Introduction

I am conducting an analysis between respiratory disease mortality rate and smoke estimate as well as its impact on public health. This analysis is really important because it connects climate change, air quality, and public health in a way that directly impacts people's lives. Additionally, healthcare is something that affects everyone's life, and the implications of health care practices extend even further to people's finances and socioeconomic status. Also, this analysis brings awareness to the growing climate change crisis as rising temperatures are no longer just an environmental issue, but a growing health concern. As temperatures rise and wildfires become more common, the problem is only going to grow, making it even more critical to study the link between smoke and respiratory disease. The goal is to raise awareness and inform policies that can help protect vulnerable populations and ease the burden on healthcare systems. Ultimately, this analysis helps us see that climate change isn't just an environmental issue, it's also a serious public health crisis that we can't afford to ignore.

Background

From Part 1: Common Analysis, I developed a smoke estimate using Wildfire data collected and aggregated by the US Geological Survey. I used this data to construct a smoke estimate (detailed more in the methodology section) for the past 60 years of data (1962 - 2021) to understand the impact of smoke exposure for land within a 650 mile radius of Dayton, Ohio. I used a time series model to forecast the next 30 years of smoke estimate data to gain a better understanding of the impact smoke will have on the future of this region. I used AQI data to successfully validate my smoke estimate. Now I want to expand my understanding of smoke impact to the domain of public health, particularly the effect of smoke impact on respiratory disease mortality rates as wildfires continue to be a growing issue in the United States.

There has been a lot of research on this topic by the United States Environmental Protection Agency (EPA). Their studies have looked at how different sources of particulate pollution, like car emissions, industrial pollution, and wildfires, affect respiratory health. Research has shown that long-term exposure to fine particles increases the risk of conditions like asthma, COPD, heart disease, and even lung cancer. Big studies tracking large groups of people

have found strong links between particle pollution and worse health outcomes, especially for people already dealing with respiratory issues.This kind of research really backs up my hypothesis that as smoke levels rise (due to wildfires or pollution), we're likely to see higher rates of respiratory disease and even more deaths. The EPA's discovery about the connection between fine particle exposure and lung cancer strengthens my idea that increased smoke exposure can lead to more severe health problems. Their research uncovers a positive association between "between fine particle exposure and lung cancer mortality." This discovery directly backs up my hypothesis that an increasing smoke estimate results in increasing rates of respiratory disease mortality rates.

Hypothesis and Research Questions

My hypothesis is that higher smoke levels lead to more deaths from respiratory diseases, which will then put more pressure on the healthcare system. Some of the big research questions I'm trying to answer are: How does smoke exposure impact respiratory disease mortality rates and to what degree? What are some steps that need to be taken (by the county or otherwise) to address these issues?

Related Work

In part 2, I had originally planned to use a healthcare model, APACHE II, that inputs specific patient labs and diagnostics to determine the likelihood of a patient being placed in the ICU. I was going to apply this data on patients afflicted by respiratory diseases and determine if there was a correlation between their likelihood of ICU admission and my smoke estimate. I would average the patient ICU likelihood data to a yearly level so that I could directly compare it with my yearly smoke estimate. This would allow me to determine if the severity of respiratory disease was associated with smoke estimates. Unfortunately, I was unable to find granular patient data for my city (Dayton, Ohio) that was required as input to the APACHE II model. In hindsight this makes sense as most patient data is protected.

Due to the fact I didn't have the input data to the APACHE II model, I decided to replace that model with an ARIMAX model to directly look at the link between smoke estimate and respiratory disease mortality rate. The ARIMAX model builds on top of the traditional ARIMA

model by incorporating an exogenous variable, the smoke estimate. This model helps directly analyze the relationship between smoke estimate and respiratory disease mortality by incorporating both past trends of respiratory disease rates and external factors (like smoke levels) that might influence these rates. By using smoke estimates as an exogenous variable, the ARIMAX model allows us to see how changes in smoke exposure impact mortality rates (the outcome variable) while accounting for other factors that might also play a role. This approach provides a clearer, more accurate picture of how smoke affects respiratory health outcomes over time. More importantly, obtaining the data was not an issue as I already had my smoke estimates from part 1 and respiratory data for Montgomery County was readily available on the CDC WONDER website.

