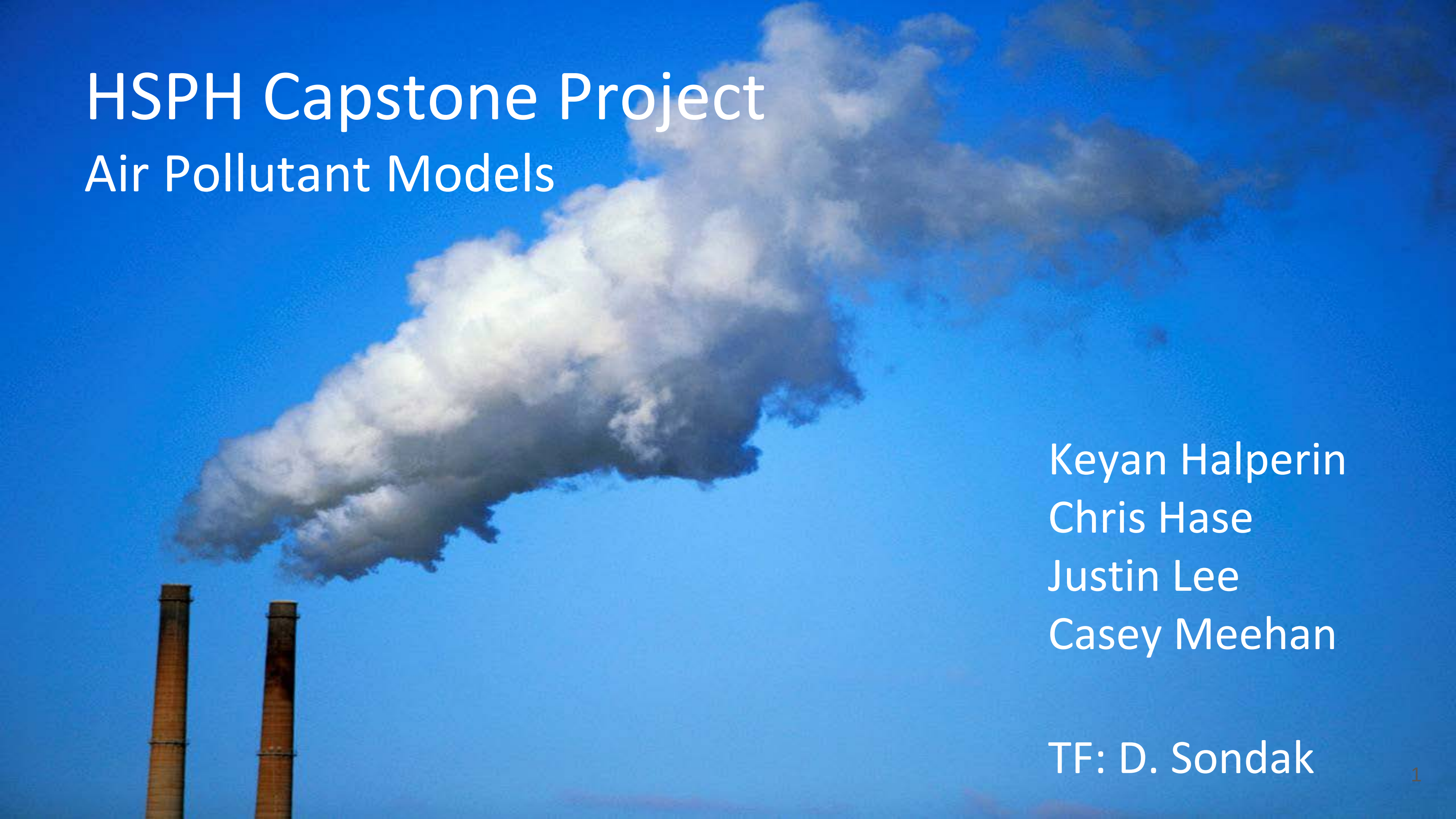


HSPH Capstone Project

Air Pollutant Models

Keyan Halperin
Chris Hase
Justin Lee
Casey Meehan

TF: D. Sondak

Our Partners

National Studies on

Air Pollution and Health (NSAPH)

PI: Prof. Francesca Dominici,

Harvard T.H. Chan School of Public Health

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

IACS
INSTITUTE FOR APPLIED COMPUTATIONAL SCIENCE
AT HARVARD UNIVERSITY


What’s the Problem?



Health Impacts

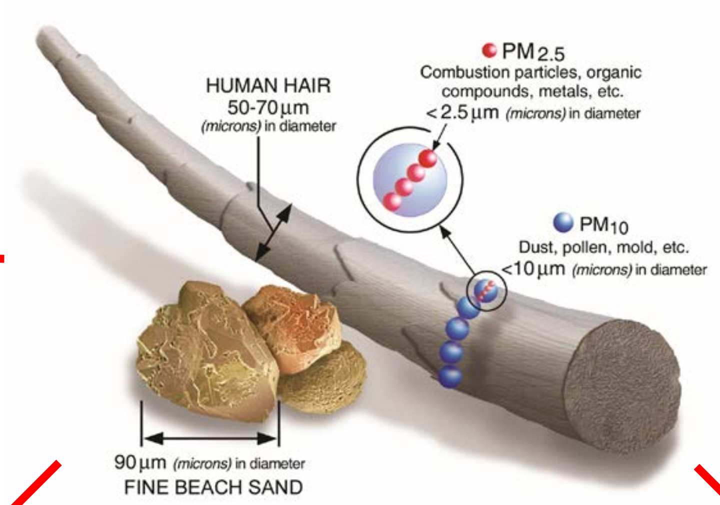
Causal Inferences

Project Scope

Pollution = Bad

Asthma

Lung Cancer



Heart Attack

Stroke

Taking Action

I

115TH CONGRESS
1ST SESSION

H. R. 3981

To establish a cost of greenhouse gases for carbon dioxide, methane, and nitrous oxide to be used by Federal agencies, and for other purposes.

IN THE HOUSE OF REPRESENTATIVES

OCTOBER 5, 2017

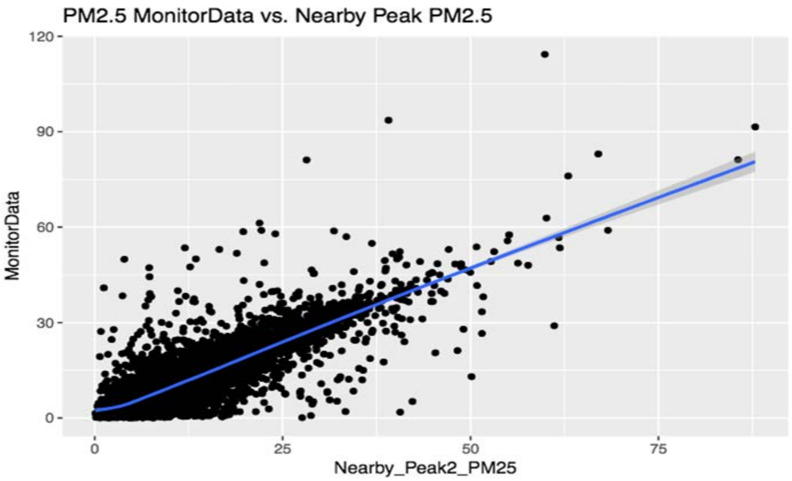
Mr. MCEACHIN introduced the following bill; which was referred to the Committee on Oversight and Government Reform, and in addition to the Committee on the Judiciary, for a period to be subsequently determined by the Speaker, in each case for consideration of such provisions as fall within the jurisdiction of the committee concerned

---

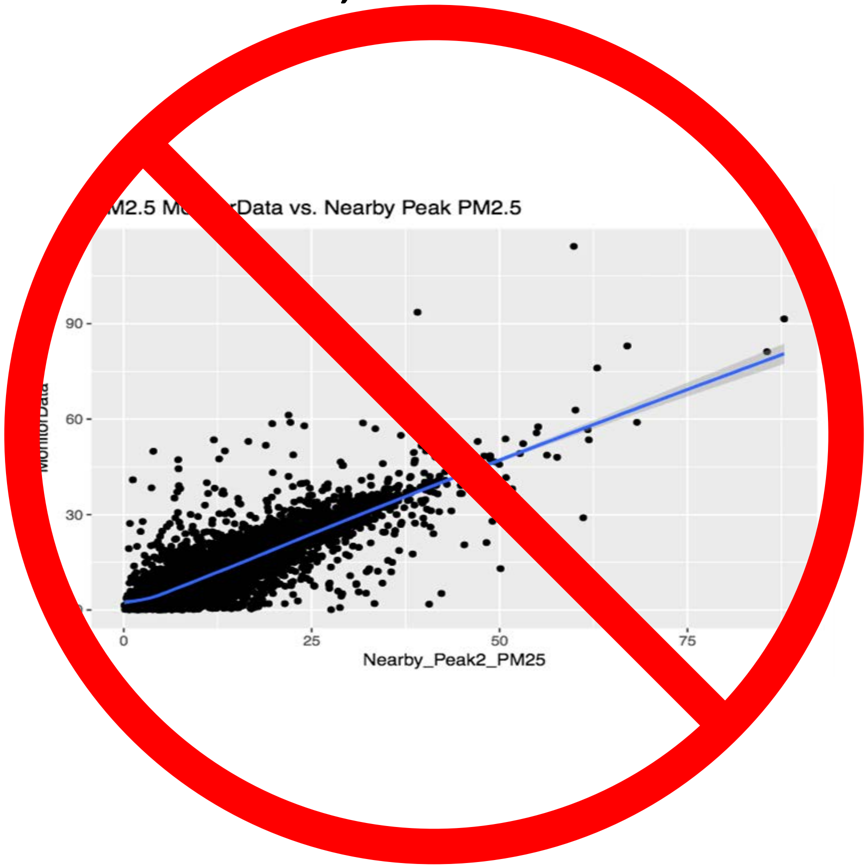
A BILL

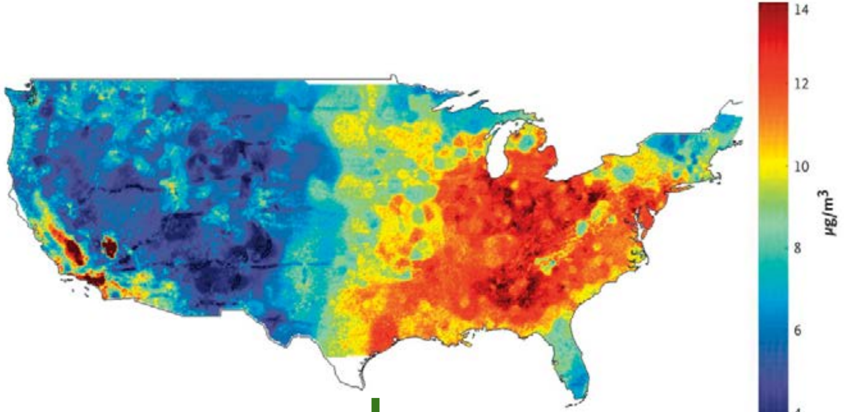
To establish a cost of greenhouse gases for carbon dioxide, methane, and nitrous oxide to be used by Federal agencies, and for other purposes.

