Science of the Total Environment 677 (2019) 131–141

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

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

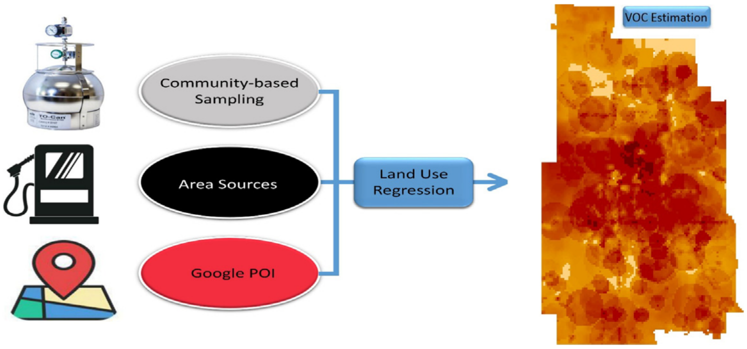
Tianjun Lu [a], Jennifer Lansing [b], Wenwen Zhang [a], Matthew J. Bechle [c], Steve Hankey [a],*

[a] School of Public and International Affairs, Virginia Tech, 140 Otey Street, Blacksburg, VA 24061, United States
[b] Minneapolis Health Department, 250 S. Fourth Street, Minneapolis, MN 55415, United States
[c] Department of Civil & Environmental Engineering, University of Washington, 201 More Hall, Seattle, WA 98195, United States

HIGHLIGHTS

- Land Use Regression (LUR) models for 60 VOC species in Minneapolis, MN
- Community-based sampling can be used for VOC LUR modeling.
- Area sources were important predictors of VOCs at small buffer sizes.
- Online mapping data could provide a useful input for LUR modeling.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 13 February 2019
Received in revised form 15 April 2019
Accepted 19 April 2019
Available online 24 April 2019

Editor: Elena Paoletti

Keywords:
Hazardous air pollutants
Volunteer-based monitoring
Local emissions
Exposure assessment

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, we incorporate area sources (e.g., gas stations) from city permitting data and Google Point of Interest (POI) data to compare model performance. We used measurements from 50 community-based sampling locations (2013–2015) in Minneapolis, MN, USA to develop LUR models for 60 VOCs. We 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 $\mu g/m^3$) 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., 25 m–500 m). Our 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.

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 locations

* Corresponding author.
E-mail address: hankey@vt.edu (S. Hankey).

https://doi.org/10.1016/j.scitotenv.2019.04.285

without monitoring data (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; 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 S1) 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 and 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., 2017a, 2017b, 2017c); only one existing LUR model (in Canada) has successfully assessed national VOC concentrations (Hystad et al., 2011a, 2011b). Previous VOC LUR models are often limited by the (1) number of monitors ($n < 40$; Atari and 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., 2017a, 2017b, 2017c; Atari and Luginaah, 2009; Carr et al., 2002; Fernández-Somoano et al., 2011; Gaeta et al., 2016; Hystad et al., 2011a, 2011b; Johnson et al., 2010; Mukerjee 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 and 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., 2017a, 2017b, 2017c). Few VOC LUR studies attempt to account for the emissions from area sources (Amini et al., 2017a, 2017b, 2017c; Hystad et al., 2011a, 2011b), 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., 2017a, 2017b, 2017c). 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 (a description of the Google POI data is below).

In this paper, we 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, our 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. We 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.

2. Materials and methods

2.1. Community-based sampling campaign for LUR development

We 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 60 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 the 60 VOCs. 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). We 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 Fig. S1).

2.2. Dependent variables for LUR

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., 2017a, 2017b, 2017c). We 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 we 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 SI. 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., 2016a, 2016b). We replaced all non-detects with half of the method detection limit of the sampling approach following U.S. EPA guidance (EPA, 1991).

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, we report LUR model results of annual-average VOC concentrations for the second year as our core model scenario. We also developed a number of alternative modeling scenarios as described in the sensitivity analysis 2.5.1. Table 1 shows the summary statistics for our core modeling scenario for the four priority VOCs. A map of sampling locations and summary statistics for all 60 VOCs are shown in Fig. S2 and Table S2.

