SI618 Final Project – Jonathan Hartman

(uniquename: jonhartm)

Motivation

I chose this topic for two reasons – firstly, I was raised to be a bit of an environmentalist, and given the dire predictions we've been getting from scientists around the world over the past few years I was curious to take a look at some of the data myself, both to perform my own analysis on it and to give myself a better grounding in the topic with some actual data. Secondly, as this is a project heavy semester for me, I was trying to make each of them as different as possible, and as I already had projects underway with political and social media datasets, an environmental set seemed like an interesting new one to add.

Project Questions

Is the general trend in airborne pollutants going up or down, and which states have the highest/lowest measures in a given year? Which are making the most/least progress?

California has some of the strictest environmental laws in the country. Does this result in a significant difference in pollutant measures when compared with other states?

How strongly correlated are the various pollutants? We would already expect a strong seasonal variation within each, but how strong are the relations between each type of pollutant measured?

The Clean Air Interstate Rule of 2005 and subsequently the Cross-State Air Pollution Rule of 2012 specifically required certain states to meet emissions targets for NO2 and SO2 – are the effects of these rules visible in the dataset?

Data Source

This project is based around public data provided by the US Environmental Protection Agency (https://aqs.epa.gov/aqsweb/airdata/download_files.html), which in turn was scraped and collected into a csv file for Kaggle (https://www.kaggle.com/sogun3/uspollution). The datafile consists of daily measures of four airborne pollutants (Nitrogen Dioxide, Carbon Monoxide, Sulphur Dioxide, and Ozone) at 204 sites between 2000 and 2016. In total, there are ~1.7m records, however most sites were sampled between 2-4 times each day, so for most purposes in this report I average together all reports from the same site on the same day. Sites are provided with some categorical labels, including State and County Codes (based on the Federal Information Processing Standard – FIPS), a unique identifier for each site, as well as text for each. For each record, there are four numerical measures for each pollutant: Mean, the daily Max, the hour of the daily Max, and the Air Quality Index for that day for that pollutant, as well as a corresponding date. For the purposes of this project, I focus on the Mean measures.

Q1 - Is the general trend in airborne pollutants going up or down, and which states have the highest/lowest measures in a given year? Which are making the most/least progress?

Method

There are two difficulties I encountered immediately upon starting with this dataset – first, the regularity of the data. There is a strong seasonal correlation within the various measures that make straightforward plotting very difficult to interpret. The second issue comes from the extremely different scales used to measure each pollutant – NO2 and SO2 are measured in Parts Per Billion, whereas O3 and CO are measured in Parts Per Million. Even when plotting them separately, the measures fall within vastly different scales – O3, for example, ranges from 0.027ppm to 0.022ppm, whereas CO covers between 0.6ppm and 0.27ppm. This make plotting

any of the measures on a chart problematic, since major variations in one will dominate the y-scale and leave the rest looking like flat lines.

Since the first part of my question is dealing specifically with the overall trend of data, I first grouped the DataFrame by date regardless of the location of the measure to get the nationwide average. To address the issue of the seasonal variations, I calculated a rolling average for each of the measures over 365 days. This results in the loss of data prior to the first year but allows us to see any trends without the seasonal waves. To deal with the second problem, I divided the DataFrame by the first non-null values, which gave me a plot of the percent change in each measure over time **(Fig. 1)** and allowed me to put them all on the same axis.

To get the top/bottom states in a given year, I used a similar method to finding the overall, just broken into years. I first made a DataFrame by grouping the data by year in order to find the average values for each per year. I then made a second DataFrame grouped by State and Year. To the second DataFrame I applied a function that took each row, retrieved the average values for the given year from the first DataFrame, subtracted the mean from the state's annual average, and divided by the yearly average to end up with the percent difference from the mean for each year and state. I then iterated through each year in the dataset, grabbing the top and bottom state from each **(Fig. 2)**.

