data; panel D is the Hybrid LUR using the national EPA data and the 6-city PPA data; panel E is the Hybrid LUR using the 6-city EPA data and the 6-city PPA data.

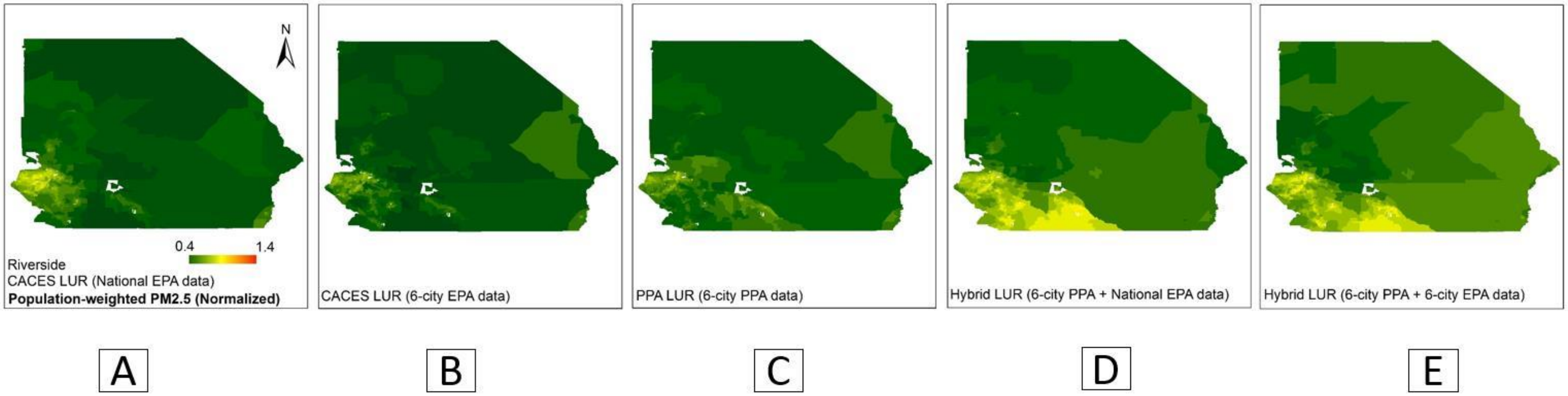

N
0.4
1.4
Riverside
CACES LUR (National EPA data)
Population-weighted PM2.5 (Normalized)
CACES LUR (6-city EPA data)
PPA LUR (6-city PPA data)
Hybrid LUR (6-city PPA + National EPA data)
Hybrid LUR (6-city PPA + 6-city EPA data)
A
B
C
D
E

Figure A4.32: Normalized population-weighted $PM_{2.5}$ concentration maps of the CACES LUR, PPA LUR, and Hybrid LUR in Riverside (Normalized to the mean concentrations of the 6 cities). Panel A is the CACES LUR developed using the national EPA data; panel B is the CACES LUR using the 6-city EPA data; panel C is the PPA LUR using the 6-city PPA data; panel D is the Hybrid LUR using the national EPA data and the 6-city PPA data; panel E is the Hybrid LUR using the 6-city EPA data and the 6-city PPA data.

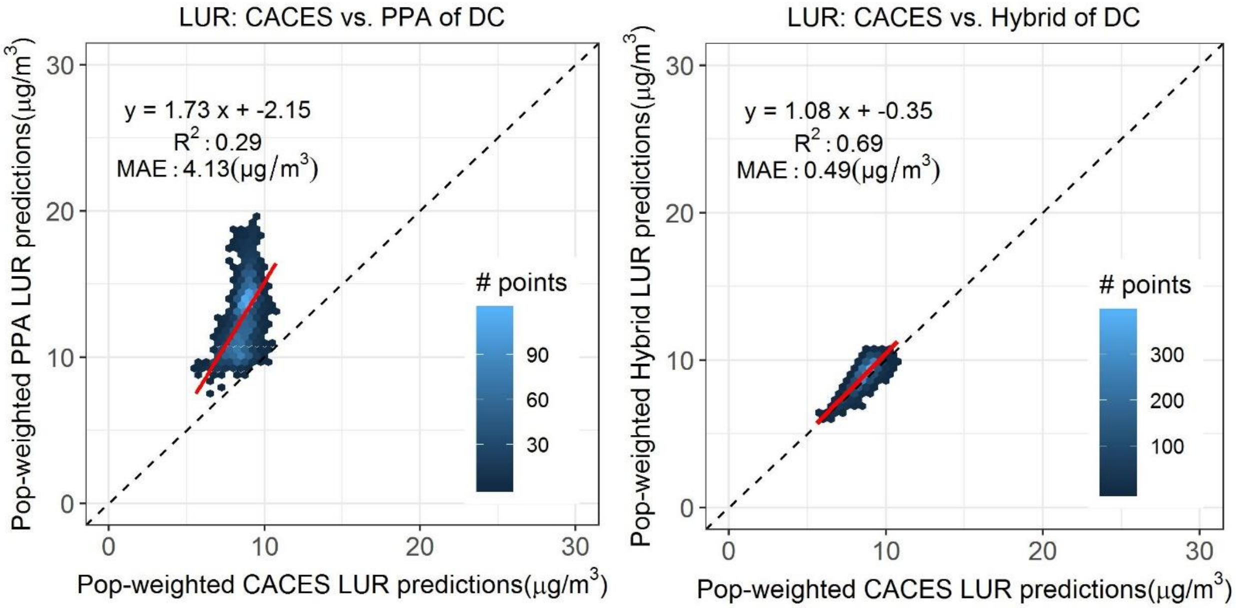


Figure A4.33. Scatterplot of the PPA LUR vs. CACES LUR and the Hybrid LUR vs. CACES LUR for DC.

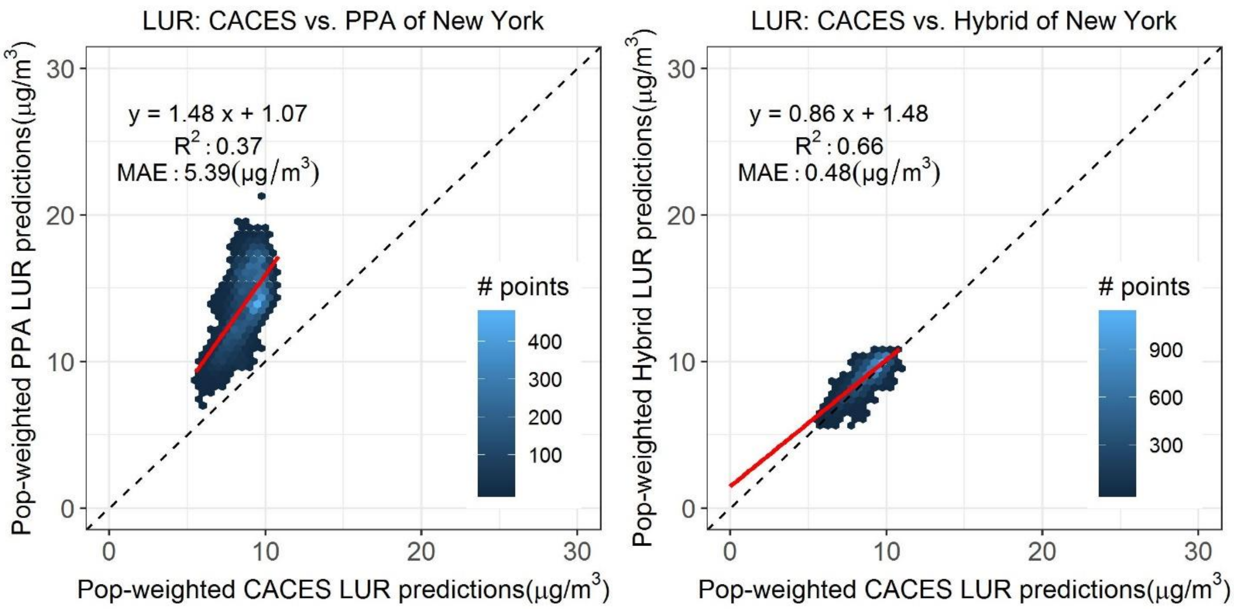


Figure A4.34. Scatterplot of the PPA LUR vs. CACES LUR and the Hybrid LUR vs. CACES LUR for New York.

