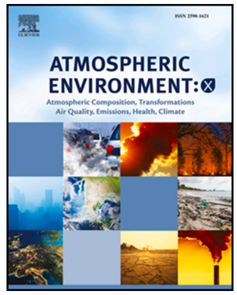
Journal Pre-proof



A spatial land use clustering framework for investigating the role of land use in mediating the effect of meteorology on urban air quality

Amir Montazeri, Achim J. Lilienthal, John D. Albertson

PII: S2590-1621(21)00026-5

DOI: https://doi.org/10.1016/j.aeaoa.2021.100126

Reference: AEAOA 100126

To appear in: *Atmospheric Environment: X*

Received Date: 29 April 2021

Revised Date: 27 July 2021

Accepted Date: 16 August 2021

CRediT author statement

Amir Montazeri: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing – Original Draft, Writing – Review & Editing, Visualization.

Achim J. Lilienthal: Conceptualization, Methodology, Writing – Review & Editing.

John D. Albertson: Conceptualization, Methodology, Writing – Review & Editing, Funding acquisition

A SPATIAL LAND USE CLUSTERING FRAMEWORK FOR INVESTIGATING THE ROLE OF LAND USE IN MEDIATING THE EFFECT OF METEOROLOGY ON URBAN AIR QUALITY

Amir Montazeri*
Sibley School of Mechanical and Aerospace Engineering
Cornell University
Ithaca, NY, USA
email: am2774@cornell.edu
*Corresponding author at: Hollister Hall, 527 College Ave, Ithaca, NY 14853

Achim J. Lilienthal
Centre for Applied Autonomous Sensor Systems (AASS)
Örebro University
Örebro, Sweden
email: achim.lilienthal@oru.se

John D. Albertson
School of Civil and Environmental Engineering
Cornell University
Ithaca, NY, USA
email: albertson@cornell.edu

August 17, 2021

ABSTRACT

Air pollution in urban areas is driven by emission sources and modulated by local meteorology, including the effects of urban form on wind speed and ventilation, and thus varies markedly in space and time. Recently, mobile measurement campaigns have been conducted in urban areas to measure the spatial distribution of air pollutant concentrations. While the main focus of these studies has been revealing spatial patterns in mean (or median) concentrations, they have mostly ignored the temporal aspects of air pollution. However, assessing the temporal variability of air pollution is essential in understanding the integrated exposure of individuals to pollutants above critical thresholds. Here, we examine the role of urban land use in mediating the effect of regional meteorology on Nitrogen Dioxide (NO_2) concentrations measured in different regions of Oakland, CA. Inspired by Land

Use Regression (LUR) models, we cluster 30-meter road segments in the urban area based on their land use. The concentration data from the resulting clusters are stratified based on seasonality and conditionally averaged based on concurrent wind speeds. The clustering analysis yielded 7 clusters, with 4 of them chosen for further statistical analysis due to their large sample sizes. Two of the four clusters demonstrated in winter a strong negative linear relationship between NO_2 concentration and wind speed ($R^2 > 0.87$) with a slope of approximately 3 ppb/m s^{-1}. A weaker correlation and flatter slope was found for the cluster representing road segments belonging to interstate highways ($R^2 > 0.73$ and slope ¡ 2 ppb/m s^{-1}). No significant relationship was found during the summer season. These findings are consistent with the concept of strong vertical mixing due to highway traffic and increased surface heat fluxes during summer weakening the relationship between wind speed and NO_2 concentrations. In summary, the clustering analysis framework presented here provides a novel tool for use with large-scale mobile measurements to reveal the effect of urban land form on the temporal dynamics of pollutant concentrations and ultimately human exposure.

Keywords Air pollution profiles · Cluster Analysis · Mobile monitoring · Land use effects · k-means · Exceedance probabilities · Unsupervised learning · Machine learning

1 Introduction

Around the globe, exposure to air pollution causes millions of premature deaths annually [1], and is associated with chronic respiratory illnesses that increase the co-morbidity risk of many viral infections [2]. Early evidence, for example, suggests exposure to air pollution may increase mortality of COVID-19 [3]. One group of pollutants with known deleterious effects on health is Nitrogen oxides (NO_x). Nitrogen dioxide (NO_2) is commonly used as the indicator for the NO_x group and NO_2 is mainly formed by burning of fuel. Exposure to NO_2 is associated with irritation of the airways, decreased lung capacity, increased mortality from coronary heart disease, and increased incidence of diabetes, hypertension, and other cardiovascular and respiratory illnesses [4, 5]. Further, in a study of 66 administrative regions in Europe, regions with chronic exposure to NO_2 were observed to experience the highest fatality rates from COVID-19 [6]. Therefore, monitoring and mitigating exposure to NO_2 is important to public health and safety.

Traditionally, air pollution has been monitored using sparse networks of fixed stations installed in urban areas with the goal of regulatory compliance. While these fixed stations offer accurate and reliable pollutant measurements, they provide very low spatial coverage. Yet, pollutant concentrations can vary sharply over short distances due to heterogeneity in emission sources and urban form [7, 8]. In fact, it has been shown that pollutant concentrations can differ more between two neighborhoods of the same city than between two distinct cities [9]. Hence, while the networks of fixed monitoring stations remain essential for air quality regulation compliance, they fail to capture the strong spatial variability in pollutant concentrations within urban areas with strong implications for epidemiology and environmental justice [8, 10, 11, 12].

Mobile measurements show promise for overcoming the limitations of fixed-site air pollution monitoring stations [9, 13, 14, 15, 16]. The spatial flexibility of mobile measurements has led to their application in characterizing regional pollutant concentrations and in locating pollution hotspots in select locales [13, 17, 18, 19]. While early local mobile campaigns were successful in describing spatial gradients in pollutant concentrations, many of these campaigns had limited spatial domains and were conducted for relatively short time periods. Recently, city-scale mobile monitoring campaigns have become more common [8, 14, 20, 21], with vehicles outfitted with state-of-the-art sensors and deployed to cover extensive parts of urban areas over several months and years, allowing for repeated sampling of visited locations. Repeated sampling coupled with data analytics algorithms grants statistical power to construct stable, long-term spatial maps of pollutant concentrations at high resolutions over large areas [8, 14, 20]. These spatial maps are useful in depicting persistent patterns in pollutant concentrations, measuring average pollution (averaged over a year) in a region, and locating air pollution hotspots. However, temporal variability in air pollution is typically not reported, despite its vital importance for identifying the time of exposure above key concentration thresholds of human health significance [2].

Temporal dynamics of pollutant concentrations within an urban area are dependent on both the regional (city-wide) meteorology for overall atmospheric boundary layer mixing and the local meteorology, as modulated by local urban form, for its control on ground level concentrations. In other words, local land use affects the temporal dynamics of air quality by mediating the relationship between regional and local meteorology (i.e. some areas more or less ventilated than others). Meanwhile, the effects of regional meteorology on air quality are known to vary between seasons [22, 23]. Therefore, the study of the temporal variability of pollutant concentrations requires local pollutant measurements over different seasons as done in large scale air quality measurement campaigns. One such campaign was the mobile measurement effort by two Google Street View Cars in Oakland, CA, sampling ambient NO_2 concentrations with a frequency of 1-Hz over a two-year period [24]. This novel dataset provides information on pollutant concentrations of all city streets within the study domain of West Oakland, downtown Oakland, and East Oakland across different seasons and under varying meteorological conditions [8].

In this paper, we investigate the role of urban land form in mediating the effect of regional meteorology on intra-urban air quality in Oakland, CA using the Google Street View air quality dataset. To the best of our knowledge, this is the first study focusing on using city-wide mobile measurements to examine spatially varying temporal patterns in air quality due to interaction between meteorology and urban form. To this end, we developed a data-driven spatio-temporal framework built upon clustering spatial locations in Oakland, CA based on land use variables. This clustering effectively increases the temporal statistical power of mobile measurements that is required for characterizing the effect of wind speed on NO_2 concentrations and investigating exceedance probabilities. Exceedance probabilities are an important measure of exposure to extreme pollutant concentrations, with clear ties to acute effects of air pollution on human health. The main contribution of this paper is providing a framework that exploits land use variables to learn about the relationship between meteorology and intra-urban air quality using limited air pollution data from mobile sensors. The second contribution is the development of a land use clustering technique consisting of the k-means algorithm and a comprehensive procedure for selecting the number of clusters. The third contribution is the application of the framework to pre-existing data from Oakland, CA and the insightful results related to how urban form modulates the effect of wind speed on intra-urban air quality.

2 Data

Multiple datasets including meteorological data, land-use data and mobile NO_2 measurements, were analyzed in this study to investigate the effect of meteorology and land-use on air pollution levels in distinct regions of Oakland, CA, with use cases of each dataset presented in Figure 1.

2.1 Data Sources

Mobile measurements of 1-Hz NO_2 concentrations were collected in Oakland, CA in a joint effort between University of Texas at Austin, Aclima Inc., Google and the Environmental Defense Fund, details of which are available in [8]. The mobile sampling emphasized three main areas within Oakland: West Oakland ($\sim$10km^2), Downtown Oakland ($\sim$5km^2), and East Oakland ($\sim$15km^2) in addition to the highways connecting these areas. West Oakland is bounded by major interstate highways, the fifth largest container port in the U.S., and associated rail and truck routes and facilities. Residential blocks are dispersed between various industries in this lower-income area of Oakland. Downtown consists of a mix of residential and commercial areas with mid to high-rise buildings. East Oakland includes both industrial and mixed-income residential areas with higher-income neighborhoods located to the north [8]. The sampling protocol involved installation of Aclima environmental intelligence fast-response pollution measurement and data integration platforms on two Google Street View mapping vehicles. These vehicles measured weekday daytime NO_2 concentrations on city streets. Measurements were collected on every road in a 30 km^2 domain, incorporating residential, commercial and industrial areas [24]. The data includes more than 2.7 million samples from two datasets with measurements from a total of 305 days from July 13, 2015 to August 31, 2017. Our data reduction ”snapping” scheme follows that of Messier et al. [25]. First, we divided a street centerline file (obtained from OpenStreetMaps.com) into more than 19,000 30-meter road segments. Next, we employed a nearest-neighbor algorithm (Python SciPy ”ckdnearest” algorithm) to ”snap” each 1-Hz measurement to its nearest road segment resulting in consistently defined locations [26].

