HSPH Capstone Project

Members: Casey, Chris, Justin, and Keyan

TF: David Sondak

Problem Statement

- Work with the National Studies on Air Pollution and Health (NSAPH) research group, within the Harvard T.H. Chan School of Public Health (HSPH) to:
    - Improve existing data imputation procedures
    - Enhance machine learning models of $PM_{2.5}$ air pollution values across the United States

Background

- HSPH aims to significantly improve pollution interpolation models in order to better understand how pollution impacts public health across the U.S.

- Even at levels below the current EPA standard, there is evidence that an increase of 10 μg per cubic meter in $PM_{2.5}$ is associated with a 13.6% increase in the risk of death.

Background - $PM_{2.5}$

- $PM_{2.5}$ (a.k.a. Fine Particulate Matter)
  - Airborne particulate matter is classified as $PM_{2.5}$ if it has a diameter of 2.5 micrometers or less
  - Originates from natural sources (volcanoes, forest fires, fields) and manmade sources (factories, industrial chemicals)

Background - $PM_{2.5}$

- $PM_{2.5}$ (a.k.a. Fine Particulate Matter)
    - Evidence suggests $PM_{2.5}$ levels are positively correlated with multiple diseases and conditions affecting the heart and lungs, with negative effects on health and quality of life

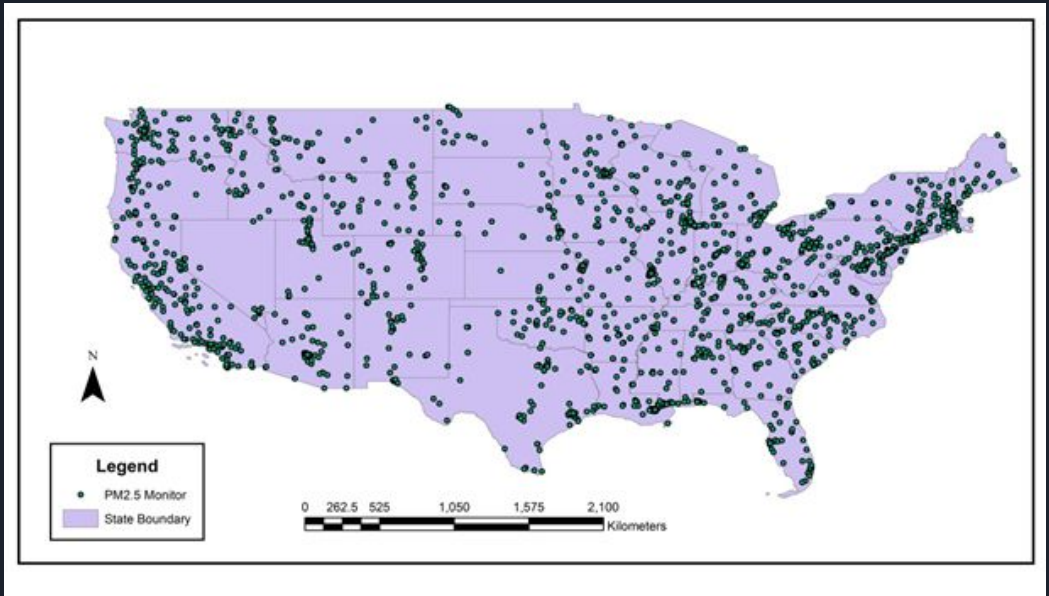
Data

- Data from 2,156 $PM_{2.5}$ monitors and 115 other variables across the US every day over the course of 16 years (2000 - 2016) for a total of 13,388,760 observations.

- Covariates include information on various weather, topographic, and satellite data.

- Significant proportion of missing data - all variables have at least some missing values; some have a majority of values missing

$PM_{2.5}$ Monitor Locations



There are many regions of the US that are hundreds of miles away from the nearest pollution sensor.

Deliverables

1. Advanced imputation methods for prediction variables
2. Augmented pollution prediction models for PM2.5
    a. Incorporate auxiliary geographic data
3. Integration with HSPH's existing software infrastructure

Project Ideas

- Develop more granular data imputation methods for sensor data with missing entries (e.g. neural nets), and more robust pollution predictions

- Find auxiliary geographic data for additional predictors of air pollution (e.g. road density, power plant proximity)

- Propose new sensor locations for bolstering predictive certainty. Given a limited budget, where are the more useful places for new sensors?

