Mining Public Datasets for Modeling Intra-City $PM_{2.5}$ Concentrations at a Fine Spatial Resolution

Yijun Lin
Spatial Sciences Institute
University of Southern California
yijunlin@usc.edu

Yao-Yi Chiang
Spatial Sciences Institute
University of Southern California
yaoyic@usc.edu

Fan Pan
Spatial Sciences Institute
University of Southern California
fanpan@usc.edu

Dimitrios Stripelis
Information Sciences Institute
University of Southern California
stripeli@isi.edu

José Luis Ambite
Information Sciences Institute
University of Southern California
ambite@isi.edu

Sandrah P. Eckel
Department of Preventive Medicine
University of Southern California
eckel@usc.edu

Rima Habre
Department of Preventive Medicine
University of Southern California
habre@usc.edu

ABSTRACT

Air quality models are important for studying the impact of air pollutant on health conditions at a fine spatiotemporal scale. Existing work typically relies on area-specific, expert-selected attributes of pollution emissions (e,g., transportation) and dispersion (e.g., meteorology) for building the model for each combination of study areas, pollutant types, and spatiotemporal scales. In this paper, we present a data mining approach that utilizes publicly available OpenStreetMap (OSM) data to automatically generate an air quality model for the concentrations of fine particulate matter less than 2.5 μm in aerodynamic diameter at various temporal scales. Our experiment shows that our (domain-) expert-free model could generate accurate $PM_{2.5}$ concentration predictions, which can be used to improve air quality models that traditionally rely on expert-selected input. Our approach also quantifies the impact on air quality from a variety of geographic features (i.e., how various types of geographic features such as parking lots and commercial buildings affect air quality and from what distance) representing mobile, stationary and area natural and anthropogenic air pollution sources. This approach is particularly important for enabling the construction of context-specific spatiotemporal models of air pollution, allowing investigations of the impact of air pollution exposures on sensitive populations such as children with asthma at scale.

CCS CONCEPTS

•**Information systems** →*Spatial-temporal systems;*

KEYWORDS

$PM_{2.5}$ Concentration Prediction, Air Quality Modeling, Geospatial Data Mining

1 INTRODUCTION

Fine particulate matter ($PM_{2.5}$) consists of particles less than 2.5 μm in aerodynamic diameter that once inhaled can penetrate the respirable region of the lungs and contribute to respiratory and cardiovascular disease. Typical primary sources of contributions to fine particulate matter include stationary and moving vehicle exhausts, burning sources (e.g., wood-burning stoves and wildfires), refineries, and power plants. Secondary $PM_{2.5}$ is also formed as a result of photochemical reactions in the atmosphere in the presence of precursor gases and solar radiation. Epidemiological studies have shown associations between exposure to $PM_{2.5}$ and various health conditions, including lung and respiratory disease [22], such as asthma [5, 7]. In the 1993 landmark air pollution "Harvard Six Cities Study" [4] and other recent studies [17], researchers reported associations between the levels of exposure to fine particulate concentrations and the risk of mortality and morbidity for cities all over the world. As a result of this scientific evidence and several other similar studies, many countries have set national ambient mass-based air quality standards for $PM_{2.5}$. In the United States, the Environmental Protection Agency (US EPA) set $PM_{2.5}$ standards and established the air quality index (AQI) to communicate relative health risk levels of current pollution levels compared to the standard, which is converted by $PM_{2.5}$ concentration. The US AQI ranges from 0 to 500 and consists of six categories: "Good", "Moderate", "Unhealthy for Sensitive Groups", "Unhealthy", "Very Unhealthy", and "Hazardous". From an ambient pollutant concentration value, one can calculate the corresponding AQI and its health risk category for each type of regulated air pollutant. For example, a 20 g/m^3 $PM_{2.5}$ measurement corresponds to an AQI of 68 and is in the "Moderate" category, which means that *"Air quality is acceptable; however, for some pollutants there may be a moderate health concern for a very small number of people who are*

SIGSPATIAL'17, November 7–10, 2017, Los Angeles Area, CA, USA
© 2017 ACM. 978-1-4503-5490-5/17/11...$15.00
DOI: 10.1145/3139958.3140013

unusually sensitive to air pollution." A 60 g/m^3 PM$_{2.5}$ measurement corresponds to an AQI of 153 and is in the "Unhealthy" category.

In the US, the EPA's ambient air monitoring network provides hourly PM$_{2.5}$ measurements at its regulatory air monitoring stations through the "Air Quality System" (AQS). These monitoring stations are established for regulatory purposes with strict siting criteria to capture regional and urban scale contributions to air pollution levels within an area. These air monitoring stations also exist in many other countries. Scientists and government agencies use measurement data from these stations to build and validate air quality models (AQMs) to explain and predict the past and future air pollution levels for unmonitored locations (e.g., [1, 3, 13, 15, 16, 19, 20, 24, 25, 28, 29]). Predictions from these models can then be used to study the associations between long-term air pollution exposure and health impact at finer spatial scales (than simply using the monitored data) [26, 27].

One popular approach to predicting long-term spatial variations in air pollution levels is land-use regression (LUR) (e.g., [2, 10, 14]) while more recent work uses machine learning techniques (e.g., [3, 12, 19, 21, 23]) and big data (e.g., [28, 29]). Existing air quality models typically consider expert-selected (unique) characteristics in a neighborhood including various types of geographical features (e.g., elevation), proximity to roadways and traffic conditions, population density, and meteorological data. The idea is that air pollutants in "nearby" locations could be spatially auto-correlated or demonstrate comparable concentrations at a given time. This is because geographically proximate locations are surrounded by similar human-made and natural features (emissions and dispersion patterns), including mountains, oceans, roads, factories, and various land-use types. However, building an air quality model that produces accurate air quality concentration predicts at a fine spatiotemporal scale to capture the intra-city air pollution surface is challenging *because there are no universal means to define and quantify location neighborhood of highest influence on local air quality, especially across various cities and regions.* Specifically, separate models require expert-selected location characteristics before the model fitting process to achieve the best regression or machine learning results. (e.g., distance to the ocean has a high correlation to air quality in San Diego but not in every coastal city). The impact of each neighboring location characteristic on air quality can vary significantly across different types air pollutants, time, and space. Moreover, some of the data used in previous studies can be difficult or expensive to obtain and are not frequently available, such as fine-scale, and real-time meteorological data and traffic volumes. (See Section 5 for a review on related work)

This paper presents a novel data mining approach that builds an accurate PM$_{2.5}$ model from publicly available geospatial data, OpenStreetMap (OSM), without using expert knowledge in selecting air quality predictors. Our approach utilizes the PRISMS-DSCIC infrastructure [18] as the data integration and analytics platform to investigate the AQS data of PM$_{2.5}$ concentrations and OSM data. The PRISMS-DSCIC (Pediatric Research using Integrated Sensor Monitoring Systems - Data and Software Coordination and Integration Center) is an NIH-NIBIB (National Institutes of Health - National Institute of Biomedical Imaging and Bioengineering) funded initiative to address pediatric asthma as a chronic disease of childhood. PRISMS-DSCIC is responsible for collecting, storing, integrating, and analyzing real-time environmental, physiological and behavioral data obtained from heterogeneous sensors and traditional data sources to help researchers to predict and prevent asthma attacks efficiently. Using publicly available data that have a global coverage with fine details (in many countries), such as the OSM data, has the advantage that the same approach can apply to many areas across the globe without manual tuning to accommodate available datasets for every study area. Similarly, a recent project using OSM data to generate patterns of human activities in Vienna, Austria demonstrated promising results [11].

