Kyle Sorstokke

Project Report

Common Analysis

In part 1 (common analysis), we set the groundwork for the rest of the project by performing initial exploration into the fire data and air quality metrics. From the exploratory data analysis, we can see that Albuquerque has a fairly high number of fires within 650 miles and a steep drop off as distance increases, likely due to the mountainous and desert regions in the area. Likely also due to regional weather and geography, we see large swings in fire acreage from year to year; contributing factors include intermittent regional droughts and shifting wind and weather patterns for which the southwest is known.

In this section, I calculated the fire impact estimate that I would use throughout the rest of the project, as a cumulative function of fire size and its inverse distance from Albuquerque. I also calculated and compared a summative estimate of annual air quality. Because of the summation and transformation process, the units largely lost meaning outside of measuring magnitude, but these values are still useful for chronological trends and comparison. We see that, while not always accurate, the fire estimate does a decent job of following the air quality at least in terms of overall trend.

I then use the fire estimate to predict fire impact going forward to 2050. The prediction model has somewhat of a smoothing effect that approaches a straight increasing trend line, without the rapid fluctuation of historical data. However, we can still see a predicted steady increase in the impact of fires on the Albuquerque area, which leads into the health impacts studied in later parts of the project.

Extension Plan

The impact focus that I selected for my extension plan is health care. I will specifically focus on the prevalence and severity of respiratory issues in relation to smoke and air quality measurements. One of the most common types of health issue is respiratory illness, and those caused or exacerbated by air quality concerns can be largely impossible to avoid. My analysis will focus on these health issues from a perspective of mitigation and preparation.

For my additional analysis component, I selected data from the National Hospital Ambulatory Medical Care Survey data collection, which contains incredibly detailed survey information on individual hospital visits. The survey data available is contained in a public

repository, with a file for each year from 2002 to 2022. Using SPSS, this information can be formatted and loaded for further analysis.

Since the dataset is from a governmental source (National Center for Health Statistics and the Centers for Disease Control and Prevention), it is subject to certain federal data governance laws. These include the Public Health Service Act and the Confidential Information Protection and Statistical Efficiency Act, which state that the data may only be used for statistical reporting and analysis. These regulations explicitly prohibit any attempt at, or combination of external data sources that leads to, the deanonymization of data subjects. No attempts may be made to learn or extract the identity or personal information of any medical patient contained in this dataset. Researchers also cannot attempt to determine any disclosure methodology or strategies that were applied to protect individual and organizational privacy standards.

I plan to include this dataset in my analysis because it contains vital health information that is not covered by our fire and air quality datasets. In order to analyze the impact that these factors have on general population health, I need additional information regarding health metrics along a corresponding time series. The dataset is incredibly detailed, with hundreds of tracked variables from a large-scale survey project of hospital patients across the country. While many of them are not relevant to our chosen area of study, the dataset contains incredibly detailed information on respiratory measurements and overall health for patients.

The relevant factors in this dataset include timestamp on each visit, region of the country, general health indicators such as temperature and respiratory rate, as well as diagnosis codes that can be used to determine severity and type of illness or health issue. As a result, we can analyze the quantity and severity of respiratory illnesses over time across the relevant region.

I will then use this dataset to develop a model for respiratory health consequences of fire smoke and air quality. While the dataset does not cover every year for which we are analyzing air quality, the given timeframe is more than enough to develop and train a model that can be applied to a larger range of years. This model will likely consider a calculated metric for health, based on the various statistics of illness type, severity, and frequency, as well as any other components of the survey data that prove to be relevant. The model will attempt to predict this value at various times, including during future fire seasons, using our pre-existing measurements of air quality and fire impact in the area.

