Scope of Work for Data Science Project

Version 1.1, 2018-02-16

Prepared by	Christopher Hase Keyan Halperin Justin S. Lee Casey Meehan	christopher_hase@g.harvard.edu keyan_halperin@g.harvard.edu justin_s_lee@g.harvard.edu casey_meehan@g.harvard.edu
Prepared for	Christine Choirat Ben Sabath	cchoirat@gmail.com mbsabath@hsph.harvard.edu
Summary of changes	• Version 1.0, 2018-02-16, Initial draft	

Background

The National Studies on Air Pollution and Health (NSAPH) research group, within the Harvard T.H. Chan School of Public Health (HSPH), use statistical methods and machine learning to research the health impacts of air pollution. One topic of interest is the health effects of Fine Particulate Matter, or $PM_{2.5}$. Airborne particulate matter is classified as $PM_{2.5}$ if it has a diameter of 2.5 micrometers or less[1]. $PM_{2.5}$ originates from both natural sources (volcanoes, forest fires, fields) and manmade sources (factories, industrial chemicals). There is much research interest in the effects of $PM_{2.5}$ on human health. Evidence suggests that $PM_{2.5}$ levels are positively correlated with multiple diseases and conditions affecting the heart and lungs, with negative effects on health and quality of life.

NSAPH has access to data on $PM_{2.5}$ concentrations collected from over 2000 sensors located across the United States. In previous work (Di, et al. 2016), using this data in conjunction with satellite data on particulate concentrations in the atmosphere, the group has trained neural networks to predict $PM_{2.5}$ levels at a given locale from geographic and atmospheric properties of the locale, as well as spatial and temporal nearby terms[2]. The group has also developed other classifiers to use in ensemble with the neural network to improve prediction robustness.

NSAPH believes that its previous work can be improved with respect to model accuracy and procedures for missing data imputation. The goal of this project is to work with NSAPH to examine its existing $PM_{2.5}$ models and create new ones with more robust performance.

[1] https://www.epa.gov/pm-pollution
[2] https://www.ncbi.nlm.nih.gov/pubmed/27023334?dopt=Abstract

Problem statement

Goal

Ultimately, we hope to make HSPH's causal inferences between pollution and health outcomes more robust. Our goal is to augment their existing efforts is two-fold:

1. Develop a more methodical data imputation scheme for existing sensor and satellite data: The partners have described their current method for imputing missing data as relatively coarse. This imputed data is being used to train pollution predictive models, which in turn are being used to make causal inferences about the effect of air pollution on health. As such, having robust imputation methods is critical for making confident claims regarding pollution's impact on public health.

2. Advance existing pollution prediction models: In an effort to measure pollution over the entire U.S., HSPH's pollution prediction models estimate pollutant content in regions without sensors. Current models use relatively limited input data for determining pollution (e.g. satellite atmospheric measurements, altitude). By introducing ulterior geographic data like road density, population density, and proximity to power plants, we hope to make these pollution predictions even more accurate. We also hope to make pollution predictions more accurate through alteration of the model structure. It will be critical to ensure the accuracy of our predictions using cross validation and testing.

Finally, time permitting, we hope to extend HSPH's project scope by making statistically driven recommendations of optimal locations for new pollution sensors. The existing map of air pollution sensors is sparse in certain regions. Given a budget for new sensors, it is critical to understand which geographic locations would optimize our pollution prediction certainty. We aim to explore how sensor location impacts the robustness of our models to make these recommendations.

Document Prepared: Christopher Hase, Keyan Halperin, Justin Lee, Casey Meehan 2018

Resources available

Data available includes:

- **Sensor Network Data**:
    - Time-series data from the U.S. pollution sensor network provides over a decade of air quality samples. These measurements contain concentrations of a variety of air pollutants at each sensor site on an approximately weekly basis. The sensor network data will be used as the pollution 'ground truth' for training pollution predictive models.

- **Atmospheric Satellite Measurements (localized to sensors)**
    - This dataset includes satellite recordings of atmospheric qualities indicative of pollutant concentration (e.g. atmospheric opacity). The satellite data currently available is synchronized and localized to the sensor network: it spans the same time-frame as the sensor readings, and only exists at the sensor locations. This data is crucial for training our pollutant predictive models. Satellite data can be considered an input to the model, and sensor readings the output.

- **Atmospheric Satellite Measurements (spanning the rest of the continent)**
    - In order to make new predictions in locations without sensors, we will use the satellite measurements that span the rest of the continent. As described above, this satellite data is currently available localized at sensor locations, but is not currently available in non-sensor locations. We hope to obtain this data soon.

- **Auxiliary Geographic Data**:
    - We hope to bolster our pollution predictions by experimenting with geographic data that may be indicative of pollution, but is not currently available in the satellite dataset. For example, it's possible that incorporating predictive variables like proximity to pollution sources (power-plants, road density) or demographic information (population density) can help advance the model's predictive ability. We are currently exploring publicly available datasets for this purpose. We will practice caution with including data that might contaminate any causal inferences based on our predictions.

Deliverables

The deliverables will be all necessary code, assets, and documentation necessary for the Client to run on their system fulfilling the following requirements:

Deliverable 1	Data imputation method which: ● Imputes missing predictor variable values
Deliverable 2	Predictive model trained on past data which: ● Predicts daily $PM_{2.5}$ air pollution values for areas of the United States with and without sensors
Deliverable 3	Integration of our imputation method and predictive model with current infrastructure used by HSPH which: ● Makes our methods seamless for HSPH use

Project timeline

Sprint ending	**Tentative milestone or goal**
2018-02-09	[Milestone 1] Get access to sensor and satellite data; EDA
2018-02-16	Explore data and literature review
2018-02-23	Read about machine learning techniques relevant to existing models (e.g. Convolutional/Recurrent Neural Networks)
2018-03-02	Finish reading existing work and try to reproduce results; prototype new models
2018-03-09	[Milestone 2] Partner Report, Midterm Presentation
2018-03-16	[Spring break]
2018-03-23	Continue work on new models
2018-03-30	Incorporate auxiliary geographic data
2018-04-06	Validation of imputation methods
2018-04-13	Validation of prediction methods
2018-04-20	[Milestone 3] Integration with HSPH infrastructure
2018-04-27	Investigate sensor recommendations framework
2018-04-30	Curate result information for partner
2018-05-06	Package deliverables and documentation together