Correlation v. Causation



Causation, not Correlation

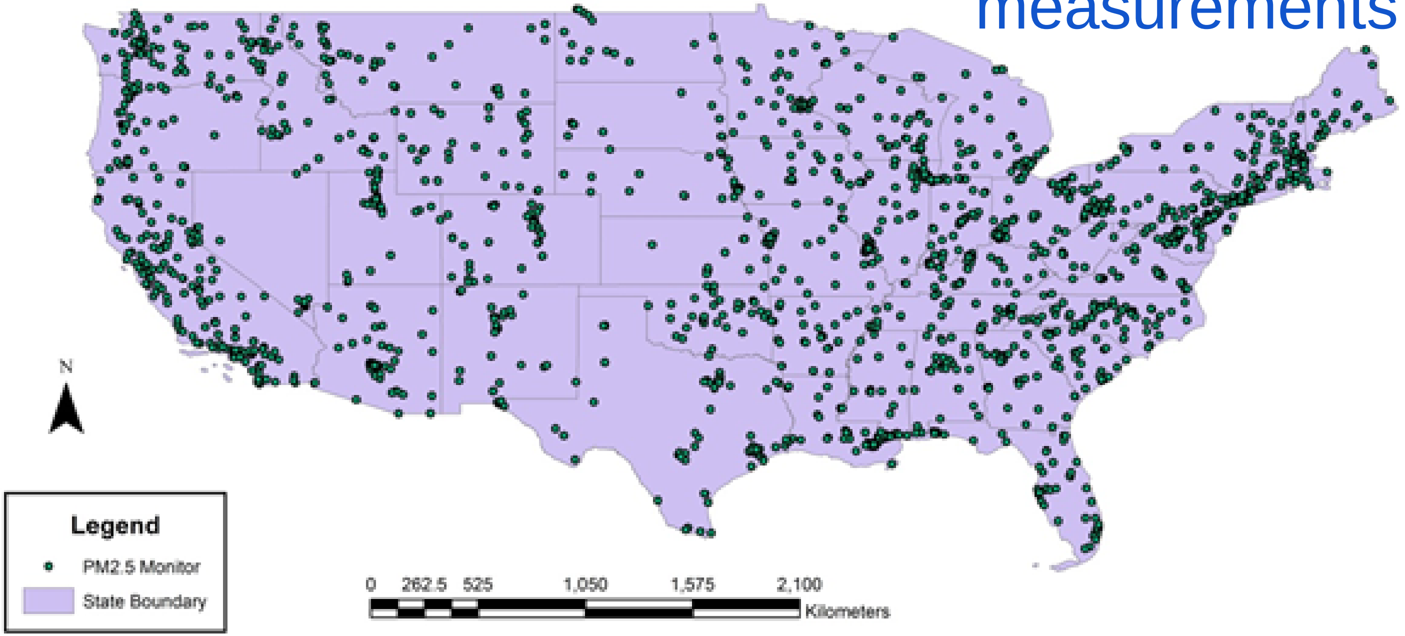




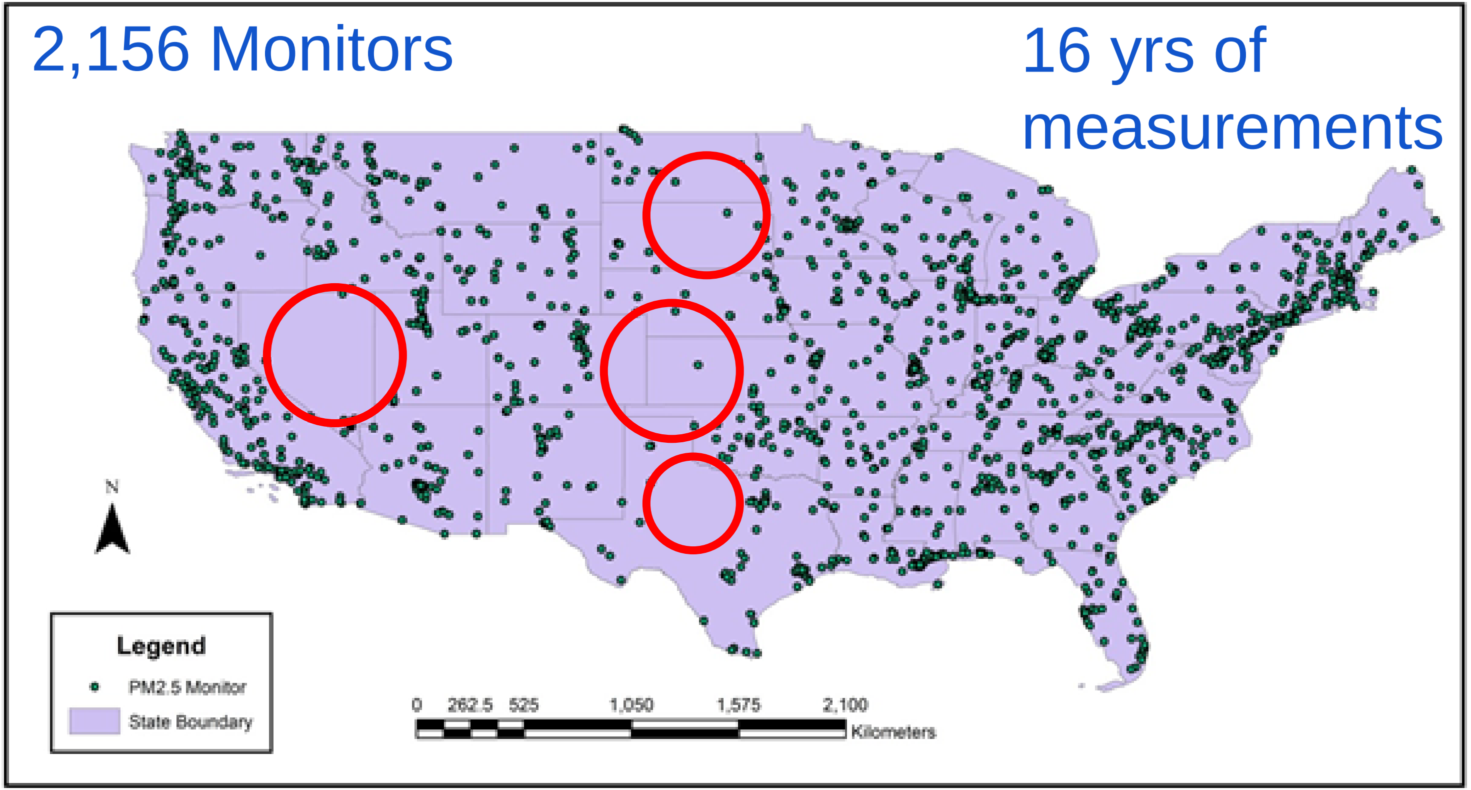
Building a Robust Map of Pollution

2,156 Monitors

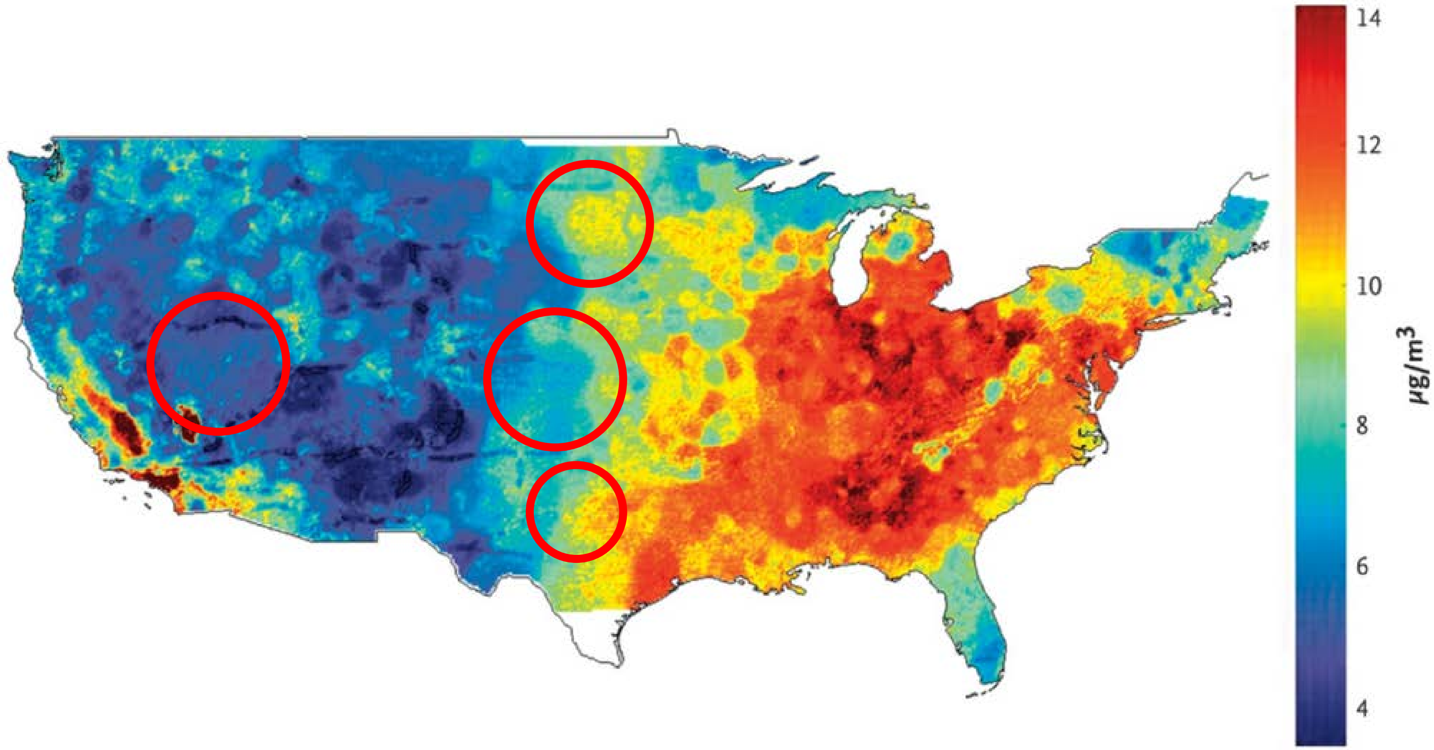
16 yrs of measurements



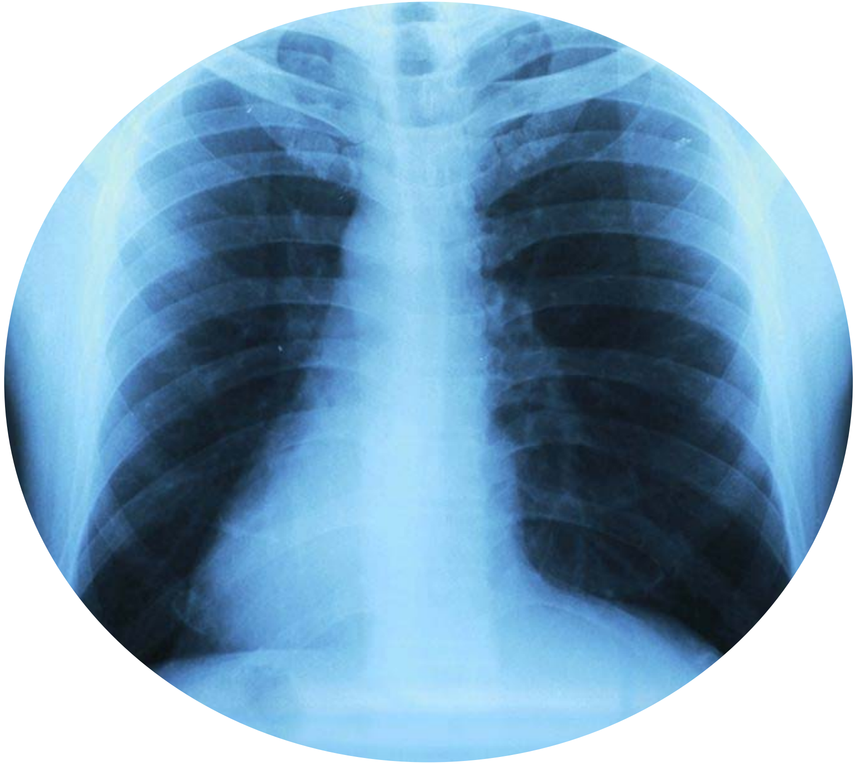
Building a Robust Map of Pollution



Building a Robust Map of Pollution



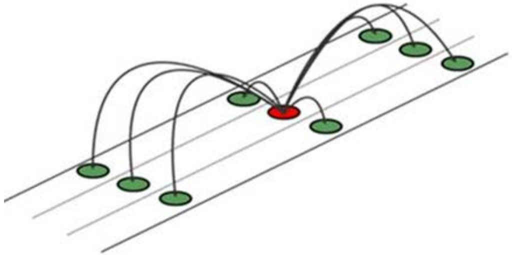
Building a Robust Map of Pollution



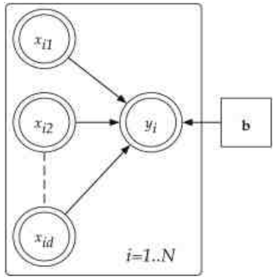
Where We Come In:

Augment our Predictors

Large-Scale Imputation



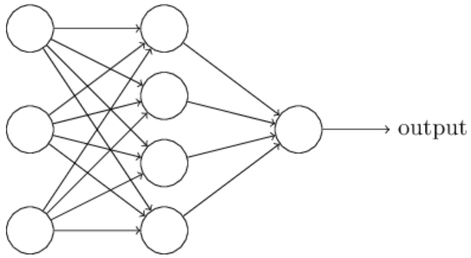
High-dim Variable Selection



Auxiliary Data

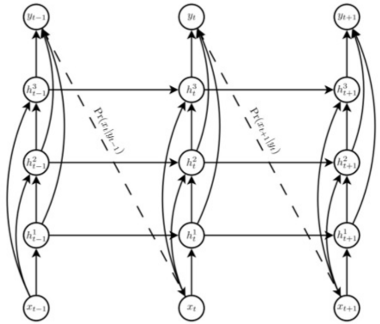
Where We Come In:

Experiment with Models

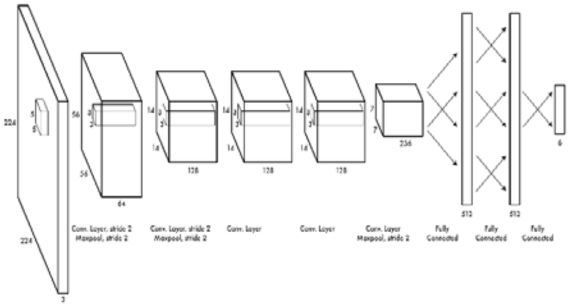


(Currently Feed-Forward)

RNNs



