Air Pollution Mapping and Prediction

Gaurav Mahamuni[1], Mingyu Wang[1] and Su Ye[2]

[1]Mechanical Engineering, University of Washington, [2]Computer Science and Engineering, University of Washington

CSE 547: Machine Learning for Big Data

ABSTRACT

- Particulate Matter (PM) analysis is important in assessing an individual's exposure to potentially harmful particles.
- Currently, PM is recorded at sparse locations in a geographical area, however, the PM level can vary dramatically over small distances.
- We map and predict PM levels at specific locations in the city of Krakow in Poland from spatio-temporal data of PM levels and meteorological data.

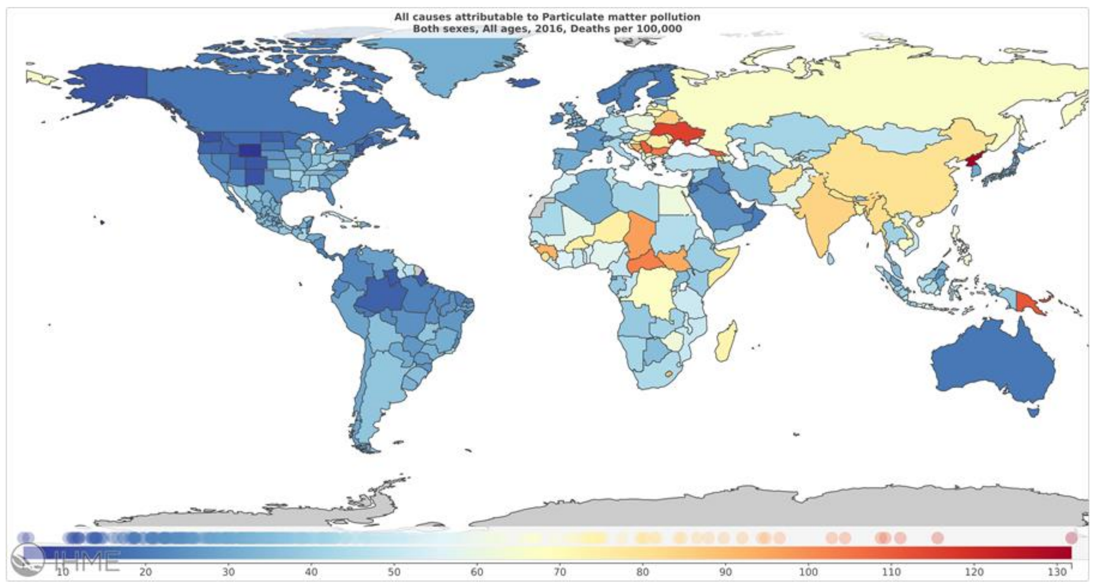


Figure 1. In the year 2016, ambient air pollution was responsible for 4.2 million deaths

DATA

We have two kinds of data in the dataset for each sensor:
1) Meteorological data: temperature, humidity and barometric pressure.
2) Air quality data: PM2.5, PM10 and PM1.

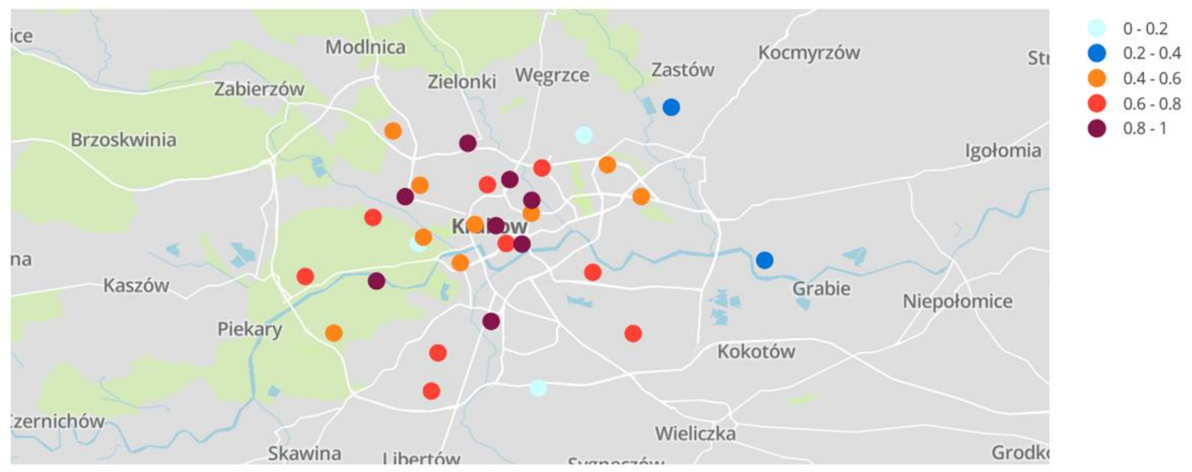


Figure 2. Overall distribution of sensors and average normalized pollution at sensor locations for 10 months in 2017

