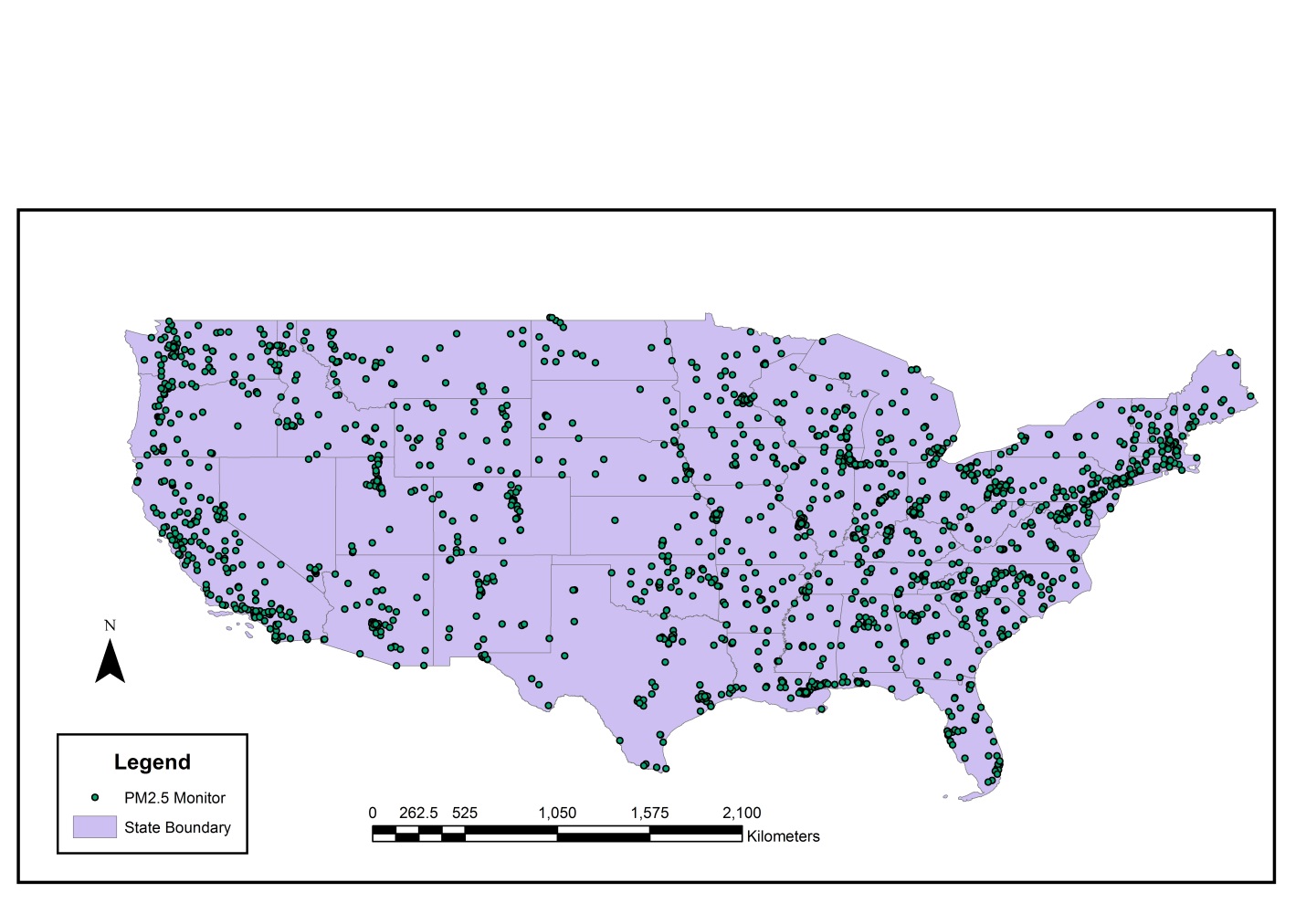
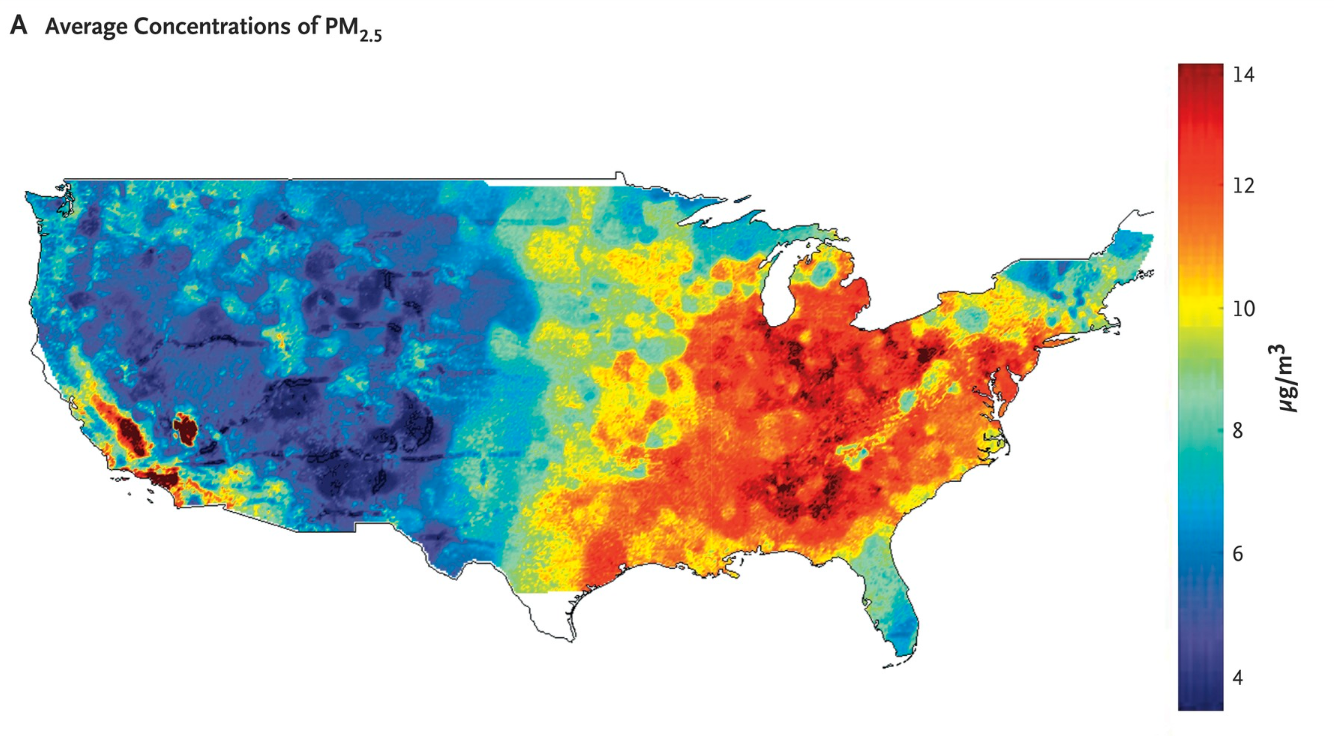
**Introduction**

