Spatial Data Analytics with Classical Data Mining and Machine Learning Algorithms

Yao-Yi Chiang

Computer Science and Engineering

University of Minnesota

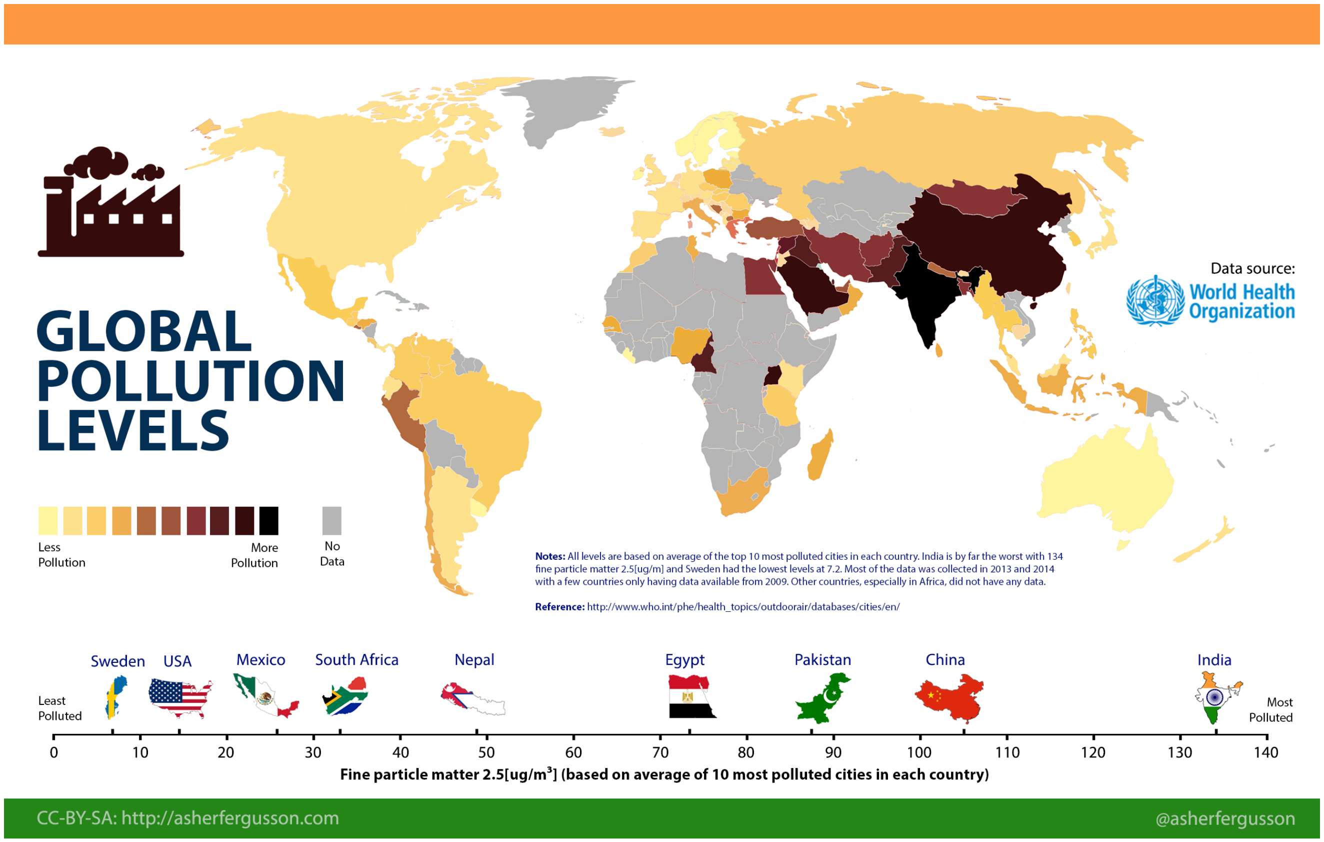
yaoyi@umn.edu

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

A motivating example

Lin Y., Pan F., Chiang Y.-Y., Stripelis D., Ambite J. L., Eckel S. P., and Habre R. (November 2017) Mining public datasets for modeling intra-city PM2.5 concentrations at a fine spatial resolution. ACM SIGSPATIAL.

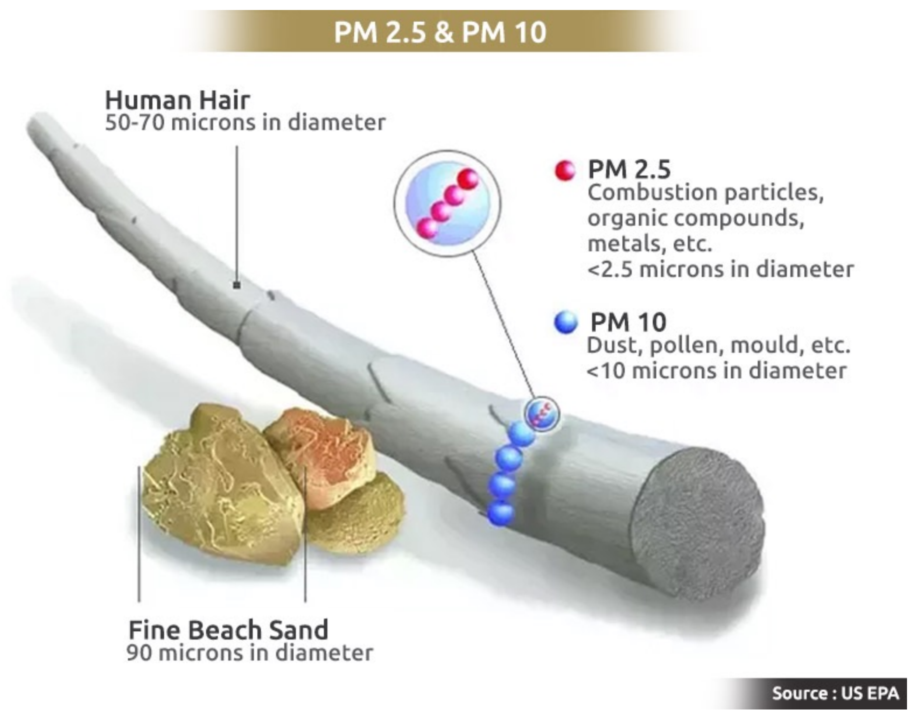
Air Pollution is a Global Problem



Air Pollutant: $PM_{2.5}$ and PM_{10}

$PM_{2.5}$: fine inhalable particles, with diameters that are generally 2.5 micrometers and smaller

United States Environmental Protection Agency



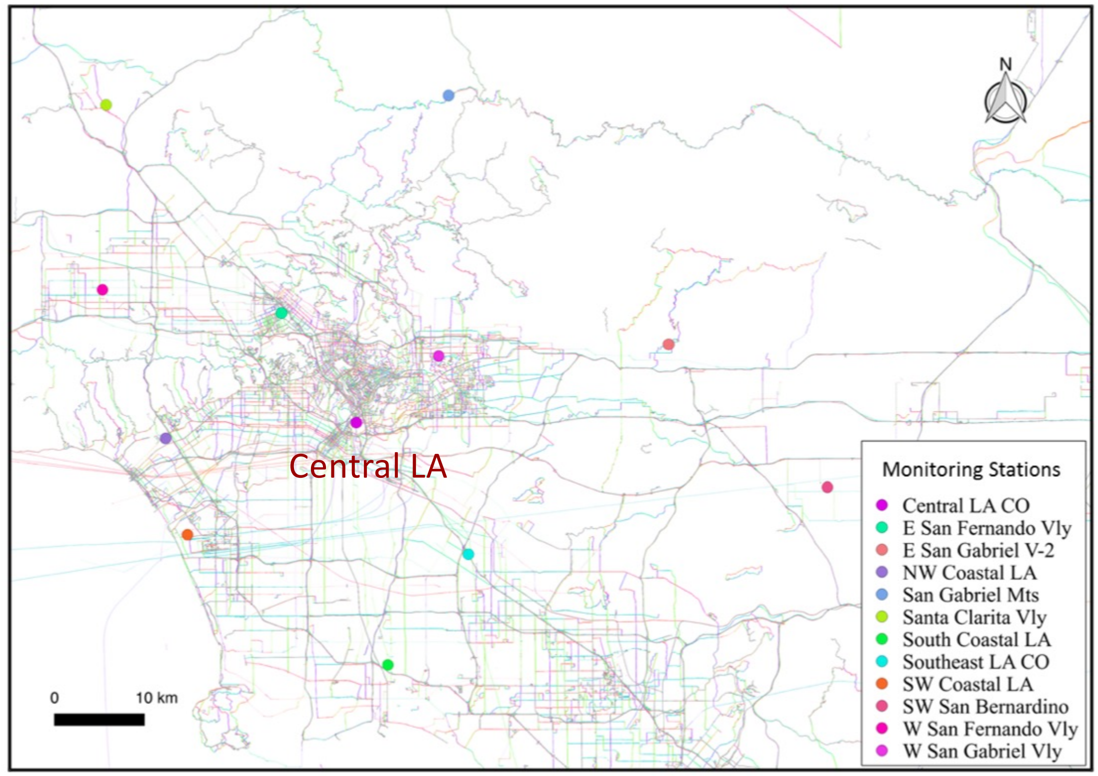
Air Quality Index

AQI: air quality index computed from a piecewise linear function of the pollutant concentration (e.g., 12.0 micrograms per cubic meter is 50 AQI for PM2.5).

