DATA 512 Project Part 2 - An Extension Plan

Name: Sagnik Ghosal

Date: November 16th 2023

Motivation/Problem Statement

The analysis of smoke fires and their impact on respiratory health in Tulare County, California, is crucial due to the escalating concern over the growing incidence of chronic respiratory diseases, increased hospitalizations, and mortality within the population. The escalating occurrences of wildfires in recent years have raised concerns about the impact of prolonged exposure to smoke on public health, especially concerning respiratory diseases like Asthma [**Ref 1**]. It's crucial to gauge the specific health implications, especially regarding hospitalizations and deaths due to respiratory diseases attributed to fire smoke. This analysis aims to offer a comprehensive understanding of the correlation between increased fire smoke activity and its adverse effects on respiratory health, providing essential insights that directly impact the community's well-being.

Tulare County residents face a unique set of challenges, given the proximity to areas prone to wildfires. Understanding the intricate relationship between smoke exposure and chronic respiratory diseases is not only a matter of public health but also holds practical implications for city planning, emergency response strategies, and healthcare resource allocation. The analysis aims to be a pivotal tool for informed decision-making by the policy makers. By shedding light on the potential future impacts of smoke on the community, the study seeks to pave the way for proactive measures to mitigate the adverse effects and safeguard the well-being of Tulare residents.

The insights derived from this analysis have direct implications for the community's well-being. Increased hospitalizations due to respiratory issues not only strain the healthcare system but also impose a significant burden on individuals and families. Moreover, the associated rise in mortality rates underscores the urgency of implementing measures to address the root causes of smoke exposure. By informing the city council and local authorities, this study aims to catalyse proactive initiatives, such as enhanced air quality monitoring, improved emergency response protocols, and public health campaigns to raise awareness about the risks and preventive measures.

From a human-centered data science perspective, this analysis holds the promise of revealing critical insights that directly impact the lives of Tulare County residents. It seeks to uncover the intricate connection between smoke exposure and chronic respiratory diseases, highlighting the tangible challenges faced by individuals due to increased hospitalizations and mortality rates. Learning how smoke fires influence respiratory health outcomes isn't merely about numbers and statistics; it's about understanding the tangible impact on individuals and families within the community. Ultimately, the analysis seeks to not just uncover correlations but to translate these findings into policies that safeguard the community against the detrimental effects of smoke fires, fostering resilience, and preparedness in the face of environmental challenges.

Impact Focus

The spotlight here is on healthcare—whether hospitalizations and deaths due to respiratory diseases are rising because of fire smoke. It's crucial to understand these impacts and inform the policy makers and residents about what might happen in the future due to smoke in their community. Having a machine learning model that predicts how hospitalizations or deaths might trend in the coming years can help plan steps to lessen the effects. While the datasets together can answer many questions about how fire smoke affects Tulare County, I'm narrowing my analysis to focus on hospitalizations and deaths linked to respiratory diseases caused by smoke or air quality. Here are a few research questions I'm aiming to answer with this analysis.

NOTE: The hypothesis below are examples, and the numbers and names shouldn’t be taken at face value.

Research Question 1:

How does fire smoke relate to the age-standardized mortality rate for respiratory diseases in Tulare County across different years?

Hypothesis 1:

- There's likely a positive link between more exposure to fire smoke and higher age-standardized mortality rates for respiratory diseases in Tulare County.
- Overall, mortality rates might increase during years with intense fire activity.

Research Question 2:

Is there a noticeable difference in respiratory disease mortality rates between sexes due to variations in exposure to fire smoke in Tulare County?

Hypothesis 2:

- Differences in fire smoke exposure might lead to a significant contrast in age-standardized mortality rates for respiratory diseases between males and females in Tulare County.

Research Question 3:

How does the prevalence of smoke fires in Tulare County connect with rates of asthma-related hospitalizations across different age groups over time?

Hypothesis 3:

- The occurrence of smoke fires in Tulare County might be connected to an increase in asthma-related hospitalizations across all age groups.
- Younger people (0-17 years) might experience, on average, a higher percentage of asthma-related hospitalizations compared to other age groups.
- Over the next 15 years, there could be a surge in asthma-related hospitalizations tied to periods of increased smoke fires in Tulare County.

Research Question 4:

To what extent do certain respiratory diseases linked to smoke-related pollution (e.g., Chronic respiratory diseases, Chronic obstructive pulmonary disease, Asthma) exhibit higher mortality rates compared to diseases not directly related to smoke-related pollution (e.g., Pneumoconiosis, Coal workers pneumoconiosis) in Tulare County from 1980 to 2014?