2.2.3. Correlation matrix for all VOCs

We developed a correlation matrix using the measurements among VOC species. We 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. We decided to model the 60 VOCs separately in this paper; however, future work might assess when it is appropriate to model species together. We 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., 2017a, 2017b, 2017c; Atari and Luginaah, 2009; Kheirbek et al., 2012; Mukerjee et al., 2012). The correlation among the priority VOCs was below 0.32. Fig. S3 shows the correlation matrix for all VOCs.

2.3. Independent variables for LUR

We 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 included sources as of November 2013. To explore an alternative data source for area sources, we 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, we used a Python script (shown in the SI) 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; we used 16 buffer lengths (25 m, 50 m, 75 m, 100 m, 150 m, 200 m, 250 m, 300 m, 400 m, 500 m, 750 m, 1000 m, 1500 m, 2000 m, 3000 m, 5000 m) following a previous LUR study in Minneapolis (Hankey and 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 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).

2.4. LUR model building

Our 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. We allowed for only one buffer size to be selected for each variable to further avoid collinearity (Wilton et al., 2010). We report model performance based on adjusted R^2, Root Mean Square Error (RMSE), and 10-fold cross validated (10-fold CV) R^2. We 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), we 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, we 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, we replaced area source city permit data with Google POI data (while still including the transportation and land use variables).

We mapped model estimates of VOC concentrations (100 m × 100 m grid) for all three types of LUR models using ArcGIS 10.6 to assess spatial patterns among models. We tabulated the independent variables (variables that were significant in our 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, we fully normalized the model coefficients by multiplying each coefficient by the 95th–5th percentile difference of the

Table 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 μg/m^3.
[b] Interquartile range; number of locations is 50.

Table 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.

independent variable divided by the 95th–5th percentile difference of the dependent variable.

2.5. Sensitivity analysis

We 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.

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

We 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 S3.

2.5.2. Seasonal LURs

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

2.6. Model validation

We examined Cook's distance to identify potential outliers that may influence our model results. We 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. Results and discussion

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

3.1. LUR model results for priority VOCs

We 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. We 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 shows model results for the priority VOCs. Table S4 shows model results for all 60 VOCs.

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. We 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 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., 2017a, 2017b, 2017c; Atari and 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). We 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., 2016a, 2016b). 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

Table 3
LUR model coefficients for the priority VOCs.

Category	Variable	Base-case: No Area Sources				Area Sources: City Permit Data				Area Sources: Google POI			
		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.

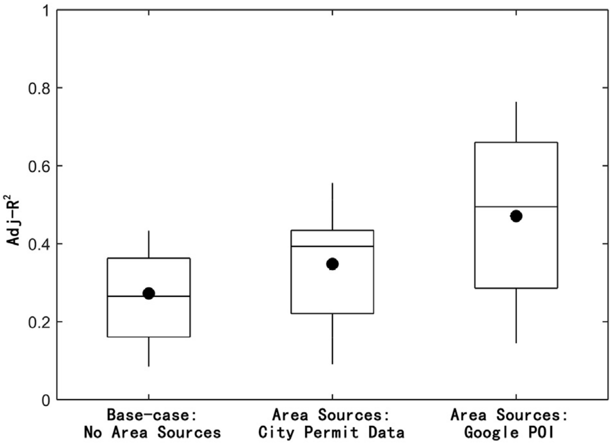


Fig. 1. LUR model performance among 60 VOC species. The three input datasets represent the addition of area source information as candidate variables.

source for LUR modeling. Fig. 1 shows a summary of model performance among all 60 VOCs.

3.1.2. City permit data vs. Google POI

We developed LUR models by including four categories of area sources from two data sources (city business permit and Google POI). We 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 we 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 500 m 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 S5 shows the average number of area source locations and coefficient of variation for the city permit data and the Google POI data.

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). This choice of spatial predictors is similar to other studies that assess these VOCs (Amini et al., 2017a, 2017b, 2017c; Atari and 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., 2017a, 2017b, 2017c) 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., 2011a, 2011b). A unique aspect of our models is the capability to compare the area sources that may be important to explain the variations in VOC concentrations. To support our findings, we 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 Fig. S4).

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 our 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 and 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.

