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

Yijun Lin, Yao-Yi Chiang, Fan Pan
Spatial Sciences Institute

Dimitri Stripelis, Jose Luis Ambite
Information Sciences Institute

Sandrah P. Eckel, Rima Habre
Department of Preventive Medicine

University of Southern California

USC Dornsife
Dana and David Dornsife
College of Letters, Arts and Sciences

Spatial Sciences Institute

Outline

- Introduction and Data Sources
- Approach and Algorithm
- Experiments and Results
- Related Work
- Conclusion and Future work

Background

?

AIR QUALITY

Existing Work

Authors	Study area	Monitor counts	Dependent variables	Independent variables	Buffer size	(Adjusted) R^2
Briggs et al. (2000)	Huddersfield (UK) Sheffield (UK) Northampton (UK)	20, 28 and 35	NO_2	Road traffic, urban land, and topography (altitudes)	300 m	0.58 to 0.76
Ross et al. (2007)	New York City (US)	28–49	$PM_{2.5}$	Traffic, land use, census	50, 100, 300, 500 and 1000 m	0.607 to 0.642
Su et al. (2008)	Greater Vancouver Regional District, (Canada)	116	NO/NO_2	Road, traffic, meteorology (wind speed, wind direction and cloud cover/insolation)	3000 m	0.53 to 0.60
Mavko et al. (2008)	Portland, (US)	77	NO_2	Traffic-related; Land use-related; Elevation; height from MSL; distance to a river; wind; direction	50, 100, 250, 300, 350, 400, 500, 750 m.	0.66 to 0.81
Rivera et al. (2012)	Girona province, (Spain)	25	Ultrafine particles (UFP)	Heavy, light and motorcy. veh in 24 h; 24 h total traffic load; length of major roads; building density; distance to bus lines, highway and intersections; land cover	25, 50, 100, 150, 300, 500 and 1000 m	0.36 to 0.72
Eeftens et al. (2012)	20 European regions	20 per area	$PM_{2.5}$, PM_{10} and PMcoarse	Traffic intensity, population, and land-use	25, 50, 100, 300, 500, and 1000 m	0.35 to 0.94
Dons et al. (2013)	Flanders, (Belgium)	63	Traffic related air pollutant black carbon	Hourly traffic streams, daily traffic volumes, total road length; population density and address density; land use variables	50, 100, 1000 m	0.44 to 0.77
Lee et al. (2014)	Taipei, (China)	40	NO_x and NO_2	Land use, no. of population and households, road length, altitude, distance to roads, ports	25, 25–50, and 50–500 m	0.63 to 0.81
Wu et al. (2015)	Beijing, (China)	35	$PM_{2.5}$	Traffic intensity, population, bus stops, restaurants, and land-use	100–3000 m	0.43 to 0.65

Expert-selected
Area-specific

e.g., $PM_{2.5}$ concentrations is high near **500 meters** of **highways** in **Los Angeles**

Source: Liu et al., 2016

Challenges and Solution

Challenges in air quality modeling

- Rely on **area-specific, expert-selected** features
  - Feature types, distance, etc.

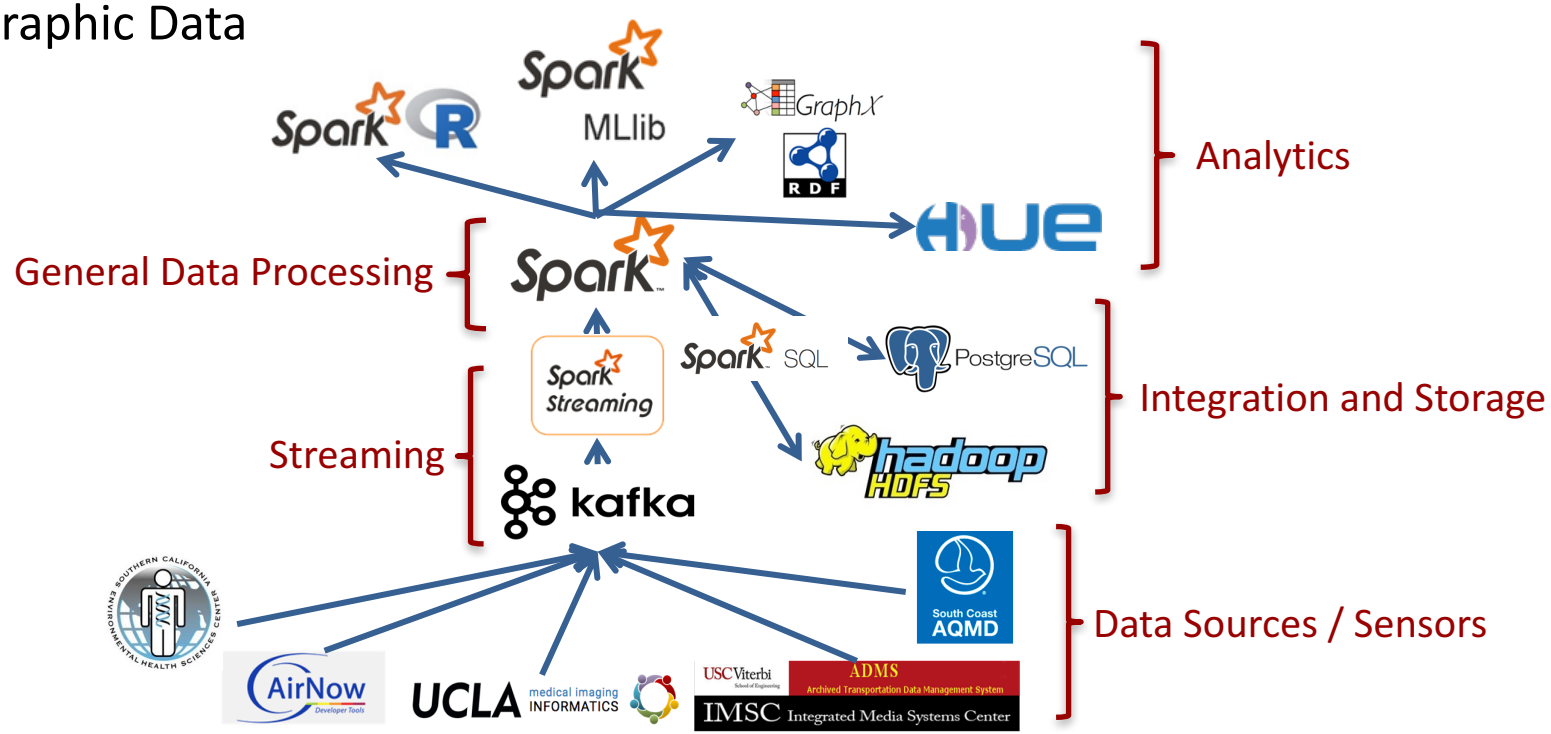
Our solution

- Quantifying the impact on $PM_{2.5}$ concentrations from a variety of geographic features
  - **Types** of geographic features (e.g., primary roads, industrial areas, parks)
  - **Distance** to geographic features (i.e., varying buffer sizes from 100m to 3,000m)
- Generating an air quality model to predict $PM_{2.5}$ concentrations with the "automatically-selected" inputs at a fine scale

