Predicting Boston Air Quality

Kaela Nelson

***Abstract*—Exposure to air pollution is a significant concern, especially in urban areas. Despite this growing issue, most major cities have four or fewer active air quality sensors. Previous studies have implemented geostatistical models using traffic count, elevation, and land cover as variables to predict pollutant levels with high accuracy. In this project, I trained geospatial and spatio-temporal models on three criteria pollutants in order to predict Boston air quality and classify land use types that contribute to poor air quality.**

I. Introduction

AIR pollutants originate from multiple sources, some anthropogenic and others from reactions in the atmosphere itself. The three criteria pollutants I focus on are particulate matter 2.5 ($PM_{2.5}$), nitrogen oxide (NO_2), and sulfur dioxide (SO_2). These pollutants are monitored and regulated by the Environmental Protection Agency (EPA). The pollutants' concentrations are measured in parts per million (ppm).

$PM_{2.5}, NO_2$, and SO_2 are mainly formed from the combustion of fossil fuels (i.e. traffic and power plants)[1]. Additional sources include particle emissions from construction and demolition, and the physiochemical transformation of gases in the atmosphere[2]. Findings from the American Heart Association indicate that $PM_{2.5}$ contributes to worsened cardiovascular health[3,7]. Both NO_2 and SO_2 are particularly harmful for those suffering from respiratory illnesses.

Modeling these pollutant concentrations can educate the public and provide more insight on how to mitigate the concentration levels of future pollutants.

II. Problem Statement

Although air pollution is a growing concern, there are only four EPA air quality sensors located within the Greater Boston area, as shown in figure 1. These sensors do not provide enough coverage over the city. In addition, well formatted data sets containing features such as land use, weather, traffic etc. are not free. As a result, I organized a data set that involved extensive cleaning and feature engineering.

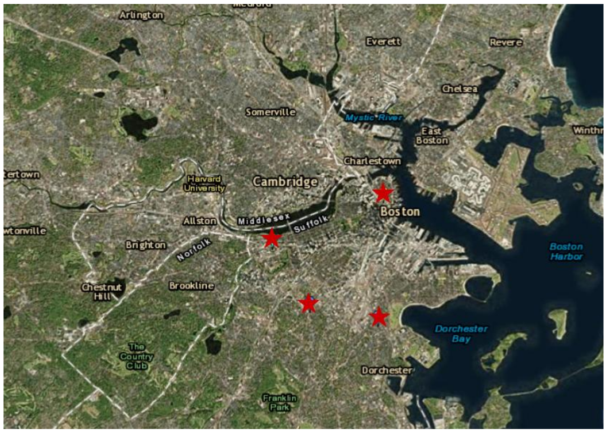

Fig. 1. Locations of EPA Air Quality Sensors in Greater Boston

My goals for the project were to predict Boston air quality and classify land use types that most contribute to high pollutant concentration. I trained land use regression, time series, and Gaussian process models to predict air quality. I implemented random forest, SVM, and logistic regression methods to classify significant land use types. In order increase accuracy, I trained the land use, Gaussian process, and classification methods on data points (site locations) from 1,948 EPA monitoring sites throughout 16 US states (398 counties). Note that these sites are primarily in urban areas to remain consistent with Boston's geographical features. Alternatively, I trained the time series model only on the four Boston sites, as it doesn't make sense to include other locations in a Boston time series model.

III. Data Collection

I collected the data this past summer with a team of five undergraduates at Harvard University. The following table contains information about the data sources.

TABLE I
DATA SOURCES

Land Use	MassGIS Oliver, USGS
Air Quality	USGS, EPA
Weather	NOAA, Weather Underground, EPA

Note that the land use data was originally formatted in GIS shape files (polygons). The Boston land use data was collected by the US Geological Survey during the period of 1970 to 1980, which unfortunately is the most up-to-date data.

IV. FEATURE ENGINEERING

A. Land use proportions

In order to create a data set containing site locations and their corresponding land use types, I created an artificial grid system over a 107.495 square mile region covering Greater Boston. I then divided the region into 50x50 grid cells. The artificial grid can be visualized as something similar to the grid in the figure below.