3.1.5. Buffer sizes of significant variables

Almost all area sources were selected at small buffer sizes (e.g., 25 m–500 m) 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., 2017a, 2017b, 2017c). 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 5000 m while TVOC selected road classifications at a buffer length at 3000 m. The choice of these large buffer sizes is consistent with a suggestion to include traffic-related variables at buffers up to 5000 m from a recent VOC review (Amini et al., 2017a, 2017b, 2017c). BTEX was also associated with lower road hierarchy (e.g., collectors) at smaller buffer lengths (e.g., 100 m), 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.

3.1.6. Mapping concentration estimates

We mapped VOC concentrations for the entire city of Minneapolis on a 100 × 100 meter grid. For the purpose 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 our 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., 2017a, 2017b, 2017c), (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). Fig. 2 shows model estimates for the priority VOCs using all three types of models in Minneapolis, MN. Fig. 3 shows scatterplots of the predicted vs. observed values for the priority VOCs.

3.2. Sensitivity analysis

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

We 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 our core model scenario (second year; $n = 50$ sampling locations; Figs. S5 and S6). Our core scenario is consistent with the adequate number of sampling locations ($n =$ ~40–80) recommended for LUR modeling over small geographic areas (Hoek et al., 2008).

3.2.2. Seasonal LUR models

We developed seasonal LUR models and compared performance to the annual-average models for the priority VOCs (see Fig. S7). 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. Model validation

3.3.1. 10-Fold CV

In general, 10-fold CV R^2 values (Table 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, our sample size was small ($n = 50$) and this result should be tested in studies with more sampling data available for VOC-specific modeling.

3.3.2. Cook's distance and spatial autocorrelation

Examination of Cook's distance for our priority VOCs confirmed that no significant outliers influenced our model results (see Table S6). Based on the Moran's I test, no significant spatial residual correlation