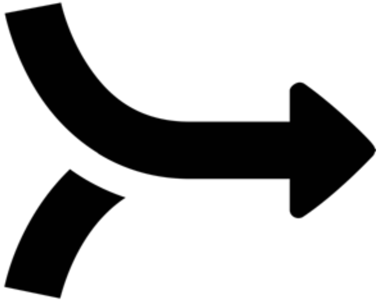
CNNs



Where We Come In:

Offer an Extensible Design

Accepts new data



Key Parameter Tweaking

What are We Working With?

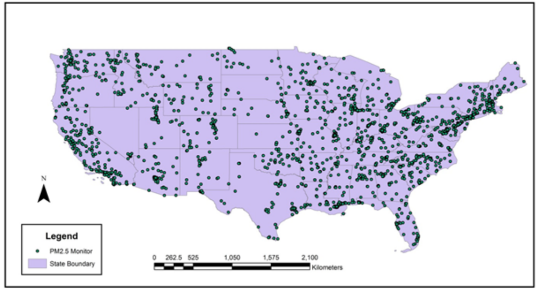
Imputation Problem

Existing Model

Given Data

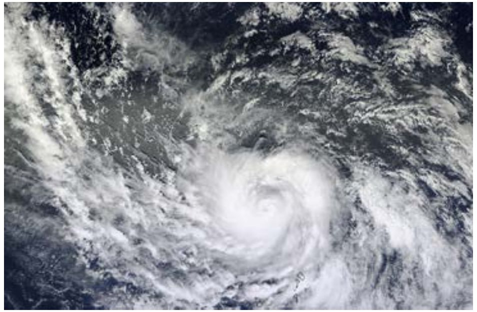
16yrs
2156 Sites

Sensor Data
(Response)



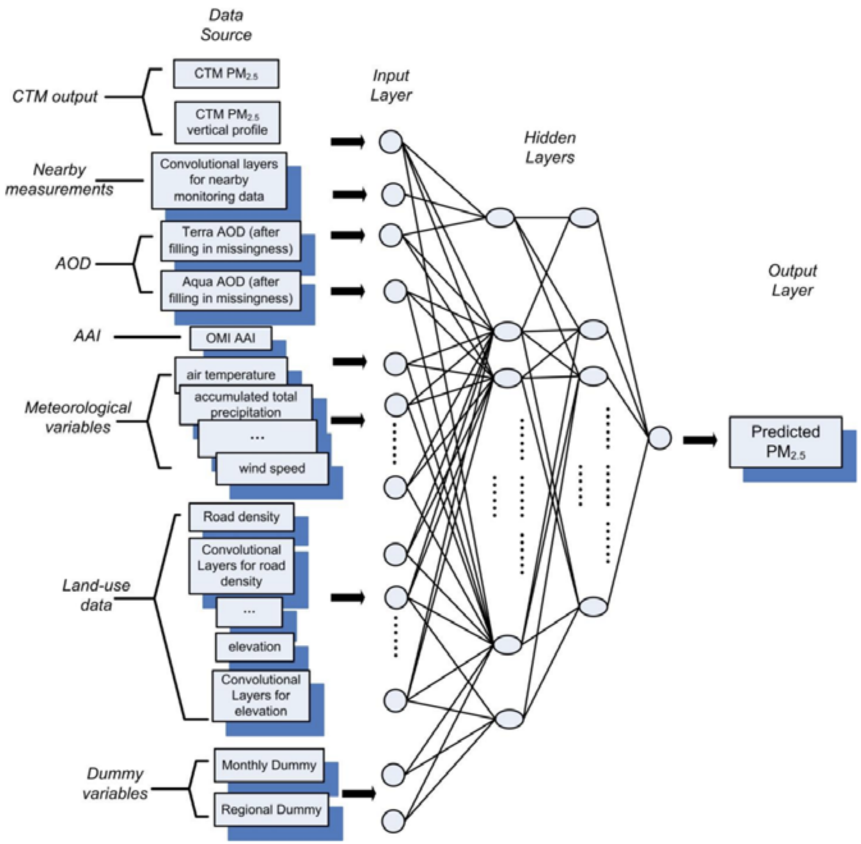
- 13M PM2.5 Sensor-Days
- ~75% sensor-days missing
    - Defunded sites (costly)

Satellite Data
(Predictors)



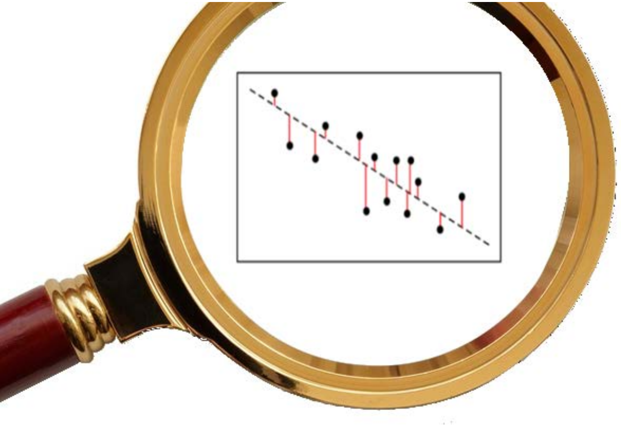
- 115 measurements/day/site
- Also severe missingness
    - Cloud cover
    - Snow reflection

Existing Model - Functional, not Optimal?



Room for improvement:

- Currently FF
    - (Could benefit from RNN)
- Convolutional layers for alternative geographical data
- Modular, legible, extensible code