Data Collection

PRISMS-DSCIC – A scalable data integration and analysis architecture

- AQS (Air Quality System) Data
- Geographic Data



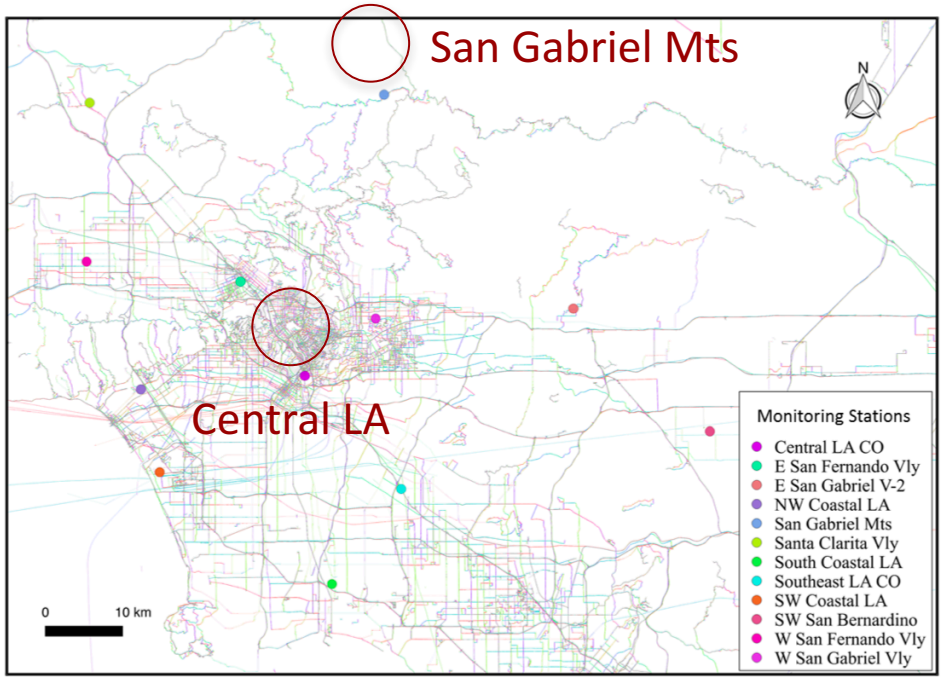
Data Sources – I

AQS (Air Quality System) Data

- Hourly $PM_{2.5}$ AQI from **12 monitoring stations** in the Los Angeles Area from 2016-10-30 00:00:00 to 2017-08-31 23:00:00

Monitoring Station	Timestamp	$PM_{2.5}$ AQI
San Gabriel Mts	2017-03-04 12:00:00	44
San Gabriel Mts	2017-03-04 13:00:00	54
Central LA	2017-03-04 12:00:00	60
Central LA	2017-03-04 13:00:00	68

Sample data



Data Sources – II

Geographic Features - OpenStreetMap (OSM)

- Land uses (67,972 polygons), Roads (544,142 lines), Water areas (11,207 polygons), Buildings (2,971,349 points), Aeroways (962 lines), etc.
- Each geographic category contains various feature types
  - E.g., types for Buildings: commercial, apartment, house, industrial, school, etc.




Outline

- Introduction and Data Sources
- Approach and Algorithm
- Experiments and Results
- Related Work
- Conclusion and Future work

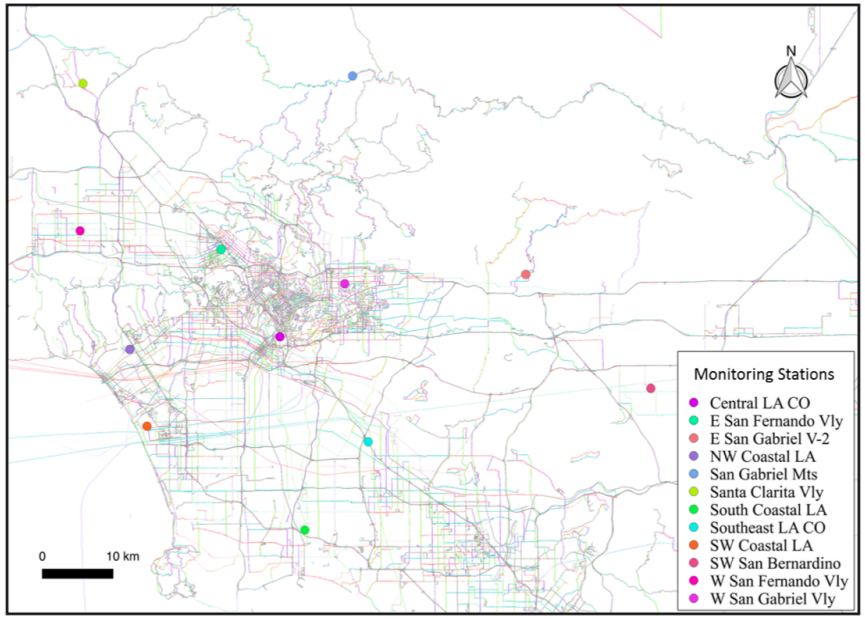
?

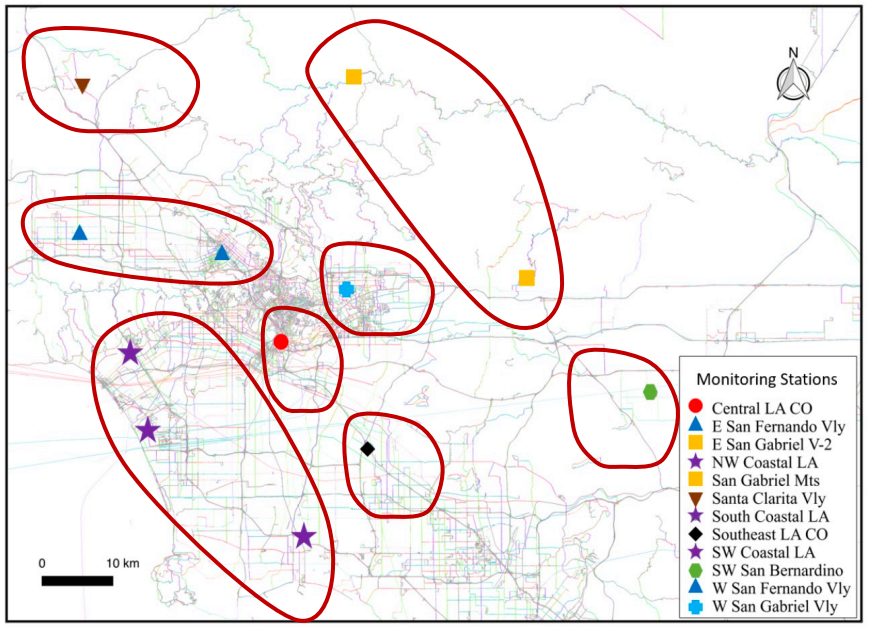
$PM_{2.5}$ AQI

Step 1. Grouping Stations on $PM_{2.5}$ AQIs

Goal?