Our approach uses the AQS data from twelve SCAQMD (South Coast Air Quality Management District) monitoring stations in the Los Angeles Metropolitan Area (LAMA) and geographic data from OSM to automatically build an air quality model. The model demonstrates on how different types of OSM features impact PM$_{2.5}$ AQIs and from what distance at a given time in LAMA. OSM contains millions of geographic features in LAMA, including points-of-interest, land-use areas, water areas, and road networks (see Section 2). Our algorithm first identifies the air monitoring stations that have a similar temporal pattern of PM$_{2.5}$ AQIs on a temporal resolution. Then using the temporal similarity, the algorithm trains a random forest model to generate the "importance" of individual OSM features (represented by points, lines, and polygons) together with their geographic distances to the monitoring stations (from 100-meter to 3,000-meter radii). For example, suppose the stations that have a similar temporal pattern of PM$_{2.5}$ AQI all have a large factory within 1,000 meters but other stations do not, then the feature-distance pair (factory, 1000-meter buffer) could have a high importance on predicting PM$_{2.5}$ concentrations. We call the geographic characteristics (e.g., factory within 1,000 meters) weighted by the importance the "geo-context". In short, the geo-context represents how each type of OSM features impact PM$_{2.5}$ AQIs in LAMA and from what distance during the period when the AQI data are available.

To predict the PM$_{2.5}$ concentration at a location, *P*, at a given time, our algorithm first generates the geo-context of *P* and the geo-context of all available monitor stations in the study area. Then the algorithm trains a second random forest model using the geo-context and the PM$_{2.5}$ AQIs at available monitor stations to predict the PM$_{2.5}$ concentration at the location *P*. This process works like a recommendation system and helps reduce the prediction errors by considering the temporal effect on the geo-context. For example, a large university campus within 1,500 meters can have a high impact on the PM$_{2.5}$ concentration during rush hours but not at night. The result is an expert-free air quality model for intra-city PM$_{2.5}$ predictions. Our findings can be used to improve air quality models that traditionally rely on geographically weighted interpolations or regressions from (spatially) sparse monitoring stations and can 1) highlight important features or nonlinear interactions amongst them that might have been previously missed with more traditional supervised approaches and 2) be incorporated into more sophisticated prediction models to select and quantify important geographic features related to air quality. This finding is particularly important in the study of air pollution and the impact on relevant populations, such as children with asthma.

The remainder of this paper is organized into four additional sections. Section 2 presents an overview of the data source. Section

3 describes our approach for modeling PM$_{2.5}$ concentrations. Section 4 presents an experiment and evaluation of the results. Finally, Section 5 concludes the paper with a discussion of future work.

2 DATA SOURCES

AQS (Air Quality System) Data

We use the AQS data collected in PRISMS-DSCIC. PRISMS-DSCIC queries the EPA's AirNow web service every hour using multiple zip codes to retrieve the AQS data. For every zip code, PRISMS-DSCIC queries the AirNow service and stores the associated spatiotemporal observations. The observations contain two parts: the environmental air quality indexes (AQI) and the pollution category (Good, Moderate, Unhealthy for Sensitive Groups, Unhealthy, Very Unhealthy, Hazardous) that each AQI measurement correspond. There are twelve monitoring stations provide observations of PM$_{2.5}$ AQI in the Los Angeles Metropolitan Area (Figure 1). In this paper, our approach uses the PM$_{2.5}$ AQI observations from 2016-10-30 12:00:00 to 2017-06-10 12:00:00 with one-hour intervals.

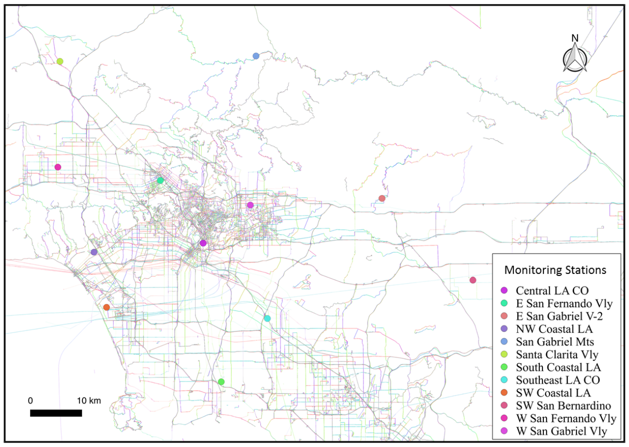


Figure 1: Monitoring Station Locations

Geographic Data

OpenStreetMap (OSM) is an open source, crowdsourced map, which allows people to edit and access global geographic data freely. OSM provides a variety of geographic data types with detailed datasets covering many areas in the world. PRISMS-DSCIC contains a copy of OSM data from Metro Extracts[1] that covers the entire Los Angeles County, including the locations of all the available PM$_{2.5}$ monitoring stations. Example OSM data types (map or geographic features) include land uses, roads, water areas, buildings, aero ways, ocean, etc.[2] OSM land-use polygons describe the primary use of land by the human, such as industrial, residential, and commercial use. OSM road lines include many types of roads, streets, or paths such as motorways, living streets, and footways. OSM water areas are bodies of water, such as lakes or ponds. OSM building types, like point locations of apartments, factories, commercial structures, could reflect the population density and traffic volumes in a local area. OSM aero ways are linear features that represent the physical infrastructure used to support aircraft, air travels, spacecraft, and space flights, which is a large air pollution source.

3 MODELING PM$_{2.5}$ CONCENTRATION

Figure 2 shows our overall approach for building a PM$_{2.5}$ concentration model from OSM and AQS data automatically. After a preprocessing step for data cleaning (Section 3.1), our approach groups available monitoring stations to identify similar temporal patterns on PM$_{2.5}$ AQIs for different time resolutions (hourly, daily, monthly) using the K-means clustering (i.e., each station is a point in the multidimensional space where each dimension is an hour/day/month) (Section 3.2). Our approach uses the clustering result in the next step to quantify the impact of a geographic feature type to PM$_{2.5}$ AQIs. Then the approach generates a "geographic abstraction" for each monitoring station automatically (Section 3.3). The geographic abstraction is a summary of various geographic features for the location using neighborhoods of various sizes. For example, the geographic abstraction can contain the length of different road types (e.g., primary and secondary roads), the counts of various location types (e.g., commercial and residential buildings), the area size of open spaces (e.g., parks and golf courses), and hydrography (e.g., rivers and ocean) within neighborhoods of 100-meter to 3,000-meter radii. Next, the approach trains a random forest model to quantify the importance of individual components in the geographic abstraction based on their supports in grouping monitoring stations of similar temporal patterns on PM$_{2.5}$ AQIs (Section 3.4). We call the geographic abstraction weighted by calculated importance the "geo-context". Finally, the approach uses the geo-context to compute the similarity of the surrounding characteristics for producing the PM$_{2.5}$ concentration prediction for locations that do not have monitoring stations (Section 3.5). The following subsections explain each step of our approach in details.

3.1 Data Preprocessing

In practice, data are generally incomplete (lacking values) and noisy (containing outliers), especially for streaming data. The AQS data quality also suffers from unknown measurement uncertainty and exceptional events that might affect the measurement process. Missing values and errors can have a large impact on the performance of analytic algorithms. Therefore, the first step of our approach is data preprocessing including removing outliers and eliminating missing values in the AQS data.