Looking at the Data

Exploratory Data Analysis

Feature Correlations

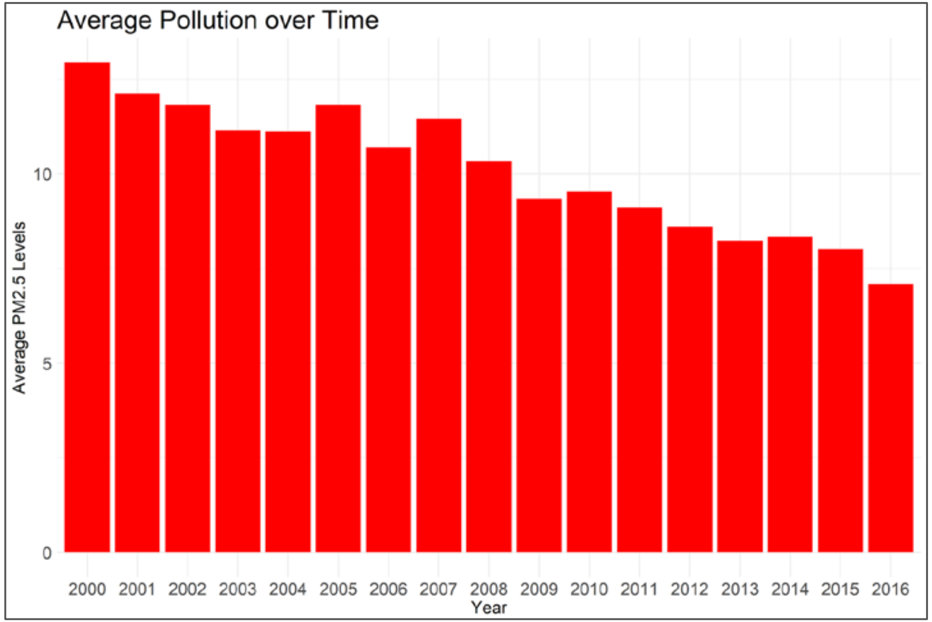
Baseline Models

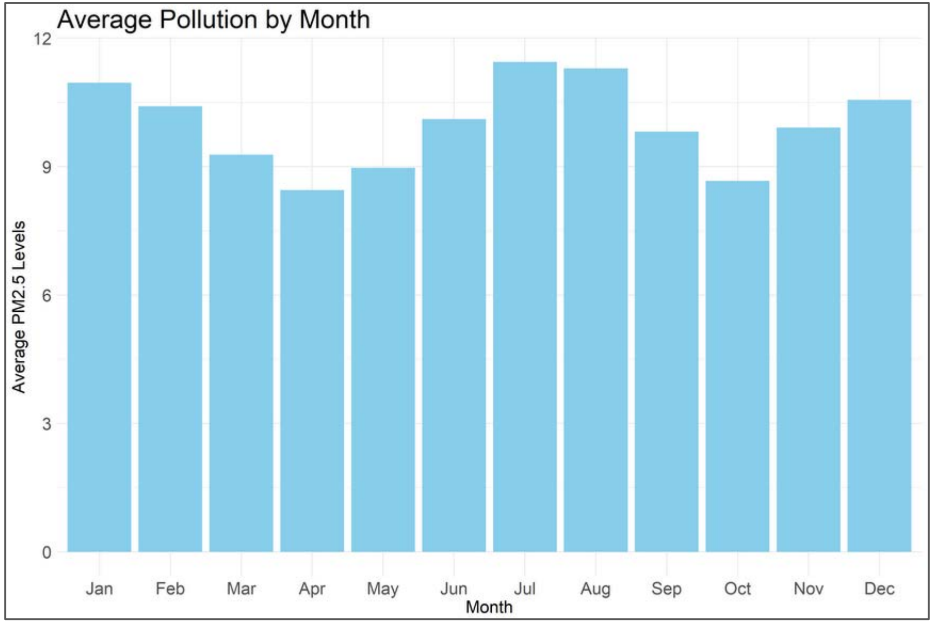
EDA - Setup

- ~13.4 million rows total in sensor data
    - 115 predictors, 1 response column

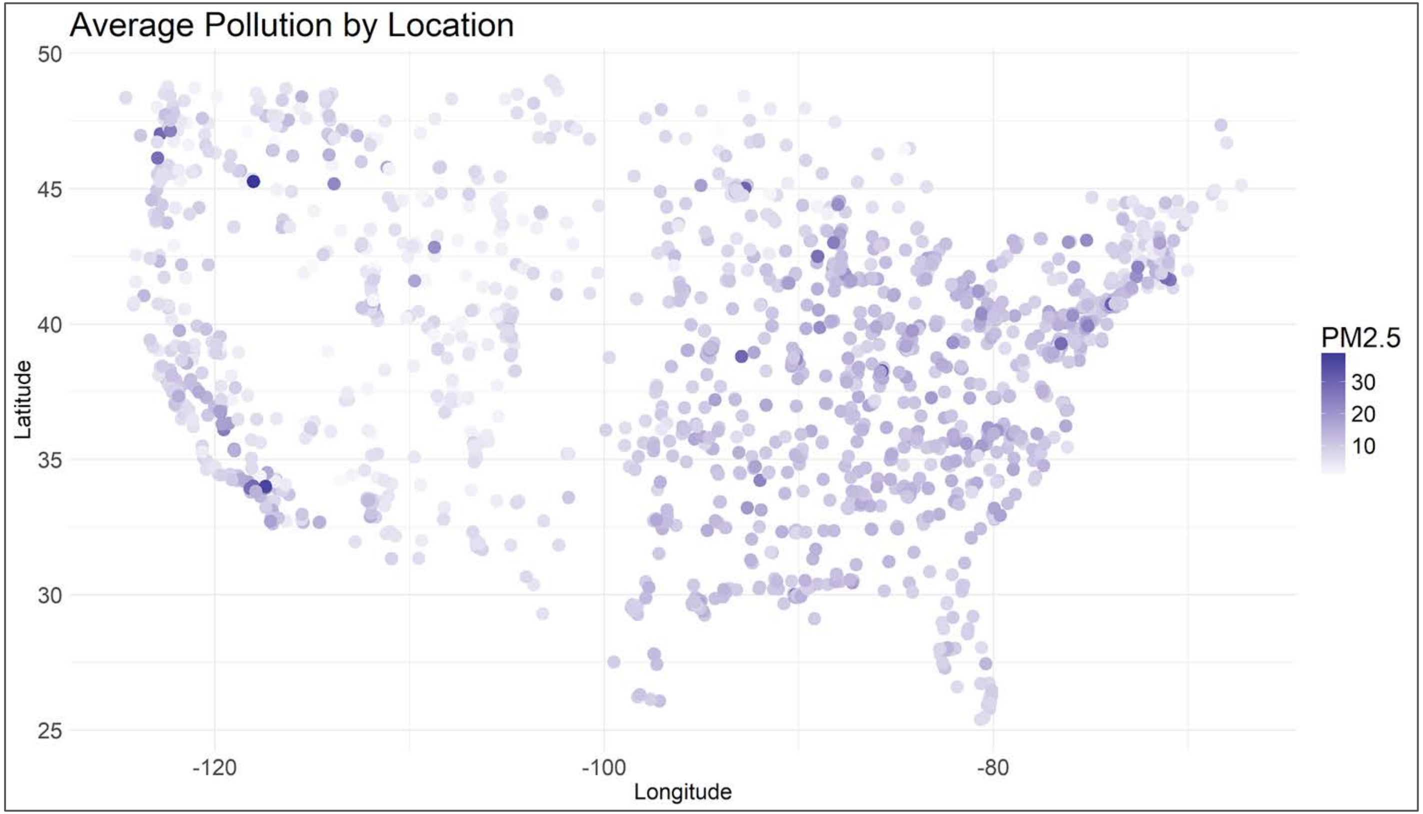
- Took a random sample of 1% of the data (~134,000 rows) in order to perform exploratory data analysis
    - Simplified preliminary work in exploring data - proportions of missing data, correlations

EDA - Pollution Over Time





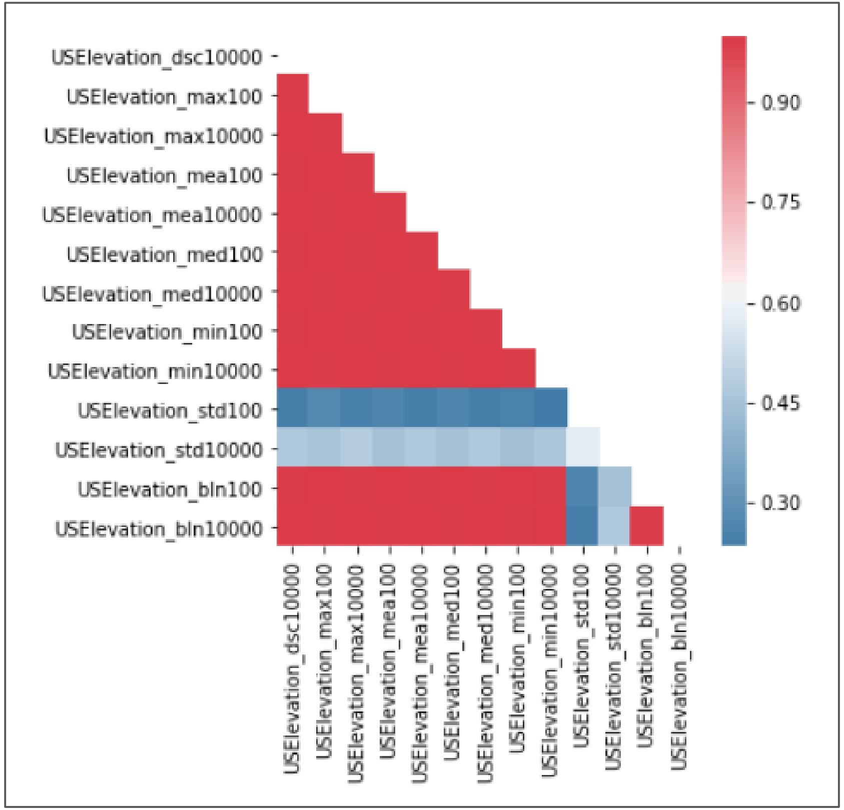
EDA - Pollution by Location

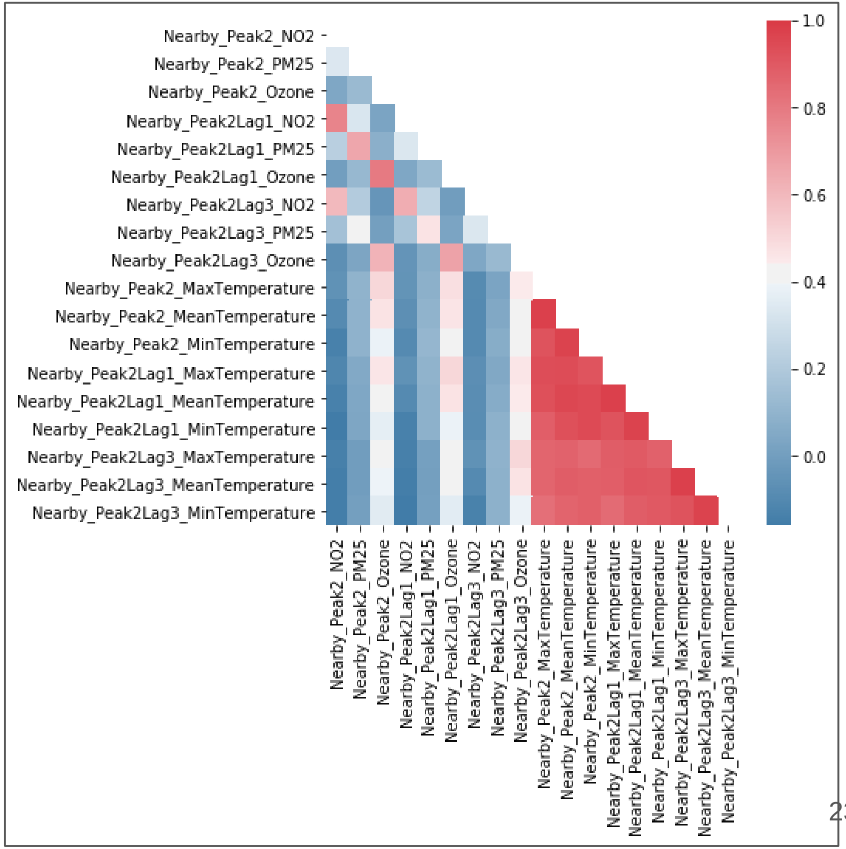


Pairwise Correlations - Groups

Category	Description	# Columns in Category
US Elevation	Statistics on elevation of site locations	13
NLCD	National Land Cover Dataset	16
Road Density	Statistics on presence of roads	5
MAIAC	Aerosol Optical Depth	6
REANALYSIS	Meteorological Data	34
MOD11A1	Surface temp, cloud cover	4
Nearby Terms	Spatial/Temporal Nearby Terms	18
OMAERO	Ozone Monitoring Instrument (OMI) Aerosol Product	3