Hurdles - Infrastructure

- Need to learn more about running jobs on Odyssey
- Need to ensure that whatever programs we run on Odyssey can run on RCE
- R/H2O has many limitations with deep learning - we would like to build our machine learning models in Python
    - Researchers on project should be able to seamlessly call our Python scripts from R

Hurdles - Model

- Data does not include satellite information for areas with no sensors
- We have been told that areas without sensors may be quite different from areas with sensors, so predictions we make may involve some degree of extrapolation
    - We will try to get a sense of this degree of extrapolation using some similarity metrics or via EDA

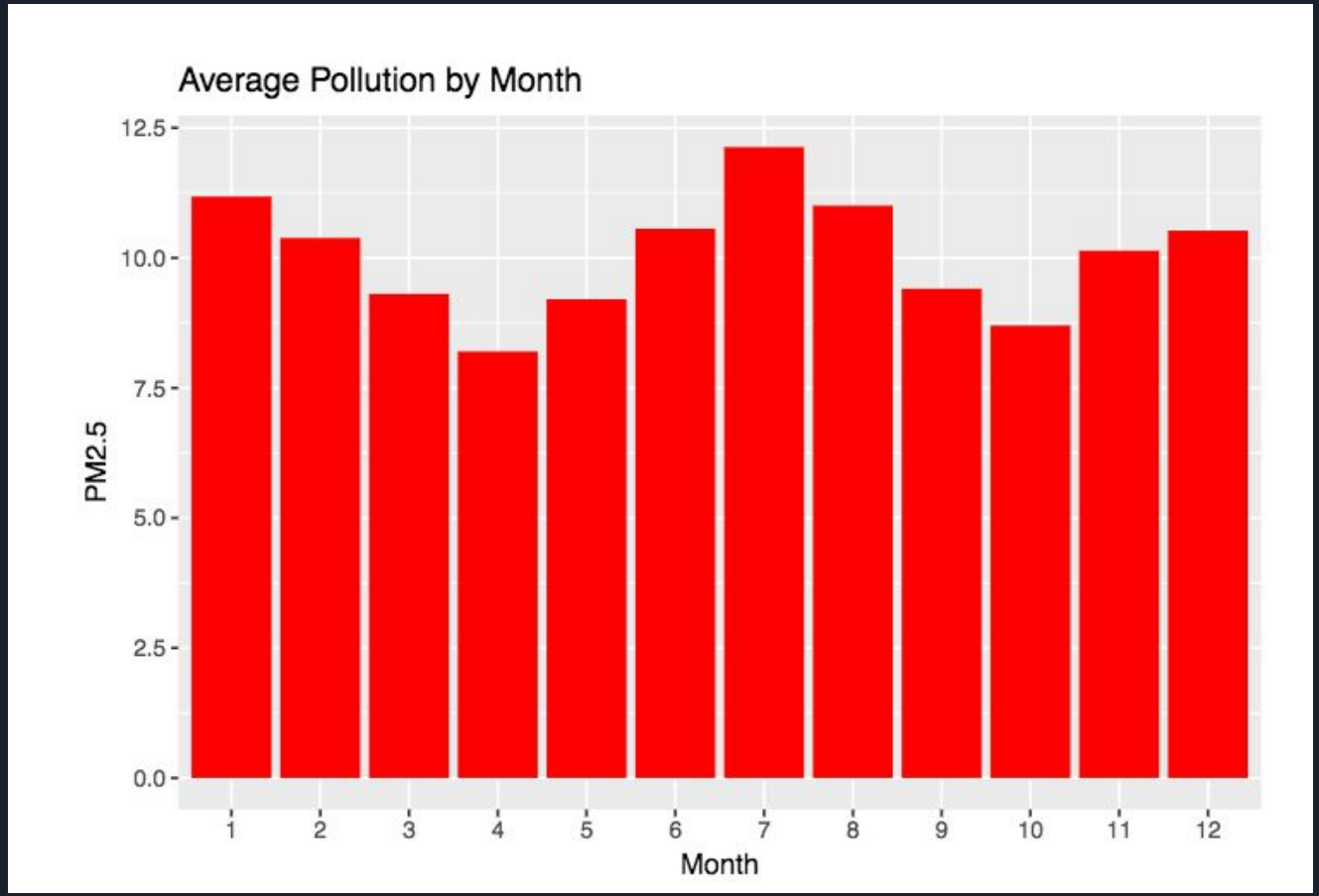
Collaboration Infrastructure

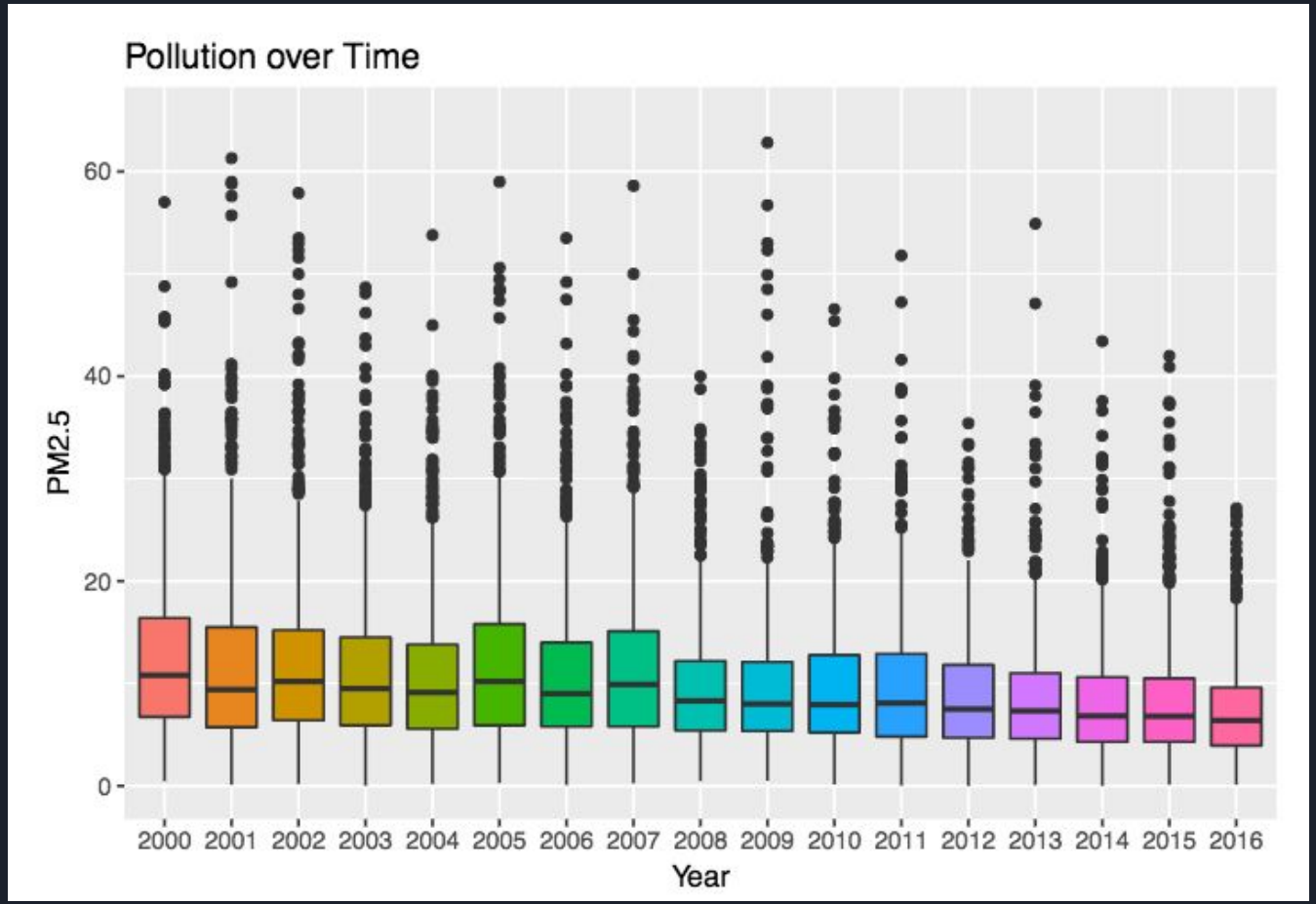
- Ben is the point of contact within HSPH
- Code will sit on GitHub
- Computing on Odyssey since we do not (yet) have compute resources on RCE
- Google Drive will contain meeting / other notes
- Group Slack channel that includes David Sondak for easy communication