3.1.1 Removing Outliers. There are several ways to remove data outliers such as computing a sliding window value, clustering to detect and remove outliers, and applying regression analysis to smooth the data. To handle streaming data with a temporal autocorrelation, using a sliding window to filter out noisy data is effective. Our approach calculates the median of a six-hour sliding window. For example, suppose we have a series of streaming PM$_{2.5}$ AQIs with the interval of one hour, $[\cdots, 20, 30, 35, 3, 50, 60, 55, \cdots]$, the sudden drop of AQI of 3 is considered as an outlier. By applying a six-hour sliding window, we replace the sudden drop by the median of the window [20, 30, 35, 3, 50, 60, 55], that is 35.

[1] https://mapzen.com/data/metro-extracts/

[2] http://wiki.openstreetmap.org/wiki/Map_Features

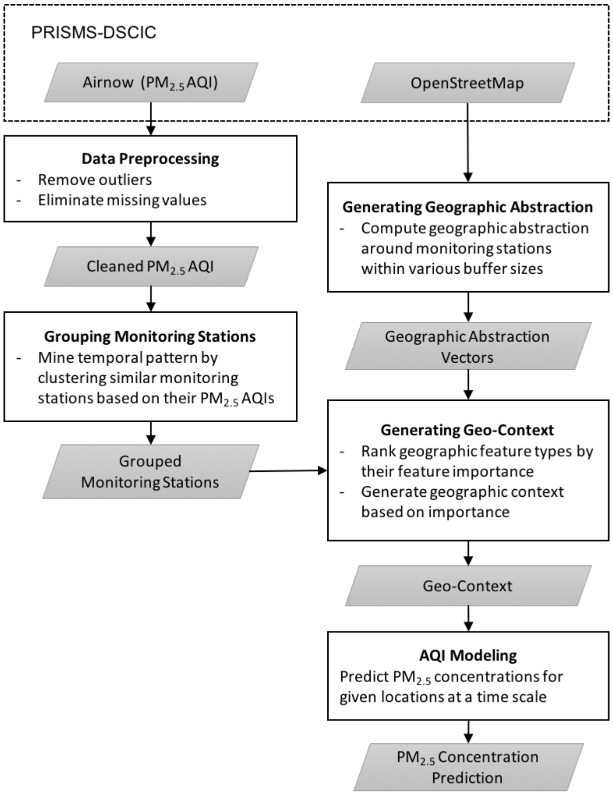

PRISMS-DSCIC
Airnow (PM2.5 AQI)
OpenStreetMap
Data Preprocessing
- Remove outliers
- Eliminate missing values
Cleaned PM2.5 AQI
Grouping Monitoring Stations
- Mine temporal pattern by clustering similar monitoring stations based on their PM2.5 AQIs
Grouped Monitoring Stations
Generating Geographic Abstraction
- Compute geographic abstraction around monitoring stations within various buffer sizes
Geographic Abstraction Vectors
Generating Geo-Context
- Rank geographic feature types by their feature importance
- Generate geographic context based on importance
Geo-Context
AQI Modeling
Predict PM2.5 concentrations for given locations at a time scale
PM2.5 Concentration Prediction

Figure 2: Overall approach for automatically building a $PM_{2.5}$ concentration model from OSM and AQS data

3.1.2 Eliminating Missing Values. The simplest way to eliminating missing values is just ignoring the data tuple when the value is missing. Imputation methods such as using the attribute mean to fill in the missing value or predicting for the missing ones by machine learning algorithms can also achieve satisfactory results, especially in building a recommendation system where lots of the dependent values are not available. Our approach eliminates the missing values by removing the timestamp that does not have a value of $PM_{2.5}$ AQI because filling missing values would require an accurate prediction of the temporal autocorrelation, which might not be robust if the input data are not representative. In our case, the timespan of our AQS data is less than one year.

3.2 Grouping Stations on $PM_{2.5}$ AQIs

In this section, our goal is to identify monitoring stations that have "similar" time-series $PM_{2.5}$ AQIs. We use this information to generate the geo-context in a later step. We define "similar" as in similar temporal pattern on the $PM_{2.5}$ AQIs.Here the temporal pattern is the AQI pattern that occurs at a certain temporal scale, e.g., hourly, daily, and monthly. Our algorithm clusters those monitoring stations with similar temporal patterns in the same group. For example, urban areas would show a higher $PM_{2.5}$ AQI during workdays than rural areas, so urban areas could be grouped together in one cluster, and rural areas are together in another.

Our approach uses K-means to cluster the available monitoring stations based on the collected time-series $PM_{2.5}$ AQIs. K-means clustering is a common method to identify groups in the dataset, with the number of groups represented by the input variable K. The algorithm works iteratively to assign each data point to one of the K groups. Thus, data points are clustered based on the similarity of their feature vector in the Euclidean space.

We construct a feature vector for each monitoring station using their time-series $PM_{2.5}$ AQIs. Table 1 shows an example of 3-hour $PM_{2.5}$ AQIs for the monitoring station in Central LA CO. From the example, our approach generates the feature vector as [50, 53, 55]. In our dataset, we have the AQS data covering 5,352 hours, so for clustering hourly $PM_{2.5}$ AQIs, each feature vector has a total of 5,352 components. In the 5,352-multidimensional space, we have twelve points where each point corresponding to a monitor station.

K-means is a type of unsupervised learning technique, and we need to define the number of groups, K, beforehand. However, the correct choice of K is often unknown in advance. Increasing K without a penalty will always reduce the amount of error in the resulting clustering, to the extreme case of zero errors if each data point is a cluster (i.e., when K equals the number of data points). In our approach, we use the elbow method to determine the value of K. The idea of the elbow method is to run K-means clustering on the dataset for a range of values of K (e.g., K from 1 to 12 in our experiment). For each value of K, we calculate the within set sum of squared errors (WSSSE), which is the sum of the distances between each point and centroid in each K partition. Then we plot a line chart of the WSSSE for each K value. The line chart would look like an arm, and the "elbow" of the arm is the best choice of K. For example, Figure 3 shows that when K equals to 8, the trend becomes slow. Therefore, we choose K equals to 8 as the number of clusters. Figure 4 shows the clustering result of twelve locations using hourly AQIs. We can find that all the coastal areas are grouping together while Central LA is itself in a group because it has a very different temporal patterns of the $PM_{2.5}$ AQI. After determining the best K, our approach uses the K-means results of the identified best K to label the monitoring stations. For example, two monitoring stations that in the same cluster will have the same group label. In the next step, our approach uses the group label of each monitoring station to quantify how each OSM feature supports the clustering result.

Table 1: Example for 3-hour $PM_{2.5}$ AQI in Central LA CO

Monitoring Station	Timestamp	$PM_{2.5}$ AQI
Central LA CO	2017-03-04 12:00:00	50
Central LA CO	2017-03-04 13:00:00	53
Central LA CO	2017-03-04 14:00:00	55

3.3 Generating Geographic Abstraction

$PM_{2.5}$ concentrations are influenced by its surrounding geographic features [2]. In this section, our approach computes a geographic abstraction to describe the surrounding environment for a location. We use the available geographic data from OpenStreetMap, which includes land use, roads, buildings, water ways, aero ways, ocean, etc. For each monitoring station, we construct a series of concentric circles (buffers) with radii from 100 meters to 3,000 meters with interval of 100 meters.

