Forecasting fine grained pollutant levels using learning algorithms from spatiotemporal datasets

Prof. Jiebo Luo
Nilesh Patil **Jiang Shang**

In every pollutant monitoring study, fine grained particulate measurements are mentioned alongside the gaseous pollutants. In general particulate matter in atmosphere is driven by natural activity and the most common source of particulate matter is still oceanic salt sprays. Other natural sources are volcanic activity, storms, forest and grassland fires etc. In recent times, human activities like increasing usage of fossil fuels has led to significant jump in anthropogenic aerosols (those made by human activity) and it now accounts for about 10% of total atmospheric aerosols.[1]

The term fine grained pollutants refers to the particles in atmosphere having size smaller than 2.5μm. In our analysis and forecasting, we are focused on PM2.5 (Particulate Matter with less than 2.5μm diameter) since they are especially dangerous with a 36% increase in lung cancer per 10 μg/m3 beyond current safety standards.[2] Since they are so small and light, they stay longer in air and this increases the chances of inhaling by humans and animals. PM2.5 are able to bypass the nose and throat and penetrate deep into the lungs and some may even enter the circulatory system. Long-term exposure to PM2.5 may lead to plaque deposits in arteries, causing vascular inflammation and a hardening of the arteries which can eventually lead to heart attack and stroke. Scientists in the study estimated that for every 10 micrograms per cubic meter (μg/m3) increase in fine particulate air pollution, there is an associated 4%, 6% and 8% increased risk of all-cause, cardiopulmonary and lung cancer mortality, respectively.[3]

Inspired from Dr. Yu Zheng's study in Beijing and Shanghai, we are focusing on building a predictive model for PM2.5 levels across USA.[4]

To begin our analysis, we are sourcing data provided by United States Environmental Protection Agency which goes back to 2009 for particulate matter and meteorological datasets from 1990 onwards. We are also attempting to collate data for New York City through a third party provider and plan to use that for the project if it becomes available.

Dr. Zheng's study uses two separate classifiers, one being spatial classifier based on an artificial neural network (ANN), the other one is temporal classifier based on a linear-chain conditional random field (CRF). Instead of inferring directly from the data, they first build these two classifiers and then use them to infer air quality. They also use co-training method, which is a semi-supervised learning technique that requires two views of the data. It assumes that each example is described by two different feature sets that provide different and complementary information about an instance [4].

Two of the most direct and important applications from pollution forecast are:

1. Health alerts: Most cities issue health alerts in USA and the better a forecast is, the more effective it is.
2. Supplementing pollution control measures: Availability of reliable forecasts provides the local authorities an opportunity to issue area specific measures for controlling particulate levels and a cost effective predictive method makes it feasible on a wide scale basis.

Problem statement:

Build predictive model for PM2.5 levels in the next 24 hours for distributed locations.

Data Source:

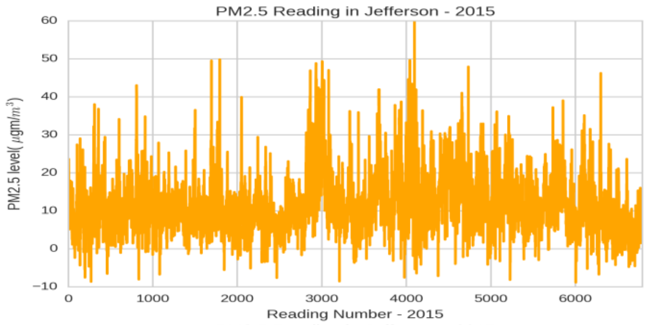
1. The Primary data source for our analysis and forecast is US-EPA[5]. EPA provides historical datasets flat files and we are using the following datasets:
   i. Particulate readings across 250+ locations in USA from 2009 onwards
   ii. Atmospheric conditions across USA from 1990 onwards
   iii. Gaseous pollutant readings from 2009 onwards
   iv. Each of the above datasets are provided at a one-hour level (i.e. each row specifies corresponding average values for one hour)
2. Supplementary data regarding each location (population, city/rural, industry etc.)
3. We are attempting to source data from a third party vendor[6] specifically for New York City, in which case, we'll be doing the exact same analysis focused on different locations within NYC. The core idea remains same in this case, but there is a better chance to analyze impact of human behavior for a single city.

