Team Air Quality Midterm Report

Team Research in Applied and Computational Mathematics
Institute for Applied Computational Science
National Science Foundation, Research Experience for Undergraduates

Anthony DePinho*, Tara Ippolito*, Biyonka Liang*, Kaela Nelson*, Annamira O'Toole*

July 13, 2017

Abstract

Air pollution is a significant concern for urban and suburban settings. The wide range of personal, industrial, and natural activities that take place in an urban space each contribute to the air pollution in their own way. This project aims to model the spatial and temporal variations in urban air quality, through widespread collection of data and application of advanced statistical models as well as computational mathematical techniques. Data for the greater Boston area will be collected from a variety of different sources pertaining to air quality and its intersection with urban living, transportation, land use, and weather, in addition to data on the air pollutants themselves. Our final deliverable will be an interface through which the results of our data analysis and modeling can be accessibly communicated and productively explored by a general audience. Such a tool can help city residents make more informed decisions on how to live a cleaner life. There will be a particular emphasis, in our interface design, on aiding people who are potentially more vulnerable to air pollution, such as people with respiratory illnesses or allergies, pedestrians, and cyclists.

1 Introduction

Air pollution data, particularly in the city of Boston, play an increasingly important role in public health and environmental analyses. While there are a number of pollutants that impact human health, the pollutants Sulfur Dioxide (SO_2), Nitrogen Dioxide (NO_2), and Fine Particulate Matter ($PM_{2.5}$) are considered particularly detrimental. These three pollutants are often constituents in a groups of atmospheric particles actively monitored by the Environment Protection Agency (EPA) called criteria pollutants. For this reason, our model in this project will be focused on these three pollutants.

While pollutants often have serious impacts on the environment - for example, Sulfur Dioxide contributes to acid rain - in this project, we focus on the human health impacts. Sulfur Dioxide comes from multiple sources such as power plants and has been linked to cardiovascular and respiratory issues. Nitrogen Dioxide, whose source is primarily combustion, affects the respiratory systems and is also an indicator of other carcinogenic chemicals that are not widely measured. Fine Particulate Matter contributes to increased cortisone levels and hormone irregularities. In addition, Fine Particulate Matter, which comes upwind from the midwest and is produced by local combustion, is one of the most damaging pollutants as the small particle size allows the particles to enter the alveolar sacs of the lungs and enter the bloodstream directly [1]. The Environmental Protection Agency sets standard levels for each of these pollutants that are deemed acceptable for public health. In our analysis, we use these standards as a baseline for our models of atmospheric pollutant concentrations.

2 Problem Statement

There are two main challenges that motivate this project. Most prominently, there are five Environmental Protection Agency (EPA) air quality sensors located in the Greater Boston Area that are responsible for recording all air quality data for all the significant pollutants in the entire