The CDC WONDER (Wide-ranging Online Data for Epidemiologic Research) page is an online database that allows users to access a variety of public health data, including mortality rates, disease statistics, and demographic information. To extract respiratory disease mortality rate data for Montgomery County, I used the system's search features to select the specific cause of death (respiratory diseases), the relevant time period (1962 - 2021), and the geographic location (Montgomery County). After specifying these parameters, I was able to download the data, which provided mortality rates for respiratory diseases in the county over the chosen period. The data on the CDC WONDER platform is primarily collected through the National Vital Statistics System (NVSS), which compiles death certificates from all 50 states and the District of Columbia. These certificates include information about the cause of death, demographic details, and geographic location. The data is processed and standardized by the CDC and mortality data, in particular, is updated regularly, typically on an annual basis, to reflect the most current public health trends. This allows researchers and policymakers to analyze trends in mortality rates, including mortality rate from respiratory diseases, at various geographic levels and across different time periods. Thus, I gathered yearly county data on deaths by respiratory disease per 100,000 people in Montgomery County.

Methodology

I calculated my smoke estimate using the following equation: GIS_Acres / (distance)^2. I calculated the smoke estimate for each fire within a 650 mile radius of Dayton, Ohio and then

added the estimates for a given year to get each year's final smoke estimate. Using the formula GIS_Acres / (distance)^2 to calculate smoke estimates is beneficial because it accounts for how smoke exposure decreases with distance from the source, reflecting the natural spread of pollutants. The GIS_Acres term represents the area affected by smoke, while the distance^2 factor adjusts for the fact that exposure intensity typically drops as you move farther away. This allows for a more spatially aware representation of smoke impact. I used an ARIMA model for my smoke estimate forecast in part 1 due to its interpretability, and continued to build on it by using an ARIMAX model for part 4, which incorporated health data along with the smoke estimate.

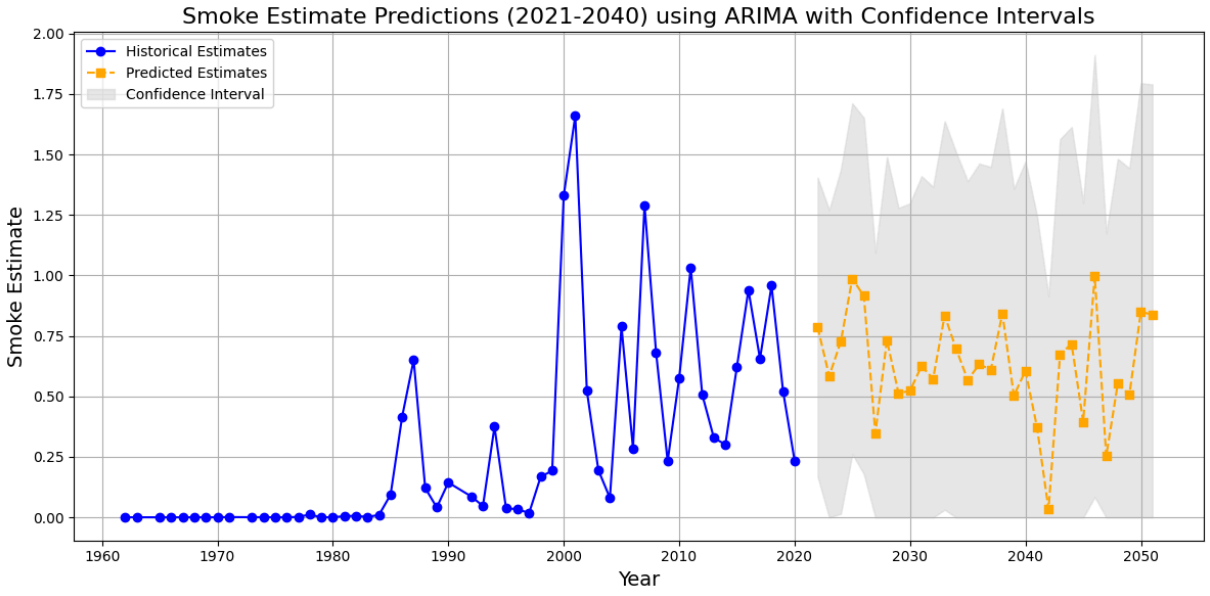
In order to validate my smoke estimate, I used yearly AQI data from the United States EPA (Environmental Protection Agency) from monitoring stations near Dayton Ohio. Most of the stations did not have data on most gaseous or pollutants, so I used the particulate pollutants which had more data available as my AQI estimate. I averaged the particulate pollutant values per year to get my yearly AQI estimate. As we can see in the next section, my AQI estimate tracks my smoke estimate pretty well. Since the values are on differing scales of magnitude, I decided to scale both estimates before plotting as I am more interested in the trend rather than the actual values.

