Common Analysis Write Up and Reflection

Tejal Kolte

Visualizations

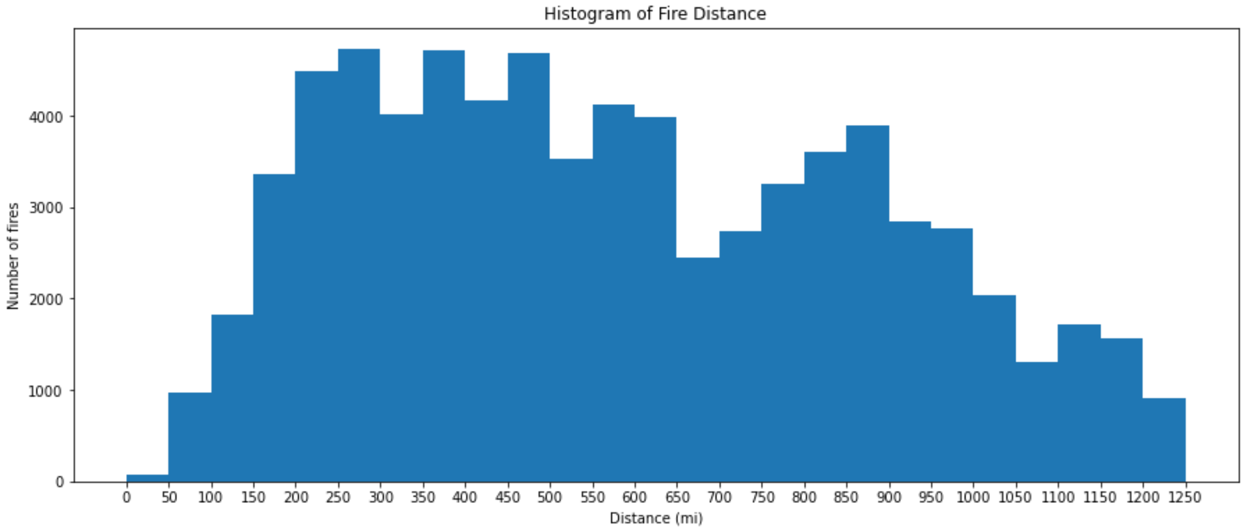

Figure 1: Histogram of Fire Distance Occurring every 50 Mile Distance

The visualization above is a histogram showing the distribution of fire distances from Longview, Washington over the last 60 years, ranging from 1963 to 2020. The maximum distance was specified to be 1,250 miles from the city. The x-axis represents the distance of the fire from the city and utilizes miles as units. The bins of the histogram have been created using increments of 50 miles. The y-axis represents the number of fires and uses a unit of one to represent one fire.

This histogram was created using the *Combined wildland fire datasets for the United States and certain territories, 1800s-Present (combined wildland fire polygons)* dataset, published by the United States Geological Survey (USGS). This dataset was first filtered to include the year range specified above. Then, geographic coordinate rings were extracted for each fire. The centroid of each fire was calculated, and the subsequent distance from Longview was determined. I chose to use the centroid of the fire region instead of another measure as I believed that this would result in the most accurate estimates. I made the assumption that fires with a centroid closer to the city would result in a higher smoke impact than fires with a centroid further from the city. Once each distance was computed, I filtered the list of fires to include those that had a distance of less than or equal to 1,250 miles. This subset of data was used to create the visualization.

From the visualization, we can see that the distribution of distances is approximately bimodal. There appears to be a peak around 400 miles, and another peak around 900 miles.