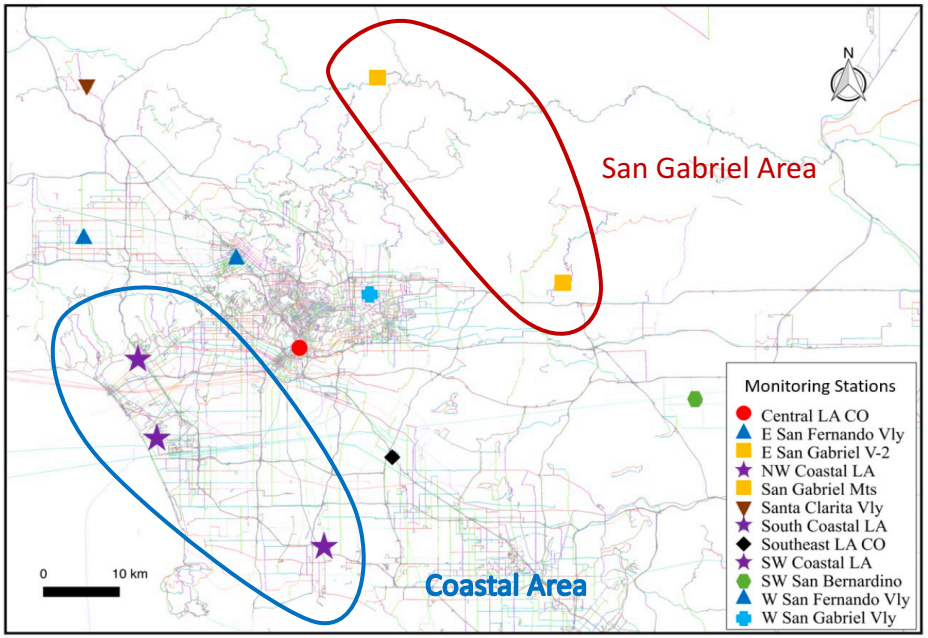
- To identify the monitoring stations that have similar temporal pattern on $PM_{2.5}$ AQIs

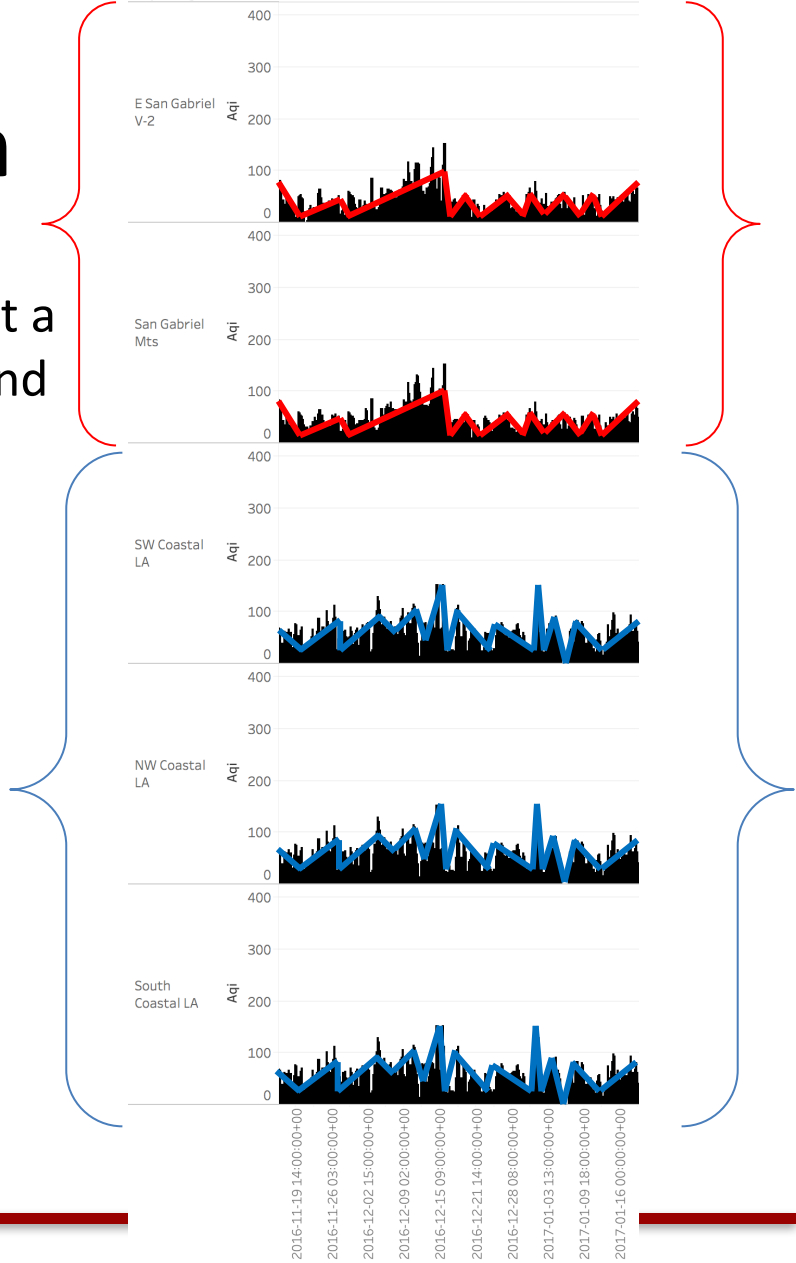




Similar Temporal Pattern

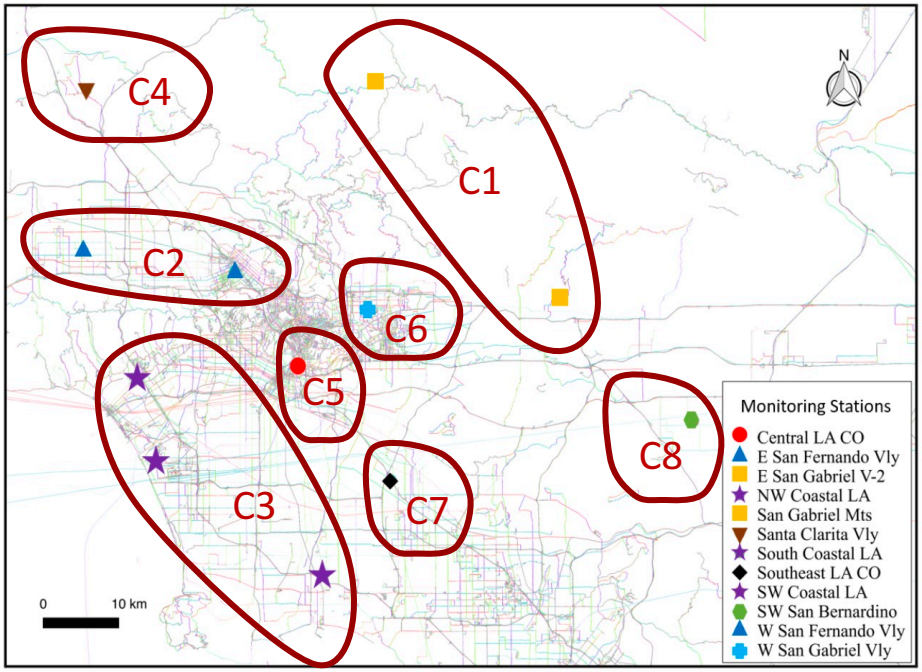
- Temporal pattern: the AQI patterns occur at a certain temporal scale, e.g., hourly, daily, and monthly.





