IACS Institute for Applied Computational Science
HARVARD SCHOOL OF ENGINEERING AND APPLIED SCIENCES

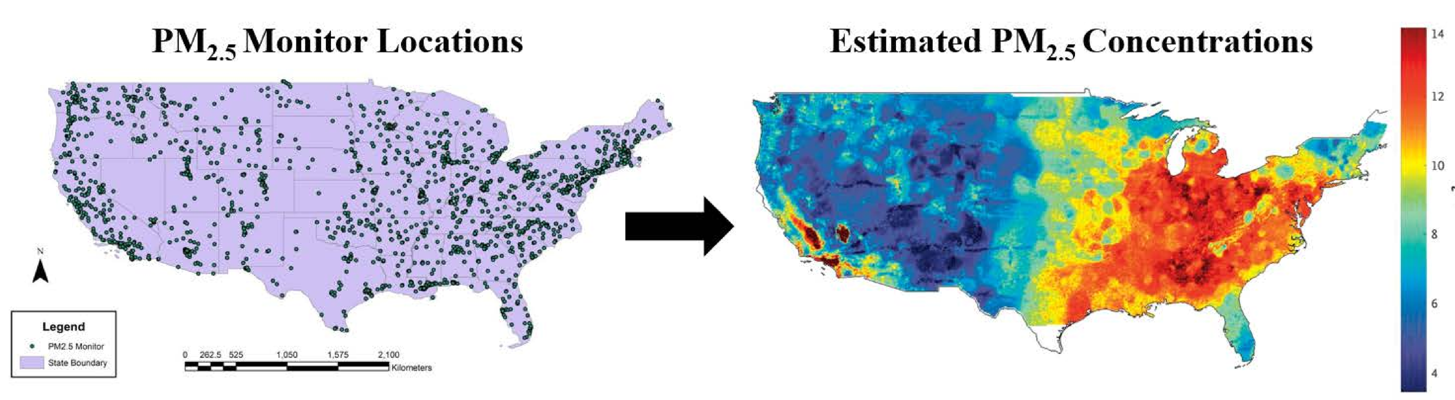
Modeling Air Pollution

HARVARD SCHOOL OF PUBLIC HEALTH

Members: Keyan Halperin, Christopher Hase, Justin Lee, Casey Meehan **Mentor:** David Sondak **Partner:** Harvard School of Public Health

Background

$PM_{2.5}$ – fine particles with a diameter of 2.5 μm or less – is measured at approximately 2,000 air pollution monitors located throughout the US. However, these monitors are costly to operate and thus, sparsely distributed. Consequently, the air quality is not known for many locations throughout the country. This is problematic for public health studies since we cannot estimate the overall effect that pollution has on health if we do not know what the pollution is in many areas with any degree of certainty.

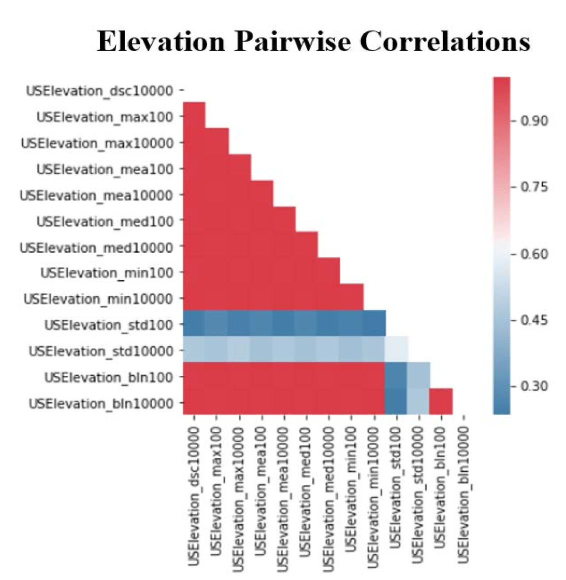


Thus, using pollution data from these sparsely distributed sensors along with several other types of data, our goal was to fit a model that accurately interpolates pollution throughout the US. With this, it would be possible to create of a continuous map of pollution on the daily level, which could be used by researchers to establish causal relationships between pollution and health outcomes, among several other potential research applications.

