Estimating Connections Between Wildfire Smoke Impact and Respiratory-Related Mortality in Fresno, California

Part 4: Written Report

Sarah Kilpatrick | DATA 512

Introduction

Fresno is a mid-sized city in the California Central Valley, an economic hub predominantly contributing to the agricultural industry. As of 2020, the city boasts over half a million inhabitants; it is geographically centered between population centers in Northern California, Southern California, and Nevada. There is evidence to suggest that wildfires are increasing in not only size but also intensity. Airborne pollutants from wildfires in forests hundreds of miles away from Fresno still pose the risk of exacerbating chronic heart and lung conditions (Environmental Protection Agency).

Wildfire data from the United States Geological Survey (USGS) includes 135,026 documented wildfires over the course of 128 years (Welty et. al). 47.92% of these fires occurred within 650 miles of Fresno. This makes Fresno a great candidate for analyzing the association between wildfire-related smoke prevalence and chronic, lower-respiratory disease-caused mortality (CLRD) in Fresno over 1989-2019.

This analysis follows current literature surrounding the impact of wildfire-related air pollution and mortality in California. However, by focusing the scope of this project on Fresno, CA, this analysis begins to fill a gap in the literature. Fresno's population is vulnerable to the effects of air pollution. The Central Valley of California often contends with placing its agricultural farm-workers in unsafe working conditions, and as Fresno is in the center of this agricultural crossroads, there is a growing concern that the city is not adequately measuring these effects or equipping its citizens with reasonable tools and programs.

Background/Related Work

Studies have shown that smoke from wildfires significantly exacerbates respiratory illnesses and even short-term exposure to wildfire-related pollution was associated with an increased risk of mortality (Chen et al.). Vulnerable populations, such as children, the elderly, and those who work outdoors, are particularly at risk. A recent study found that wildfires releasing large amounts of particulate matter have led to a change in aerosol compositions in the stratosphere, emphasizing the broader atmospheric impacts of wildfire smoke (e.g., pyrocumulonimbus events) and potential long-term health implications (J. M. Katich et al.).

Research from Johnson and Garcia (2023) reports that models for smoke impact perform best when they integrate wind patterns and total area burned into their models. Additionally, meteorological stations report that the prevailing wind direction in California's Central Valley during fire season comes from the northwest. Therefore, I chose to integrate not only a wildfire's total acres burned, but also the cardinal direction of such fires relative to Fresno.

Furthermore, the existing literature suggested this analysis test the hypothesis that discrepancies in mortality over the course of multiple years also follows the multi-year seasonality of wildfire intensity: that is, intense fire years are often followed by a few less intense years, as vegetation grows back, which exacerbates fire size and the cycle begins again. We hypothesize this cycle is also visible in the mortality caused by Chronic, Lower Respiratory Diseases (CLRD) with most deaths lagging after particularly intense fire years.

Model Documentation

Holt-Winters Exponential Smoothing Model: This model is relatively simple, fast, known to work well with trend and seasonality and has low parameter tuning requirements. However, the lack of flexibility for irregular data limited this model's capability to capture the historical data's variability.

Seasonal Autoregressive Integrated Moving Average with Exogenous Regressors (SARIMAX) Model: The most flexible out of the three for handling seasonality and non-stationary data; however, the forecasted results were not very statistically significant.

Convolutional Neural Network (CNN): this model was the most promising. It was adapted for time series data, as found on the Tensorflow website. This method is promising but there are opportunities to improve its ability to generalize to unseen data. It can handle non-linear patterns, flexible input, but it is most useful for large datasets, rather than univariate data over only 30 years. This model was chosen due to its best success relative to the other models in assessing unseen test data. The first model proved to be unadaptable to the historical data and the second model's parameter tuning proved to be beyond the project's timeline.

Data used in all models consists of the independently-generated smoke estimates and public domain data provided by the California Department of Health and Human Services.

Analytical Methodology

The links between increased wildfire-related pollution and increased mortality in California are well-documented (Chen et al., Gwon et. al). This analysis creates a predictive estimate for the impact of smoke caused by wildfires within 650 miles of Fresno. Then, data on deaths caused by chronic lower respiratory diseases inform the predictive estimate to quantify future impacts of smoke from a public health perspective. This project presents a refinement of the preliminary smoke estimate and discusses the use of non-periodic time series forecasting techniques. This project generates correlations between the refined smoke estimate and Fresno's deaths caused by respiratory conditions. In turn, calling for a study detecting causal connections between the two phenomena.