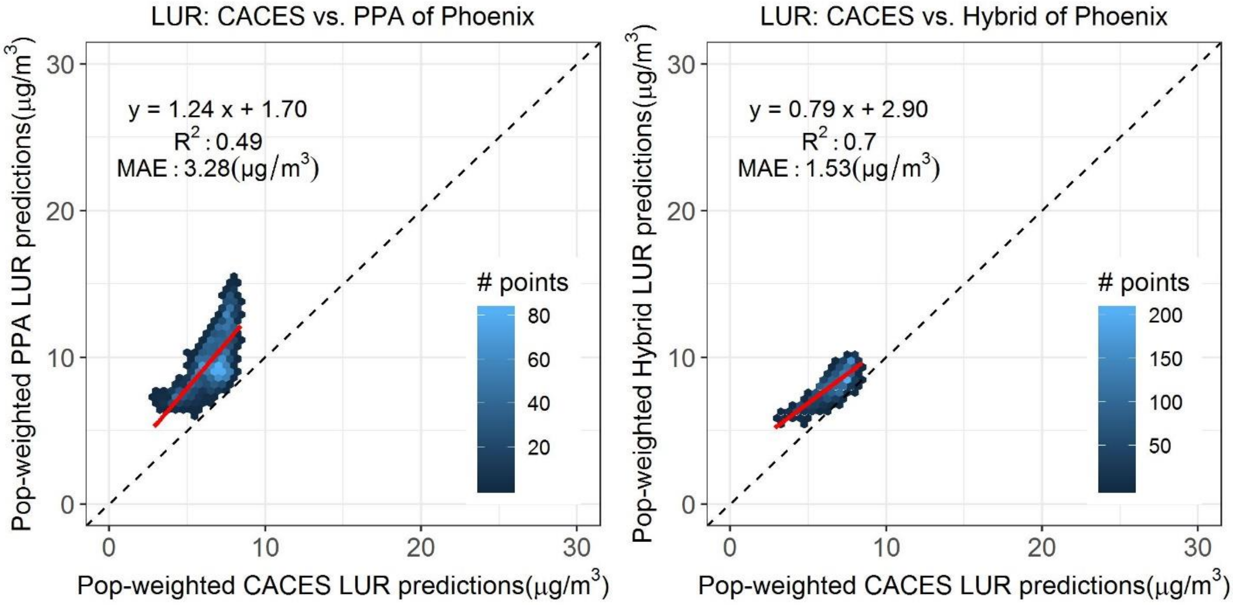


Figure A4.35. Scatterplot of the PPA LUR vs. CACES LUR and the Hybrid LUR vs. CACES LUR for Phoenix.

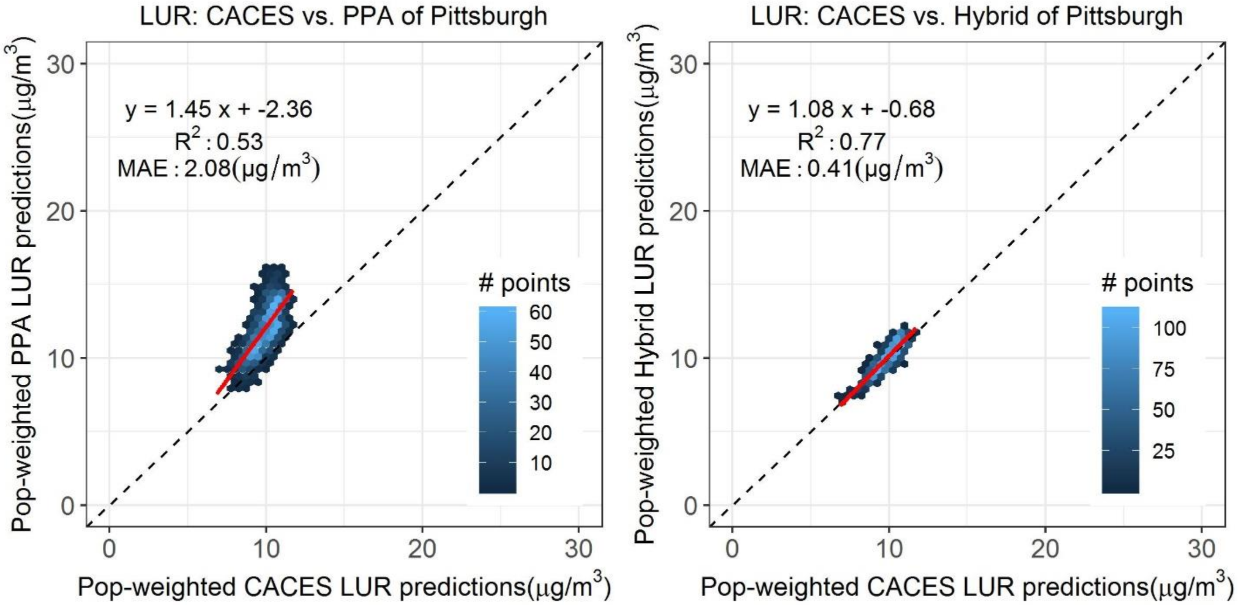


Figure A4.36. Scatterplot of the PPA LUR vs. CACES LUR and the Hybrid LUR vs. CACES LUR for Pittsburgh.

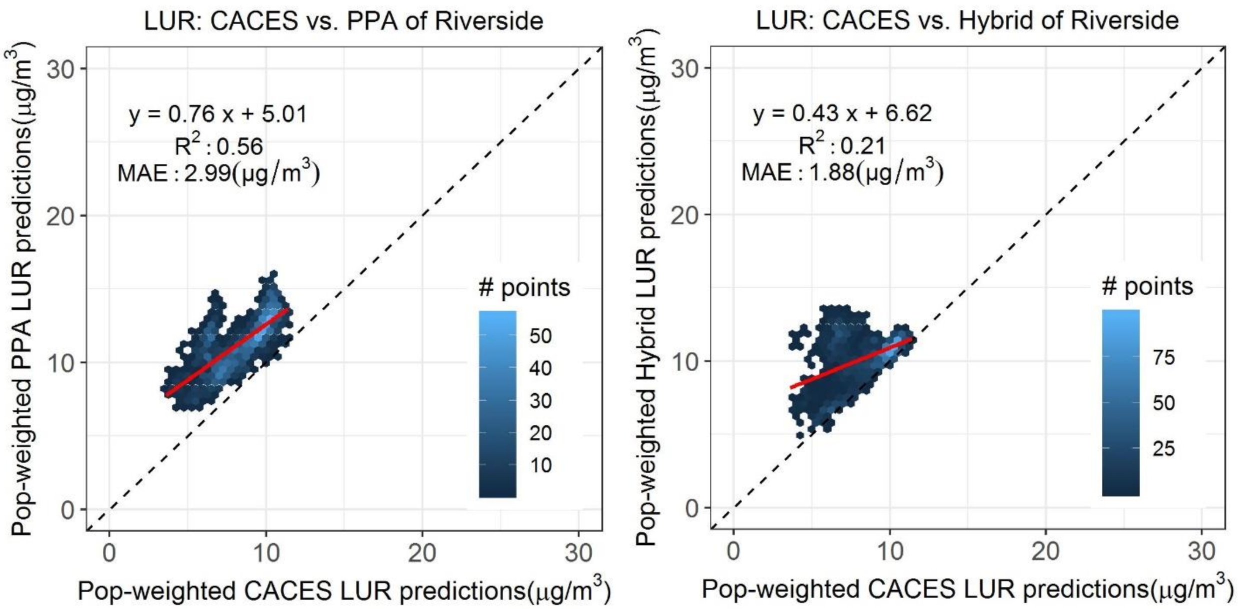


Figure A4.37. Scatterplot of the PPA LUR vs. CACES LUR and the Hybrid LUR vs. CACES LUR for Riverside.