We also gathered land use data for each 30 meter road segment in the form of 32 binary and continuous geographic covariates following the methods of Messier et al. [25]. The full list of the geographic covariates alongside collection details are presented in Table S1 of the supporting information document.

Surface meteorological observations, including hourly temperature, wind speed and direction, and precipitation, for Oakland International Airport were acquired from the National Oceanic and Atmospheric Administration (NOAA) Automated Surface Observing Systems (ASOS) through Iowa Environment Mesonet (IEM) portal maintained by Iowa State University (`https://mesonet.agron.iastate.edu/request/download.phtml`; accessed November 1, 2020). A major strength of the ASOS is the consistency of measurements in reporting wind data which is a crucial

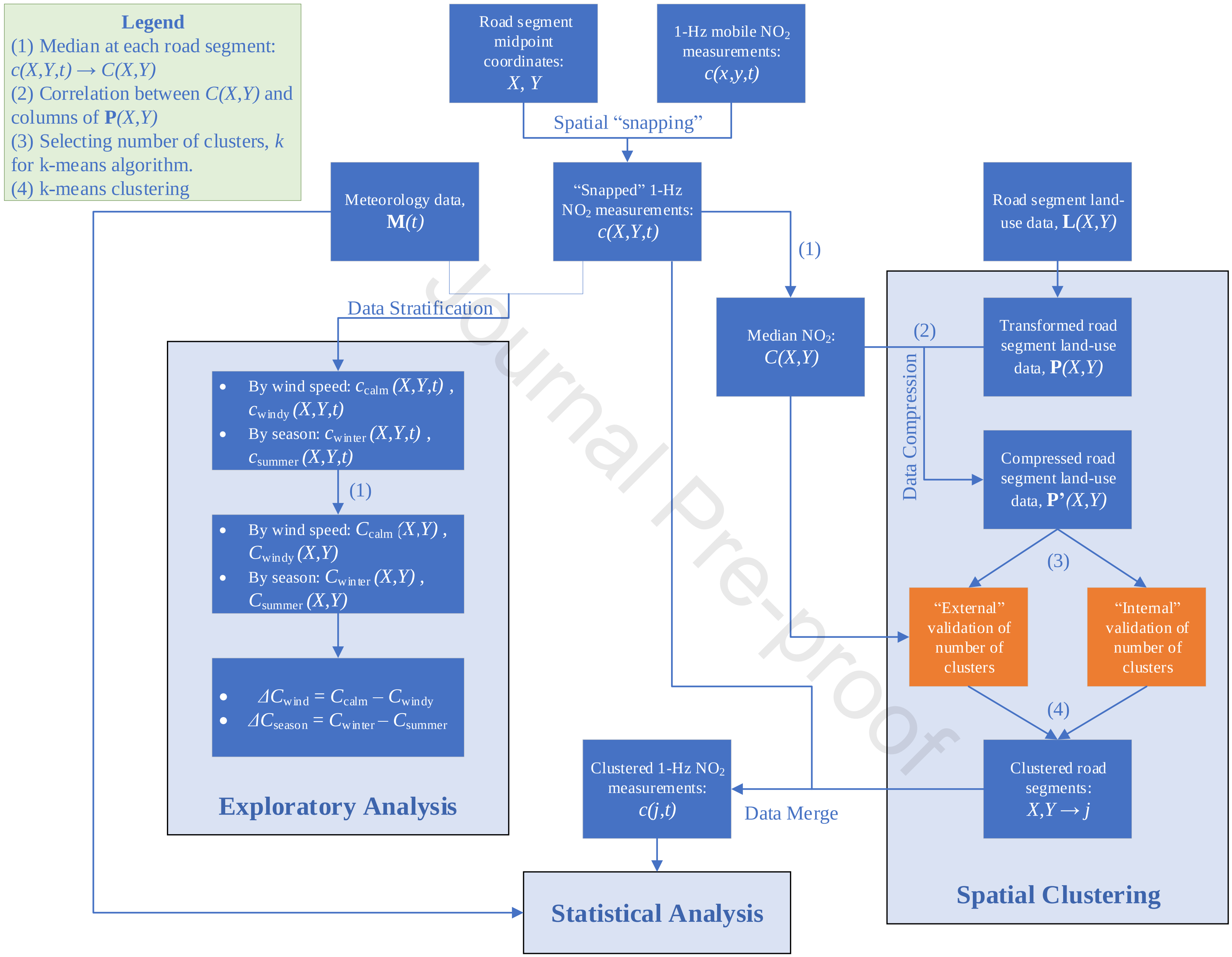


Figure 1: Flow diagram depicting the evolution of the data throughout the study. Dark blue rectangles correspond to data and orange rectangles correspond to data processing steps. Variables in bold refer to data matrices and non-bold variables indicate vectors. Light blue rectangles refer to the methodologies used in this study. (Color should be used for any figures in print)

variable in this study. Hourly solar radiation data in the form of Global Horizontal Irradiance (GHI) was obtained from Solcast (`https://solcast.com/`; accessed November 5, 2020) using the Solcast API, a source for satellite-derived solar irradiance data. The data was obtained for a location of $37°48'54''$N $122°16'57''$W and is within the core of the West Oakland/Downtown domain of NO_2 mobile measurements and coincides with the fixed-site regulatory monitor located at the Oakland West site managed by the Bay Area Air Quality Management District (BAAQMD). We then used linear interpolation to convert the observations to match the 1-Hz measurements of the mobile campaign, therefore augmenting the NO_2 observations with surface meteorological measurements and satellite-driven radiation measurements.

2.2 Selection of temporal variables

NO_2 concentrations in urban areas are affected by regional meteorological variables. Strong inter-dependencies between different meteorological variables, complicate the relationship between these variables and pollutant concentrations. Establishing links between regional meteorology and pollutant concentrations is further complicated by the role of local urban land form in mediating the effect of regional meteorology on the local mixing within the urban area. Therefore, prior to our statistical analysis, we apply a variable selection procedure driven by the regional meteorological conditions during the measurement period and unique to the study area of Oakland, CA. In particular, the temporal variables are selected based on two conditions. First, temporal controls with high variability are retained such that robust statistical inferences between NO_2 concentrations and the variables can be made. Second, correlation between temporal variables is used as a selection criterion such that variables with high correlations with each other are discarded. This allows us to isolate the effect of the remaining variables and safely assume a cause and effect relationship between the controls and the observed concentrations.

The climate in Oakland is characterized by dry, warm summers and mild, wet winters. However, during the measurement campaign precipitation data was reflective of prevailing drought conditions (zero precipitation for more than 99% of all study hours). In addition, the prevailing wind direction was found to be from the West for approximately 85% of all study hours. Due to the low variability observed in wind direction and precipitation during the study period, the effects of these parameters on intra-urban NO_2 pollution are not pursued here.

While high daily temperatures have been previously linked to higher concentrations of NO_2, increases in global radiation have been shown to correlate with reduced NO_2 concentrations [22]. The lack of nighttime measurements coupled with a moderate positive correlation (Spearman's correlation coefficient $= 0.57$) observed between temperature and radiation during the study period, leads to the conclusion that isolating the effect of each of these variables is not viable in our analysis. On the other hand, pollutant concentrations, including NO_2, are known to be seasonal [22, 27]. Henceforth, we assume that investigating the seasonality in the data indirectly accounts for the effects of emission seasonality, temperature and radiation. Therefore, temperature and radiation are excluded from the analysis and instead a seasonal stratification of concentration data as described in section 3 is adopted.

A secondary variable with known effects on atmospheric dispersion that can be calculated from the available data (radiation and wind speed) is atmospheric stability. In urban areas however, the increased drag force caused by roughness obstacles (e.g. buildings and other structures) leads to larger friction velocities than in open areas. Therefore, stability over urban areas is biased towards neutral (adiabatic) conditions [28]. As a result, the effects of atmospheric stability on intra-urban air pollution are not pursued in this study, due to low variability in stability conditions.

In this study, we primarily investigate the effects of wind speed on intra-urban NO_2 concentrations, as it has been established as an important meteorological parameter in affecting NO_2 pollution by previous studies [22, 29, 30]. In addition, seasonality of NO_2 concentrations in Oakland is studied. The exploratory analysis in section 3 further validates the choice of wind speed as an important meteorological parameter controlling NO_2 concentrations across the city of Oakland.

3 Exploratory Data Analysis

Prior to clustering, we conducted a preliminary analysis to examine the relationship between the selected variables in section 2.2 and NO_2 observations on 30-m road segments. The analysis relies on data stratification which refers to partitioning the concentration data into distinct and non overlapping groups of independent variable states. Two distinct stratifications are applied to the data separately to identify effects of wind speed and seasonal changes on NO_2 concentrations, respectively. Wind speed stratification is carried out by dividing all 1-Hz NO_2 measurements into two groups: wind speeds below 3.5 m/s (calm) and above 5.5 m/s (windy). The threshold values of 3.5 and 5.5 m/s are chosen for the following reasons: 1) similar sample sizes between the two groups, and 2) a wind speed buffer of 2 m/s prevents misclassification as the accuracy of the ASOS monitoring system is 1 m/s. After stratification, each group is analyzed separately to calculate the median of 1-Hz NO_2 measurements (C_{calm}, C_{windy}) for those 30-m road segments that have been visited on at least 10 distinct days, noting that 10 distinct measurement days ensure stable estimations of median concentrations [8]. Lastly, the local differences in median NO_2 concentrations (ΔC_{wind}) between calm and windy measurements are computed as $\Delta C_{wind} = C_{calm} - C_{windy}$ for each 30-m road segment (Figure 2a). The spatial distribution shows the contrast between the median concentrations, with the mean (median) $\pm$ standard deviation of $\Delta C_{wind} = 8.0\ (7.6) \pm 5.8$ ppb. It is worth noting that an increase of 5.3 ppb in long-term NO_2 concentrations (averaged over one year or more) has been associated with all-cause mortality with hazard ratios of $1.01 - 1.03$ (95% CI), highlighting the significance of the computed ΔC_{wind} [31].