Smoke Estimate Methodology

The Smoke Impact Score provides an estimation of the potential air quality impact of wildfires on Fresno, CA. The inclusion of the median cardinal weight (*W*) accounts for Fresno's vulnerability to smoke due to prevailing winds in California's Central Valley. During the fire season (May 1–October 31), winds frequently originate from the northwest, amplifying the transport of wildfire smoke into the region. The scaling factor (*F*) ensures that the resulting score is intuitive and comparable to established air quality indices, making it a practical tool for understanding and communicating wildfire smoke impacts.

The use of C_l is useful because distance correlation detects both linear and nonlinear associations, capturing more complex relationships that may exist between the preliminary smoke estimate and deaths due to CLRD. It is multiplied by 10 to give it an amplified effect to the smoke estimate where a higher statistically significant correlation between the non-informed smoke estimate and Fresno's observed CLRD deaths means a higher, and therefore more intense, final smoke estimate. The lag in years (l) can be set between 0-5 to measure the association between a wildfire in year x and deaths in year $x+l$.

This expression uses the inverse square law, a principle modeling diminishing intensity of a physical effect with the square of the distance, reflecting how smoke disperses over greater distances. By adding 1 to the distance in miles (*D)*, the formula avoids division errors in rare cases where a wildfire's perimeter intersects Fresno's city boundaries. The use of medians (*A*, *W*, and *D*) instead of averages improves the robustness of the estimate by reducing the influence of outliers, such as unusually large fires or extreme distances; this is an update to the preliminary smoke estimate which used averages.

Findings

Results graphed from the dataset provided by the U.S. Geological Survey show us that throughout the decades, one thing is clear: wildfires pose a close and increasing threat to Fresno. For example, **Fig. 1** shows that nearly 50% of all 130,000 fires tracked in the United States are within 650 miles of Fresno, realistically impacting the city's air quality. Wildfires are also becoming more expansive and destructive. **Fig. 2** shows that even at a wider scale, within 1,800 miles of the Central Valley, the average total acres burned is only increasing as time goes on. In 2008, the total acres burned surpassed 5M, and 2020 was the first year where more than 7M acres were burned in one year.

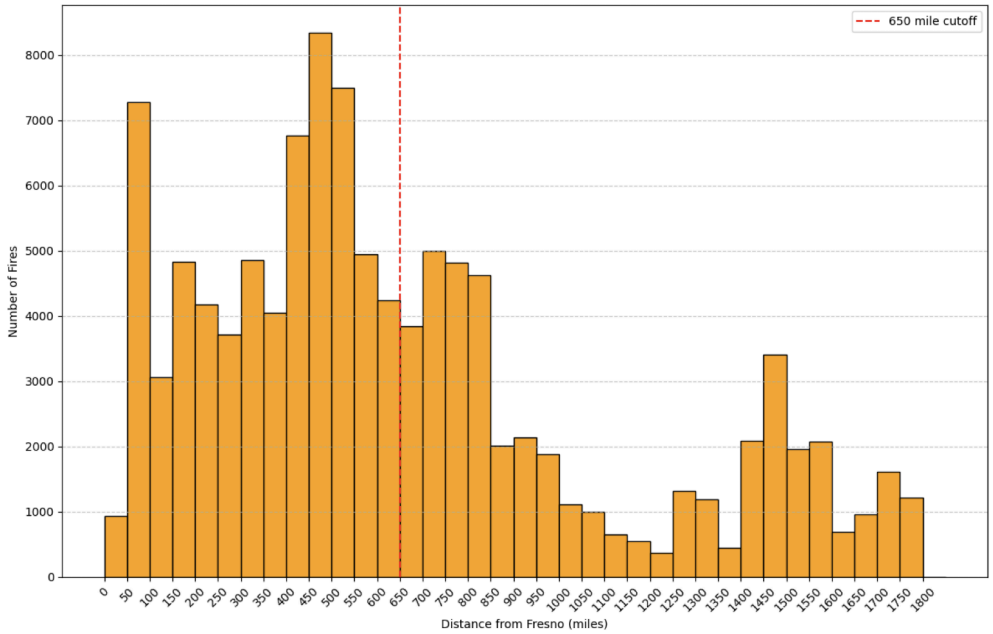

Fig. 1 Wildfire data from the United States Geological Survey (USGS) includes 135,026 documented wildfires over the course of 128 years (Welty et. al). 47.92% of these fires occurred within 650 miles of Fresno.