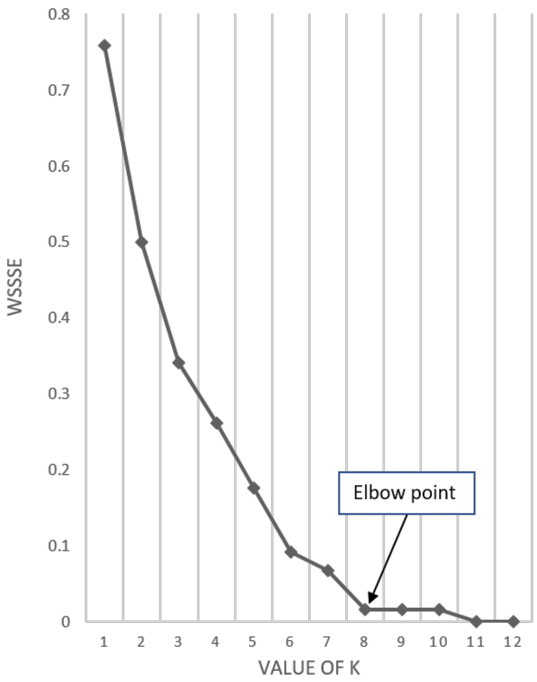


Figure 3: The elbow point for choosing the best K

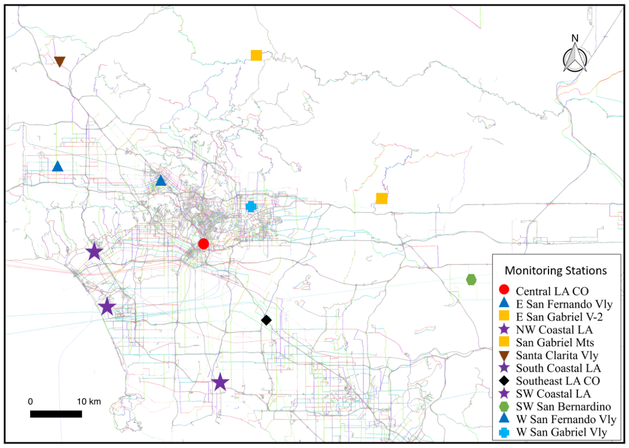


Figure 4: Clustering result of twelve monitoring stations

Length of Line Features

Our approach computes the sum of lengths of different feature types to create a geographic abstraction for line geographic features from OSM, e.g., roads and aero ways. Figure 5 shows an example of roads around the monitoring station A with the 100-meter and 200-meter buffers. In the example, there is a total of three roads of two types, the pedestrian and motorway roads. (Both “Pedestrian” and “Motorway” are OSM feature types.) For each type, we sum up the length of road segments within the buffer. As Figure 5 shows, within the 100-meter buffer, the monitoring station A contains 23-meter (m) Pedestrian and 30m Motorway. Within the 200-meter buffer, it has 43m Pedestrian and 200m Motorway. Thus, our approach generates the components for the geographic abstraction vector for the station A as:

[23, 30, 43, 200]

Each component represents an abstraction of a unique geographic feature type within a specific distance to the monitoring station. The example contributes four components to the abstraction vector: the “Pedestrian” road length in the 100-meter buffer, the “Motorway” length in the 100-meter buffer, the “Pedestrian” road length in the 200-meter buffer, and the “Motorway” length in the 200-meter buffer. Our approach iterates through all available line OSM features to generate an abstraction for every feature type for each buffer size.

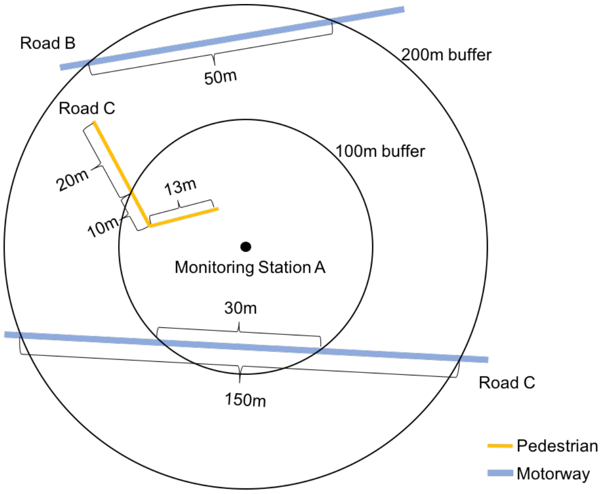


Figure 5: Example for roads in the 100-meter and 200-meter buffers

Area of Polygon Geographic Features

For polygon geographic features such as land uses and water areas, our approach computes the sum of the overlapping areas between each type of the features and the buffers. Figure 6 shows an example of the land use around the monitoring station A within the 100-meter and 200-meter buffers. In this example, there is a total of four area features of two land-use types, park and industrial land-use. (Both “Park” and “Industrial” are OSM feature types.) For each type, our approach calculates the sum of the overlapping areas of the feature type and the buffers (i.e., we only compute the area located within the buffer). As in Figure 6, for the 100-meter buffer, the station A contains 500-square-meter (m^2) of park areas. For the 200-meter buffer, it has 950m^2 park areas and 740m^2 industrial areas. Thus, our approach generates the components for the geographic abstraction vector for the station A as:

[500, 0, 950, 740]

The example contributes four components to the abstraction vector: “Park” areas in the 100m buffer, “Industrial” areas in the 100m buffer, “Park” areas in the 200m buffer, and “Industrial” area in the 200m buffer. Our approach iterates through all available polygon OSM features to generate an abstraction for every feature type for each buffer size.

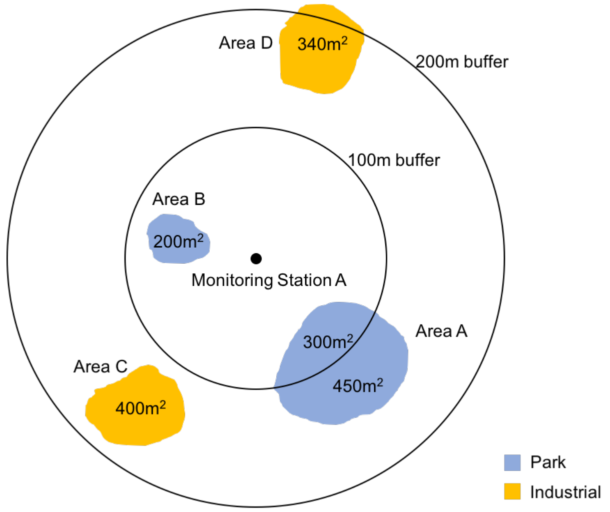


Figure 6: Example for land use in the 100-meter and 200-meter buffers

Count for Point Features

Our approach computes the count of individual types of point features (e.g., building types) to represent the geographic abstraction for point OSM features. The number of buildings in an area could reflect population density and traffic patterns. Figure 7 shows an example of some buildings around the monitoring station A within the 100-meter and 200-meter buffers. In the example, there is a total of twelve buildings of two building types, apartment and factory buildings. (Both "Apartment" and "Factory" are OSM feature types.) In this example, there are two apartments within the 100-meter buffer and eight apartments and three factories within the 200-meter buffer. Thus, our approach generates the components for the geographic abstraction vector for the station A as:

$$[2, 0, 8, 3]$$

The example contributes four components to the abstraction vector: "Apartment" counts in the 100m buffer, "Factory" counts in the 100m buffer, "Apartment" counts in the 200m buffer, and "Factory" counts in the 200m buffer. Our approach iterates through all available point OSM features to generate an abstraction for every feature type for each buffer size.

Distance to Ocean

The geographic abstraction vector also includes a component of the geographic distance from a location to the ocean. For example, suppose the distance from the monitoring station A to ocean is 4000m, the approach generates a feature vector component as:

$$[4000]$$

Generating Vector as Geographic Abstraction

Our approach generates a vector as geographic abstraction for each location. For example, to construct a geographic abstraction vector for the monitoring station A, we combine all the components mentioned above to form a new vector as,

$$[23, 30, 43, 200, 500, 0, 950, 740, 2, 0\ , 8, 3, 4000]$$

Each column of the vector represents the value of a unique geographic feature type with a specific buffer size. Our approach creates buffers from 100 meters to 3,000 meters with an interval of 100 meters. There are more than 3,500 columns in each geographic abstraction vector. Our approach generates the vector for each monitoring station and together the vectors constitute a matrix (Figure 8). In the next step, our approach quantifies the importance of individual components in the geographic abstraction vectors (column in the matrix).