Data

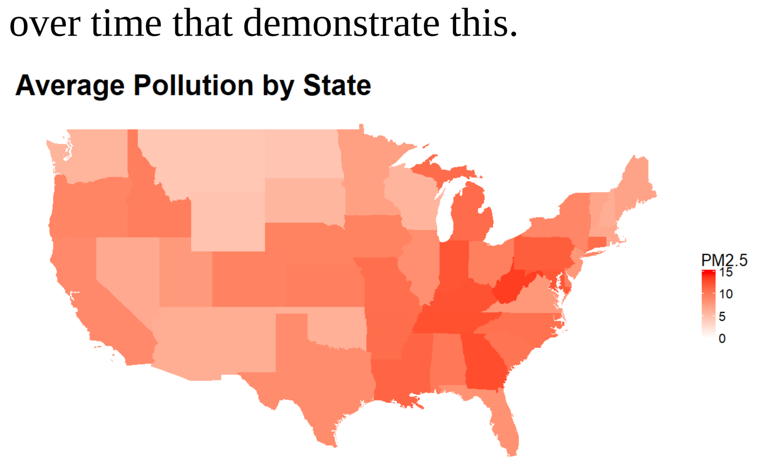
Monitor Data	Pollution readings from approximately 2,000 $PM_{2.5}$ sensors each day for the past 16 years.
Satellite Data	Various meteorological, geographical, and topographical variables collected from satellites and other instruments.
US Census Data	Demographic variables on the ZIP Code level as collected by the US Census.

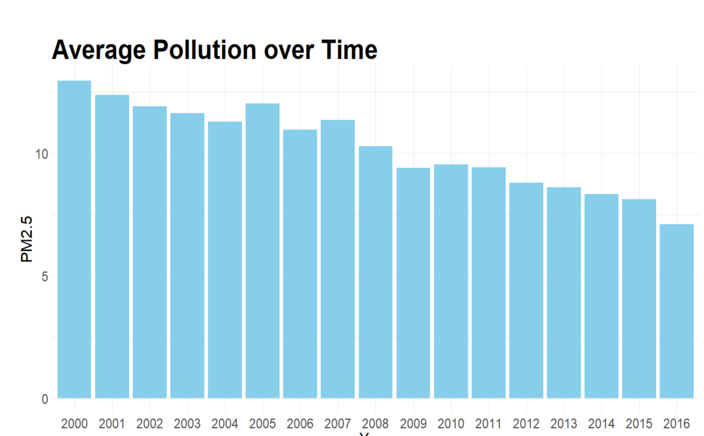
Because of the high number of variables, we decided to perform variable selection. Having extraneous or redundant variables not only increases the computation time, but it can also harm the predictive performance of machine learning models. To reduce dimensionality, for each set of highly correlated features, we removed all but one. We also removed variables that had non-significant partial correlations with $PM_{2.5}$ after controlling for nearby $PM_{2.5}$. On the right you can see a set of variables for which there were many high pairwise correlations.



Exploratory Data Analysis

We also conducted a thorough exploratory data analysis to gain insights that would help us in modeling. Most notably, we discovered that $PM_{2.5}$ has strong spatial and temporal dependencies. Below are plots of average $PM_{2.5}$ levels by state and average $PM_{2.5}$ levels over time that demonstrate this.