Intro

In this document, I will define my plan for the analysis extension of my work surrounding wildfire impact and air quality measurements for Albuquerque, New Mexico. From a scientific perspective, it is interesting because it will dive into the relationships between wildfires, air quality, and population health. From a practical perspective, it will provide insight to city leadership and healthcare organizations in order to better understand and prepare for respiratory health crises. Albuquerque, like many American cities, faces frequent threats from large scale wildfires (1). Especially in the wake of COVID-19, we are incredibly aware of the impact that respiratory health can have on a population at large; smoke inhalation and air pollution are often linked to these health issues. My analysis is relevant because it will provide further understanding of these connections and their impact on our health; the target community of health and policy workers in Albuquerque will care because it directly impacts their ability to protect the people of their communities and maximize healthcare functionality.

Related Work

The related work for my project centers primarily around the health data I obtained from the National Hospital Ambulatory Medical Care Survey. This study covered more than two decades, with records of individual hospital visits tracking up to a thousand fields of data per patient visit. The patients are randomly selected from hospitals across the country in a four-week reporting period each year, with proper anonymization procedures implemented given the sensitive medical nature of this information. While the anonymization restrictions mean that data is defined at a regional level instead of city specificity, our fire calculations are already considering the larger region around Albuquerque, so this does not raise a large issue.

Due to the unstructured, qualitative nature of the hospital data and my combination of two disparate fields of data collection, I chose to create my own models for analysis. Any existing models I considered may have been useful when applied to fire predictions or to health data on their own, but original models felt more appropriate for expanding into a new problem space. The hospital data is all text based, having been transformed and decoded for readability; this means I had to manually select and sort data, the unpredictability of which stands in the way of applying a prebuilt model.

Methodology

For the fire impact estimate, I started by acquiring the dataset of fires through API calls. Once collected, the fires were filtered for proximity to Albuquerque. For each fire season, I calculated an aggregate impact estimate by scaling each fire using its size and the inverse of its distance from the city, then adding it to the season's total. With these values, we were able to estimate the perceived severity of fire season in Albuquerque, with most of the impact coming from large fires in the closely surrounding area. The smaller and farther fires, while less impactful, still have the potential to affect the city due to wind patterns, so the aggregate value reflects this.

For air quality, I used a similar API call structure to obtain observation data from stations across the country. I filtered to the relevant stations in my area of interest and narrowed values down to only the particulate and gaseous pollutants in which we are interested. Using these daily average values, I aggregated a seasonal air quality total to match the scale and timeframe of the fire season impact estimates from above. With some small linear scaling adjustments to the already abstracted units, I was able to align my air quality and fire impact estimates on a similar scale. From here, I overlaid plots of the two and compared trends over time, with a moderately visible relationship appearing in the timeseries plots.

When it came to predicting fire impact in future seasons, I decided to create my own algorithm as opposed to using existing time series forecast models. I wanted the freedom to finely tune the impact of past years and determine what window of years the model would consider and to what degree it would scale each past year. By using decreasing weights for each year going back from the most recent, I created a prediction model that takes into account both recent history and the greater historical trend. As a result, we see an increasing fire estimate in our future predictions, although the system of averaging past years did eventually converge and smooth the predictions farther into the future.

The incorporation of my external respiratory health data proved to be the most labor-intensive part of my methodology. The hospital survey published one dataset per year of the study, with each contained in a very large file containing tens of thousands of observations, each containing hundreds of text-based fields. This required me to create one large dictionary with each year mapping to its corresponding dataset that had been extracted from zip files. Because of the previously described nature of the data being largely unformatted and converted for readability over analyzability, I had to create a manual codebook of keywords to search through symptoms and diagnoses for every hospital visit. Because the ordered codes had already been replaced with verbal descriptions in these categories before the datasets were published, I had to search for

words like “respiratory” and “asthma” among many others to identify respiratory related patient visits.

Once the health data was properly loaded, cleaned, and tagged, I had to create a metric for tracking overall prevalence of respiratory issues. Each patient had multiple columns containing diagnoses and reported symptoms, so I used these with weighted scoring based on primary and secondary status; I then aggregated the respiratory results for each year as a whole. This annual composite statistic became my metric for overall respiratory health issue prevalence, which I would go on to use in my analysis of the impact of fires on this area of health.