Air Quality Index Levels of Health Concern	Numerical Value	Meaning
Good	0 to 50	Air quality is considered satisfactory, and air pollution poses little or no risk.
Moderate	51 to 100	Air quality is acceptable; however, for some pollutants there may be a moderate health concern for a very small number of people who are unusually sensitive to air pollution.
Unhealthy for Sensitive Groups	101 to 150	Members of sensitive groups may experience health effects. The general public is not likely to be affected.
Unhealthy	151 to 200	Everyone may begin to experience health effects; members of sensitive groups may experience more serious health effects.
Very Unhealthy	201 to 300	Health alert: everyone may experience more serious health effects.
Hazardous	301 to 500	Health warnings of emergency conditions. The entire population is more likely to be affected.

Limited Air Quality Observations

- Monitoring stations are usually sparse – 12 stations for $PM_{2.5}$ in Los Angeles



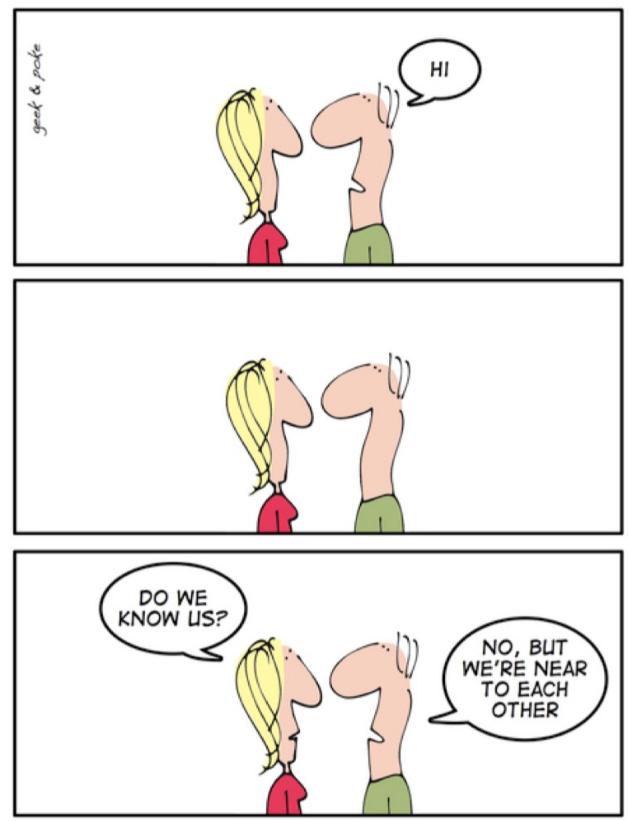
Nearby locations would have similar air quality

- Tobler's 1st law of geography:

"all things are related, but nearby things are more related than distant things"

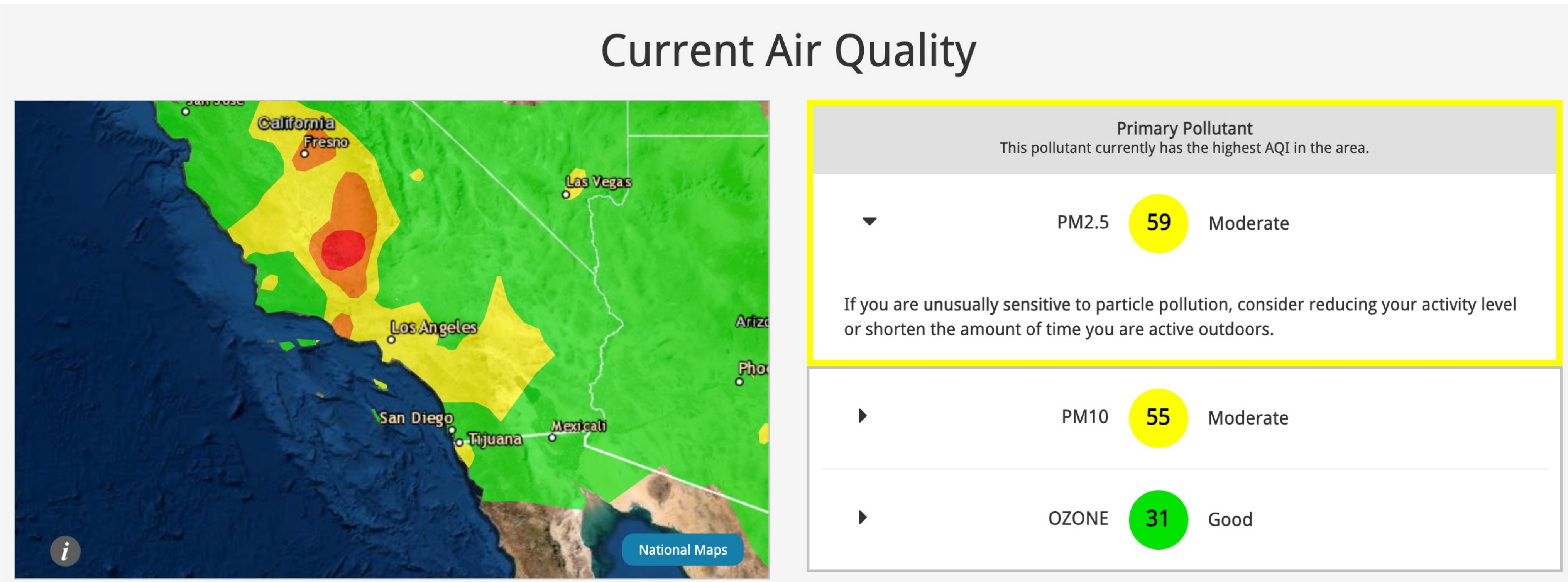
Tobler – 1970

- Spatial interpolation? IDW?



Vandecasteele & Devillers, Improving volunteered geographic data quality using semantic similarity measurements

Typical Air Quality Prediction Result (AirNow)

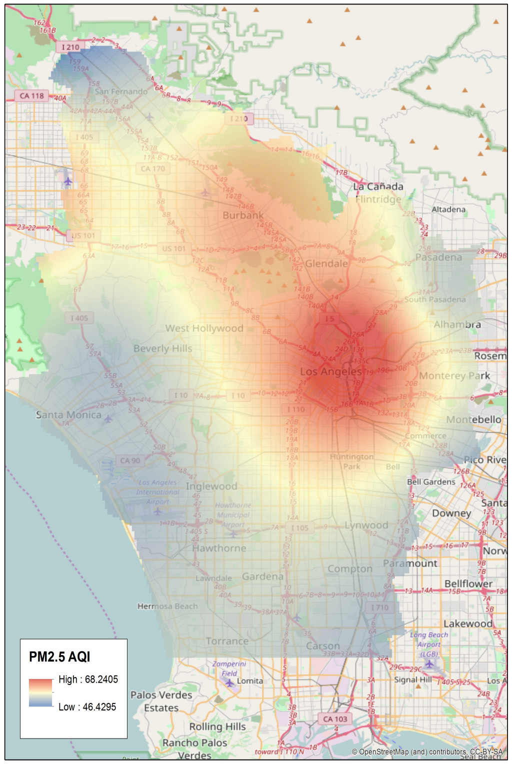


https://www.airnow.gov/?reportingArea=Central%20LA%20CO&stateCode=CA

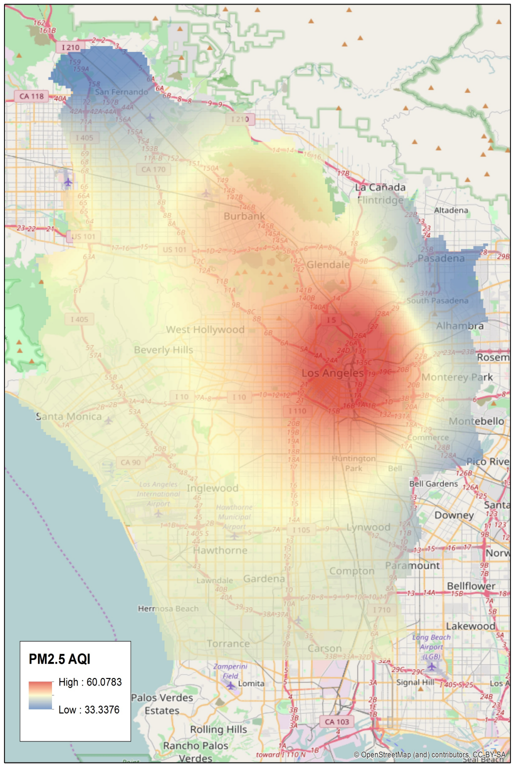
Inverse Distance Weighting

- Why are the results so smooth over space?
- Recall IDW:

$$u(\mathbf{x}) = \begin{cases} \dfrac{\sum_{i=1}^{N} w_i(\mathbf{x}) u_i}{\sum_{i=1}^{N} w_i(\mathbf{x})} \\ u_i, \end{cases}$$



