Researching the Causes of Air Pollution Across the U.S.

Shaheer Aslam, Data Science Intern

August 2020

Abstract

The goal of this research is to tackle the issue of air pollution within the United States, using visualizations of data to aid us in our analysis. We investigate measures of air pollution across the fifteen largest metropolitan statistical areas (MSA) to understand the varying immutable and mutable factors that contribute to dirty air. Areas include large metros such as the New York MSA and the Los Angeles MSA. This study considers the implications of the climate, types of energy consumption, production and GDP, and commuting. According to the Environmental Protection Agency (EPA), these are the most common causes of polluted air. We overlay these factors alongside air quality data and present them using a variety of time series graphs and visualizations to reveal any possible correlations. To develop this research, we create functions which clean, format, and plot data for all of the MSAs to easily compare and picture the trends for analysis. With these graphs, we can then begin to answer our question: what are the differences in air quality across the largest MSAs across the United States, and what factors seem to be the most correlated to higher pollutant concentrations?

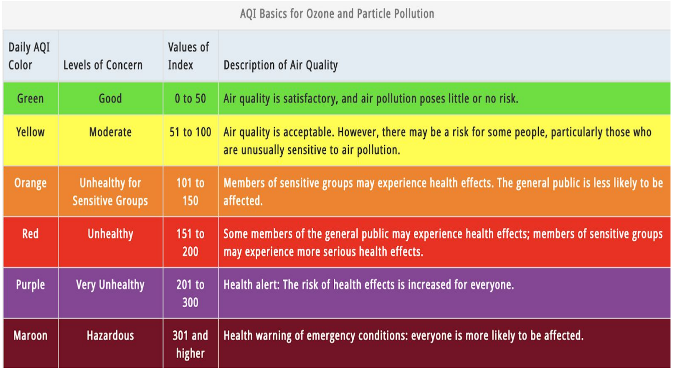
Keywords: air quality, climate, energy, metropolitan statistical area, time series, correlation

1. Introduction

Air pollution is caused by harmful pollutants and substances finding their way into the atmosphere. The Environmental Protection Agency (EPA) holds responsibility within the United States for establishing legislations to protect the health of individuals and the environment. The Clean Air Act (CAA) of 1970 permits the EPA to establish National Ambient Air Quality Standards (NAAQS) for regulation of emissions of air pollutants. The five major pollutants are ground-level ozone, fine particulate matter, carbon monoxide, sulfur dioxide, and nitrogen dioxide. For the purpose of this study, we will be looking at ground-level ozone (O3) and fine particulate matter (PM2.5).

Ozone and particulate matter form differently. According to the EPA, O3 is a result of sunlight striking nitrogen oxides and other hazardous air pollutants, which are primarily produced by industry, factories, and power plants. PM2.5 is primarily caused by a complex reaction between nitrogen and sulfur oxides which are produced mainly by construction sites, industries, and automobiles. These two pollutants are also the most prominent causes of unhealthy air.

The EPA measures air quality by the Air Quality Index (AQI). This can be seen in Figure 1. The air quality health safety varies depending on how harmful it is to individuals in the area. We use this air quality index to understand the differences between areas of our study. We break down the fifteen MSAs by region: Northeast, Central, Southeast, Southwest, and Other. It allows us to understand regional differences and compare their air quality index based on these disparities.



AQI Basics for Ozone and Particle Pollution

Daily AQI Color	Levels of Concern	Values of Index	Description of Air Quality
Green	Good	0 to 50	Air quality is satisfactory, and air pollution poses little or no risk.
Yellow	Moderate	51 to 100	Air quality is acceptable. However, there may be a risk for some people, particularly those who are unusually sensitive to air pollution.
Orange	Unhealthy for Sensitive Groups	101 to 150	Members of sensitive groups may experience health effects. The general public is less likely to be affected.
Red	Unhealthy	151 to 200	Some members of the general public may experience health effects; members of sensitive groups may experience more serious health effects.
Purple	Very Unhealthy	201 to 300	Health alert: The risk of health effects is increased for everyone.
Maroon	Hazardous	301 and higher	Health warning of emergency conditions: everyone is more likely to be affected.

Figure 1: Air Quality Index (EPA)

The Northeast region encompasses humid subtropical areas that have high population densities. This includes the New York, Boston, Philadelphia, and Washington D.C. MSAs. The Southeast holds humid tropical areas of the Miami and Atlanta MSAs. The central region is composed of the Detroit, Chicago, St. Louis and Minneapolis MSAs which have more continental climates, while the Southwest region holds the Los Angeles, Phoenix, and Dallas metro areas which have dry desert climates. The Outlier region incorporates the Denver and Seattle MSAs, which hold unique climates and regional locations that fail to fit within the other designated regions. This breakdown allows us to differentiate between divisions of the United States and discover similarities between the metros and the factors that contribute to air quality.