PM2.5 - 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.



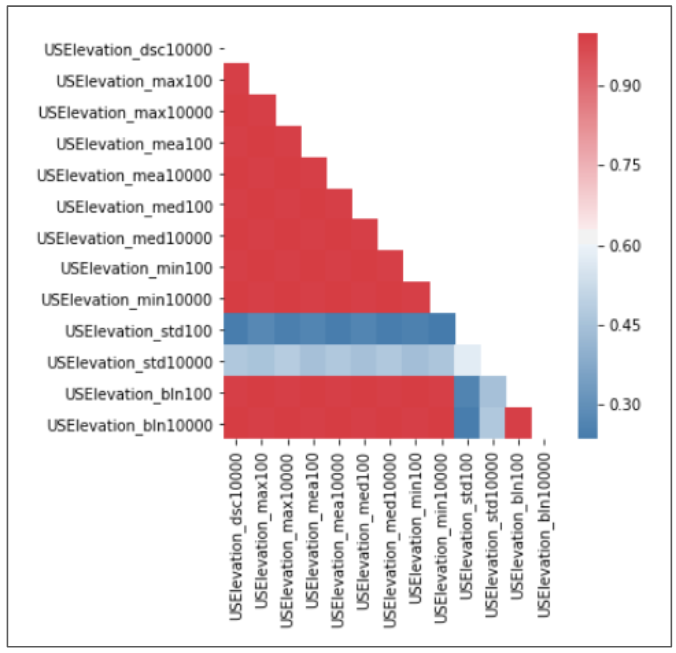
Thus, using pollution data from these sparsely distributed sensors along with several other types of data, our goal was to fit a model that accurately interpolates pollution throughout the entire US. With this, it would be possible to create of a continuous map of pollution on the daily level. This map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