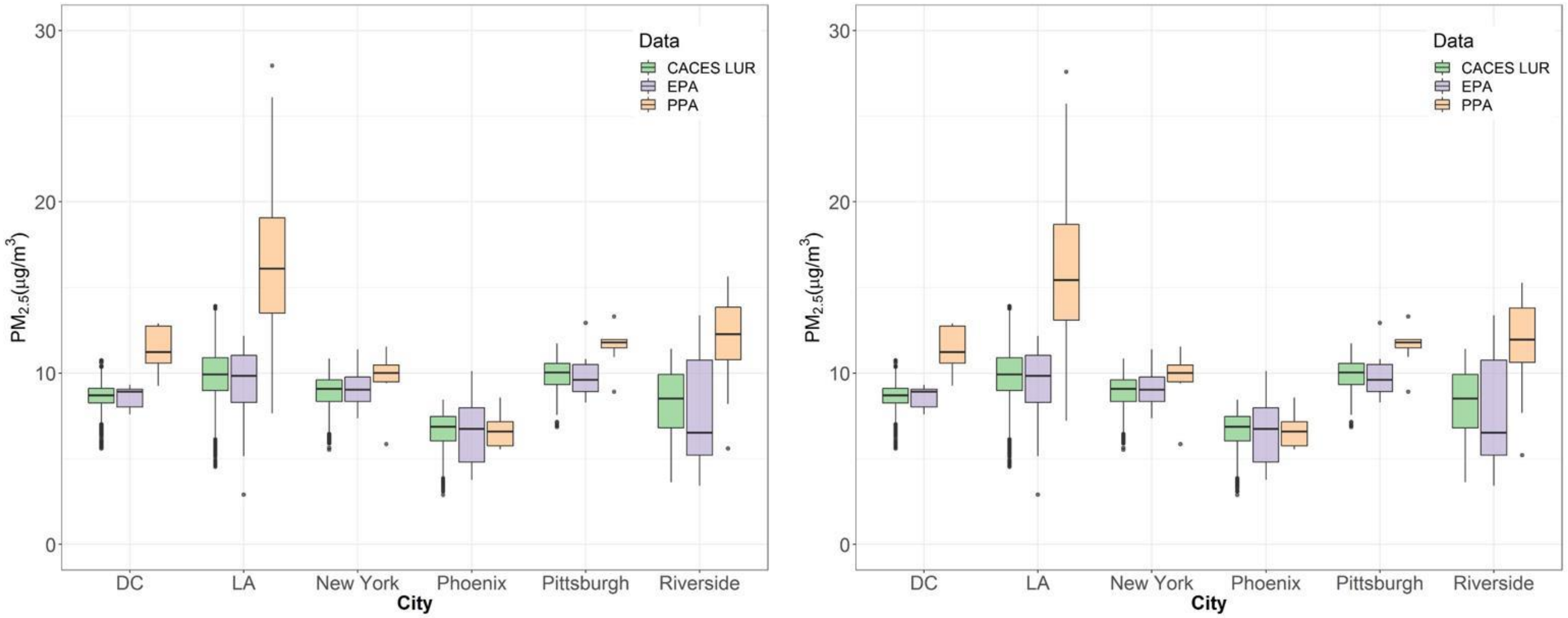


Figure A4.38. Boxplots of $PM_{2.5}$ concentrations: full-year version (left panel) vs. a filtered version (right panel; excluding the PPA data of December 2017 in LA and Riverside.

Table A5.1 Candidate independent variables of satellite products

Pollutant	**Years**	**Resolution**	**Instrument**	**Level**	**Source**
$PM_{2.5}$	1998 – 2014	0.1°	Multiple instruments	Surface	http://fizz.phys.dal.ca/~atmos/martin/?page_id=140
NO_2[a]	2004 – 2015	0.1°	OMI[c]	Column	http://www.temis.nl/airpollution/no2.html
SO_2	2005 – 2016	0.25°	OMI	Column	https://disc.gsfc.nasa.gov/datacollection/OMSO2_CPR_003.html
HCHO[b]	2005 – 2016	0.1°	OMI	Column	https://disc.gsfc.nasa.gov/datasets/OMHCHO_V003/summary
CO	2001 – 2016	0.25°	MOPITT[d]	Surface	https://eosweb.larc.nasa.gov/datapool

[a] both 1-year and 3-year averages calculated;

[b] long term (12-year) average only;

[c] Ozone Monitoring Instrument;

[d] Measurements of Pollution in the Troposphere.

Table A5.2 Categories of point of interest (POI) data

POI name	**POI name**	**POI name**
accounting	electrician	night_club
airport	electronics_store	painter
amusement_park	embassy	park
aquarium	fire_station	parking
art_gallery	florist	pet_store
atm	funeral_home	pharmacy
bakery	furniture_store	physiotherapist
bank	gas_station	plumber
bar	gym	police
beauty_salon	hair_care	post_office
bicycle_store	hardware_store	real_estate_agency
book_store	hindu_temple	restaurant
bowling_alley	home_goods_store	roofing_contractor
bus_station	hospital	rv_park
cafe	insurance_agency	school
campground	jewelry_store	shoe_store
car_dealer	laundry	shopping_mall
car_rental	lawyer	spa
car_repair	library	stadium
car_wash	liquor_store	storage
casino	local_government_office	store
cemetery	locksmith	subway_station
church	lodging	supermarket
city_hall	meal_delivery	synagogue
clothing_store	meal_takeaway	taxi_stand
convenience_store	mosque	train_station
courthouse	movie_rental	transit_station
dentist	movie_theater	travel_agency
department_store	moving_company	veterinary_care
doctor	museum	zoo

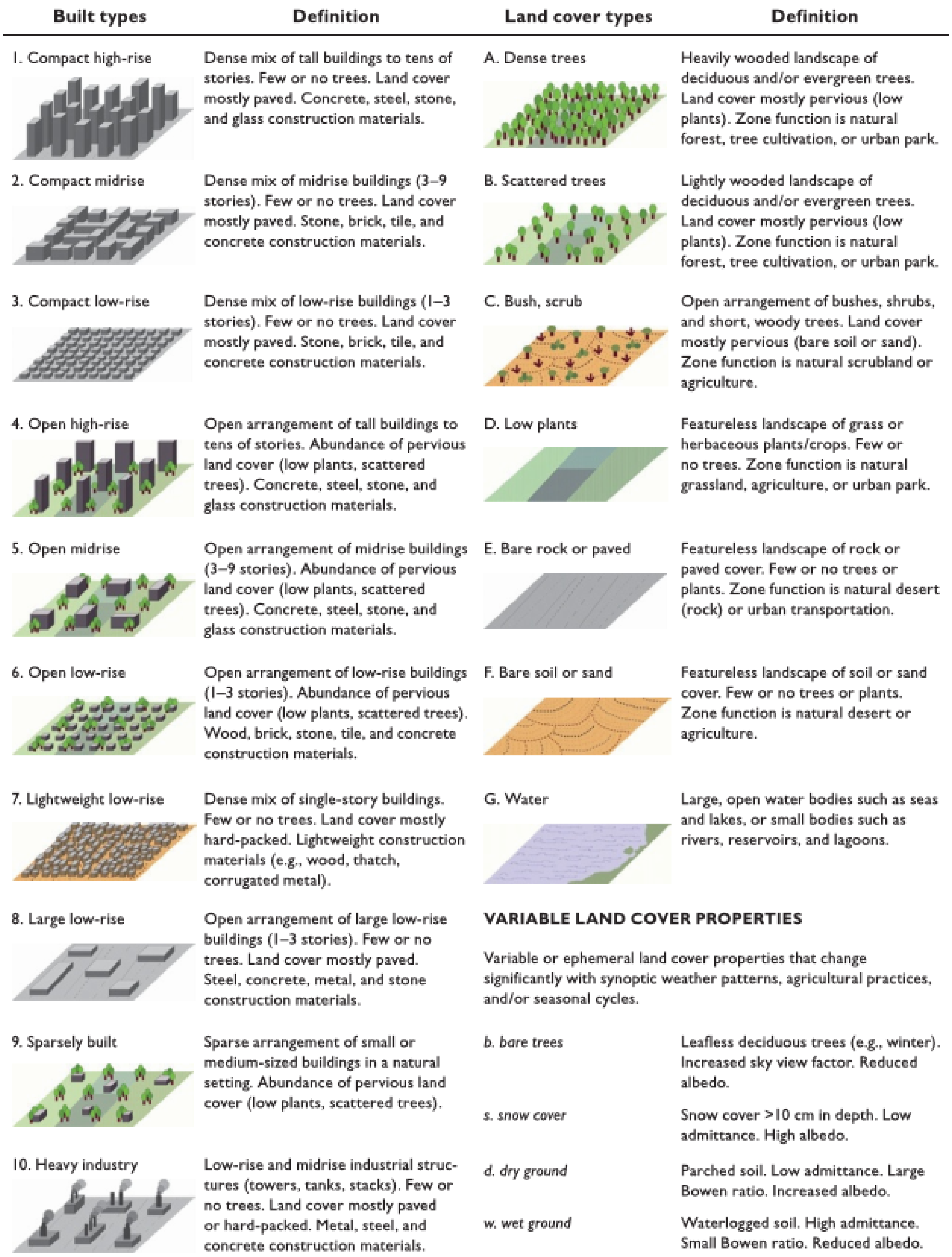
Table A5.3 Categories of Google street view imagery (GSV) data used in LUR models