The third part of the question deals with progress. I decided that this could be interpreted as the slope of the regression line through all of a state's measures over time, where a negative slope would indicate progress in lowering airborne pollutants. I calculated the regression line for each state in turn for each measure, calculated the mean slope, and plotted the result on a choropleth map **(Fig. 3)**. I did run into the issue here of five states which had no data but leaving them as "NaN" values was problematic for the plot. I corrected this by filling them with "0" when plotting, then making a second plot on top, with only those states colored grey.

Analysis

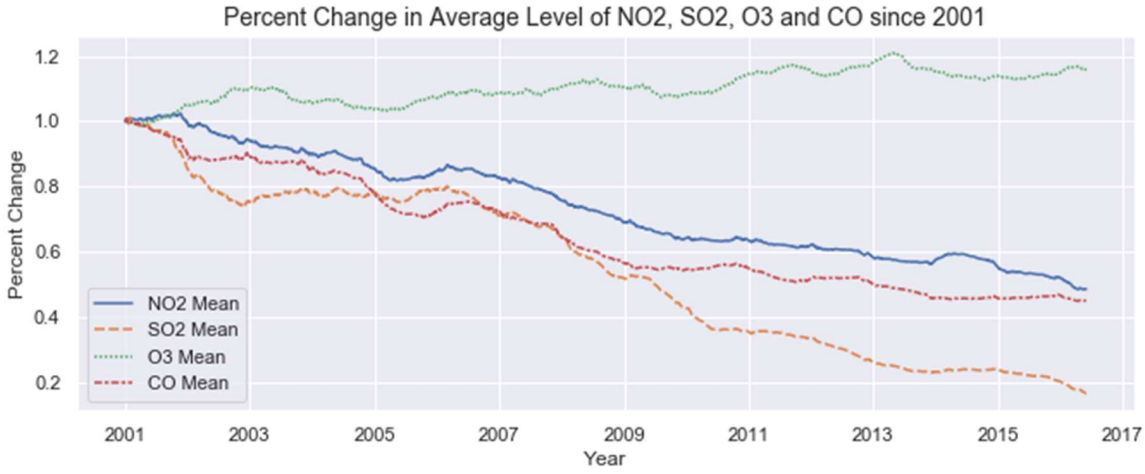

Figure 1 – Percent change of nationwide average measures for all four pollutants from 2001 to 2017. Based on a 365-day rolling average

Year	***Bottom State***	***Top State***
2000	District Of Columbia	Oklahoma
2001	District Of Columbia	Nevada
2002	District Of Columbia	Nevada
2003	District Of Columbia	Nevada
2004	District Of Columbia	Oklahoma

2005	District Of Columbia	Oklahoma
2006	Indiana	Iowa
2007	District Of Columbia	Maine
2008	District Of Columbia	North Dakota
2009	Country Of Mexico	South Carolina
2010	District Of Columbia	South Carolina
2011	Country Of Mexico	North Dakota
2012	Kansas	North Dakota
2013	Kansas	Wyoming
2014	Alaska	Wyoming
2015	Alaska	North Dakota
2016	Utah	Wyoming

Figure 2 – The top and bottom states for overall pollutant concentrations by year. Based on the difference of each state from the mean for that year.