In the modeling process, I was fairly limited in my options by the small size of data points I had to work with. Because I could not work at any granularity beyond the annual level, which I will discuss in the limitations section, I did not have enough values for each metric to use complex timeseries modeling algorithms. As a result, I chose to work with polynomial regression, since linear regression failed to capture the variability of the data. The first model I tested was a direct use of my fire estimate to predict respiratory health scores. While this method, shown in the plots below, managed to show some of the bigger trends in this relationship, it did not do a good job of reflecting smaller changes and fluctuations from year to year.

I was originally concerned that other causes of health issues may be prevalent enough to obfuscate any relationship between fires and respiratory health, so I refined my model using past respiratory scores as well as fire estimates. By standardizing the fire estimate in relation to the average across the entire time period, I was able to create a scaling factor that closely follows the change in respiratory health each year. By starting from the most recent year’s health score and adjusting based on the current fires or lack thereof, we see a much more accurate prediction of the coming year’s respiratory health. However, the recursive nature of this model’s reliance on last year’s data means it has compounding uncertainty when predicting multiple years in the future.

Findings

In part one of the project, we found that Albuquerque is likely to see continued increases in the impact of fire smoke, as well as increases in air quality concerns. The model may have a non-negligible amount of uncertainty in predictions, especially farther in the future, but it is still fairly confident in the overall upward trend. This conclusion lays the

groundwork for our further health analysis; the expected rise in fire impact makes it all the more important to understand how it relates to health and what can be done to prepare.

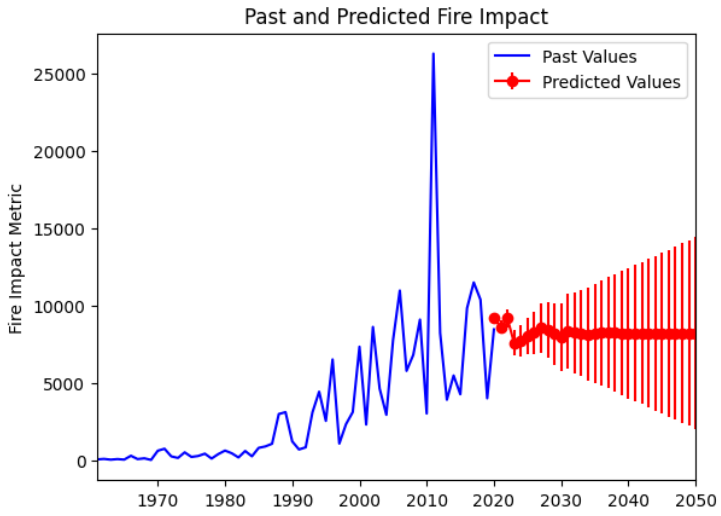

Fig 1.1: Trends in observed and predicted fire impact estimates

Although our prediction model does not reflect the huge variability in annual fire totals, this trend is inherently tied to the climate change that is also driving the overall rise in fires (3). Prediction models inherently use averages to smooth and expose trend lines, but we can expect that this variability will also continue in the future.

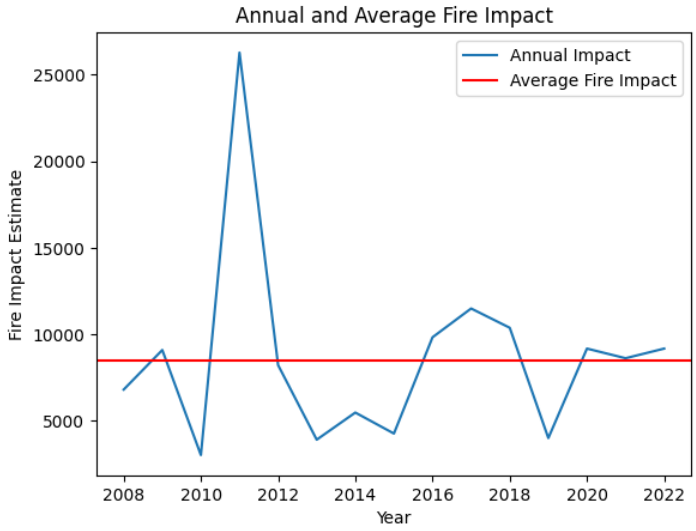

Fig 1.2: Highly variable yearly fire impact estimates compared to average across the timeframe