2. Data

The air quality data comes from the EPA, which provides daily air quality measurements for hazardous pollutants in metropolitan statistical areas from 1985 to 2020. We have annual average temperature measures by state collected by the National Oceanic and Atmospheric Administration (NOAA) also from 1985–2020. The study also includes measures of GDP of industries from the Bureau of Economic Analysis U.S. Department of Commerce, annually from 2001-2018 which is measured by metropolitan statistical area. Alongside GDP, production data from Datausa expresses the strongest economies by metropolitan area, measuring their worth in millions of dollars in the years of 2012-2015, as well as 2020. Also from Datausa, we have commuting data which has risen since the introduction of coronavirus to measure how much people are traveling, which is the daily trend from February 15 to July 31st. The energy data we analyze comes from the Energy Information Association, which holds data by year of the different types of energy and their usage from 2001-2019.

3. Data Visualization & Observation

We use Python 3.7 alongside Jupyter Notebooks as our IDE to visualize and understand the data. We graph every air quality, temperature, energy, production, and commuting graph for each MSA, but we have selected the most significant ones for this paper.

3.1. Air Quality (1985-2020, EPA)

We open this study by graphing air quality trends by MSA. Figure 2 displays the air quality trend for the New York MSA, portraying the general decline in both the seasonal and annual trends of O3 and PM2.5. The annual trend is the average by year of air quality measured by the index, while the seasonal trends measure by monthly rolling averages. In order to see and compare the metro areas together rather than individually, however, we develop a map to get a general understanding of the regional differences.

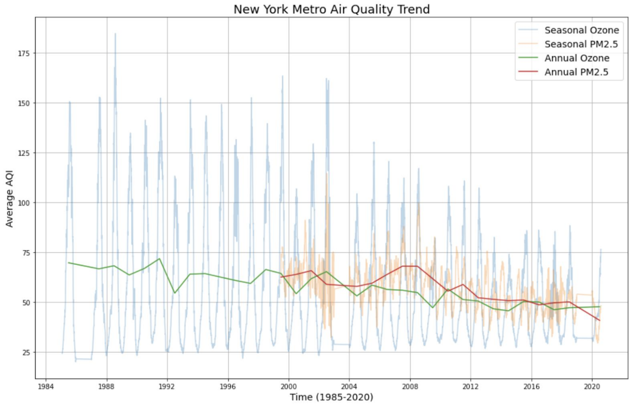


Figure 2: New York MSA Air Quality Trend of O3 and PM2.5, 1985–2020

Figure 3 displays maps of MSAs and their O3 and PM2.5 in the year of 2020. From the graphs, we can see differences in metropolitan area air quality. The Southwest Region metropolitan statistical areas generally have higher concentrations of pollutants in comparison to others. We can see that the Seattle MSA in the top left corner of the Outlier region seems to be the healthiest. The Denver MSA in Colorado also seems to have a larger concentration of O3 in comparison to other areas, but its particulate matter emissions are low. The Central Region generally has a higher amount of particulate matter, while the Northeast is moderate. The Southeast is also moderately healthy.

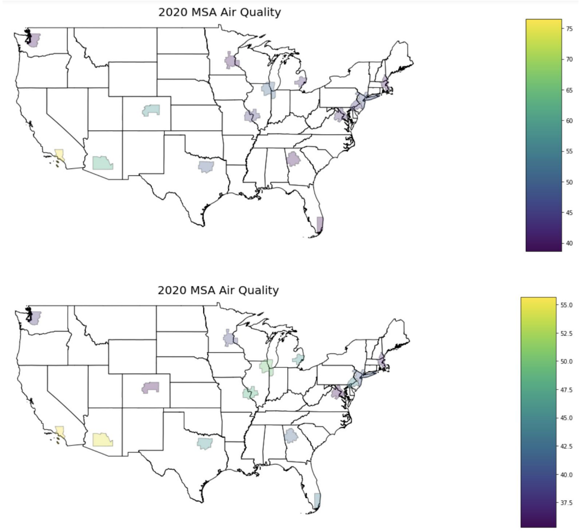


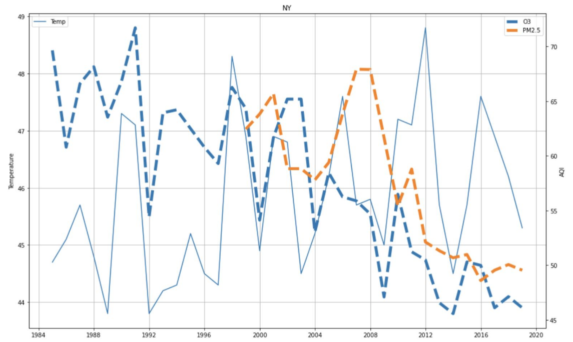
Figure 3: Map visualizations of O3 (top) and PM2.5 (bottom) across the United States MSAs in 2020

We can already see similarities and differences across the metropolitan areas, but now we have to understand what might be causing these differences. Why are certain regions higher in pollution of either O3 or PM2.5 in comparison to others?