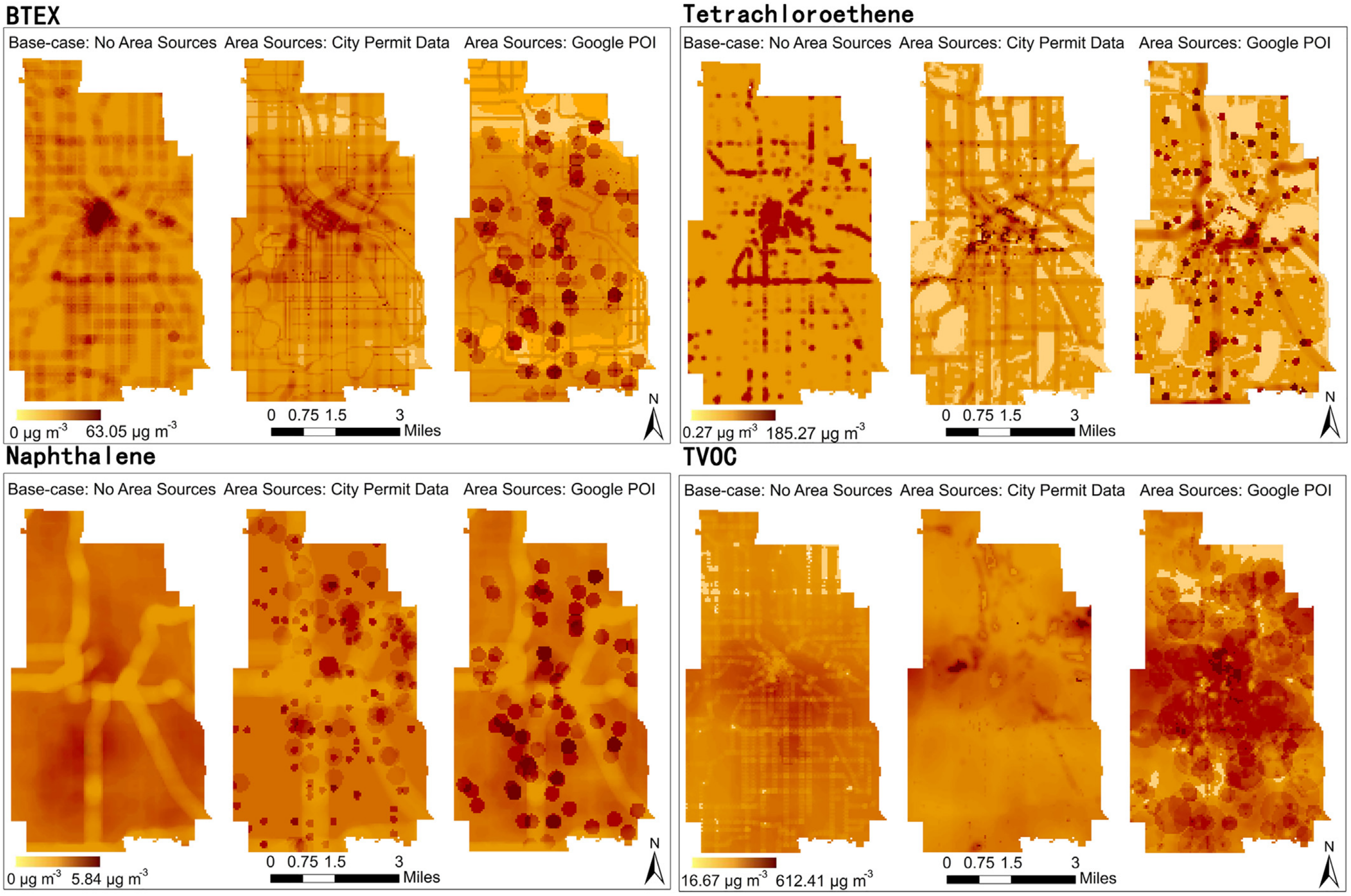


Fig. 2. LUR model estimates for the priority VOCs among model types in Minneapolis, MN.