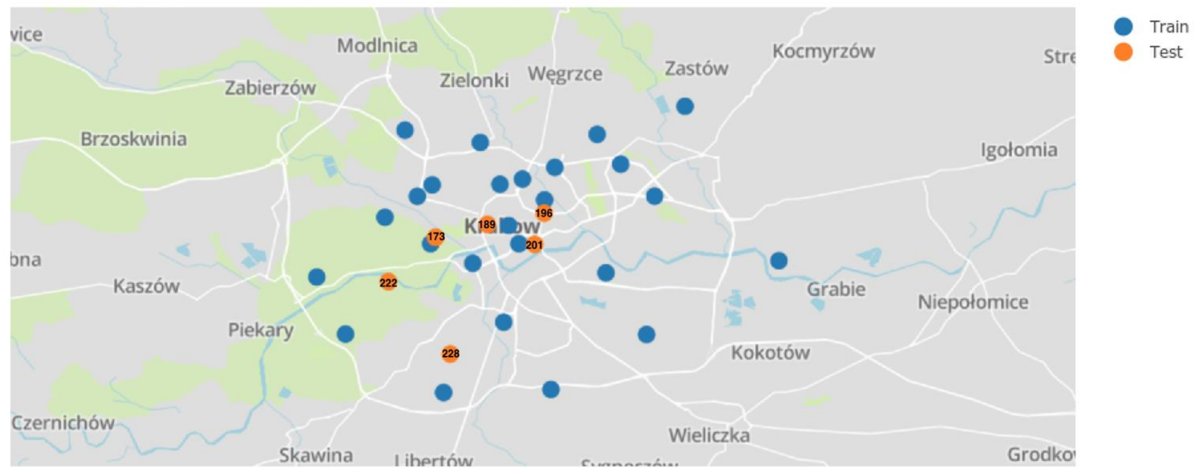


Figure 3. The relative position for test data with respective to all other training sensors

MODELS / ALGORITHMS

Bellkor recommendation system

$$E[R,P,Q] = \sum_{(i,l)\in records} (R_{il} - q_i * p_l)^2 + \lambda\left(\sum_{l=1}^{n} \|p_l\|_2^2 + \sum_{i=1}^{m} \|q_i\|_2^2\right) \quad (1)$$

Algorithm 1: Stochastic Gradient Descent Latent Factor Model

Inputs: Training dataset $D = D_l \cup D_{ul}$, where D_l consists sensors with geographical data and given value for PM2.5, D_{ul} contains sensors with only geographical data.
Initialization: Initialize P, Q matrix with initial value $\sqrt{100/k}$
for *<i = 1...number of iterations>* **do**
  for *each data point* $v_s t$ **do**
    $\epsilon_s t \leftarrow 2(v_{st} - q_t \cdot p_s)$
    $q_t \leftarrow q_t + \mu(\epsilon_{st} p_s - 2\lambda q_t)$
    $p_s \leftarrow p_s + \mu(\epsilon_{st} q_t - 2\lambda p_s)$
  end
end

Semi-supervised Classification using L_1-regularized Logistic Regression

$$\hat{\theta} = \underset{\theta}{argmax}\left(log\left(\prod^{n} p(x_i;\theta)^{y_i}(1 - p(x_i;\theta)^{y_i})\right)\right) \quad (2)$$

$$p(x;b,w) = e^{(b+wx)}/(1 + e^{(b+wx)}) \quad (3)$$

Algorithm 2: Semi-supervised Logistic Regression

Inputs: Training dataset $D = D_l \cup D_{ul}$, where D_l consists of labeled samples and D_{ul} contains unlabeled samples
Initial Estimates: Build initial classifier (L_1-regularized Logistic Regression + MLE) from the labeled training samples, D_l. Estimate initial parameter θ using MLE.
while *log likelihood increases* **do**
  E-step: Use current classifier to estimate the class membership of each unlabeled sample, that is, the class with maximum probability that the sample belongs to that particular class (see (3)).
  M-step: Re-estimate the parameter, $\hat{\theta}$, given the estimated label of each unlabeled sample (see (2))
end
Output: An MLE classifier that takes the given sample (feature vector) and predicts a label.