Seasonal stratification is carried out by dividing the 1-Hz NO_2 measurements into two groups: November 1st until February 28th are labeled winter measurements and May 1st until August 31st are labeled summer. Following similar steps as the wind speed analysis, the local differences in median NO_2 concentrations (ΔC_{season}) between winter and summer are computed as $\Delta C_{season} = C_{winter} - C_{summer}$ (Figure 2b) for each 30-m road segment. The spatial distribution of ΔC_{season} indicates higher median concentrations during winter which is in agreement with our analysis of hourly NO_2 observations from the fixed site monitoring site in West Oakland (Figure S1 in Supporting information document). The mean (median) $\pm$ standard deviation of ΔC_{season} is 8.0 $(7.1) \pm 5.1$ ppb.

Our exploratory analysis reveals the effect of wind speed on NO_2 concentrations through a two-group stratification (windy and calm), because a multi-group stratification would not be appropriate as very few road segments would pass the 10 distinct day selection criterion. Furthermore, a mixed stratification based on wind speeds and seasons leading to 4 groups (e.g. winter and windy, summer and calm, etc.) would not be viable for the same reason. Therefore, we propose an approach that uses cluster analysis to group together road segments that are similar in terms of land use to investigate the effect of each temporal control separately and with finer granularity (i.e. more wind speed intervals). This clustering approach increases the statistical power of our temporal analysis, because of significantly larger sample sizes of each cluster compared to individual road segments.

4 Methodology

In this section, we introduce the methodology for using city-wide mobile measurements to examine spatially varying temporal patterns in air quality due to interaction between meteorology and urban form. A summary of the developed data-driven spatio-temporal framework is as follows. First, inspired by findings of Messier et al. (2018), we cluster the spatial locations in Oakland, CA based on land use covariates (as surrogates for emission sources and urban form) using the k-means clustering algorithm [25, 32]. This clustering effectively reduces the spatial fidelity of the data, but increases its statistical power by producing clusters with large sample sizes. The increase in statistical power is required for successful data stratification based on wind speed and season. Subsequently, we use conditional averaging to characterize the effect of wind speed on NO_2 concentrations in each cluster. We note that the focus on wind speed as an effective temporal variable in modulating NO_2 concentrations and the need for clusters with large sample sizes were discussed in detail in our exploratory analysis described in section 3. The analysis is concluded with the study of exceedance probabilities under varying seasons and wind speed conditions. Exceedance probabilities are an important measure of exposure to extreme pollutant concentrations, with clear ties to acute effects of air pollution on human health.

4.1 Spatial Clustering

A popular approach for quantifying intra-urban variation in air pollution is land use regression (LUR) [33, 34, 35]. LUR models are mainly used to depict spatial variation of air pollution and do not give any information on temporal variations of air quality. Furthermore, time series analysis of the mobile measurements is not feasible as the data are collected along spatio-temporal paths (cars traversing the city). In addition, the size of the dataset is inadequate to resolve the effects of all the factors influencing pollutant concentrations at every 30-m road segment.

Inspired by LUR models which suggest that locations with similar land use characteristics have similar pollutant concentrations, we aim to overcome the sample size issue by clustering the 30-m road segments based on their land use

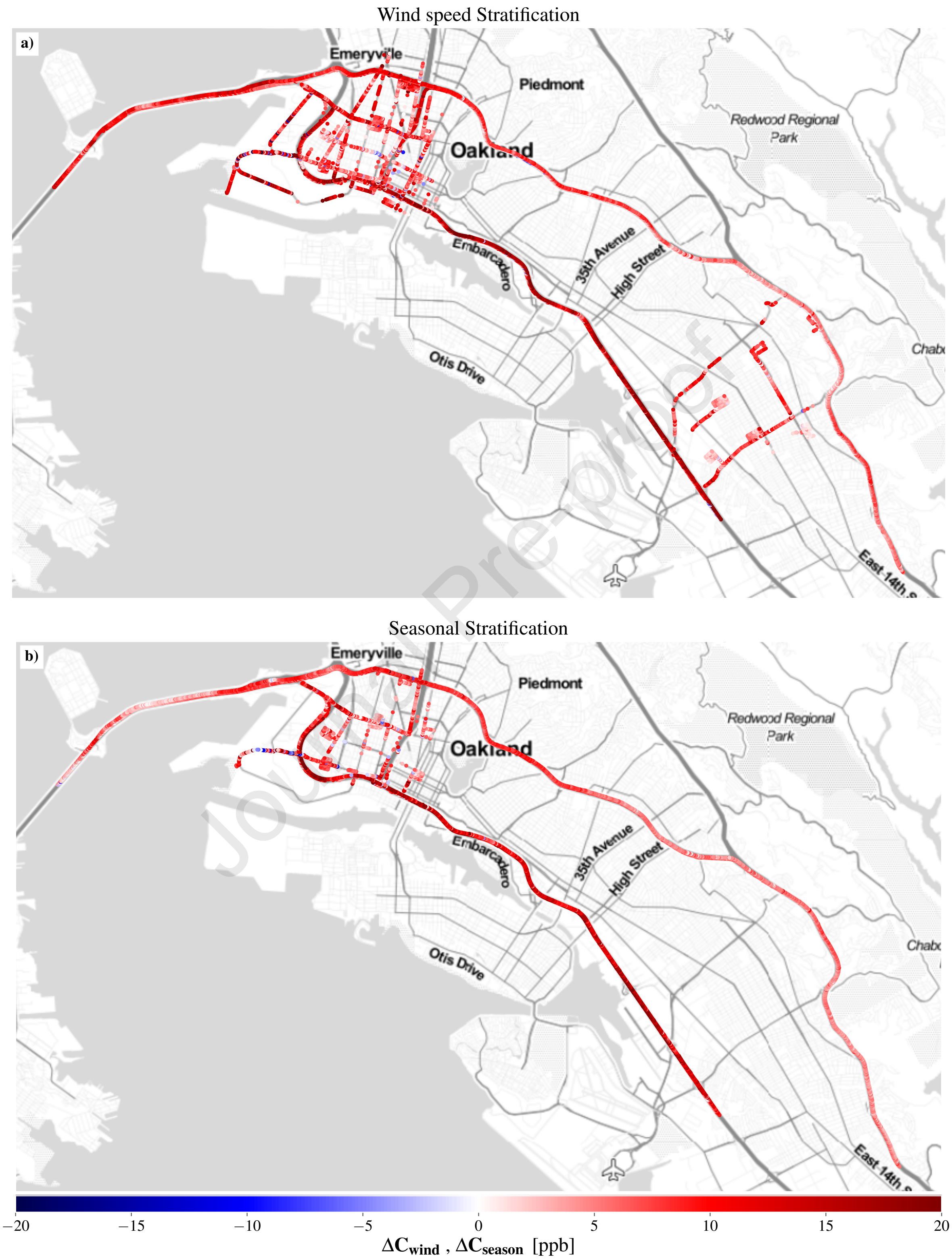


Figure 2: Difference in median NO_2 concentrations between **(a)** calm and windy and **(b)** winter and summer observations. Map tiles by Stamen Design. Map data by OpenStreetMap. (Color should be used for any figures in print)

covariates, and then study the temporal evolution of NO_2 concentrations within each cluster. This allows us to examine how land use modulates the effect of regional meteorology on local air quality dynamics.

Clustering is an unsupervised learning method for grouping a set of objects in a way that objects in the same group are more similar to each other than to those in other groups. The similarity of objects is assigned by the features that clustering is based on. In this study, we cluster 30-m road segments in the city of Oakland, CA, by using land use covariates of these road segments as features. As discussed in section 2.1 a total of 32 land use covariates are considered. Furthermore, it is desirable that road segments that are geographically close to each other fall in the same cluster, as we expect the effects of emission sources and local meteorology to be similar for adjacent road segments. Therefore, the latitude and longitude of the center point of individual road segments are also included as features in the clustering algorithm bringing the total feature count to 34.

4.1.1 Data Pre-processing

Performance of clustering algorithms are generally improved when the number of features are lowered [36]. First, we lower the number of features using a principal component analysis (PCA). Feature reduction using PCA is appropriate in the land-use context, because the land use variables considered are highly correlated with each other, containing redundant information that is detrimental to the performance of clustering algorithms. Prior to PCA, the features are standardized by subtracting the feature mean and rescaling the feature variance to unity. The standardized features are then stored in an $n \times 34$ matrix, with n being the number of unique road segments. Performing PCA on this preliminary matrix leads to a new $n \times 34$ matrix that we label matrix $\mathbf{P}$:

$$\mathbf{P} = (\mathbf{p_1}, \mathbf{p_2}, \ldots, \mathbf{p_{34}}) = \begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,34} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,34} \\ \vdots & \vdots & \vdots & \vdots \\ p_{n,1} & p_{n,2} & \cdots & p_{n,34} \end{bmatrix} \tag{1}$$

where each column vector $\mathbf{p_j}$ corresponds to the newly formed principal components (PCs) that are linearly uncorrelated with each other. The PCs are ordered based on amount of variance in the original variables accounted for by each component, with PC_1 accounting for the most variance and PC_{34} accounting for the least. The first 13 PCs account for approximately 80% of the variance in land-use variables. To further reduce the number of features, out of the first 13 PCs, we retain those PCs that are correlated with median NO_2 concentrations computed for each road segment. Therefore, we calculate the Pearson correlation coefficients of the columns of $\mathbf{P}$ and the column vector $\mathbf{C}$:

$$\mathbf{C} = (C_1, C_2, \ldots, C_n)^T \tag{2}$$

with C_i computed as median of NO_2 concentration at road segment i. Labeling the Pearson correlation coefficient between $\mathbf{p_i}$ and $\mathbf{C}$ as ρ_i, we only retain those PCs that satisfy $|\rho_i| > 0.1$. This analysis results in the retainment of 4

PCs that account for approximately 60% of the variance, therefore, greatly reducing the number of features prior to clustering.