The findings of the second half of this project clearly indicate the potential for fire strength metrics to be used as a factor in predicting respiratory health crises. The combined model that used both fire season statistics as well as recent historical respiratory data performed very well when predicting our calculated respiratory health score. While the fire impact estimate on its own was not a hugely significant predictor of respiratory health, I found that it was very useful as a factor in determining health trends; that is, a strong or weak fire season was largely indicative of whether respiratory health issues would worsen or decrease in the coming year.

As shown below, the model based on fire estimate only managed to capture big picture trends in respiratory health but was not nearly as accurate as the model using fire estimates alongside historical health data.

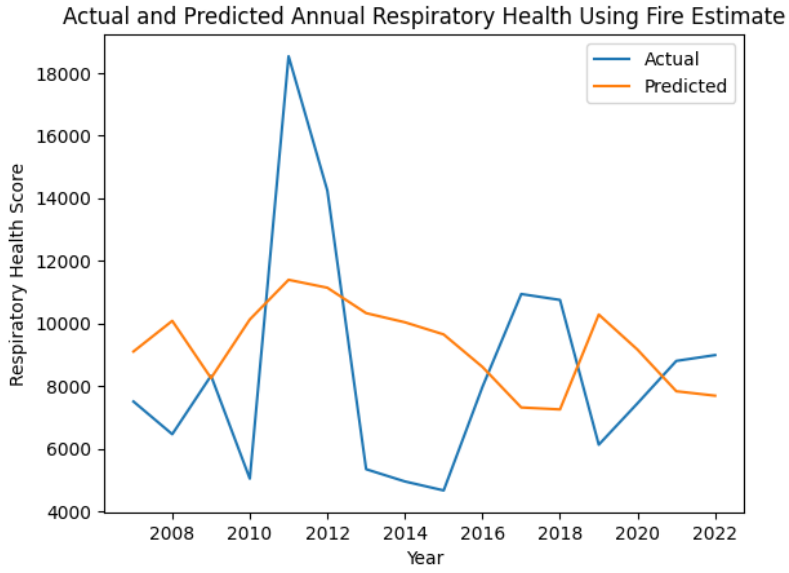

Fig 1.3: Respiratory health predictions based solely on fire estimates

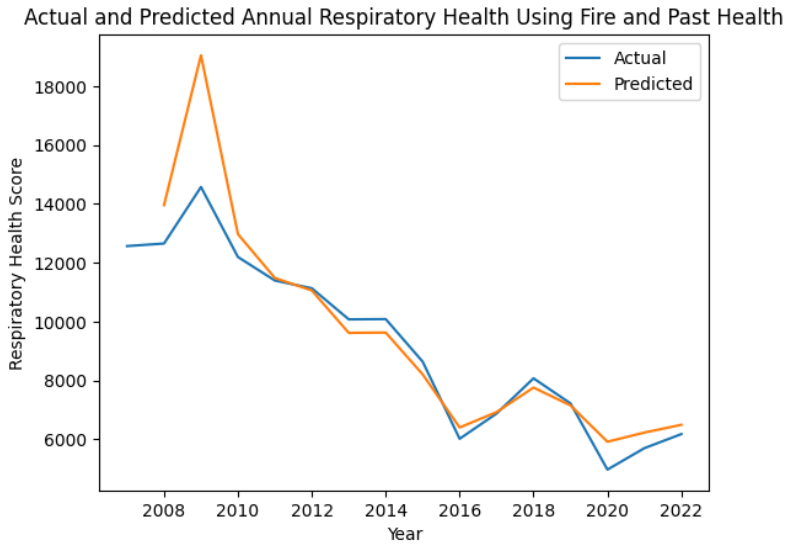

Fig 1.4: Respiratory health predictions based on past respiratory health along with current fire estimates