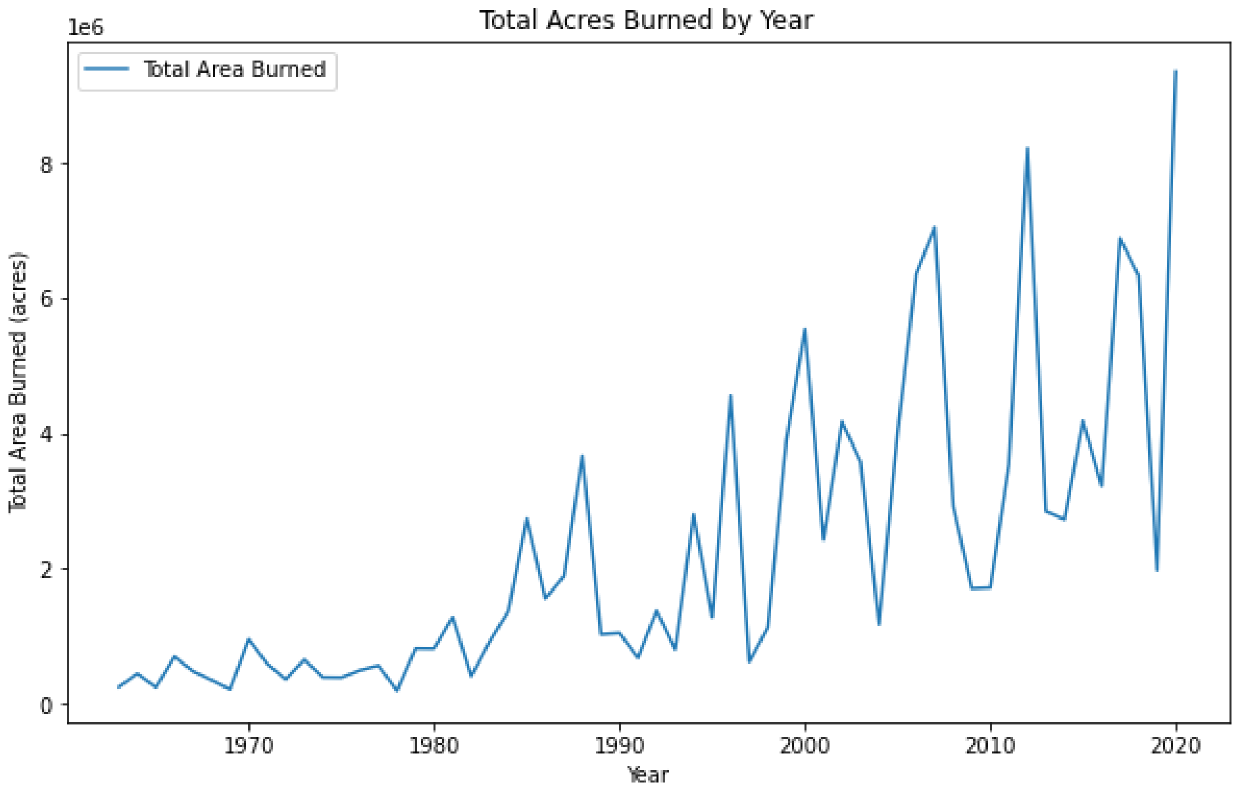

Figure 2: Time Series Graph of Total Acres Burned per Year

This visualization is a time series graph looking at the total acres burned per year by the fires within a 1,250-mile radius of Longview, Washington. The underlying data used to create this graph was also from the United States Geological Survey's *Combined wildland fire datasets for the United States and certain territories, 1800s-Present (combined wildland fire polygons)* dataset. Once the dataset was filtered to include fires within the specified distance of the city, I created an aggregated table that summed the total acres burned by year. I used these yearly summations to plot the data above. The x-axis of the graph represents the year and ranges from 1963 to 2020. The y-axis of the graph represents the total area burned and uses the units of 1 million acres.

The figure shows that the number of total acres burned per year has been steadily increasing over time, though there is great fluctuation with many peaks and troughs. Towards the later years, the peaks appear to occur periodically, with an approximate frequency of 10 years, as shown by peaks around 1999, 2009, and 2019. Additionally, it is important to consider variations in data quality within this dataset. According to NASA[1], there is large uncertainty in wildfire data prior to the early 1980s, when satellites began to map

[1] Voiland, Adam. "Building a Long-Term Record of Fire." *NASA Earth Observatory*, NASA, earthobservatory.nasa.gov/images/145421/building-a-long-term-record-of-fire. Accessed 8 Nov. 2023.

fires. Fire-spotting airplanes were used before this and could not produce results with the same accuracy. This fact could explain why the influx in total acres burned begins after 1980.