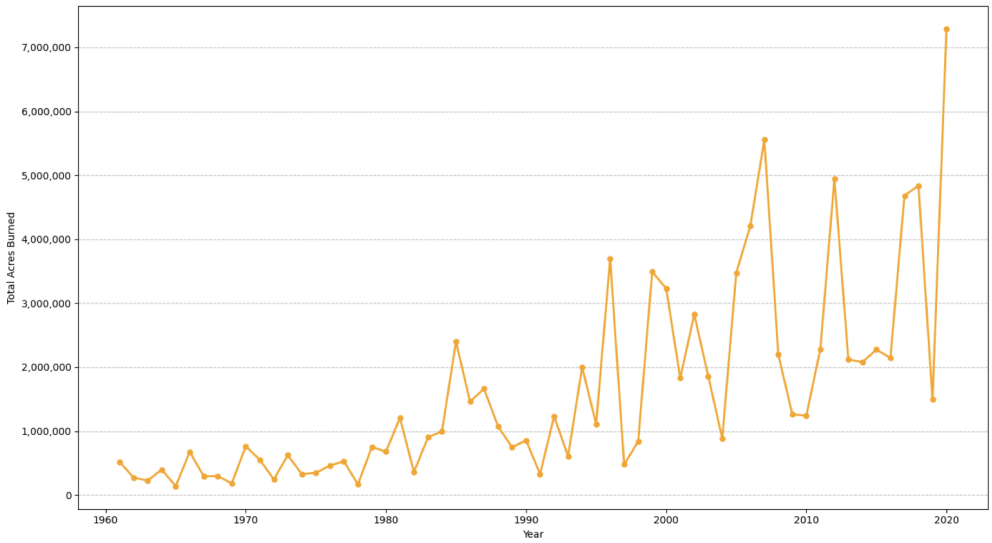

Fig. 2 Total Acres Burned Per Year (1961-2020). Total acres burned in North America have become increasingly variable over the years. While large fire years are not inherently harmful to fire-dependent ecosystems, the rise in particularly large, destructive fires is likely to significantly impact respiratory health (Donato et al.).