GSV category	**GSV category**	**GSV category**
wall	railing	airplane
building	box	dirt.track
sky	signboard	pole
tree	sand	land
road	skyscraper	van
windowpane	path.1	ship
grass	runway	fountain
sidewalk	river	canopy
person	bridge	swimming.pool
earth	flower	waterfall
mountain	hill	tent
plant	palm	minibike
car	boat	food
water	hovel	pot
house	bus	animal
sea	truck	bicycle
field	tower	lake
fence	awning	sculpture
rock	streetlight	traffic.light

Table A5.4 Summary statistics of other criteria pollutants in 2015

Statistics	**CO**[a]	**PM$_{10}$**[b]	**SO$_2$**[c]
Number of monitors	196	456	367
Mean concentrations	0.3	16	1.2
Std	0.1	7.9	0.9
Min	0.1	3.2	0.1
Q1	0.3	10.9	0.7
Median	0.3	15.6	1
Q3	0.3	19.7	1.4
Max	0.7	46.9	6.4

[a] Concentration unit is ppm;
[b] Concentration unit is μg/m^3
[c] Concentration unit is ppb.



Built types	Definition
1. Compact high-rise	Dense mix of tall buildings to tens of stories. Few or no trees. Land cover mostly paved. Concrete, steel, stone, and glass construction materials.
2. Compact midrise	Dense mix of midrise buildings (3–9 stories). Few or no trees. Land cover mostly paved. Stone, brick, tile, and concrete construction materials.
3. Compact low-rise	Dense mix of low-rise buildings (1–3 stories). Few or no trees. Land cover mostly paved. Stone, brick, tile, and concrete construction materials.
4. Open high-rise	Open arrangement of tall buildings to tens of stories. Abundance of pervious land cover (low plants, scattered trees). Concrete, steel, stone, and glass construction materials.
5. Open midrise	Open arrangement of midrise buildings (3–9 stories). Abundance of pervious land cover (low plants, scattered trees). Concrete, steel, stone, and glass construction materials.
6. Open low-rise	Open arrangement of low-rise buildings (1–3 stories). Abundance of pervious land cover (low plants, scattered trees). Wood, brick, stone, tile, and concrete construction materials.
7. Lightweight low-rise	Dense mix of single-story buildings. Few or no trees. Land cover mostly hard-packed. Lightweight construction materials (e.g., wood, thatch, corrugated metal).
8. Large low-rise	Open arrangement of large low-rise buildings (1–3 stories). Few or no trees. Land cover mostly paved. Steel, concrete, metal, and stone construction materials.
9. Sparsely built	Sparse arrangement of small or medium-sized buildings in a natural setting. Abundance of pervious land cover (low plants, scattered trees).
10. Heavy industry	Low-rise and midrise industrial structures (towers, tanks, stacks). Few or no trees. Land cover mostly paved or hard-packed. Metal, steel, and concrete construction materials.

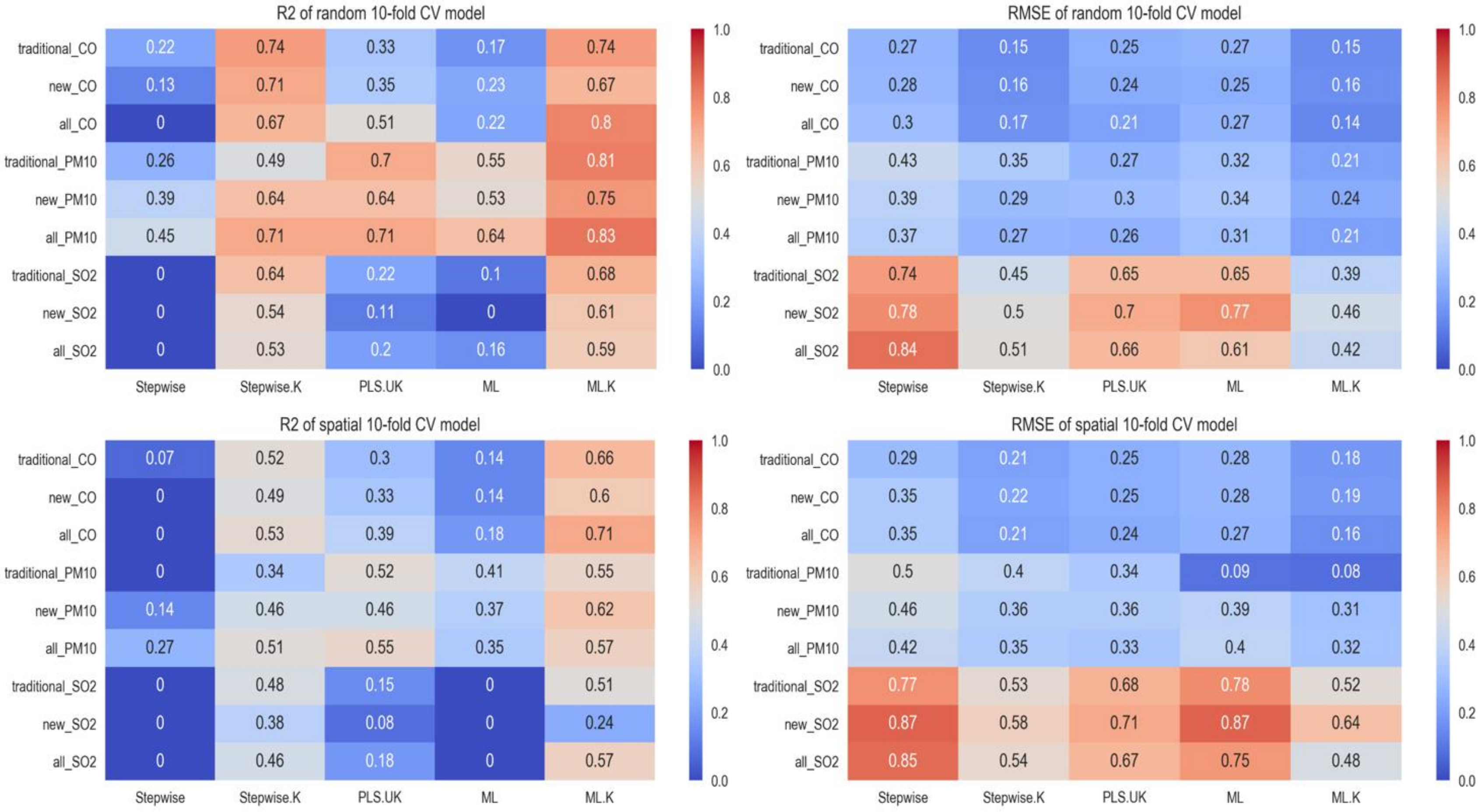
Land cover types	Definition
A. Dense trees	Heavily wooded landscape of deciduous and/or evergreen trees. Land cover mostly pervious (low plants). Zone function is natural forest, tree cultivation, or urban park.
B. Scattered trees	Lightly wooded landscape of deciduous and/or evergreen trees. Land cover mostly pervious (low plants). Zone function is natural forest, tree cultivation, or urban park.
C. Bush, scrub	Open arrangement of bushes, shrubs, and short, woody trees. Land cover mostly pervious (bare soil or sand). Zone function is natural scrubland or agriculture.
D. Low plants	Featureless landscape of grass or herbaceous plants/crops. Few or no trees. Zone function is natural grassland, agriculture, or urban park.
E. Bare rock or paved	Featureless landscape of rock or paved cover. Few or no trees or plants. Zone function is natural desert (rock) or urban transportation.
F. Bare soil or sand	Featureless landscape of soil or sand cover. Few or no trees or plants. Zone function is natural desert or agriculture.
G. Water	Large, open water bodies such as seas and lakes, or small bodies such as rivers, reservoirs, and lagoons.