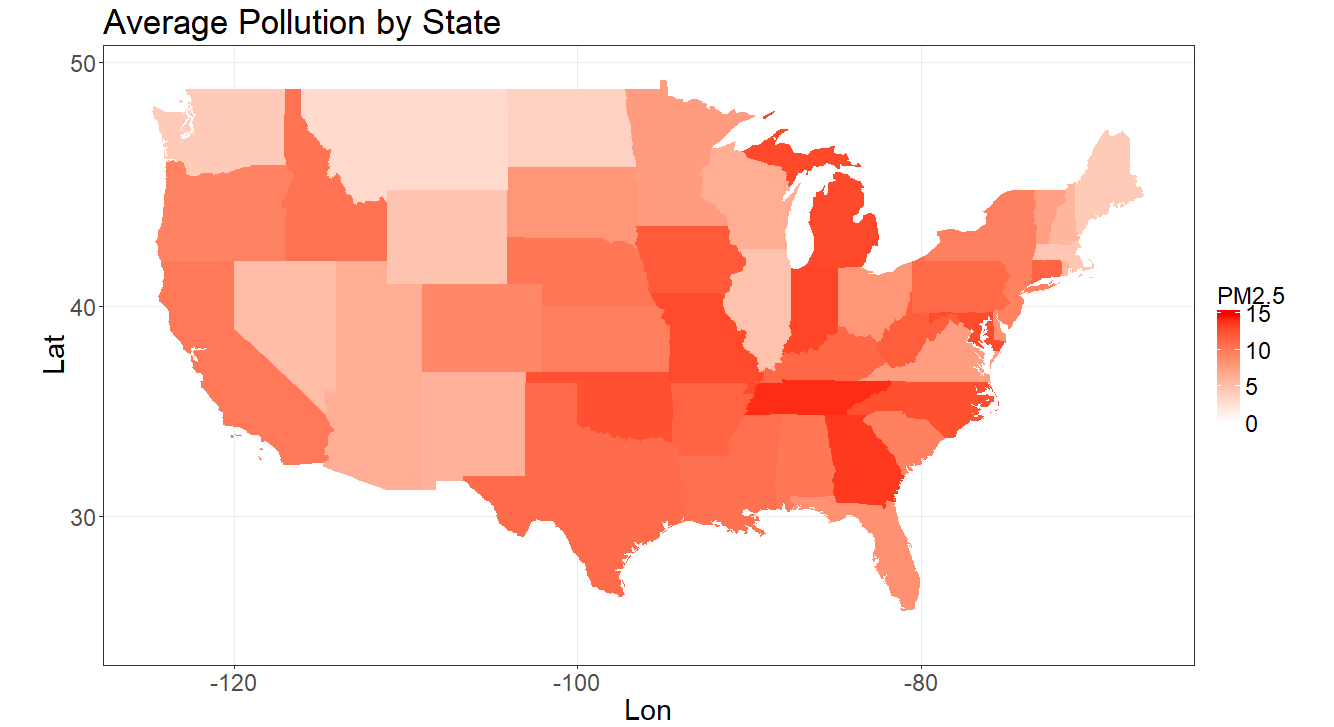
**Data**

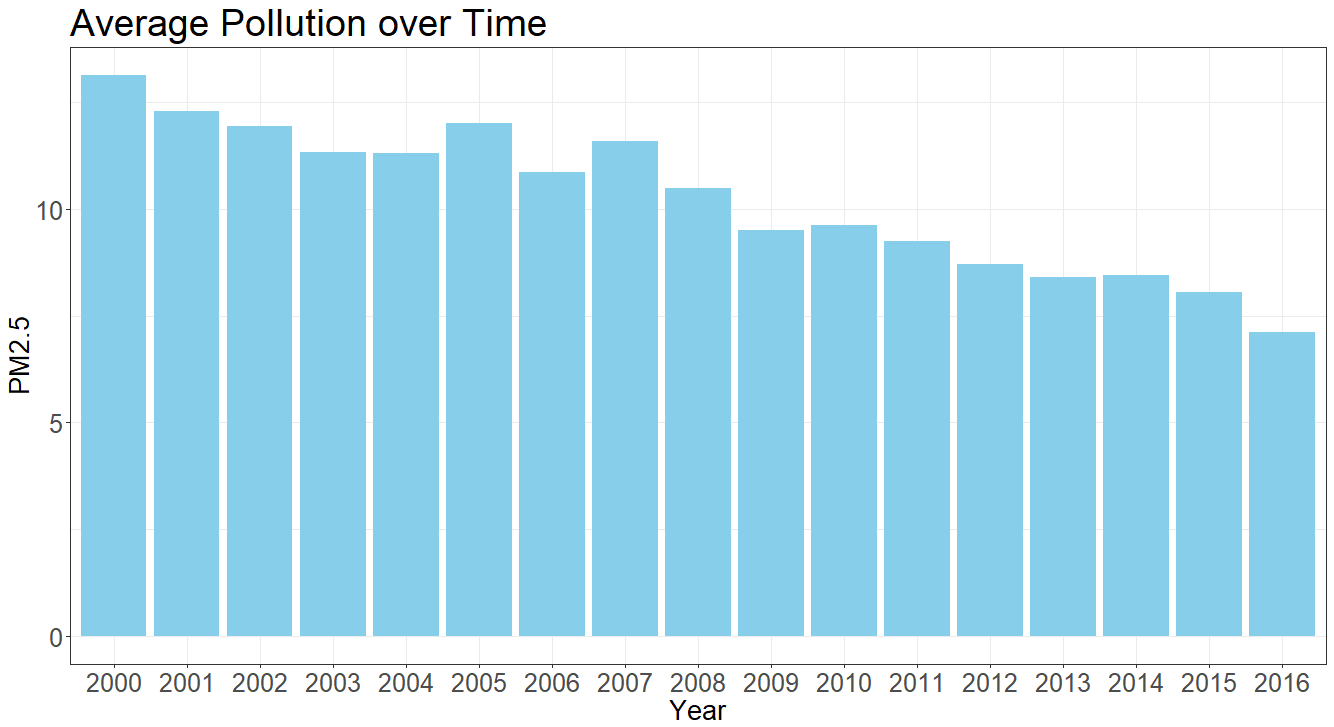
In order to model PM2.5, we used satellite, meteorological, land coverage, and nearby pollution related variables as features. We also incorporated of zip code level census data in and created several features that we thought might be useful in estimating PM2.5. In total, to begin with, we had just under 200 variables for each sensor each day for the past 16 years, giving us approximately 13 million rows of data (this amounted to about 20 GB).