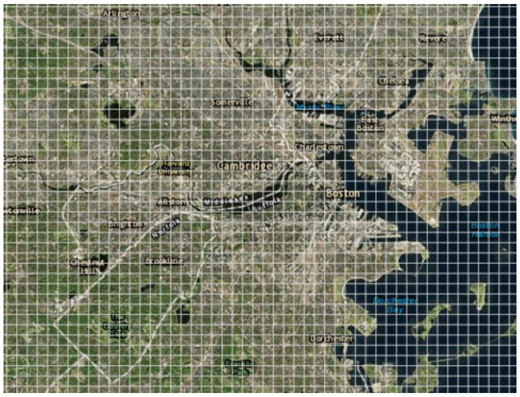

Fig. 2. Grid over Boston

I then uniformly sampled 100 random points within each grid cell. Each point is classified by the land use polygon within which it was located. This created a percentage break down of land use types for each grid cell. I found that Boston's land use types were distributed as such in fig. 3, where the x-axis represents the land use proportions.

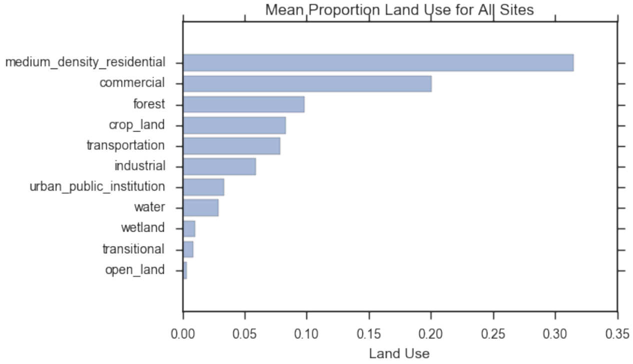

Fig. 3. Distribution of Land Use in Site Data

B. Feature engineering for classification methods

To make my data set a classification problem, I binned the ppm values into different categories for each air pollutant. The $PM_{2.5}$ ppm values were binned by the categories: low, low medium, medium, medium high, high, and highest. I binned the NO_2 values into the categories: low, low medium, medium, medium high, and high. Lastly, I binned the SO_2 ppm values into the categories: low, medium, high. Note, the SO_2 had fewer categories because the ppm levels lied within a smaller range.

V. MATHEMATICAL MODELING

A. Land Use Regression

Land Use Regression (LUR) is a linear regression model most commonly used in past studies to predict air pollutant concentration. The following land use types were included in my LUR model: industrial, commercial, medium density residential area, open space, crop land, water, wetland, transitional, forest, transportation, and urban public space. I also included weather data measured by 3 metrics: outdoor temperature, solar radiation and wind speed. The form of the LUR models is as follows

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \epsilon,$$

where the dependent variable y is the pollutant concentration of a given area. The variables $X_1 ... X_n$ represent the land use and weather features.

B. Variable Selection and Validation for LUR

The three different metrics I used for variable selection are p-value, R^2 and AIC. I implemented an 8-fold cross validation for the R^2 variable selection. Using backwards stepwise elimination, the set of predictor variables from the model with the worst metric, i.e. highest mean p-value, the lowest validation R^2 and the highest AIC, was eliminated at each step. Note that for all the metrics, variable selection reduced the predictor set to 8-12 variables.

Tables II through IV detail the results of variable selection on all three LUR models.

TABLE II
VARIABLE SECTION FOR PM2.5

Metric	MSE	Training R^2	Test R^2
P-value	9.910	0.211	0.222
AIC	21.970	0.231	0.234
R^2	16.955	0.209	0.203

TABLE III
VARIABLE SECTION FOR NO2

Metric	MSE	Training R^2	Test R^2
P-value	0.569	0.635	0.593
AIC	3.015	0.613	0.573
R^2	2.107	0.639	0.562

TABLE IV
VARIABLE SECTION FOR SO2

Metric	MSE	Training R^2	Test R^2
P-value	0.273	0.431	0.368
AIC	1.074	0.445	0.403
R^2	0.443	0.439	0.374

By these metrics, levels of NO_2 are most correlated with geospatial variations. The predictors that appear most often in the final subsets of variable selection are shown in Figure 4. I analyzed these feature importances in a further section.