Pairwise Correlations



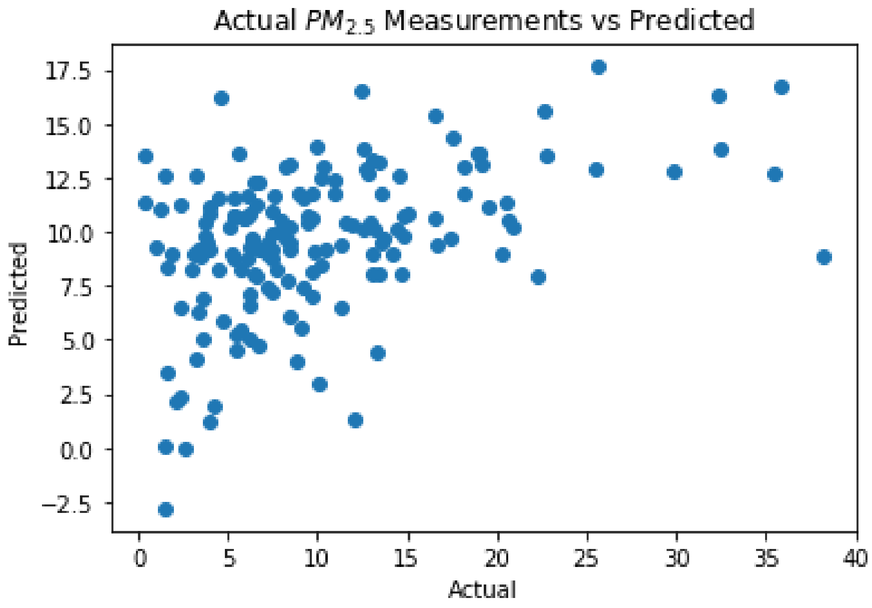


Baseline Model - LASSO

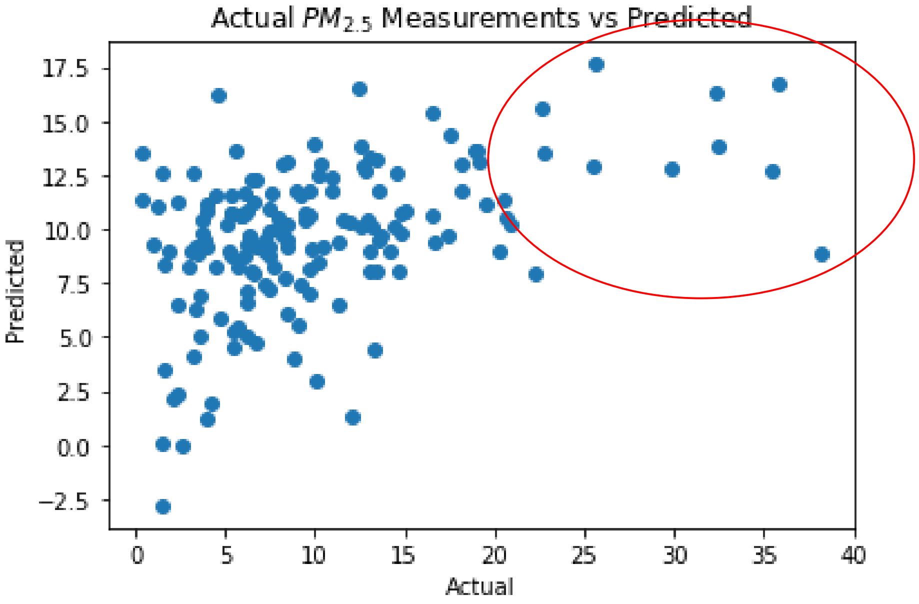
- Only data from 2010 with non-missing response

- Considered a subset of columns with low within-group pairwise correlation and low missing data proportion, plus location and month

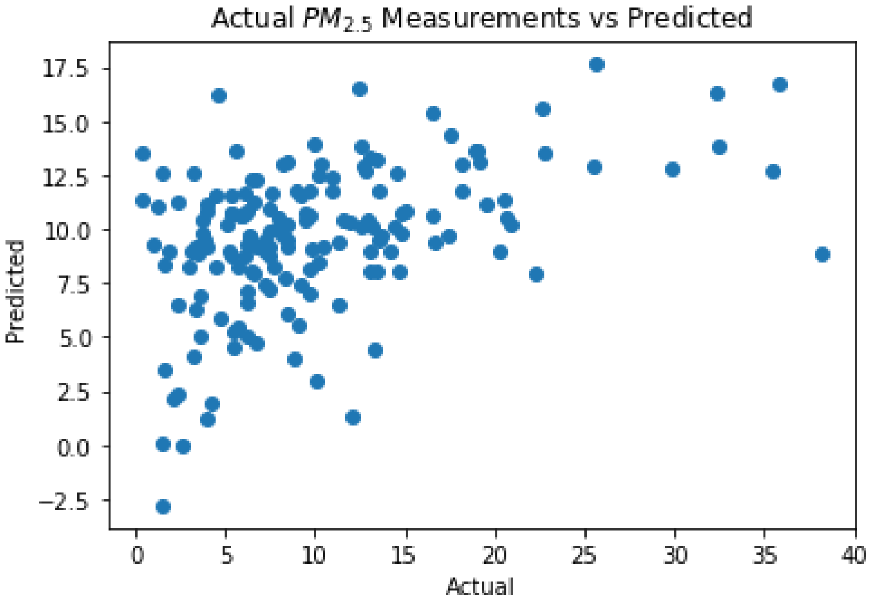
- Test R^2 of 0.192

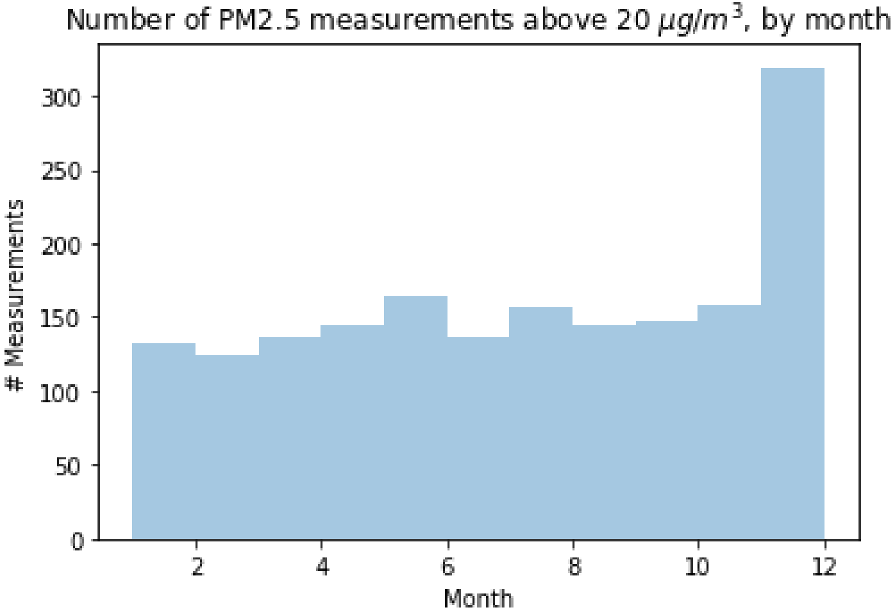


Searching for Patterns



Searching for Patterns



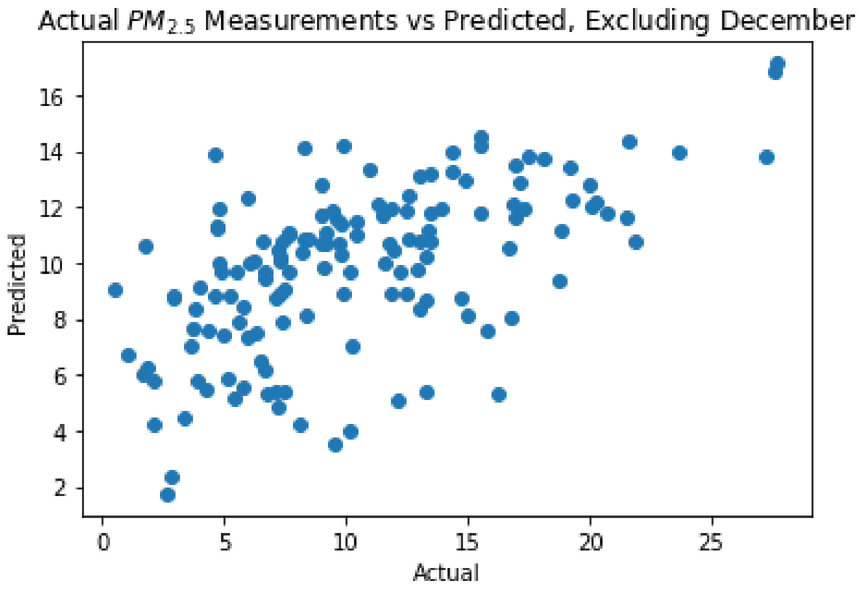


Baseline Model - LASSO ex. December

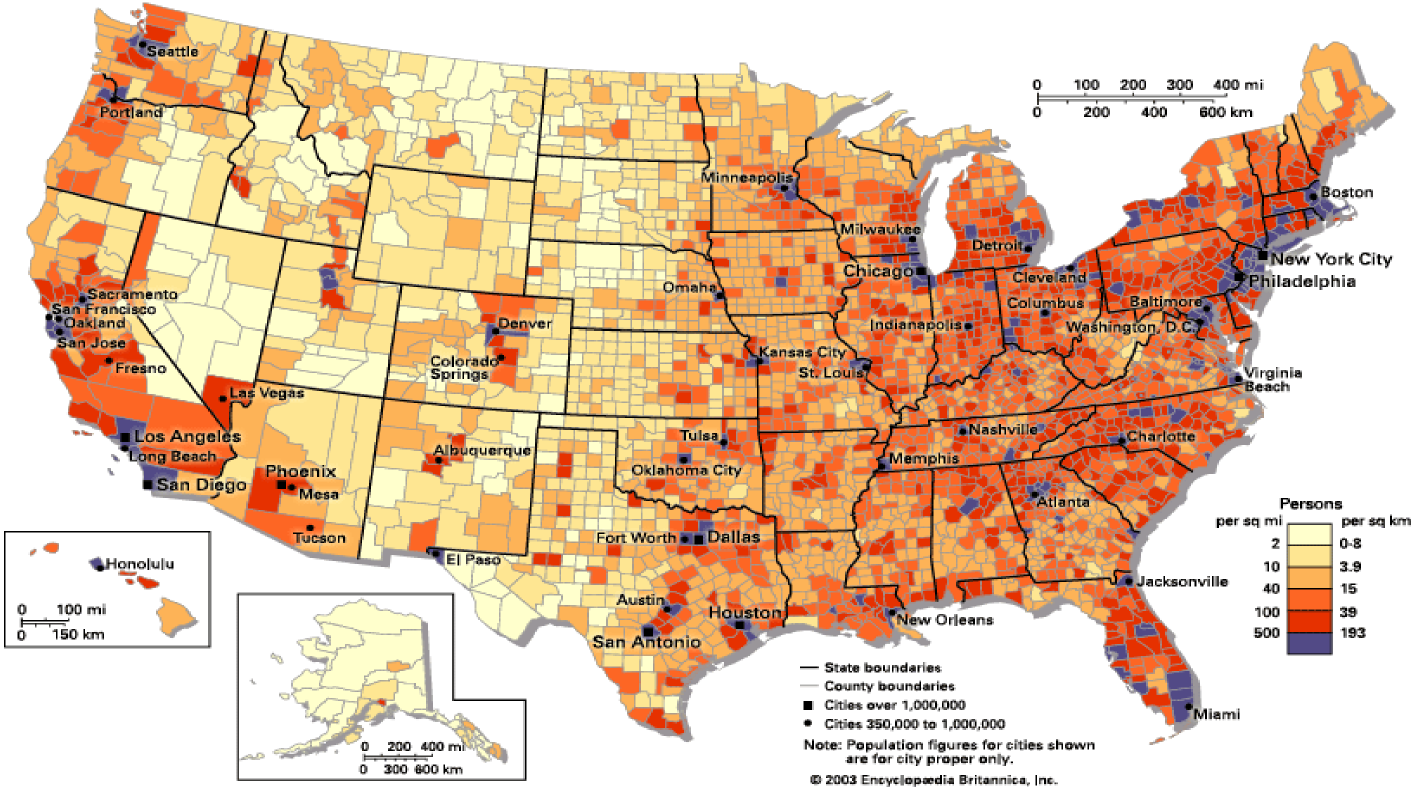
- Same dataset as before, but excluding measurements from the month of December