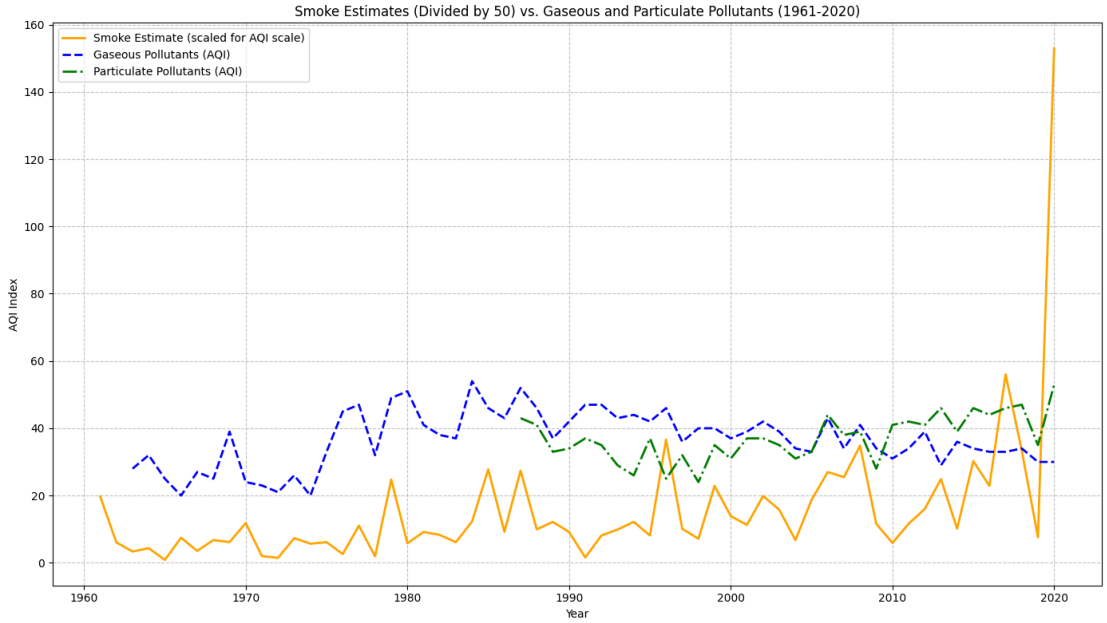

Fig. 3 Preliminary Smoke Estimate compared with EPA AQI data. The preliminary estimate, scaled for visual alignment with the AQI data, is actually not dissimilar to the Gaseous and Particulate pollutants. Notice how the Gaseous pollutants and the smoke estimate are similar in the late 1960s, the 1980s, and even the late 2000s. However, with only a modest Pearson's correlation of around 0.3, this preliminary smoke estimate has room to be improved.

In a similar vein to increasing wildfire prevalence, there is a noticeable uptick in mortality due to CLRD in Fresno in recent years. As seen in **Fig. 4,** the past 7 years have recorded the closest mortality due to CLRD to the 30-year maximum than the past 30 years. Statistical methods detecting linear and nonlinear links can tease out meaningful associations between two wildfire smoke estimates and deaths due to CLRD.

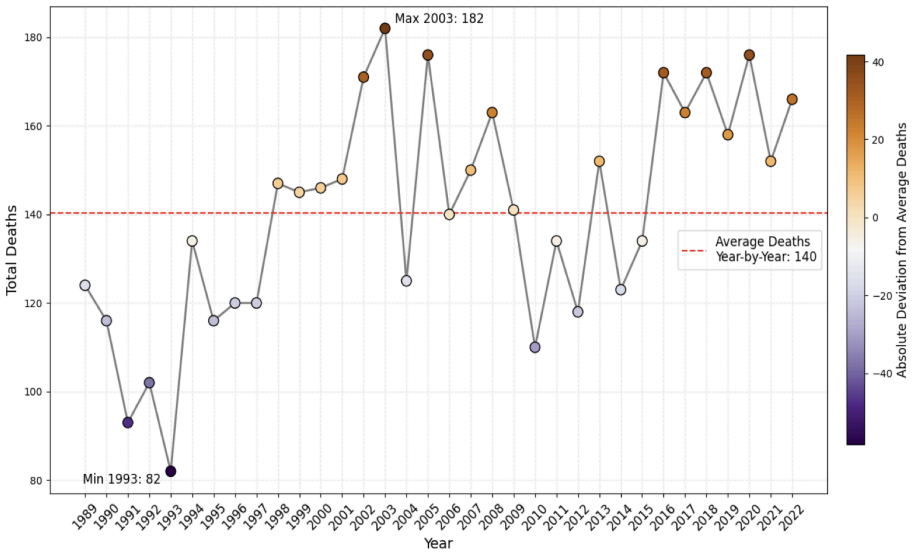

Fig. 4 Fresno, CA (1989-2019) Deviation from Average Deaths Due to CLRD per 100,000 people.

Furthermore, a geographic assessment of Fresno's mortality due to CLRD (1989-2019) demonstrates that 10 out of 19 zip codes experience a higher-than-average rate of deaths. **Fig. 5** shows that some zip codes have even experienced higher than 800 deaths per 100,000 people in the CalHHS record.

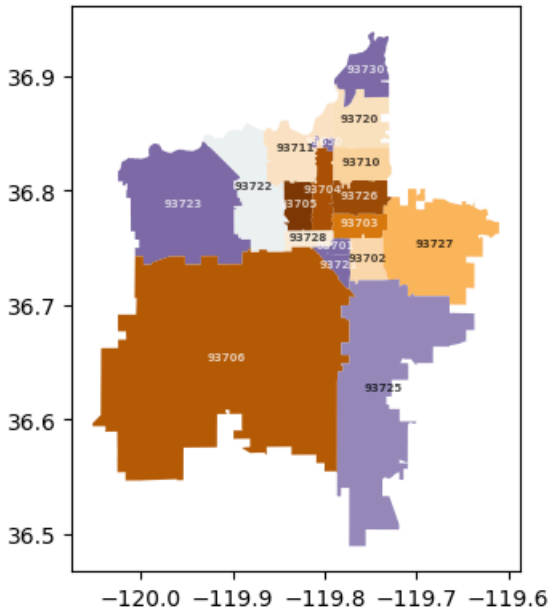

Fig. 5 Total deaths due to CLRD by year (1989-2019) provided by California HHS.