4.1.2 Clustering Method

We apply the k-means algorithm developed by Hartigan and Wong (1979) to cluster the 30-m road segments [32]. This algorithm seeks to partition n points (30-m road segments) in D dimensions (4 PCs in this case) into k clusters. It iteratively searches for a local solution that minimizes Euclidean distance between the points and cluster centers. The initial cluster centers in the k-means algorithm can be chosen randomly, by the user or by randomized techniques. Here, we utilize the popular "k-means++" initializing algorithm as it seeks to spread out the cluster centers, a desirable property in this study [37]. The main advantages of k-means are its ease of implementation, computational efficiency, and reduced sensitivity to outliers compared to hierarchical clustering methods.

4.1.3 Selecting the Number of Clusters

In k-means clustering the main required hyper parameter is the number of clusters (k) which is often not known a priori. The number of clusters can be assigned by either pre-existing knowledge of the data that is not available from the dataset itself, or by providing a descriptive statistic for ascertaining the extent to which the observations comprising the dataset fall into natural distinct groupings [38]. In short, the number of clusters can either be assigned solely through the dataset (Data-based or internal methods) or by additional knowledge obtained externally (External methods). In this study, we apply both internal and external methods to select the optimal value of k and validate the clustering analysis. To select the number of clusters, clustering solutions are first found for a sequence of consecutive k values between 5 and 15. These solutions are then compared to each other using internal and external methods to find the optimal number of clusters.

Internal method The gap statistic approach originally introduced by Tibshirani et al. is among the standard data-based methods for choosing the number of clusters in a dataset [39]. This method utilizes the total "within-cluster dispersion", which is defined as the sum of the distance between each data point (road segment features) in the cluster and the cluster center. For each value of k, the k-means algorithm is applied to the observed data and a randomly generated data set that uniformly spans the feature space and has the same size as the observed data. The gap function, Gap(k), is then computed as the difference between the sum of the total within-cluster dispersion for the observed and random data (generated 100 times in this analysis). The optimal number of clusters for the given data set is the smallest k such that

$$\text{Gap}(k) \geq \text{Gap}(k+1) - s_{k+1} \tag{3}$$

where s_{k+1} is the standard deviation of the total within-cluster sum of squares of the randomly generated data.

External method In regards to applying additional knowledge to assign the number of clusters, we consider the cluster average of variability of median NO_2 concentration for all 30-m road segments within each cluster. Variability labeled V, is calculated as the standard deviation from the mean of median daytime concentrations for 30-m road segments within each cluster:

$$V(j) = \left(\frac{1}{n_j} \sum_{i=1}^{n_j} \left(C_i^{(j)} - \bar{C}^{(j)} \right)^2 \right)^{1/2} \tag{4}$$

where n_j is the number of road segments in cluster j, $C_i^{(j)}$ is the median NO_2 concentration observed at the i'th road segment belonging to cluster j and $\bar{C}^{(j)}$ is the mean of median NO_2 concentrations observed at all road segments belonging to cluster j. Average cluster variability, labeled S, is then calculated as follows:

$$S(k) = \frac{1}{k} \sum_{j=1}^{k} V(j) \tag{5}$$

At first glance, solutions with lower average variability may be judged to be superior to those with higher average variability. However, average variability within clusters generally tends to decrease with increasing number of clusters. Therefore, we create a "benchmark" for every value of k, and judge the superiority of solutions based on their distance from this benchmark. For each k, the benchmark is created by first sorting 30-m road segments by their corresponding value of median NO_2 concentrations and then grouping the road segments into k equally-sized clusters. We then find the number of clusters that minimizes the difference between average variability of the median NO_2 concentrations of the original clustering using k-means algorithm, $S(k)$ from Eq. 5, and the average variability of median concentrations of the benchmark, $S^*(k)$, for k values between 5 and 15:

$$\underset{k \in [5,15]}{\arg\min} \left[S(k) - S^*(k) \right] \tag{6}$$

4.2 Statistical Analysis

Once the road segments are clustered, the effects of wind speed and seasonality on 1-Hz NO_2 concentrations corresponding to road segments in each cluster are investigated. Similar to section 3, NO_2 concentrations in each cluster are stratified into two groups based on the measurement season. Following this division, conditional averaging based on wind speed is employed to quantify the effect of wind speed on NO_2 concentrations for each cluster/season combination. Further, probabilities of NO_2 exceeding pre-determined thresholds are calculated through a two-step sampling process for every cluster, season and wind speed condition.

4.2.1 Conditionally Averaged Concentration

Every NO_2 concentration measurement coincides with a wind speed measurement as described in section 2.1. The concentration values are organized based on the wind speed such that multiple concentration values are grouped together within a given wind speed interval, U. The conditionally averaged NO_2 concentration value, denoted $\langle c|u \rangle$, is calculated within designated wind speed intervals as shown:

$$\langle c|u \rangle = \frac{1}{N_U} \sum_{u_i \in U(u)} c\left(u_i\right) \tag{7}$$

where c represents 1-Hz NO_2 measurements, $U(u) = \{u_i : -\Delta u/2 \leq u - u_i < \Delta u/2, \forall i = 1, 2, \ldots, N_U\}$ and N_U is the total number of data points within the given wind speed interval U. In this analysis, Δu is set at $1m/s$. This choice of the wind speed intervals is driven by the accuracy of $1m/s$ of the ASOS monitoring system and the available sample size of NO_2 measurements coinciding with each given interval. In addition, conditional probability distribution functions (PDFs) of concentration are also constructed to calculate the conditional interquartile range in a similar manner to the conditional averages.

4.2.2 Exceedance Probabilities

Exceedance probabilities are calculated by computing empirical cumulative distribution functions (ECDFs) of NO_2 concentrations for every cluster, season and wind speed condition. Due to the streaming nature of mobile measurements, observations recorded on any given day are correlated, particularly if the observations were recorded over a short period of time (e.g. one hour). Furthermore, the number of measurements on each day varies widely across different days, especially after cluster, season and wind speed stratifications. Therefore, direct calculation of the ECDFs using raw 1-Hz measurements gives extra weight to days with high number of measurements and biases calculated exceedance probabilities. To overcome this issue, we utilize the following two-step sampling strategy to compute ECDFs and exceedance probabilities. For each cluster, season and wind speed condition, the steps are as follows:

1. Randomly select a day with replacement from the days with at least 100 mobile measurements for the given cluster, season and wind condition.
2. Randomly sample $N = 100$ NO_2 measurements with replacement from the selected day.
3. Repeat the first two steps $N_D = 10$ times to create an ECDF with $N_D \times N = 1000$ samples.
4. Calculate exceedance probability as: $\mathbb{P}_E(T) =$ (**Number of samples with concentrations** $> T$)/$(N_D \times N)$

with T corresponding to the concentrations threshold chosen for NO_2. A robust estimate of the exceedance probability is then computed by repeating the steps above 1000 times to account for variability introduced through the random selection. We note that the data corresponding to days with less than 100 measurements account for less than 5% of all the data for a given cluster, season and wind condition, and therefore unlikely to have a significant effect on the calculated probabilities. In addition, $N_D = 10$ is chosen since there are at least 10 unique measurement days with at least 100 measurements for each cluster, wind and season condition.

5 Results and Discussion

5.1 Spatial Clustering

After pre-processing the land use data corresponding to individual 30-m road segments described in section 4.1.1, we select the number of clusters k, using both data-based and external methods.

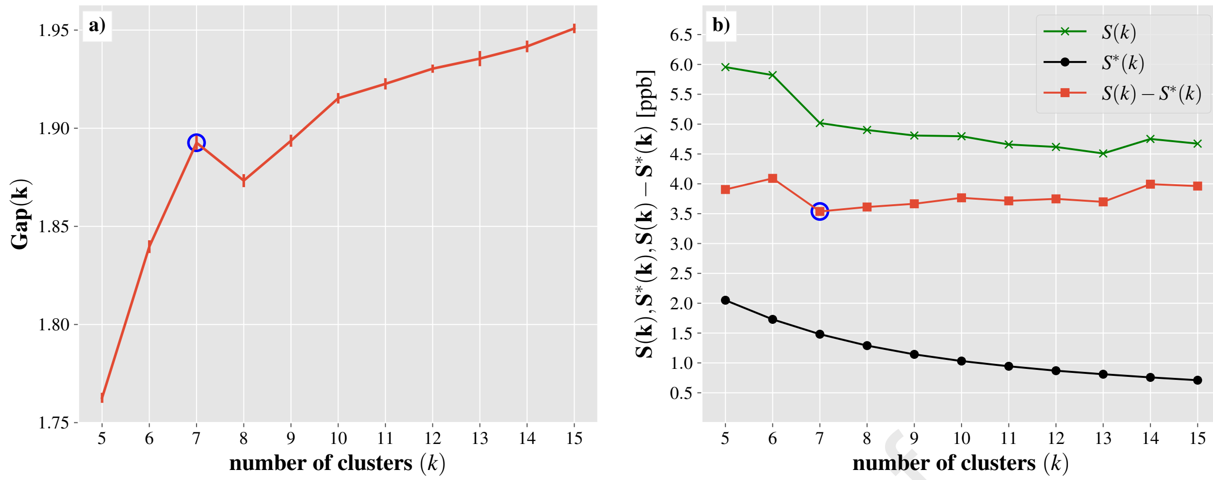


Figure 3: Selecting optimal number of clusters through **(a)** gap-statistic as an internal method suggesting 7 clusters as the optimal choice for k, with the vertical lines corresponding to s_k and **(b)** comparison of average within-cluster variability of daytime median NO_2 concentrations between the clustering solution and clustering benchmark as an external method, suggesting 7 clusters.

Internal method We computed the gap statistic for clustering solutions between 5 and 15 clusters to find the optimal number of clusters suggested by this method. The gap statistic for these solutions are shown in Figure 3a with the vertical error bars corresponding to the standard error, s_k. Based on equation 3, this method assigns 7 clusters as the optimal value for k.

External method As discussed in section 4.1.3, we computed the statistics required to select the optimal number of clusters k using information external to land-use and location data. The results are shown in figure 3b where $S(k)$, $S^*(k)$ and their differences are plotted for clustering solutions between 5 and 15 clusters. Since the goal is to minimize $S(k) - S^*(k)$, this methodology indicates that the optimal choice for k is 7 clusters.