The ARIMAX model takes in past respiratory disease rates and the smoke estimate values as the predictors to make accurate predictions on future respiratory disease rates (response). I chose to use an ARIMAX model to forecast the data as it is more human centered and interpretable. The model focuses on identifying patterns in time series data that are easy for most people to find and understand, like seasonality. It also places emphasis on past data points (auto regression) and trends over time (integration), which are both concepts that most people who have seen a time series plot are familiar with, even if they don't know the exact terminology for it. Therefore, using a model that encapsulates the idea of using past trends to predict future outcomes seems relatively intuitive, especially in comparison to black box models. Overall, the relatively high interpretability of the ARIMAX model was the main reason I chose to incorporate it into my analysis.

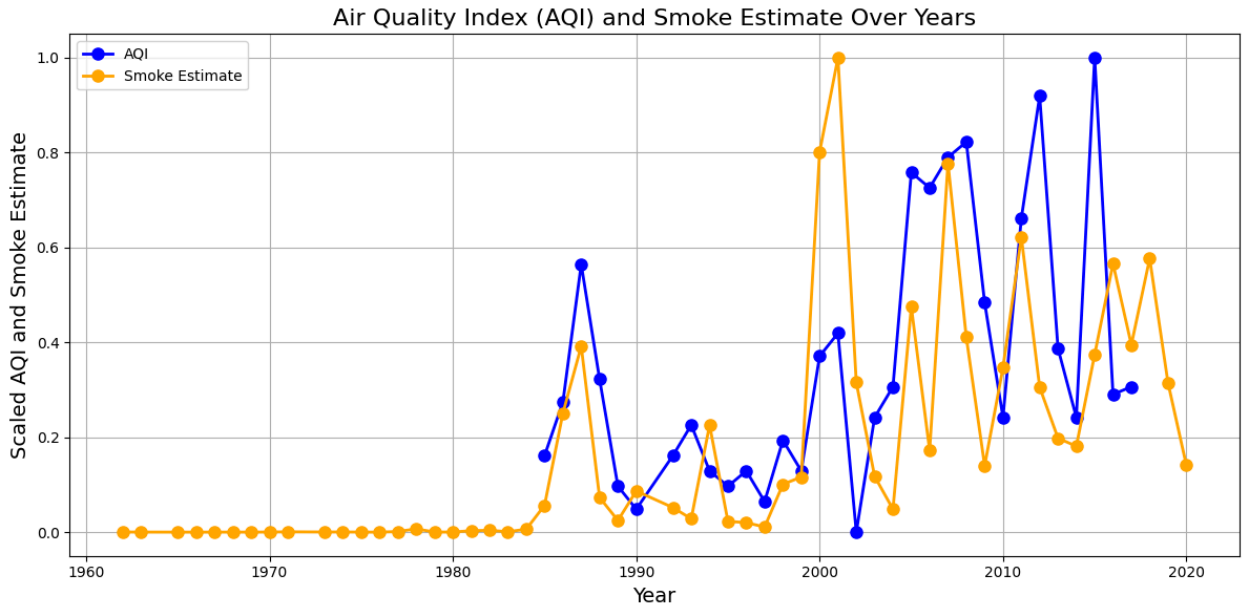
The methodology behind the ARIMAX model is an extension of the ARIMA, except that it adds on exogenous variables, additional factors that might influence the data being analyzed. Like ARIMA, ARIMAX looks at past data, removes trends, and smoothes fluctuations. The key difference is that ARIMAX also considers external variables (such as smoke levels or weather patterns) that might impact the outcome being predicted, like respiratory disease mortality. This allows the model to give more accurate forecasts by accounting for factors beyond just past trends in the data. I also chose to add noise to my model forecasts just to reflect the random noise that is apparent in real world events, especially events related to weather, such as my smoke estimate. I also used the Augmented Dickey Fuller test to verify that my data meets the assumption of stationarity before I could appropriately use it on my ARIMA model. Also, I checked the ACF (autocorrelation function) and PACF (partial autocorrelation function) of my smoke estimate and health data to check for stationarity and lag effects.