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 (correlation > 0.9) features, we removed all but one. We also removed variables that had non-significant correlations with PM2.5 after controlling for nearby PM2.5 – the most important predictor of PM2.5. Below you can see a set of variables for which there were many high pairwise correlations.



We also conducted a thorough exploratory data analysis to gain insights that would help us in modeling. Most notably, we discovered that PM2.5 has strong spatial and temporal dependencies. Below are plots of average PM2.5 levels by state and average PM2.5 levels over time that demonstrate this.

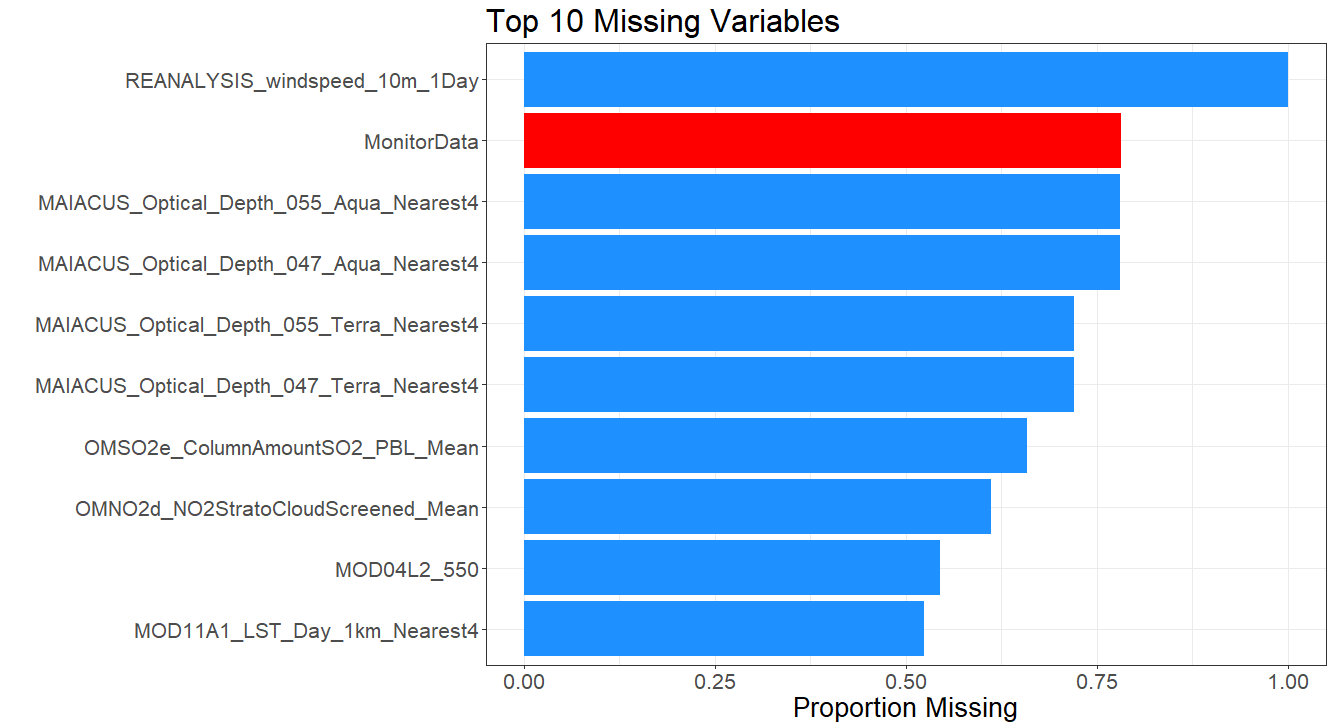




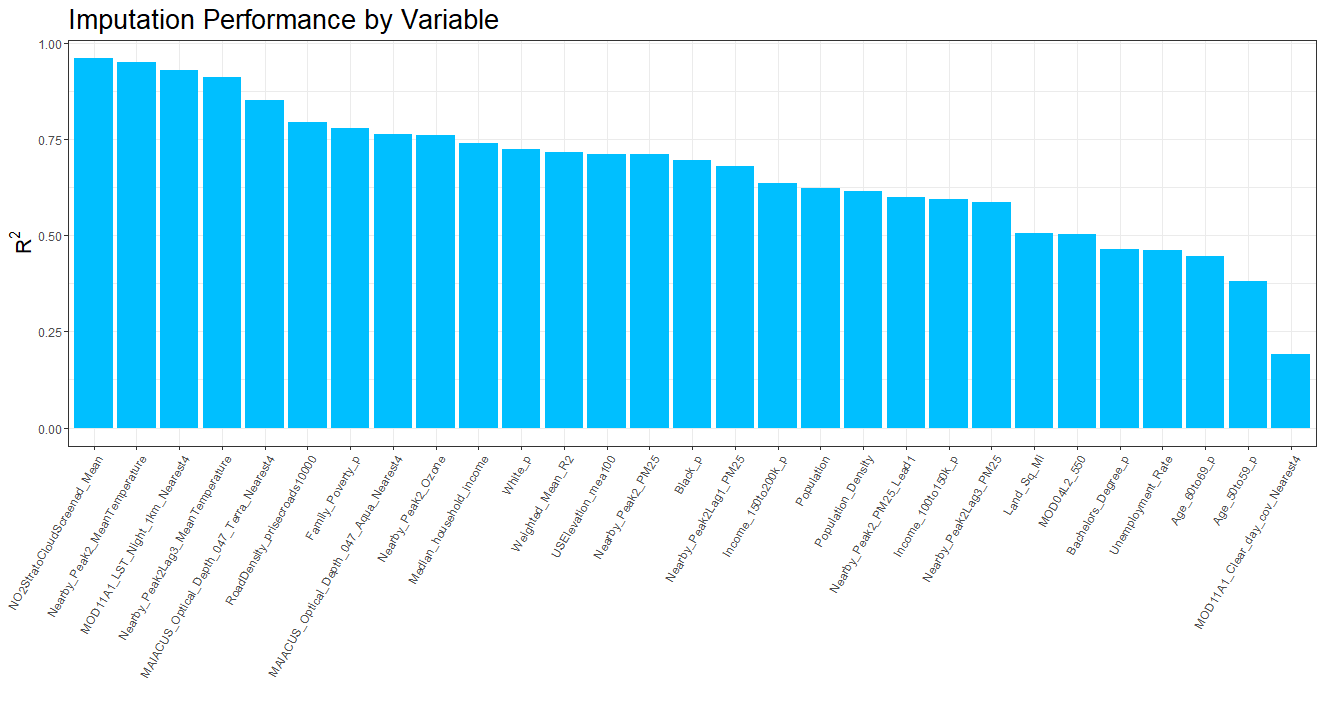
Thus, there is strong evidence to suggest that we should implement models that incorporate take into account spatial and temporal relationships.