Dec 2016



Jan 2017

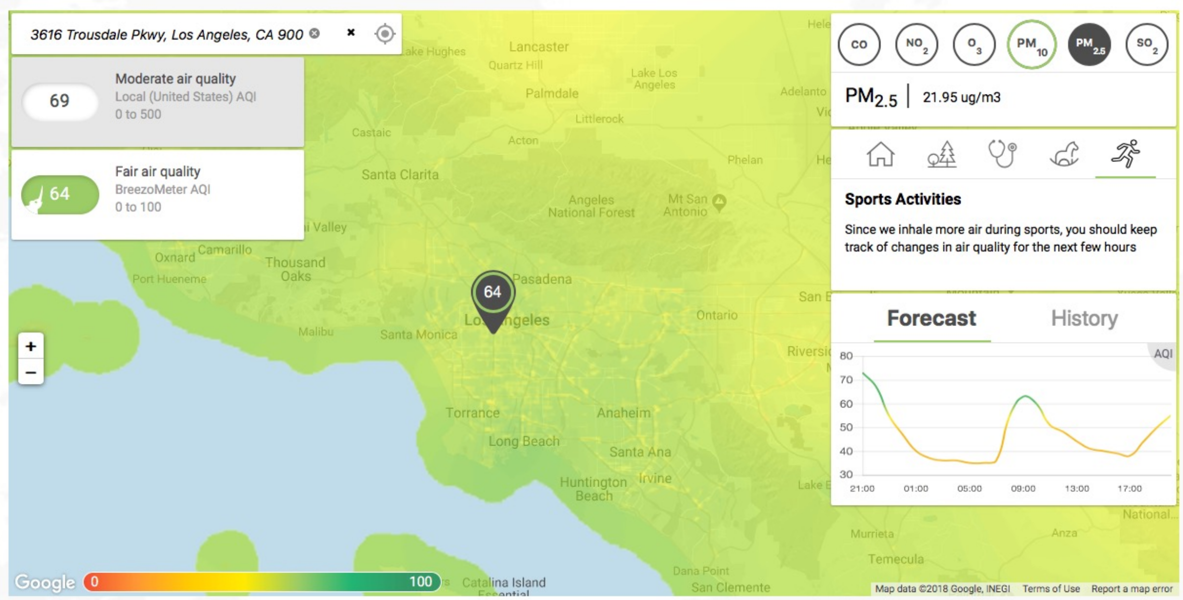
Machine Learning Methods

- Some prediction variations in space, e.g., the road network is obvious

BreezoMeter
Because Breathing Matters

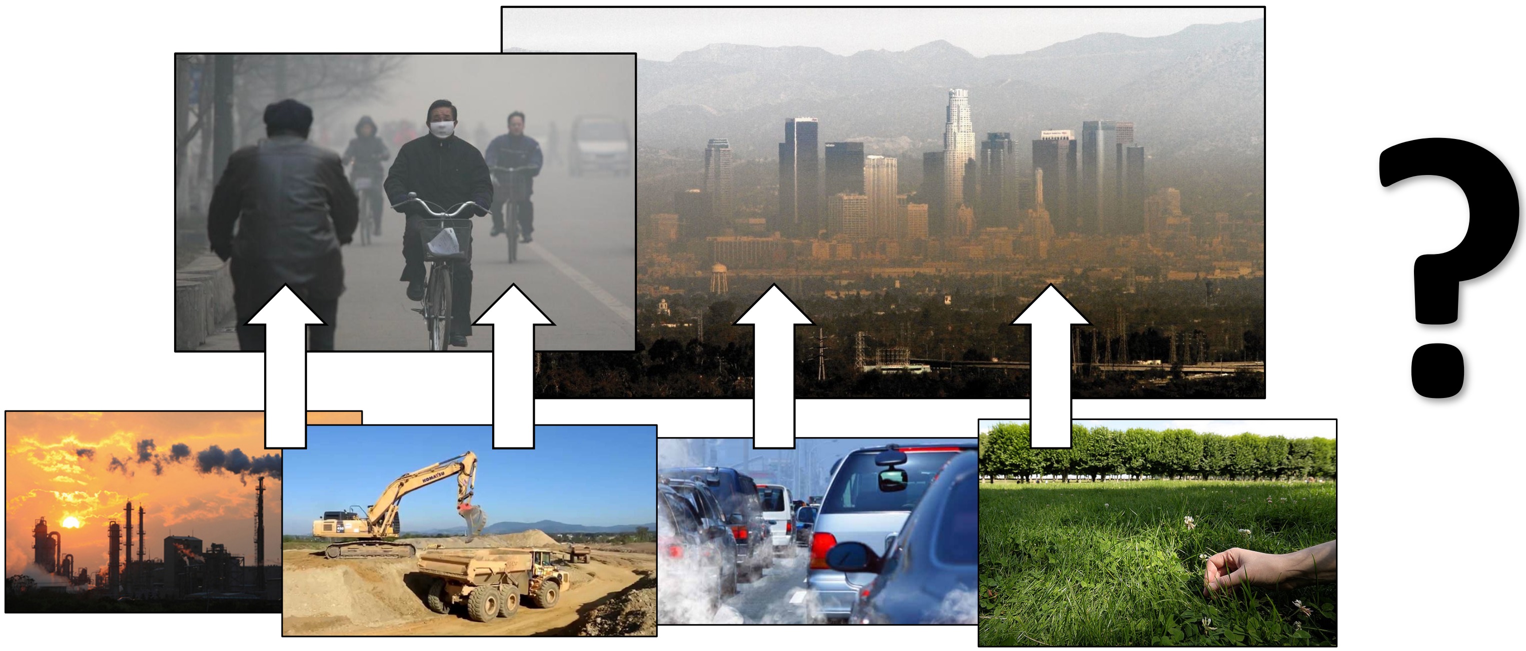
Los Angeles, United States
Air Quality

"Powered by proprietary algorithms, strict QA & machine learning."



Existing Work for Air Quality Modeling

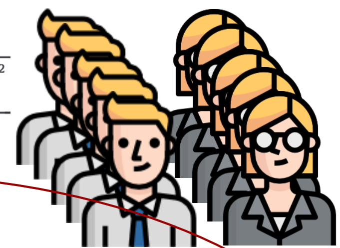
The built environment has a strong impact on air quality but *how*?



Land-use Regression Models (LUR)

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~~) Taiwan	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

Source: Liu et al., 2016



Expert-selected & Area-specific

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

LUR Limitations

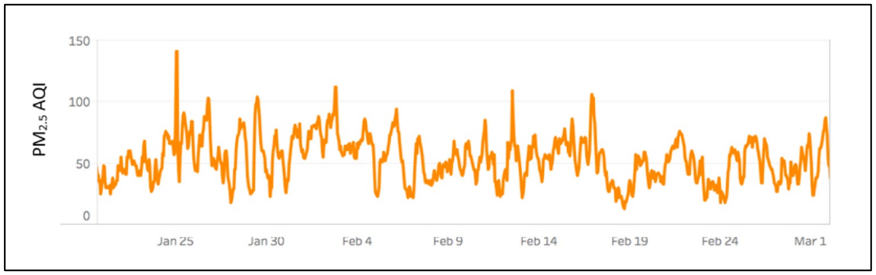
- Experts are expensive
- Do not scale well for predictions at various spatial and temporal resolutions
- Sometimes rely heavily on datasets that are not easy to obtain
  - e.g., traffic

Can we do better?