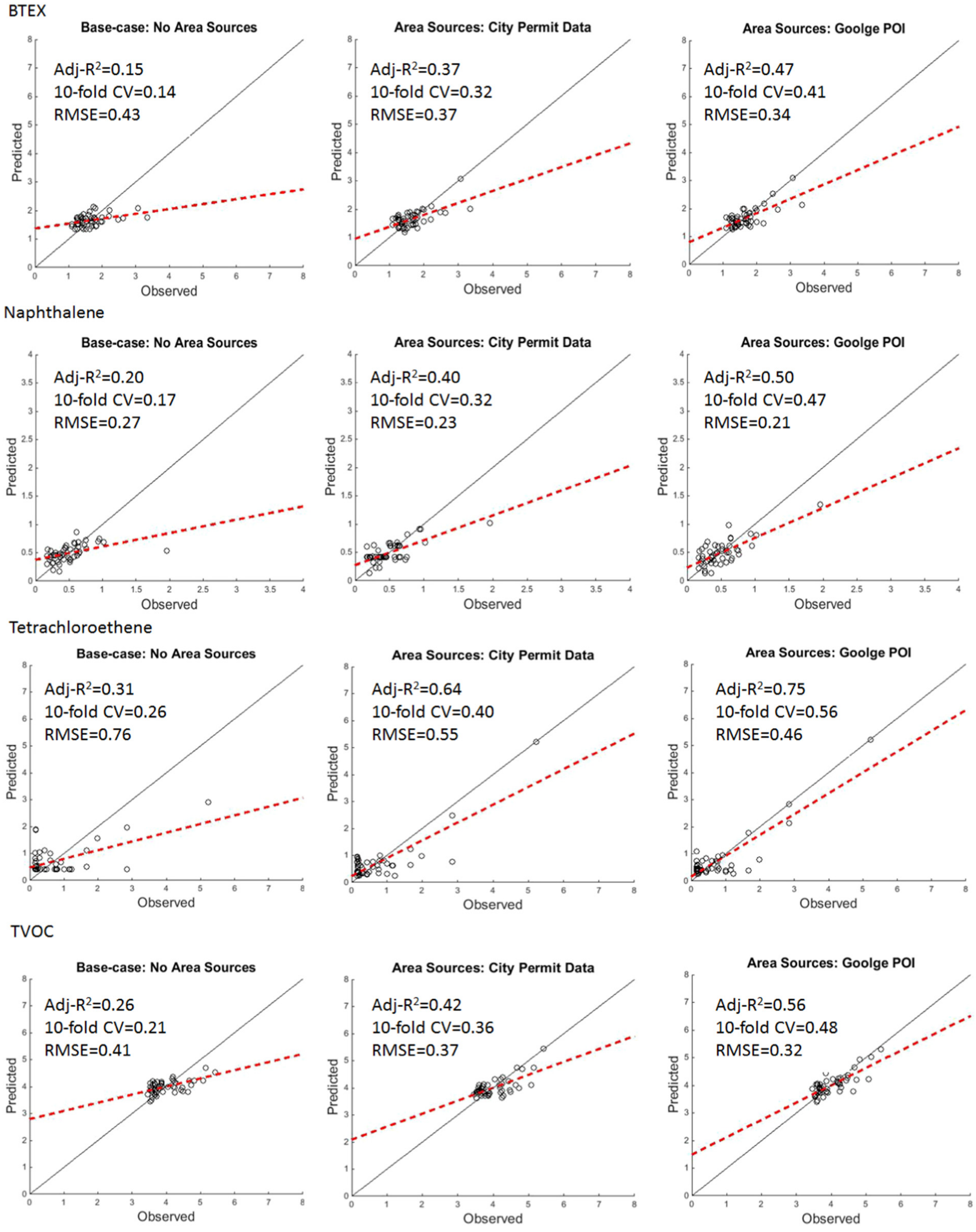


Fig. 3. Scatterplots of predicted vs. observed values for the priority VOCs. Solid black lines represent the 1:1 line; dashed red-lines represent the best fit line.

was found in the LUR models for the priority VOCs except for some instances in the BTEX models (Table S7). 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). We 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 Fig. S7). 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.

3.4. Implications for developing VOC LUR models

3.4.1. Implications for modeling 60 VOCs

To our 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 S1). We were able to model 60 VOCs to explore how different VOCs show varying spatial patterns. Our 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 our targeted modeling strategy for individual VOCs in certain cases (Lansing et al., 2016).

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. Our 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 our study is that out of 186 total locations, we were only able to use 50 locations for modeling due to a lack of sampling data during specific sampling events among locations; our community-based sampling may also be limited by the fact that its original purpose was not for LUR. Our 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., 2017a, 2017b, 2017c; Atari and Luginaah, 2009; Kheirbek et al., 2012; Su et al., 2010; Wheeler et al., 2008).

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., 2017a, 2017b, 2017c). Comparatively, our study analyzed 60 VOCs and presented a more comprehensive assessment of different VOC species in our LUR models.

A contribution of our models is that adding area sources helps to assess whether these small-scale sources are correlated with VOC concentrations. We were 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, we 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 the 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., 2017a, 2017b, 2017c); 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., 2017a, 2017b, 2017c).

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. We used two measures of area source data (Google POI and city permit data) to explore the potential benefits of each dataset. We 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. Our LUR models show that online mapping data could provide a useful input for LUR modeling. We only included variables from the Google POI database that were closely matched to the categories of area sources we had from the city business permit database; future work could replicate and expand our approach by including all data available in the Google POI database.

Our modeling approach of combining an open dataset (Google POI) for 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., 2011a, 2011b). Our 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 and 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.

4. Conclusion

We 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. We were able to assemble three sets of independent variables to develop our 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. We found that models with the Google POI area source data performed better as compared to the base-case model and the permit data model. We 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. Our work suggests that community-based sampling could be used as a valuable input for LUR models to estimate VOC concentrations. Our study explores the spatial patterns of a wide breadth 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. Our work could be used to inform planning policies to reduce emissions from area sources.

Acknowledgments

This publication was developed as part of the Center for Air, Climate, and Energy Solutions (CACES), which was supported under Assistance Agreement No. R835873 awarded by the U.S. Environmental Protection Agency. It has not been formally reviewed by EPA. The views expressed in this document are solely those of authors and do not necessarily reflect those of the Agency. EPA does not endorse any products or commercial services mentioned in this publication. The authors acknowledge Patrick Hanlon for providing the VOC sampling data and funding from Minneapolis Health Department.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2019.04.285.

References

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. Environ. Sci. Technol. 42 (3), 815–821. https://doi.org/10.1021/es0715492.

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

Amini, H., Schindler, C., Hosseini, V., Yunesian, M., 2017b. Land use regression models for alkylbenzenes in a Middle Eastern megacity: Tehran Study of Exposure Prediction for Environmental Health Research (Tehran SEPEHR). Environ. Sci. Technol. 51, 8481–8490. https://doi.org/10.1021/acs.est.7b02238.