**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 PM2.5 itself was often missing since the monitors do not collect pollution data every day.

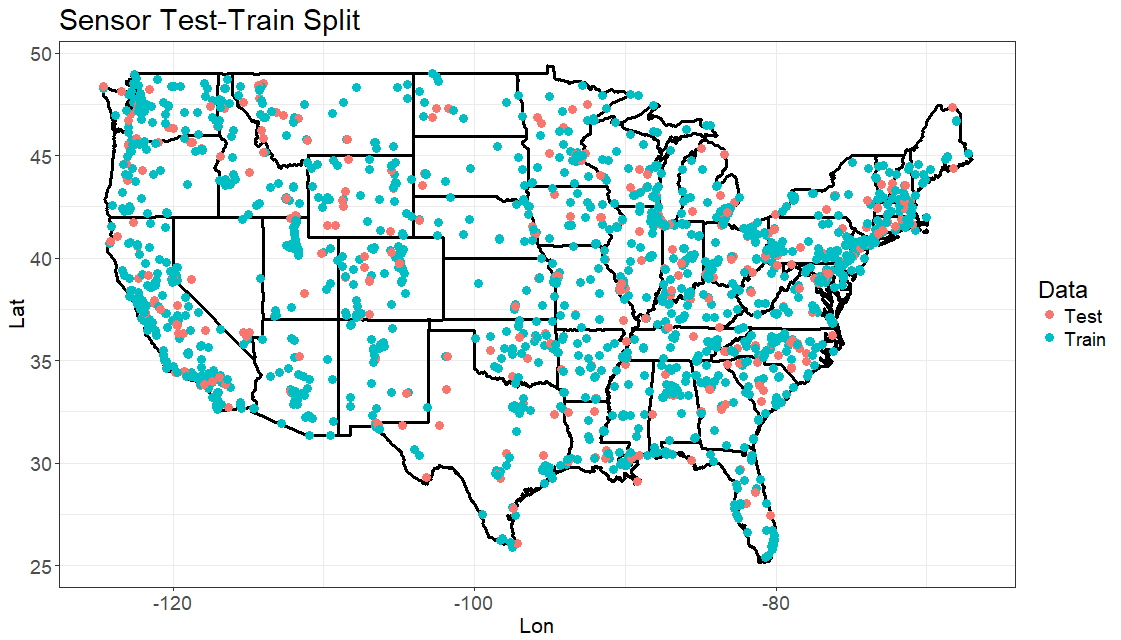


Because of the volume of data that was missing and its potential importance for modeling PM2.5, we could not afford to ignore it. After some research and experimentation, we decided to implement an iterative imputation algorithm called MissForest (Stekhoven & Bühlmann, 2012). After some modifications to the algorithm and creating a scheme for validating the quality of our imputations, we obtained accurate imputations. Below is a plot of the imputation performance for each post variable selection feature with missing values as measured by R2. **(we will go into more detail on MissForest for final draft)**