C. Prophet Time Series

Prophet is a time series model with adjustable parameters used to increase or decrease the flexibility of the fit of the model. One of its parameters is a change point coefficient ranging from 0 - 1; the closer the coefficient is to 1, the more flexible the model is. It also is very useful for modeling non-daily data, i.e. time frequency, seasonality etc. In addition, the model fills in missing data by assuming a sparse prior based on the data's trend [3].

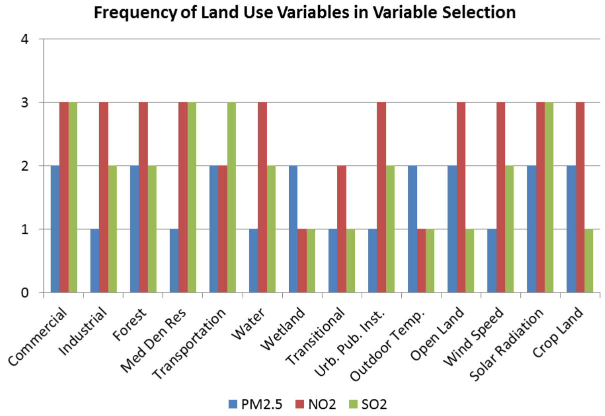

Fig. 4. Land use type frequency in variable selection

This model works well for modeling air quality data because it can easily adjust the fit to hourly data, and it fills in missing data more precisely than I could. It also is easily interpretable. I fit a model on each air pollutant for each of the four Boston EPA sensors. The features I used are ppm levels and sample collection date. I predicted a 48 hour forecast of air quality readings. I analyzed the predictions, the overall trend, and the daily trend of $PM_{2.5}$, NO_2, and SO_2 for one of the four Boston site locations. The other locations have similar trends for each air pollutant.

Note in the $PM_{2.5}$ daily trend plot (fig. 5), there are peaks around 3 AM, 1 PM, and 7 PM. Even more interesting, the change points (changes in rate) occur at 6 AM and 5 PM. This makes sense because traffic picks up around those times when people commute to and from work.

The NO_2 average daily trend (fig. 6) shows an increase in concentration from 3 AM - 7 AM followed by a decrease in concentration from 7 AM - 11 AM. This seems feasible because of high traffic in the morning.

Notice in fig. 7 that the average daily trend in NO_2 is different from $PM_{2.5}$. The NO_2 concentration consistently increases from 5 AM - 1 PM, where a rapid decrease in concentration occurs from 1 PM to 6 PM. It is interesting that NO_2 remains

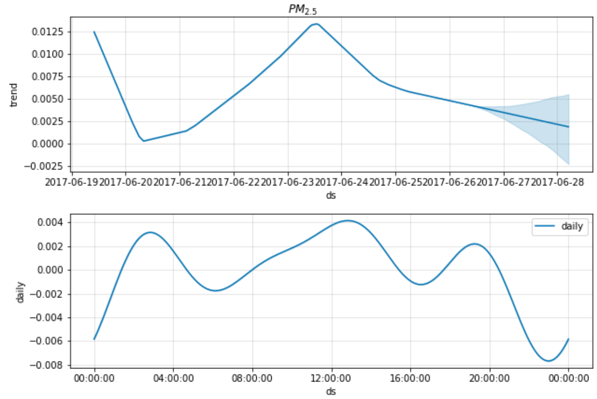

Fig. 5. $PM_{2.5}$ trends

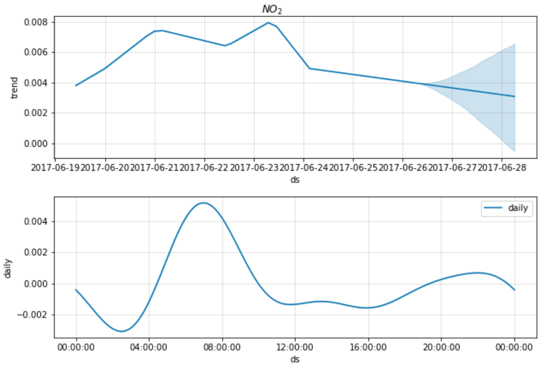

Fig. 6. NO_2 trends

relatively low during the same time period that SO_2 remains relatively high. Although NO_2 and SO_2 are formed by similar sources, daily power plant emissions may contribute to SO_2, which could explain this difference in trend.

