Evan Yip

Data 512

8 November 2023

Part 1 - Common Analysis

Part II Figures

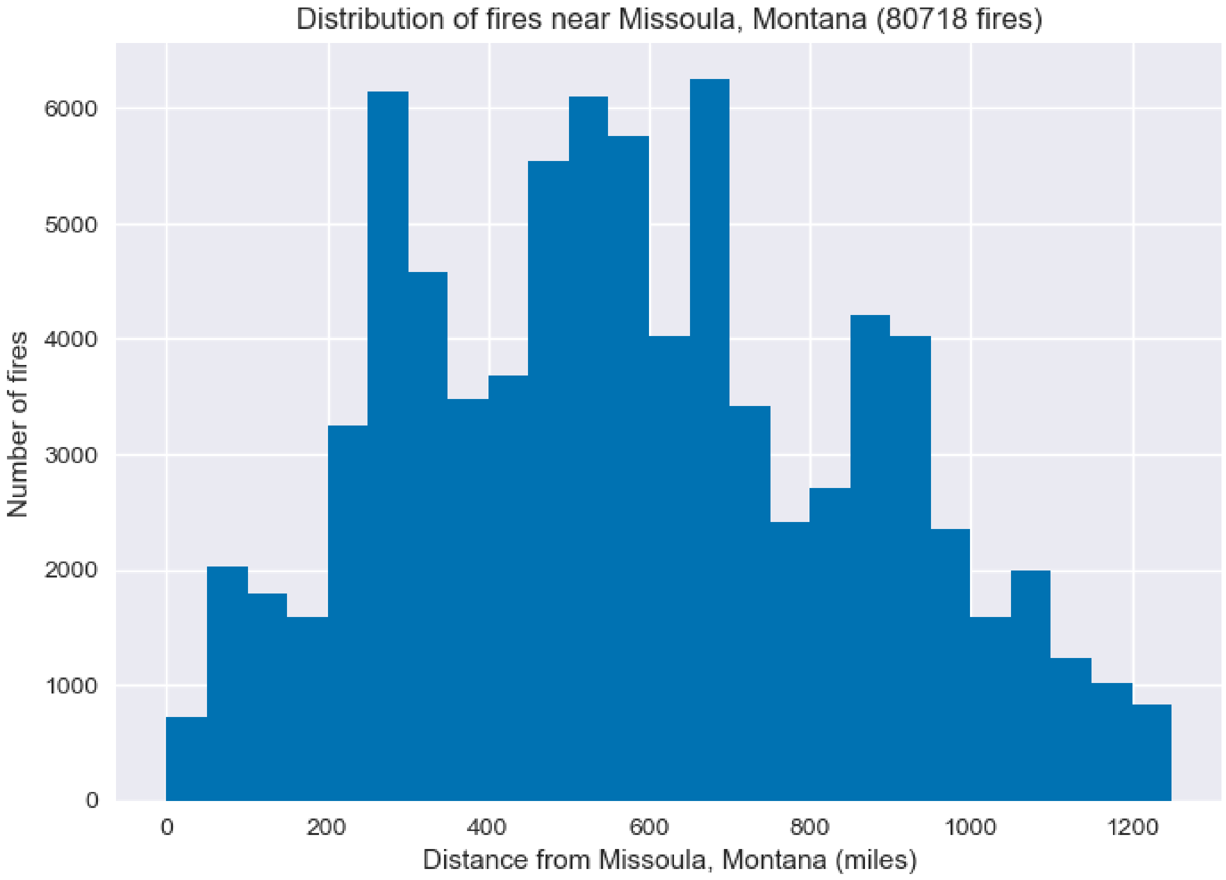

Figure 1: Distribution of fires binned by distance from Missoula, Montana (50 mile bins).

The figure above displays the distribution of the number of fires, both wildfires and prescribed fires that occurred within 1250 miles from Missoula, Montana. Each bin is 50 miles wide and the y-axis displays the frequency, or number of fires that fell within that distance range from Missoula. From the visualization we can see that the fires are roughly normally distributed around about 550 miles from Missoula, Montana. The fire data was collected from a compiled data source from the Combined wildland fire datasets for the United States and certain territories, 1800s-Present (combined wildland fire polygons) dataset. In order to compute the distances from Missoula, we computed the shortest distance from Missoula's coordinates of [46.8721, -113.9940] to the estimated outer perimeter of the fire.