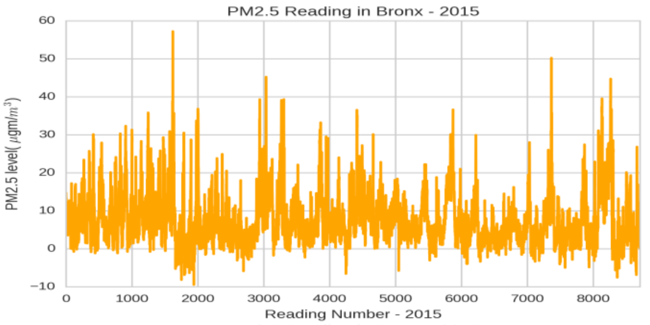
Feature Engineering:

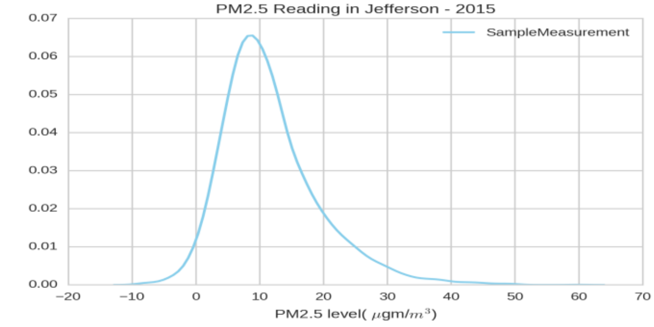
1. Our biggest challenge is the total data size and the features to use.
2. To start with, we are focused on gathering location specific parameters and the atmospheric conditions associated with these locations (Temperature, Humidity, pressure, number of rainy days)
3. We'll be using historical pollutant readings (Average – Last week, month, year) to check how significantly it impacts current readings – keeping in mind the ease to use such a variable. (e.g. it is difficult to implement a 1 day/week window but much easier to take yearly average into account as it is a onetime computation)

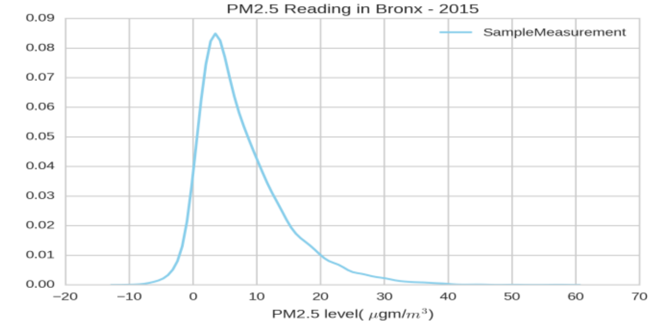
Comparing Bronx and Jefferson counties

1. We have plotted PM2.5 readings in 2015 for Jefferson and Bronx counties and it is clear that even though both counties are below current regulations (annual average should be less than 35.5µm), there a definitely high reading days which adversely affect health and businesses.
2. This is much more frequent in Jefferson and this prolonged exposure to higher levels of PM2.5 may lead to severely increased health risks.









Exploratory Data Analysis:

1. We'll start with one-hour crosslinked dataset and perform EDA to understand our raw variables.
2. We plan to check variable pairs to determine correlations and to create transformations which might prove useful.
3. We'll explore our variables by following aggregations:
    a. 8-hour variation
    b. 24-hour variation
    c. Weekly variation
    d. Monthly variation
    e. Seasonal variation
4. Depending on this analysis, we'll decide which variables to proceed with and whether we should convert our problem into a classification problem (binning PM2.5 level into 5 zones) or use it as its (Numeric)

Modeling approach:

1. As described in exploratory analysis, we are going to test how the wind speed/ temperature/ pressure/ humidity in last 12 hours ago/24 hours /48 hours/ 1 week/month/year affect a given temporal point's PM2.5 level.
2. We are adding window aggregates of PM2.5 as our input for the same time windows
3. Using these variables in conjunction with city attributes (population, core industries, area classification etc) provides us a feature rich set for building a predictive model
4. We have divided data for each city into 70:30 train: test set.

Modelling algorithms:

To begin our analysis, we are exploring the following algorithms[7]:

1. OLS regression
2. Random Forests (Randomized tree building and vote based prediction)
3. Support Vector Regression
4. Gradient Boosted Models
5. Neural networks using available implementations
6. Stochastic Gradient Descent
7. Bayesian linear regression

Validation:

We'll be using k-fold cross validation to estimate model accuracy and its prediction for the test set to determine performance on unseen data. For our model, we'll also be using subjective measure like model complexity of model to provide our final measures.

References:

[1] *Hardin, Mary; Kahn, Ralph."Aerosols and Climate Change"* *http://earthobservatory.nasa.gov/Features/Aerosols/*

[2] *Particulate matter air pollution contributes to lung cancer incidence in Europe. Ole Raaschou-Nielsen; et al. (July 10, 2013)*

[3] Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution.
Pope CA 3rd1, Burnett RT, Thun MJ, Calle EE, Krewski D, Ito K, Thurston GD

[4] *Y. Zheng, X. Yi, M. Li, R. Li, Z. Shan, E. Chang, T. Li, "Forecasting fine-grained air quality based on big data", Proc. ACM SIGKDD Conf. Knowl. Discovery Data Mining, pp. 2267-2276, 2015*

[5] https://www.epa.gov/outdoor-air-quality-data

[6] https://breezometer.com/

[7] http://scikit-learn.org/stable/supervised_learning.html#supervised-learning