I chose to conduct a correlation analysis between respiratory disease rates and smoke estimates because it offers a straightforward way to identify if there is a significant relationship between increased smoke estimates and rising respiratory disease rates. This method allows me to analyze patterns and draw meaningful conclusions without having to rely on complex assumptions. Ethical considerations were central in my decision to use this approach, as it ensures a data-driven, unbiased examination of the issue. By focusing on correlation, I avoid jumping to conclusions (such as causation) or misinterpreting data, especially when dealing with sensitive health information.

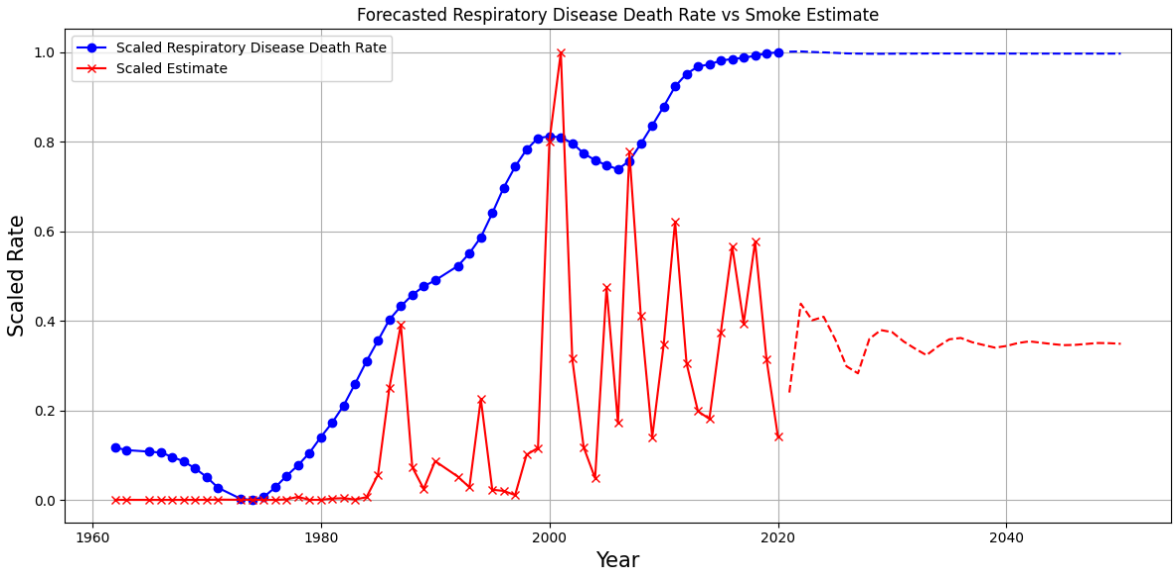
Findings



The time series chart above details the historical smoke estimate and its forecasted values for the next 30 years using an ARIMA model. I added some noise to my forecast values to reflect the random noise that is apparent in real world events, especially events related to weather, such as my smoke estimate. We can see a sharp uptick in smoke estimates after 1985 and a major increase around 2001. The forecast remains in a pretty stable window between 0.45 to 1 from 2022 to 2040 where it takes a sharp nosedive. Regarding forecasted smoke estimate data, we have to keep in mind predicting weather data so far in the future is not reliable because the weather is very unpredictable and influenced by many factors that are hard to forecast accurately. While short term forecasts (within the next week) are pretty accurate, the further ahead you try to predict, the more difficult it gets. Unexpected events or changes in the environment can throw off predictions. So, long-term weather forecasts are often just guesses and can be wrong.