Integrating the association between wildfire smoke prevalence and mortality into our final smoke estimate requires creating a set of first-round smoke estimates: one that uses averages, and one that uses medians. These first-round estimates involve a numerator multiplying the average/median acres burned in one year by the average/median cardinal direction of fires in that same year. Then, this numerator is divided by an expression for average/median distance of fires in that year, in miles, away from Fresno, using the inverse square law.

Then, a variety of variables (including the preliminary estimate using averages, and the updated estimate using medians) are included in a set of distance correlations computed with mortality in the years (0-5) after the wildfire years. **Fig. 6** presents the results:

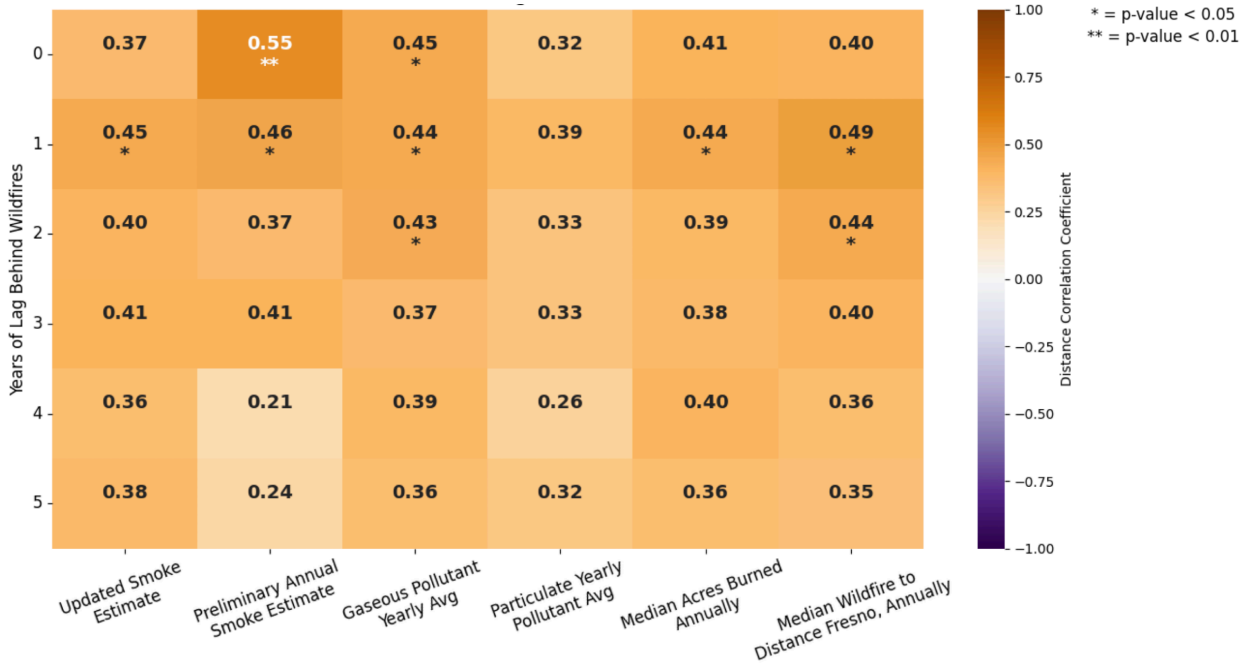

Fig. 6 Correlation between two smoke estimates and associated variables with deaths due to CLRD 0-5 years following the wildfire years. Boxes with asterisks are statistically significant with p-values less than 0.05, and boxes with two asterisks are more statistically significant with p-values less than 0.01.

The correlation matrix above reveals a few important findings:

1. The updated smoke estimates are more associated (positive correlation) with mortality in the 12-24 months following a wildfire year than they are in any other time frame.
2. The preliminary smoke estimates are most associated (positive correlation) with mortality in the same year as the wildfire year.
3. Gaseous pollutants, median acres burned annually, and median wildfire distance to Fresno are each most positively correlated with mortality in the 12-24 months following a wildfire year.
4. Gaseous pollutants in 0-2 years after a wildfire year are more positively correlated with mortality due to CLRD than particulate pollutants.