Source: Liu et al., 2016

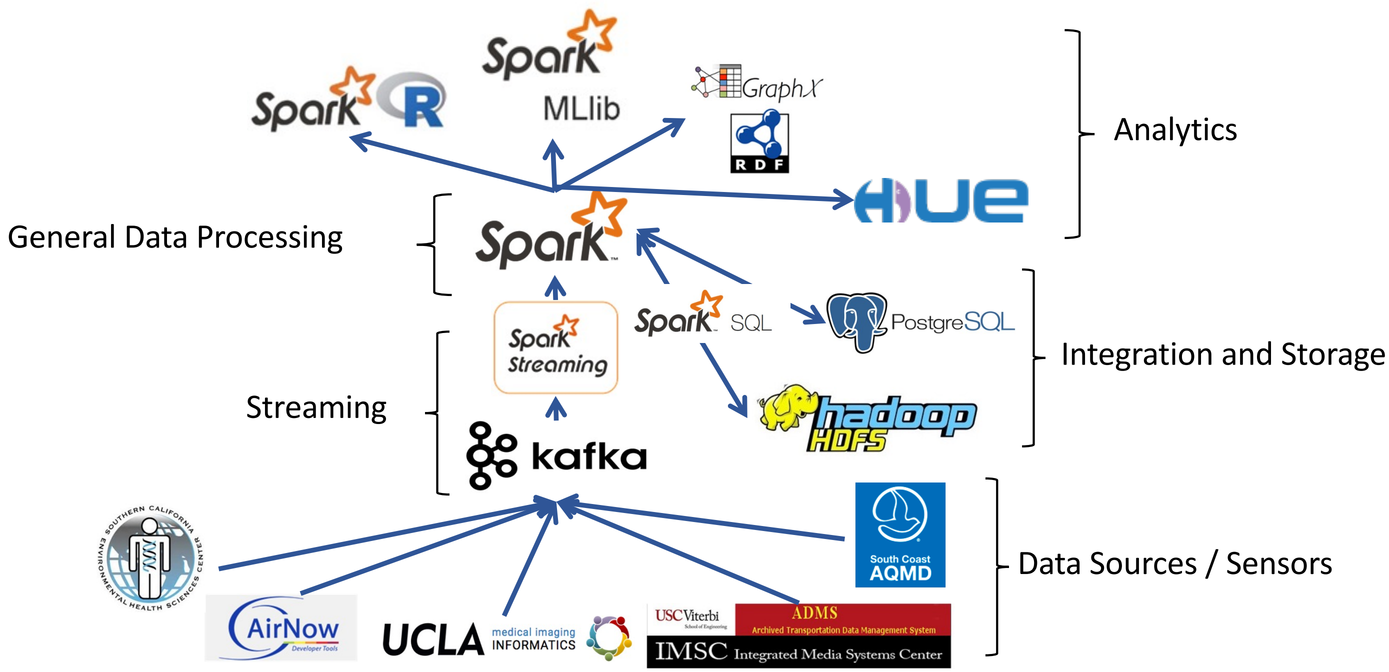
JonSnow: Data-Driven Air Quality Prediction at Fine-Spatial Scale

- Problem
  - Given some sensors and their locations, predicting air quality for locations that do not have a sensor
- Hypothesis
  - Similar environments should have a similar air quality



Data Collection

PRISMS-DSCIC – A scalable data integration and analysis architecture



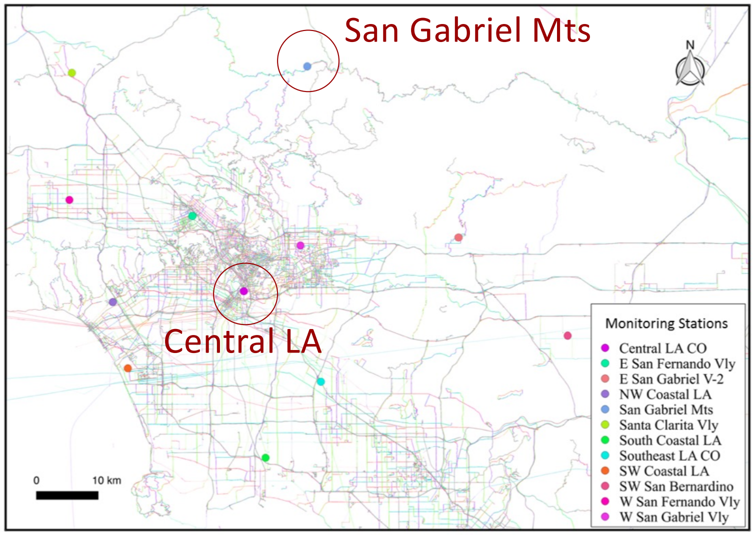
Data Sources – I

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



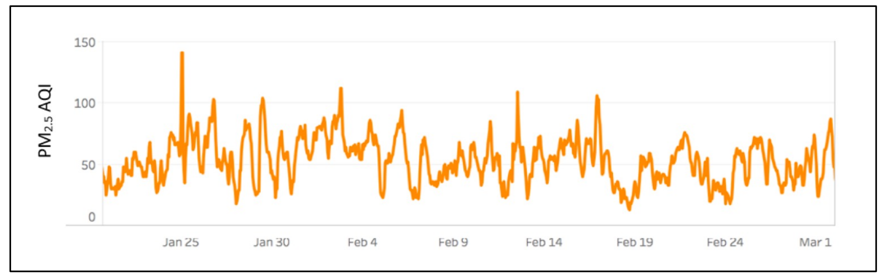
Data Sources – II

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 subtypes
  - e.g., subtypes for “Buildings”: commercial, apartment, house, industrial, school, etc.

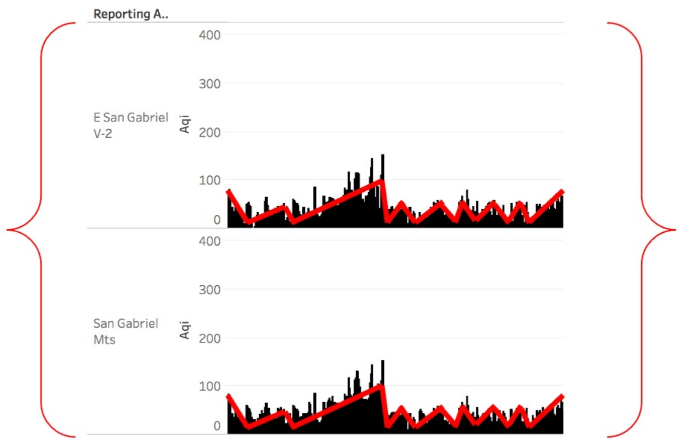
Recall: JonSnow: Data-Driven Air Quality Prediction at Fine-Spatial Scale

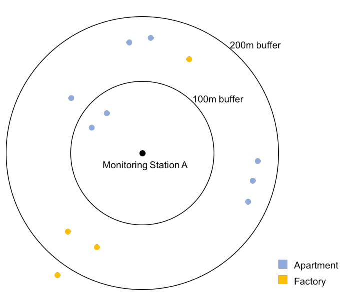
- Problem
  - Given some sensors and their locations, predicting air quality for locations that do not have a sensor
- Hypothesis
  - Similar environments should have a similar air quality



Approach Overview

- Similar environments should have a similar air quality
  - How to quantify "similar air quality"
    - Clustering of air quality measurements
    - K-Means, hierarchical clustering, dimension reduction
  - How to quantify "similar environments"
    - Train an interpretable machine learning model using geographical context to predict whether two locations would have "similar air quality"
    - Random Forest





	Pedestrian 100m	Motorway 100m	...	Apartment 200m	Factory 200m
Monitoring Station 0	0.1	5.28		2	0.3

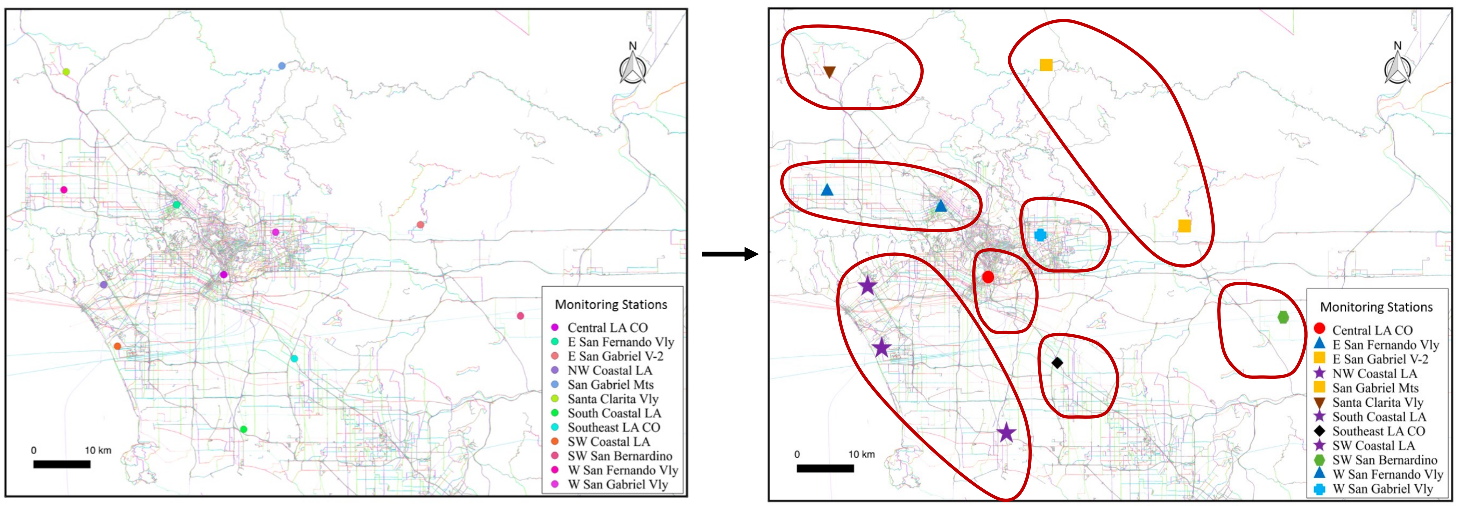
	Pedestrian 100m	Motorway 100m	...	Apartment 200m	Factory 200m
Monitoring Station 1	0.0	3.27		0.041	0.432