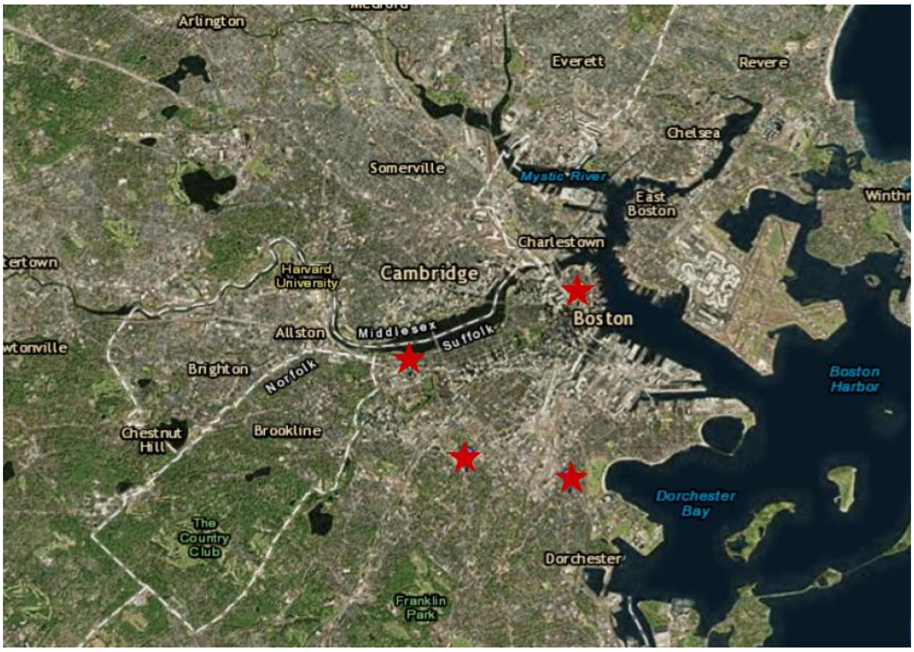

Figure 1: Locations of EPA Air Quality Sensors in Greater Boston [4]

city. Spatially speaking, there are not enough EPA sensors to provide an adequate assessment of variations in air quality conditions for the entire city, as large portions of the city are outside the range of existing sensors. Since the installation of new sensors to provide more widespread coverage of land is a difficult task, our main objective is to implement statistical modeling techniques such as Land Use Regression and the Gaussian Process to model air pollution in areas that sensors do not cover. The independent variables impacting pollutant concentration that we will consider include land use, weather, bus routes and traffic. The specifics of data collection and modeling will be discussed later in this report.

Another challenge that this project plans to tackle is the lack of accessible presented scientific data and findings, especially when such information affects the health and welfare of urban residents. The EPA makes all of its air quality data publicly available, but it does so in large data sets that use scientific language which may be inaccessible to the majority of citizens with no scientific expertise. Thus, a second motivation of this project is to design an easy-to-read, interactive interface that will provide a more intuitive visualization of significant air quality data. The interface will provide users with a tangible way to understand air quality conditions on both a widespread and granular level, so the findings of our research can be communicated succinctly and able to be interpreted in simple ways by anyone who uses the interface. In the end, we hope that our interface would serve both as an educational tool for the public and as a tool for scientists to better visualize pollution.

3 Data Collection and Cleaning

Our goal is to implement a statistical model that will predict air quality in areas of Boston lacking sensors. Our independent variables will include static or geospatial data (land use, open space, and bus routes) and dynamic or temporal data (traffic and weather).

Sources: We pulled land use and green space data using the MassGIS OLIVER tool. Air quality data was pulled from the Environmental Protection Agency (EPA). We found Boston area traffic data from MassDOT Highway Division and Boston Region Metropolitan Planning Organization. Topological data was pulled from Harvard's Center for Geographical Analysis (CGA). Weather data came from the National Oceanic and Atmospheric Administration (NOAA) and Weather Underground. Lastly, public transit route data and schedules came from MBTA (dynamic) and MassGIS's OLIVER tool (static). If time permits, we will also gather data on population movement within Boston through the US Census.

Format and Cleaning: Our land use data was downloaded as GIS shapefiles. Each land use