Hypothesis 4:

- Diseases primarily linked to smoke-related pollution might show, on average X% higher age-standardized mortality rate compared to diseases attributed to non-smoke related pollution in Tulare County.

Data To Be Used

Besides the USGS Wildland Fire Combined Dataset from Combined wildland fire datasets for the United States and certain territories, 1800s-Present (combined wildland fire polygons) we used in Part 1, I'll be including hospitalization and mortality rate data from the Institute for Health Metrics and Evaluation (IHME) and the California Health and Human Services Open Data Portal (CalHHS) for our analysis. This extra information will help us see how more fires close to Tulare affect the air quality and, consequently, impact the residents. We'll look at how this situation leads to more hospital visits for respiratory problems and higher rates of deaths caused by these diseases.

Data Source 1

- **Name of dataset:** IHME Mortalities from 1980 to 2014 from Respiratory Diseases in California Counties.

- **Link to dataset:** https://ghdx.healthdata.org/record/ihme-data/united-states-chronic-respiratory-disease-mortality-rates-county-1980-2014
- **Description of dataset:** This dataset comprehensively records data on causes, years, sexes, posterior mean estimates, 2.5th percentile estimates, and 97.5th percentile estimates. It delineates the age-standardized mortality rate (deaths per 100,000 population) across different sexes and for both sexes combined, spanning the years 1980-2014 for all counties within California. Focused on Tulare County, the dataset underwent filtering using FIPS=6107, resulting in 1050 entries across 16 columns. The dataset includes both string-based descriptors (measure_name, location_name, cause_name, sex, age_name, metric) and integer-based identifiers (measure_id, location_id, FIPS, cause_id, sex_id, age_id, year_id, mx, lower, upper). It's worth noting that the dataset stands clean, devoid of any null values across all columns. The table below provides detailed descriptions of the 16 columns.

COLUMN NAME	DESCRIPTION
measure_id	Unique numeric identifier for the measure generated and stored in an IHME database of data dimensions
measure_name	The measure (indicator) of the estimate
location_id	Unique numeric identifier for the location generated and stored in an IHME database of data dimensions
location_name	Location of the estimate
FIPS	The Federal Information Processing Standards (FIPS) code, a unique identifier for states and counties in the United States
cause_id	Unique numeric identifier for the cause of disease or injury generated and stored in an IHME database of data dimensions
cause_name	Cause of disease or injury of the estimate
sex_id	Unique numeric identifier for the sex generated and stored in an IHME database of data dimensions
sex	Gender for the estimate
age_id	Unique numeric identifier for the age group generated and stored in an IHME database of data dimensions
age_name	Age group estimated
year_id	Time period of estimate
metric	Metric/unit of measure for the estimate
mx	Posterior mean estimate
lower	2.5% percentile estimate
upper	97.5% percentile estimate

- **How the Dataset Contributes:** This dataset's vital columns—year_id, cause_name, sex, mx, lower, and upper—will serve as the focal points for my analysis. By juxtaposing this information with the fire activity dataset previously constructed in Part 1 of our analysis, I aim to discern any discernible patterns indicating an escalation in diseases associated with air pollution, particularly stemming from increased fire smoke. Should this correlation hold true, the subsequent step involves scrutinizing the impact on different demographic groups, specifically exploring whether there's a disproportionate effect on either males or females. Additionally, I intend to conduct an in-depth examination of mortality rates, seeking to establish a correlation with fire activity.
- **Terms of Use:** Data made available for download on IHME Websites can be used, shared, modified or built upon by non-commercial users in accordance with the IHME FREE-OF-CHARGE NON-COMMERCIAL USER AGREEMENT.

Data Source 2

- **Name of dataset:** Asthma Hospitalization Rates by County from 2015 to 2020.

- **Link to dataset:** https://data.chhs.ca.gov/dataset/asthma-hospitalization-rates-by-county
- **Description of dataset:** This comprehensive dataset encompasses counts and rates (per 10,000 residents) of asthma-related hospitalizations across all counties in California. The dataset is organized by age groups (all ages, 0-17, 18+, 0-4, 5-17, 18-64, 65+) and racial/ethnic categories (white, black, Hispanic, Asian/Pacific Islander, and American Indian/Alaskan Native). Sourced from the Department of Health Care Access and Information Patient Discharge Data, the information is further refined by filtering exclusively for Tulare County, resulting in 76 entries and 9 columns. These columns include County name, Year, Age group, Number of hospitalizations, among others.
- **How the Dataset Contributes:** This dataset serves as a pivotal resource for assessing trends in asthma occurrences, particularly from 2015 onwards, and investigating their potential correlation with smoke activity. It also offers a robust avenue for analyzing whether specific age demographics are disproportionately affected by asthma-related hospitalizations.
- **Terms of Use:** Data available for download through the CalHHS Open Data Portal is open for utilization, sharing, modification, or expansion by non-commercial users, aligning with the CHHS Terms of Use.