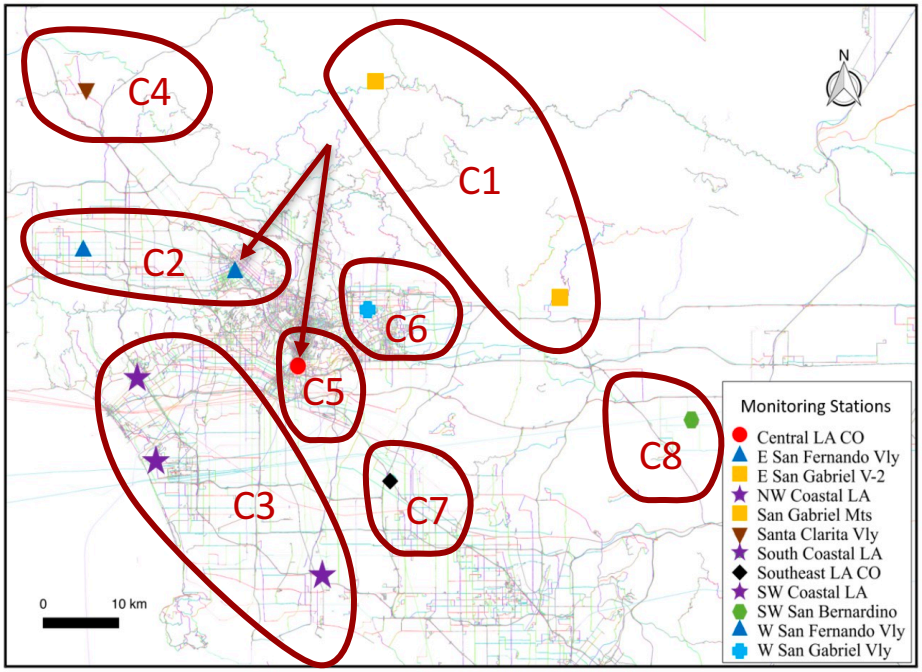
Outputs of Clustering

- Identifying the clusters by using cluster method (K-means)
- Next, we will find out **what specific geographic feature types** (e.g., primary roads, industrial areas, parks) and **from what distance** have the most impact on the clustering result.



Outputs of Clustering

- Identifying the clusters by using cluster method (K-means)
- Next, we will find out **what specific geographic feature types** (e.g., primary roads, industrial areas, parks) and **from what distance** have the most impact on the clustering result.



Step 2. Generating Geographic Abstraction

Length of line features

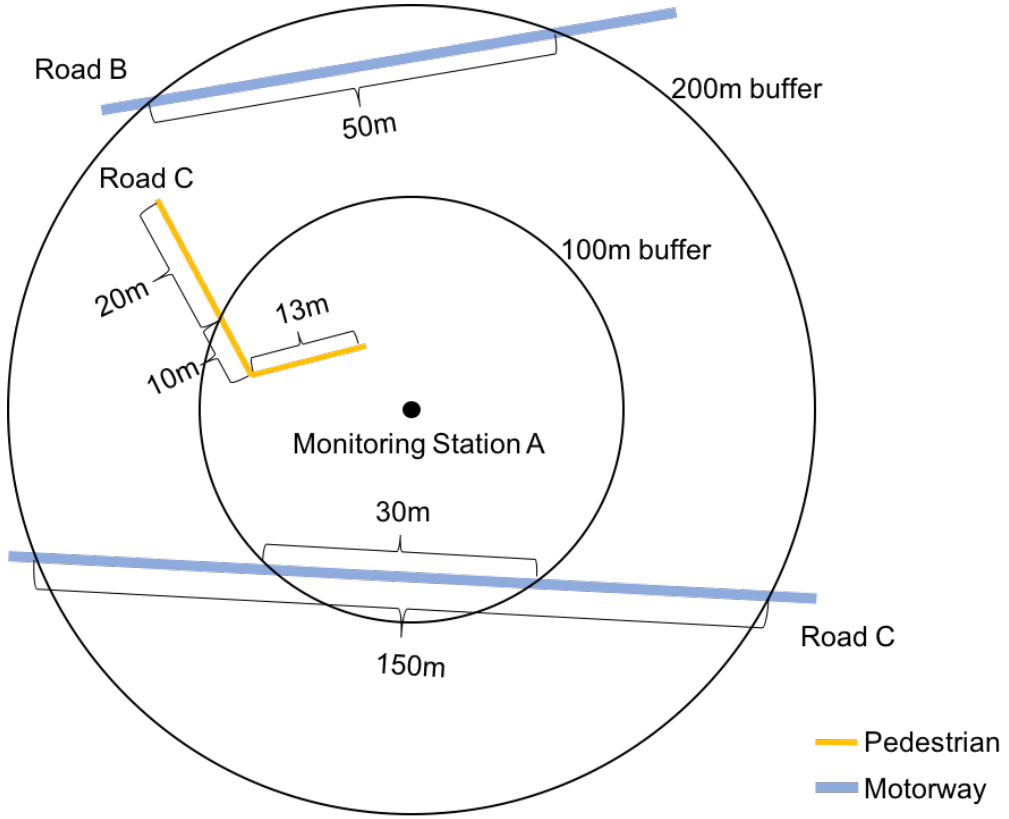
- e.g., Roads, Aeroways

Example for Roads

	100m	200m
Pedestrian	23	43
Motorway	30	200

↓

[23, 30, 43, 200]



Step 2. Generating Geographic Abstraction

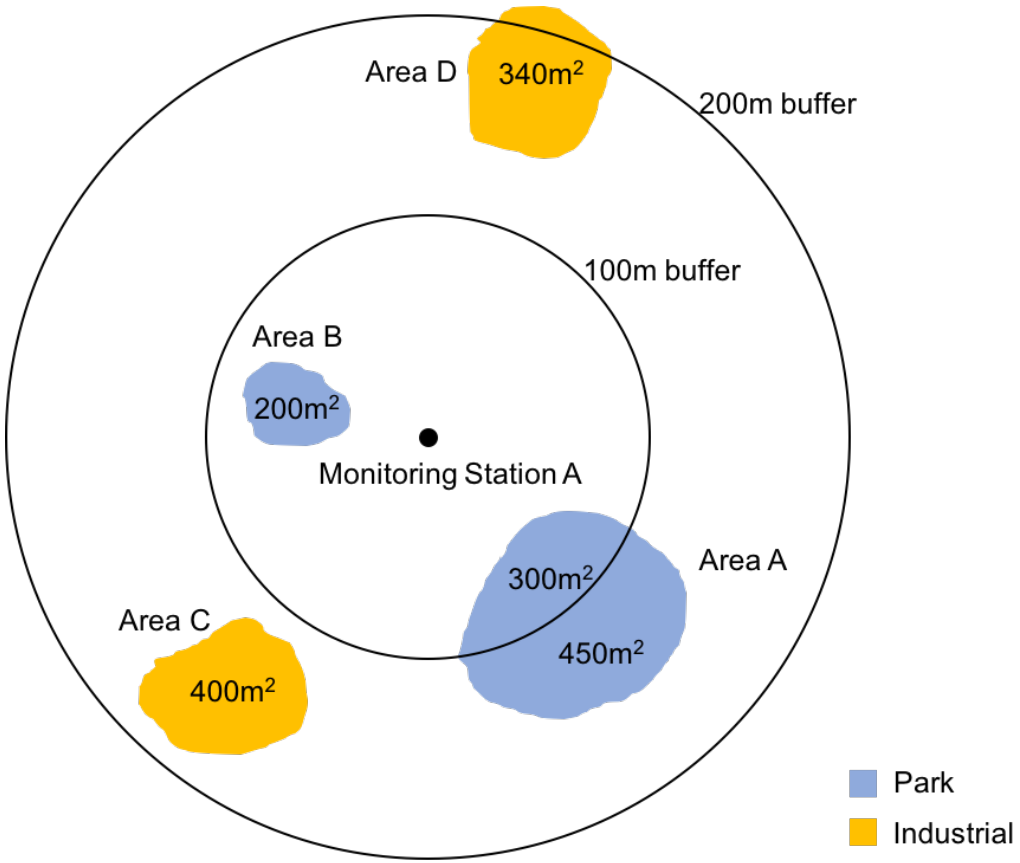
Area of polygon features

- e.g., Land uses, Water areas

Example for Land uses

	100m	200m
Park	500	950
Industrial	0	740

[500, 0, 950, 740]