VARIABLE LAND COVER PROPERTIES

Variable or ephemeral land cover properties that change significantly with synoptic weather patterns, agricultural practices, and/or seasonal cycles.

b. bare trees	Leafless deciduous trees (e.g., winter). Increased sky view factor. Reduced albedo.
s. snow cover	Snow cover >10 cm in depth. Low admittance. High albedo.
d. dry ground	Parched soil. Low admittance. Large Bowen ratio. Increased albedo.
w. wet ground	Waterlogged soil. High admittance. Small Bowen ratio. Reduced albedo.

Figure A5.1. Local climate zones (LCZ) classification (Stewart & Oke, 2012).



R2 of random 10-fold CV model

	Stepwise	Stepwise.K	PLS.UK	ML	ML.K
traditional_CO	0.22	0.74	0.33	0.17	0.74
new_CO	0.13	0.71	0.35	0.23	0.67
all_CO	0	0.67	0.51	0.22	0.8
traditional_PM10	0.26	0.49	0.7	0.55	0.81
new_PM10	0.39	0.64	0.64	0.53	0.75
all_PM10	0.45	0.71	0.71	0.64	0.83
traditional_SO2	0	0.64	0.22	0.1	0.68
new_SO2	0	0.54	0.11	0	0.61
all_SO2	0	0.53	0.2	0.16	0.59

RMSE of random 10-fold CV model

	Stepwise	Stepwise.K	PLS.UK	ML	ML.K
traditional_CO	0.27	0.15	0.25	0.27	0.15
new_CO	0.28	0.16	0.24	0.25	0.16
all_CO	0.3	0.17	0.21	0.27	0.14
traditional_PM10	0.43	0.35	0.27	0.32	0.21
new_PM10	0.39	0.29	0.3	0.34	0.24
all_PM10	0.37	0.27	0.26	0.31	0.21
traditional_SO2	0.74	0.45	0.65	0.65	0.39
new_SO2	0.78	0.5	0.7	0.77	0.46
all_SO2	0.84	0.51	0.66	0.61	0.42

R2 of spatial 10-fold CV model

	Stepwise	Stepwise.K	PLS.UK	ML	ML.K
traditional_CO	0.07	0.52	0.3	0.14	0.66
new_CO	0	0.49	0.33	0.14	0.6
all_CO	0	0.53	0.39	0.18	0.71
traditional_PM10	0	0.34	0.52	0.41	0.55
new_PM10	0.14	0.46	0.46	0.37	0.62
all_PM10	0.27	0.51	0.55	0.35	0.57
traditional_SO2	0	0.48	0.15	0	0.51
new_SO2	0	0.38	0.08	0	0.24
all_SO2	0	0.46	0.18	0	0.57

RMSE of spatial 10-fold CV model

	Stepwise	Stepwise.K	PLS.UK	ML	ML.K
traditional_CO	0.29	0.21	0.25	0.28	0.18
new_CO	0.35	0.22	0.25	0.28	0.19
all_CO	0.35	0.21	0.24	0.27	0.16
traditional_PM10	0.5	0.4	0.34	0.09	0.08
new_PM10	0.46	0.36	0.36	0.39	0.31
all_PM10	0.42	0.35	0.33	0.4	0.32
traditional_SO2	0.77	0.53	0.68	0.78	0.52
new_SO2	0.87	0.58	0.71	0.87	0.64
all_SO2	0.85	0.54	0.67	0.75	0.48

Figure A5.2. Random and spatial 10-fold CV results of criteria pollutants (CO, PM_{10}, and SO_2).

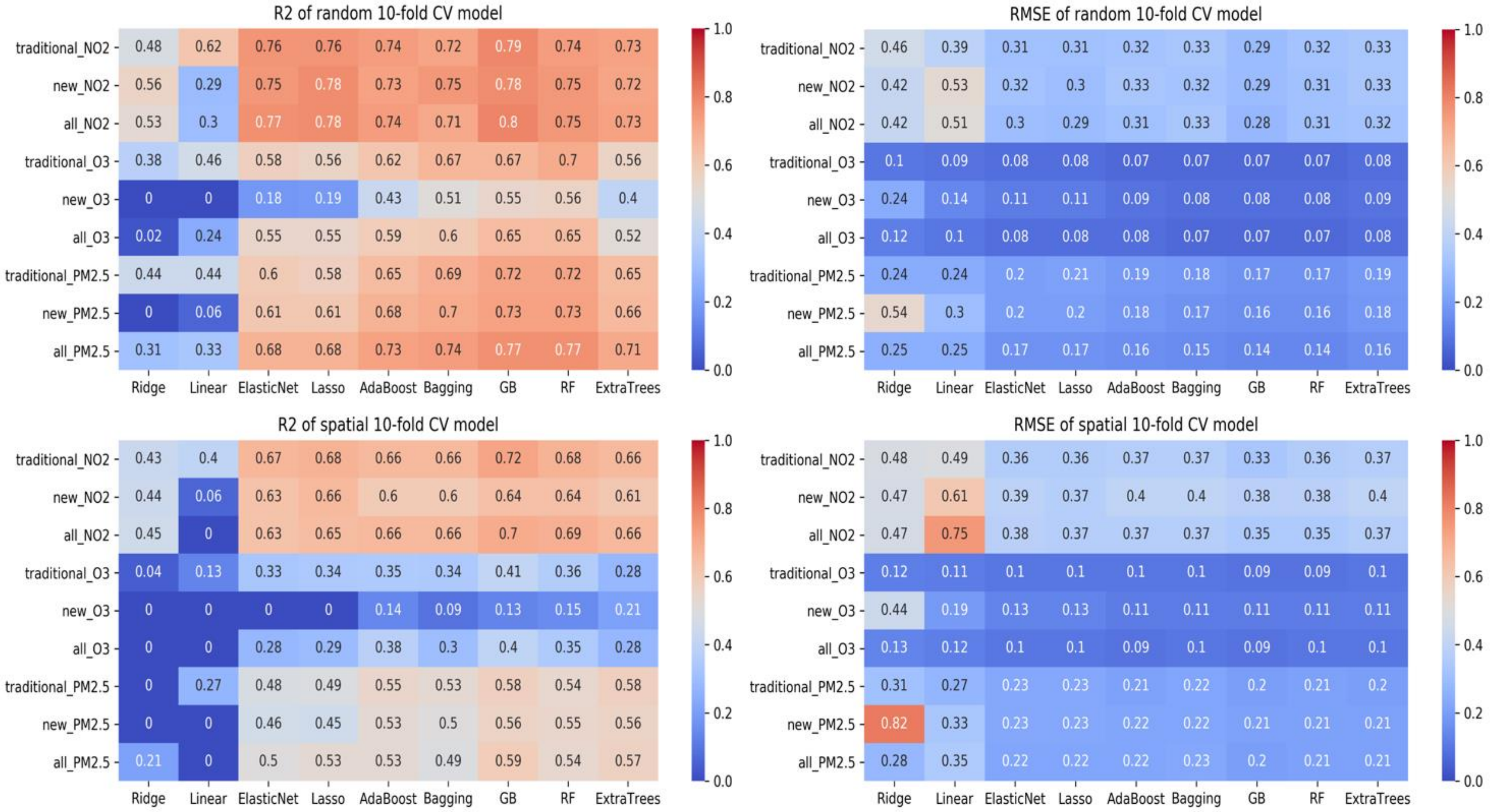


Figure A5.3. Random and spatial 10-fold CV results of major criteria pollutants of the nine ML algorithms (NO_2, O_3, and $PM_{2.5}$). GB stands for Gradient Boosting while RF stands for Random Forest.

