Air Quality in the United States

Sam Kidd   
*Wentworth*

*Abstract*—Air Quality is an essential part of human life, and so it needs to be studied and understood. This paper looks at various factors that influence air quality and determines how much of an impact they have. It looks at the correlation between wildfires and air quality, the impact human emissions in cities have on their local air quality, and how seasonal effects create patterns in air quality. For each of these factors a different statistical or machine learning model was employed to demonstrate the relationship each of the factors has with air quality. In order for air quality to improve, then the causes of poor air quality need to be understood so steps can be taken to improve or fix those specific problems.

# Introduction

Air quality rates in the United States varies throughout months and years. The quality of the air that we breathe has massive impacts on our health, and so we should try and ensure that it is as clean as possible. Air quality is measured by the amount of harmful pollutants that exist in the air. A part of ensuring that the air we breathe is clean is understanding what factors contribute to lowering air quality. Factors such as industrial production, natural wildfires, and the changes in the seasons all affect the quality. Those factors need to be investigated to see how they correspond to the air quality in the surrounding regions to see how major of a problem they are.

# Datasets

## Source of dataset

In this project I use 3 different datasets to do the analysis of air quality. The first dataset I found is called “U.S. Pollution Data 2000 - 2023”, and it comes from Kaggle [1]. Kaggle is a reliable source for downloading datasets. The author of the dataset created it using data provided by EPA, which is a US government body that works to protect the environment. The second dataset that I used is one called “1.88 Million US Wildfires”, and it also comes from Kaggle [2]. The author used reports from federal, state, and local fire departments to compile the list of all the wildfires that occurred in the United States from 1992 to 2015. The final dataset that I use is called “Emissions Inventory”, and it comes from data.gov [3]. It is specifically from the District of Colombia, and the EPA tracked all sources of emissions starting in 2008 to get a report of how much each source emitted, and what kind of pollutant it is.

## Character of the datasets

All the datasets are CSV files with their entries listed. The first dataset “U.S. Pollution Data 2000 – 2023” is very large with 665,414 entries, as it tracks the number of pollutants in the air for multiple counties in all 50 states over a period of multiple years. It tracks various factors of O3, CO, SO2, and NO2, but of particular importance is the Air Quality Index or AQI of each. This tracks the quality of the air based on those pollutants, and it is what I use to measure the air quality in my analysis. I had to do some preprocessing of this data, as the format of the dates for each entry were different than the other datasets, I wanted to compare, so I just changed the format that the date was written in. The second dataset “1.88 Million US Wildfires” is also very large with 55,366 entries. Each entry includes the size of the fire in acres, the date it was discovered, and the state the fire occurred. I had to do some preprocessing of the data to change the format that the state names are listed. In the dataset it lists each state name as the abbreviation, so they needed to be changed to the full name. It also includes Puerto Rico, which none of the other datasets do, so that state needed to be dropped. The third dataset that I use, “Emissions Inventory” has 52,695 entries of all the different sources of pollution in DC.

# Methodology

## Pearson Correlation

I run a Pearson correlation test on the wildfire frequency vs the O3 AQI in the state of California. This method is a statistical measure to see if there is a linear relationship between the two variables and then quantify that relationship. I want to see if there is a clear relationship between the number of wildfires that are occurring, and the amount of pollution in the air. It is interesting to see what the correlation between wildfire and air pollution is. I used SciPy, specifically SciPy.stats module to get the function pearsonr that calculates the Pearson correlation. It is a simple module to use that only needs one function to find the result. I combined it with a graph of the data to visualize the results.

## Linear Regression

I set up a linear regression to test the relationship between total emission of Criteria Air Pollutant, and the amount measured mean of NO2 found in the atmosphere. The linear regression will try to fit a line to the data to see if there is a correlation. Linear regression seems to be a good choice for this kind of data, as it makes sense to try and test the relationship between emissions into the atmosphere, and the average air quality measured. I use scikit-learn to calculate the linear regression, as it is very simple to generate the model using the built-in LinearRegression function. For this model I only focused on the DC area, as the emissions inventory that is used for the data is only from DC.

## Time Series Model

In this model I want to test and predict how different times of year affect air quality. Different times of year have different air quality, and it is interesting to see those trends. It’s a simple time series model that is developed using scikit-learn linear regression and lag features to try and create a predictive model, as scikit-learn does not have a built-in time series feature. Using a linear regression model for this use case assumes that there is a linear relationship between air quality and time, which may not work perfectly, so the results will need to be interpreted to keep this in mind.

# Results

## Wild Fire Results

When wildfires burn, they produce a massive amount of pollution in the form of particulate matter that is all released into the atmosphere. Not all parts of the country experience wildfires, but the areas that have them frequently can be expected to see reflected in their air quality.

A blue dotted line with a black border

Description automatically generated

Figure

The graph of a Pearson correlation here demonstrates the relationship between wildfire frequency and the average O3 air quality index. O3 is a pollutant that is released when wildfires are burning. This specific correlation is only looking at wildfires that occur in California, and the air quality in the region in the time after the wildfire. When running the Pearson correlation, it returned a coefficient of .6, which means there is a moderately strong positive correlation between the frequency of wildfires and the average O3 AQI. When wildfire frequency is below 20 wildfires, there is a large spread of air quality measures, from very good air quality below 20, up to 60, which is much worse. As the number of wildfires increases, the average air quality gets worse, as there are almost no data points below 40 O3 AQI once the frequency passes 20 wildfires. This suggests that there are many factors that contribute to the amount of O3, so it can be hard to predict the exact air quality if there have not been many wildfires. Once the number of wildfires increases though, there are no more days when the O3 AQI is low, so larger numbers of wildfires must be contributing a significant amount to the levels of pollutants in the air. Along with the Pearson coefficient of .6, the p-value from the graph is very small at 7.77e-19, which means that the correlation is statistically significant, and there is a linear relationship between the frequency of wildfires and the amount of O3.