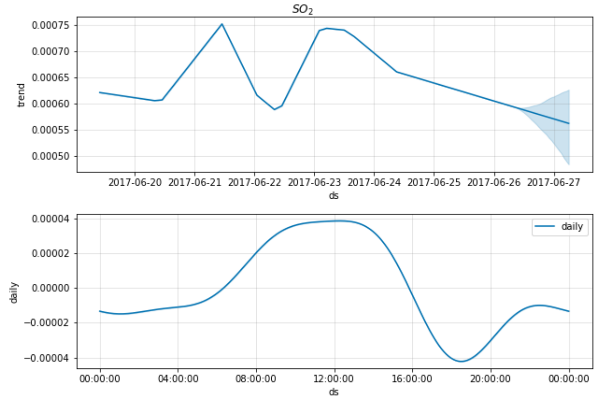

Fig. 7. SO_2 Trends

After a grid search, I fit the $PM_{2.5}$ model (fig. 8) using a change point coefficient value of 0.5; higher values do not significantly change the fit. In addition, I specified an uncertainty width of 0.5 which provides upper and lower bounds for my predictions.

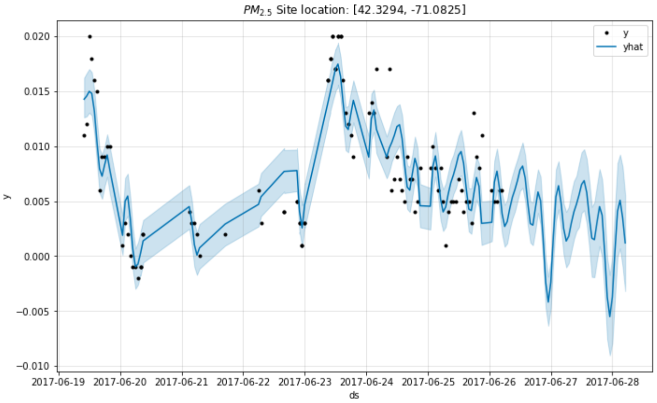

Fig. 8. $PM_{2.5}$ Predictions

The Prophet model fit the SO_2 more easily (fig. 9), with a change point coefficient of 0.2. However, I did include a larger uncertainty width of 0.7 in order to account for the noise around the fit of the model.

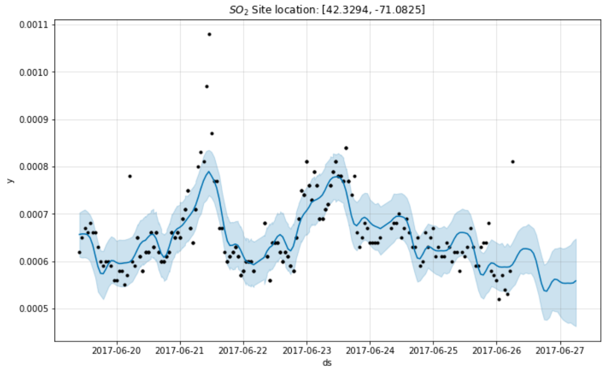

Fig. 9. SO_2 Predictions

The predicted NO_2 trend (fig. 10) looks the most cyclic compared to the $PM_{2.5}$ and SO_2 trends. The overall growth of NO_2 over this time period also remains relatively steady. I used a change point coefficient of 0.4 to aid in modeling the cycles; including more flexibility does not greatly affect the fit of the model and may result in over fitting.