Data-Driven Discovery of Partial Differential Equations (PDE)

Algorithm 3: $\mathrm{STRidge}(\mathbf{\Theta}, \mathbf{U}_t, \lambda, tol, \mathrm{iters})$

$\hat{\xi} = arg\,min_{\xi} \|\mathbf{\Theta}\xi - \mathbf{U}_t\|_2^2 + \lambda\|\xi\|_2^2$ # ridge regression
bigcoeffs $= \{j : |\hat{\xi}_j| \geq tol\}$ # select large coefficients
$\hat{\xi}[\sim$ bigcoeffs$] = 0$ # apply hard threshold
$\hat{\xi}$[bigcoeffs] $= \mathrm{STRidge}(\mathbf{\Theta}$[:, bigcoeffs]$, \mathbf{U}_t, tol, \mathrm{iters} - 1)$
recursive call with fewer coefficients
return $\hat{\xi}$

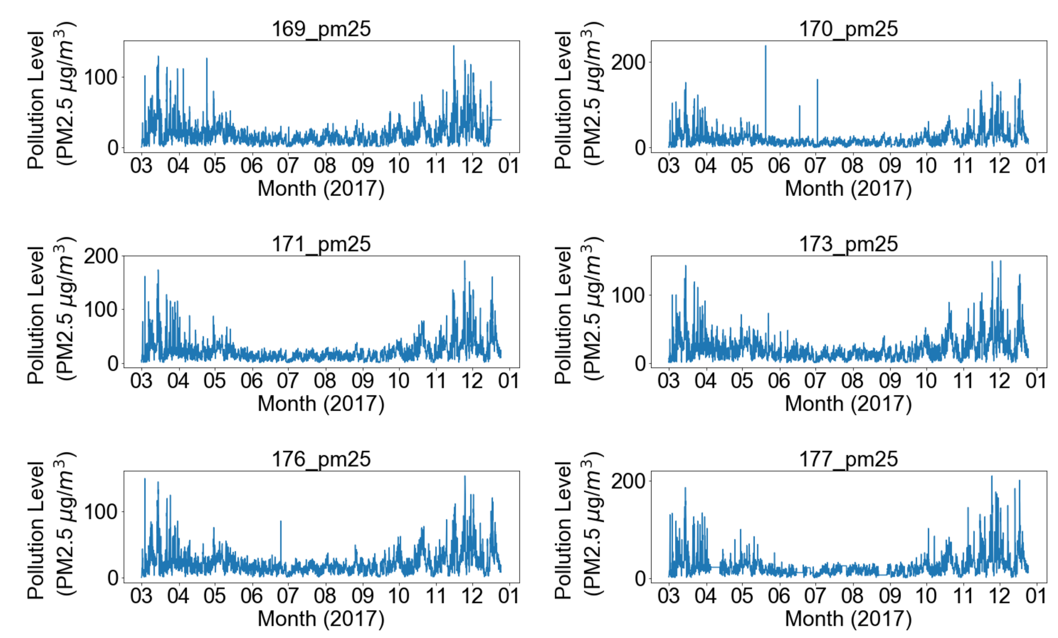


Figure 4. Pollution data over 10 months for 6 sensors. The pollution levels are higher in the fall and winter months.

BELLKOR RECOMMENDATION RESULTS

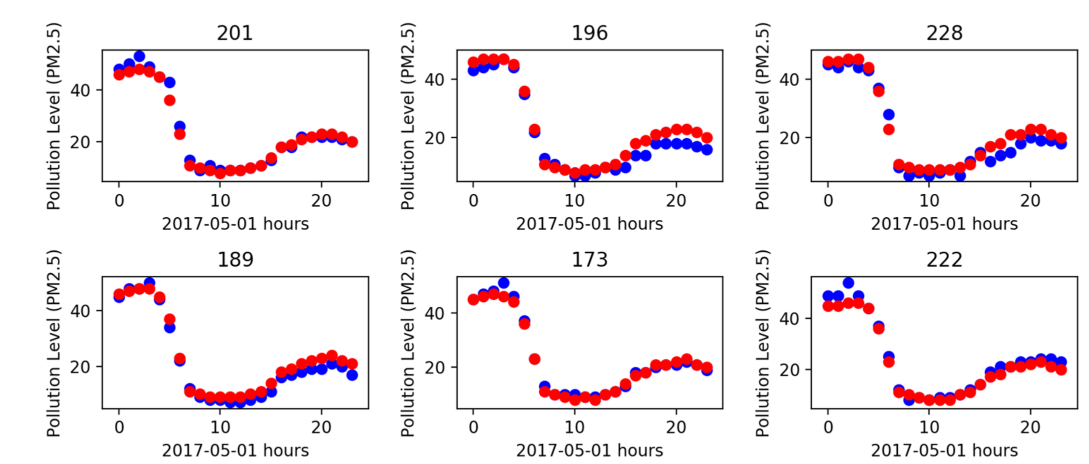


Figure 5. Comparison of latent factor model results (red) vs true records (blue) .