3.2. Temperature Trends (Annual 1985-2020, NOAA)

As we understand from the EPA that heat generally causes O3, we take a look at temperature trends in the United States.

Figure 4 displays the temperature trend of the New York MSA (NY-NJ-PA) and compares state level to the air quality of the metropolitan area. We can see the significant spikes and drops in temperature, which can be attributed to the El Niño and La Niña weather cycle. This weather cycle is made up of two



opposite phases: El Niño is the warm phase, and La Niña is the cold phase. However, we can see that the ozone levels follow the spikes and decline of temperature, which is consistent among every single temperature to air quality graph.

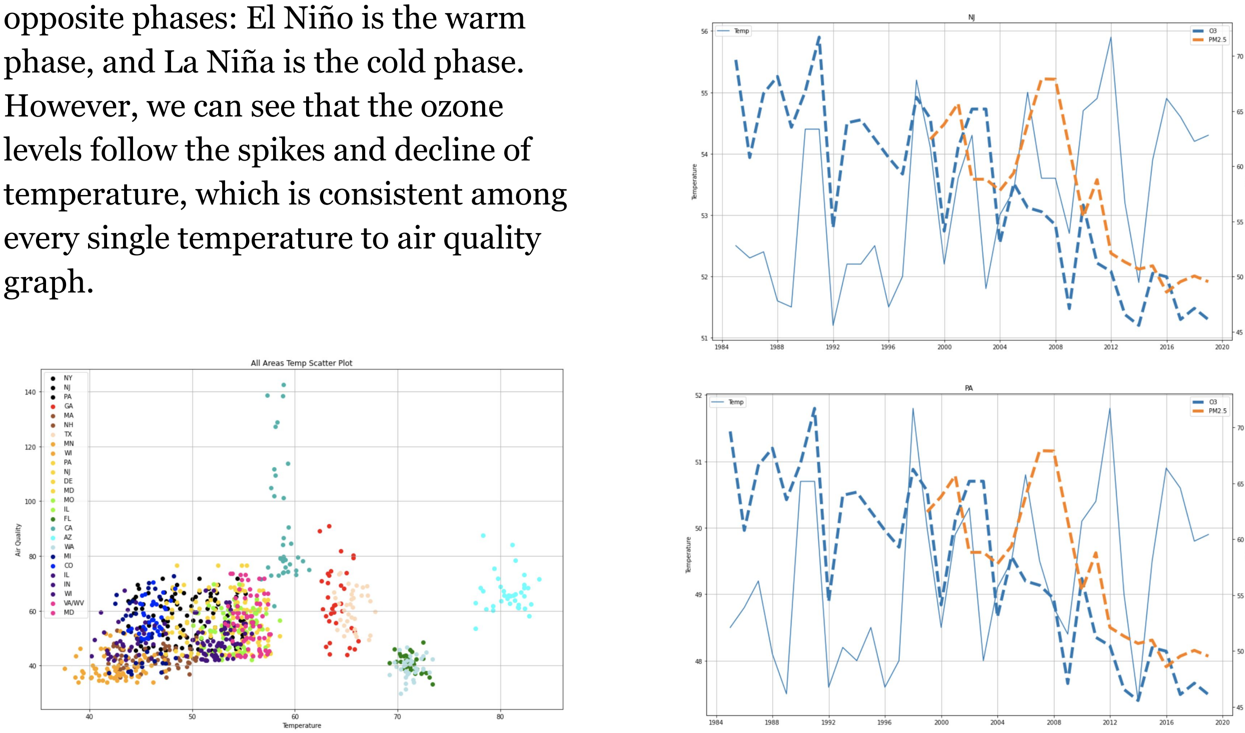


Figure 4: Line plots of NY-NJ-PA and an all metro scatter plot of temperature to air quality across the MSAs from 1985-2020

We can see with the scatter plot that generally, the higher the temperature, the worse the air quality. A place like MN-WI (orange) has a much lower temperature and lower ozone levels while AZ (light blue) has the highest temperature and a higher ozone level. The Los Angeles MSA (CA) presents itself as a clear outlier with poor ozone levels and a moderate temperature, while Seattle (WA) and Miami (FL) have low ozone levels but a higher temperature. However, the slight curve upwards from MN-WI to AZ across the other MSAs as well as the individual graphs that reveal ozone spikes and drops alongside temperature confirms a strong correlation between O3 and temperature.