Ethical Considerations

For the above two datasets, differential privacy has been used, enabling high-quality results without identifying anyone. To further protect people’s privacy, it is ensured that no personal information or individual search queries are included in the dataset. The IHME dataset, encompassing mortality rates from respiratory diseases across California counties from 1980 to 2014, strictly maintains data integrity, ensuring anonymity and confidentiality. Similarly, the Asthma Hospitalization Rates dataset, spanning 2015 to 2020 and delineating asthma occurrences across various demographics, abides by California's stringent data privacy regulations.

These additional datasets—the IHME Mortalities from Respiratory Diseases and the Asthma Hospitalization Rates by County—play pivotal roles. The IHME dataset allows for a nuanced exploration of patterns correlating air pollution-related diseases, particularly those exacerbated by increased fire smoke. This dataset's comprehensive insights, when combined with the fire activity data from Part 1, aim to unravel potential connections between escalated respiratory ailments and heightened smoke activity. Moreover, the Asthma Hospitalization Rates dataset, detailing asthma occurrences from 2015 onwards and segmented by demographics, offers a vital avenue to scrutinize correlations between asthma trends and smoke activity. By examining specific age groups and ethnicities, this dataset enriches the analysis, shedding light on potential vulnerabilities to smoke-induced respiratory issues among distinct demographic segments. Both datasets, offer crucial insights into the impact of fire smoke on respiratory health within Tulare County.

Unknowns and Dependencies

The research exploring how fire smoke links to respiratory health faces a few uncertainties and things that depend on each other, which could affect the study. Even though the datasets give us a lot of details about deaths and hospital visits due to asthma, there might be some gaps or inconsistencies. Sometimes, the data might mostly come from one area, making it not so representative of the whole picture. And in some cases, parts of the data might need to be taken out to make sure people's information stays private and protected. This missing or incomplete data could make it harder to dig deep into the analysis, possibly making it less reliable when trying to connect fire smoke to respiratory problems. Also, how accurate the data is and how it's reported, especially when compared to fire activity info, might make it tough to find clear connections.

Moreover, health issues, especially ones affected by things like fire smoke, can be influenced by many different factors. Things not in the datasets, like changes in how healthcare works, rules about the environment, or even how society is, can also play a part in how healthy people are. Trying to figure out how much fire smoke specifically affects respiratory health might be tricky because of these other things that can also have an impact. So, to make the analysis, we're making some guesses by assuming these other things stay the same to focus on how fire smoke might affect breathing problems.

Timeline to Completion

Below are the timelines and project tasks:

Nov 17, 2023 – Problem statement submission (Part 2 - Extension plan)
Finalize data collection and document the motivation, hypotheses, and methodology.

Nov 19, 2023 – Data extraction and exploratory data analysis
Extract county-specific data and conduct exploratory analysis to grasp overall trends in hospitalizations and mortality.

Nov 21, 2023 – Regression and Correlation analysis
Extend the analysis from Part 1 by examining significant search patterns highly linked to the progression of smoke fires, hospitalizations, and fatalities in the county.

Nov 23, 2023 – Model diagnostics and inference analysis
Validate model assumptions, conduct diagnostic checks, and derive conclusions from the analysis.

Nov 25, 2023 – Data Visualization
Create visual representations of strongly correlated hospitalization and mortality trends.

Nov 30, 2023 – Presentation of key takeaways and findings (Part 3)
Craft a compelling narrative presenting the key takeaways and findings from the regression and correlation analyses.

Dec 11, 2023 – Final report documentation (Part 4)
Submit the comprehensive final report documenting the analysis, conclusions, and potential future steps for the project.

References

Ref 1: Tiotiu, A.I., Novakova, P., Nedeva, D., Chong-Neto, H.J., Novakova, S., Steiropoulos, P. and Kowal, K., 2020. Impact of air pollution on asthma outcomes. International journal of environmental research and public health, 17(17), p.6212.