- Test R^2 of 0.331

- Shows the need for a model that captures temporal dependencies



Incorporating External Data

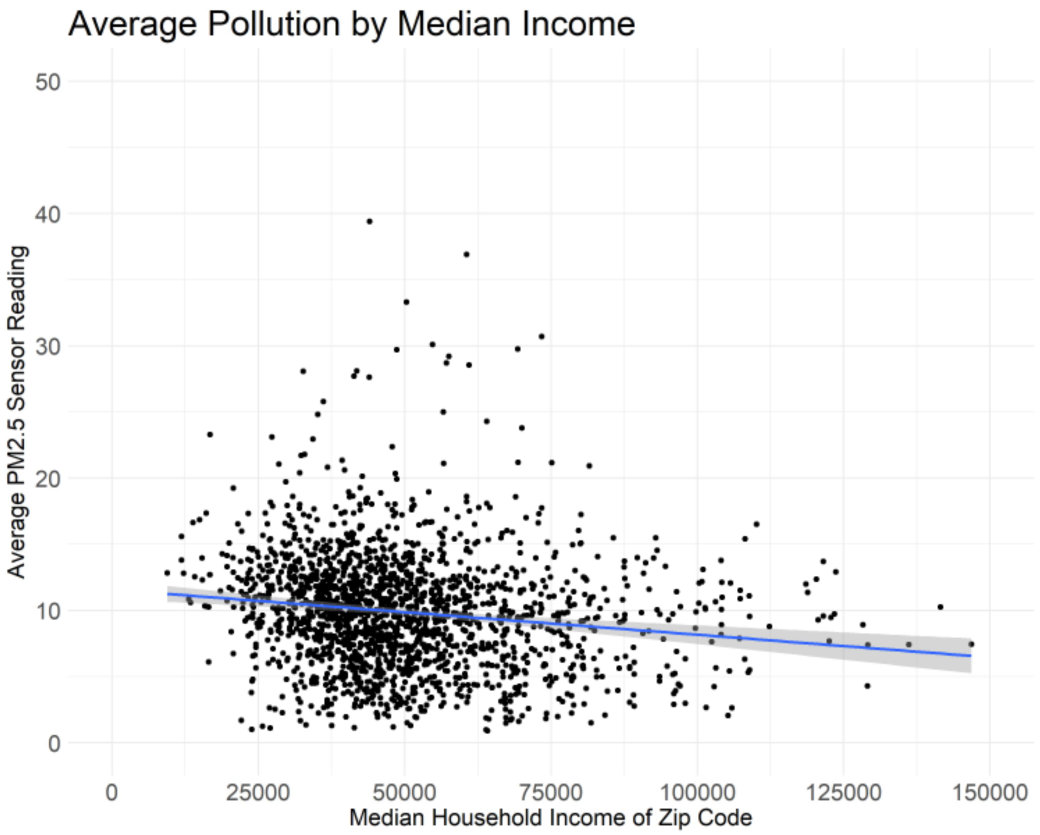


Incorporating Census Data

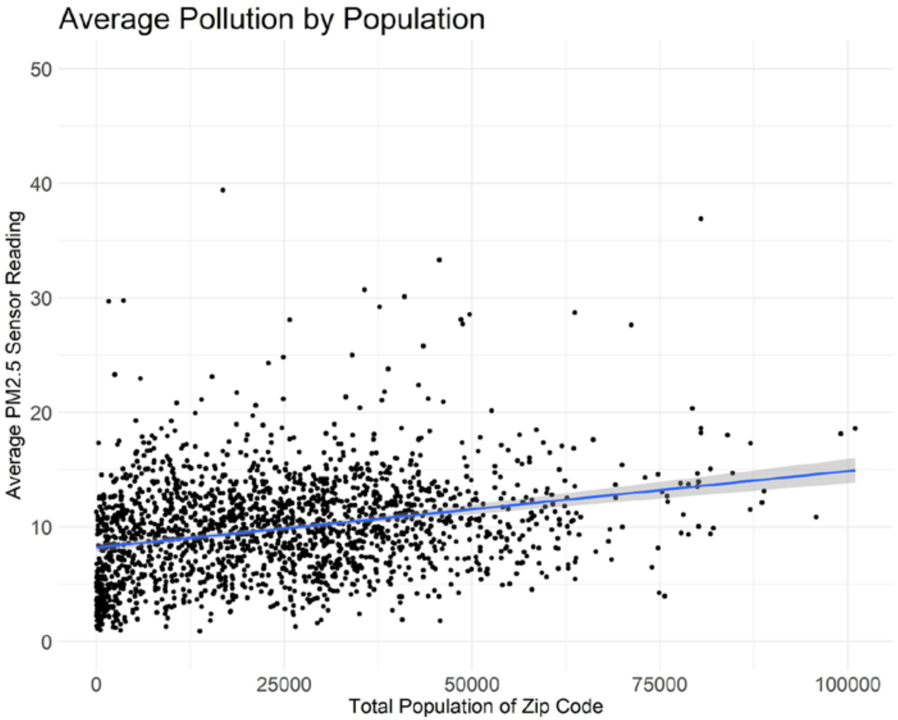
- We were able to obtain two things:
  - The longitude and latitude of each pollution sensor
  - US Census data by Zip Code

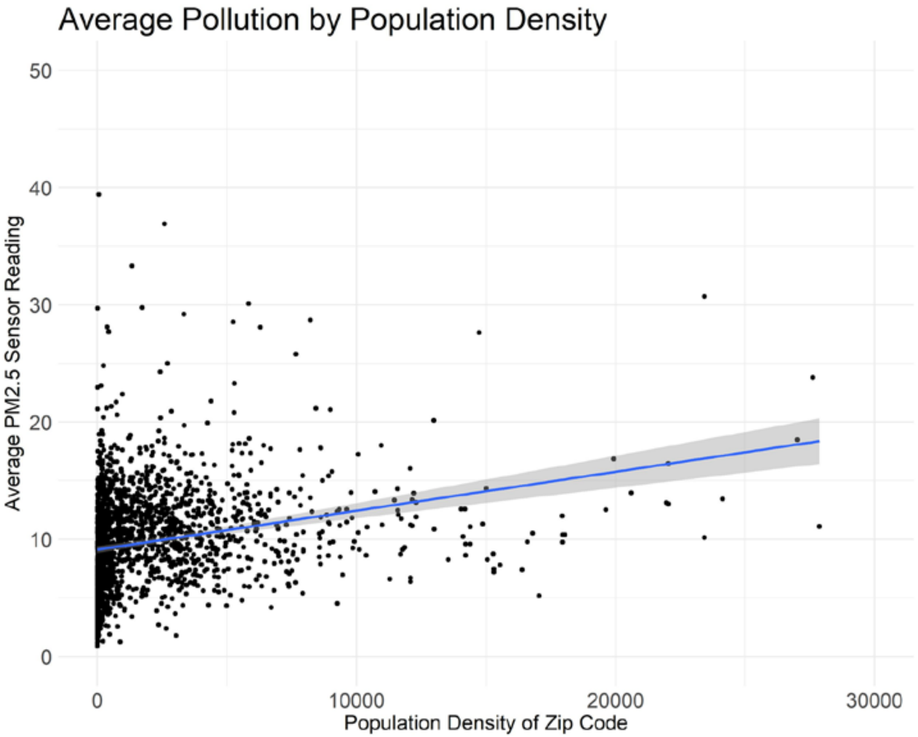
- We then reverse geocoded the sensor locations to find which Zip code they were located in, in order to merge the pollution data with the Census data