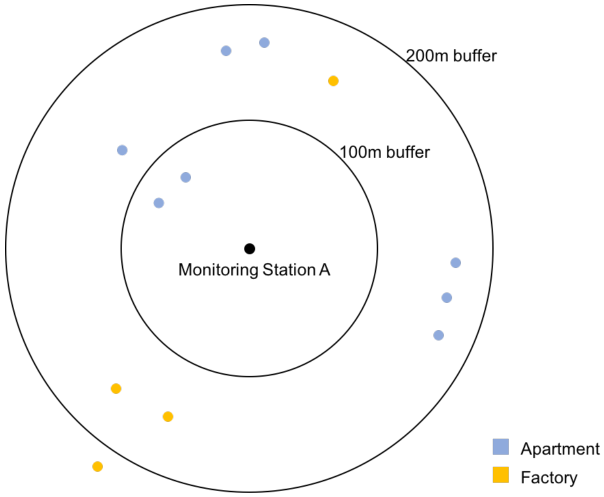
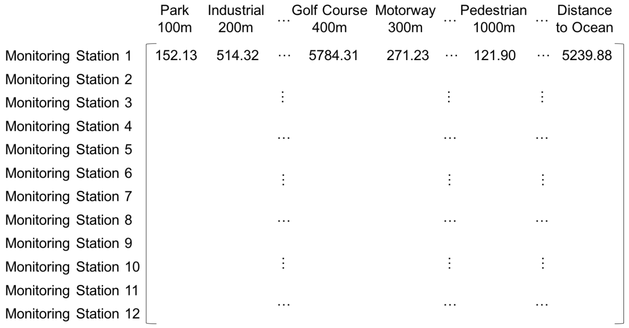


Figure 7: Example for buildings in the 100-meter and 200-meter buffers



	Park 100m	Industrial 200m	…	Golf Course 400m	Motorway 300m	…	Pedestrian 1000m	…	Distance to Ocean
Monitoring Station 1	152.13	514.32	⋯	5784.31	271.23	⋯	121.90	⋯	5239.88
Monitoring Station 2									
Monitoring Station 3			⋮			⋮		⋮	
Monitoring Station 4			⋯			⋯		⋯	
Monitoring Station 5									
Monitoring Station 6			⋮			⋮		⋮	
Monitoring Station 7									
Monitoring Station 8			⋯			⋯		⋯	
Monitoring Station 9									
Monitoring Station 10			⋮			⋮		⋮	
Monitoring Station 11									
Monitoring Station 12			⋯			⋯		⋯	

Figure 8: Geographic abstraction matrix

3.4 Computing Geo-Context

3.4.1 Computing Geographic Importance. In many cases, building an air quality model requires air quality experts to decide which geographic types and what buffer sizes should be considered in the modeling process. However, this process is expensive and time consuming. In this section, we present a method to automatically identify which geographic types with what buffer size have the most impact on $PM_{2.5}$ concentration.

Our approach uses the random forest technique to quantify the importance of individual components in the geographic abstraction vector. Random forest is an ensemble learning method for classification and regression, which consist of multiple single-decision-trees.

When classifying a new object, each tree gives a classification (i.e., the tree "votes" for that class). A random forests classifier chooses the classification that has the most votes. It also provides an easy way to assess feature importance for classification or regression tasks. Our approach uses the grouped monitoring stations as the label (the dependent variable) and their geographic abstractions as the predictor features to train a random forest model. We use the random forest implementation provided in Spark MLlib to derive the feature importance for each component in the geographic abstraction. The idea is that the feature components with higher importance indicate higher impact on the clustering result (based on the $PM_{2.5}$ AQI temporal patterns of the monitoring stations used in the K-means process). The features with zero importance means that they are not important at all in classifying the $PM_{2.5}$ AQI temporal patterns.

3.4.2 Constructing Geo-Context. The feature importance helps us identify which types and what buffer size matters in predicting $PM_{2.5}$ AQI. For each component in the geographic abstraction, we multiply its value by its importance. In this way, we can reward those important features and penalize unimportant ones. For instance, suppose there is a large university area (e.g., 3000m^2) in a 100-meter buffer, but it has zero importance (i.e., it has no relationship with the similarity of $PM_{2.5}$ AQI at different locations), we eliminate its value as it does not exist. We call this weighted geographic abstraction the "geo-context". The geo-context replaces the original geographic abstraction and become a description of the geographic environment around a location for predicting $PM_{2.5}$ concentration.

3.5 Predicting $PM_{2.5}$ Concentration

To predict $PM_{2.5}$ concatenation at a certain time for a target location that does not have air quality sensor, our approach trains a second random forest model with the geo-context (as the predictors) and the $PM_{2.5}$ AQI (as the dependent variable) at that time from all available monitoring stations. Then we construct the geographic abstraction (Section 3.2) for the location and compute the geo-context by applying the feature importance (Section 3.3). Next, we use the trained random forest model to predict the $PM_{2.5}$ AQI for the targeting location and finally convert the predicted $PM_{2.5}$ AQI into $PM_{2.5}$ concentration.

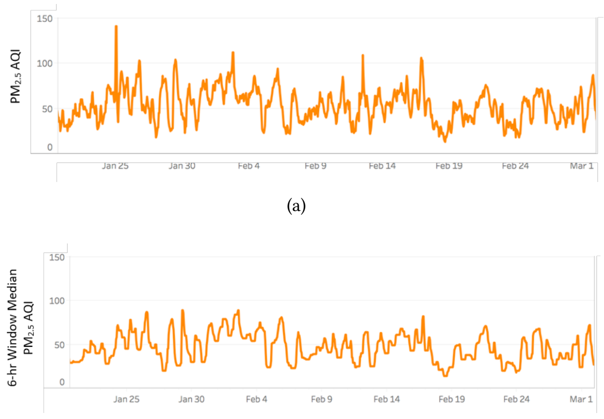
4 EXPERIMENT

We utilized the AQS data (AirNow) and OpenStreetMap data collected in PRISMS-DSCIC for the experiment. We conducted the experiment using the Apache Hue interface, which operates on an interactive session with the Spark cluster on PRISMS-DSCIC. All geospatial computing was done in PostGIS and statistical analysis was done in Scala, version 2.11.8 and Spark MLlib, version 2.1.0.

We performed data preprocessing on the AQS data including removing outliers, eliminating missing values, and aggregating data to lower temporal resolutions (from hourly to daily and monthly). The timespan is seven entire months, 233 days, and 5,352 hours. Figure 9 shows an example result before (a) and after (b) removing outliers using a sliding window of six hours. Our approach also removed the timestamp that did not contain $PM_{2.5}$ AQIs. To conduct our experiment for different temporal resolutions, we computed the mean for daily and monthly $PM_{2.5}$ AQIs for each monitoring station to generate the daily and monthly data.

4.1 Experimental Settings

In the experiment, we tested the performance of both the geographic abstraction and geo-context for generating hourly, daily, and monthly predictions of $PM_{2.5}$ in Los Angeles Metropolitan Area. We verified our results using leave-one-out cross-validation and compared our results with the inverse-distance weighing (IDW) method. We started by taking one monitoring station out (i.e., the target station) and using the remaining 11 stations to predict $PM_{2.5}$ concentration at the hourly, daily, and monthly temporal resolution. We then used the left-out station as the ground truth to calculate the prediction accuracy.