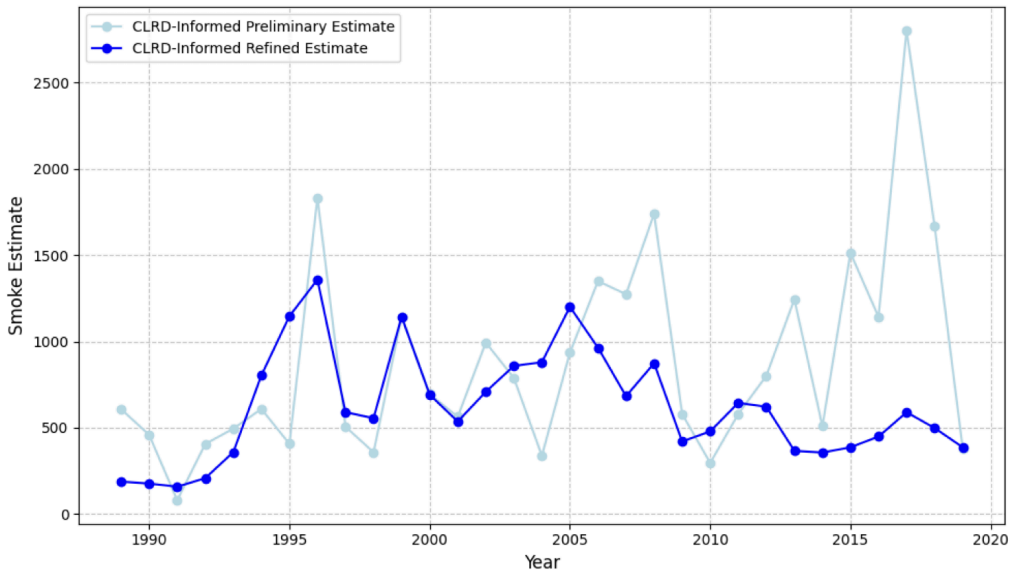

Fig. 7 Comparison of final smoke estimate using averages and final smoke estimate using medians.

Lastly, comparing the two final smoke estimates integrating CLRD deaths reveals a few new insights: the updated (refined) estimate using medians is much less spiky, because medians are more robust to outliers. However, outliers can be useful, especially when these wildfire outliers have the capacity to greatly risk the health of Fresno's residents.

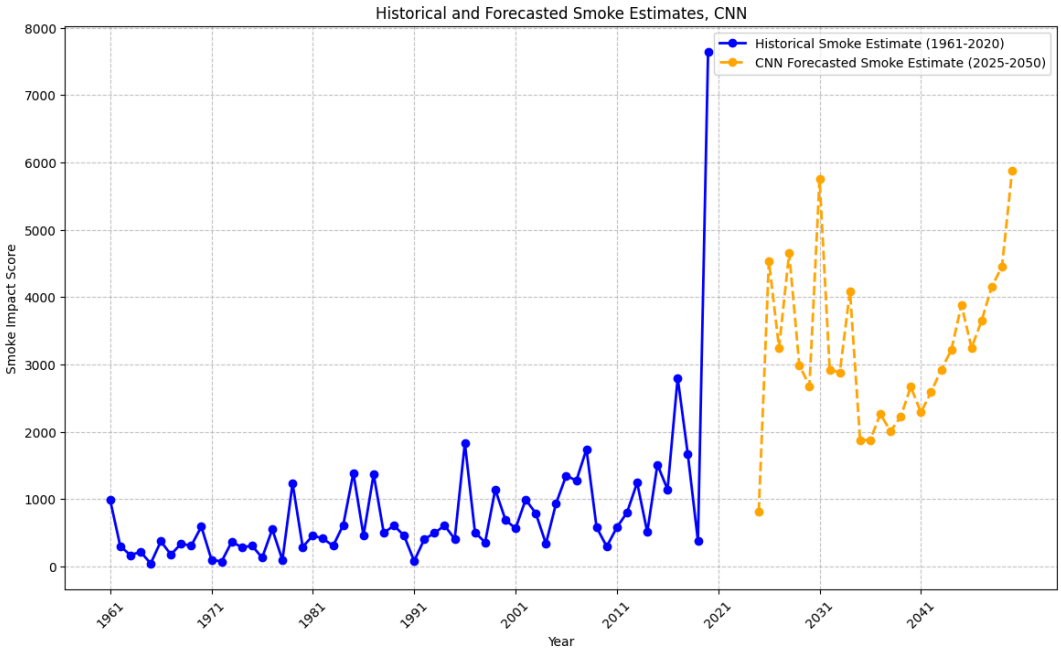

Fig. 8 Forecasted smoke estimate in orange shows a growing impact over time.