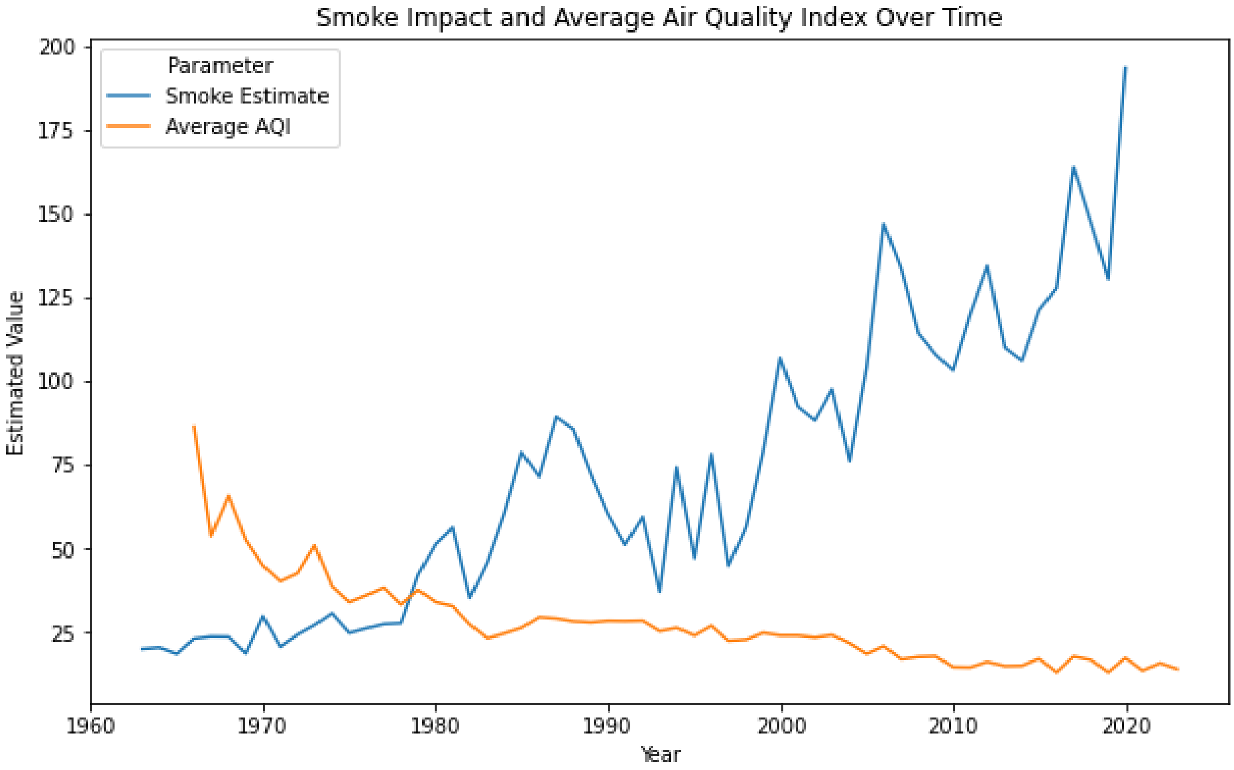

Figure 3: Time Series Graph of Smoke Impact and AQI Estimates

This plot is a time series graph that portrays my Smoke Impact estimates and AQI estimates for each year from 1963 to 2020. The x-axis represents the year, and ranges from 1963 to 2020. The y-axis represents the estimated value.

I created my smoke impact estimates using the USGS wildfire dataset detailed above. To create my estimates, I first assigned a score to each fire in my subset of data (previously filtered by the specified year range and distance radius from Longview). I gave each fire a fire a score between 0 and 100, by scaling the distance of the fire. This score was based on an inverse linear relationship; fires that occurred at a closer distance received a higher score, and fires that occurred further away received a lower score.

I scaled this estimate and considered 3,000 miles to be the greatest possible distance, even though we only considered fires that occurred within 1,250 miles of Longview. This is because, an article by NPR[2] described how smoke from recent fires back in 2021 affected people up to 3,000 miles away. Next,

[2] Fischels, Josie. “The Western Wildfires Are Affecting People 3,000 Miles Away.” *NPR Environment*, NPR, 21 July 2021, www.npr.org/2021/07/21/1018865569/the-western-wildfires-are-affecting-people-3-000-miles-away.

I assigned each fire a score from 0 to 100 based on the area covered by the fire. In this case, I used a linear relationship, as I assumed that fires with a greater area would result in a greater smoke impact. An article published by USA Today[3] outlined the biggest wildfires in the history of the United States. As the 2020 wildfires in California were the largest and covered an estimated 4.4 million acres, I considered this value to be the greatest area.

I combined the distance score estimate and area score estimate to assign an overall fire estimate for each fire, ranging from 0 and 200. I chose to weight the distance and area measures equally, as I believed they both contributed equally to smoke impact. I then computed an average score by year.

I calculated an estimator for the yearly number of fires by scaling the number of fires to a value from 0 to 2.5. According to a Wikipedia article[4] regarding the California wildfires (2020) mentioned above, there were 9,639 fires that took place. Thus, I rounded up and considered 9,650 to be the largest number when scaling. I multiplied my earlier fire score by this estimator, as I believed that the number of fires that occurred in the area had a multiplicative impact on annual smoke. As our fire score estimates ranged between 0 and 200, our final estimator ranged between 0 and 500. I chose to use a scale of 0 to 500, as this matched the AQI scale.

To obtain AQI data, I utilized the U.S. Environmental Protection Agency's Air Quality System API. I set a bounding box of 100 miles to access data from monitors within a 50-mile radius of Longview. I collected data for the following parameters: *Carbon monoxide, Sulfur dioxide, Nitrogen dioxide (NO2), Ozone, PM10 Total 0-10um STP, PM2.5 - Local Conditions,* and *Acceptable PM2.5 AQI & Speciation Mass*. I averaged each of these measures to estimate the Average AQI per year. I plotted this measure and my smoke estimate measure by year.