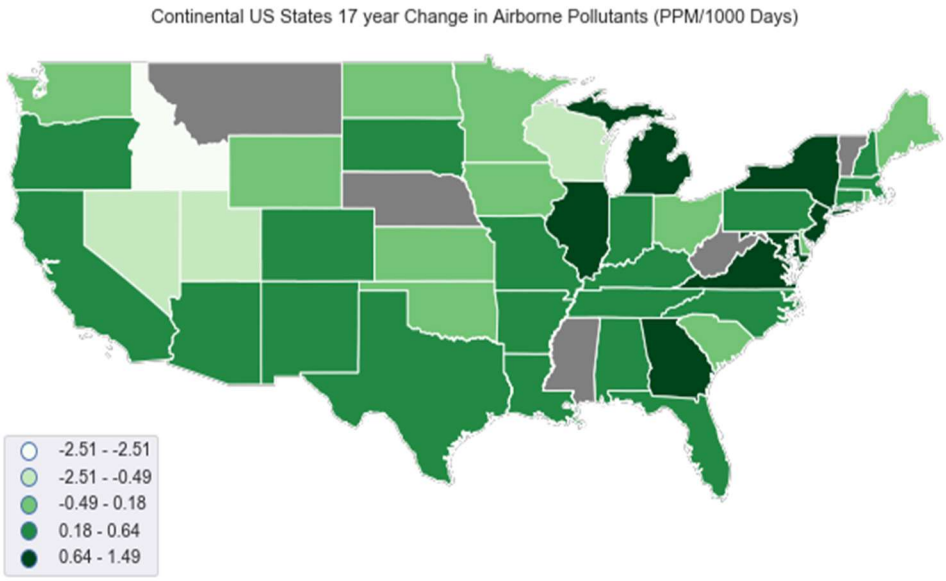

Figure 3 – Choropleth map of the continental US. Darker shades indicate a steeper drop in airborne pollutants. (Grey indicates no data)

Regarding the general trend in airborne pollutants, of the four measures, three do appear to have been dropping nationwide over the last 16 years. Only O3 doesn't appear to have dropped and has actually increased over the time period covered by the dataset. The other three appear to have been dropping at roughly similar rates. All measures seem to plateau a bit between 2003 and 2006 before starting to decline.

I was a little surprised at the results to the second part of this question – Washington D.C. doesn't have a reputation for air quality like some other parts of the country, so I'm not sure what to make of the number of consecutive years it makes it onto this list. The only hypothesis I can put forward is that these four measures are directly related to vehicle emissions, and the DC area is known for having some of the worst traffic in the country. That doesn't go to explain why Alaska, Kansas and Utah are also in this table however. The top list is more what I was expecting – typically rural states without many large population centers.

The results of the third part regarding which states are making the most progress might have a few explanations – since the slope doesn't take into consideration the starting point of a given state, it may just be that Idaho, Nevada and Utah started with better measures than other states and are just becoming average. It also might be that low-level pollution travels with weather patterns, and that west coast states are contributing to pollution levels in states directly to the east of them. (I opted to remove Alaska and Hawaii from this map, despite being included in the dataset, as the shapefiles I was able to find for GeoPandas distorted the contiguous states in order to show them.)

Q2 - California has some of the strictest environmental laws in the country. Does this result in a significant difference in pollutant measures when compared with other states?

Method

One of the caveats with answering this question has to be that California is vastly over-represented in this dataset. Of the 204 reporting sites in the data, 49 of them are located in California. As a result, although drawing conclusions from California's records may be representative, states such as Washington or North Dakota with only a single reporting site may have more variation in their results. To address this, I decided that rather than compare California against each other state, I would compare against the daily average across the US without California. This also helps in a few areas where there are incomplete measures, e.g. Missouri is missing entries between 2009 and 2013.

I also ran into an issue where Seaborn's regplot would not take a pandas DateTime variable as an x-axis. I got around this by creating a new column from the DateTime that was the Unix Timestamp equivalent. As a continuous variable, this was accepted by the regression plot. The drawback of this, of course, is that Unix timestamps are not readily understood by a human observer, so although the general trend of the data is visible, it is difficult to specify when any particular part of the plot is occurring. As I was only interested in comparing the two trends, I decided that this was an acceptable tradeoff.

To address this question, I first split the dataset into two halves – one that contained the running averages of California alone, and the other which was the combined average of all observations not in California. As we still have the scale issue preventing us from directly comparing all four measures at once, I ended up making four plots – one for each measure.

