Air Pollution and Its Relation to Weather

Introduction:

Air Pollution has gained a lot of traction in the years that we have studied our planet. There has been a steady rise in pollution levels that has been reported all over the world. According to the World Health Organization the outdoor pollution has risen 8% in the past five years with billions of people at risk of being exposed to dangerous levels of air pollution. These substances have led to our air being more and more unbreathable. The United Nations has started counting it among the 17 Sustainable Development Goals (SDG) for the planet. Air pollution comes under the SDG-15, Life On Land. The World Health Organization also estimates that there are 7 million premature deaths related to air pollution in a year. Though we have seen a consistent improvement in the air quality in the US we need better means to see what is causing these changes. Whether we are adequate in our efforts to control this irreversible change we are causing or do we need to put in more efforts.

With this particular goal in mind we are looking at the air pollution data all over the US and comparing them to weather patterns. We are trying to see if the changes in the air pollution parameters that we are witnessing are due to the weather patterns or are we successful in our efforts to reduce the effects of our activities on the environment. We are trying to find which weather factors if any affect the pollution factors extensively. We will also predict the air quality in the future to see if we are meeting the air quality estimates made by our model. All this is to know our current and future standing in this battle against the air pollution.

Background:

We conceived the idea of this project after coming across another study. In December 2019 Rafael Borge, Weeberb J.Requia, Carlos Yague, Iny jhun and Petros Koutrakis came up with a study of air pollution in Spain over the period of 25 years and their effect on the health. This paper studied the effects of weather on the air pollutants. They have considered some of the major pollutants like C_6H_6, CO, NO_2, NO_x, O_3, PM_{10}, $PM_{2.5}$ and SO_2. They have considered the fact that the pollutant concentration is not just an effect of the emissions and weather. They have also done the analysis in specific geographical and seasonal data to account for the variability of the seasons.

The model used in the study to [predict the air quality in the future was a regression model. They have used a regression model with years, months and days of week as the features in the model. They have assigned penalties relating to the weather. They have not used the weather as a separate feature. Instead they have opted to account for the weather by assigning penalties in the regression model. Their study revealed significant penalties for the weather patterns. Hence we felt we needed to investigate the correlations of weather patterns with air pollution parameters. We have decided to predict the Air Quality just based on the weather pattern. Hence we elected for an LSTM model which can take the weather patterns as inputs and give us predictions for the air quality in the future.

Data:

We were building the model specifically to concentrate on US air pollution. We got data that was specific to the US and had good resolution of data geographically as well as temporally. We used the United States Environmental Protection Agency's website to gather our data on air pollution. We selected this source to gather the data as the data available from this source was well structured and met the

requirements for our application very well.

Our source had data worth 17 GB to be exploited for our analytical purposes. We had data for each pollutant and each meteorological parameter in separate tables for separate years. We have used the air pollution parameters like CO, SO_2, NO_2, O_3 and Air Quality Index (AQI). The table for AQI has 10 features and about 1 million observations. The other air pollution parameter tables have 29 features and about 1 million observations. The weather patterns that we used were temperature, pressure, wind and humidity. All the parameters have their own tables with 29 features and about 1 million observations each.

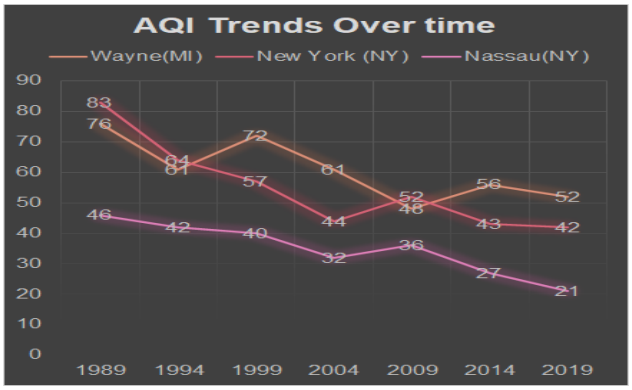

Figure 1: AQI trends from 1989 to 2019.