Since both validation methods yield the same result regarding the optimal number of clusters, 7 was chosen as the number of clusters. Figure 4a shows the clustering solution utilizing the k-means algorithm with $k = 7$ as a spatial map of Oakland, CA. Meanwhile, Figure 4b presents the histograms of median NO_2 concentrations at each road segment belonging to each of the 7 clusters. This clustering solution shows that cluster 1 is a mixture of highways and major roads in industrial areas closer to East Oakland, cluster 2 covers residential areas in East Oakland that are located at higher elevations (¿100m higher than sea level), cluster 3 mostly includes both major and narrow roads in industrial zones of West Oakland and Downtown, cluster 4 covers highways that are truck prohibited, cluster 5 mostly covers residential zones and roads located in East Oakland, cluster 6 mostly consists of interstate highways that allow truck passage and cluster 7 covers residential areas in West Oakland and Downtown. Based on these findings, the clusters will be referred to using the following labels:

- Cluster 1 - Industrial East Oakland
- Cluster 2 - Elevated residential East Oakland

- Cluster 3 - Industrial West Oakland
- Cluster 4 - Truck prohibited highways
- Cluster 5 - Residential East Oakland
- Cluster 6 - Truck-route highways
- Cluster 7 - Residential West Oakland

With geographically similar road segments grouped together in clusters with a significant number of mobile NO_2 measurements available within each cluster, mobile measurements within each cluster can be investigated with regards to wind speed and seasonal changes.

5.2 Effects of Wind Speed on Concentrations

For each cluster, effects of wind speed on NO_2 concentrations during each season are examined through conditionally averaged concentrations and are shown in Figures 5 and 6 for winter and summer, respectively. The results are shown for 4 of the 7 clusters including Industrial and residential West Oakland and inter-state highways (i.e. clusters 3, 4, 6 and 7) for the following reasons: 1) These regions cover highways, industrial and residential zones where the population lives, works and commutes, 2) the results allow for comparisons between residential/industrial zones, truck-route/truck-prohibited highways, and highway/non-highway roads, and 3) the majority of mobile measurements were made in these regions and therefore sample sizes are large enough for statistically significant analyses.

During winter, the West Oakland clusters follow a similar downward trend as measured by a linear fit to the conditionally averaged concentrations, even though concentrations are generally higher in the industrial cluster. While the concentrations on truck-route highways also drop with increasing wind speed, the drop is smaller than West Oakland. A plausible explanation for this behaviour is the additional turbulence on the highways caused by moving traffic which increases vertical mixing of the pollutants with the clean air above even in the absence of wind. This additional turbulence in turn leads to a smaller marginal effect of wind speed on NO_2 concentrations. Concentrations on truck prohibited highways do not follow a significant downward trend which is likely due to traffic turbulence and the topography of this cluster, located at higher elevations compared to other investigated clusters.

In the summer, the conditionally averaged concentrations do not follow a significant trend in any of the clusters, suggesting that wind speed is a less important predictor of NO_2 concentrations in the summer compared to winter. One possible explanation for this behaviour is increased vertical mixing in the summer caused by increased radiation and surface heat fluxes that leads to overall lower concentrations in the summer as investigated in section 3. We note that the concentrations observed for each cluster during summer is consistently lower than those observed in the winter, as evident through a comparison between figures 5 and 6 which is in agreement with Figure 2b. These results also explain the minor differences observed between concentrations corresponding to calm and windy conditions in Figure 2a, since summer and winter measurements were not separated in the analysis of wind speed in section 3.

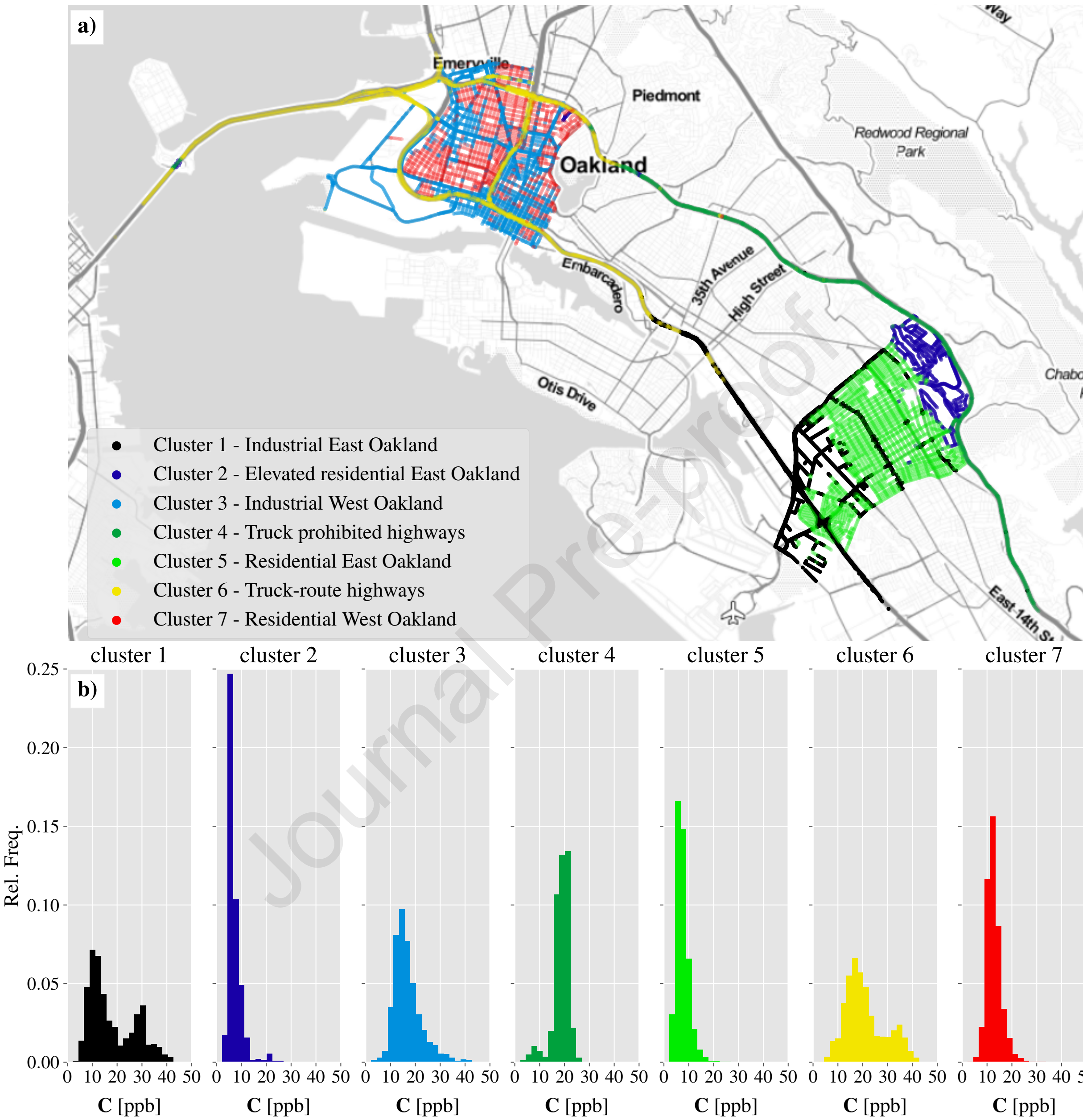


Figure 4: Clustering 30-m road segments into $k = 7$ clusters. **(a)** Spatial map of 30-m road segments, color coded based on cluster numbers, and **(b)** histograms of daytime median NO_2 concentrations for each cluster. Cluster 1 is a mixture of highways and major roads in industrial areas closer to East Oakland, cluster 2 covers residential areas in East Oakland that are located at higher elevations (¿100m higher than sea level), cluster 3 mostly includes both major and narrow roads in industrial zones of West Oakland and Downtown, cluster 4 covers highways that are truck prohibited, cluster 5 mostly covers residential zones and roads located in East Oakland, cluster 6 mostly consists of interstate highways that allow truck passage and cluster 7 covers residential areas in West Oakland and Downtown. Map tiles by Stamen Design. Map data by OpenStreetMap. (Color should be used for any figures in print)