Amini, H., Yunesian, M., Hosseini, V., Schindler, C., Sarah, B., Künzli, N., 2017c. A systematic review of land use regression models for volatile organic compounds. Atmos. Environ. 171, 1–16. https://doi.org/10.1016/j.atmosenv.2017.10.010.

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

Atari, D.O., Luginaah, I.N., 2009. Assessing the distribution of volatile organic compounds using land use regression in Sarnia, “Chemical Valley”, Ontario, Canada. Environ. Health 14, 1–14. https://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). Environ. Sci. Technol. 32 (3), 405–412. https://doi.org/10.1021/es970008h.

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. Sci. Total Environ. 407 (6), 1852–1867. https://doi.org/10.1016/j.scitotenv.2008.11.048.

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. Atmos. Environ. 44 (36), 4614–4621. https://doi.org/10.1016/j.atmosenv.2010.08.005.

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.

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

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. Environ. Res. 118 (2), 111–118. https://doi.org/10.1006/enrs.2002.4393.

Chang, S.J., Chen, C.J., Lien, C.H., Sung, F.C., 2006. Hearing loss in workers exposed to toluene and noise. Environ. Health Perspect. 114 (8), 1283–1286. https://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. J. Environ. Manag. 166, 440–449. https://doi.org/10.1016/j.jenvman.2015.10.050.

Conrad, C.C., Hilchey, K.G., 2011. A review of citizen science and community-based environmental monitoring: issues and opportunities. Environ. Monit. Assess. 176 (1–4), 273–291. https://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. https://doi.org/10.1016/j.geoforum.2009.01.003.

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). Atmos. Environ. 45 (29), 5240–5246. https://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. 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. Atmos. Environ. 131, 254–262. https://doi.org/10.1016/j.atmosenv.2016.01.052.

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. https://doi.org/10.1097/01.ede.0000082001.05563.e0.

Guerreiro, C.B.B., Foltescu, V., de Leeuw, F., 2014. Air quality status and trends in Europe. Atmos. Environ. 98, 376–384. https://doi.org/10.1016/j.atmosenv.2014.09.017.

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. Environ. Sci. Technol. 49, 9194–9202. https://doi.org/10.1021/acs.est.5b01209.

Hochadel, M., Heinrich, J., Gehring, U., Morgenstern, V., Kuhlbusch, T., Link, E., ... Krämer, U., 2006. Predicting long-term average concentrations of traffic-related air pollutants using GIS-based information. Atmos. Environ. 40 (3), 542–553. https://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. https://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. Atmos. Environ. 42 (33), 7561–7578. https://doi.org/10.1016/j.atmosenv.2008.05.057.

Hystad, P., Setton, E., Cervantes, A., Poplawski, K., 2011a. Creating national air pollution models for population exposure assessment in Canada. Environ. Health Perspect. 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., 2011b. Creating national air pollution models for population exposure assessment in Canada. Environ. Health Perspect. 119 (8), 1123–1129. https://doi.org/10.1289/ehp.1002976.

Jerrett, M., Arain, A., Kanaroglou, P., Beckerman, B., 2005. A review and evaluation of intraurban air pollution exposure models. J. Expo. Anal. Environ. Epidemiol. 15, 185–204. https://doi.org/10.1038/sj.jea.7500388.

Jia, C., Batterman, S., 2010. A critical review of naphthalene sources and exposures relevant to indoor and outdoor air. Int. J. Environ. Res. Public Health 7 (7), 2903–2939. https://doi.org/10.3390/ijerph7072903.

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. Atmos. Environ. 44 (30), 3660–3668. https://doi.org/10.1016/j.atmosenv.2010.06.041.

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. Environ. Health 11 (1), 51.

Kim, Y.M., Harrad, S., Harrison, R.M., 2001. Concentrations and sources of VOCs in urban domestic and public microenvironments. Environ. Sci. Technol. 35 (6), 997–1004. https://doi.org/10.1021/es000192y.

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. Environ. Sci. Technol. 40 (13), 4074–4082. https://doi.org/10.1021/es051828u.

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.

Larson, T., Su, J., Baribeau, A.M., Buzzelli, M., Setton, E., Brauer, M., 2007. A spatial model of urban winter woodsmoke concentrations. Environ. Sci. Technol. 41 (7), 2429–2436. https://doi.org/10.1021/es0614060.