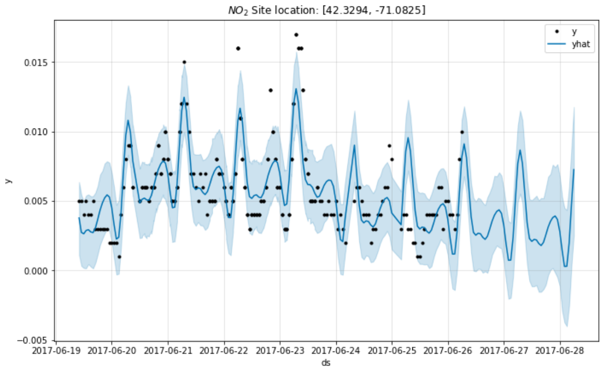

Fig. 10. NO_2 Predictions

D. Gaussian Processes

Next, I fit a Gaussian process model to my data. A Gaussian process models the non-linearity of the data by assuming a Gaussian prior on the observed (y) and predicted values ($\hat{y}$), i.e.

$$(y, \hat{y}) \sim \mathcal{N}(\mu, \Sigma).$$

Note, Σ is a covariance matrix where each entry is filled with a value that represents the relationship between its input variables. These values are calculated by a kernel function:

$$K(x, x*) = \sigma_f^2 \exp(\frac{-(x - x*)^2}{2\ell^2}) + \sigma_n^2 \delta(x, x*).$$

In my model, σ_f^2 is the amplitude of the air quality approximation, ℓ is the length scale, and σ_n^2 is the noise variance. I included a constant noise level of 0.001 to account for uncertainty in the data. The features are land use types and weather (outdoor temperature, solar radiation and wind speed). The data points are site the coordinates of my US data set.

The performance of the Gaussian process models is detailed in Table V.

TABLE V
Simulation Parameters

Pollutant	MSE	Training R^2	Test R^2
NO2	0	0.497	0.453
SO2	0	0.314	0.332
PM 2.5	0	0.206	0.199

E. Classification

In addition to predicting air quality, I am interested in classifying which sites and land use types correlate to low, medium, and high ppm levels of air pollution. To do this, I implemented random forest, Support Vector Machine (SVM), and logistic regression models. For each model (of each pollutant), the features are Boston land use types and weather (outdoor temperature, solar radiation and wind speed). The data points are the different ppm classifications described above.

Logistic regression models are a simple, yet effective algorithm when the data is linearly separable. Thus, I first trained a logistic regression model on my data to test linear separability of my data. The results are shown in table VI.

TABLE VI
Logistic Regression

Pollutant	R^2
NO_2	0.56
SO_2	0.71
$PM_{2.5}$	0.37

Note that logistic regression seemed to classify SO_2 relatively well, however the R^2 values for NO_2 and $PM_{2.5}$ are relatively low. This suggests that the data for these pollutants is not linearly separable. I then implemented an SVM because it uses a kernel trick to classify data that is not linearly separable. I found that using a linear kernel results in better classification of my data for each pollutant. SVM's can also be adjusted through a C value, which penalizes the error coefficient. I choose a C value of 1, because it is smaller and allows a small amount of misclassification. The results are shown in table VII.

TABLE VII
SVM

Pollutant	R^2
NO_2	0.57
SO_2	0.65
$PM_{2.5}$	0.33

Lastly, the random forest classifier is a useful method to finding feature importances by running an exhaustive search through decision trees on random subsets of the data. Note that my data set does not have more than $10,000$ data points, so I felt that using other tree classification algorithms, such as XGBoost, would over-fit my data. The results are shown in table VIII.

TABLE VIII
RANDOM FOREST

Pollutant	R^2
NO_2	0.54
SO_2	0.68
$PM_{2.5}$	0.352

Note that among all three methods, SO_2 was easily classified while $PM_{2.5}$ was the hardest to classify. Also, SVM performed the better than the other methods used, but not by much.

Lastly, I used the random forest feature importances parameter to find the most significant land use types. Commercial, transportation, and forest were significant to $PM_{2.5}$. The most significant land use types for NO_2 were industrial, transportation, and wetland. Lastly, the most significant land use types for SO_2 were industrial, wetland, and forest.

VI. ANALYSIS OF PREDICTION METHODS

A. Land Use Regression

Since a primary source of all three pollutants is the combustion of fossil fuels, I expected to see variable selection preserve ”industrial”, ”transportation”, ”commercial”, and ”medium density residential”. However, there are some variables that do not have known scientific relationships to the pollutants; ”wetland” for $PM_{2.5}$ is one example. Since the R^2 scores are on the lower end, I suspected that the time collection inconsistency of the land use data could affect the fit of the model. Recall, the most current land use types available from MassGIS is from the time period of 1980. This problem highlights why it is hard to model air quality with out-of-date data.