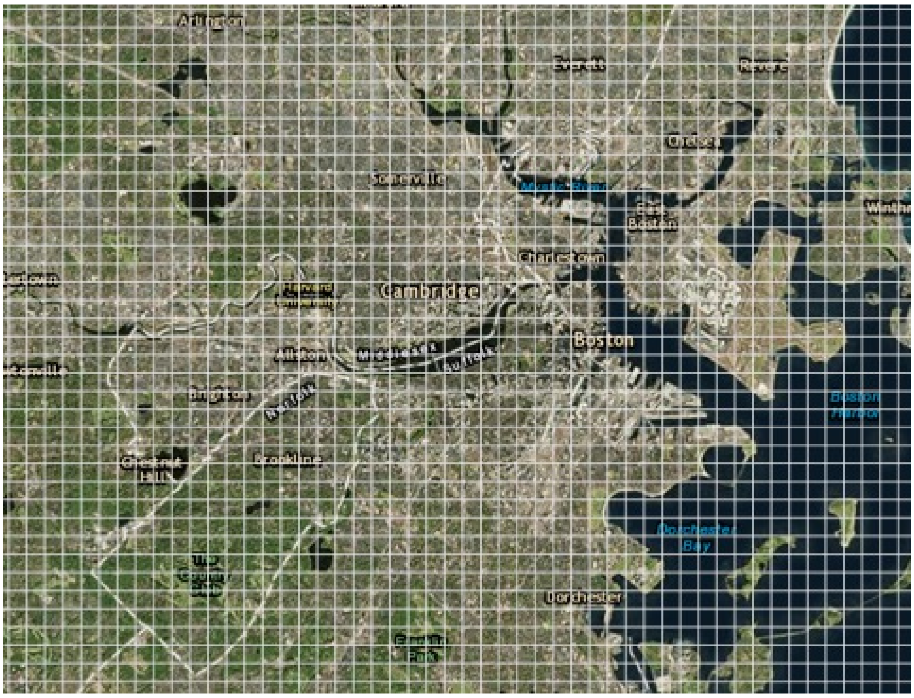

Figure 2: Visual representation of the Greater Boston area grid system [4]

type is described by a string format of polygons and multipolygons (polygons within polygons). We manually rastered this data set to a csv format, which is the same format in which the air pollutant, weather, and traffic data appear. After formatting, each layer of data was prepared to be rastered using a 50x50 grid covering the Greater Boston Area.

Rastering: Once we created data layers with a consistent format, we implemented a grid system. Within a 107.495 square mile region in Massachusetts that includes Boston, we created 2500 cells. Within each cell, we sampled 100 random points that we use to find proportions of land use within a particular cell. Our proportion system also captures green space, weather, number of bus stops within each cell. These proportions are then used as predictor variables in our models.

4 Computational Resources

Our primary computing environment is Jupyter Notebook, using a Python3 kernel. For statistical modeling we used SciPy, scikit-learn and we use Matplotlib for visualization. For GIS, we use the Shapely and PyShp libraries. The web application is developed in D3, Leaflet, CSS and HTML. Amazon Web Services is used for to large scale data collection and processing. Finally, GitHub is used for project collaboration, organization and versioning.

5 Mathematical Modeling

5.1 Land Use Regression

The Land Use Regression is a linear regression model for air pollutant concentration whose predictors include geospatial variables such as land use and average (static) weather conditions like wind speed and air pressure. Our training data consists of a collection of U.S. locations where these land use attributes are known and air pollution concentrations are also known. Using our fitted model, we will predict air quality in areas where the same land use attributes are known, but air pollution concentrations are unknown. We will perform variable selection and analysis to reason about the impact of geospatial variables on the concentration of atmospheric pollutants. The predicted air pollution concentration values from the Land Use Regression will be used as a prior for our Gaussian Process model.

We will have a separate model for each EPA criteria pollutant in our study.

5.1.1 Land Use Regression: Multiple Linear Regression

Multiple Linear Regression (MLR) is a type of regression analysis that is used to predict the value of a response variable (air quality) from of a set of multiple predictors (land use, traffic, weather, etc.). The main motivation of MLR is the belief that a set of n $>$ 1 predictors is believed to be related to a response variable y. In MLR we assume that this relationship is linear, and a combination of predictors (the land use attributes) multiplied by some unknown and fixed regression coefficients.

A linear model is parametric, and thus has a formatted equation. Let $x_1, x_2, \ldots x_n$ be a set of n predictors believed to be related to y, our response. In the context of our project, these are land use attributes. We will store our regression coefficients in the set corresponding coefficients $\beta_1, \beta_2 \ldots \beta_n$. There is also an intercept term β_0, and a noise term written as ϵ.