3.3. Energy Trends (2001-2019)

As fossil fuels are commonly known to produce heat, we can look at energy consumption across the metropolitan areas to understand how energy correlates with air quality. Below, we display certain energy visualizations which will help us dissect our question.

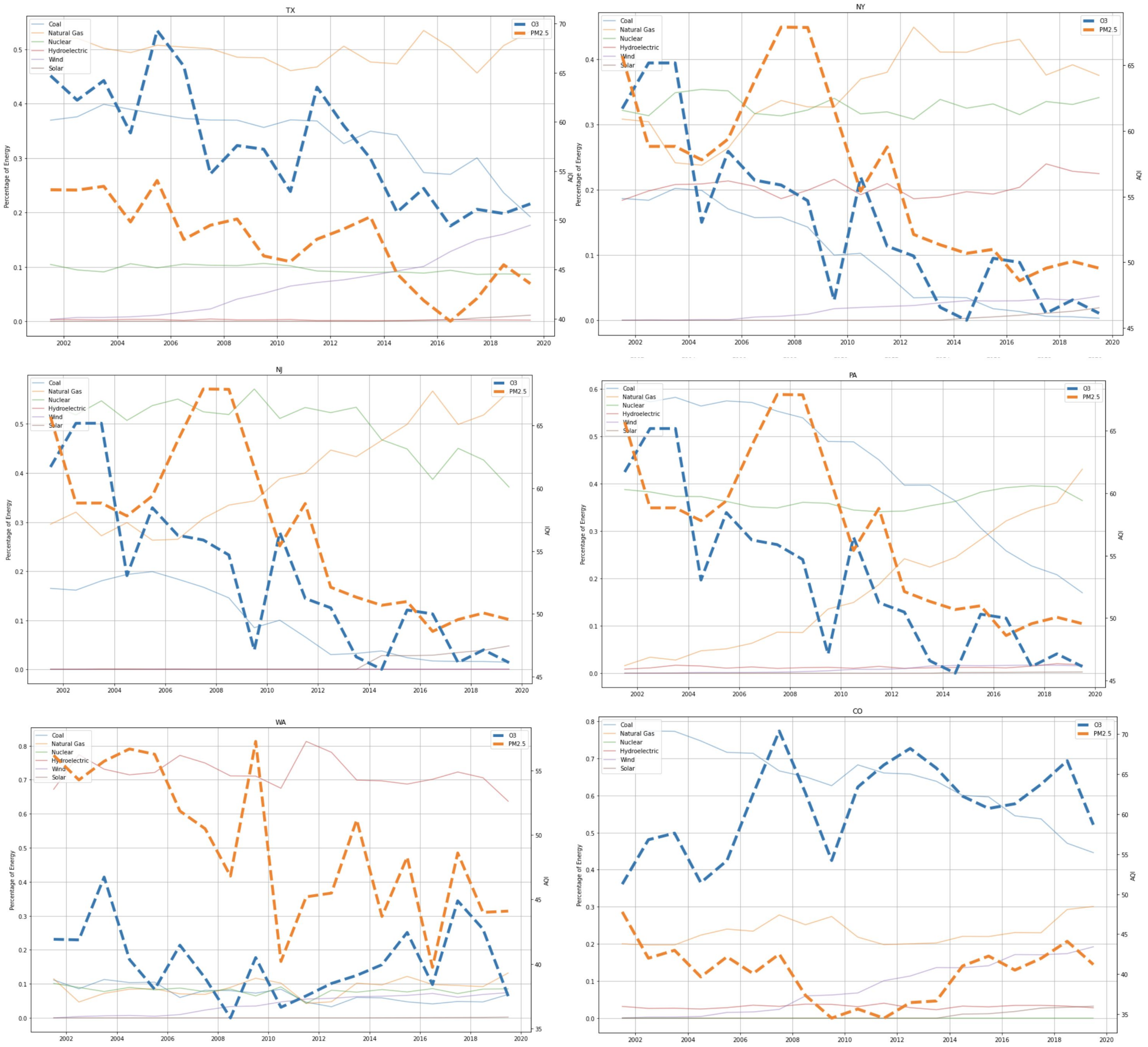


Figure 5: Energy trends of the Dallas, New York, Seattle, and Denver MSAs, 2001-2020

Figure 5 displays the Dallas and New York MSAs, where in each graph we can see that a decrease in coal consumption is followed by a decrease in ozone. We can see that generally, O3 shows a decline alongside a decline in coal. Although PM2.5 declines as well, the Seattle MSA portrays that coal is already low to begin with where the area relies heavily on hydroelectric power, indicating that the already low ozone parallels the already low coal energy percentage. However, looking at the

Denver, CO MSA, we can see that ozone is increasing while coal is decreasing.