(a)

(b)

Figure 9: Example result of removing outlier before (a) and after (b) for Monitoring Station in Central LA

4.2 Experimental Results and Discussion

We compared the result of our approach (both geographic abstraction and geo-context) to IDW on the same dataset. IDW (Inverse Distance Weighting) is the most frequently used deterministic models in spatial interpolation. We evaluated the performance using RMSE and MAE. RMSE (Root-mean-square deviation) measures the differences between values predicted by a model and the actual values. MAE (Mean absolute error) measures the absolute difference between two continuous variables. We tested on seven months with monthly data, 233 days with daily data, and 168 hours (one day, including 24 hours, randomly chosen from every month) with hourly data. We computed the overall RMSE and MAE for all target locations (twelve monitoring stations). Table 2 presents the evaluation result for month, daily, and hourly, respectively. Figure 10 shows the monthly prediction error for all monitoring stations using three methods. Our approach achieves the best performance with the smallest errors.

Our approach using either geo-context or geographic abstraction generated competitive low RMSE and MAE as other expert-curated models (See Section 5). Using geo-context generated more accurate results from using the geographic abstraction, and both methods were comparable to IDW while IDW cannot provide fine-scale predictions (see next paragraph). We also performed the paired t-test and found that the geo-context MAE results and IDW are statistically different with 95% confidence. For hourly, daily, and monthly predictions, the t-test results were p = 1.73212E-15, p = 1.31243E-05, and p = 0.002, respectively.

Table 2: Result for prediction evaluation

	Geo-Context	Geo-Abstraction	IDW
RMSE (Monthly)	2.53984	2.62391	2.88263
MAE (Monthly)	1.86657	1.93673	2.18675
RMSE (Daily)	4.33786	4.35857	4.10172
MAE (Daily)	3.26140	3.28176	3.10185
RMSE (Hourly)	7.38823	7.59260	6.66106
MAE (Hourly)	5.06559	5.12406	4.54779

To demonstrate our results in predicting fine scale predictions of $PM_{2.5}$ concentration, we created a 1-mile apart fishnet covering most of the City of Los Angeles (604 points). We used our approach to predict the $PM_{2.5}$ AQI monthly mean for each point on the fishnet. Our approach generated a list of feature importance based on the monitoring stations. Table 3 shows the top 15 features ranked by importance. "Motorway", "primary", and "tertiary" are roads that reflect traffic volume. "Village_green", "farmland", and "pitch" are open spaces of green area. "University", "residential", and "retail" are the places attract traffic and people. "Wetland", and "industrial", "garages" are the sources of water pollutants and air pollutants. The results demonstrate that the identified feature types with high importance using the geo-context are similar to other studies in analyzing $PM_{2.5}$ concentration (See Section 5). By automatically quantifying the importance of individual geographic feature types, we could easily explain those geographic feature types affect $PM_{2.5}$ concentrations and from what distance. Figure 11 shows the $PM_{2.5}$ AQI predictions of our approach and IDW for Dec. 2016 (a) and Jan. 2017 (b). As expected, IDW could not generate fine-scale predictions while our approach successfully identified intra-city areas where the air quality is typically poor (e.g., the south part of the city near the port of San Pedro and downtown Los Angeles).

5 RELATED WORK

There exists an abundant literature on air quality modeling and prediction. (The reader is referred to [10] for a review on land-use regression (LUR) methods and [26, 27] for comprehensive reviews on air quality predictions using various methods). The basic and the least computationally expensive methods use spatial interpolations, such as inverse distance weighting (IDW) and Kriging. The methods do not explicitly consider neighborhood characteristics and cannot generate results at a fine-scale with sparse monitoring stations.

Sophisticated and more accurate air quality modeling and prediction methods typically include two steps. First, a domain expert uses knowledge in previous studies on air quality models (AQMs) and statistical methods to test and select the independent variables (predictors). This variable selection step includes choosing a predictor type (e.g., the length of the primary roads and regional average humidity) and a spatial distance. This step largely depends on the availability of a dataset and previous studies of similar pollutants. For example, some studies used crowdsourced data [8] or area specific data (e.g., dense sensors on public transport vehicles [9]) that