Analysis

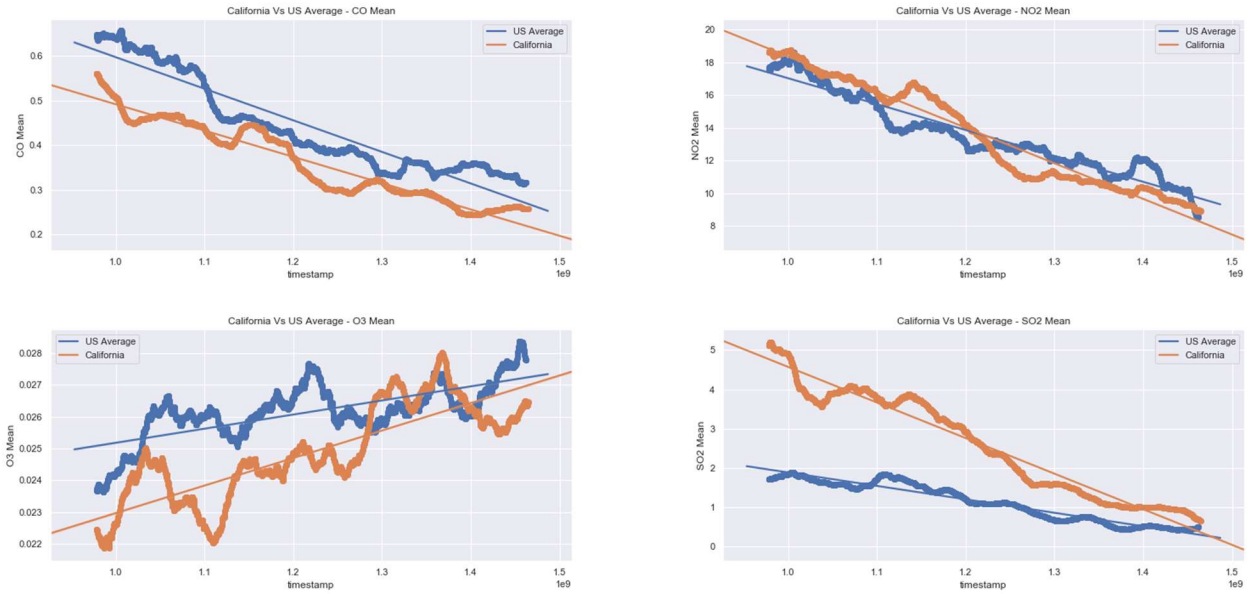

Figure 4 – Regression plots of each of the four pollutants. The blue plots indicate the mean measures from California, the orange plot is the remainder of the US states.

Overall, California does seem to be doing slightly better than the US average. Visually comparing the regression lines, it appears that in three of the measures – CO, NO2, and SO2, California is outperforming the US average, and in the case of SO2, it is doing significantly better. O3 measures are still an outlier, as they were when looking at the nationwide average. Despite having some of the stricter environmental laws, California only appears to be matching the national average in NO2 and CO, while slightly outstripping the average in O3, to the point where, despite starting lower than the average, California looks to surpass it soon if it hasn't already (the dataset ends in 2016). SO2 levels, on the other hand, began at nearly twice the national average in California, and are down to nearly the same level at this point in time. Of the four, SO2 is probably the most concerning pollutant, as it is both a major health hazard as well as a primary factor in rain acidification.

Q3 - How strongly correlated are the various pollutants? We would already expect a strong seasonal variation within each, but how strong are the relations between each type of pollutant measured?

Method

In my analysis of the data, this was actually the first step, although it was the third of my questions. My first approach was to simply create a Pairplot of the four measures. It was readily apparent that this was a less than ideal approach, as the plot not only took several minutes to create, but the resultant plots were far too busy to easily compare them. I decided to group the data on the Date collected, which left me with around 5000 readings to compare, but still managed to capture the general appearance of the data. **(Fig. 5)**

In addition to the Pairplot, I included a HeatMap **(Fig. 6)** of the correlation matrix to help show how strongly the variables were related numerically. I then took the top three most correlated pairs of measures and created JointPlots **(Fig. 7)** to get a better sense of the specific relations between those two measures.