Figure 6 displays the temperature trend in the Denver metropolitan statistical area. We can see that temperature here appears to be increasing alongside O3. But why is this the case?

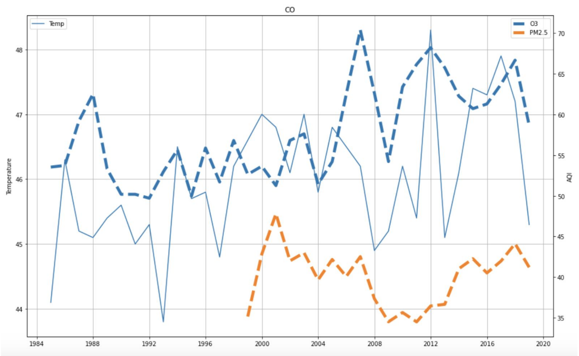


Figure 6: Denver temperature to air quality trend

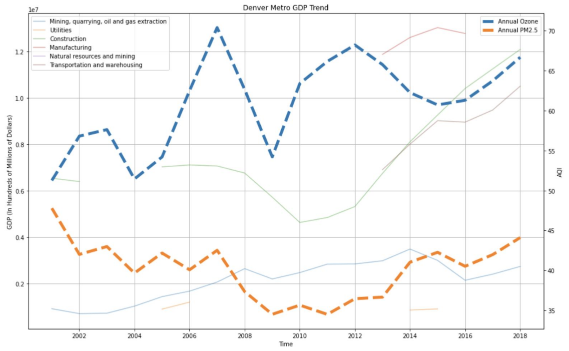
3.4. Production and GDP

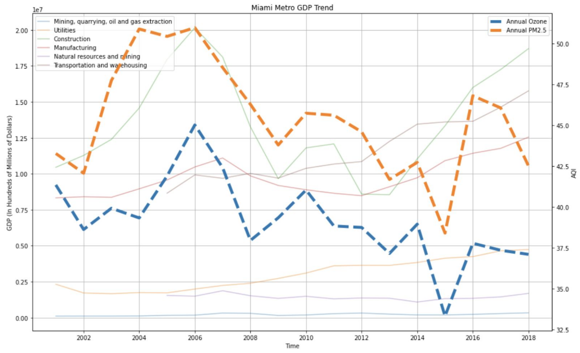
Production and GDP may possibly be another cause for the ozone increase we see in Denver. Let’s take a look at GDP trends and prominent industries within MSAs.

3.4.1. GDP Trends (2001-2018)

The first graph in Figure 7 displays the GDP trend for Denver. We can see that the economy, especially in recent years, has been on the rise: construction has grown significantly, alongside transportation and warehousing as well as manufacturing within the few years of data we have for it. Their growing economy could be a possible reason for the increase in ozone, however we can see that the particulate matter trend follows construction. We know construction is a commonly known producer of particulate matter, and we can see this in the Miami MSA GDP trend graph as well. PM2.5 follows construction very closely.

Looking at the Detroit MSA, the PM2.5 trend seems to decline alongside the manufacturing industry and slightly increases as manufacturing increases again in 2010. Manufacturing includes the





production of motorized vehicles, and Detroit has a prominent automobile industry. Automobiles are also commonly known to produce pollution, specifically particulate matter emissions.

3.4.2. Prominent Production (2012-2015, 2020)

Figure 8 displays the energy production for the Detroit MSA, in millions of dollars spent. The most prominent industry is motorized vehicles. This lines up with the GDP trend in manufacturing, indicating that automobiles hold a strong economy in Detroit. The metro area is also home to the “Big Three” automobile companies: General Motors, Ford, and Chrysler. If motorized vehicle production has a correlation to PM2.5, how has the quarantine of 2020 and the significant drop in commuting affected emissions of particulate matter?