Census Data - Relationships



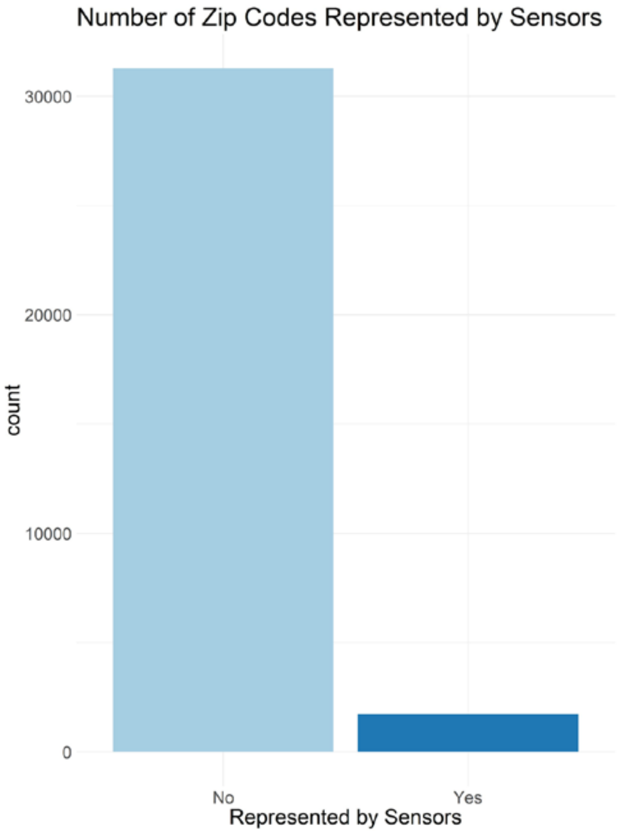
Census Data - Relationships



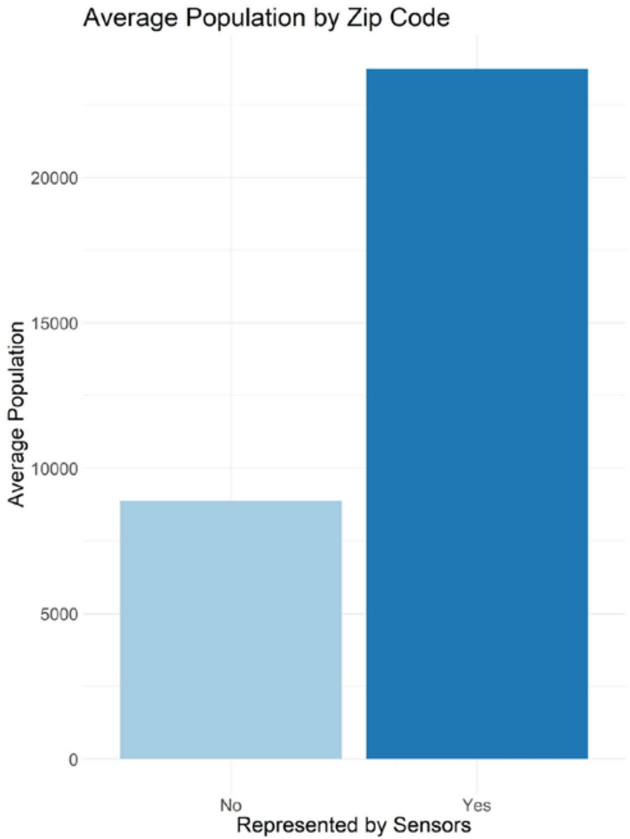


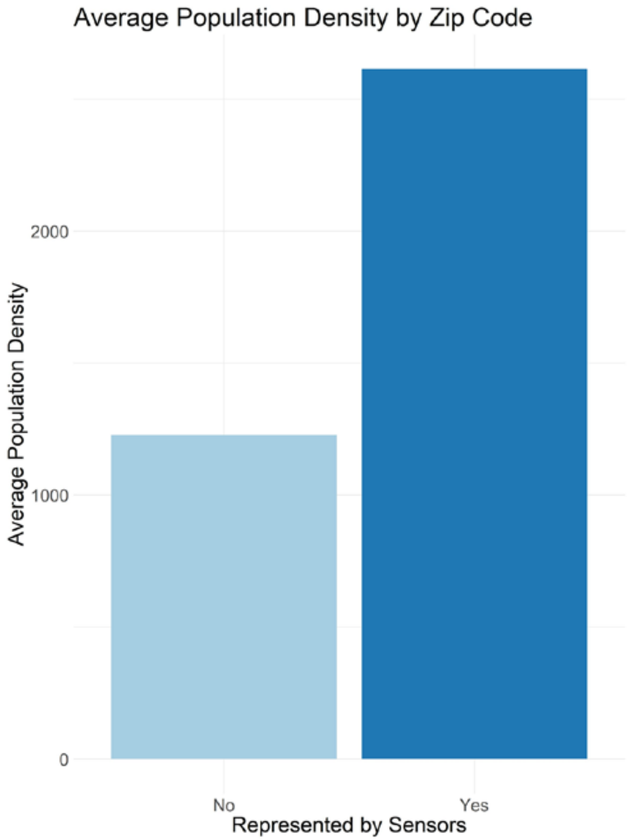
Pollution Sensor Locations

Census Data - Representativeness



Census Data - Representativeness





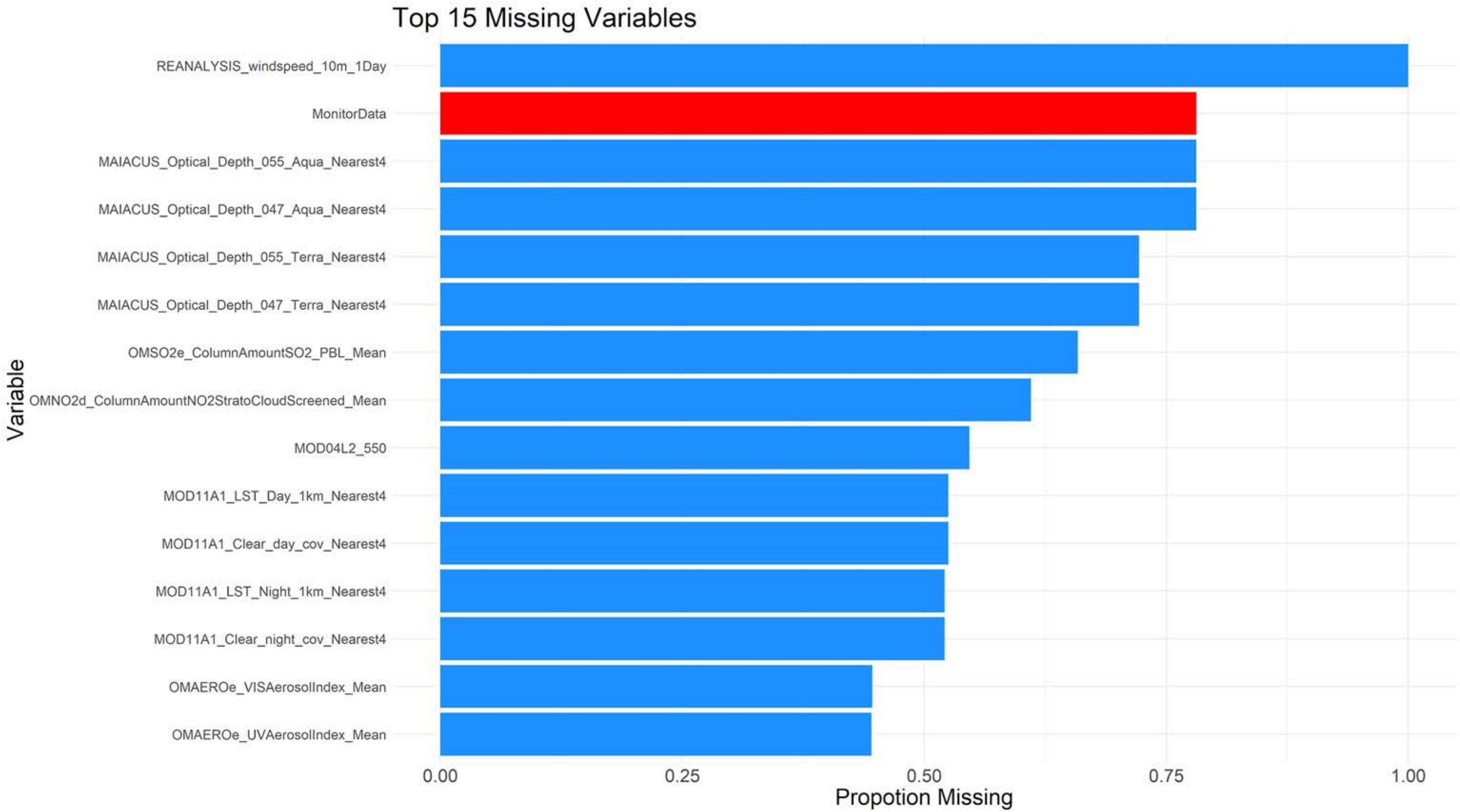
Data Preprocessing

Missing Data

Dimensionality Reduction

Imputation

Missing Data

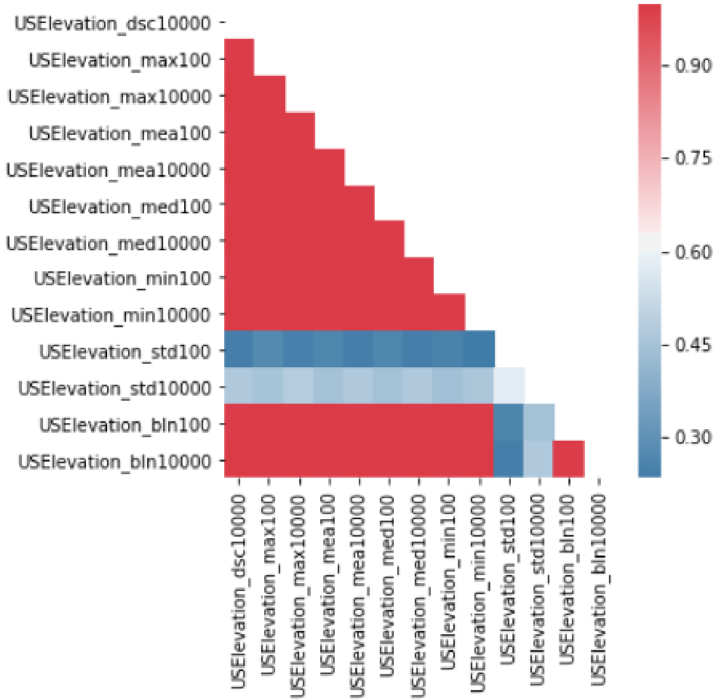


Missing Data

Variable <chr>	Correlation <dbl>
Nearby_Peak2_PM25	0.8571
Nearby_Peak2Lag1_PM25	0.5917
MAIACUS_Optical_Depth_047_Terra_Nearest4	0.4155
MAIACUS_Optical_Depth_055_Terra_Nearest4	0.4062
Nearby_Peak2Lag3_PM25	0.3775
Nearby_Peak2_NO2	0.3409
MAIACUS_Optical_Depth_047_Aqua_Nearest4	0.3316
Nearby_Peak2Lag1_NO2	0.3252
MAIACUS_Optical_Depth_055_Aqua_Nearest4	0.3250
REANALYSIS_hpbl_DailyMean	-0.2924

Dimensionality Reduction

- For each pair of variables with greater than 0.9 correlation, drop one based on:
  - Correlation with response
  - Amount of missingness

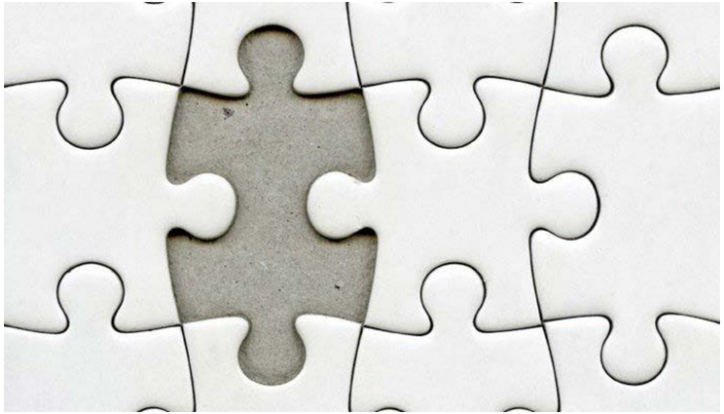


- Allowed us to drop 30 predictors
- 133 predictors remaining

Imputation Methods

- Over 100 predictors with missing values need to be imputed prior to modeling
- Currently, HSPH using linear models with a fixed subset of predictors that have little no missingness

- Our attempts thus far:
  - Mean imputation
  - Iterative random forest imputation



Pollutant Modeling

Setup

Models Tested

Performance

Insights

Improvements

Modeling Setup

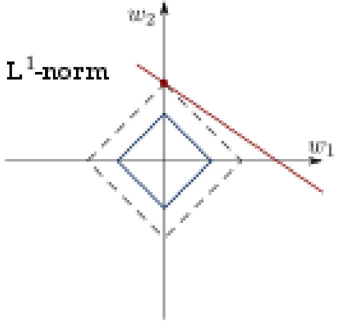
- Randomly split sensors into train/test
  - ~1500 train sensors yields ~12,000 train observations
  - ~300 test sensors yields ~2,500 test observations

- 10-fold cross-validation with train data to tune hyperparameters
  - Optimizing for R^2