Analysis

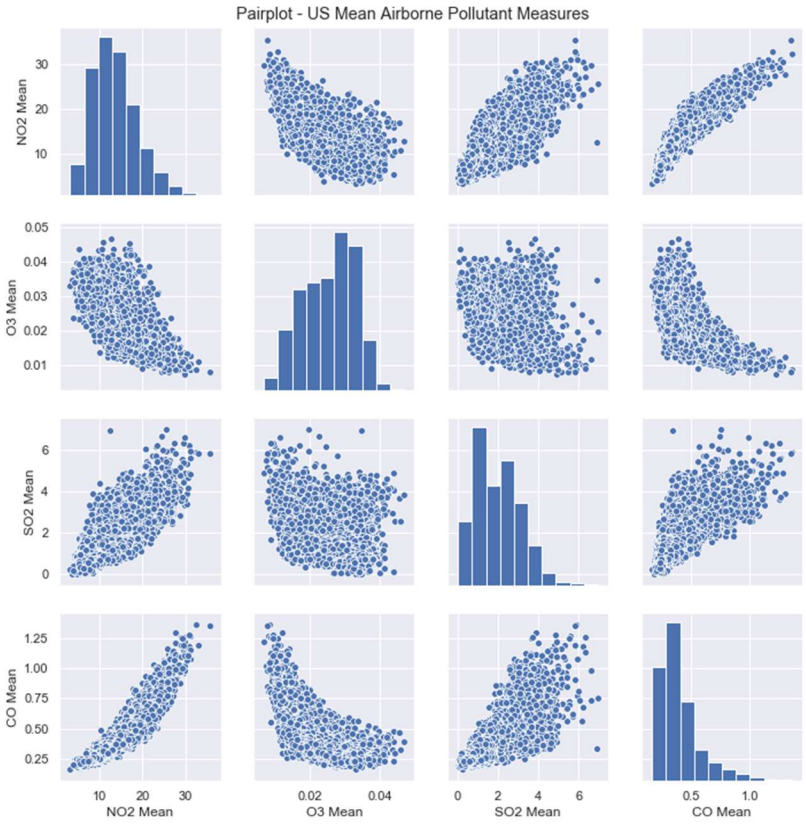

Figure 5 – Pair plot of all four measures

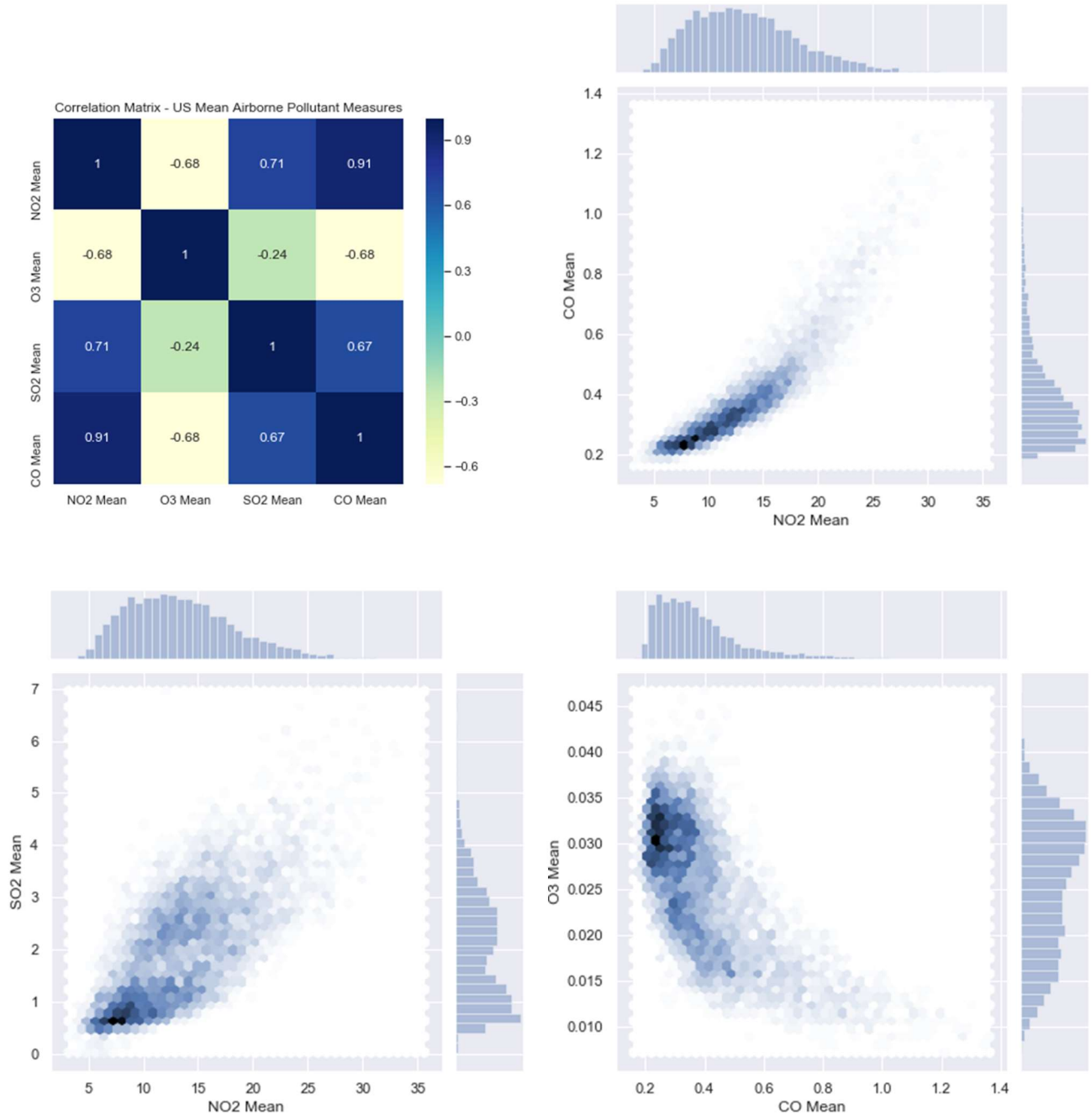

Figure 6 – Heatmap version of the correlation matrix and Joint plots of three of the stronger correlated pairs of measures.

Of the six potential pairs, there is a very strong correlation between two – CO & NO2 and CO & SO2 – and a weaker but still somewhat significant relation between two others – SO2 & CO and O3 & CO. CO & NO2 I think is the most interesting, since the relation is very strong but it doesn't look quite linear. I've read the EPA's explanations of these two pollutants and I'm not sure why this kind of relationship would be present. There is a similar non-linear appearance to the CO & O3 plot, though with a negative correlation.

In both of these cases, the strongest part of the pair plot, in which most measurements fall, could be considered linear, while it's the sparser part of the plot which begins to curve. The O3 & CO plot could be a result of two trends in the O3 levels – looking back at Figure 1, there seems to be little to no trend until 2011, when it begins to increase slightly. My only other hypothesis would be that as there is a time component to

each measure, that one decreased at a higher rater than the other for a period of time, before the other began to decrease at a similar rate.

Q4 - The Clean Air Interstate Rule of 2005 and subsequently the Cross-State Air Pollution Rule of 2012 specifically required certain states to meet emissions targets for NO2 and SO2 – are the effects of these rules visible in the dataset?

Method

Since this question is asking about comparing three separate timeframes, I used a combination of the methods I applied to earlier questions. I started by filtering the dataset into three separate groups – one from 2001-2005, one from 2005 to 2009, and one from 2012 to 2016. This way I had three sets of data covering four years of measures I could compare. I also filtered the data in each set to only the 28 states for which the laws applied. I again applied the rolling average in order to account for the seasonal variation, then created a new calculated column for each measure based on the Date of each measure – I converted each time to a Unix Timestamp, then subtracted the Timestamp value of the first day in that set and divided each by the number of seconds in a day. This resulted in a new column which represented the number of days since the implementation of a law for each measure. Finally, I divided each column by the first value to get the percent change over time. Plotting each measure on its own plot avoids some messiness, and more clearly demonstrates the differences between the two measures. I also decided to omit the entirety of 2010 and 2011 in order to provide comparable four-year spans of data.