Step 2. Generating Geographic Abstraction

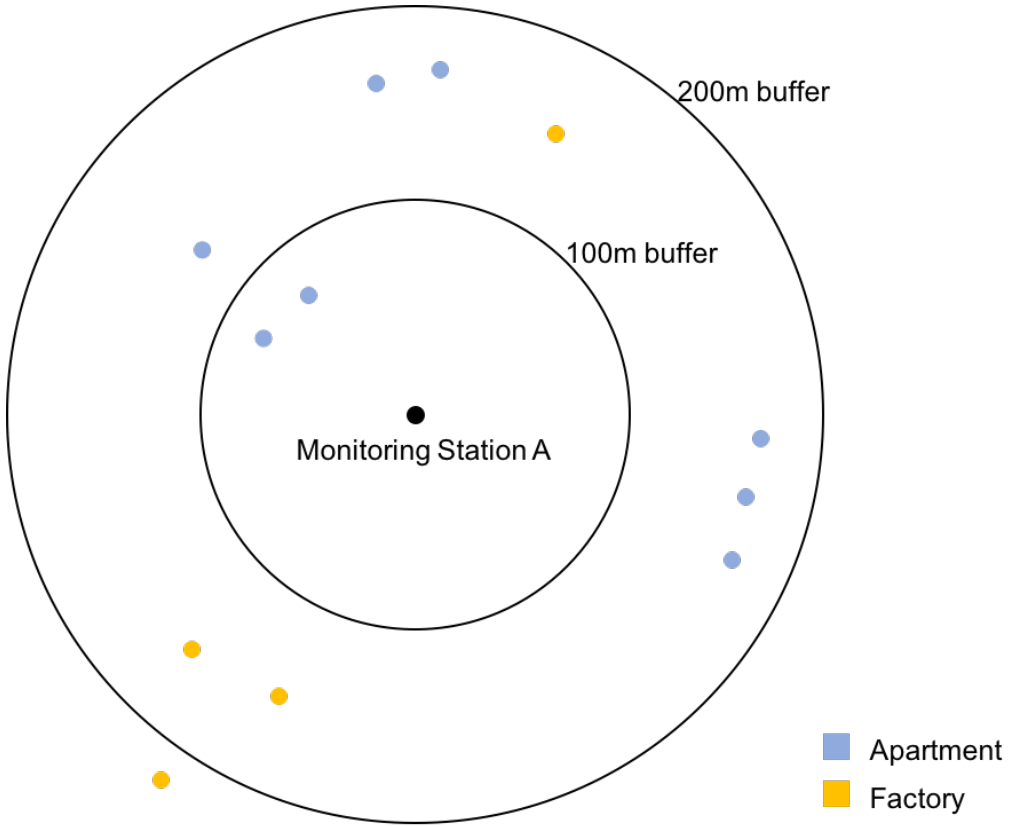
Count of point features

- e.g., Buildings

Example for Buildings

	100m	200m
Apartment	2	8
Factory	0	3

[2, 0, 8, 3]



Step 2. Generating Geographic Abstraction

- Generating a vector for each monitoring station as Geographic Abstraction

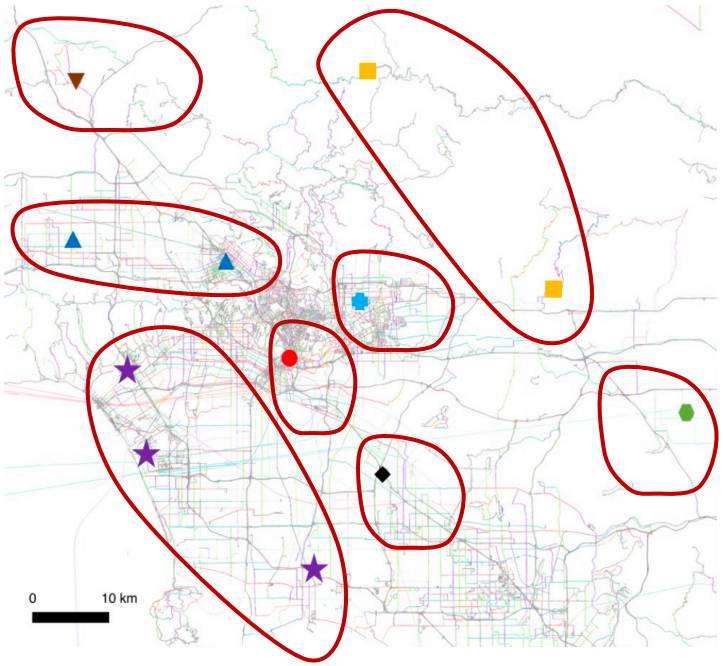
	Pedestrian 100m	Motorway 100m	Pedestrian 200m	Motorway 200m	Park 100m	Industrial 100m
Monitoring Station A	23	30	43	200	500	0

Park 200m	Industrial 200m	Apartment 100m	Factory 100m	Apartment 200m	Factory 200m	Distance to Ocean
950	740	2	0	8	3	4000

- In practice, we creates buffers from 100 meters to 3,000 meters with an interval of 100 meters.

Geographic abstraction → How? → Clustering result

	Pedestrian 100m	Motorway 100m	...	Apartment 200m	Factory 200m
Monitoring Station 1	23	30		2	3
Monitoring Station 2	10	25		1	3
Monitoring Station 3	56	100		8	0
.....					
Monitoring Station 12	67	10		8	0