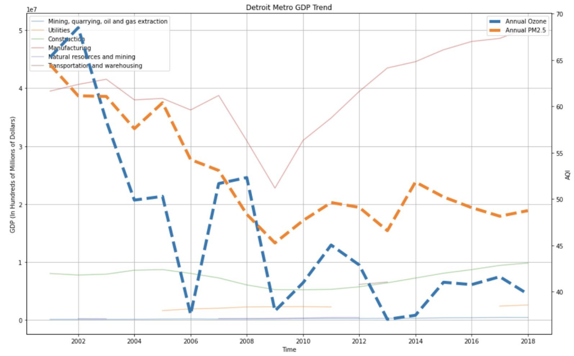


Figure 7: GDP Trends of the Denver, Miami, and Detroit MSAs, 2001-2018

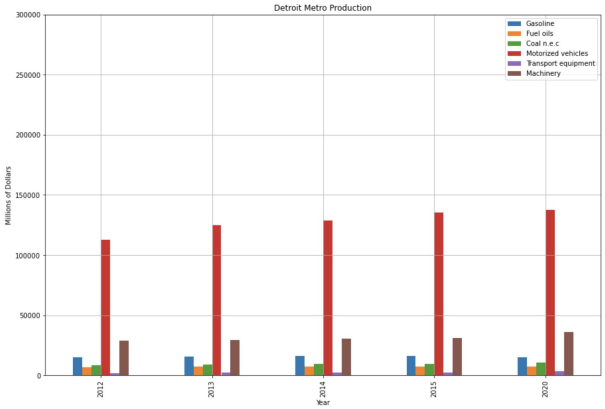


Figure 8: Detroit MSA Production Bar Chart (2012-2015, 2020)

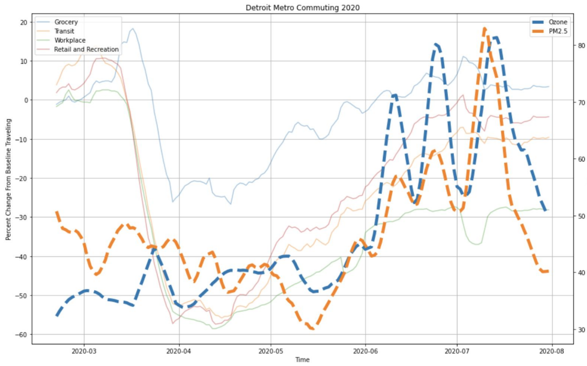
3.5. Commuting Trends (February 15 to July 31st of 2020)

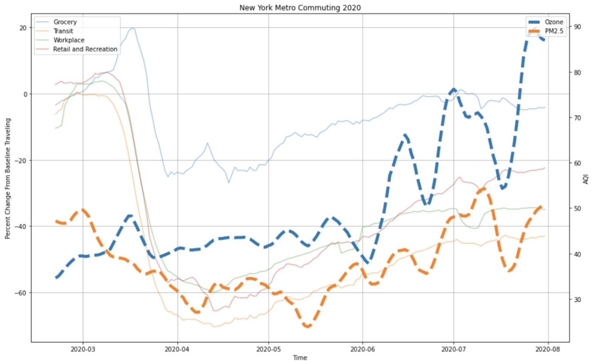
The Detroit MSA in Figure 9 displays a moderate decline in PM2.5 alongside the drop in commuting, and then an increase in both variables. Among the MSAs studied, the Detroit MSA ranks 8/15 in highest population density with 1,104 people per square mile, and produces a significant amount of motorized vehicles.

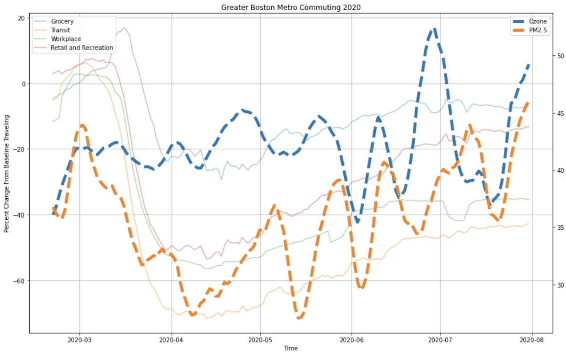
The New York MSA similarly experiences a decline in PM2.5 once commuting drops, and an increase when the area begins to reopen. This metro holds the highest population density with 5,318 people per square mile. It is also commonly known that New York has a significant amount of traffic, indicating that the concentration of people and vehicles may support why PM2.5 would decrease alongside commuting.

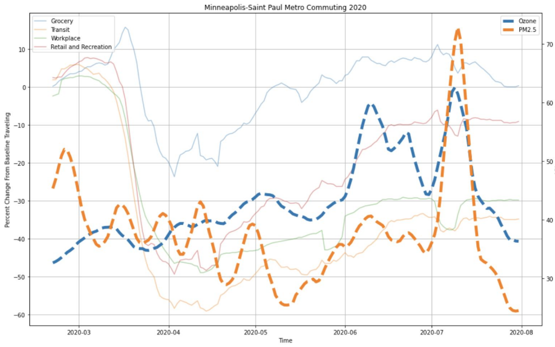
The Greater Boston area also saw a decline in PM2.5 in early quarantine, and ranks 6/15 on population density. Although there are fluctuations in fine particulate matter, it seems to generally increase alongside the reopening of the area.