Required Technologies

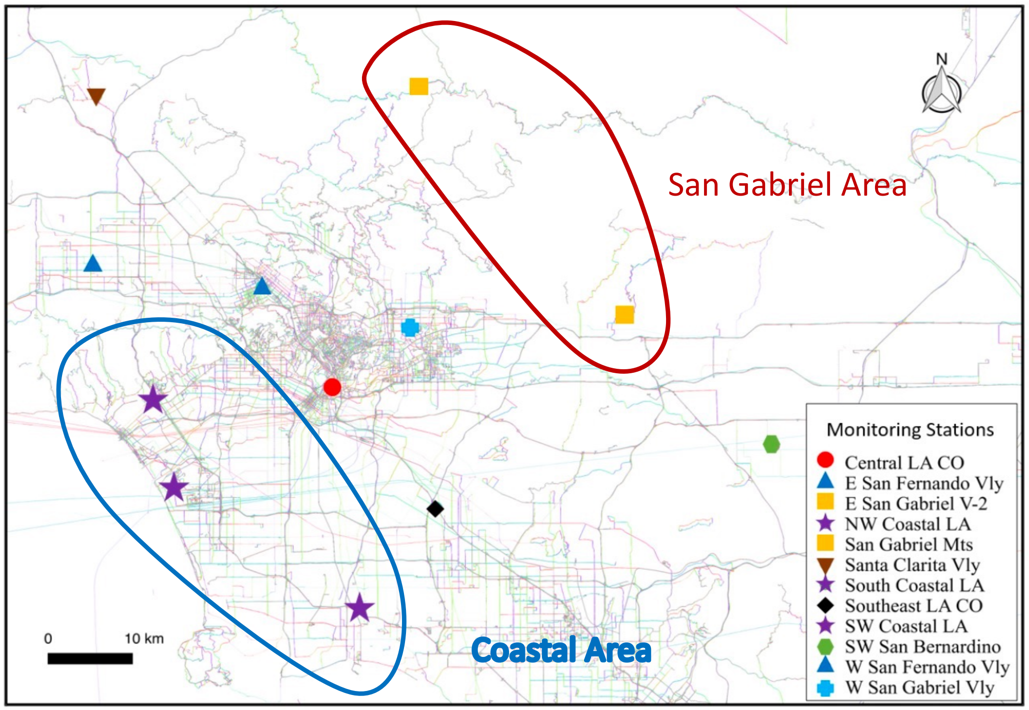
- Clustering
  - K-Means, Hierarchical Clustering
- Dimension Reduction
  - SVD (Singular Value Decomposition)
- Interpretable Machine Learning Method
  - Random Forest

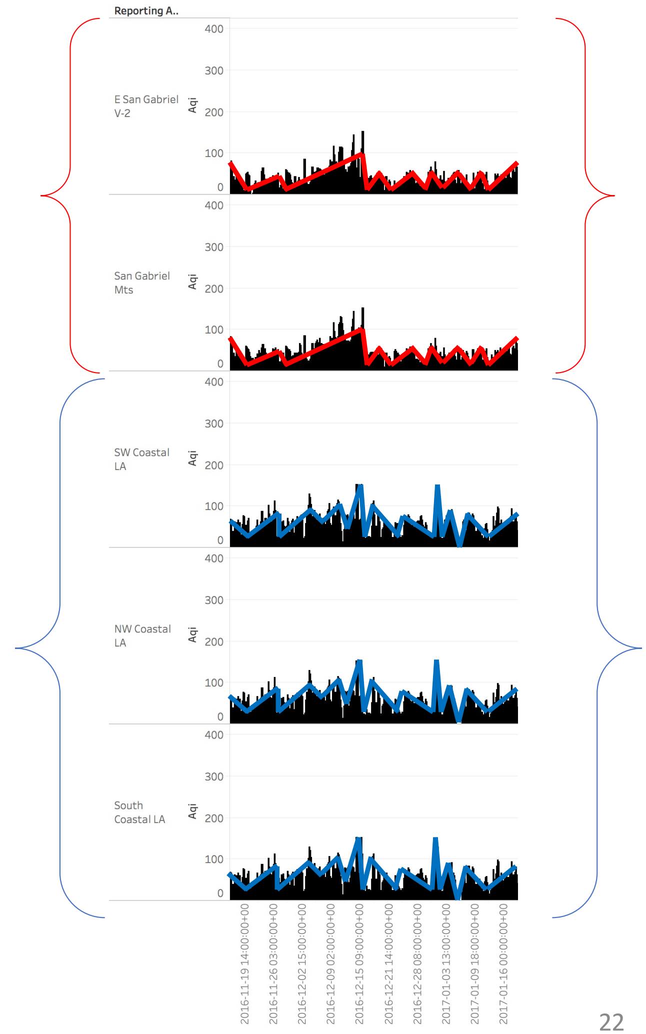
Step 1. Grouping Stations based on their $PM_{2.5}$ AQIs

- To identify the monitoring stations that have similar temporal pattern on $PM_{2.5}$ AQIs
- These monitoring stations should have a similar environment.