Step 3. Geo-context

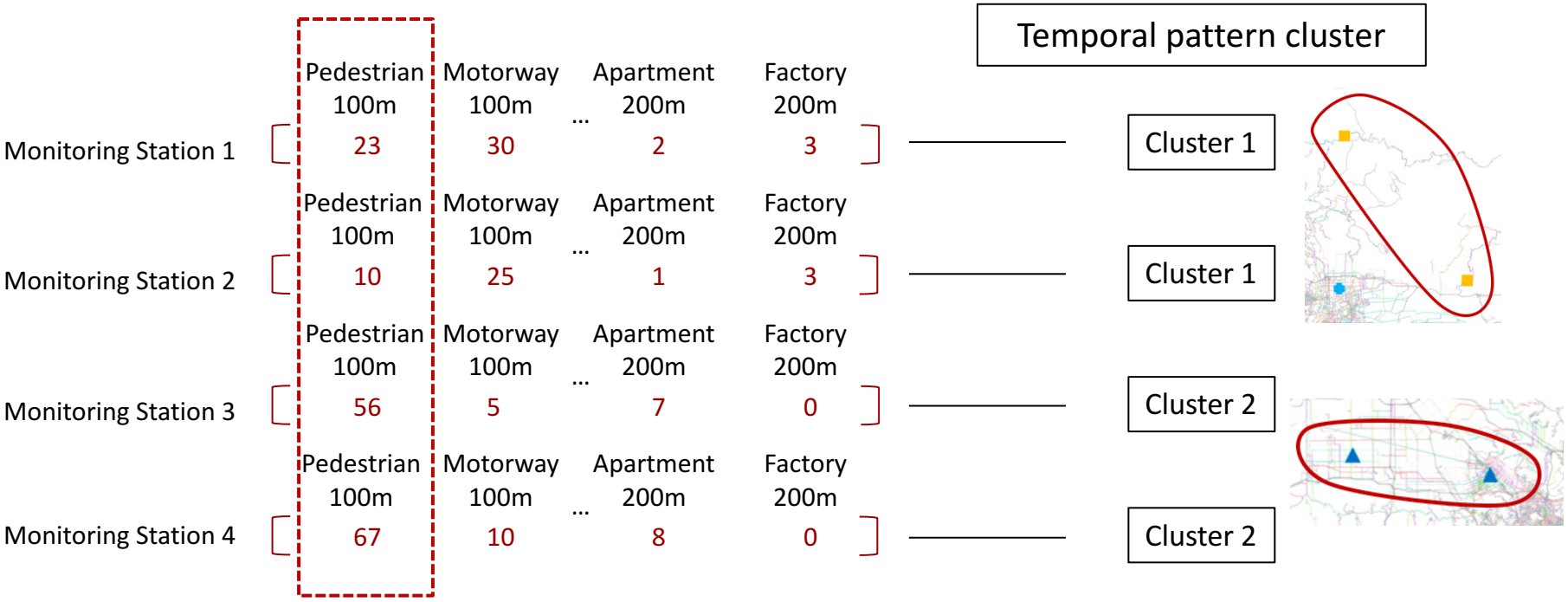
What is geo-context?

- An updated vector that describes how each feature type within a certain distance (each column) in Geographic Abstraction affects the Temporal Pattern ($PM_{2.5}$ AQI).

- It rewards the important (relevant) features and penalizes others

Step 3. Computing Feature Importance

- Compute feature importance by training a random forest model, i.e., quantify the impact of each geographic feature type within certain distance (column) on the temporal pattern



Step 3. Generating Geo-context

- Multiplying each geographic abstraction value by its feature importance to generate geo-context

$$Geographic\ Abstraction\ Vector\ \boldsymbol{A} = [a_1, a_2, ..., a_n]$$

$$Importance\ Vector\ \boldsymbol{I} = [i_1, i_2, ..., i_n]$$

$$Geo\text{-}Context\ Vector\ \boldsymbol{C} = \boldsymbol{A} * \boldsymbol{I}$$

	Pedestrian 100m	Motorway 100m	⋯	Apartment 200m	Factory 200m
Monitoring Station 1 (Geographic Abstraction)	23	30		2	3
↓					
Monitoring Station 1 (Geo-context)	0.0	3.27	...	0.041	0.432

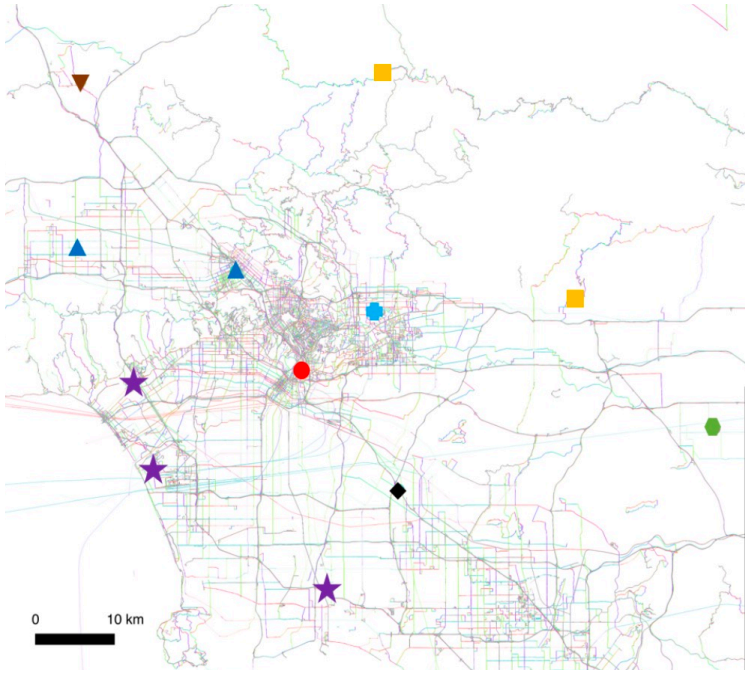
Example of Importance

Geo-feature	Importance
Pedestrian 100m	0.000
Motorway 100m	0.109
...	...
Apartment 200m	0.041
Factory 200m	0.144
...	...
Total	1.0

Step 4. Predicting $PM_{2.5}$ AQI

To predict $PM_{2.5}$ AQI for a target location at time T

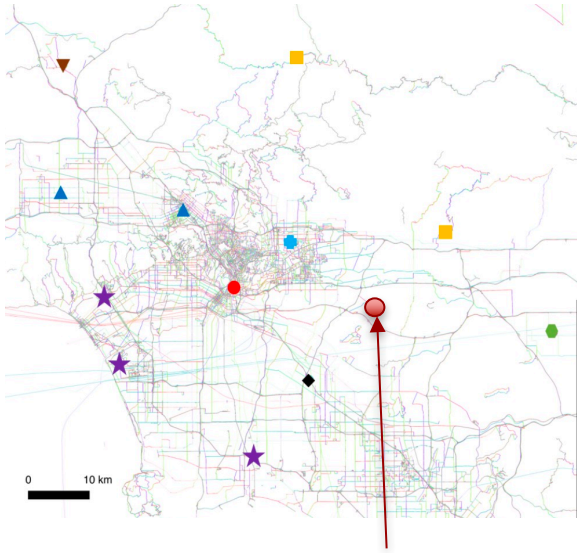
[Geo-context, AQI] for each monitoring stations at time T



Step 4. Predicting $PM_{2.5}$ AQI

To predict $PM_{2.5}$ AQI for a target location at time T

[Geo-context, AQI] for each monitoring stations at time T

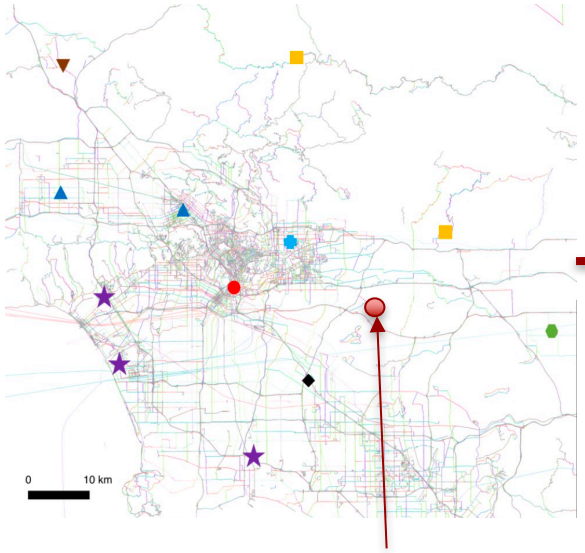


[Geo-context] for target location

Step 4. Predicting $PM_{2.5}$ AQI

To predict $PM_{2.5}$ AQI for a target location at time T

[Geo-context, AQI] for each monitoring stations at time T



Regression Model

AQI prediction of target location

[Geo-context] for target location

Outline

- Introduction and Data Sources
- Approach and Algorithm
- Experiments and Results
- Related Work
- Conclusion and Future work