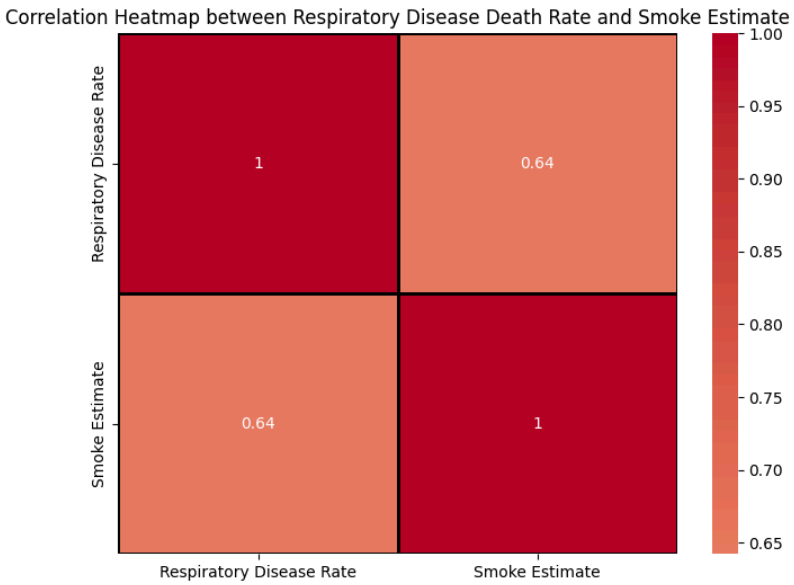


Another interesting finding from part 1 was my yearly AQI data and smoke estimate data tracked each other pretty well (as shown above) from 1986 to 2019, indicating there is some correlation between the two. As the smoke estimate increases, so does the yearly AQI. This makes sense because the AQI measures how polluted the air is, and smoke is a major source of air pollution. As smoke levels rise, from wildfires in our case, it adds more particles and pollutants to the air, which directly worsens air quality. So, when the smoke estimate goes up, it means there's more smoke in the air, leading to a higher AQI.



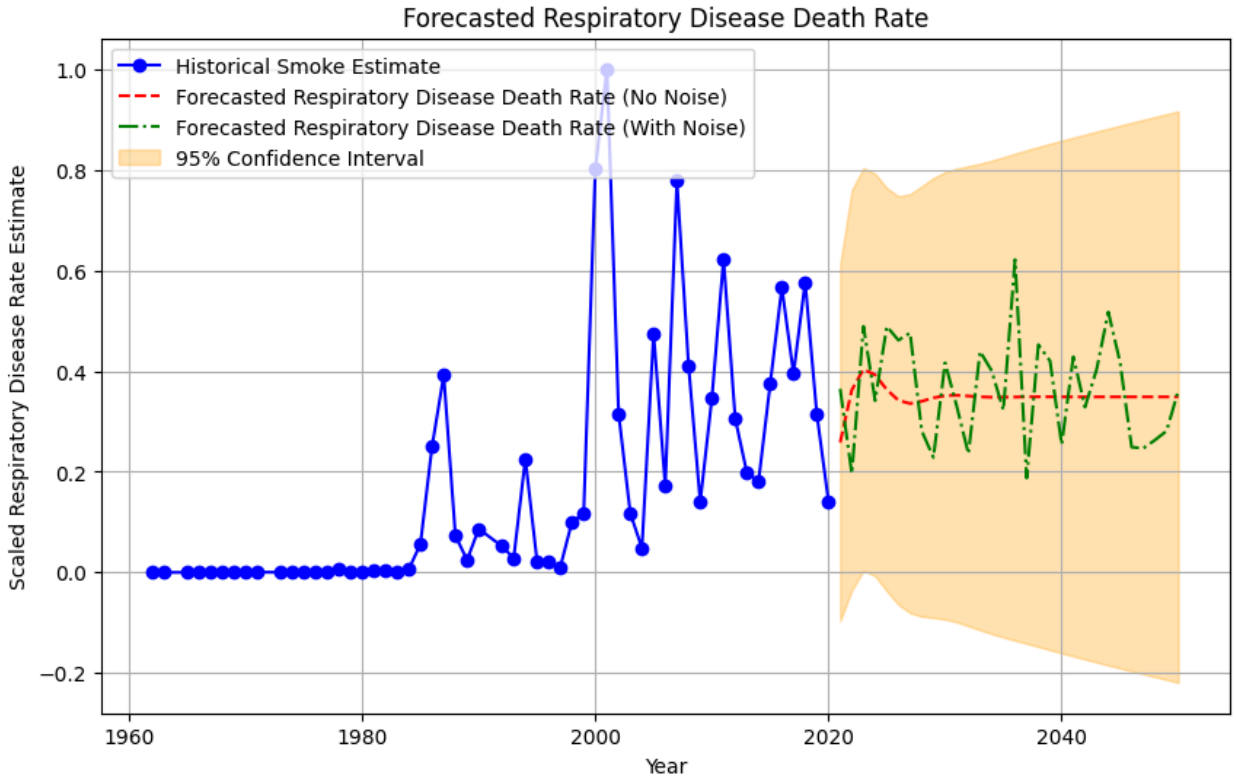
Next, I compared the time series between the smoke estimate and the respiratory disease mortality rate to see how the trends and forecasts compare. Based on the chart shown above, we can see the two lines do not track each other as well as the AQI and smoke estimate did in the chart shown above. However, there is a steady increase in both the smoke estimate and respiratory disease mortality rate after 1985. Also, we see an increase in the respiratory disease mortality rate around the year 2000, which is also when the smoke estimate peaks. Similarly, there is a drop in respiratory disease mortality around 2005 and we see the smoke estimate drop around a year before that indicating there may be some sort of lag effect. The 30 year forecast for the respiratory disease mortality rate remains constant at its peak, indicating the prevalence and urgency of the issue. The smoke estimate (without any added noise) also reaches a constant rate around 2035. Overall, the forecast indicates there may be some association between both trends. More importantly, the forecast suggests respiratory disease mortality rate remains a constant, pressing issue throughout the next 30 years.