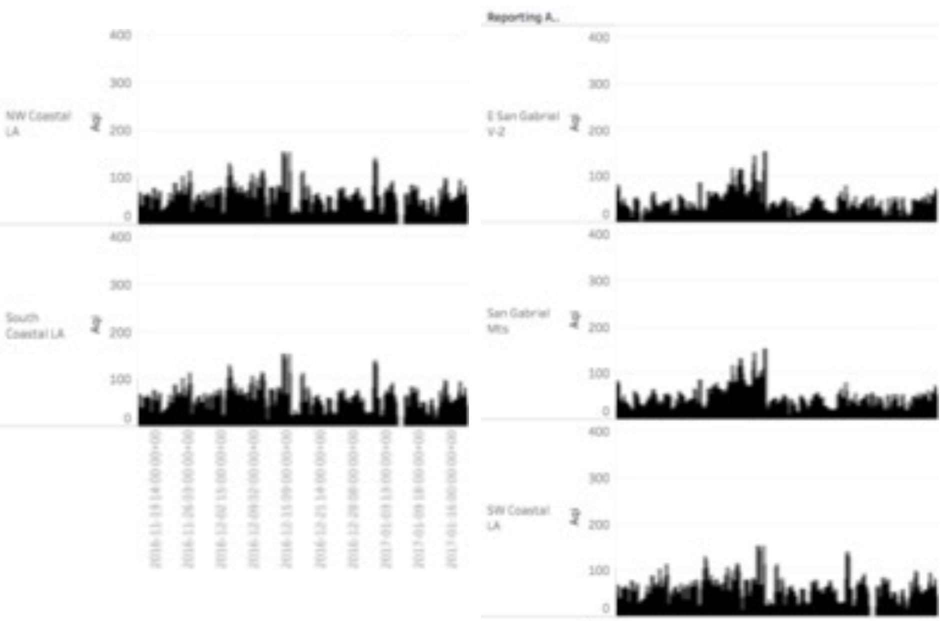
Similar Temporal Pattern



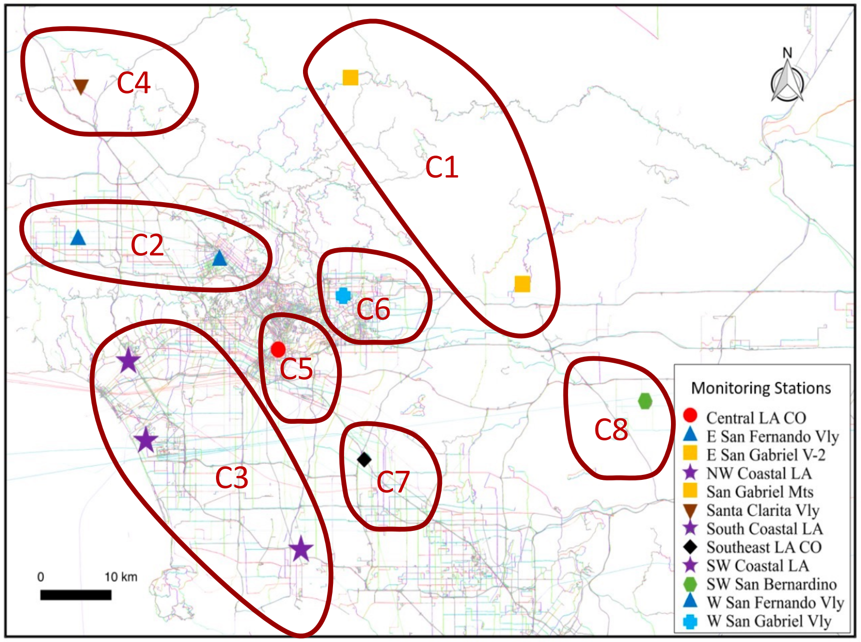


K-means Clustering

- Input: time-series observations at each station



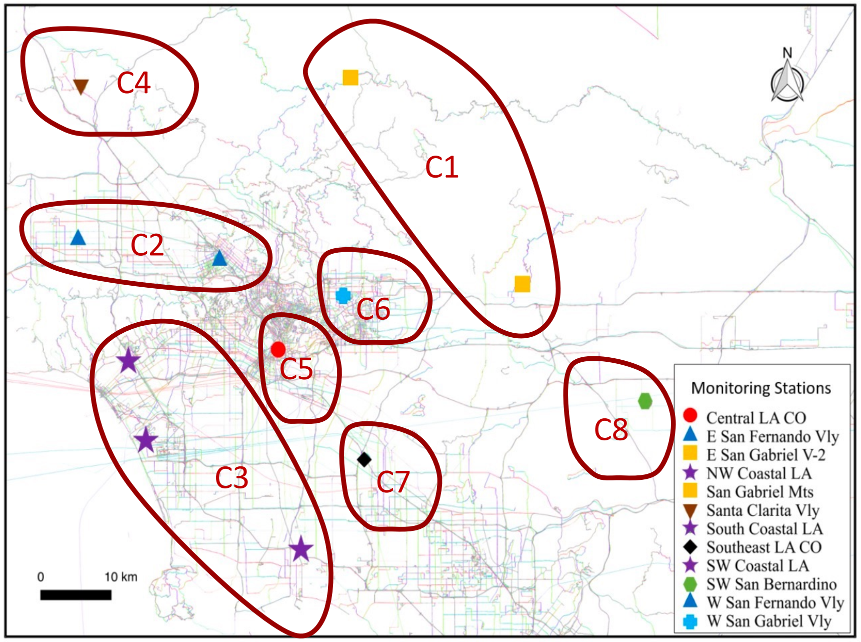
- Output: clusters of stations having a similar temporal pattern



K-means Clustering

- Recall: Hypothesis
  - Similar environments should have a similar air quality
- Stations in the same cluster have a similar temporal pattern
- How to quantify "similar environment"
  - what specific geographic feature types (e.g., primary roads, industrial areas, parks)
  - from what distance have the most impact on the clustering result?

- Output: clusters of stations having a similar temporal pattern



Step 2. Generating Geographic Abstraction

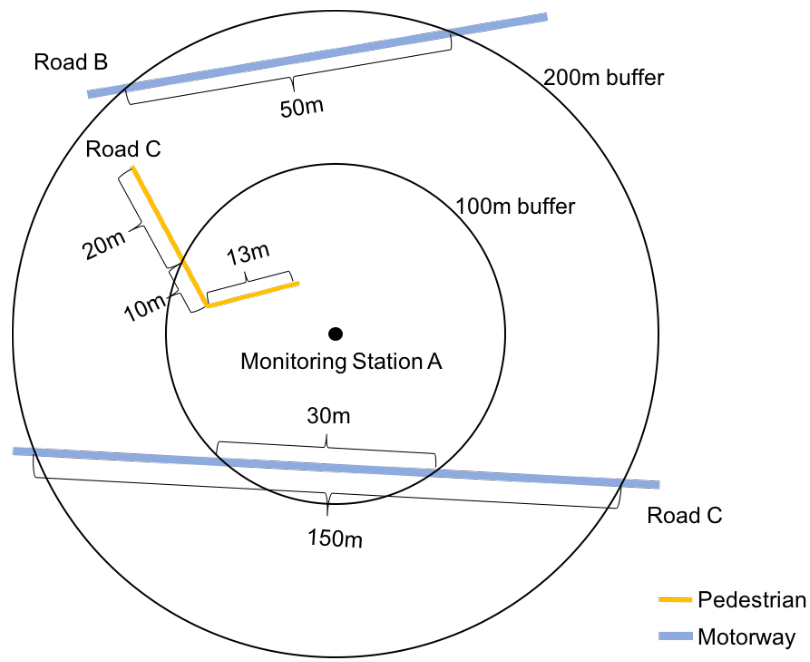
Length of line features

- e.g., Roads, Aeroways

Example Roads

	100m	200m
Pedestrian	23	43
Motorway	30	200

[23, 30, 43, 200]



Step 2. Generating Geographic Abstraction

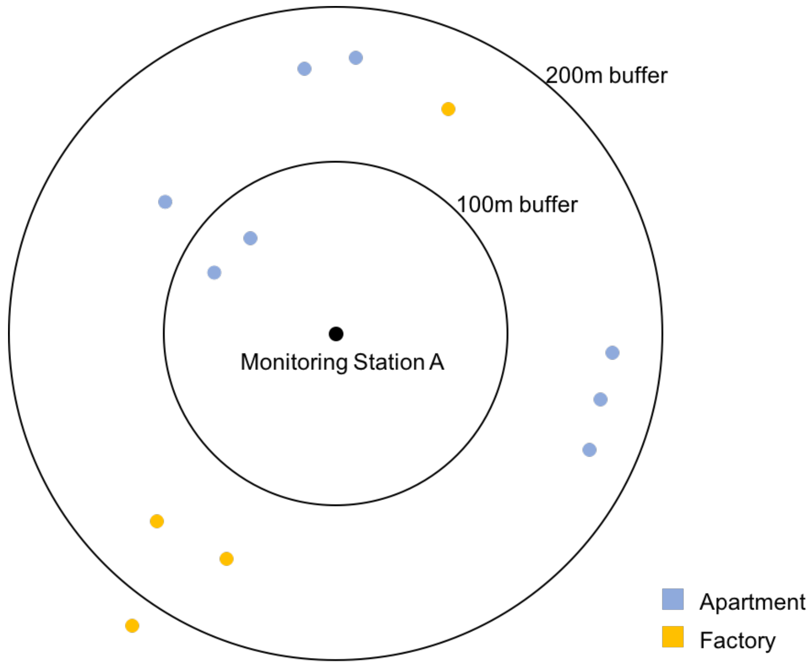
Count of point features

- e.g., Buildings

Example Buildings

	100m	200m
Apartment	2	8
Factory	0	3

[2, 0, 8, 3]



Step 2. Generating Geographic Abstraction

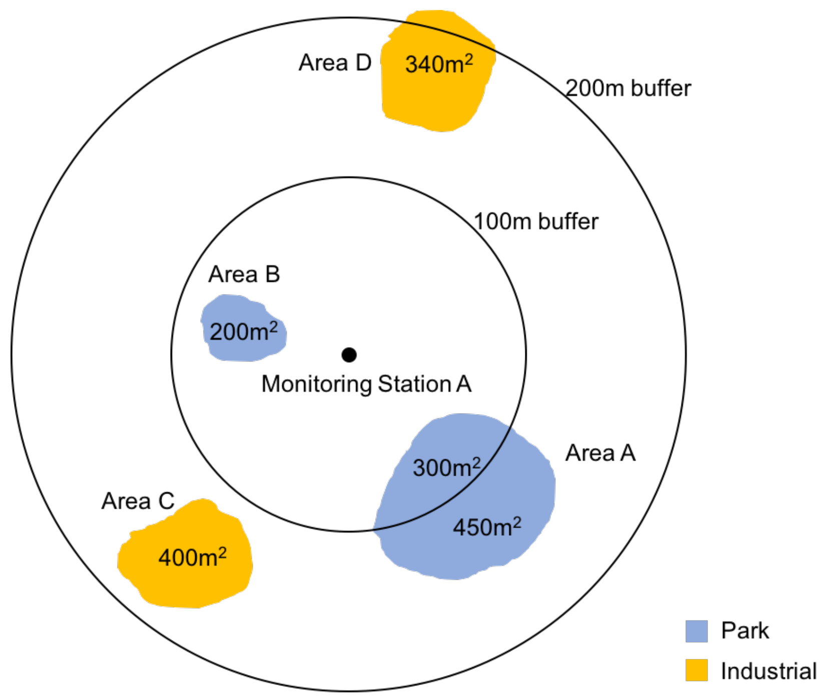
Area of polygon features

- e.g., Land uses, Water areas

Example Land uses

	100m	200m
Park	500	950
Industrial	0	740

[500, 0, 950, 740]