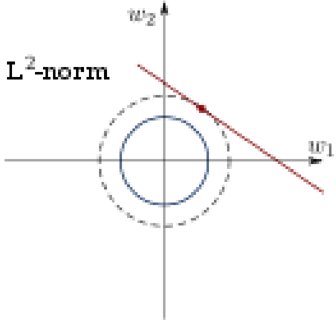
- Compare models' test R^2

Models Tested

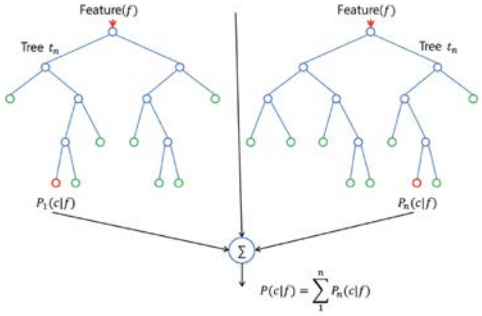
Lasso



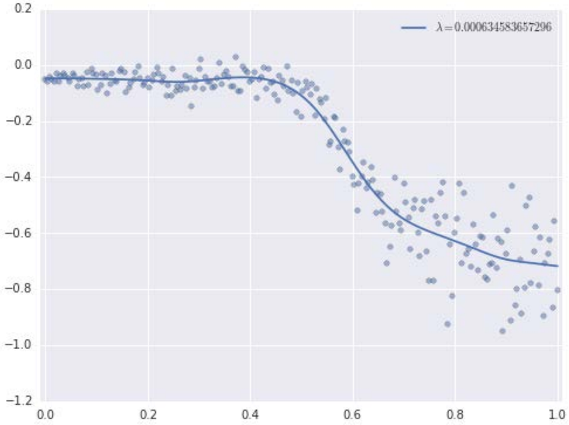
Ridge



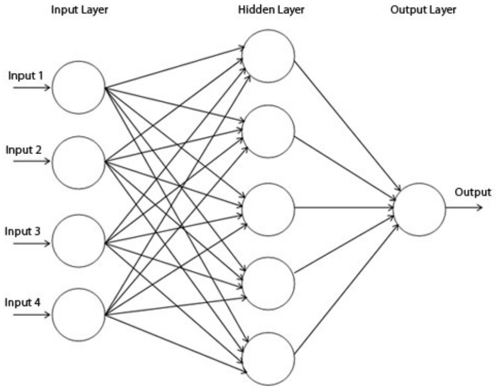
Random forest



Generalized additive model



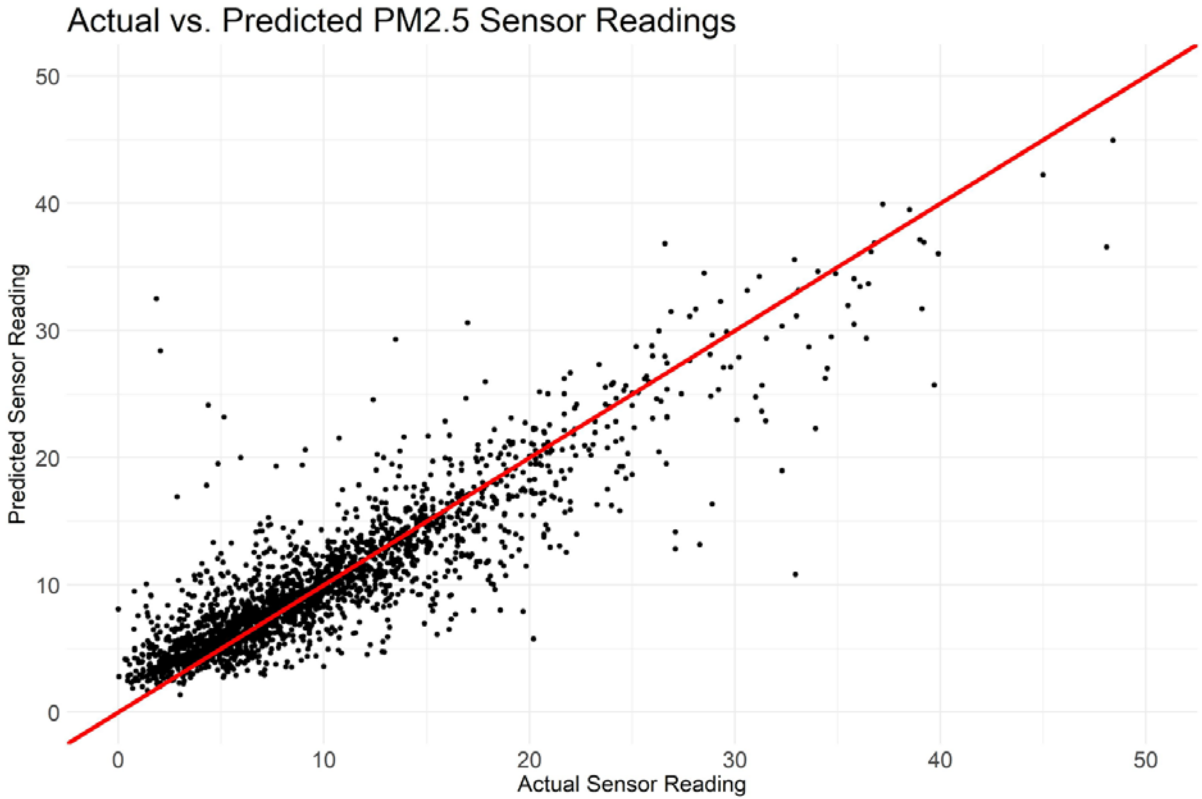
Feed-forward NN



Model Test R^2 Results

Model / Imputation Method	Mean	Random Forest
Ridge	0.765	0.767
Lasso	0.764	0.767
Generalized Additive	0.771	0.772
Random Forest	0.782	**0.787**
Feed-forward NN	0.735	0.74

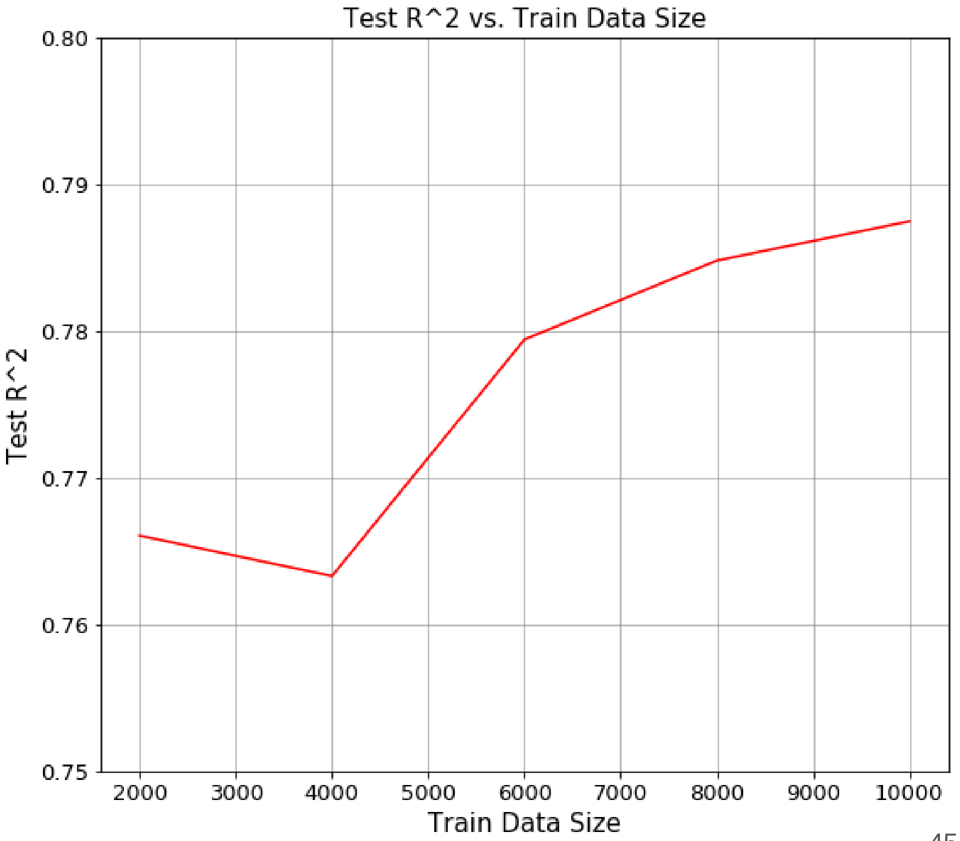
Best Model - Actual vs. Predicted



Modeling Insights

- Important predictors
  - Vegetation of nearby areas

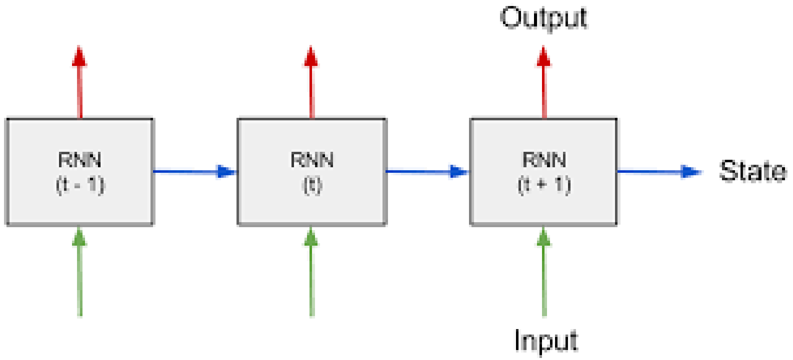
  - Nearby Peak Ozone



Improvements

- Use ‘big data’
  - Evidence to suggest that the model R^2 will improve by using more data
  - Info within sensor sequence can help for imputation and modeling

- Model complexity
  - RNNs and CNNs can learn complex temporal relationships within sensor sequences
  - CNNs can learn complex spatial relationships between nearby sensors



Looking Ahead

Deliverables

Nice to Have’s

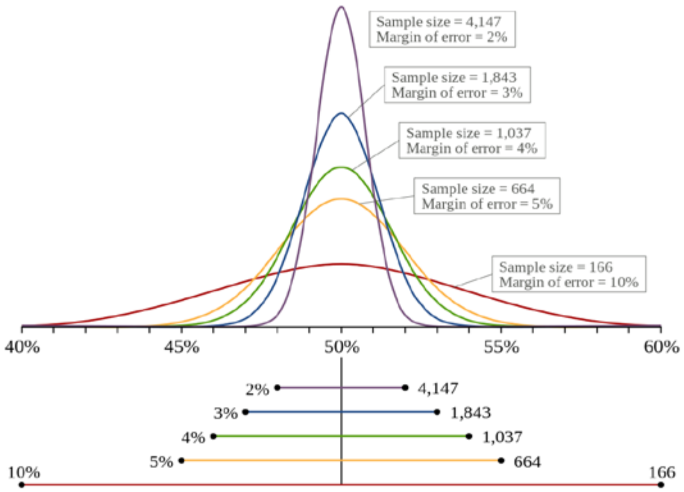
Timeline

Deliverables

1. Optimal imputation software for overcoming serious missingness
2. Improved predictive model, incorporating auxiliary data
3. Extensible package for HSPH to use moving forward

Nice to Have's

1. Uncertainty quantification of model across map
2. Considerations of optimal new sensor locations



Timeline

- **April 1:** Synopsis of Model Improvements
- **April 30**: Model Accuracy Report
- **May 6:** Packaged SW & Documentation

Thank you!