B. Prophet time series

The three metrics I used to measure accuracy of the Prophet time series models are proportion of true values within the uncertainty interval of predicted values (PIR), mean squared error (MSE), and mean absolute error (MAE). I calculated the PIR in order to get a good estimate for the accuracy of the fit. Note that MSE and MAE are commonly used to evaluate time series models. Recall, I analyzed the results for one of Boston’s sensor sites, which are recorded in the table IX.

TABLE IX
TIME SERIES MEASUREMENT METRICS

Pollutant	PIR	MSE	MAE
NO_2	0.696	1.691e-05	0.0024
SO_2	0.632	2.072e-08	4.599e-05
$PM_{2.5}$	0.415	7.956e-05	0.0035

The PIR for NO_2 and SO_2were both high. This implies that the models for these pollutants were able to capture a majority of the trend. The PIR for $PM_{2.5}$ was low; only 41.5 percent of the true values fell within the uncertainty range of predicted values.

The results of both MSE and MAE implied that the error between the SO_2 true values and predicted values was the smallest. I honestly expected the MSE and MAE to be lower in NO_2; in figure 10, the trend is more consistent than the trend in figure 9. I suspected that SO_2 has noisier data than NO_2. I am not surprised that $PM_{2.5}$ had higher MSE and MAE scores, because its trend (figure 5) has irregular patterns. There is also missing data in the $PM_{2.5}$ data that the model had to fill in with its default sparse prior.

C. Gaussian Process Model

Although the Gaussian process model introduced more flexibility, it did not perform significantly better than the regression models. I suspect that further feature engineering would increase accuracy, i.e. including distances from the EPA sensor sites to the city center, or including features such as total length of road segments contained in grid, traffic data, etc. Another possibility is that the land use data included a combination of 1970’s and 80’s Boston land use data along with a combination of data from 16 states collected at different time periods up to 2017.

VII. ANALYSIS OF CLASSIFICATION METHODS

I implemented logistic regression, SVM, and random forest models for SO_2, NO_2, and $PM_{2.5}$.

I analyzed the results for SO_2 in depth since the methods classified this air pollutant the best. I then summarized the results of NO_2, and $PM_{2.5}$.

A. Logistic Regression

As we see from table VI, the logistic regression model classified SO_2 relatively well. This is interesting because logistic regression is a simpler model. This infers that SO_2 ppm levels are linearly separable. To analyze this further, we look at its confusion matrix.

The confusion matrix evaluates the accuracy of the model's classification. Each row corresponds to a predicted class, whereas each column corresponds to the actual class. For example, along the first row of table X, 119 sites in total have low low ppm. However, LR classified 103 sites with low ppm, 16 with medium ppm, and 0 high ppm. LR classified sites with low ppm well, but was less accurate with classifying sites with medium and high ppm. It looks like sites with medium and high ppm are rarer than sites with low ppm. This is a common issue that occurs when working with data. I decided to use other classification methods in hopes to mitigate this issue.

TABLE X
LR CONFUSION MATRIX FOR SO_2

	Low	Medium	High
Low	103	16	0
Medium	46	10	0
High	1	0	0

Similarly to SO_2, LR easily classified sites with low ppm values for $PM_{2.5}$. However, I included more classes for $PM_{2.5}$, i.e. low, low medium, and medium. Because there were more classes, LR was able to classify sites more accurately between these three categories. Similar results to $PM_{2.5}$ held true for NO_2, as I also added an additional classification ”low medium”.

B. SVM

I tested the SVM with three different kernel methods and I found that the linear kernel classifies the data the best. I analyzed SVM's classification accuracy through looking at the confusion matrix for SO_2 (table XI).

TABLE XI
SVM CONFUSION MATRIX FOR SO_2

	Low	Medium	High
Low	92	27	0
Medium	32	23	0
High	2	0	0

Note that SVM was able to classify which data points had low ppm relatively well. Also, SVM did a better job than LR in classifying sites with medium ppm. Although, its classification could still be improved in sites with medium ppm and high ppm. This again could be due to rare occurrence of sites with high SO_2 ppm levels.