From the visualization, it is clear that my smoke impact measure greatly varies from the AQI measures. My smoke estimate steadily increases year over year, whereas the AQI measures appear to be decreasing over time. This is likely because my smoke estimator places a greater weight on the number of fires that occurred each year, which was steadily increasing.

In addition, my smoke estimate and AQI measures are not based on the same factors. As mentioned earlier, my smoke estimate is based on the number of fires that occurred each year, the distance of each fire from Longview, and the acres of land the fire burned. On the other hand, AQI measures focus on a subset of pollutants. According to the World Health Organization[5], wildfire smoke is comprised of particulate matter, nitrogen dioxide, ozone, aromatic hydrocarbons, and lead. Our AQI measures do not consider pollutants such as aromatic hydrocarbons and lead.

[3] Fine, Camille, and Chris Lange. "The 10 Biggest Wildfires in US History." *USA Today*, Gannett Satellite Information Network, 8 June 2023, www.usatoday.com/story/news/nation/2023/06/08/largest-us-wildfires-history/70302872007/.

[4] "List of California Wildfires." *Wikipedia,* Wikimedia Foundation, 1 Nov. 2023, en.wikipedia.org/wiki/List_of_California_wildfires.

[5] "Wildfires." *World Health Organization*, World Health Organization, www.who.int/health-topics/wildfires#tab=tab_1. Accessed 8 Nov. 2023.

Although both measures range from 0 to about 500, the categories vary as well. AQI measures are categorized by the following groups: 0 to 50 is considered *Good*, 51 to 100 is considered *Moderate*, 101 to 150 is considered *Unhealthy for sensitive groups*, 151 to 200 is considered *Unhealthy*, 201 to 300 is considered *Very Unhealthy*, and 301 and higher is considered *Hazardous*. As my smoke measure was linearly scaled from 0 to 500 using measures of the most severe historical fires, I considered values around 500 to be hazardous, rather than values starting at 300.

Another factor to consider is the reliability of the earlier AQI data. According to the Environmental Protection Agency, data starting in 1980 should be used to track trends as this marked the start of nationally consistent operational and quality assurance procedures for air quality monitoring. Although data from previous years is available, there is uncertainty in the results.

Reflection on Collaboration

The ability to collaborate in this assignment greatly helped and influenced my thinking of the problem at different stages in the process. Firstly, it was very helpful to discuss improvements in ways to obtain and clean the data. Wanwei Huang shared the idea of using the tqdm package in my code. As I was running many for-loops, some of which took a very long time to complete, this package allowed me to track my progress and ensure that my code was running properly. Using this package, I could showcase a progress bar at the bottom of my code chunk and avoid using print statements that would otherwise clutter my notebook. This technique especially came into use when I converted my geographic rings from the ESRI:102008 coordinate system to the EPSG:4326 coordinate system, a process that took approximately two hours.

In addition, when running the Air Quality System API, I struggled with getting the JSON results data in a workable format. Jenny Wong gave me the idea to convert my API responses into pandas data frames. This method made it much easier to filter and manipulate the data in the later steps, as each feature was now stored in a column.

It was also very insightful to discuss different approaches for constructing our own smoke impact estimates and prediction models. Although we all created our own estimators, I discussed different ideas with Wanwei Huang, Jenny Wong, and April Gao. We considered the pros and cons of using various variables in the dataset, and ways to weight each variable to create an overall estimate. It was interesting to hear different perspectives on which variables deserved a greater weight, and whether or not to include a multiplicative factor in the overall estimate. I personally decided to use a weighted average for the distance and area variables, and a multiplicative factor for the number of fires. On the other hand, my classmates explained how they planned to use simple linear regression models to create their own estimates.

As for prediction models, Wanwei Huang introduced me to the Meta Prophet machine learning model. It was fascinating to see how machine learning models could be used for time series analysis and predictions. As I had previously thought about the prediction model through a univariate lens, this discussion showed me how the problem could be approached using multiple variables. I also shared my idea to use the ARIMA time series model, as I wanted to use my smoke estimate and use auto-regressive and moving average parameters. I had previously taken a time series class that covered these methods, so it was nice to share my prior knowledge with my classmates.

Overall, the possibility of collaboration significantly affected my approach to this project. It was great to share code snippets and packages to improve the efficiency of obtaining and cleaning the source data. As the tasks were open-ended, I also greatly enjoyed discussing different methods with my classmates. Even though we all decided to use different estimators and prediction models, I learned a lot about different statistical methods and the reasoning behind them. It was also helpful to explain my own decisions and processes during my discussions with my classmates. In answering their questions, I considered potential weak points in my own decision-making process and strengthened my justifications to take certain actions.