Table 1: R^2 measurement for all test sensors

	189	201	173	196	222	228
R^2 scores	0.935	0.915	0.912	0.906	0.822	0.778

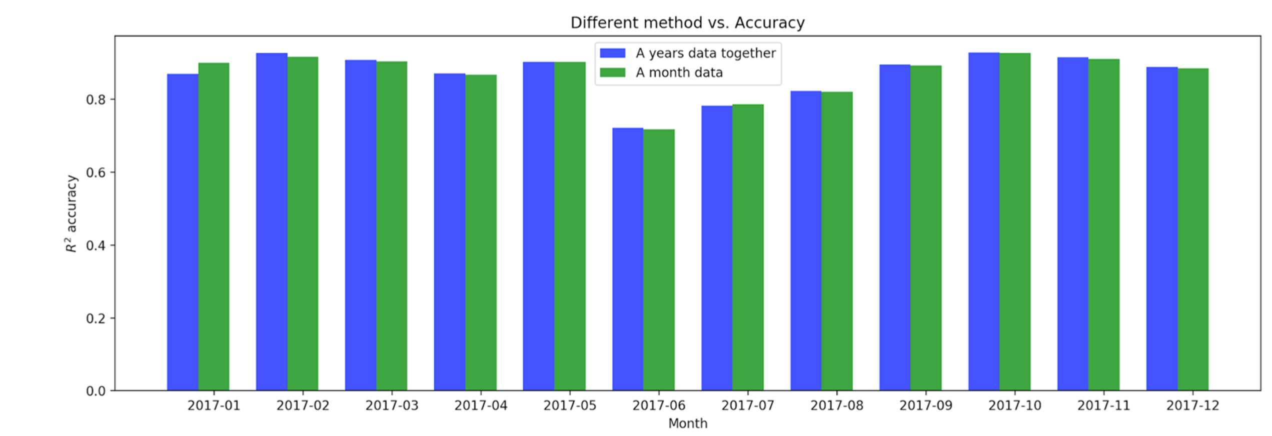


Figure 6. Accuracy considering year vs month data.

- This model can measure the overall trend with R^2=0.928.
- The model does not perform well for current time trend (best R^2 = 0.484).
- The model is not suitable for prediction due to low accuracy which might be due to missing features in data.

SEMI-SUPERVISED LOGISTIC REGRESSION RESULTS

$$l(f(x),y) = 1(f(x) \neq y) \quad (5)$$

$$\%accuracy = 100\left(1 - \left(\sum_{i=1}^{n} l(f(x),y)/n\right)\right) \quad (6)$$

Table 2: Prediction accuracy using 0/1 loss for semi-supervised classification

	0th Hour	1st Hour	2nd Hour	3rd Hour	4th Hour
L_1-regularized Logistic Regression	69.4%	61.4%	57.5%	54.7%	51.7%