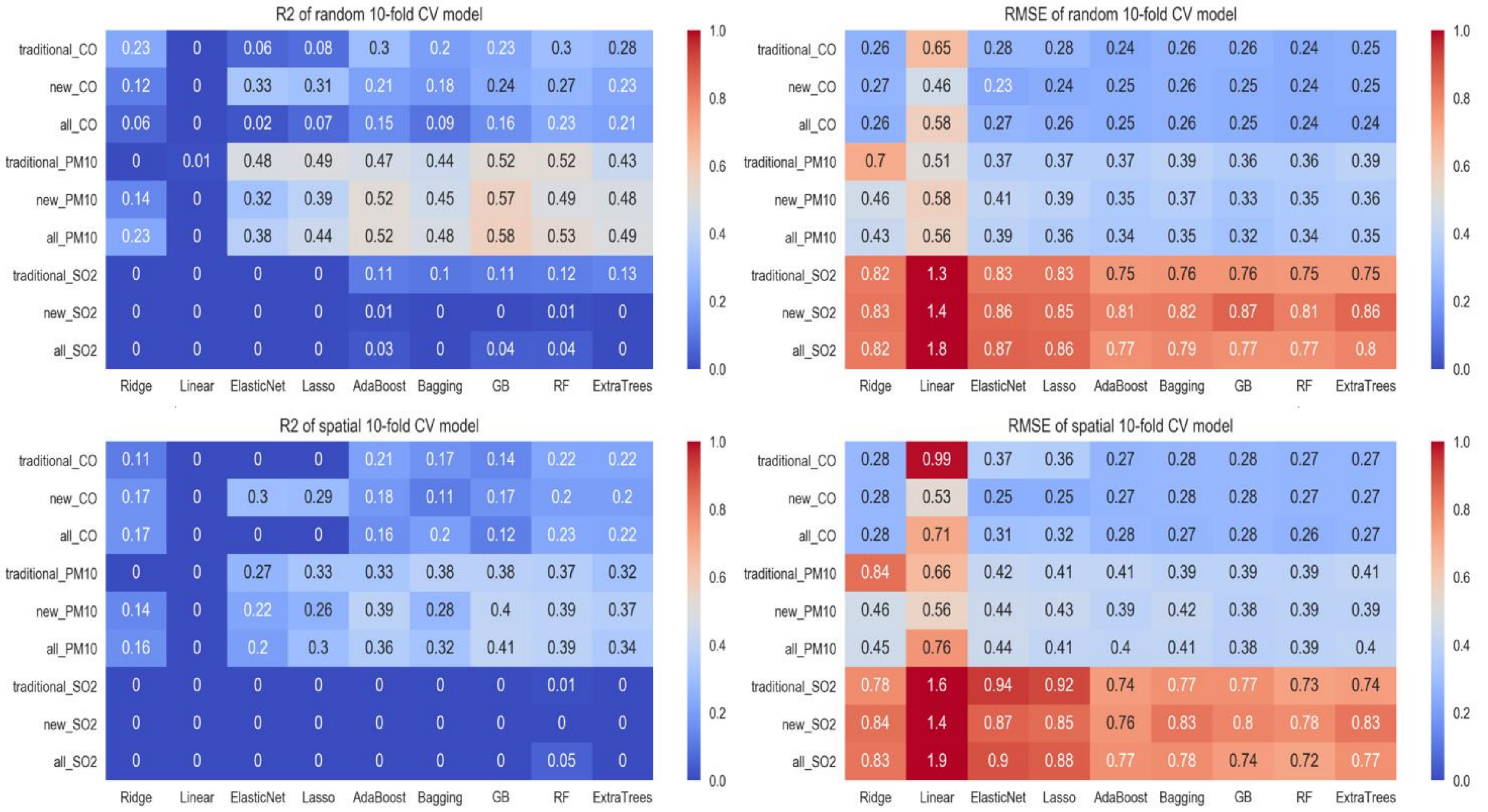


Figure A5.4. Random and spatial 10-fold CV results of criteria pollutants of the nine ML algorithms (CO, PM_{10}, and SO_2). GB stands for Gradient Boosting while RF stands for Random Forest.

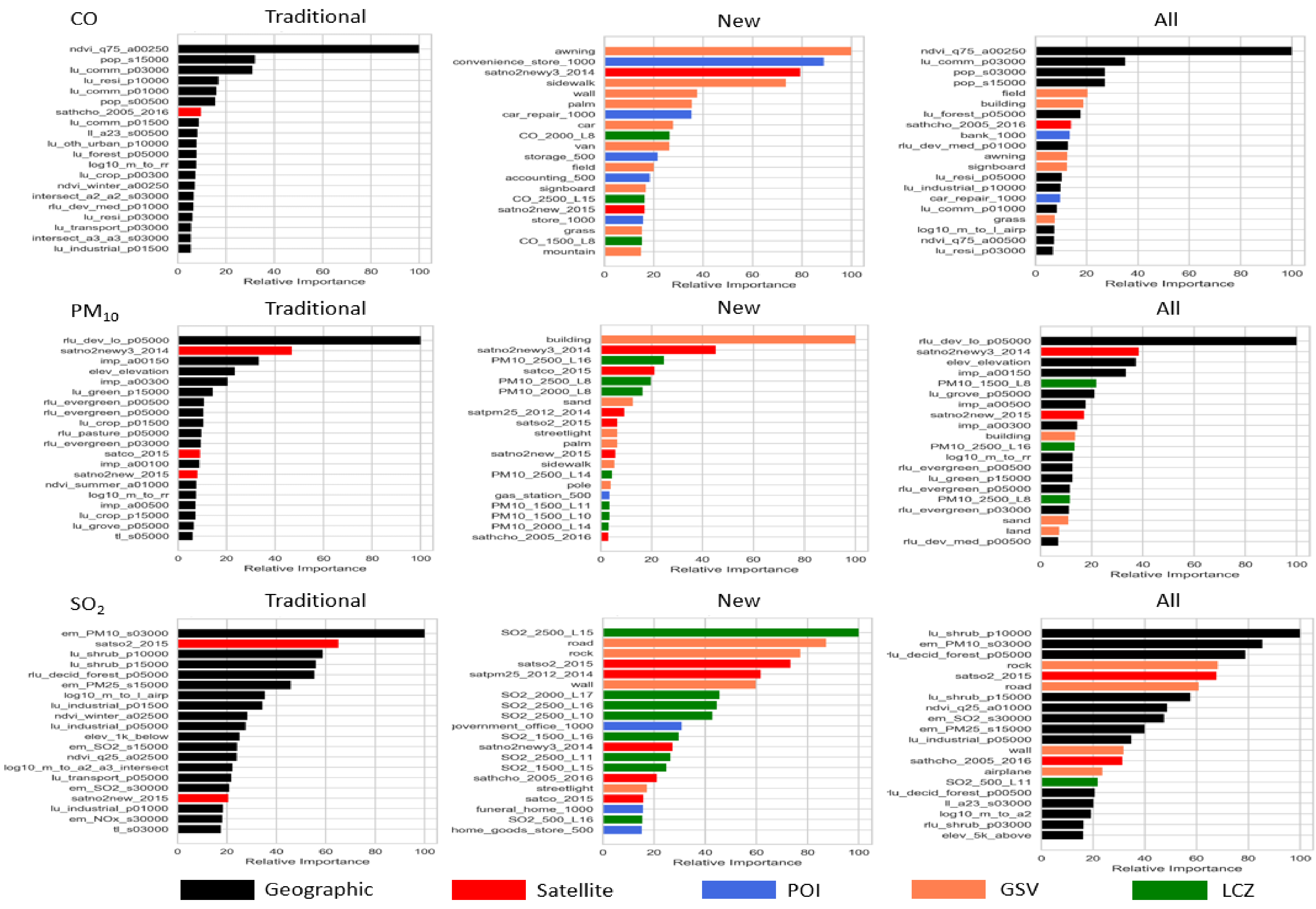


Figure A5.5. Top 20 most important features of the ML models for CO, PM_{10}, and SO_2.