## Total Emissions Results

Cities produce massive amounts of pollution from a wide range of sources. Industry, automotives, and residential buildings all are using power or burning materials that produce various polluting gases. NO2 is a gas that is primarily produced from burning fossil fuels, and it can be dangerous for people to breathe, so it is tracked as a part of air quality.A red line with blue dots

Description automatically generated

Figure

Fig. 2 is a linear regression that demonstrates the relationship between total emission and NO2 in the District of Colombia. Each data point on the graph represents a year that the data was measured, and it shows that generally in the years where total emissions were higher there was a higher mean NO2 recorded for the year. The value for this linear regression was only .2831, which suggests that this model is not very strong. The is quite low, however this can be partly explained by the complexity measuring the environment and determining what factors influence it. It is clear that there is some relationship between the total emissions, and NO3 in the atmosphere, but the emissions do not explain the entire picture.

## Air Quality Over Time

Earth’s climate is a complex system that is constantly changing. The changes in weather throughout the year can potentially influence the air quality at different times. Differences in temperature and wind patterns change how the air moves, which can have measurable effects on the air quality.

A red and blue graph

Description automatically generatedFigure

Fig. 3 is a time series model that shows the air quality over time in blue, along with the predicted air quality in red based on the training of the data. The graph demonstrates the seasonal fluctuation of the O3 AQI, with it peaking in the midsummer, and dropping in midwinter. The time series model using linear regression can predict the peaks and valleys of the data, however it does not show the peaks as high as they are in reality and does not show the minimum value as low as they are in reality. This suggests that the seasonal effects on air quality are present and play a significant role, however factors other than time also contribute to pollution in very significant amounts. The model does not account for any of these other factors, but it does show that predictions can be made on air quality based on the time of year. Seasonal effects can change how much certain pollutants remain in the atmosphere, particularly O3, as higher temperatures can encourage the formation of it near the ground through reactions with sunlight. This tracks with the data and predictions that show summers as having the highest levels of O3.

# Discussion

Each of the three models that were used to analyze the data showed interesting results with some limitations. The linear regression analysis of the total emissions could be improved with some simple additions, such as more data. The data available only had five years available, even though there was a significant amount of data for each of the years, taking the average of the data for each year gave the best results. Additional years in the data could be useful to see if the trend continues. The time series model had some significant limitations on it, some of which came from the use of Scikit-learn, which does not have any time series functionality. Some future work in this field should include a better time series model that does not rely on a linear regression to approximate it. For all the models it would also be very useful to include more pollution types in the analysis, as only looking at O3 or NO2 introduces blind spots to the other pollutants that may have different effects on the environment and human health, as well as different sources where they are produced. Overall, this project was a good introduction to the study of the environment and the atmosphere, but much more work can be done to improve the results.

# Conclusion

The environment and air that people breathe is affected by numerous factors which are constantly changing. This can make it difficult to try and find trends and sources for pollution in the air, but there are a few key areas that can be studied. Wildfires burn through thousands of acres of land, the entire time they are releasing gas and harmful particulate matter into the atmosphere. Tracking the AQI of O3 and the number of wildfires reveals a moderately strong linear relationship between the two. As the frequency of wildfires increases the average AQI of O3 gets worse, which makes sense as the products of the wildfires burning are harmful to humans and so we should try and limit the number of wildfires that occur. There is also a link between the total emissions of a city, and the amount of NO2 in the atmosphere. Cities are major centers of production, and other human activity, so they produce very large amounts of pollutants, and the mean amount of NO2 can be correlated with emissions over any given year. If people want to improve the quality of the air they are breathing, then the emissions of the city will need to be reduced, such as a smaller number of automotive vehicles, or developing more clean methods of industry. Human emissions are not the only factor in air quality, as the time of year can play a major role. Differences in temperature, sunlight, and windspeed all play roles in the different pollutants in the atmosphere at any given time. Differences in wind can disperse the pollutants leading to cleaner air, or the higher temperatures can facilitate the formation of harmful compounds. The air we breathe is one of the most important aspects of long-term health, and so we should do what we can to make the air as clean as possible, and the first step in ensuring the air is clean is to understand why air quality worsens in the first place.

##### References

1. “U.S. Pollution Data 2000 - 2023”, [https://www.kaggle.com/datasets/](https://www.kaggle.com/datasets/guslovesmath/us-pollution-data-200-to-2022) [guslovesmath/us-pollution-data-200-to-2022](https://www.kaggle.com/datasets/guslovesmath/us-pollution-data-200-to-2022)
2. “1.88 Million US Wildfires”, [https://www.kaggle.com/datasets/](https://www.kaggle.com/datasets/rtatman/188-million-us-wildfires)

[rtatman/188-million-us-wildfires](https://www.kaggle.com/datasets/rtatman/188-million-us-wildfires)

1. “Emissions Inventory”, https://catalog.data.gov/dataset/emissions-inventory