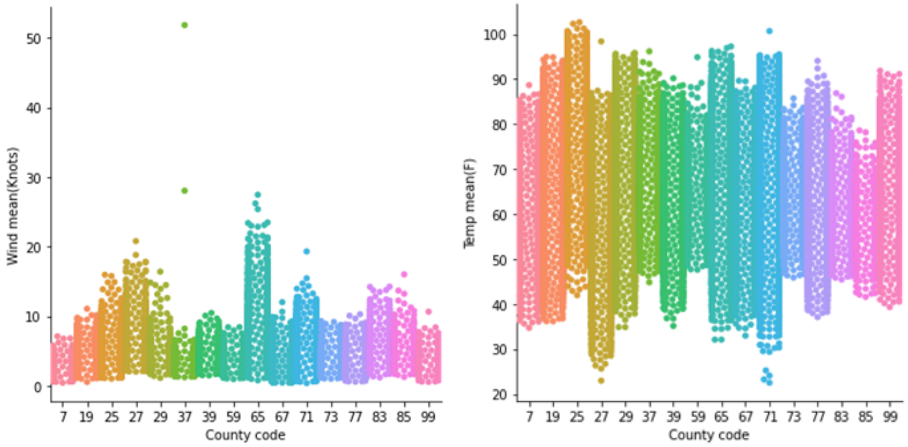

Figure 2: Plot of Weather pattern data segregated by county code

We can see the visualizations of the AQI data over the course of many years in the figure 1. We can see the AQI has been on a consistent downward trend for a very long time. This is a good sign since lower AQI is more desirable. We want to know if the efforts put by humans to curb the air pollution is yielding effects or does it have more to do with the weather pattern and their change over time. The figure 2 shows us the plot of the weather pattern data available to us. We can see that we have weather data which is spread over a very wide range and can be used to give AQI predictions for every weather imaginable in the US. We can also see that there is considerable grouping in the data when arranged by county code.

Methods:

Task 1: Multivariate Regression with Fixed and Random Modelling

The goal of the first task of the project was to determine if there is any correlation between meteorological features like wind, temperature, pressure and humidity with pollutants like CO, SO2, Ozone and NO2. The data used for this task consisted of daily measurements of weather features and pollutant concentration for each of the 50 states in the United States. We calculated the correlation values for each of the 50 states and organized the data state-wise so that each of the regression task could be run in parallel on Spark.

Since each state consists of different number of county and it might be possible that each county has its own individual characteristics other than the weather features which might be contributing to the pollutant concentration, therefore we needed to control for these omitted features which we have not included in our independent variables so that the estimated coefficients are not biased. To do this, we modelled county as a fixed effect i.e. we included dummy variables for all the counties in a state except for the first county which is the reference county. Fig. 1A shows fixed effect modelling where each county has the same slope but different intercept i.e. fixed effect modelling does not take into account the difference in meteorological features between different counties. In order to control for this as well we modelled county as a random effect so that we are taking into account the difference in weather features between different counties as well. Fig. 1B shows the random effect modelling where each county has different intercept and different slop. Since our data was spread over a duration of 10 years, we also needed to take into account seasonality. To do these we included dummy variables for year and month except for the first year and first month which are the reference year and month.

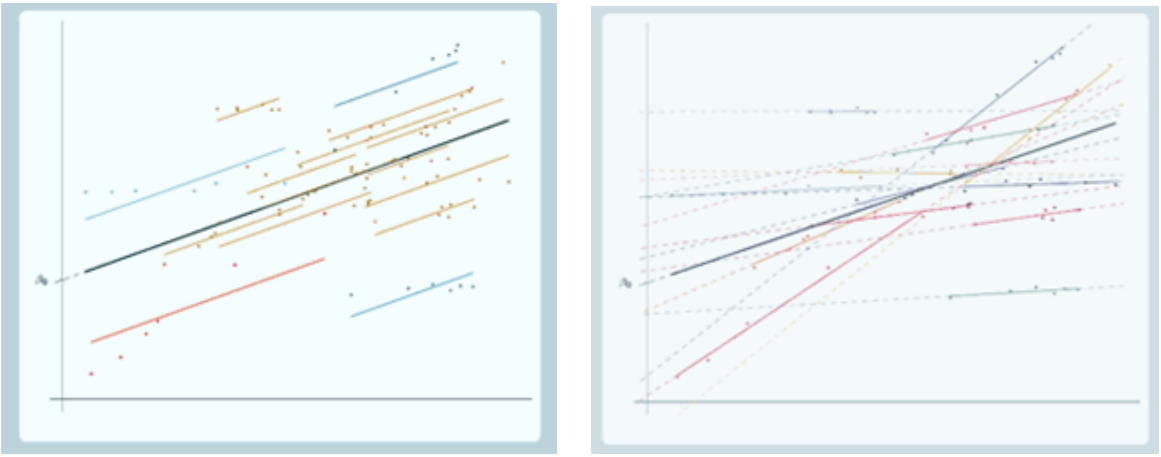

Fig 3. Image A represents fixed modelling with county as a fixed effect and Image B represents random modelling with county modelled as a random effect.

Task 2: AQI prediction using LSTM

To predict AQI values based on a history of wind parameters and previous AQI values, a LSTM Recurrent Neural Network was developed in Keras. The predictions were to be made on data seperated by state, with one data point per day. To organize the data into this structure, and remove unnecessary features, Pandas was used to clean the initial data and prepare it to train the LSTM network. Pandas was used to strip the unnecessary columns, concatenate input csv files by year, intersect csv files by State Code and Date for different parameters, and eliminate rows of duplicate State Code and Date. Resulting in the State Code separated data, with one set of measurements per date.