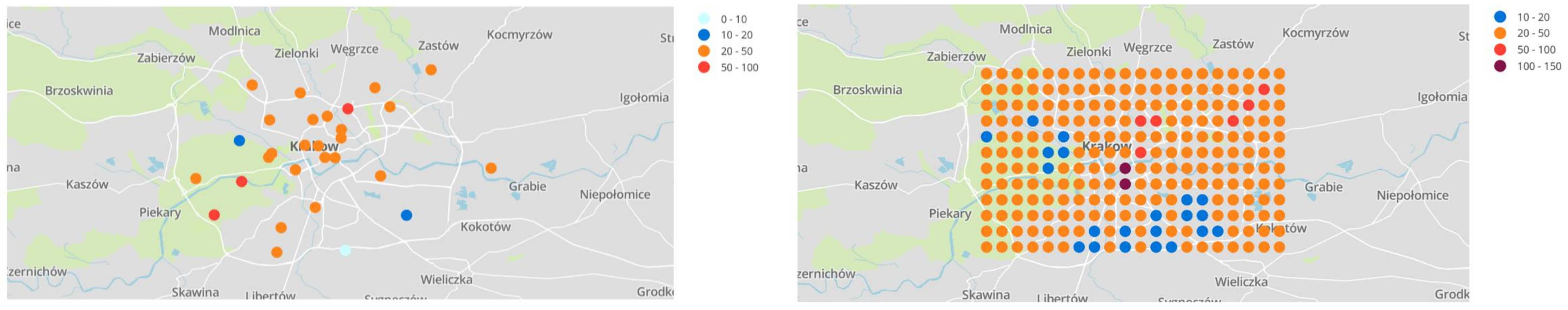
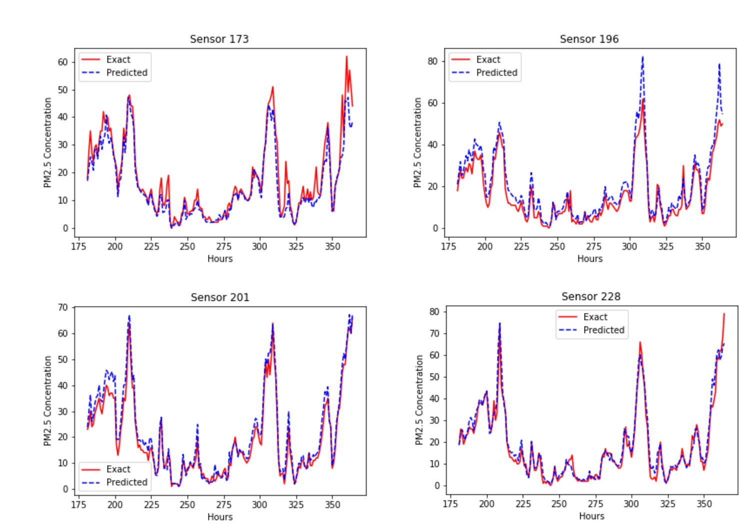


Figure 7. PM2.5 concentration labels for 7th March 6:00 AM at all 29 sensor locations (left) and PM2.5 concentrations mapped by semi-supervised L_1-regularized logistic regression model.

DATA DRIVEN DISCOVERY OF PDE

Prediction Based on Radial Basis Function Interpolation



Sensor	173	196	201	208
R^2 score	0.895	0.849	0.956	0.957

Training data Matrix

$$\underbrace{\begin{bmatrix} u_t(x_0,t_0) \\ u_t(x_1,t_0) \\ u_t(x_2,t_0) \\ \vdots \\ u_t(x_{n-1},t_m) \\ u_t(x_n,t_m) \end{bmatrix}}_{} = \underbrace{\begin{bmatrix} 1 & u(x_0,t_0) & u_x(x_0,t_0) & \dots & u^5 u_{xxx}(x_0,t_0) \\ 1 & u(x_1,t_0) & u_x(x_1,t_0) & \dots & u^5 u_{xxx}(x_1,t_0) \\ 1 & u(x_2,t_0) & u_x(x_2,t_0) & \dots & u^5 u_{xxx}(x_2,t_0) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & u(x_{n-1},t_m) & u_x(x_{n-1},t_m) & \dots & u^5 u_{xxx}(x_{n-1},t_m) \\ 1 & u(x_n,t_m) & u_x(x_n,t_m) & \dots & u^5 u_{xxx}(x_n,t_m) \end{bmatrix}}_{\text{Example } \Theta \text{ for real valued function in one spatial dimension}} \begin{bmatrix} \xi \end{bmatrix}.$$

Partial Differential Equation Generated By Algorithm 3

$$U_t = -1.47U_y + 2.2U_{xy} + 0.13U + 0.03hU_y$$

CONCLUSIONS

- Measurement of the PM level trend using Bellkor recommendation system achieved overall R^2 =0.928.
- We classify PM concentrations into 6 classes using semi supervised L_1-regularized logistic regression. The model has 69.4% mapping accuracy and 61.5 % - 51.7 % prediction accuracy for 1 - 4 hrs.

FUTURE WORK

- Generating an algorithm that can accurately calculate the derivative of the interpolated data for data driven discovery of PDE.
- Improving feature selection using different algorithms in semi-supervised classification.

ACKNOWLEDGMENTS

This project was a part of CSE547 Machine Learning for Big Data Course. We thank Prof. Tim Althoff for his feedback.

Gaurav Mahamuni gauravsm@uw.edu Mingyu Wang myw518@uw.edu Su Ye yes23@uw.edu