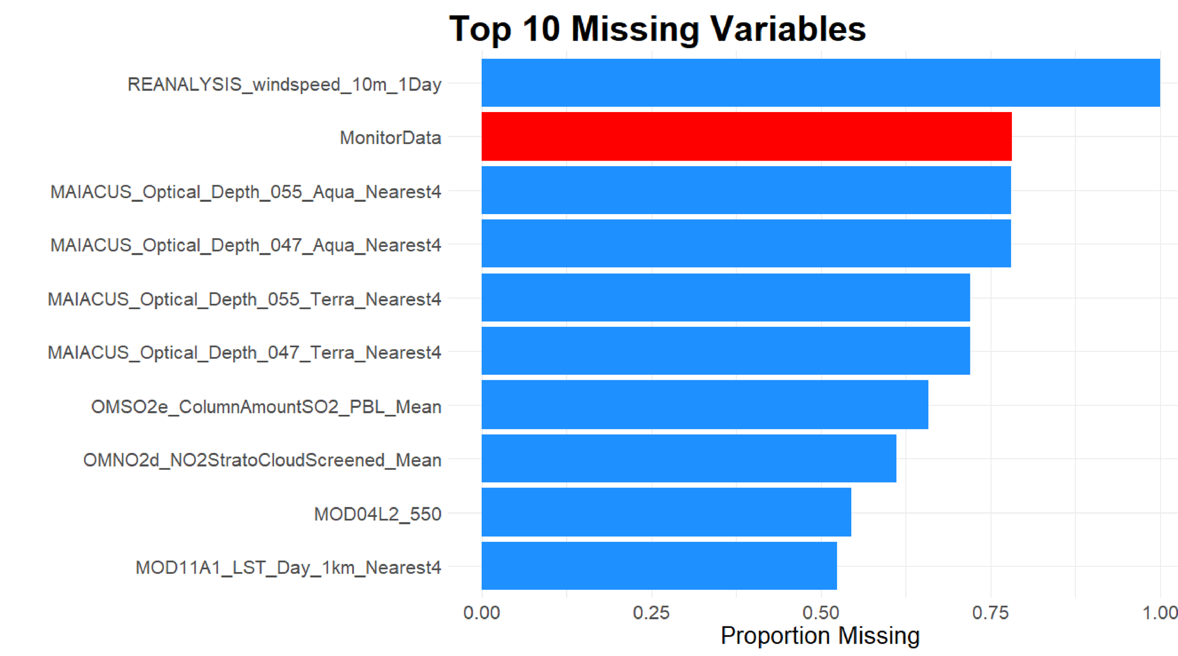




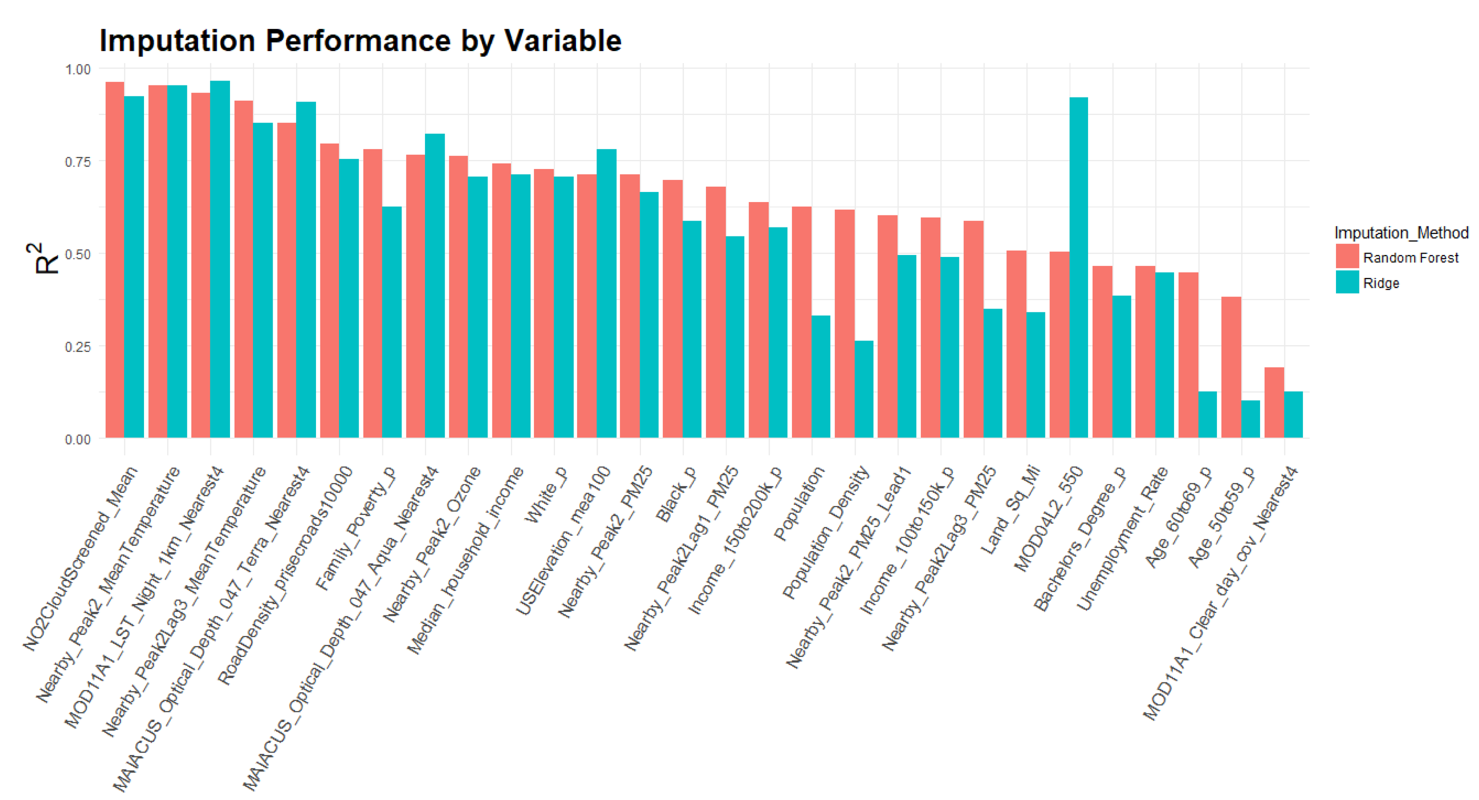
We also explored the relationship between each feature and $PM_{2.5}$. By far, the feature that had the highest correlation with $PM_{2.5}$ was nearby $PM_{2.5}$, with a correlation of about 0.857. Thus, there is strong evidence to suggest that we should implement models that take into account both spatial and temporal relationships.