Step 2. Generating Geographic Abstraction

- Generating a large vector for each monitoring station

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

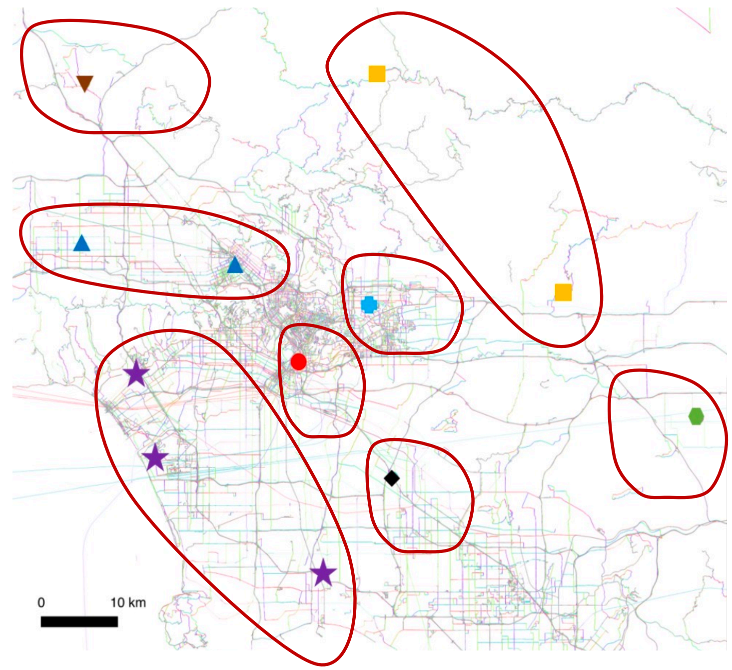
Park 200m	Industrial 200m	Apartment 100m	Factory 100m	Apartment 200m	Factory 200m	Distance to Ocean
950	740	2	0	8	3	4000

- In practice, we creates buffers from 100 meters to 3,000 meters with an interval of 100 meters
  - 3,500+ components in a vector

How to quantify “similar environment”

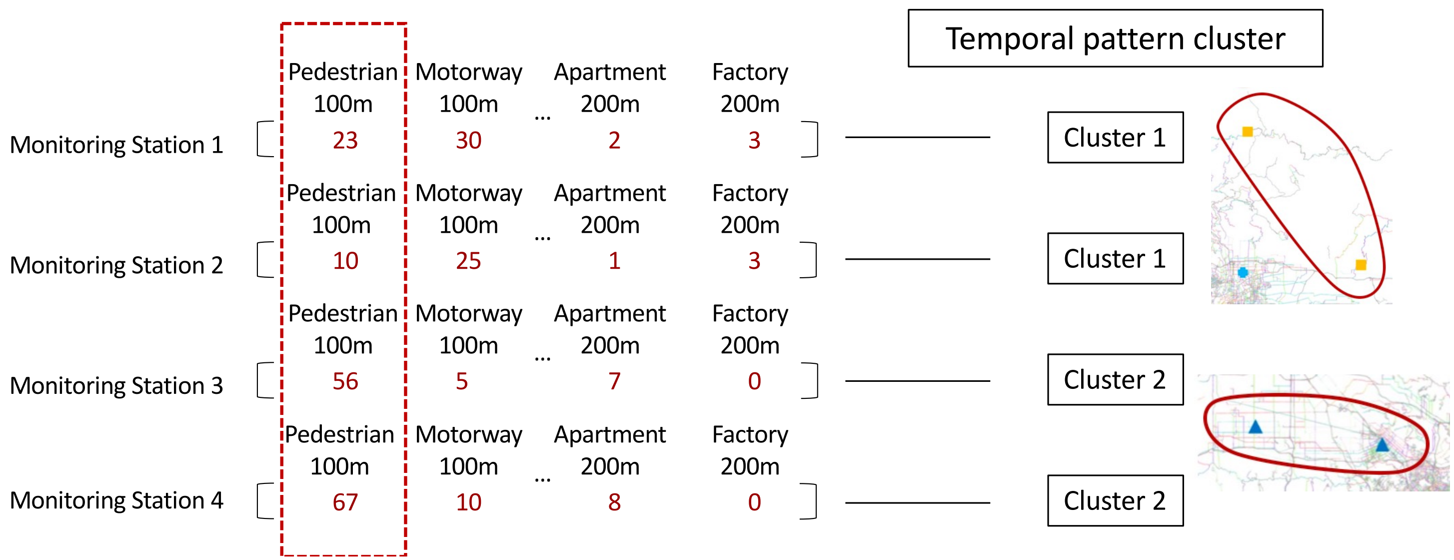
Geographic abstraction → ? → Clustering result

	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. Computing Feature Importance

- Training **a random forest model** to
  - predict cluster label using the geographic context
    - each feature component represents **a geographic feature type** within **certain distance**
  - quantify the **impact** of **each feature component**



Step 3. Generating Geo-context

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

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

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

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

Monitoring Station 1 (Geographic Abstraction)

Pedestrian 100m	Motorway 100m	…	Apartment 200m	Factory 200m
23	30		2	3

Monitoring Station 1 (Geo-context)

Pedestrian 100m	Motorway 100m	…	Apartment 200m	Factory 200m
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 3. Geo-context

- Geo-context is an updated vector from geo-abstract for describing
  - how **each feature type** within **a certain distance** (a feature component) in Geographic Abstraction affects the Temporal Pattern ($PM_{2.5}$ AQI)
- Reward important (relevant) features and penalize others

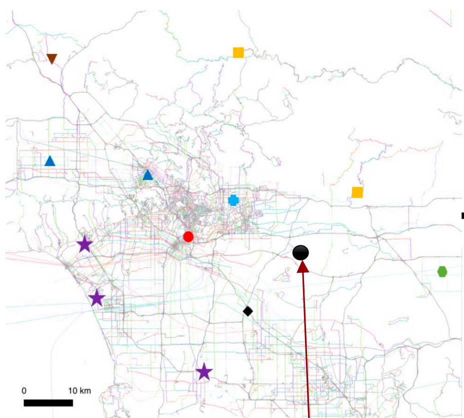
	Pedestrian 100m	Motorway 100m	...	Apartment 200m	Factory 200m
Monitoring Station 1 (Geographic Abstraction)	23	30		2	3

	Pedestrian 100m	Motorway 100m	...	Apartment 200m	Factory 200m
Monitoring Station 1 (Geo-context)	0.0	3.27		0.041	0.432

Step 4. Predicting $PM_{2.5}$ AQI

Train a regression model to predict $PM_{2.5}$ AQI for a target location at time T

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



Regression Model

AQI prediction of target location

[Geo-context] for target location

Experiments

Leave-one-out cross-validation method

- Predict $PM_{2.5}$ AQI for the removed station by using other 11 stations
- Compare our approach with baseline methods

Predicting at a fine scale

- Predict $PM_{2.5}$ AQI of each point on an 1-mile-apart fishnet covering most of the Los Angeles area (604 points)
- Visualize the fine-scale prediction results

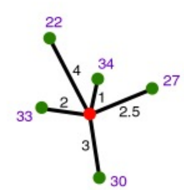
Experiment & Result – I

Leave-one-out cross-validation method

- Tested 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)
  - **RMSE** - root-mean-square error; **MAE** - mean absolute error

	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



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

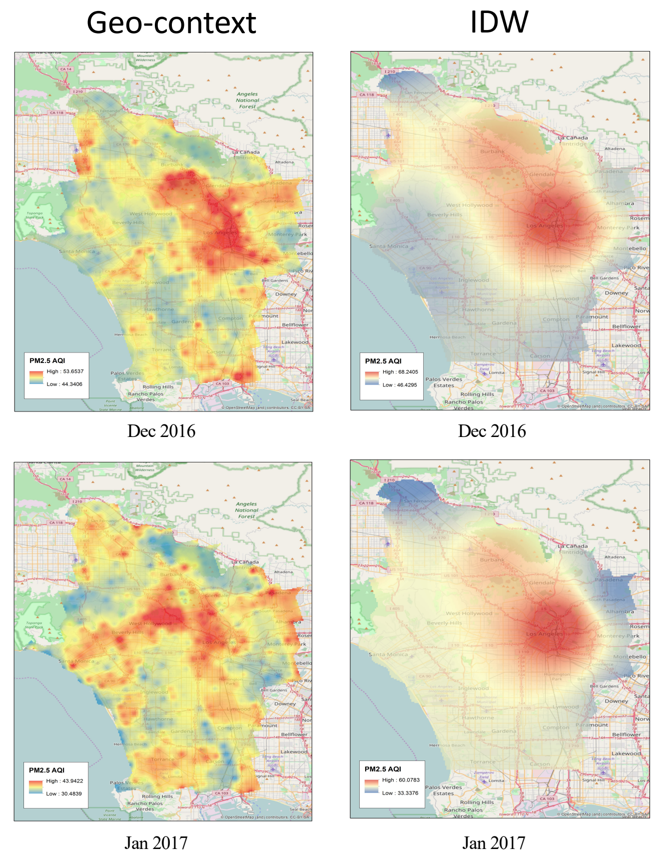
All within 10% error margin; Significant different with 95% confidence (paired t-test)