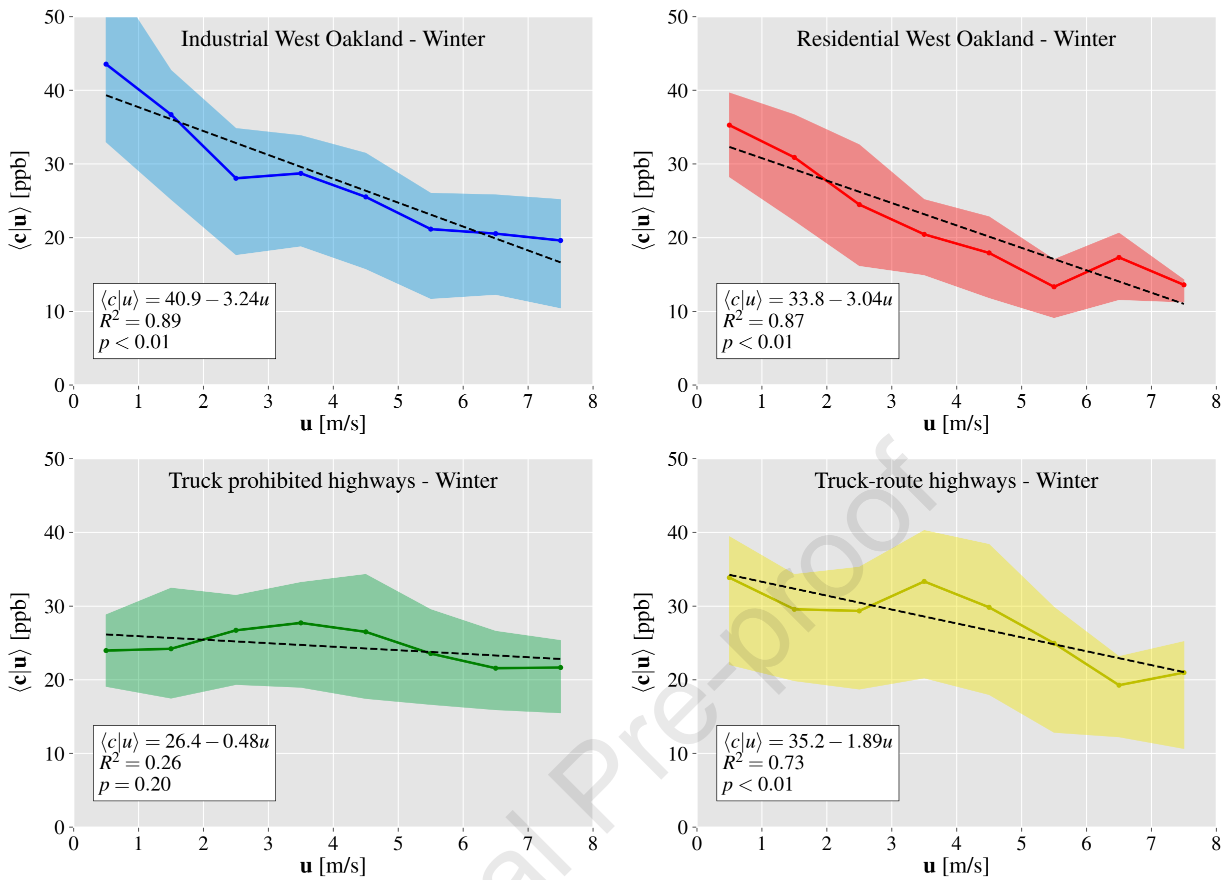


Figure 5: Effect of wind speed on NO_2 concentrations for each cluster during Winter, with statistically significant decay of concentrations observed in three clusters consisting of Industrial West Oakland, Residential West Oakland and Truck-route highways. Statistically significant trends were not found between the concentrations and wind speeds for the Truck prohibited highways. The colored solid lines correspond to conditionally averaged concentrations found through Eq. 7. Shaded regions correspond to the interquartile range of conditional concentration distributions. The black dashed lines correspond to a linear fit to the curve with details of the fit described in the text boxes, where coefficient of determination is represented by R^2 and the significance of the slope of the linear fit is quantified through t-tests with the p-values shown.

As mentioned in section 2.2, we found that for more than 85% of the study period (more than 90% during winter) the prevailing wind direction was from the West. Hence, there are few measurements in each cluster during winter where the wind is from other directions, leading to high uncertainties when making inferences. Additionally, with mobile concentration measurements, the alignment between polluting sources and the sensor is constantly changing within each cluster. Therefore, at the spatial resolution of our analysis, wind direction does not provide us with additional information regarding the NO_2 concentration patterns. These reasons have led us to refrain from providing a wind rose alongside Figure 5.

5.3 Exceedance Probabilities

For each cluster, the probability of observing NO_2 concentrations above the threshold of 40 ppb (95th percentile of concentrations observed for the investigated clusters) are calculated under four conditions based on wind speed

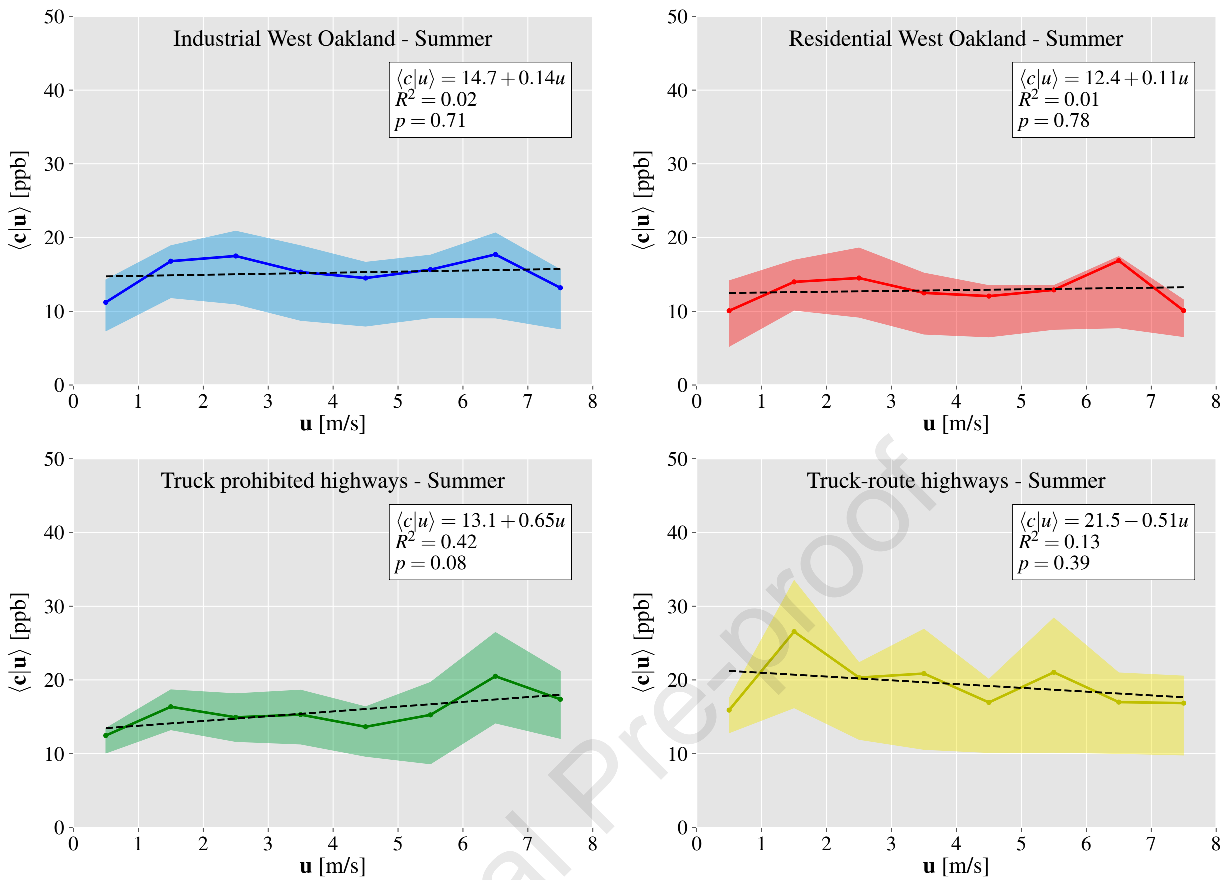


Figure 6: Effect of wind speed on NO_2 concentrations for each cluster during Summer. Statistically significant trends were not found between the concentrations and wind speeds for any of the clusters. As in Figure 5, the colored solid lines correspond to conditionally averaged concentrations found through Eq. 7. Shaded regions correspond to the interquartile range of conditional concentration distributions. The black dashed lines correspond to a linear fit to the curve with details of the fit described in the text boxes, where coefficient of determination is represented by R^2 and the significance of the slope of the linear fit is quantified through t-tests with the p-values shown.

and seasonality as depicted in Figure 7. The four conditions are obtained through a mixed data stratification process following the steps described in section 3. The truck-route highways cluster shows a sharp drop in exceedance probabilities during windy conditions compared to calm conditions with a 53% drop during winter and a 84% drop in the summer. One possible explanation for this sharp drop is tied to traffic density and speed of cars on the highway. Considering that high NO_2 are often due to high traffic during which cars are moving slowly, therefore not contributing to turbulence and mixing of the pollutants. In these conditions wind can be an effective tool for creating additional turbulence that leads to the mixing of the pollutants and lowers pollutant concentrations. The significant difference between the probabilities of the two highway clusters highlights the effect of trucks and high emitting vehicles on high NO_2 concentrations. In addition, almost all of the measurements on truck prohibited highways during summer fall below the 40 ppb threshold, leading to very small exceedance probabilities. The trend observed for the industrial West Oakland cluster is similar to that found in section 5.2, with exceedance probability dropping under windy conditions

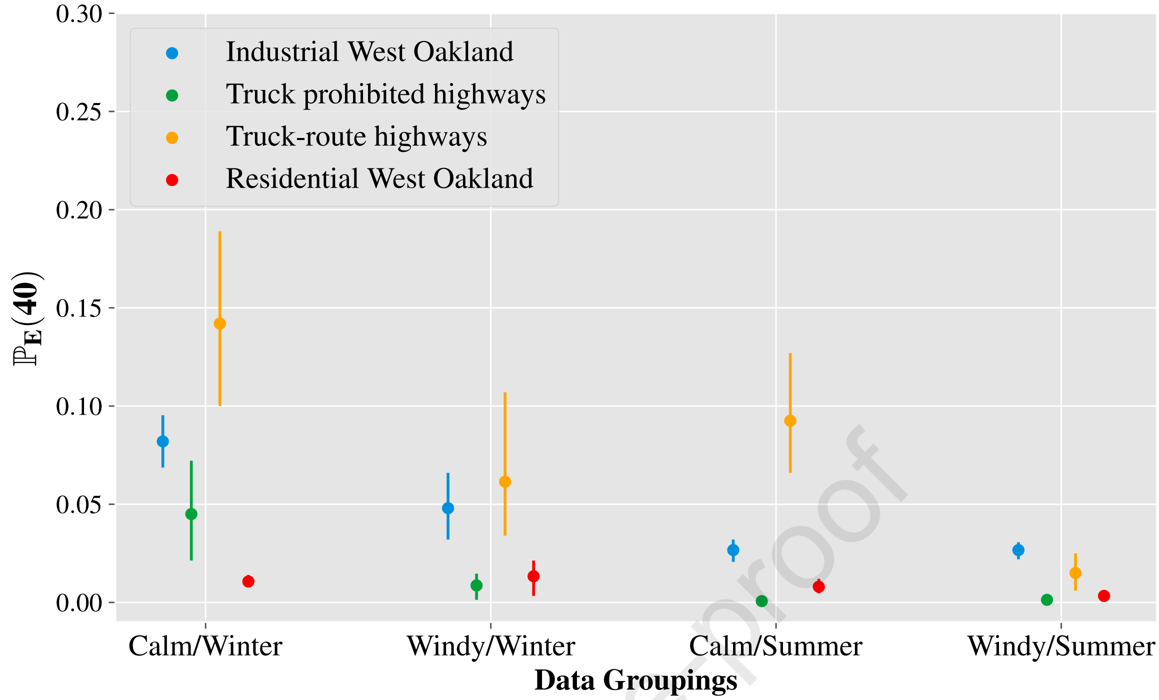


Figure 7: Probability of observing NO_2 concentrations above 40 ppb for groupings based on cluster, season and wind speed. Exceedance probabilities are calculated as the average of 1000 sampling simulations shown as filled circles, with vertical lines corresponding to the 25th-75th percentile ranges. (Color should be used for this figure in print)

and lower values observed during summer. Moreover, there is a perceptible difference between the two West Oakland clusters, highlighting the correlation between land use and pollutant concentrations.