In order to plan ahead, Fresno city officials can take advantage of a variety of predictive models. The Convolutional Neural Network (CNN), for instance, provides a readable and visually striking vision of the coming years: in the near future, Fresno can expect smoke impact on CLRD-related deaths to continue to be highly variable. Then, as we approach the mid-21st century, Fresno will see consistently-rising impacts.

Discussion/Implications

These results provide valuable and statistically-significant associations between wildfire smoke and mortality due to CLRD. Not only are these results in line with the current literature, but the linkages between the two phenomena raise questions to any causal connections between air pollution and CLRD mortality.

Fresno city residents can take advantage of state and federal programs providing resources for supporting respiratory health. Chronic Lower-Respiratory Diseases do their damage over a long period of time, and integrating healthy lifestyle changes will help Fresno residents to support their health in a changing future. City leaders can take advantage in investing their time, attention, and possibly their budget into further research on the causal association between air pollution in the Central Valley and mortality due to CLRD.

Fresno would benefit from looking into these linkages immediately. There is evidence in the literature suggesting that creating a concrete plan for supporting social, economic, and environmental programs would help their residents as soon as possible. It is never too early to begin implementing further research findings into social and economic measures.
Human-centered data science principles are as important in the initial analytical process as they are in communicating the results. This analysis maintains a commitment to reproducibility by using publicly-available data and open-source statistical methods. This frames my code not as a black box but as an "expression of human reasoning" (Knulth). Furthermore, the notebook is designed to be human-readable, organized, and modular, for replicability purposes. Then, this analysis takes necessary steps to reflect Fresno's population in good faith; this is seen in not only the discussion of limitations of the geographic analysis, but also the highlighting of opportunities to deepen the study.

Additionally, by looking into a variety of descriptive statistics and predictive models, this analysis aims to minimize the amplification of bias by assessing all of these findings holistically. Lastly, considering various iterations of the chosen CNN model moves away from the traditional, unidirectional workflow identified in Amershi et al. As with all academic investigation, this research remains open-ended. By making my analysis and findings open-source, I contribute to human-centered data science by inviting future researchers to use my research as a stepping stone for further investigation.

Limitations

All data is public, and as it stems from government sources, is expected to remain public. The data from the California Health and Human Services (CalHHS) contains only deaths due to CLRD from hospitals, which obfuscates the true amount of deaths due to CLRD. Additionally, the scope of the CalHHS data was only from 1989-2019, which restricts our analysis by nearly 50%. Furthermore, the data quality from the United States Geological Survey is inconsistent throughout the decades: wildfires were not measured the same way in the early 60s as in the 21st century. This could potentially confound the input data.

Missing values for CalHHS data between 1961-1988 were not imputed. This was due to the principle that forecasting future results should not come from computer-generated data, but rather from more faithfully-reported, real information.

Both the Holt-Winters Exponential Smoothing model and the SARIMAX model assume stationary data. The input data was pre-processed to be stationary before using the models to ensure forecasting integrity.

The Distance Correlation does not assume a linear association, which makes it more desirable over the commonly-seen Pearson or Spearman correlation. It is also a non-parametric metric, so it does not assume any specific distribution for the data. It does, however, assume the existence of variance in the data and that the two phenomena are measurably distinct.

Conclusion

When The U.S. Geological Survey reports an intense wildfire year, Fresno has historically seen a spike in mortality the following year. This finding is substantiated by current literature assessing the impact of air pollution on respiratory health and the growing amounts of air pollutants generated by increasingly large wildfires in the United States. The smoke estimates take advantage of a unique set of variables, including an independently-created metric for prevailing winds through the California Central Valley. Then, the statistically-significant positive correlation established in the findings presents an opportunity to establish a causal relationship between air quality impacted by wildfire smoke and mortality in Fresno caused by Chronic, Lower-Respiratory Diseases.

This study serves as an example of data science-related research following Guidelines for Human-AI interaction (Amershi et al.) and using Python as a “Literate Programming Tool” (Kery et al.). This study also contends with common ethical blind spots in machine learning-related research by providing a transparent discussion about the models chosen and an in-depth look at the methods and their limitations. Ultimately, it enriches the reader’s understanding of what human-centered research entails.