My next step was to do a correlation analysis between the smoke estimate and respiratory disease mortality rate. After calculating Pearson's correlation coefficient and constructing a correlation matrix, as shown to the right, we see a



relatively strong positive correlation of 0.64. This indicates as smoke estimates increase, respiratory disease mortality rates also increase. While the correlation coefficient suggests there is a clear link, other factors might also be influencing respiratory disease mortality rate, not just smoke.

Lastly, I applied the ARIMAX model (as shown below) to forecast respiratory disease mortality rates. I created a version of the forecast with noise added to it so as to reflect the randomness in real world data. For the forecast with noise, we see some fluctuation but the no noise forecast portrays a bump around 2022. Overall, the fact that we do not see the trend decreasing, is enough to indicate that deaths by respiratory illnesses are a prevalent issue and should be closely monitored and addressed. Also, with growing carbon emissions, smoke exposure rates do not seem to be improving, indicating respiratory diseases will still remain a major health issue.



Discussions and Implications

My findings are important because they show that respiratory disease rates are not decreasing, even as we might hope they would with improvements in healthcare and air quality. This suggests that factors like smoke from wildfires or pollution are continuing to harm public health, and these issues aren't going away on their own. The city council, mayor, and residents need to take action now to reduce smoke exposure, whether by improving air quality regulations, increasing awareness about health risks, or investing in better healthcare for affected communities. They don't have much time, maybe a year maximum, because smoke-related health problems are growing, and immediate steps are needed to protect vulnerable groups like children, the elderly, and those with pre-existing conditions.

In this project, I focused on using interpretable models to make my analysis more human-centered and accessible, especially for non-technical audiences. I chose methods like correlation analysis and the ARIMAX model because they provide clear, understandable results that can be easily explained. For example, the correlation coefficient of 0.64 directly shows the relationship between smoke levels and respiratory disease rates, which can be easily communicated to city officials and residents without getting into complex technical details. By using these models, I aimed to make the data more actionable for policy makers, to convey a sense of urgency and take steps to protect public health. This approach ensures that my findings aren't just useful to experts, but also to the broader community, helping everyone understand the real world impact of air quality on their health.

Human-centered data science played a major role in how I approached the problem. Instead of simply looking at numbers or trends, I focused on how the findings could inform decisions that have a direct impact on people's health. This means considering the broader context: how environmental issues like rising smoke from wildfires are worsening air quality and affecting people. By framing the problem this way, I ensured that my findings were not just technical but actionable and easily communicable (through interpretable models) and relevant to the needs of residents, healthcare providers, and policymakers. Human-centered design also helped me think across different domains: environmental, health, and social, and consider how they all intersect in people's daily lives. This approach influenced my decision to make sure the

analysis was easily understandable, so city leaders and residents could act on the findings. For instance, knowing that worsening air quality is linked to increasing health risks, it was clear that actions to combat climate change, such as reducing carbon emissions and improving air quality, should be prioritized.