The Multiple Linear Regression model for the j-th sample unit (measurement) has the form:

$$Y_j = \beta_0 + \beta_1 X_(j, 1) + \beta_2 X_(j, 2), + \ldots + \beta_n X_(j, n) + \epsilon_j$$

As a result, a linear relationship is developed from more than one predictor variable, as each predictor is multiplied by a coefficient in a linear form. In analysis of this model, the coefficients that are found work together to generate a model to find as small a difference between the observed and predicted values as possible. In our Land Use Regression we assume that the noise is normally distributed, which holds in a Multiple Linear Regression as well. Further, the covariance of the noise terms of two samples (i.e, ϵ_j and ϵ_{j+1}) are zero, which shows that the error of one measurement is related to the error of another.

5.1.2 Land Use Regression: Bayesian Formulation

After implementing the land use regression in a multiple linear regression, we computed a corresponding coefficient for each predictor variable. The coefficients determine the importance of variables and their weight in our model. This raised an issue of over fitting when we began to model with a less robust data set.

By using a Bayesian regression model, we can account for this issue by incorporating priors on the coefficients the model. Incorporating priors allows us to inject bias into the model that will help us prioritize the variables that most affect air pollutant levels. The two types of priors, or regularizations, we experimented with are Lasso Regression, which focuses on a L1 regularization, and Ridge Regression, which uses a L2 regularization. L1 regularization assumes a Laplace distribution and its regularization term is the sum of weights, while L2 regularization assumes a Gaussian distribution and its regularization term is the sum of the square of the weights. Since Lasso regression creates sparse matrices, we focused on implementing Ridge Regression.

With Ridge Regression, we want to maximize the likelihood is defined as probability of the data given the model. The prior times the likelihood yields the posterior for the Bayesian model, which is the probability of the model given the data (Bayes rule). The end goal of creating these priors is to maximize the posterior. Maximizing the posterior is equivalent to minimizing the mean squared error with regularization i.e. the mean squared model with a prior.

5.2 Gaussian Processes

The above models are parametric models, which assume a fixed form. For example, the models above follow a linear form, $f(x) = a_1x_1 + a_2x_2 + ...$ Though these models are effective when we suspect $f(x)$ to follow a certain form, they lack flexibility when fitting data with more complex relationships. We now consider a non-parametric model called Gaussian Processes, in which the form of the predictor function varies with the data. This approach allows the data to guide the form of the model, rather than manipulating a fixed model to fit the data. Therefore, Gaussian Processes allow us to consider a non-linear model that is more flexible and likely more accurate than a high-degree polynomial model.

Specifically, a Gaussian Process refers to a collection of random variables $W = w_1, \ldots, w_n$ such that any finite subset of W is jointly Gaussian. For example, let $y = y_1, y_2, ...y_n$ be the NO2 levels for various states across America. We also have sets of data that we believe has some relationship with NO2 readings, such as weather, traffic and land use. Let us call this dataset X. Let $y* = y_{n+1}, \dots y_m$ be the NO2 levels for each of our grids in Boston. We want to predict the NO2 levels in Boston using the model we trained on X and y. Therefore, we can treat each element in $y*$ as a random variable such that the joint distribution of y and $y*$ are multivariate normal, $[y, y*]\ N(\mu, \Sigma)$. We can then use the properties of multivariate normals to find the conditional distribution of each $y*$ element given our y values. Because the conditionals of multivariate Gaussian distributions are also multivariate Gaussian, a Gaussian Process will output a unique Gaussian distribution for each element in $y*$. These Gaussian distributions are useful because they encode beliefs about expected value and confidence. In the figure below, we see that both distributions have an expected value of 0. However, the left distribution has a wider spread. Therefore, we would be more confident in our prediction if the right distribution described our prediction for some y_i^* in $y*$.

The assumption that our data must follow some multivariate Gaussian distribution may seem rigid at first glance, but Gaussian Processes are actually universal. Consider this: we can think of any data set $(y_1, ...y_n)$ as a point sampled from some n-th dimensional multivariate normal distribution. Since we can associate every dataset to some multivariate Gaussian distribution, intuitively, we can apply the Gaussian Process to it.