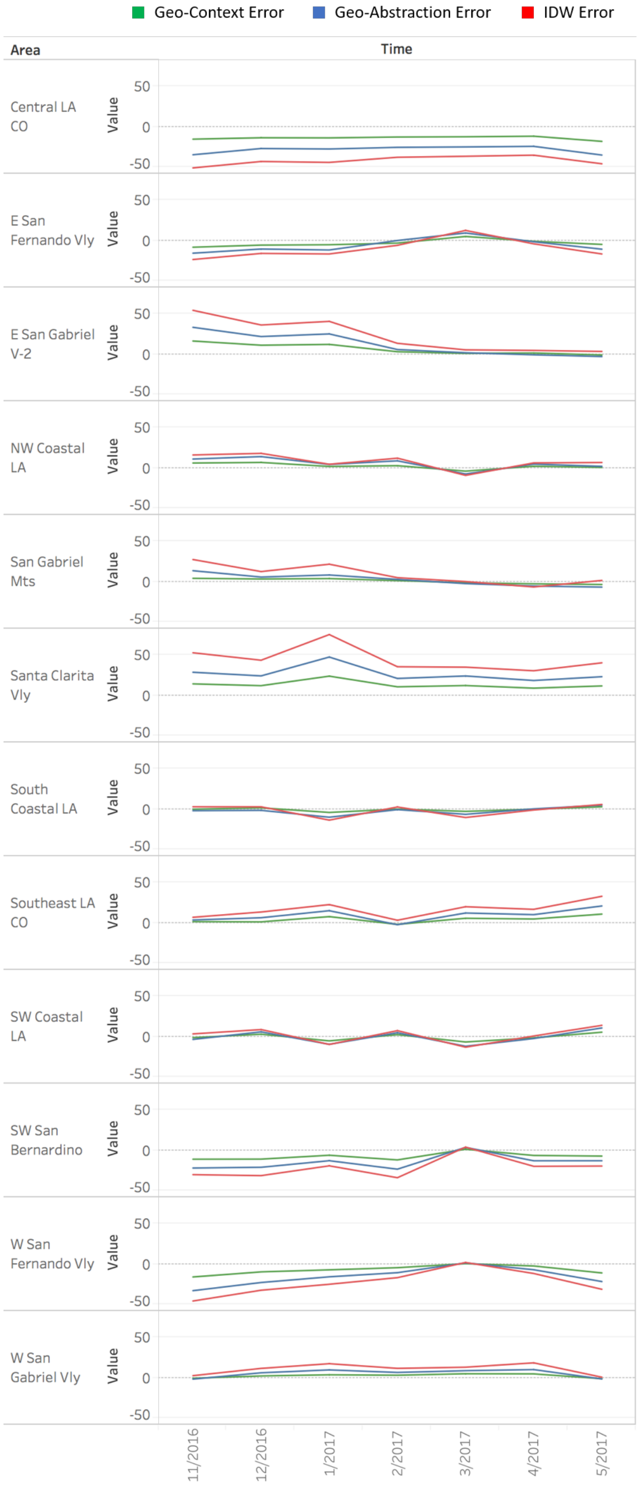


Figure 10: Prediction errors for all monitoring stations

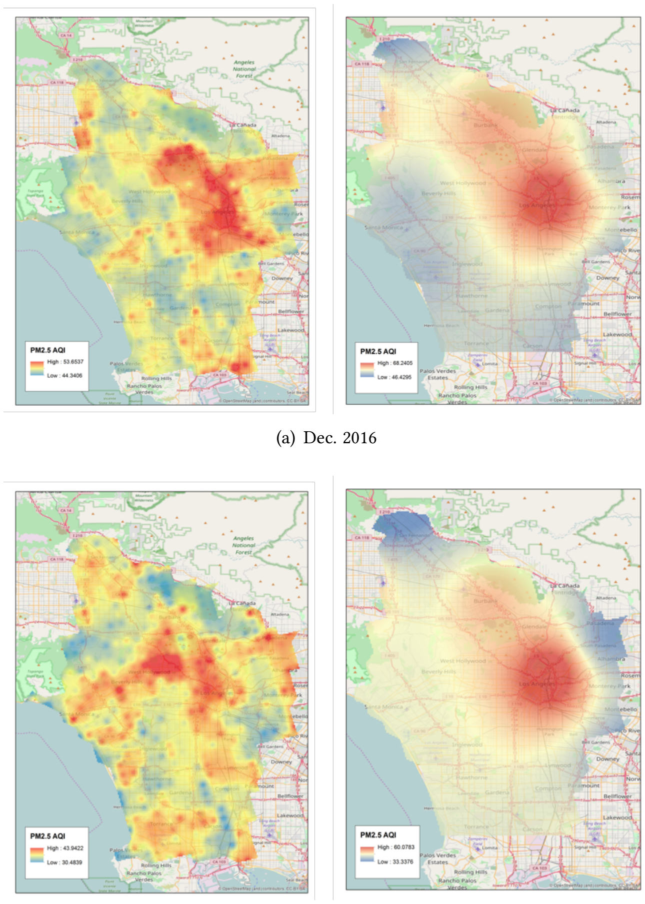
Table 3: Example for feature importance

Geo Name	Buffer Size (meter)	Geo type	Importance (%)
land use	1100	*wetland*	0.0051177
land use	1300	*university*	0.004450
road	600	*rail*	0.0044327
land use	1200	*village_green*	0.0037241
road	700	*primary*	0.0035520
land use	1900	*farmland*	0.0031458
land use	2700	*village_green*	0.0030063
road	800	*residential*	0.0028980
building	2000	*retail*	0.0027980
building	900	*industrial*	0.0027576
road	500	*tertiary*	0.0027357
land use	900	*pitch*	0.0026613
building	2900	*school*	0.0025681
building	1700	*garages*	0.0025361
road	1300	*motorway*	0.0023724

are not available elsewhere. If a dependent variable is not available for a study area, a common approach is to compute surrogate variables. For example, if traffic dynamics are not available, the domain expert would use remotely sensed data or available road data as the surrogate [10]. This step needs to repeat for each study areas and pollutant types [10].

Once the predictor variables are chosen, the second step is to build the prediction models. The mainstream methods include the classical dispersion models, LUR models, and more recently machine learning and data mining models. Dispersion models often require very detailed data (e.g., building heights and distances between neighboring buildings) and area specific parameters [26], which is difficult to generalize and transferred to other locations. Also, dispersion models are usually computationally expensive. In comparison, LUR models have advantages that 1) the results are human-explainable and 2) they have less computational requirement (than dispersion models and machine learning methods).

Since the first LUR study on air pollution modeling in 1997 [2], many LUR models and features are used to study air pollution modeling and predictions (e.g., [10, 14]). However, they heavily rely on expert-selected predictors including predictor types and their finite spatial radii, and every study area requires a domain expert to select and fine-tune the variables. For example, the same radius selected for transportation features in one area might mean something else in another. (Road density within 500 meters in Los Angeles likely captures very different processes than the same variable in rural Montana.) The result is that LUR models in the literature demonstrated significantly different error ranges in their predicts (see [10, 26, 27]). For example, Liu et al. [13] reported high R^2 for their NO and $PM_{2.5}$ LUR models for Shanghai, China, but their RMSE for the 35 verified locations was 194.59 (g/m^3). In Hoek et al. [26], their RMSE ranged from 1.6 to 9.8 (g/m^3) for various types of air pollutants in study areas across the globe, and their temporal resolutions are commonly low (e.g., seasons). Moreover, LUR models rarely deal with spatial effects (e.g., spatial non-stationarity) except a more recent study that built a wind model to improve traditional LUR and had a 10-20% improvement on the prediction [1].



(a) Dec. 2016

(b) Jan. 2017

Figure 11: $PM_{2.5}$ AQIs prediction for Dec. 2016 (a) and Jan. 2017 (b) using geo-context (left) and IDW (right)

With more datasets and software tools becoming available, many studies start to adopt machine learning techniques for building air quality models or predictors [3, 19, 21, 23, 28, 29]. The advantage of using machine learning techniques include 1) the capability to handle large volumes and varieties of data types and formats (e.g., categorical and numerical data), 2) having more accurate prediction results because machine learning methods are less influenced by the choice of parameters or specific dataset (e.g., see [3] for a comparison of LUR and Random Forest), and 3) requiring less expert efforts in selecting input features. Among others, a notable work is the Microsoft Urban Air system [28, 29] that generates air quality predictions covering large areas. While these machine learning methods could achieve more accurate results than the popular LUR models, the price to pay is that the machine learning models are often not easily translatable to policy makers or urban planners (e.g., prediction results from multiple machine learning models). Also, many of the existing studies tested with region specific data sets that are difficult to obtain.

In comparison, our approach is similar to LUR models in that the results are explainable (i.e., the geo-context), but our approach does not require expert-selected predictors. Contrast to dispersion models and the advanced machine learning models (e.g., [28, 29]), our approach is less computationally expensive but generate less accurate results because of the limitation in our predictors. For example, using a static geographic data source, currently our model captures spatial variability but not temporal variability, and future extensions of this work will aim to incorporate meteorology to capture temporality better. More types of globally available data such as the WorldClim (global climate data with 1 km^2 resolution) and satellite imagery could be helpful in improving prediction results of our approach in the future. In sum, the previous studies typically rely on expert-selected and regional available predictors, and our approach is expert-free and can generate an accurate model for predicting intra-city $PM_{2.5}$ concentrations from OSM data.

6 DISCUSSIONS AND FUTURE WORK

This paper presented a data mining approach to build an accurate model to predict $PM_{2.5}$ concentration by automatically selecting important geographic features without using expert knowledge. The advantages of our approach include 1) it can quantify the influence of geographic features on air quality, which helps us do geographic feature selection for air quality analysis without using the domain knowledge; 2) we use the easily accessible OpenStreetMap to construct geographic abstraction instead of using data that is expensive and difficult to obtain; 3) the model performed well in predicting $PM_{2.5}$ concentration and could generate fine-scale predictions. We plan to improve the work presented in this paper in several ways. First, we are going to test our approach with the Esri StreetMap Premium dataset for the same study area, since the data quality of OpenStreetMap cannot be assured [6]. Then we will be able to compare the prediction results from using both the Esri and OSM datasets and learn how data quality affects the air quality model built with our approach. We also plan to compare the work in this paper with our most recent work that uses expert-selected features for air quality modeling [12]. Second, we plan to test the approach for other cities (e.g., Salt Lake City). Third, we plan to incorporate other time-series data, such as weather information to tackle the challenges in modeling spatial effects.

ACKNOWLEDGMENTS

We thank Yi Chen for generating the maps of $PM_{2.5}$ predictions. This work is supported by the NIH grant 1U24EB021996-01.

REFERENCES

[1] S. Bertazzon, M. Johnson, K. Eccles, and G. G. Kaplan. 2015. Accounting for spatial effects in land use regression for urban air pollution modeling. *Spatial and spatio-temporal epidemiology* 14 (2015), 9–21.