Opportunities for Continuation

Account for annual income and occupation, key predictors of health outcomes.
Incorporate evidence suggesting that wildfire-induced air pollution in the stratosphere may significantly influence pollutant dispersion alongside tropospheric winds (J.M. Katich et al.).
Investigate the confounding effects of air pollutants originating from the northwestern San Francisco Bay Area, which may impact Fresno independently of wildfire activity.
Incorporate a more refined dispersive mathematical model for wildfire smoke.
Include or restrict different time periods for a more specific analysis.
Include Canadian wildfires into this analysis.

References

Daniel C. Donato, Joshua S. Halofsky, Derek J. Churchill, Ryan D. Haugo, C. Alina Cansler, Annie Smith, Brian J. Harvey, Does large area burned mean a bad fire year? Comparing contemporary wildfire years to historical fire regimes informs the restoration task in fire-dependent forests, Forest Ecology and Management, Volume 546, 2023, 121372, ISSN 0378-1127, https://doi.org/10.1016/j.foreco.2023.121372.

"Fires and Your Health." Fires and Your Health | AirNow.Gov, AirNow.gov, U.S. EPA, www.airnow.gov/air-quality-and-health/fires-and-your-health/. Accessed 15 Nov. 2024.

Gongbo Chen et al., Mortality risk attributable to wildfire-related PM2·5 pollution: a global time series study in 749 locations, The Lancet Planetary Health, Volume 5, Issue 9, 2021, Pages e579-e587, ISSN 2542-5196, https://doi.org/10.1016/S2542-5196(21)00200-X.

Gwon JG, Lee SA, Park KY, Oh SU, Kim JS, Seo HM. Long-Term Exposure to Air Pollution and Incidence of Venous Thromboembolism in the General Population: A Population-Based Retrospective Cohort Study. J Clin Med. 2022 Jun 19;11(12):3517. doi: 10.3390/jcm11123517. PMID: 35743587; PMCID: PMC9224855.

Jaffe, D. A., O'Neill, S. M., Larkin, N. K., Holder, A. L., Peterson, D. L., Halofsky, J. E., & Rappold, A. G. (2020). Wildfire and prescribed burning impacts on air quality in the United States. Journal of the Air & Waste Management Association, 70(6), 583–615. https://doi.org/10.1080/10962247.2020.1749731.

J. M. Katich et al., Pyrocumulonimbus affect average stratospheric aerosol composition. Science 379, 815-820(2023). DOI:10.1126/science.add3101.

Johnson MM and Garcia Menendez F (2023) International Journal of Wildland Fire doi:10.1071/WF22172.

Makowski, Dominique, et al. *BayestestR: A Suite of Functions to Report and Support Bayesian Analysis*. 2019, dominiquemakowski.github.io/publication/makowski2019bayestestr/makowski2019bayestestr.pdf.

"Time Series Forecasting : Tensorflow Core." *TensorFlow*, www.tensorflow.org/tutorials/structured_data/time_series. Accessed 2 Nov. 2024.

Welty, J.L., and Jeffries, M.I., 2020, Combined wildfire datasets for the United States and certain territories, 1878-2019: U.S. Geological Survey data release, https://doi.org/10.5066/P9Z2VVRT.\

Data Sources

Death Profiles by Zip Code (1989-2022), Hosted on the California Health and Human Services Open Data Portal, provided by the California Department of Public Health.
Dataset:
CSV Files of:

- 2019-2022 Final Deaths by Year by ZIP Code
    - "This data table reports the annual number of deaths of California residents by the ZIP Code of residence regardless of where the death occurred (by residence). The cause of death categories are based solely on the underlying cause of death as coded by the International Classification of Diseases." (CalHHS)
- 2009-2018 Final Deaths by Year by ZIP CodeCSV Popular
- 1999-2008 Final Deaths by Year by ZIP CodeCSV Popular
- 1989-1998 Final Deaths by Year by ZIP CodeCSV Popular
- Data Dictionary - Deaths by ZIP CodeCSV Popular

License/Terms of Use: Open for public use contingent on following the CalHHS terms of use.

Combined Wildfire Datasets for the United States and Certain Territories (1878-2019), Hosted on the ScienceBase Catalog. Map services and data available from the U.S. Geological Survey, National Geospatial Program.
Dataset:

- JSON Files of:
    - Wildfires 1878-2019 in the Contiguous US as Rasters

License/Terms of Use: Open-use and in the public domain with appropriate citation.