The Gaussian Process is based on the belief that the relationship between weather data, land use data, and traffic data, etc, is similar to the relationship between country-wide air quality data and Boston area air quality data (y and $y*$ respectively). We capture the nature of the relationship between our input variables with a covariance matrix. The covariance matrix is filled in by a kernel function, which for every position in the matrix takes in two sets of input data, and outputs the value of variance between those two sets. This results in the diagonal of the covariance matrix representing the variance of one input variable with itself, for example, the variance of our weather data set. All other values not along the diagonal express the covariance of one input with another, for example, how weather varies with traffic. With the covariance matrix we mathematically capture the relationship between weather, land use, traffic, etc. Since we do not have air quality data for Boston, we use the covariance matrix from our input data in combination with the air quality data we have from other cities around the United States to approximate it.

In our Gaussian Process, we use the radial basis function (RBF) kernel function:

$$K(x, x^*) = \sigma_f^2 \exp(\frac{-(x - x^*)^2}{2\ell^2}) + \sigma_n^2 \delta(x, x^*)$$

Where σ_f^2 is the amplitude of the air quality approximation, ℓ is the length scale, and σ_n^2 is the noise variance.

Currently, we have trained a Gaussian Process model using land use and pollution levels data from many countries around the US. However, this data was collected in the 1970s. Outdated data hinders us from making accurate predictions about current pollution levels. In the next two weeks, we will add weather and sea level data to our model, which were collected in the past decade. In order to improve the accuracy and granularity of our model, we aim to add real-time traffic data and a time element in our Gaussian Process by the end of week 9.

6 Interface

Our goal for our web interface is to make it an interactive and educational experience for all users. We want to implement a tour that educates the user on the effects of air pollution and our research findings. The other setting in our interface will be an interactive advanced view setting, which will include air pollutant in relation to health effects, visual of the different geographical layers, and a time bar of the level of air pollutants that vary throughout the day.

We have currently made progress in the following areas. We have used HTML as a tool to set up the format and outline of our interface. This includes assigning space allocated to the map, sidebar, and footer. We used CSS as our stylesheet for our interface. Here, we implemented the color scheme, the translucent sidebar and footer, and the two buttons (tour and advanced view). Lastly, we used D3, a javascript library, to attach an interactive map that is draggable and zoomable, implement the on-click reaction on the buttons, and create a interactive drop down menu on the advanced view button. We are currently working on attaching a grid with our different layers of data onto this map. From there, we will start working on our advanced view features.

7 Acknowledgements

Weiwei Pan, Harvard Institute for Applied Computational Science
Pavlos Protopapas, Harvard Institute for Applied Computational Science
Gary Adamkiewicz, Harvard T.H. Chan School of Public Health
Jaime Hart, Harvard T.H. Chan School of Public Health

8 References

[1] Much of the information on the health impacts of the pollutants we are studying comes from Dr. Jaime Hart from the Harvard T.H. Chan School of Public Health.

[2] Hasenfratz, David, Olga Saukh, Christoph Walser, Christoph Hueglin, Martin Fierz, Tabita Arn, Jan Beutel, and Lothar Thiele. "Deriving High-resolution Urban Air Pollution Maps Using Mobile Sensor Nodes." Pervasive and Mobile Computing 16 (2015): 268-85. Web.

[3] Hankey, Steve, Greg Lindsey, and Julian D. Marshall. "Population-Level Exposure to Particulate Air Pollution during Active Travel: Planning for Low-Exposure, Health-Promoting Cities." Environmental Health Perspectives 125.4 (2016): n. pag. Web.

[4] Satellite map images and air quality sensor locations were obtained from the Environmental Protection Agency (EPA) AirData Map website. Web. 12 July 2017. https://epa.maps.arcgis.com/apps/webappviewer/index.html?id=5f239fd3e72f424f98ef3d5def547eb5&extent=-146.2334,13.1913,-46.3896,56.5319