Experiments

Leave-one-out cross-validation method

- Predict $PM_{2.5}$ AQI for the removed station by using other 11 stations

Predicting at a fine scale

- Predict $PM_{2.5}$ AQI of each point in a 1-mile-apart fishnet covering most of Los Angeles area (604 points)

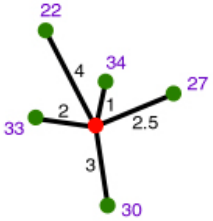
Experiment & Result – I

Leave-one-out cross-validation method

- Predicted with three methods on three temporal scales
  - Geo-context, Geo-abstraction, IDW (Inverse distance weighting)
  - Monthly (7 months), daily (233 days), and hourly (168 hours)

	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

IDW method



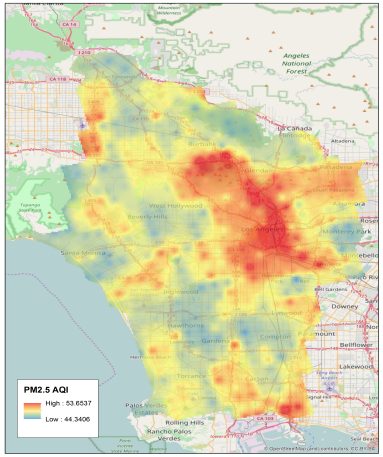
$$Z(x) = \frac{\sum w_i z_i}{\sum w_i} = \frac{\frac{34}{1^2} + \frac{33}{2^2} + \frac{27}{2.5^2} + \frac{30}{3^2} + \frac{22}{4^2}}{\frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{2.5^2} + \frac{1}{3^2} + \frac{1}{4^2}} = 32.38$$

Experiment & Result – II

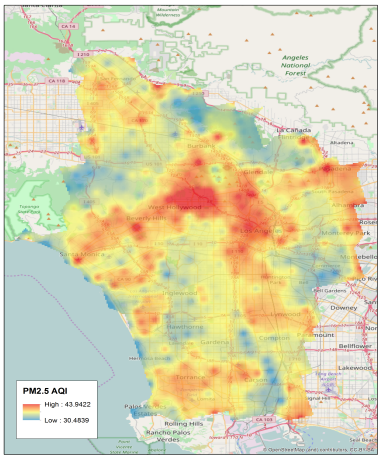
Predicting $PM_{2.5}$ AQIs at a fine scale

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

Geo-context

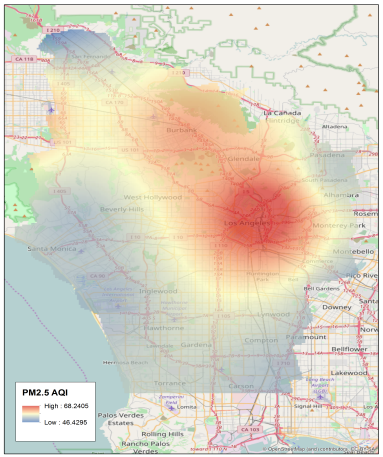


Dec 2016

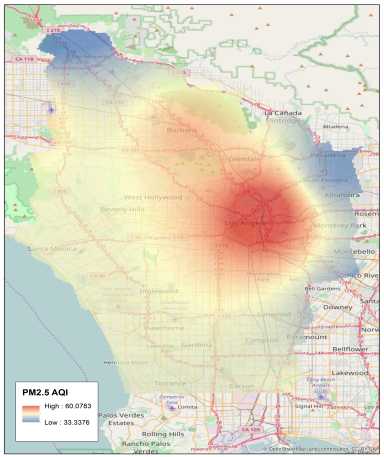


Jan 2017

IDW



Dec 2016



Jan 2017

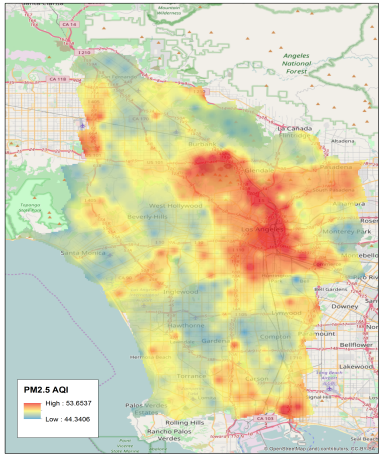
Experiment & Result – II

Predicting $PM_{2.5}$ AQIs at a fine scale

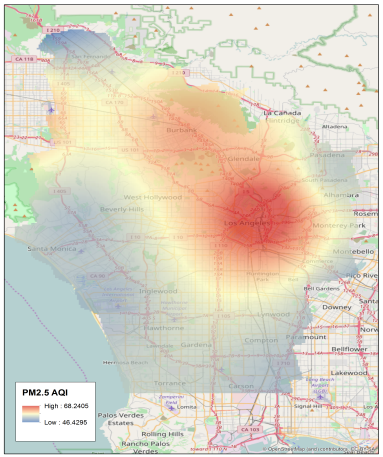
Geo Name	*Buffer Size (meter)*	*Geo type*	*Importance (%)*
land use	1100	*wetland*	0.0051177
land use	1300	*university*	0.004450
road	600	*rail*	0.0044327
land use	1200	*village_green*	0.0037241
road	700	*primary*	0.0035520
land use	1900	*farmland*	0.0031458
land use	2700	*village_green*	0.0030063
road	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