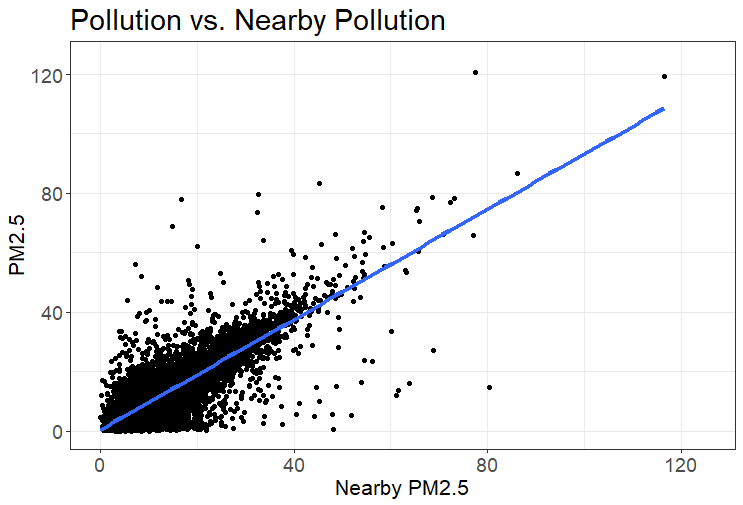
**Modeling**

After pre-processing and imputation, we proceeded with the model tuning, training, and testing process. We trained our models on a random 80% sample of sensors, and we reserved the remaining sensors to test the performance of our models. We also tuned all hyper-parameters using K-fold cross-validation.



**Linear Regression**

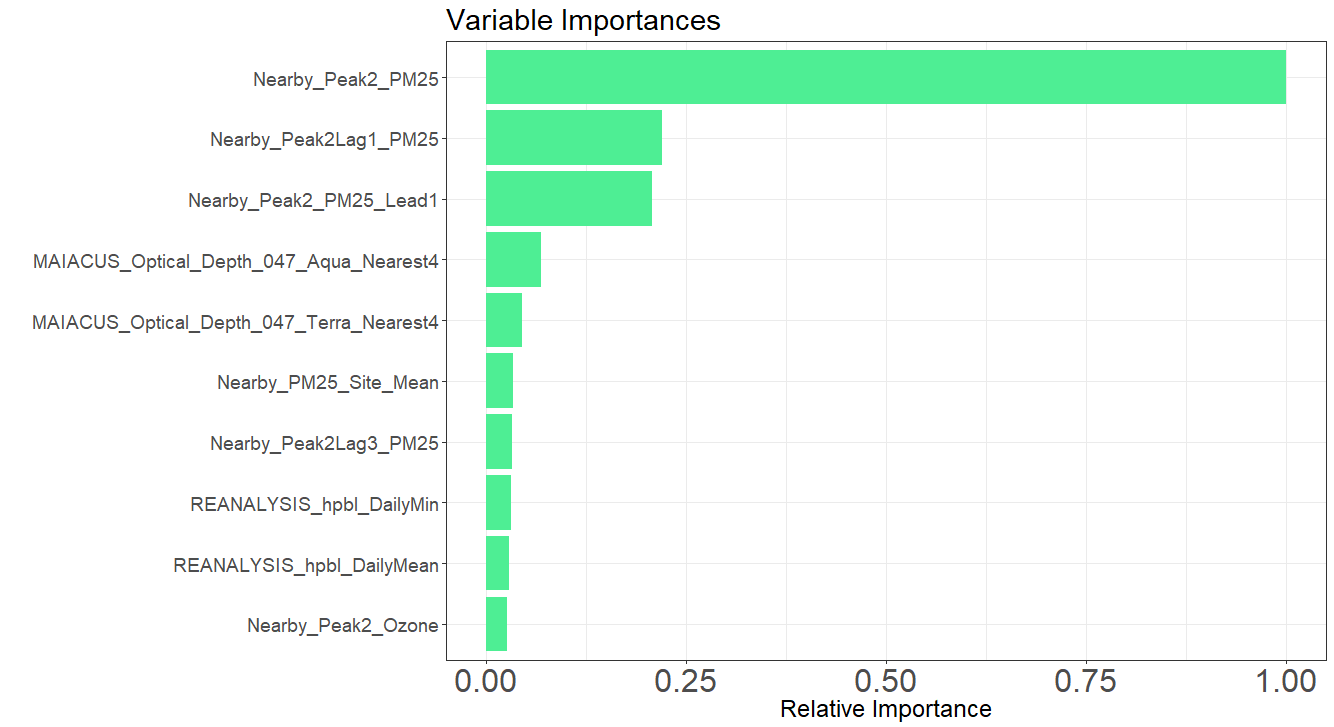
Based on the correlations between each of the predictors and PM2.5, by far the most important linear predictor of PM2.5 was nearby PM2.5. Thus, our baseline model was just a simple linear regression with nearby PM2.5 as the only predictor. Despite its simplicity, this model does surprisingly well with a test R2 of 0.712. Below is a scatterplot of the relationship between PM2.5 and nearby PM2.5 on a random 1% subset of the data.