Experiment & Result – I (Cont'd)

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

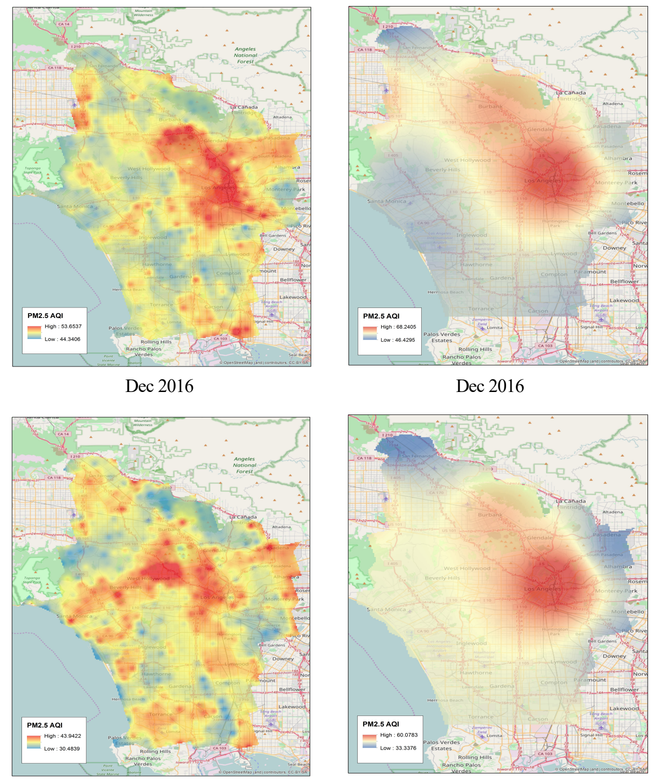


Experiment & Result – II

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

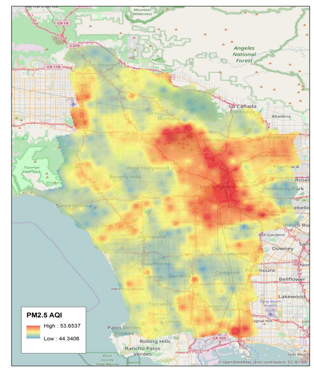
Experiment & Result – II

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

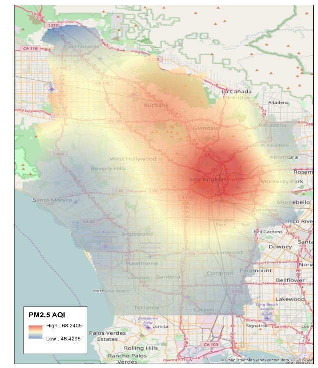
Geo Name	Buffer Size (meter)	Geo type	Importance (%)
land use	1100	wetland	0.0051177
land use	1300	university	0.004450
road	600	rail	0.0044327
land use	1200	village_green	0.0037241
road	700	primary	0.0035520
land use	1900	farmland	0.0031458
land use	2700	village_green	0.0030063
road	800	residential	0.0028980
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

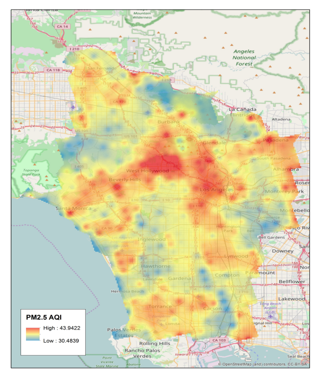
IDW



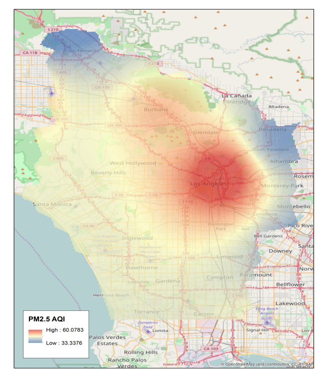
Dec 2016



Dec 2016



Jan 2017



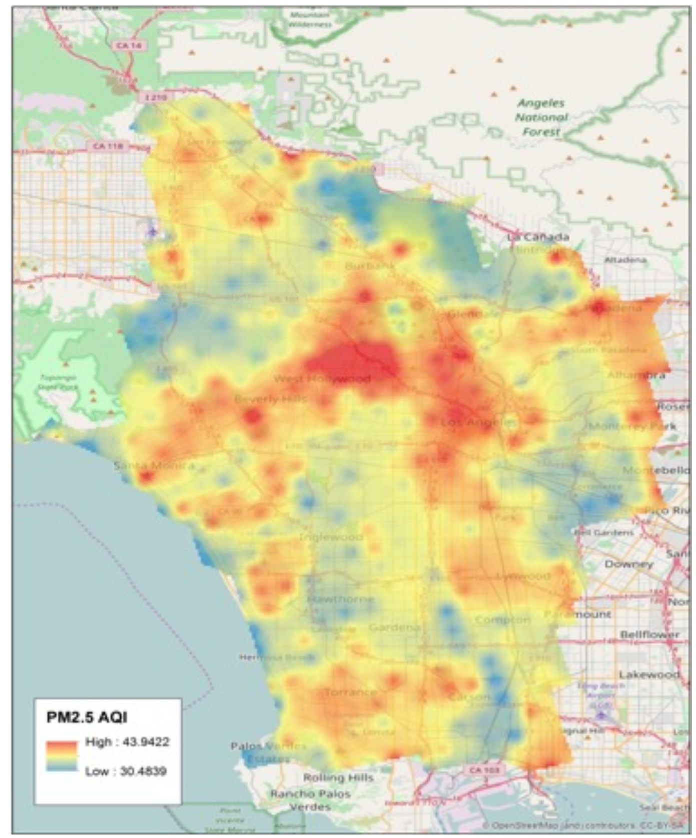
Jan 2017

Related Work

	Limitations	**Advantages of our method**
Spatial interpolation methods, e.g., IDW and Kriging	Not considering neighborhood characteristic	With 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

- A spatial data mining approach to build an accurate model to predict $PM_{2.5}$ concentrations at a fine scale by
- Automated selection of important geographic features without using expert knowledge.



Additionally, Air Quality Forecasting

Goal

Build a general approach for **location-dependent time-series data** forecasting

Challenges:

Existing approaches do not handle spatial correlation well

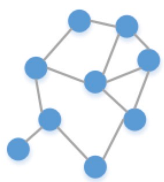
e.g., Auto-Regression Integrated Moving Average (ARIMA), Kalman filtering, Artificial Neural Network (ANN)

Our approach

- We are building a Diffusion Convolutional Recurrent Neural Network for forecasting location-dependent time series data.
- Continuously forecasting air quality index (AQI) in next 24 hours at a fine scale using data on the PRISMS-DSCIC

Lin, Y., Mago, N., Gao, Y., Li, Y., Chiang, Y.-Y., Shahabi, C., and Ambite, J. L. (November 2018). Exploiting Spatiotemporal Patterns for Accurate Air Quality Forecasting using Deep Learning. In *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 359 – 368, Seattle, WA, USA

DCRNN – Diffusion Convolutional Recurrent Neural Network



Graph Construction

- Each point in the graph represents the time series at the station
- The link between points would be the proximity between stations (e.g., distance, geographic similarity)

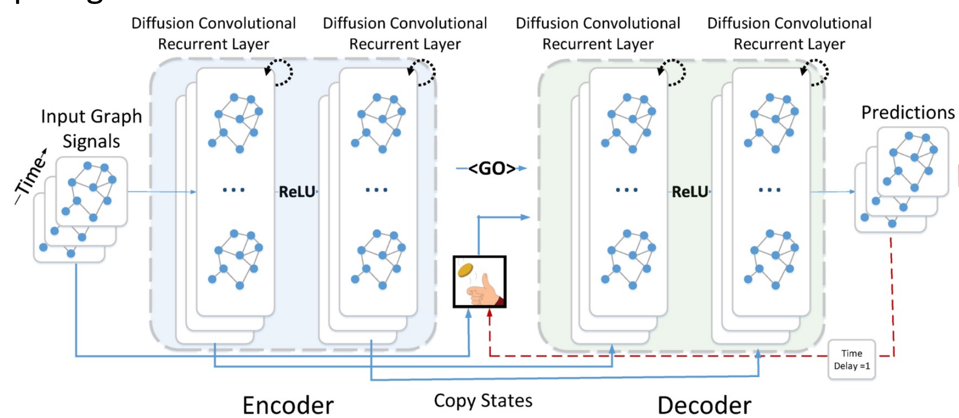
Spatial Dependency Modeling

- Use diffusion convolution to learn a function that maps historical graph signal to future graph signal

Temporal Dependency Modeling

- Use Recurrent Neural Networks

Air quality time series



Forecasting result

Acknowledgements

- Gil, Yolanda (Ed.) Introduction to Computational Thinking and Data Science. Available from http://www.datascience4all.org