Imputing Missing Data

A major obstacle that we had to overcome was the amount of missing data. Many of the variables and almost all observations had some missing data. Even $PM_{2.5}$ itself was often missing since the monitors do not measure pollution each day. Below is a graph of the top 10 most missing variables with $PM_{2.5}$ monitor data highlighted in red.

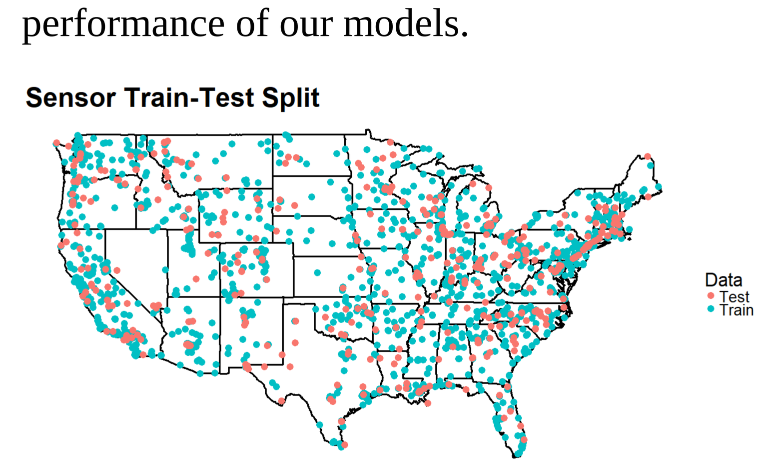


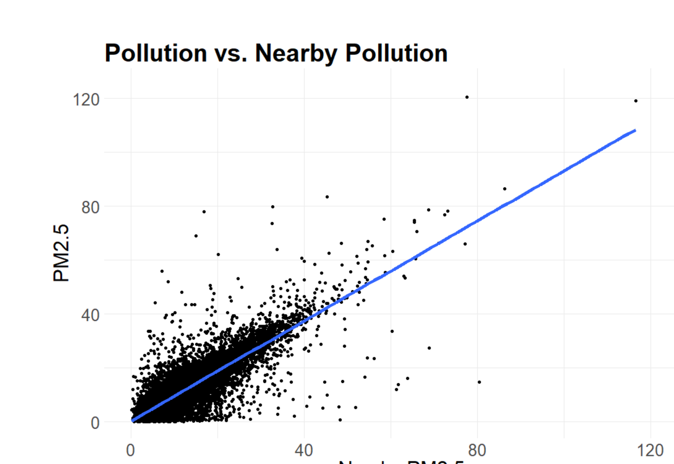
Because of the volume of data that was missing and its potential importance for modeling $PM_{2.5}$, we could not afford to ignore it. After some research and experimentation, we decided to implement an iterative imputation algorithm called MissForest (Stekhoven & Bühlmann, 2012). Allowing for other supervised learning models to be used instead of random forest enabled us to compare two MissForest variations – one that used random forests as in the original paper and one that used ridge regressions. The scheme we created for evaluating the quality of our imputations indicated that our imputations were accurate. The random forest variation achieved a weighted R^2 of 0.717 and the ridge variation achieved a weighted R^2 of 0.774. Below is a plot of the imputation performance of both variations of this algorithm for each variable as measured by R^2.