Conversely, for $PM_{2.5}$, SVM classified sites correctly with medium ppm values more often then for SO_2. Although, there were still a large number of sites misclassified, so $PM_{2.5}$ was not significantly better. Note that the results for NO_2 mirrored the results of SO_2.

C. Random Forest

After fitting a random forest classifier to my data, I found that it did not classify much better than logistic regression. I then analyzed the confusion matrix (table XII) for random forest classification on SO_2.

TABLE XII
RF CONFUSION MATRIX FOR SO_2

	Low	Medium	High
Low	119	0	0
Medium	56	0	0
High	1	0	0

From table XII, we can see that RF classified low ppm really well, in fact perfectly. However, RF completely misclassified medium and high ppm. RF produced similar classifications for both $PM_{2.5}$ and NO_2. Overall, it seems that all the classification methods I used misclassified sites with higher ppm values. This is an issue, as I am particularly interested in finding land use types corresponding to sites with higher air pollution (higher ppm values). I learned that introducing more classification categories improves error in misclassification. However, introducing more classifications can make the results less interpretable, and result in over fitting.

D. Feature Importances

I analyzed the coefficient terms in SVM and LUR, as well as the feature importances determined by my random forest model. I will discuss which features were important in classifying air pollutant concentration in the following paragraphs.

The most significant features for classifying $PM_{2.5}$ ppm were commercial, transportation, and forest. Since $PM_{2.5}$ is a fine matter that comes from a combination of sources, it makes sense that higher levels exist in commercial areas and in transportation areas. Traffic historically has highly correlated with polluted areas. The land use type that I was surprised by was forest. It could be that recreational activities, such as camping and hunting, affect $PM_{2.5}$ levels to a more significant degree.

The most significant land use features in classifying NO_2 were industrial, transportation, and wetland. The most significant land use types for SO_2 were industrial, wetland, and forest. Recall that the primary source of SO_2 and NO_2 is fossil fuel combustion. So, it makes sense that industrial and transportation are significant land use types. In addition, SO_2 is largely produced by power plants; there are about 10-15 power plants that were included in the grid. I was more surprised to find wetland to be significant feature for both SO_2 and NO_2. The only intuition I have for this is that the wind carries pollution over from the coast of Boston, aiding in higher urban pollution levels.

VIII. Conclusion

Air quality measurements, while not widely available or understandable, are crucial for understanding the behavior of pollution levels. This is a relevant issue as it can create a great deal of insight on how to increase public health. Given that the average person is unaware of the air quality in the area they live, I modeled the intra-urban pollution variations in the Boston area in order to attempt to create more awareness for air quality conditions. I implemented a time series model, which created helpful predictions and visuals on overall and daily trends. The Gaussian models helped me understand that air pollution most likely does not have a linear relationship; it is likely correlated with time dependent features, such as traffic. Lastly, I implemented three classification models in order to analyze further which land use types correlate with higher levels of air pollution. Land use types such as industrial, commercial, and transportation were important in classifying ppm values.

Overall, air quality is a rising issue that many people are trying to model in order to mitigate its affects on public health. It is an intriguing, and complex issue with many variables, both spacial and temporal, involved.

Acknowledgment

Anthony DePinho*, Tara Ippolito*, Biyonka Liang*, Annamira O'Toole*

Dr. Weiwei Pan, Harvard Institute for Applied Computational Science

Dr. Pavlos Protopapas, Harvard Institute for Applied Computational Science

Dr. Gary Adamkiewicz, Harvard T.H. Chan School of Public Health

Dr. Jaime Hart, Harvard T.H. Chan School of Public Health

*Represents those who equally contributed to data collection, cleaning, feature engineering, and LUR and Gaussian Process Modeling

References

[1] the American Heart Association. Circulation, vol. 121, no. 21, Oct. 2010, pp. 23312378., doi:10.1161/cir.0b013e3181dbece1.

[2] Brook, R. D. Air Pollution and Cardiovascular Disease: A Statement for Healthcare Professionals From the Expert Panel on Population and Prevention Science of the American Heart Association. Circulation, vol. 109, no. 21, Jan. 2004, pp. 26552671., doi:10.1161/01.cir.0000128587.30041.c8.

[3] https://facebook.github.io/prophet/docs/trend_changepoints.html.