The same LSTM model is trained and tested for each State Code. The model consists of 2 layers: A LSTM layer of 50 LSTM neurons, and a fully connected layer of 1 neuron. The system is rather shallow, but because of this, training and testing were very fast. The fully connected layer is the output layer and the LSTM layer is the input layer. The inputs to the input layer are wind speed, pressure, temperature, and humidity. The output of the output layer is the predicted AQI values. Below is a diagram of the model:

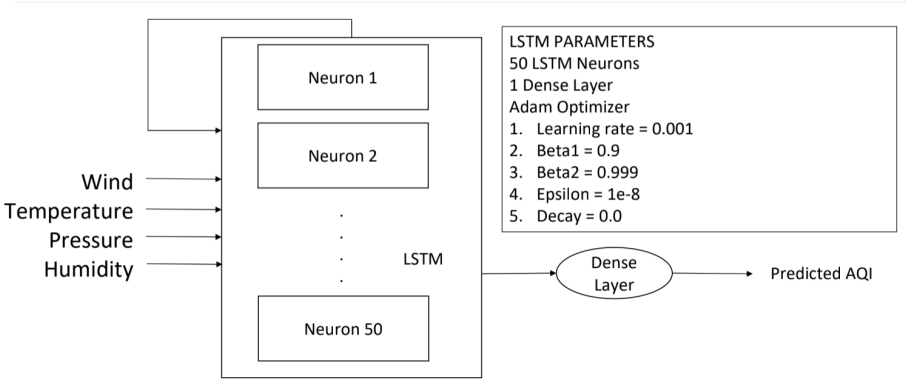

Fig 4. LSTM Architecture

The data for any given State Code is split 80% for training and 20% for testing. The rows of earlier dates are used for training the model, while the rows of later dates are used for testing. Many hyperparameter combinations were considered when optimizing the network. The following hyperparameters were varied: Learning rate, momentum, batch size, number of LSTM neurons, number of epochs, decay, betas, epsilon, and back-propagation optimizers. The Adam optimizer and Stochastic Gradient Descent (SGD) optimizers were considered. But for no combinations of momentum and learning rate was the SGD optimizer better than the Adam for this application, so the Adam optimizer was used for back-propagation, using absolute mean error. Although the error/cost functions were not varied throughout optimizations. Epochs larger than 25, batch sizes larger than 72, and LSTMS larger than 50 seemed to no longer yield significant improvements in test accuracy. A learning rate of 0.001, betas of 0.9 and 0.999, epsilon of 1e-8, and decay of 0.0, were discovered to yield the best test accuracies.

Evaluation/Results:

Task 1:

We found that in general that temperature has very high correlation with pollutants specially ozone this may be because when temperature is higher more cooling systems are running and hence more ozone concentrations. The general trend for humidity we observed was that it was negatively correlated to pollutant and this may because since all the pollutants are water soluble they may not be measured by the sensors.

We observed that results for fixed and random modelling varied highly for some states this may be because fixed modelling requires that there should not be any correlation between heterogeneity of different county in states. We found that mixed modelling would have been a better approach for this problem as it would take into account county specific heterogeneity as well as the effect of differences in weather features between different county.

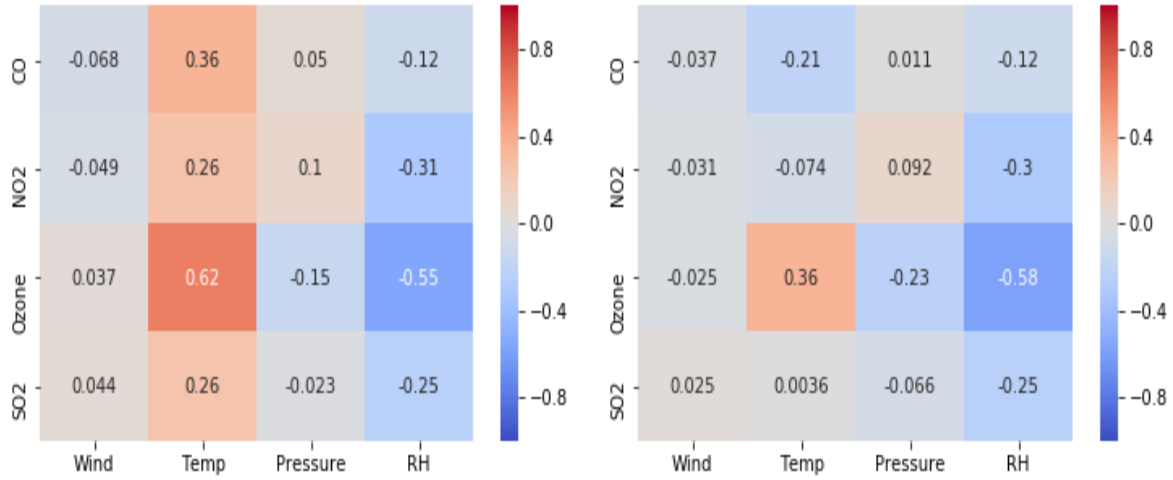

Fig 5. Image A represents the correlation coefficients for fixed modelling and Image B represents the correlation coefficients for random modelling.