The New York MSA and Greater Boston MSA are both a part of the Northeast Region, where the other MSAs also have strong correlations between commuting and PM2.5. We can see in Figure 10, the correlation matrix has a strong relation between PM2.5 and types of commuting while O3 has a much lower number. The O3 in all of the commuting graphs seem to increase, however this most likely is due to the rise in temperature of summertime. The Northeast areas all rank within the top 6 MSAs for population density, and they have the









strongest correlations between drops in commuting and particulate matter.

We also take a look at the Los Angeles Metro and the Minneapolis Metro, as they are the other two MSAs within the top 6 in population density. They both have a decline in PM2.5 alongside the drop in commuting.

Los Angeles in particular, however, is known for its “LA Traffic,” which may indicate why it’s particulate matter trend may be more smooth in comparison to Minneapolis. This area also has a prominent motorized vehicle industry: as seen with Figure 11, motorized vehicles are heavily produced within the metro.

In comparison to the other MSAs in the Southwest Region, Los Angeles has a greater density, motorized vehicle production, and is generally known for its traffic. The Dallas MSA ranks 11/15 in population density while the Phoenix MSA ranks 15/15. They also have lower amounts of motorized vehicle production, and their commuting correlation to PM2.5 is extremely weak.

Generally, we begin to see some consistencies across regions. Detroit’s motorized vehicle production and relative density might contribute to why its PM2.5 emissions relate to commuting. Los Angeles expresses similarities, where it also produces automobiles while having a high density,

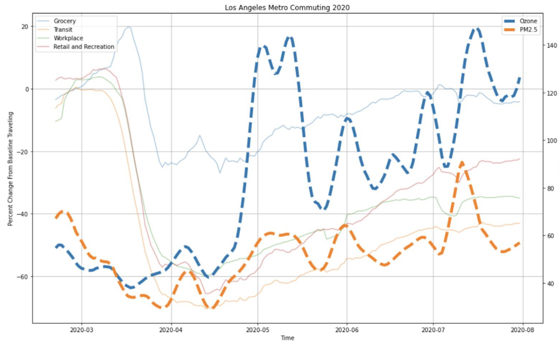


Figure 9: Commuting alongside air quality trends from February 15 to July 31st of 2020

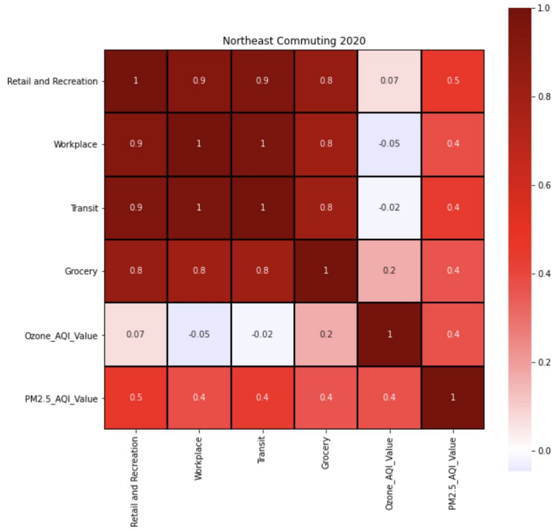


Figure 10: Correlation matrix of commuting and air quality across the Northeast Region

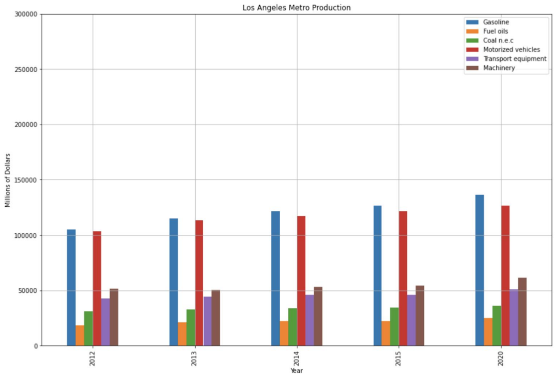


Figure 11: Los Angeles MSA Production Bar Chart (2012-2015, 2020)

and is also known for its congested traffic. The New York MSA similarly has a high amount of traffic and has the highest population density across the MSA. Alongside the other metro areas in the Northeast which all have strong densities, we can see that there may be some consistency between PM2.5, motorized vehicles, population density, as well as general traffic.