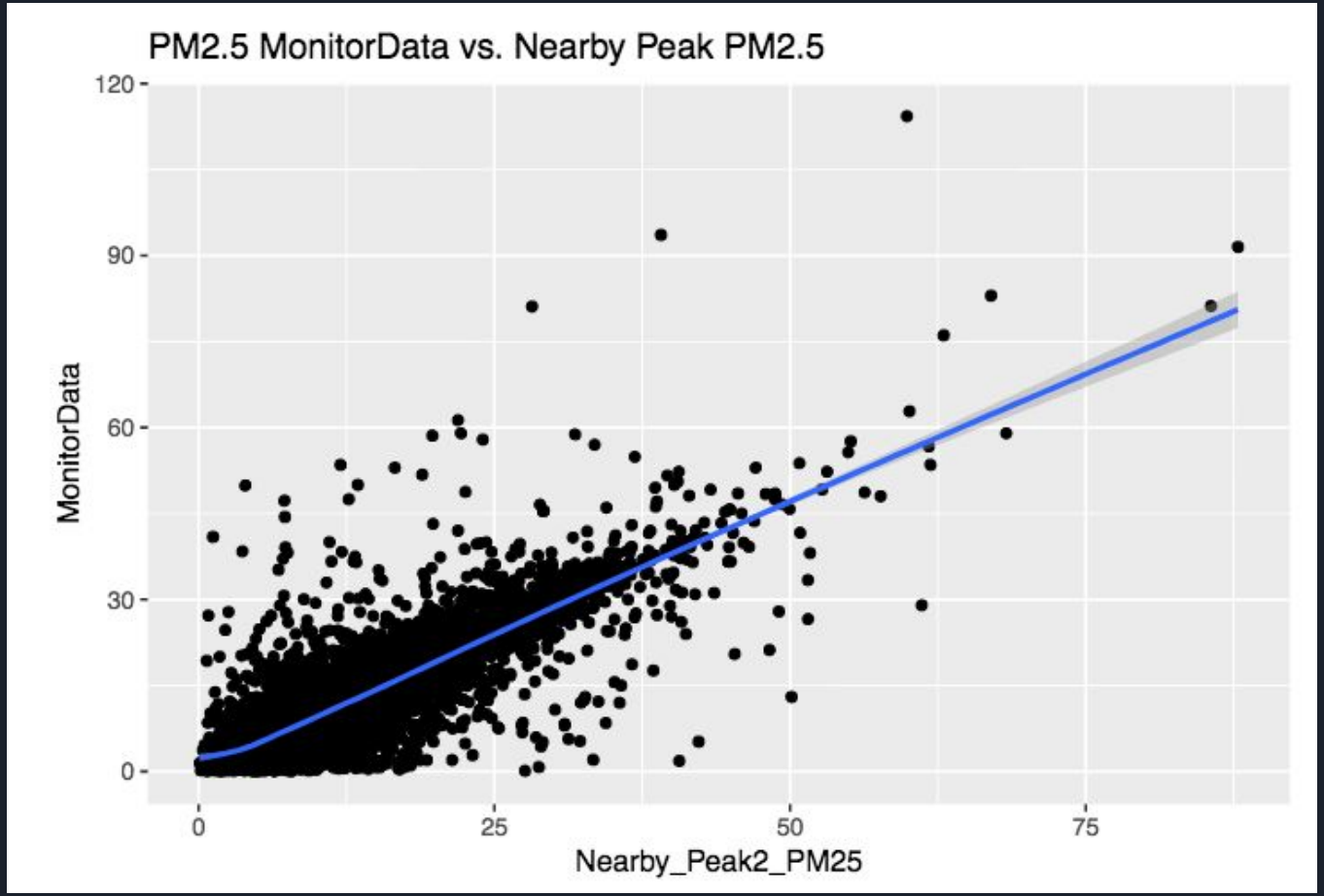
Next Steps

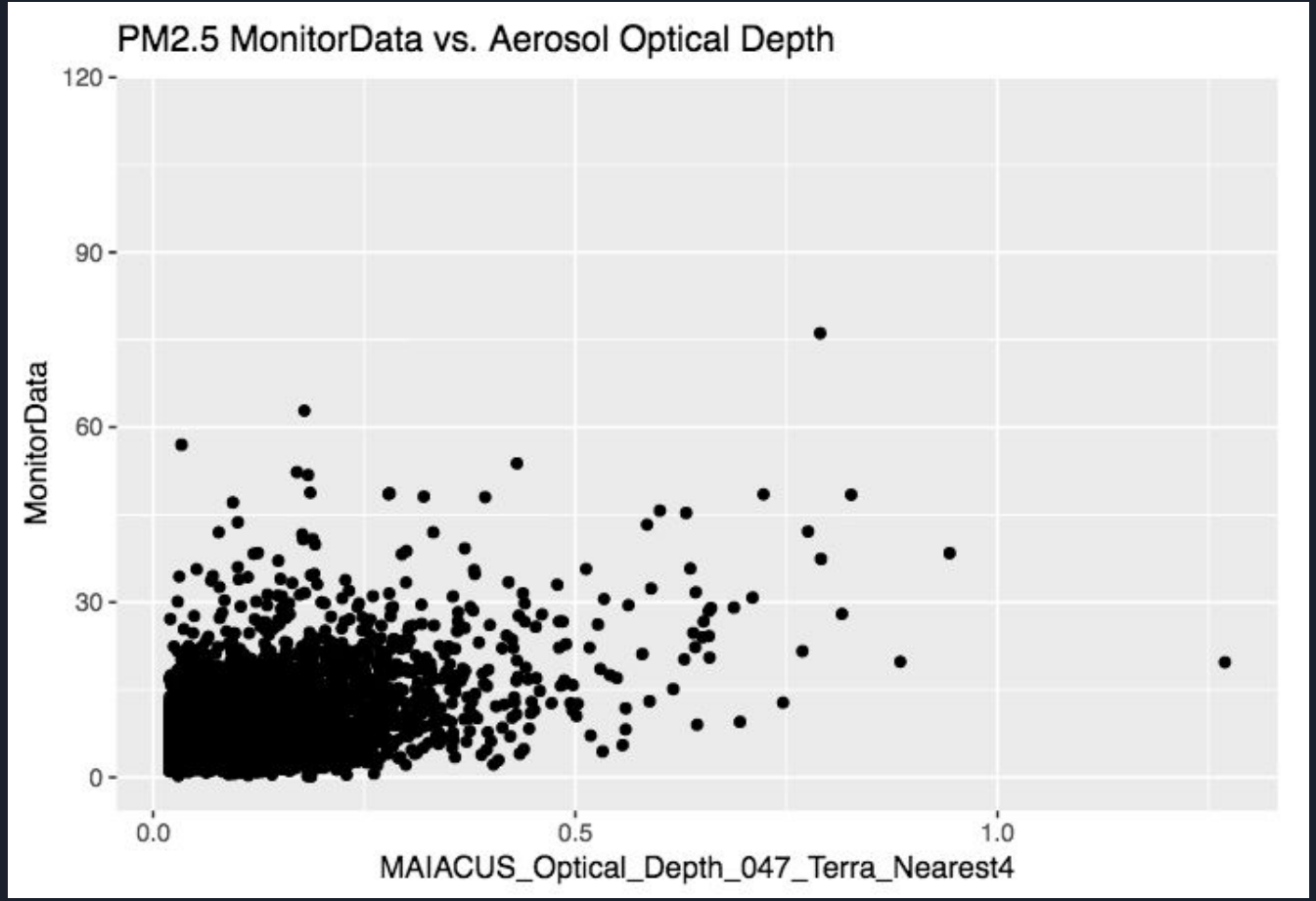
- Access satellite data for remainder of map
- Review existing imputation and interpolation methods
- Implement existing accuracy tests
- Survey alternative, more effective prediction models

Appendix: Exploratory Data Analysis


Average Pollution by Month
PM2.5
12.5
10.0
7.5
5.0
2.5
0.0
1
2
3
4
5
6
7
8
9
10
11
12
Month


Pollution over Time
PM2.5
60
40
20
0
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
Year


PM2.5 MonitorData vs. Nearby Peak PM2.5
120
90
60
30
0
MonitorData
0
25
50
75
Nearby_Peak2_PM25


PM2.5 MonitorData vs. Aerosol Optical Depth
MonitorData
120
90
60
30
0
0.0
0.5
1.0
MAIACUS_Optical_Depth_047_Terra_Nearest4