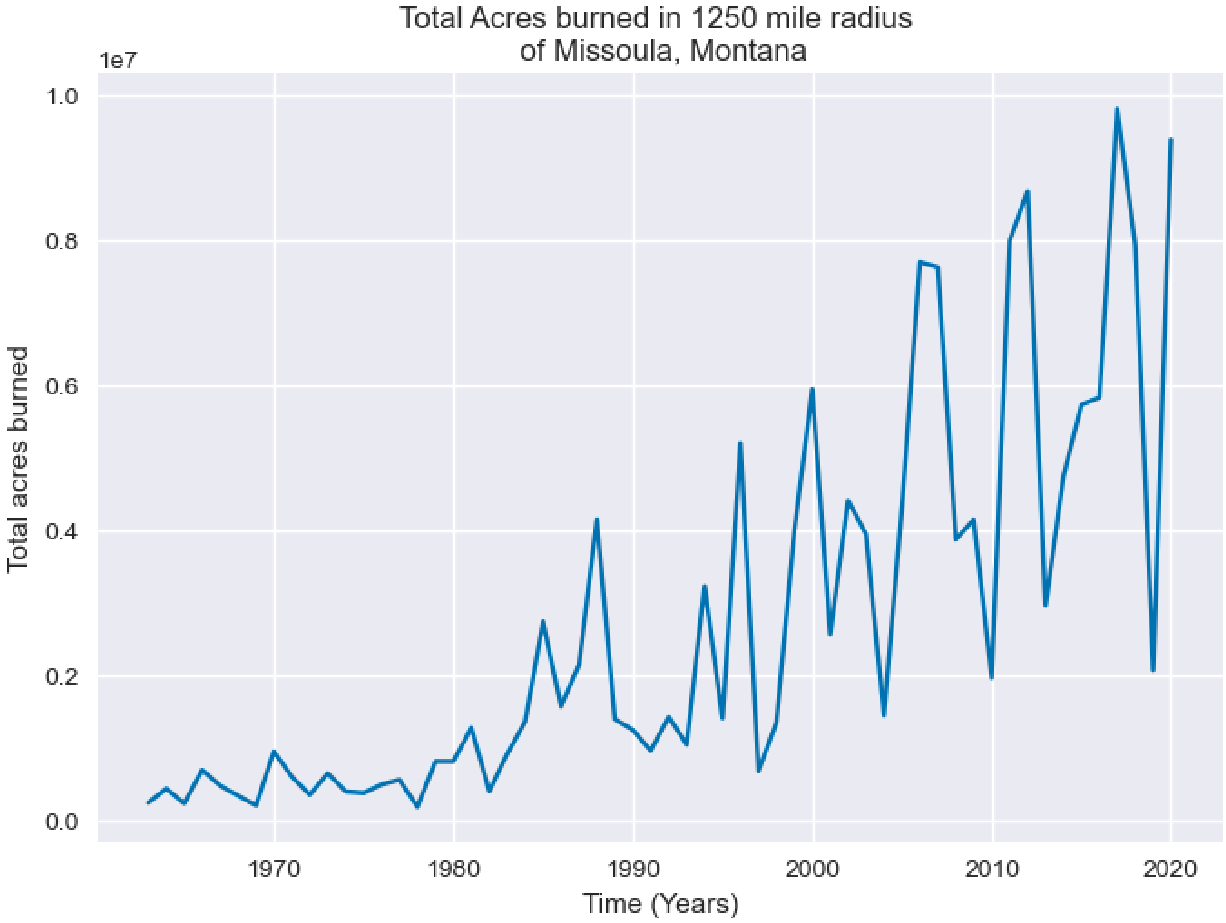

Figure 2: Time series graph of the total number of acres burned per year within a 1250 mile radius of Missoula, Montana.

The figure above displays the time series data of the total number of acres burned in a 1250 mile radius of Missoula, Montana from 1963 to 2023. Looking at the graph we can see that we can see the general trend that it appears the total acreage being burned has been increasing at a seemingly exponential rate. The y-axis in this plot is total acres burned, it is important to note that 1.0 in this case represents 10,000,000 acres as the scale is 1e7. This data comes from the same combined dataset that Figure 1 utilized. In this case we simply needed to aggregate the data on years and sum up all of the acres that were burned to transform the data to this format.