4. Conclusions

We can see that O3 and PM2.5 vary across the MSAs in the United States. Our input variables differentiate depending on the area, whether we look at temperature, industry, energy, production, or commuting. However, consistencies can be drawn between specific variables and the major pollutants. O3 is highly correlated with the spikes and drops in temperature. These spikes are most likely due to the weather cycle of El Niño and La Niña.

We then looked at energy consumption as it is commonly known to produce heat, and a strong correlation between coal and O3 can be seen. Generally, as coal energy production declines so does ground-level ozone, and we can see in a place like Seattle where ozone is less variable, it may be because the hydroelectric is the most prominent energy source. We saw Denver's O3 seemed to be increasing while coal decreased, so we graphed the temperature trend and saw that temperature was increasing alongside O3.

To investigate this further, we visualized the production of the Denver MSA which revealed the increase in industry which may correlate to the gradual growth in ozone emissions. We also saw, however, that PM2.5 followed the construction industry, which is commonly known to produce particulate matter as well. Miami expressed similarities, where particulate matter followed the construction industry. However, looking at Detroit, we can see that the GDP of manufacturing seems to be decently correlated with particulate matter. The production of automobiles falls under manufacturing, and when the bar chart reveals the prominence of motorized vehicles within Detroit, it nudged us toward looking at commuting trends, specifically during the coronavirus pandemic.

The drop in commuting with the quarantine reveals that particulate matter drops in areas where there seems to be more traffic, higher population densities, as well as greater motorized vehicle production. We see consistencies and differences across the regional breakdown, leading us to understand what the primary contributors to O3 and PM2.5 may be, and understanding that they vary.

5. Study Limitations

Certain limitations we can address would be the lack of traffic volume data. We assume that places like LA and New York have significant traffic, but we do not have the data to support this. We also have the limitation of comparing state-level to MSA-level data, which may not portray the most consistent or accurate description of the energy and temperature trends. Another limitation would be the time range, where time isn't consistent across all variables, and we have to compare them individually at times. For example, the production data from Datausa only has the years 2012-2015, and the year of 2020. Local geographical effects were also not included within the study: Los Angeles' air quality is harder to improve due to the geography, as LA remains in a low basin surrounded by mountains that trap the atmosphere above the city.

6. Future Investigations

We cover the strongest correlations we see across the metropolitan areas, but other input variables could be investigated for specific areas. For example, traffic volume could be looked at. Other climate variables such as the amount of sunshine, precipitation, or snowfall could be viewed to also see whether there is a correlation between unique weather patterns and air quality.

For specifics, we could look at Arizona and question why the solar energy consumption in the area is extremely low. Is it the cost, or is there legislation or other reasons for not leveraging the Valley of the Sun's light consistency? We could also ask why Los Angeles has hit a plateau in air quality, where it declined significantly but has remained stagnant at a relatively bad air quality.

7. Acknowledgements

I want to firstly thank Jon Stelman, Data Scientist at Renaissance Learning, for leading this internship and making this research project possible. I also want to thank the Learning Sciences & Innovation Team as well as the entire company of Renaissance Learning for showing me what it is like to be a part of a community motivated by education. This experience has been inspirational, furthering my knowledge not only about computer science and research, but also to understand the importance of tackling real issues with data just as the company does. I wish Jon, the company, and the other interns the best of luck in the future.

8. References

"Air Data - AQI Plot." *United States Environmental Protection Agency*, 18 Dec. 2018, www.epa.gov/outdoor-air-quality-data/air-data-aqi-plot. Accessed 31 July 2020.

"Climate at a Glance." *National Oceanic and Atmospheric Administration*, www.ncdc.noaa.gov/cag/statewide/time-series/28/tavg/12/12/1985-2020?base_prd=true&begbaseyear=1985&endbaseyear=2020&filter=true&filterType=binomial. Accessed 20 Aug. 2020.

Datausa. datausa.io. Accessed 27 July 2020.

"GDP by County, Metro, and Other Areas." *Bureau of Economic Analysis U.S. Department of Commerce*, 14 May 2020, www.bea.gov/data/gdp/gdp-county-metro-and-other-areas. Accessed 12 Aug. 2020.

"State Outlines." *ArcGIS Online*, ESRI, www.arcgis.com/home/item.html?id=e9338909b71c4b6389ea2e2407d7b46b&view=list&sortOrder=true&sortField=defaultFSOrder.

"TIGER/Line Shapefile, 2019, Nation, U.S., Current Metropolitan Statistical Area/Micropolitan Statistical Area (CBSA) National." *Data Catalog*, catalog.data.gov/dataset/tiger-line-shapefile-2019-nation-u-s-current-metropolitan-statistical-area-micropolitan-statist.