Geo-context

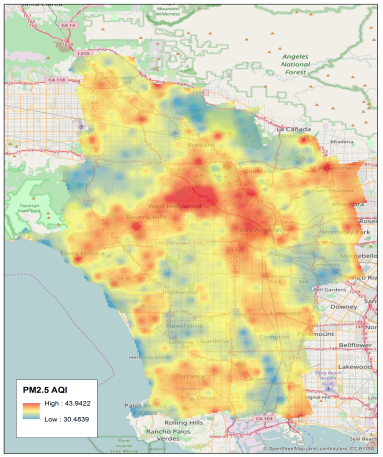
IDW



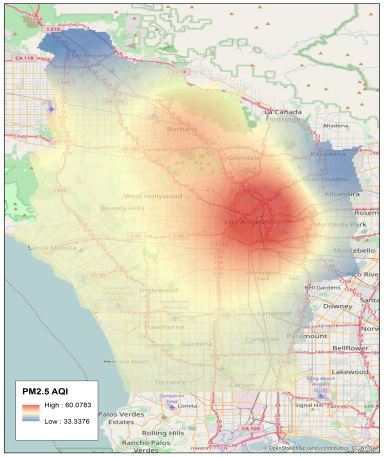
Dec 2016



Dec 2016



Jan 2017



Jan 2017

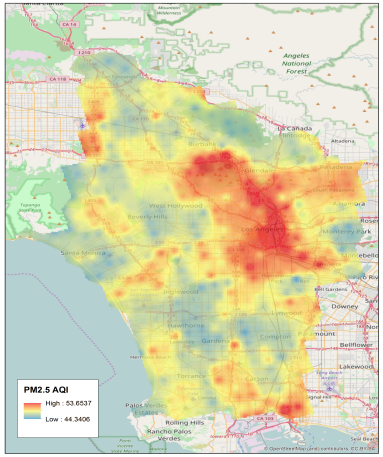
Experiment & Result – II

Predicting $PM_{2.5}$ AQIs at a fine scale

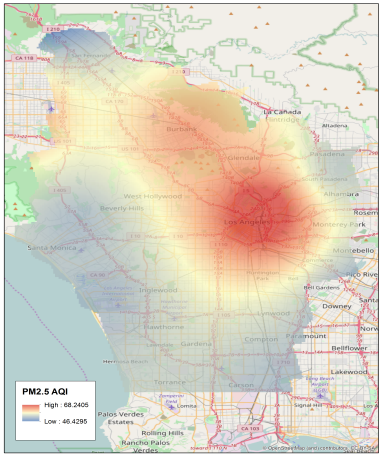
Geo Name	*Buffer Size (meter)*	*Geo type*	*Importance (%)*
land use	1100	*wetland*	0.0051177
land use	1300	*university*	0.004450
road	600	*rail*	0.0044327
land use	1200	*village_green*	0.0037241
road	700	*primary*	0.0035520
land use	1900	*farmland*	0.0031458
land use	2700	*village_green*	0.0030063
road	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

Geo-context

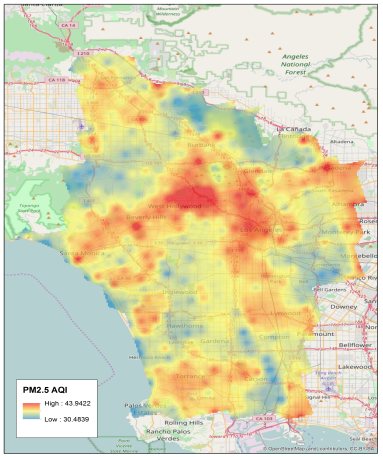
IDW



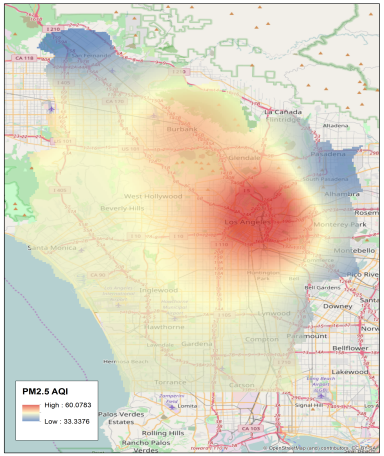
Dec 2016



Dec 2016



Jan 2017



Jan 2017

Outline

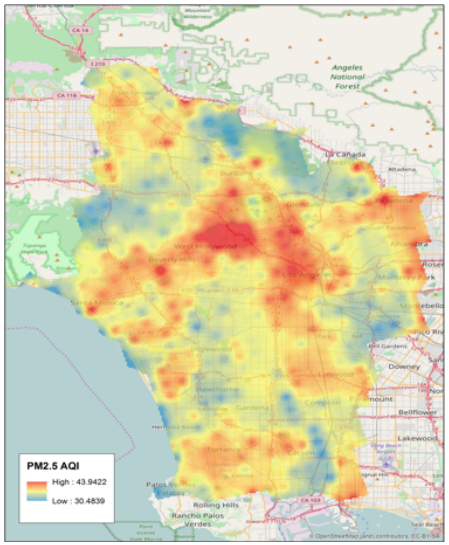
- Introduction and Data Sources
- Approach and Algorithm
- Experiments and Results
- Related Work
- Conclusion and Future work

Related Work

	Limitations	Advantages of our method
Spatial interpolation methods, e.g., IDW and Kriging	No consideration on neighborhood characteristic	Consider neighboring geographic features
	Cannot generate a fine scale result with sparse monitoring stations	Can generate accurate result in a fine scale
Dispersion models	Require detailed data (e.g., building heights and distance between neighboring buildings)	Use easily accessible datasets (OpenStreetMap)
Land-use regression (LUR) methods (e.g., Hoek (2008))	Rely on expert-selected predictors, including types and spatial radii	Expert-free feature selection

Summary

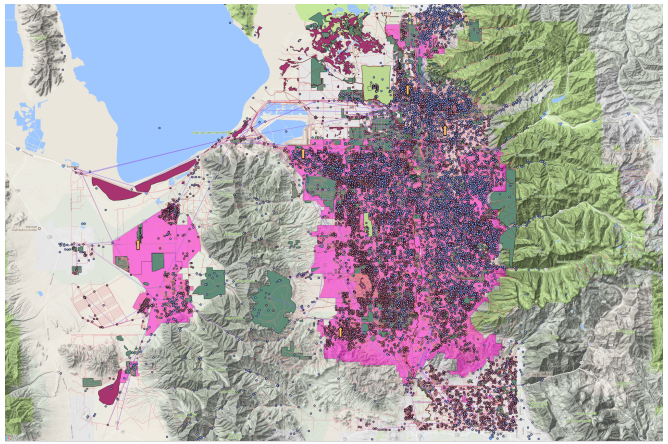
- We presented a data mining approach to build an accurate model to predict $PM_{2.5}$ concentrations at a fine scale by automatically selecting important geographic features without using expert knowledge.



Future Work

- Test our approach for more cities
  - Salt Lake City
  - Morocco
- Incorporate other time-series data, such as weather information, to improve the air quality model
  - Dark Sky

OpenStreetMap of Salt Lake City



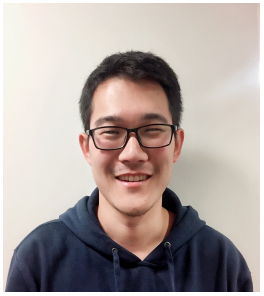

```
"latitude": 42.3601,
"longitude": -71.0589,
"timezone": "America/New_York",
"hourly": {
  "summary": "Snow (6–9 in.) and windy starting in the afternoon.",
  "icon": "snow",
  "data": [
    {
      "time": 255589200,
      "summary": "Mostly Cloudy",
      "icon": "partly-cloudy-night",
      "precipIntensity": 0,
      "precipProbability": 0,
      "temperature": 22.8,
      "apparentTemperature": 16.46,
      "dewPoint": 15.51,
      "humidity": 0.73,
      "pressure": 1026.78,
      "windSpeed": 4.83,
      "windBearing": 354,
      "cloudCover": 0.78,
      "uvIndex": 0,
      "visibility": 9.62
    },
```


Request from Dark Sky

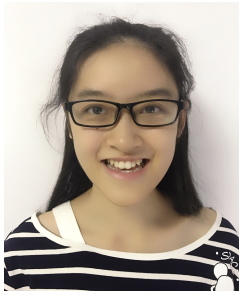
Our Team

Yijun Lin
Research Staff
USC Spatial Sciences Institute

Yao-Yi Chiang
Associate Professor (Research)
USC Spatial Sciences Institute



Yuanbin Cheng
Graduate Student
USC Computer Science



Xin Yu
Graduate Student
USC Spatial Informatics

Mengyue Huan
Graduate Student
USC Spatial Informatics

spatial computing INNOVATION
COMPUTER SCIENCE × SPATIAL SCIENCE

USC Dornsife
Dana and David Dornsife
College of Letters, Arts and Sciences

Spatial Sciences Institute

THANK YOU

Yijun Lin yijunlin@usc.edu
Applying for Phd Program this year

Spatial Sciences Institute
University of Southern California