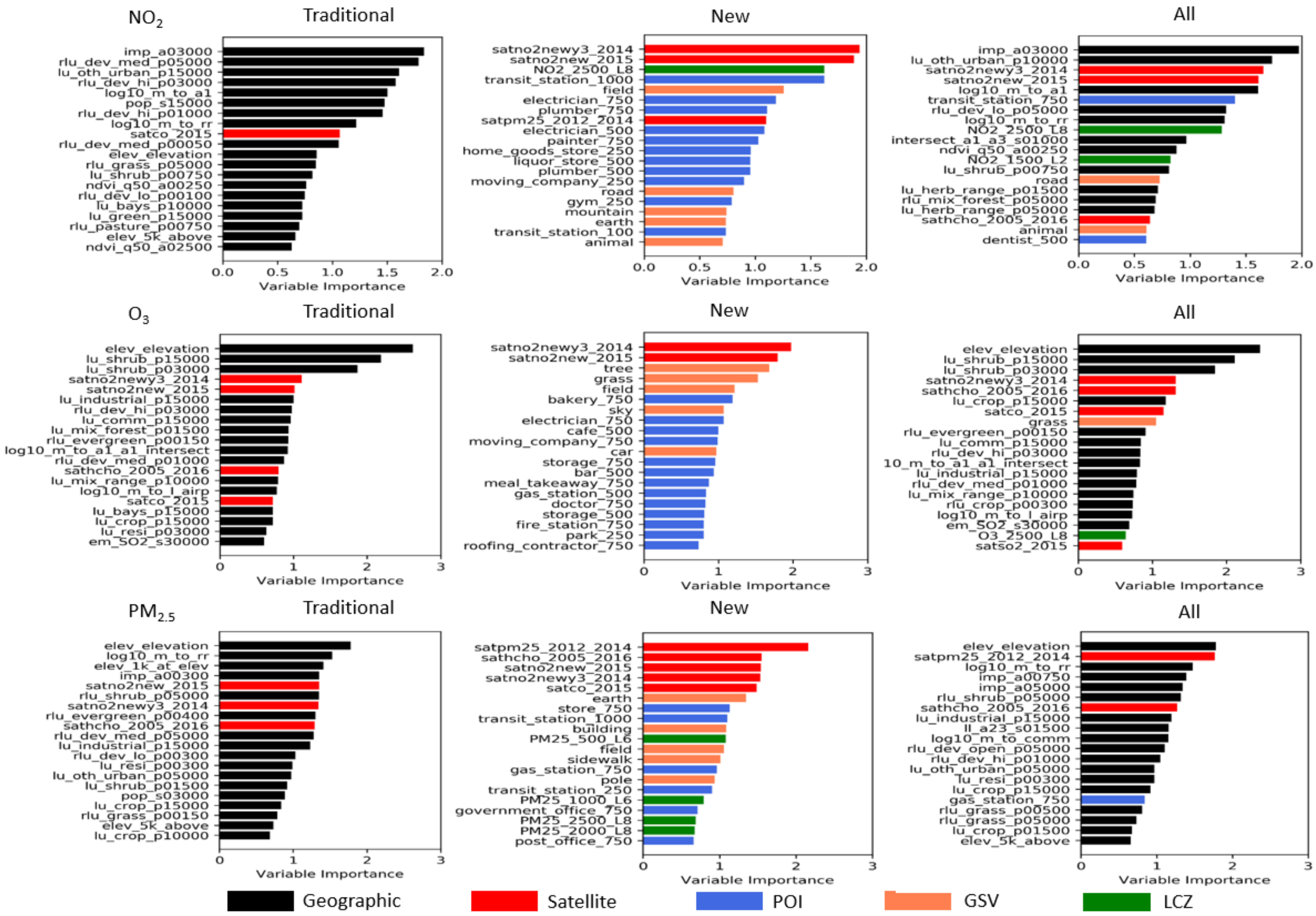


Figure A5.6. Top 20 most important features of the PLS-UK models for NO_2, O_3, and $PM_{2.5}$.

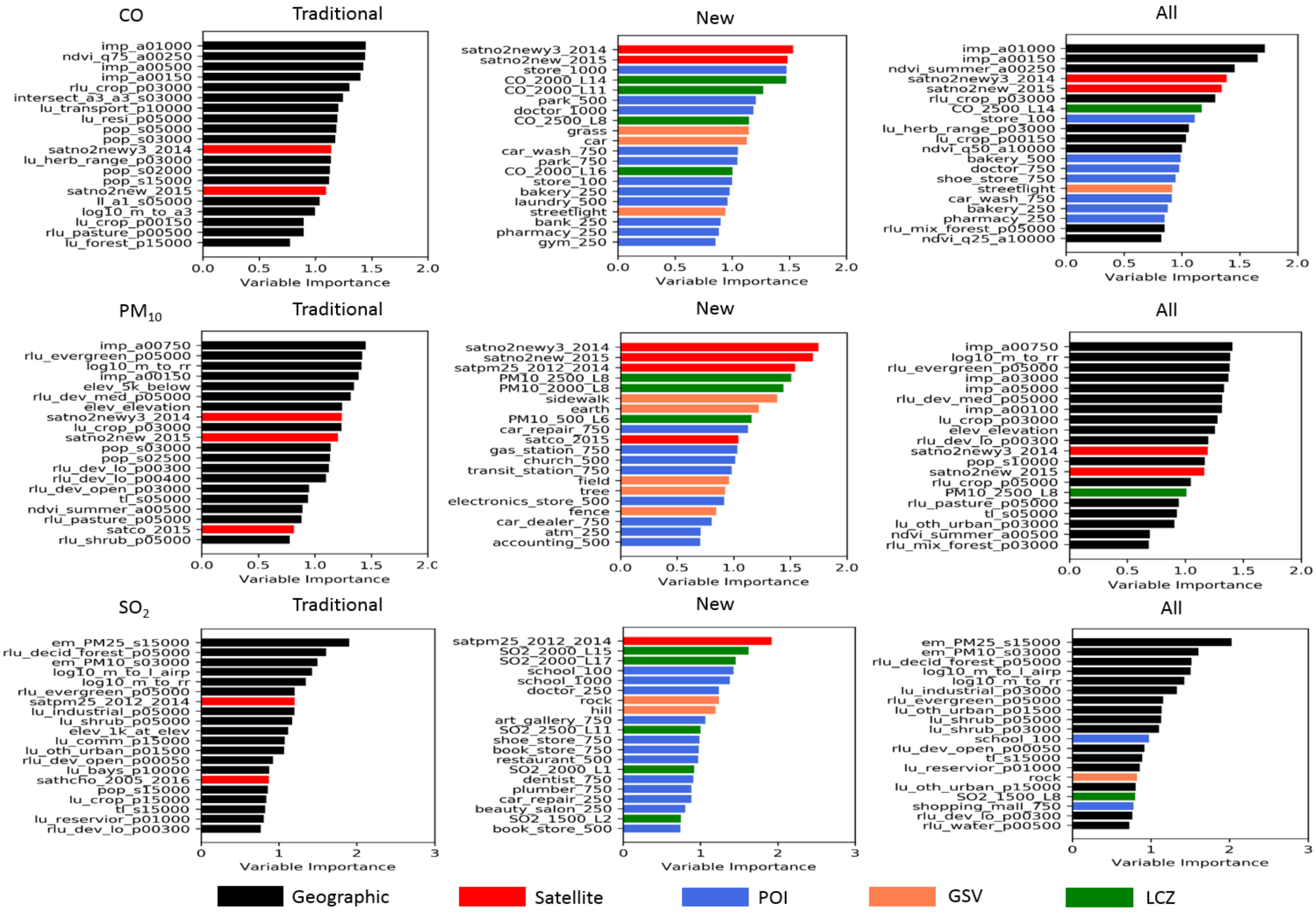


Figure A5.7. Top 20 most important features of the PLS-UK models for CO, PM_{10}, and SO_2.