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. Am. J. Epidemiol. 159 (3), 294–303. https://doi.org/10.1093/aje/kwh043.

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. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 185–194 (doi: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. Atmos. Environ. 41 (35), 7500–7511. https://doi.org/10.1016/j.atmosenv.2007.05.039.

Marshall, J.D., Nethery, E., Brauer, M., 2008. Within-urban variability in ambient air pollution: comparison of estimation methods. Atmos. Environ. 42, 1359–1369. https://doi.org/10.1016/j.atmosenv.2007.08.012.

Mečiarová, L., Vilčeková, S., Burdová, E.K., Kiselák, J., 2017. Factors effecting the total volatile organic compound (TVOC) concentrations in Slovak households. Int. J. Environ. Res. Public Health 14 (12). https://doi.org/10.3390/ijerph14121443.

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. Sci. Total Environ. 407 (16), 4642–4651. https://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. Sci. World J. 2012. https://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. Atmos. Environ. 30 (20), 3457–3470. https://doi.org/10.1016/1352-2310(95)00487-4.

Oiamo, T.H., Johnson, M., Tang, K., Luginaah, I.N., 2015. Assessing traffic and industrial contributions to ambient nitrogen dioxide and volatile organic compounds in a low pollution urban environment. Sci. Total Environ. 529, 149–157. https://doi.org/10.1016/j.scitotenv.2015.05.032.

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. Atmos. Environ. 37 (36), 5023–5046. https://doi.org/10.1016/j.atmosenv.2003.08.006.

Piccot, S.D., Watson, J.J., Jones, J.W., 1992. A global inventory of volatile organic compound emissions from anthropogenic sources. J. Geophys. Res. Atmos. 97 (D9), 9897–9912. https://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. J. Occup. Environ. Med. 57 (12), 1291–1298. https://doi.org/10.1097/JOM.0000000000000604.

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. Atmos. Environ. 41 (11), 2255–2269. https://doi.org/10.1016/j.atmosenv.2006.11.012.

Singh, D., Kumar, A., Kumar, K., Singh, B., Mina, U., Singh, B.B., Jain, V.K., 2016a. Statistical modeling of O3, NOx, CO, PM2.5, VOCs and noise levels in commercial complex and associated health risk assessment in an academic institution. Sci. Total Environ. 572 (x), 586–594. https://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., 2016b. Spatial and temporal variability of VOCs and its source estimation during rush/non-rush hours in ambient air of Delhi, India. Air Qual. Atmos. Health 9 (5), 483–493. https://doi.org/10.1007/s11869-015-0354-3.

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. Atmos. Environ. 40, 3773–3787. https://doi.org/10.1016/j.atmosenv.2006.02.036.

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. Environ. Monit. Assess. 128 (1–3), 369–379. https://doi.org/10.1007/s10661-006-9320-8.

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. J. Environ. Monit. 13, 999–1007. https://doi.org/10.1039/c0em00724b.

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

Su, J.G., Jerrett, M., Beckerman, B., 2009. A distance-decay variable selection strategy for land use regression modeling of ambient air pollution exposures. Sci. Total Environ. 407 (12), 3890–3898. https://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. Atmos. Environ. 44 (29), 3529–3537. https://doi.org/10.1016/j.atmosenv.2010.06.015.

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 Qual. Atmos. Health 6 (1), 85–94. https://doi.org/10.1007/s11869-011-0148-1.

Sun, J., Wu, F., Hu, B., Tang, G., Zhang, J., Wang, Y., 2016. VOC characteristics, emissions and contributions to SOA formation during hazy episodes. Atmos. Environ. 141, 560–570. https://doi.org/10.1016/j.atmosenv.2016.06.060.

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%20RECORD/N60191_000280.pdf.

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. Am. J. Epidemiol. 179 (4), 443–451. https://doi.org/10.1093/aje/kwt289.

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

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. Environ. Res. 106 (1), 7–16. https://doi.org/10.1016/j.envres.2007.09.004.

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. Sci. Total Environ. 408 (5), 1120–1130. https://doi.org/10.1016/j.scitotenv.2009.11.033.

World Health Organization (WHO), 2000. Air Quality Guidelines for Europe. 2nd edition. https://doi.org/10.1007/BF02986808.