Limitations

One of the main limitations of this analysis is the quality and completeness of the data. The data on respiratory disease mortality rates and smoke estimates might have missing values or inconsistencies, which could affect the results. Some regions or years may have incomplete data, and if not handled properly, this can lead to inaccurate conclusions. While I worked with the data available, any gaps could limit the ability to get a fully accurate picture, especially if certain socioeconomic groups of people were not accounted for when gathering this data.

Another limitation is how the smoke estimate was calculated. I used a simple model that takes into account the area affected by smoke and the distance from the source. While this helps estimate the spread of smoke, it doesn't capture factors like wind patterns or weather changes that affect smoke movement. So, the smoke estimate might not fully reflect actual exposure levels, especially in places with varying environmental conditions.

The correlation analysis assumes that the relationship between smoke and respiratory disease rates is linear, meaning it assumes that as smoke increases, disease rates increase at a steady rate. However, the relationship might be more complicated. For example, there could be non-linear effects, or other factors like healthcare access or socioeconomic status could play a bigger role. The correlation coefficient of 0.64 shows a moderate relationship, but it's possible that other factors are influencing the results too.

The ARIMAX model also has certain assumptions. It assumes that past data can predict future trends, and that external factors like smoke estimates are properly accounted for. However, the data needs to be stable (stationary) for this model to work well, which I validated. Still, it's possible that this didn't fully remove all the underlying trends, which could affect the accuracy of the predictions. Also, many other factors, like air pollution from other sources, healthcare access, or lifestyle choices, can also affect respiratory health. These factors weren't included in the

models, which means the analysis might not show the complete picture of what's influencing respiratory disease rates.

Finally, while the data used in this study is publicly available, there are still ethical considerations around the use of health and geographic data. It's important to use this data responsibly and make sure it's not misinterpreted in ways that could harm people. In particular, ensuring that the analysis is fair and doesn't overlook vulnerable populations is crucial.

Conclusion

The main research questions in this study were: How does smoke exposure impact respiratory disease mortality rates and to what degree? What are some steps that need to be taken (by the county or otherwise) to address these issues? My hypothesis was that higher smoke levels would lead to more deaths from respiratory diseases and put more pressure on healthcare systems.The results showed a moderate positive correlation (0.64) between smoke estimates and respiratory disease mortality rates. This means that as smoke levels increase, the death rates from respiratory diseases also tend to rise, supporting my hypothesis. While the relationship isn't perfect, it's clear that smoke is a significant factor in respiratory health. The ARIMAX model also showed that past smoke exposure can help predict future mortality trends which are not decreasing, further supporting the link between smoke and respiratory disease outcomes and urgency to instill change.

This study highlights key ideas in human-centered data science by focusing on the real-world impact of smoke on people's health. I used clear and understandable models like correlation analysis and ARIMAX, making the findings easy to communicate to non-technical audiences, such as city officials or residents. Human-centered design guided me to consider the interconnectedness of environmental, health, and social factors, and how they all impact people's everyday lives. This approach led me to focus on making the analysis clear and accessible, so that city leaders and residents could use the results to take meaningful action. For example, understanding the link between poor air quality and rising health risks highlighted the importance of taking immediate steps to address climate change, such as reducing carbon emissions and improving air quality.

References

Centers for Disease Control and Prevention. *CDC WONDER*. U.S. Department of Health and Human Services, 2024, https://wonder.cdc.gov/cmf-icd8.html.

National Oceanic and Atmospheric Administration. *Climate Change: Atmospheric Carbon Dioxide*. U.S. Department of Commerce, 2024, https://www.climate.gov/news-features/understanding-climate/climate-change-atmospheric-carbon-dioxide#:~:text=Carbon%20dioxide%20concentrations%20are%20rising,in%20just%20a%20few%20hundred.

U.S. Environmental Protection Agency. *How AQI is Calculated*. U.S. Environmental Protection Agency, 2024, https://www.epa.gov/outdoor-air-quality-data/how-aqi-calculated.

U.S. Environmental Protection Agency. *Particle Pollution and Respiratory Effects*. U.S. Environmental Protection Agency, 2024, https://www.epa.gov/pmcourse/particle-pollution-and-respiratory-effects.

Data Sources

The Wildfire data was collected and aggregated by the US Geological Survey. The source data is from https://www.sciencebase.gov/catalog/item/61aa537dd34eb622f699df81

CDC Respiratory Disease Mortality Rate Data : https://wonder.cdc.gov/cmf-icd8.html