We then decided to use ridge regression with all of our selected features since including all of our predictors in an ordinary least squares model may have led to overfitting. This noticeably improved our performance on the test set with an R2 of 0.733.

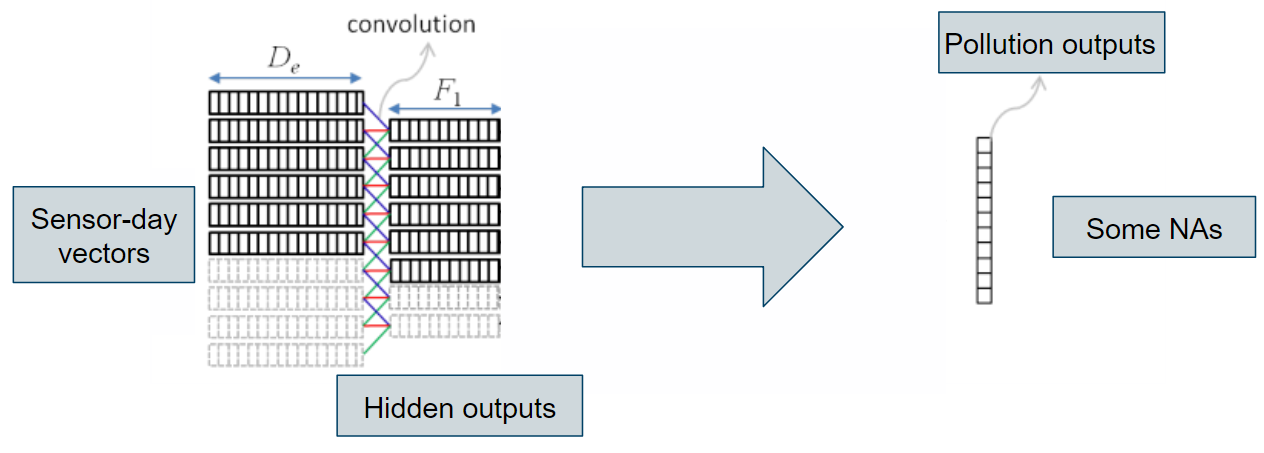
**Tree-Based Methods**

We then decided to implement some tree-based methods - in particular, random forest and XGBoost - since these methods implicitly account for non-linear relationships and interactions between variables without explicitly specifying them, and they often have strong prediction performance. Below are the feature importances for our tuned random forest model. As we had previously discovered, nearby PM2.5 in all of its variations are disproportionately important predictors of PM2.5. Both random forest and XGBoost performed well with an R2 of 0.780 and 0.776, respectively.



**Neural Networks**

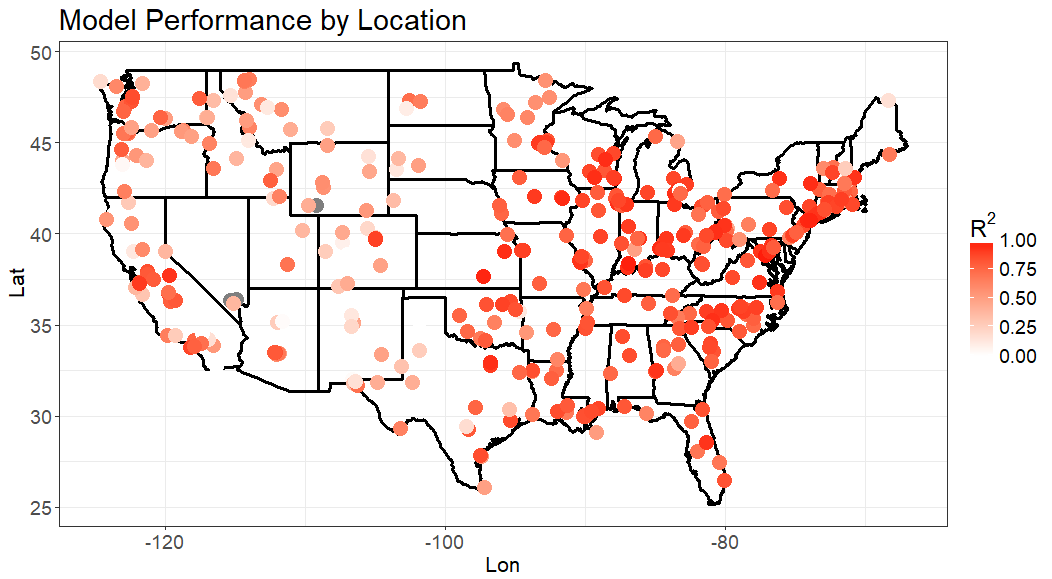
Because days that are close together are likely to be related in ways that are relevant to pollution, we decided to implement a CNN. In particular, we used a kernel width of size 3 so that features from the previous day, current day, and day after are used for predicting pollution on any given day. After tuning the relevant parameters, we obtained an R2 of x, giving us our best results! **(still getting final results, will go into more detail for final draft)**