Analysis

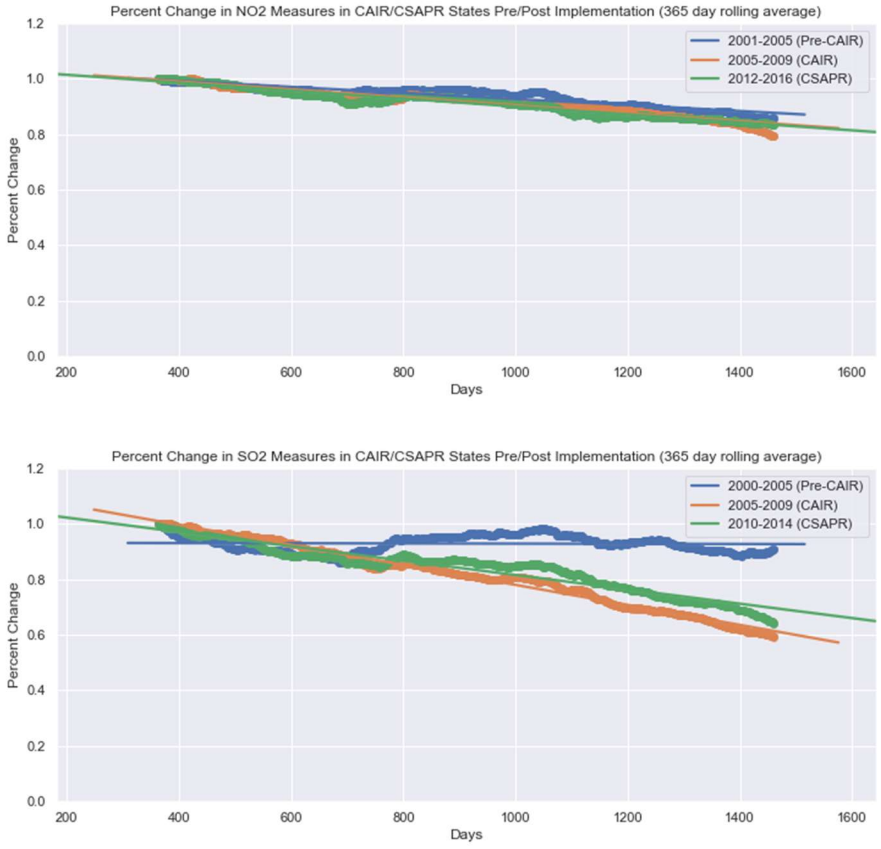

Figure 7 – Regression plots showing the percent change in NO2 and SO2, among states that fall under CAIR/CSAPR regulations. Each line represents a four-year period immediately before or after a law was implemented.

Based on this dataset, it looks as though neither law has had a particularly dramatic effect on the NO2 levels. They are generally decreasing, which is of course the end goal of these regulations, but there doesn't appear to be much acceleration in change. The CAIR and CSAPR years are just slightly better than the pre-2005 regression line, but it's a negligible difference. The SO2 measures on the other hand, are significantly different. The pre-2005 regression line is almost level (statsmodels.linregress returns a slope of -.0000036), whereas the CAIR and CSAPR regressions are clearly trending downward. This isn't to suggest that legislation and regulation aren't important in environmental protection, but I think it's valuable to be able to attempt to relate the outcomes of a particular regulation with it's apparent effects.

These results might be a little misleading, as they're based on percent change rather than the actual value. The dataset is missing any records prior to 2000, so we can't really determine what the trajectory of each of these measures was before this dataset begins. If NO2 measures were already trending down in these states, then a continued decrease at a similar pace might be all that we would expect.