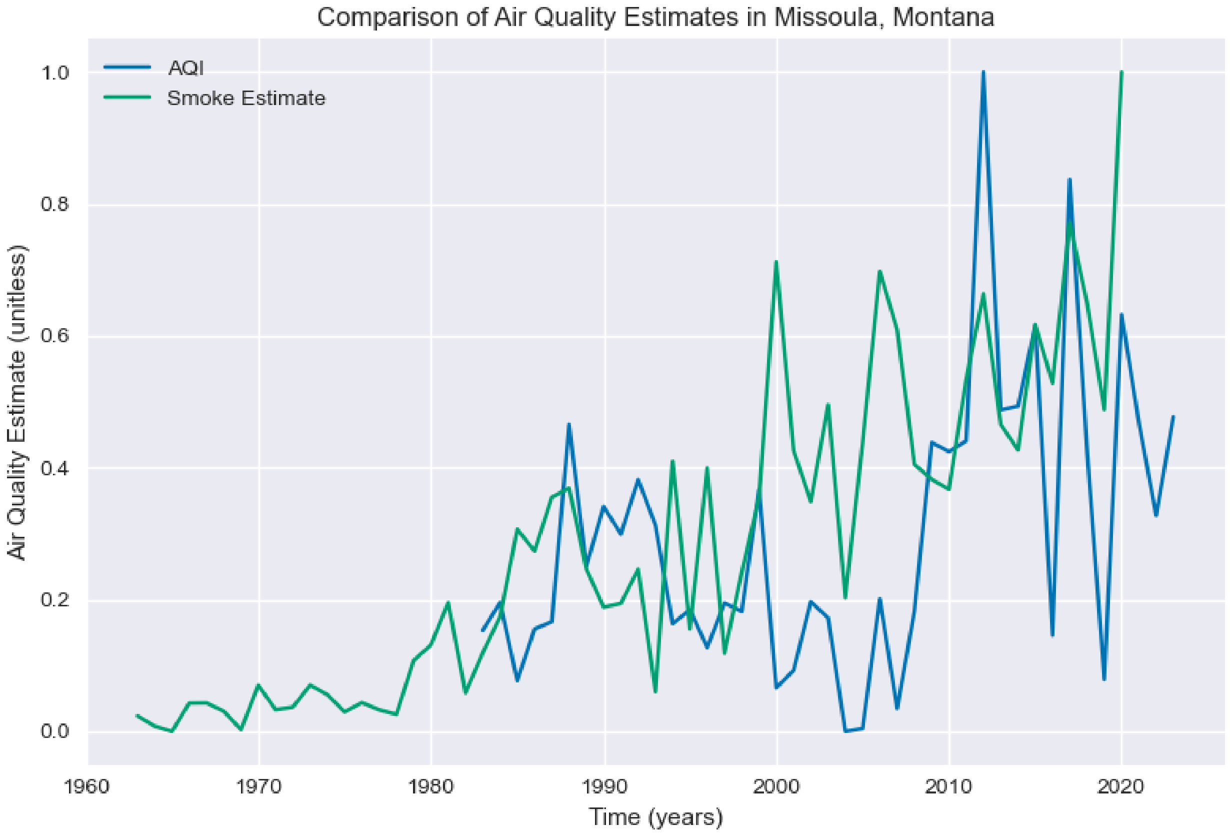

Figure 3: Time series graph of my calculated smoke estimate compared to the AQI over time.

In the figure above, I have plotted my computed smoke estimate against time compared to the EPA AQS air quality index (AQI) over time. While the x-axis is relatively simple, representing the time in years, the y-axis is a unitless measure of air quality. The AQI and smoke estimates were both scaled via standard scaling and min-max scaling methods respectively to put them on the same axes. The scaling methods were chosen after looking at the distribution of values. AQI, which was distributed more normally, was scaled via standard scaling, whereas the smoke estimates which were right-skewed were scaled via max-min scaling. The smoke estimate was derived as a function of four features the year, acres burned, distance from Missoula, and the fire type (wild or prescribed). The formula was as follows:

$$SE(i) = \begin{cases} \sum_{j=0}^{n}(1 * \text{acres}_{ij}/\text{distance}_{ij}) & \text{;if } \mathit{wildfire} = 1 \\ \sum_{j=0}^{n}(0.25 * \text{acres}_{ij}/\text{distance}_{ij}) & \text{;if } \mathit{wildfire} = 0 \end{cases}$$

In this equation, *i* represents the year and *j* represents a unique identifier for each wildfire. Additionally the acres and distance features were scaled via standard scaler and then put through a sigmoid function to put their values between 0 and 1. Thus in more simple terms, for each year we will iterate through all of the wildfires, dividing the scaled acres burned by the scaled distance from Missoula and multiply by the wildfire type constant. Finally, we will sum up all of these computed components for our smoke estimate for that year. The reasoning behind this equation is that I wanted to capture the inverse relationship between distance and smoke but also capture the multiplicative relationship between the size of the fire and how close it was to the city. The constant that was chosen (0.25) to represent prescribed fires was chosen with the assumption that prescribed fires are likely to produce less smoke than wildfires. The reason for this is because prescribed fires are often planned and controlled in a manner that should have a minimal impact on surrounding cities.

Looking at my smoke estimate we can see that it roughly follows the same trend as the AQI. One notable difference is the higher levels predicted by my smoke estimate between 2000 and 2010. We can also see that we have more smoke estimates for the early years (1963 - 1980) than AQI, this is because there likely weren't any air quality monitoring stations back then.

Collaboration Reflection

In this assignment, I learned how to iterate and develop my own mathematical model, apply it to data, and evaluate its performance. Coming from a bioengineering background, this is a familiar process. As an engineer we often build models to try to build simple models to try and capture the behavior of reality. The process involved thinking about the relationships between covariates and the response variables. The really interesting part of this project was being able to extract and compile all the data, build the model, and evaluate it against other proxies (AQI). It truly was an end to end data science project that required a lot of different skills.

For this assignment I collaborated with three different classmates, Mark Qiao, James Joko, and John Michael. Although neither of us shared any code snippets or raw code we had frequent conversations working through the processes of the assignment. It was helpful to give each other tips and strategies for tackling the different problems. For example, at times we talked about tips at using the starter code for using the EPA API to extract the air quality index data. Being able to validate and check processes for how other students were tackling these problems helped give me confidence that I was on the right track. Additionally, we would talk about high level what our ideas were for composing our smoke estimates. Though we did not arrive at the same equation or conclusion it was helpful to think about how other students were thinking about how certain features may be correlated with the response variable.