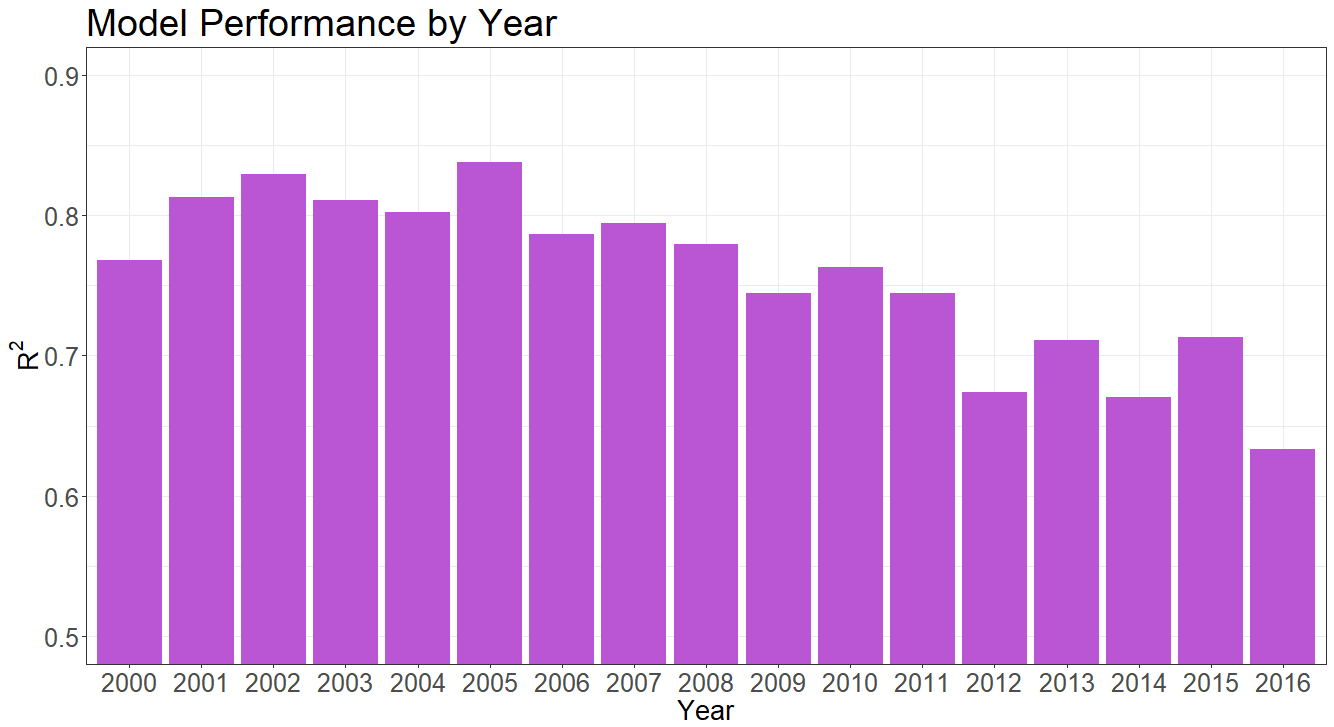
**[Insert table comparing model performance]**

**Model Diagnostics**

Below is the performance of our best model by location. Unsurprisingly, our model has much better predictive performance in regions with more sensors nearby. This highlights the need for to have more monitors installed.



Despite our best efforts, the accuracy of our predictions seemed to be inconsistent over time. In general, the performance of the model as measured by R2 appears to be decreasing over time.



**Discussion**

One result we have consistent throughout our analyses and modeling is the disproportionate importance of one variable, nearby PM2.5. Despite having many additional variables, the strong performance of our imputation procedure, and the complex models that we implemented, we were unable to improve model performance substantially beyond a model that uses nearby PM2.5 as its only feature. We think that this is because the other available data is just not particularly predictive of PM2.5 after having already taken into consideration nearby PM2.5. Consequently, we believe that in order to obtain accurate predictions of PM2.5 throughout the US, it is absolutely essential for more pollution monitors to be installed, especially in regions where there are few.

We have evidence to suggest that our imputation procedure provides high quality imputations. The procedure is quite easy to implement, so we recommend that HSPH consider using it in the future.