Given the uncertainty in our long-term fire prediction model, along with the recursive nature of our respiratory health model, my analysis is likely not hugely meaningful for predicting outcomes far into the future. However, it is extremely useful for predicting the short-term future, and its accuracy will continue to improve as we test it against incoming real-world data. The biggest piece of information from the respiratory health analysis is that fire estimates can predict the change in respiratory health outcomes; knowing this, we can get ahead of health crises as we see fire seasons spike going forward.

Implications

As indicated above, this analysis has potentially vital ramifications on health system preparedness in relation to respiratory issues. While the long-term projections are not fully reliable, the short-term models show a very clear indication that past health statistics, in conjunction with fire impact analysis, can provide insight into current and coming respiratory health conditions. For the city of Albuquerque, this is incredibly useful in allocating resources and preparing to handle any health issues that may arise. With this type of forecasting, city leadership and medical infrastructure can get a jump start on preparing for incoming issues, potentially boosting the healthcare system and improving the health of the community at large. As we saw during the onset of the COVID-19 pandemic, one of the most damaging effects of health crises can be the initial surge that

overwhelms our established care systems; analysis that warns of incoming issues, even in the current short-term format, can mitigate or avoid this overload on our vital infrastructure. Especially in a smaller city like Albuquerque that lacks the resources of larger cities, preparedness is key to surviving and thriving throughout changing environmental health conditions.

Limitations

The primary limitations of my analysis lie in the realm of data quality. The hospital survey data is incredibly comprehensive but largely unformatted and inconsistent; each year they used a different subset of over a thousand potential variables. On top of this, the fire data was inconsistent and spotty when it comes to exact time data. When compounded with the health data's source being a four-week window each year, it was impossible to create a more granular analysis than the annual level. As a result, there are limited data points available, and the sample size of my analysis is fairly constrained. With cleaner and more complete data, monthly or even weekly analysis could be used for much more accurate and fine-tuned results. Another consequence of this limitation in total data points for health and fire estimates is that I did not have enough points to separate out a designated test set. Because of the small sample size, the modeling process would be impacted significantly by the exclusion of any data points, so I could not test on a truly blind set of values to analyze the performance of my models. However, as future data comes in and the models are applied, the accuracy can be tracked, and the functionality can be adjusted to meet the concerns raised by live prediction accuracy.

Conclusion

I came into this project with the intent to find a connection between my calculated fire impact estimate and respiratory health concerns in the city of Albuquerque. I believe my analysis has demonstrated that, in combination with recent respiratory health history, this measurement of wildfire impact can accurately predict short-term health concerns, with relevant but decreasing usability as the model extends farther into the future. Going forward, this type of analysis can be vital in helping city leadership prepare for and manage the respiratory issues that accompany the climate change induced increase in fire severity and variability.

From a human-centered data science perspective, this analysis establishes a direct connection between theoretical modeling and real-world benefits. The health of an entire

community can potentially be improved by the insights gained from a statistical analysis of fire and health trends. When analyzing human health concerns such as respiratory illness, the project is inherently human centered; it revolves around the intent to improve infrastructure and management of individual and societal well-being.

References

(1): Las Conchas Fire: https://www.nps.gov/band/learn/nature/lasconchas.htm

(2) NHAMCS: https://www.cdc.gov/nchs/nhamcs/about/index.html

(3) Climate impact: https://interactive.carbonbrief.org/attribution-studies/index.html

Data sources

National Hospital Ambulatory Medical Care Survey:
https://ftp.cdc.gov/pub/Health_Statistics/NCHS/dataset_documentation/nhamcs/spss/

US Geological Survey
https://www.sciencebase.gov/catalog/item/61aa537dd34eb622f699df81

US EPA Air Quality System

https://www.epa.gov/aqs