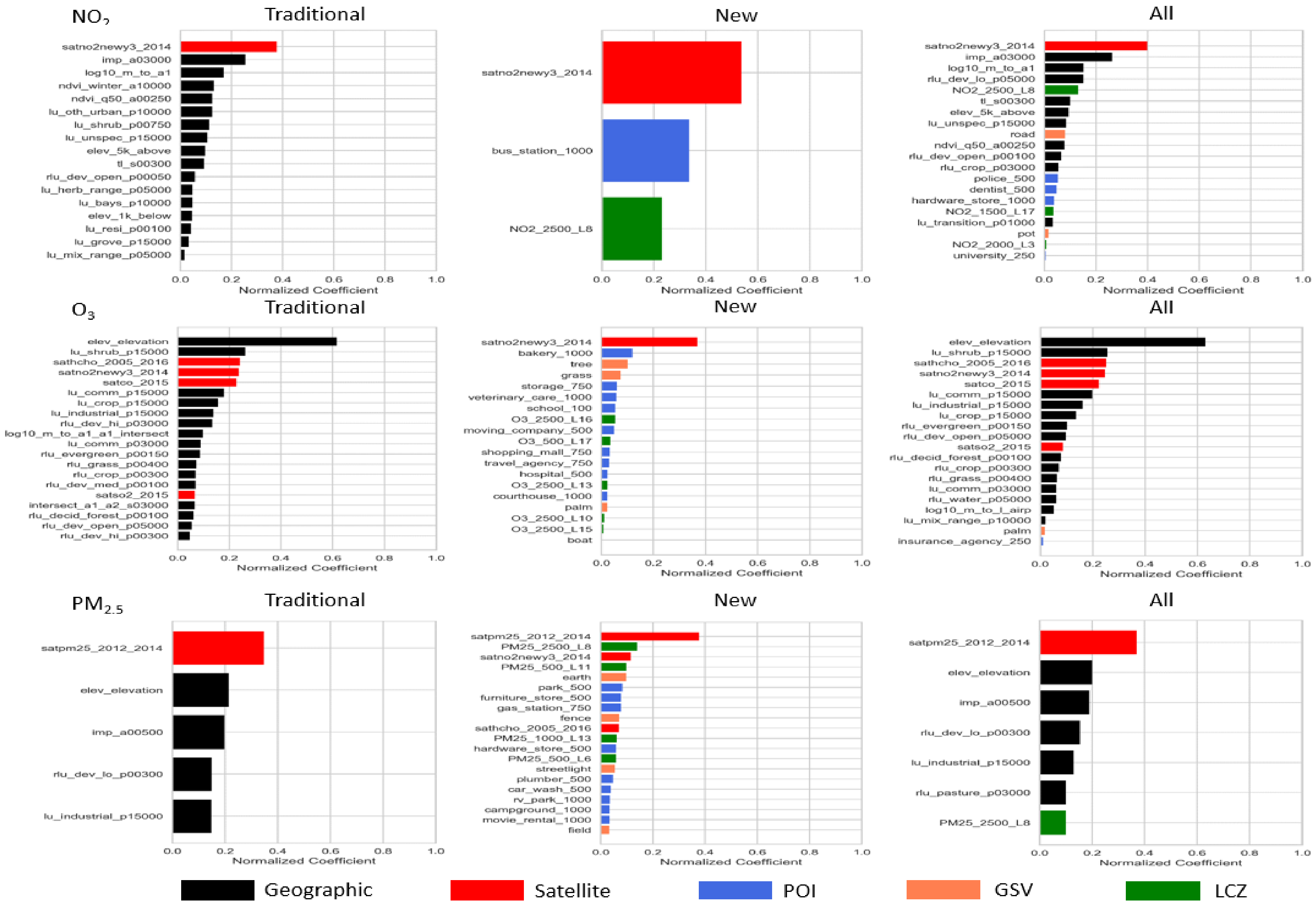


Figure A5.8. Top 20 most important features of the stepwise regression models for NO_2, O_3, and $PM_{2.5}$.

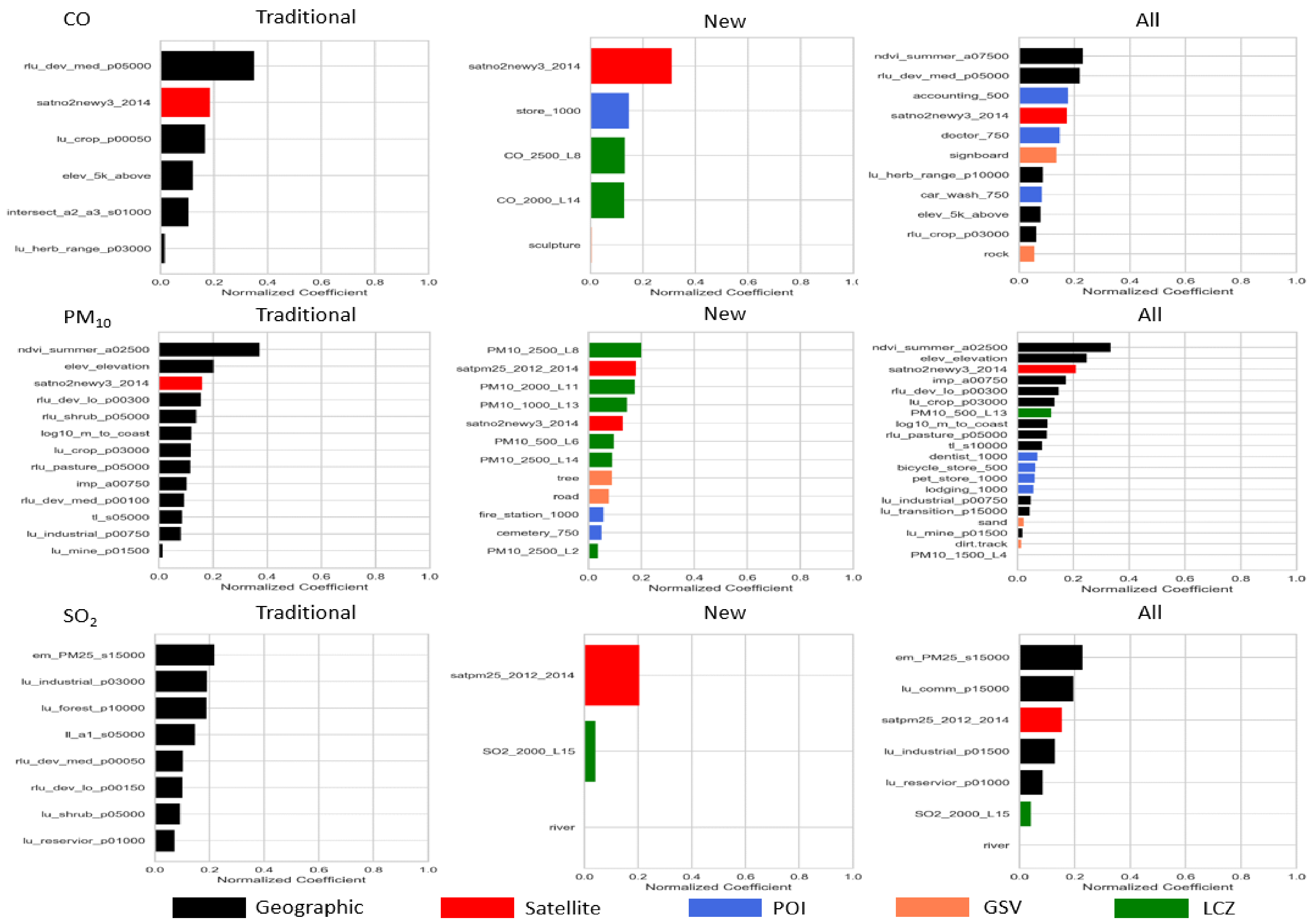


Figure A5.9. Top 20 most important features of the stepwise regression models for CO, PM_{10}, and SO_2.