It is worth noting that the 40 ppb threshold is smaller than regulatory limits for short term exposure. Nevertheless, the exceedance probability analysis was worthwhile as it showed that the response of the tails of the concentration distribution to wind speed differed from the response of the mean concentrations. Furthermore, NO_2 levels are correlated with other pollutant concentrations highlighting the importance of an exceedance probability analysis in the context of exposure to other air pollutants in addition to NO_2 [40].

6 Sensitivity Analysis

6.1 Sensitivity of Wind Speed Effects to Wind Speed Intervals

The linear fits to the conditionally averaged concentrations found in Section 5.2 are subject to the chosen wind speed intervals. As such we repeated the analysis to compute the slope of the linear fit to the conditionally averaged concentrations for different lengths of the wind speed intervals, Δu, varying between $0.5m/s$ and $1.5m/s$. The calculated slopes for different wind speed intervals for each cluster during winter are provided in Table 1, indicating that the magnitude of the calculated slopes depend on the wind speed intervals. Nevertheless, these results confirm

Table 1: Slope of linear fit to conditionally averaged NO_2 concentrations for 4 clusters during winter. Numbers in brackets refer to the p-values of the slope significance t-tests and are shown for p-values above 0.05. The boldface row corresponds to the analysis of section 5.2.

Δu (m/s)	**Industrial West Oakland**	**Residential West Oakland**	**Truck-Prohibited Highways**	**Truck Route Highways**
0.5	-3.16	-3.04	-0.61 (0.07)	-2.12
0.6	-3.07	-2.96	-0.51 (0.18)	-2.02
0.7	-3.00	-2.98	-0.49 (0.22)	-1.84
0.8	-3.01	-2.85	-0.49 (0.22)	-1.81
0.9	-3.42	-3.17	-0.63 (0.23)	-1.95
1.0	**-3.24**	**-3.04**	**-0.48 (0.20)**	**-1.88**
1.1	-3.17	-2.69	-0.53 (0.34)	-1.71
1.2	-3.37	-2.97	-0.38 (0.50)	-2.09
1.3	-3.03	-2.83	-0.73 (0.22)	-1.87
1.4	-3.11	-3.16	-0.54 (0.46)	-1.82 (0.11)
1.5	-2.92	-2.94	-0.44 (0.43)	-1.79 (0.07)

that the effects of wind speed are less pronounced on NO_2 concentrations on highways compared to residential and industrial regions in West Oakland.

6.2 Exceedance Probabilities

The two-step sampling process used to compute the exceedance probabilities, requires two parameters: Number of randomly selected days, N_D, and the number of samples per day, N. Here, we investigate the dependence of the calculated exceedance probabilities on these two parameters, N_D and N, respectively.

Sensitivity to number of randomly selected days, N_D The exceedance probabilities were calculated as described in section 4.2.2 for number of randomly selected days between 10 and 20 days. For each N_D, the average exceedance probabilities for 1000 simulations were computed for each cluster under each wind/season conditions. The resulting average exceedance probabilities showed very little dependence on N_D with all values staying within 10% of the original average exceedance probabilities plotted in Figure 7.

Sensitivity to number of samples per day, N Similarly exceedance probabilities were calculated with varying number of samples per day between 100 and 500 with increments of 50. There was no observable change in exceedance probabilities when number of samples per day was increased, suggesting that the original sampling of 100 samples per day was sufficiently large and therefore did not influence the exceedance probabilities.

7 Conclusions

An understanding of the interaction between urban form and the temporal dynamics of air pollutants is crucial for characterizing the effects of urban development and climate change on urban air quality, and especially for understanding how different settings in a given city can be subject to different health risks. In this study, a spatio-temporal framework consisting of a spatial clustering analysis and a robust statistical analysis of wind speed effects on pollutant concentrations was presented. The framework was used to study the influence of wind speed in the reduction of NO_2 concentrations in different regions of Oakland, California during different seasons. The analysis showed a negative correlation between wind speed and NO_2 concentrations in industrial and residential regions bounded by highways during winter, with increasing wind speeds leading to lower concentrations. However, it was found that increased vertical mixing of pollutants caused by sources other than wind speed (e.g. moving traffic and increased surface heat fluxes during summer) can lower the effectiveness of wind speed in lowering NO_2 concentrations. Furthermore, an analysis of exceedance probabilities showed that the response of the tails of the concentration distribution differs from that of the mean concentrations. These findings coupled with projections of climate and urban development can be used as predictive tools for future air quality in urban areas. For example, if reductions in wind speeds and increases in periods of stability as observed over the past few decades continue (through either climate or urban density changes) [41], on the basis of the current level of emissions poorer air quality is expected in residential and industrial areas of Oakland during winter. The large discrepancies between the exceedance probabilities observed on truck-route and truck prohibited highways suggest that stricter truck emission standards can potentially lead to substantial decreases in exposure to traffic related pollutants. It is worth noting that the truck-route highways surround the lower-income residential neighborhoods of West Oakland, while the truck-prohibited highways are bounded by higher-income regions to the north. Hence, truck-route designations can be considered by policymakers to address disparities in exposure to air pollution. Moreover, a study of the health of the commuters using truck-route highways versus truck-prohibited highways can be informative regarding these acute effects on the health of the Oakland population.

The application of the proposed framework to mobile measurements in Oakland has been insightful in comparing the effects of wind speed on NO_2 concentrations across different clusters. However, the findings presented here are particular to the measurement domain of Oakland, and generalizing the findings to other urban areas should be done with care. An additional consideration for interpreting our results is the choice of the pollutant: NO_2 is a secondary pollutant forming through photochemical conversion from Nitrogen Oxide and is dominated by local traffic. Moreover, for epidemiological analyses, it is necessary to relate the on-road concentrations investigated here to true exposures at residential and work addresses. On the other hand, the proposed framework can be applied to other urban areas with less consistent meteorology than Oakland, to study the effects of other prominent meteorological parameters on air quality as mediated by local land use. The framework could be applied to study the response of other major air pollutants such as ozone (O_3) and $PM_{2.5}$ to meteorological conditions as influenced by varying urban land form. Other well-known clustering algorithms such as DBSCAN, HDBSCAN, and hierarchical clustering could also lead to potential improvements in the presented framework. It is worth noting that while the developed framework has not been used as a tool to predict NO_2 concentrations at locations not measured by the mobile monitors, prediction is possible if certain conditions are met. In particular, if road segments without NO_2 measurements are incorporated into the spatial clustering scheme and clustered into one of the existing clusters with sufficient measurements, predictions regarding NO_2 concentrations can be made based on the prevailing wind conditions. Although further investigation is required to quantify the prediction performance of this methodology, we believe that predictions will be highly uncertain with respect to instantaneous measurements but will likely be more accurate in the mean.

By utilizing the meteorological data from one station, we captured the effect of urban form in mediating the effect of regional meteorology on intra-urban air quality. We note that an improved measurement campaign could deploy meteorological stations in the measurement area (e.g. in each cluster) or integrate anemometers onto the measurement vehicle for real-time wind speed measurements [42]. In that case, an even more robust spatio-temporal analysis can be designed to study the relationship between air quality and meteorological conditions at the neighborhood scale. Furthermore, coupled meteorological and air quality measurements can also be utilized in emission source characterization, similar to efforts in characterizing methane emission sources using mobile sensors in the oil and gas industry [43].

Declaration of competing interest

The authors declare they have no actual or potential competing financial interests.

Acknowledgments

This project was supported by David R. Atkinson Center for a Sustainable Future (ACSF) at Cornell University, and DOE ARPA-E's Methane Observation Networks with Innovative Technology to Obtain Reductions (MONITOR) program under grant DE-AR0000749.

References

[1] P. Das and R. Horton, "Pollution, health, and the planet: time for decisive action," *The Lancet*, vol. 391, pp. 407–408, Feb. 2018.

[2] F. Dominici, R. D. Peng, M. L. Bell, L. Pham, A. McDermott, S. L. Zeger, and J. M. Samet, "Fine Particulate Air Pollution and Hospital Admission for Cardiovascular and Respiratory Diseases," *JAMA*, vol. 295, pp. 1127–1134, Mar. 2006.

[3] X. Wu, R. C. Nethery, B. M. Sabath, D. Braun, and F. Dominici, "Exposure to air pollution and COVID-19 mortality in the United States: A nationwide cross-sectional study," *medRxiv*, p. 2020.04.05.20054502, Apr. 2020.

[4] W. Q. Gan, H. W. Davies, M. Koehoorn, and M. Brauer, "Association of Long-term Exposure to Community Noise and Traffic-related Air Pollution With Coronary Heart Disease Mortality," *American Journal of Epidemiology*, vol. 175, pp. 898–906, May 2012.

[5] S. Shin, L. Bai, T. H. Oiamo, R. T. Burnett, S. Weichenthal, M. Jerrett, J. C. Kwong, M. S. Goldberg, R. Copes, A. Kopp, and H. Chen, "Association Between Road Traffic Noise and Incidence of Diabetes Mellitus and

Hypertension in Toronto, Canada: A Population-Based Cohort Study," *Journal of the American Heart Association*, vol. 9, p. e013021, Mar. 2020.

[6] Y. Ogen, "Assessing nitrogen dioxide (NO2) levels as a contributing factor to coronavirus (COVID-19) fatality," *Science of The Total Environment*, vol. 726, p. 138605, July 2020.

[7] G. S. W. Hagler, M.-Y. Lin, A. Khlystov, R. W. Baldauf, V. Isakov, J. Faircloth, and L. E. Jackson, "Field investigation of roadside vegetative and structural barrier impact on near-road ultrafine particle concentrations under a variety of wind conditions," *Science of The Total Environment*, vol. 419, pp. 7–15, Mar. 2012.