[2] D. J. Briggs, S. Collins, P. Elliott, P. Fischer, S. Kingham, E. Lebret, K. Pryl, H. van Reeuwijk, K. Smallbone, and A. Van Der Veen. 1997. Mapping urban air pollution using GIS: a regression-based approach. *International Journal of Geographical Information Science* 11, 7 (1997), 699–718.

[3] C. Brokamp, R. Jandarov, M. B. Rao, G. LeMasters, and P. Ryan. 2017. Exposure assessment models for elemental components of particulate matter in an urban environment: A comparison of regression and random forest approaches. *Atmospheric Environment* 151 (2017), 1–11.

[4] D. W. Dockery, C. A. Pope, X. Xu, J. D. Spengler, J. H. Ware, M. E. Fay, B. G. Ferris Jr, and F. E. Speizer. 1993. An association between air pollution and mortality in six US cities. *New England journal of medicine* 329, 24 (1993), 1753–1759.

[5] J. Fan, S. Li, C. Fan, Z. Bai, and K. Yang. 2016. The impact of $PM_{2.5}$ on asthma emergency department visits: a systematic review and meta-analysis. *Environmental Science and Pollution Research* 23, 1 (2016), 843–850.

[6] M. F. Goodchild and L. Li. 2012. Assuring the quality of volunteered geographic information. *Spatial statistics* 1 (2012), 110–120.

[7] M. Guarnieri and J. R. Balmes. 2014. Outdoor air pollution and asthma. *The Lancet* 383, 9928 (2014), 1581–1592.

[8] D. Hasenfratz, O. Saukh, S. Sturzenegger, and L. Thiele. 2012. Participatory air pollution monitoring using smartphones. *Mobile Sensing* (2012), 1–5.

[9] D. Hasenfratz, O. Saukh, C. Walser, C. Hueglin, M. Fierz, and L. Thiele. 2014. Pushing the spatio-temporal resolution limit of urban air pollution maps. In *IEEE International Conference Pervasive Computing and Communications*. 69–77.

[10] G. Hoek, R. Beelen, K. De Hoogh, D. Vienneau, J. Gulliver, P. Fischer, and D. Briggs. 2008. A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric environment* 42, 33 (2008), 7561–7578.

[11] J. Jokar Arsanjani, M. Helbich, M. Bakillah, J. Hagenauer, and A. Zipf. 2013. Toward mapping land-use patterns from volunteered geographic information. *International Journal of Geographical Information Science* 27, 12 (2013), 2264–2278.

[12] L. Li, F. Lurmann, R. Habre, R. Urman, E. Rappaport, B. Ritz, J.-C. Chen, F. D. Gilliland, and J. Wu. 2017. Constrained Mixed-Effect Models with Ensemble Learning for Prediction of Nitrogen Oxides Concentrations at High Spatiotemporal Resolution. *Environmental Science & Technology* 51, 17 (2017), 9920–9929.

[13] C. Liu, B. H. Henderson, D. Wang, X. Yang, and Z.-R. Peng. 2016. A land use regression application into assessing spatial variation of intra-urban fine particulate matter ($PM_{2.5}$) and nitrogen dioxide (NO 2) concentrations in City of Shanghai, China. *Science of The Total Environment* 565 (2016), 607–615.

[14] D. K. Moore, M. Jerrett, W. J. Mack, and N. Künzli. 2007. A land use regression model for predicting ambient fine particulate matter across Los Angeles, CA. *Journal of Environmental Monitoring* 9, 3 (2007), 246–252.

[15] Z. Ross, P. B. English, R. Scalf, R. Gunier, S. Smorodinsky, S. Wall, and M. Jerrett. 2006. Nitrogen dioxide prediction in Southern California using land use regression modeling: potential for environmental health analyses. *Journal of Exposure Science and Environmental Epidemiology* 16, 2 (2006), 106–114.

[16] Z. Ross, M. Jerrett, K. Ito, B. Tempalski, and G. D. Thurston. 2007. A land use regression for predicting fine particulate matter concentrations in the New York City region. *Atmospheric Environment* 41, 11 (2007), 2255–2269.

[17] J. M. Samet, F. Dominici, F. C. Curriero, I. Coursac, and S. L. Zeger. 2000. Fine particulate air pollution and mortality in 20 US cities, 1987–1994. *New England journal of medicine* 343, 24 (2000), 1742–1749.

[18] D. Stripelis, J. Ambite, Y.-Y. Chiang, S. P. Eckel, and R Habre. 2017. A Scalable Data Integration and Analysis Architecture for Sensor Data of Pediatric Asthma. In *Data Engineering*. 1407–1408.

[19] W. Sun, H. Zhang, A. Palazoglu, A. Singh, W. Zhang, and S. Liu. 2013. Prediction of 24-hour-average $PM_{2.5}$ concentrations using a hidden Markov model with different emission distributions in Northern California. *Science of the total environment* 443 (2013), 93–103.

[20] D. Wilton, A. Szpiro, T. Gould, and T. Larson. 2010. Improving spatial concentration estimates for nitrogen oxides using a hybrid meteorological dispersion/land use regression model in Los Angeles, CA and Seattle, WA. *Science of the total environment* 408, 5 (2010), 1120–1130.

[21] X. Xi, Z. Wei, X. Rui, Y. Wang, X. Bai, W. Yin, and J. Don. 2015. A comprehensive evaluation of air pollution prediction improvement by a machine learning method. In *Service Operations And Logistics, And Informatics*. 176–181.

[22] Y.-F. Xing, Y.-H. Xu, M.-H. Shi, and Y.-X. Lian. 2016. The impact of $PM_{2.5}$ on the human respiratory system. *Journal of thoracic disease* 8, 1 (2016), E69.

[23] W. Xu, C. Cheng, D. Guo, X. Chen, H. Yuan, R. Yang, and Y. Liu. 2014. PM2. 5 Air Quality Index Prediction Using an Ensemble Learning Model. In *International Conference on Web-Age Information Management*. Springer, 119–129.

[24] X. Yang, Y. Zheng, G. Geng, H. Liu, H. Man, Z. Lv, K. He, and K. De Hoogh. 2017. Development of $PM_{2.5}$ and NO_2 models in a LUR framework incorporating satellite remote sensing and air quality model data in Pearl River Delta region, China. *Environmental Pollution* 226 (2017), 143–153.

[25] L. Zhai, B. Zou, X. Fang, Y. Luo, N. Wan, and S. Li. 2016. Land Use Regression Modeling of $PM_{2.5}$ Concentrations at Optimized Spatial Scales. *Atmosphere* 8, 1 (2016), 1.

[26] Y. Zhang, M. Bocquet, V. Mallet, C. Seigneur, and A. Baklanov. 2012. Real-time air quality forecasting, part I: History, techniques, and current status. *Atmospheric Environment* 60 (2012), 632–655.

[27] Y. Zhang, M. Bocquet, V. Mallet, C. Seigneur, and A. Baklanov. 2012. Real-time air quality forecasting, part II: State of the science, current research needs, and future prospects. *Atmospheric Environment* 60 (2012), 656–676.

[28] Y. Zheng, F. Liu, and H.-P. Hsieh. 2013. U-air: When urban air quality inference meets big data. In *Proceedings of the 19th ACM international conference on Knowledge discovery and data mining*. 1436–1444.

[29] Y. Zheng, X. Yi, M. Li, R. Li, Z. Shan, E. Chang, and T. Li. 2015. Forecasting fine-grained air quality based on big data. In *Proceedings of the 21th ACM International Conference on Knowledge Discovery and Data Mining*. 2267–2276.