State	p-val(wind)	p-val(temperature)	p-val(pressure)	p-val(humidity)
Maryland	0.00525	0.175	0.617	0.000236
Florida	0.1377	1.39e-12	0.1494	2.95
North Carolina	3.09e-14	2.019	2.35e-5	0.33

Table1. Shows the hypothesis testing result for 3 states for CO pollutant w.r.t weather features. P-vals highlighted with red are rejected while with blue are accepted.

Task 2:

To evaluate the results of the LSTM model, the average Root Mean-Square Error (RMSE) for predictions during the testing phase were measured. The average RMSE across all State Codes was 13%. This gives a good insight that by using the history of weather parameters, the AQI of a region can be predicted with considerable accuracy. Although the worst case average RMSE for a certain State Code was 45%. This is suspected to be due to problems occuring during the data cleaning with Pandas. The results for a 200-day prediction for a certain State Code can be observed below.

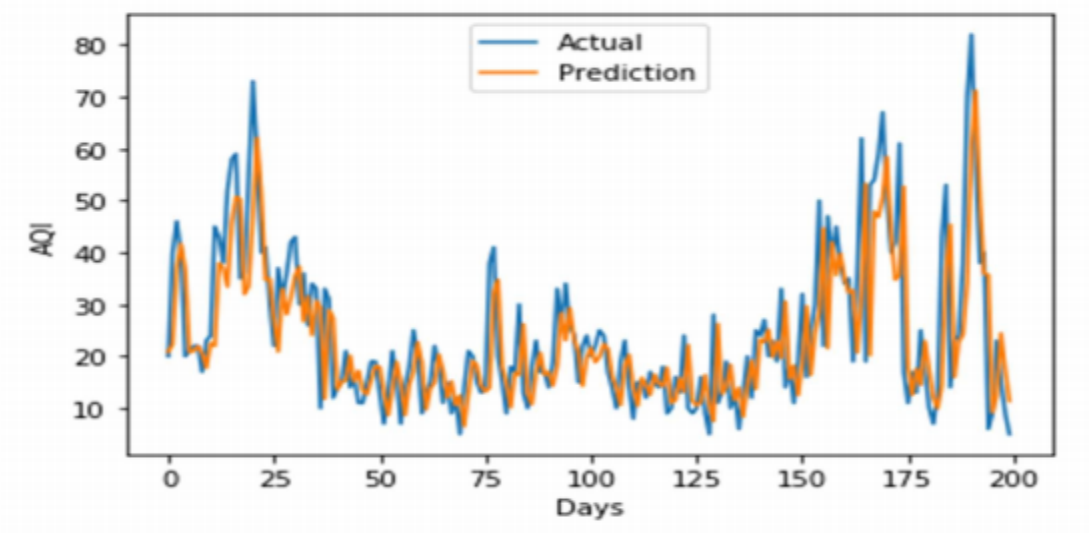

Fig 5. AQI prediction over a period of 200 days

The resulting AQI values in this example closely predict the actual AQI values for this region (with about 11% average RMSE). Although a limitation of the model is that it relies on only a few weather parameters. If given more parameters, such as rain and snow measurements, the model may be able to make more accurate predictions.

Conclusion:

Correlations with AQI and predictions of AQI were successfully made by using weather pattern data in this experiment. Pollutants have strong correlations with temperature and humidity, and AQI values can be predicted accurately (averaging 13% error) using previous weather data for a region. Using these correlations, federal governments can decide the location of industries such that they are mainly located where there is negative correlation between pollutant and meteorological features. The LSTM model predictions can be used to further assess how we can expect the AQI of a region to change in the future, and track our progression towards the SDG. These conclusions can be made with confidence, considering the vast amount of data used (and hypothesis testing).

References:

1) https://www.theguardian.com/environment/2016/may/12/air-pollution-rising-at-an-alarming-rate-in-worlds-cities
2) https://www.sciencedirect.com/science/article/pii/S016041201930248X
3) https://aqs.epa.gov/aqsweb/airdata/download_files.html#Annual
4) https://www.kaggle.com/ojwatson/mixed-models
5) http://www.bristol.ac.uk/cmm/learning/videos/random-slopes.html
6) https://machinelearningmastery.com/multivariate-time-series-forecasting-lstms-keras/