Baseline Model

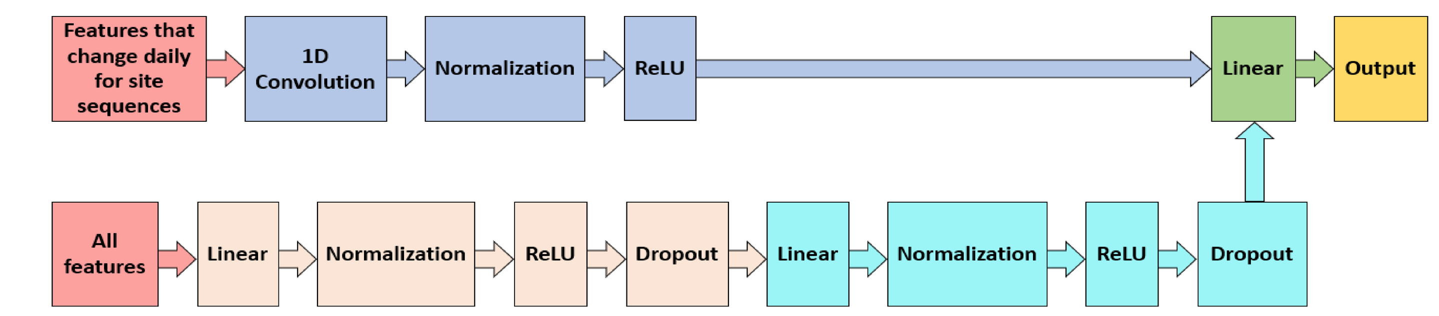
After pre-processing and imputation, we tuned our models using K-fold cross-validation on a random 80% of the sensor site sequences and reserved the remaining sensors to test the performance of our models.





Because of the very high correlation between $PM_{2.5}$ and nearby $PM_{2.5}$, we decided that our baseline model should be a simple linear regression with nearby $PM_{2.5}$ as the only feature. Despite its simplicity, this model performs surprisingly well with a test R^2 of 0.712. Above is a scatterplot of the relationship between $PM_{2.5}$ and nearby $PM_{2.5}$ on a random 1% subset of the data.

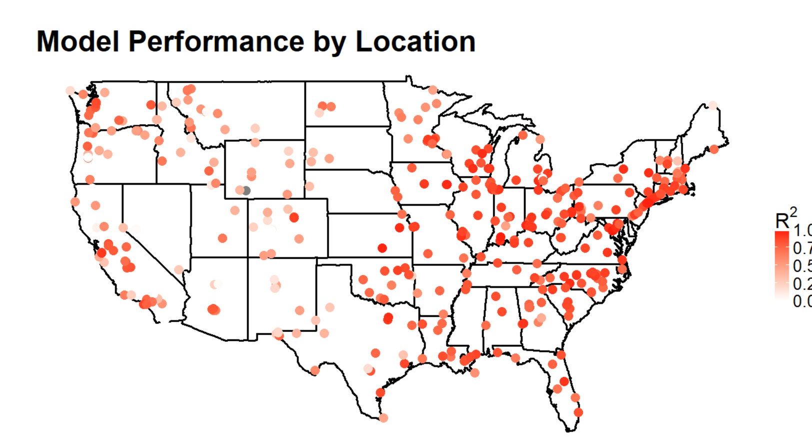
Modeling



Because consecutive days are likely to be related in ways that are relevant for predicting pollution, we decided to implement a CNN. In particular, we used a kernel width of size 3 so that features from the previous day, current day, and the following day are used for predicting pollution for the current day. We are still in the process of tuning the CNN, but our best result so far is an R^2 of 0.775.

	OLS	Ridge	RF	XGBoost	CNN	Ensemble
R^2	0.712	0.733	0.780	0.776	0.775	**0.784**

We also implemented various standard machine learning models to compare their performances. Although random forest outperformed the other models, our best results came from an ensemble of random forest, XGBoost, and CNN.



On the left is the performance of our ensemble by location as measured by R^2. Unsurprisingly, the ensemble produces much more accurate predictions in regions with more sensors. This highlights the need to install more monitors in locations where monitors are sparse.

Discussion

Because of the disproportionate importance of nearby $PM_{2.5}$, we believe that it is absolutely essential for more pollution monitors to be installed, especially in regions where there are few. We also have evidence to suggest that our imputation procedure provides high quality imputations, and since the procedure is quite easy to implement, we recommend that HSPH use it in the future.

Going forward, in addition to continuing to improve model performance, we believe that an important task is to quantify the uncertainty of the $PM_{2.5}$ predictions. This would be important because (1) it would give more insight into model performance in areas that are far away from sensors (2) it would allow for more accurate variance estimates of any associated causal effects and (3) it could be used in determining which locations should be prioritized for new sensor installations. Gaussian process regression may be a good tool for quantifying the prediction uncertainty.

Software Package

The scripts written for this project were all run on Harvard's Odyssey Research Computing Cluster, and instructions for running our code were written with Odyssey in mind. This is because of the substantial amount of RAM and processing power needed to load and fit models on the nearly 20 GB of pollution data. Python and R scripts for pre-processing and imputation as well as model validation, training, and testing were wrapped in Bash scripts formatted to be scheduled using Slurm. All code was made available on a GitHub repository so that we could turnover our work to HSPH for use and continued development.