[8] J. S. Apte, K. P. Messier, S. Gani, M. Brauer, T. W. Kirchstetter, M. M. Lunden, J. D. Marshall, C. J. Portier, R. C. Vermeulen, and S. P. Hamburg, "High-Resolution Air Pollution Mapping with Google Street View Cars: Exploiting Big Data," *Environmental Science & Technology*, vol. 51, pp. 6999–7008, June 2017.

[9] P. Deshmukh, S. Kimbrough, S. Krabbe, R. Logan, V. Isakov, and R. Baldauf, "Identifying air pollution source impacts in urban communities using mobile monitoring," *Science of The Total Environment*, vol. 715, p. 136979, May 2020.

[10] C. A. Pope, "Epidemiology of fine particulate air pollution and human health: biologic mechanisms and who's at risk?," *Environmental Health Perspectives*, vol. 108, pp. 713–723, Aug. 2000.

[11] R. Morello-Frosch, M. Pastor, and J. Sadd, "Environmental Justice and Southern California's "Riskscape": The Distribution of Air Toxics Exposures and Health Risks among Diverse Communities," *Urban Affairs Review*, vol. 36, pp. 551–578, Mar. 2001.

[12] G. A. Millett, A. T. Jones, D. Benkeser, S. Baral, L. Mercer, C. Beyrer, B. Honermann, E. Lankiewicz, L. Mena, J. S. Crowley, J. Sherwood, and P. Sullivan, "Assessing Differential Impacts of COVID-19 on Black Communities," *Annals of Epidemiology*, pp. 37–44, May 2020.

[13] G. S. W. Hagler, E. D. Thoma, and R. W. Baldauf, "High-Resolution Mobile Monitoring of Carbon Monoxide and Ultrafine Particle Concentrations in a Near-Road Environment," *Journal of the Air & Waste Management Association*, vol. 60, pp. 328–336, Mar. 2010.

[14] D. Hasenfratz, O. Saukh, C. Walser, C. Hueglin, M. Fierz, T. Arn, J. Beutel, and L. Thiele, "Deriving high-resolution urban air pollution maps using mobile sensor nodes," *Pervasive and Mobile Computing*, vol. 16, pp. 268–285, Jan. 2015.

[15] J. Wallace, D. Corr, P. Deluca, P. Kanaroglou, and B. McCarry, "Mobile monitoring of air pollution in cities: the case of Hamilton, Ontario, Canada," *Journal of Environmental Monitoring*, vol. 11, pp. 998–1003, May 2009.

[16] C. E. Kolb, S. C. Herndon, J. B. McManus, J. H. Shorter, M. S. Zahniser, D. D. Nelson, J. T. Jayne, M. R. Canagaratna, and D. R. Worsnop, "Mobile Laboratory with Rapid Response Instruments for Real-Time Measurements of Urban and Regional Trace Gas and Particulate Distributions and Emission Source Characteristics," *Environmental Science & Technology*, vol. 38, pp. 5694–5703, Nov. 2004.

[17] H. L. Brantley, G. S. W. Hagler, E. S. Kimbrough, R. W. Williams, S. Mukerjee, and L. M. Neas, "Mobile air monitoring data-processing strategies and effects on spatial air pollution trends," *Atmospheric Measurement Techniques*, vol. 7, pp. 2169–2183, July 2014.

[18] M. Van Poppel, J. Peters, and N. Bleux, "Methodology for setup and data processing of mobile air quality measurements to assess the spatial variability of concentrations in urban environments," *Environmental Pollution*, vol. 183, pp. 224–233, Dec. 2013.

[19] L. M. Zwack, C. J. Paciorek, J. D. Spengler, and J. I. Levy, "Modeling Spatial Patterns of Traffic-Related Air Pollutants in Complex Urban Terrain," *Environmental Health Perspectives*, vol. 119, pp. 852–859, June 2011.

[20] R. Hagemann, U. Corsmeier, C. Kottmeier, R. Rinke, A. Wieser, and B. Vogel, "Spatial variability of particle number concentrations and NOx in the Karlsruhe (Germany) area obtained with the mobile laboratory 'AERO-TRAM'," *Atmospheric Environment*, vol. 94, pp. 341–352, Sept. 2014.

[21] Environmental Defense Fund, "Why new technology is critical for tackling air pollution around the globe." `https://www.edf.org/airqualitymaps`. Accessed February 20, 2021.

[22] J. L. Pearce, J. Beringer, N. Nicholls, R. J. Hyndman, and N. J. Tapper, "Quantifying the influence of local meteorology on air quality using generalized additive models," *Atmospheric Environment*, vol. 45, pp. 1328–1336, Feb. 2011.

[23] J. P. Dawson, P. J. Adams, and S. N. Pandis, "Sensitivity of ozone to summertime climate in the eastern USA: A modeling case study," *Atmospheric Environment*, vol. 41, pp. 1494–1511, Mar. 2007.

[24] Google, "Raw air quality data from google / aclima." `https://goo.gl/q4TRtt`. Accessed November 1, 2020.

[25] K. P. Messier, S. E. Chambliss, S. Gani, R. Alvarez, M. Brauer, J. J. Choi, S. P. Hamburg, J. Kerckhoffs, B. LaFranchi, M. M. Lunden, J. D. Marshall, C. J. Portier, A. Roy, A. A. Szpiro, R. C. H. Vermeulen, and J. S. Apte, "Mapping Air Pollution with Google Street View Cars: Efficient Approaches with Mobile Monitoring and Land Use Regression," *Environmental Science & Technology*, vol. 52, pp. 12563–12572, Nov. 2018.

[26] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. v. d. Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, I. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, and P. v. Mulbregt, "SciPy 1.0: fundamental algorithms for scientific computing in Python," *Nature Methods*, vol. 17, pp. 261–272, Mar. 2020.

[27] D. Roberts–Semple, F. Song, and Y. Gao, "Seasonal characteristics of ambient nitrogen oxides and ground–level ozone in metropolitan northeastern New Jersey," *Atmospheric Pollution Research*, vol. 3, pp. 247–257, Apr. 2012.

[28] R. E. Britter and S. R. Hanna, "Flow and Dispersion in Urban Areas," *Annual Review of Fluid Mechanics*, vol. 35, no. 1, pp. 469–496, 2003.

[29] E. S. Kimbrough, R. W. Baldauf, and N. Watkins, "Seasonal and diurnal analysis of NO2 concentrations from a long-duration study conducted in Las Vegas, Nevada," *Journal of the Air & Waste Management Association (1995)*, vol. 63, pp. 934–942, Aug. 2013.

[30] J. Richmond-Bryant, M. Snyder, R. Owen, and S. Kimbrough, "Factors associated with NO2 and NOX concentration gradients near a highway," *Atmospheric environment*, vol. 174, pp. 214–226, Nov. 2017.

[31] R. W. Atkinson, B. K. Butland, H. R. Anderson, and R. L. Maynard, "Long-term Concentrations of Nitrogen Dioxide and Mortality," *Epidemiology*, vol. 29, pp. 460–472, July 2018.

[32] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-Means Clustering Algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[33] D. J. Briggs, S. Collins, P. Elliot, P. Fischer, S. Kingham, E. Lebret, K. Pryl, H. V. Reeuwijk, K. Smallbone, and A. V. D. Veen, "Mapping urban air pollution using GIS: a regression-based approach," *International Journal of Geographical Information Science*, vol. 11, pp. 699–718, Oct. 1997.

[34] M. Jerrett, A. Arain, P. Kanaroglou, B. Beckerman, D. Potoglou, T. Sahsuvaroglu, J. Morrison, and C. Giovis, "A review and evaluation of intraurban air pollution exposure models," *Journal of Exposure Science & Environmental Epidemiology*, vol. 15, pp. 185–204, Mar. 2005.

[35] X. Xie, I. Semanjski, S. Gautama, E. Tsiligianni, N. Deligiannis, R. T. Rajan, F. Pasveer, and W. Philips, "A Review of Urban Air Pollution Monitoring and Exposure Assessment Methods," *ISPRS International Journal of Geo-Information*, vol. 6, p. 389, Dec. 2017.

[36] C. Ding, X. He, H. Zha, and H. Simon, "Adaptive dimension reduction for clustering high dimensional data," Tech. Rep. LBNL-51472, Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States), Oct. 2002.

[37] D. Arthur and S. Vassilvitskii, "k-means++: the advantages of careful seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, SODA '07, (USA), pp. 1027–1035, Society for Industrial and Applied Mathematics, Jan. 2007.

[38] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer Series in Statistics, New York, NY: Springer New York, 2009.

[39] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63, no. 2, pp. 411–423, 2001.

[40] R. J. Delfino, R. S. Zeiger, J. M. Seltzer, D. H. Street, and C. E. McLaren, "Association of asthma symptoms with peak particulate air pollution and effect modification by anti-inflammatory medication use.," *Environmental Health Perspectives*, vol. 110, pp. A607–A617, Oct. 2002.

[41] R. Vautard, J. Cattiaux, P. Yiou, J.-N. Thépaut, and P. Ciais, "Northern Hemisphere atmospheric stilling partly attributed to an increase in surface roughness," *Nature Geoscience*, vol. 3, pp. 756–761, Nov. 2010.

[42] D. Belušić, D. H. Lenschow, and N. J. Tapper, “Performance of a mobile car platform for mean wind and turbulence measurements,” *Atmospheric Measurement Techniques*, vol. 7, pp. 1825–1837, June 2014.

[43] J. D. Albertson, T. Harvey, G. Foderaro, P. Zhu, X. Zhou, S. Ferrari, M. S. Amin, M. Modrak, H. Brantley, and E. D. Thoma, “A Mobile Sensing Approach for Regional Surveillance of Fugitive Methane Emissions in Oil and Gas Production,” *Environmental Science & Technology*, vol. 50, pp. 2487–2497, Mar. 2016.

A spatial land use clustering framework for investigating the role of land use in mediating the effect of meteorology on urban air quality

Amir Montazeri, Achim J. Lilienthal, John D. Albertson

Highlights

- Clustering framework developed for spatio-temporal analysis of mobile measurements
- Domain-related procedure for selecting number of clusters in k-means is presented
- Effect of meteorology on pollutant levels as mediated by land-use is investigated
- Wind speed is only effective in reducing pollutant levels in some urban regions

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: