Trends of Air Pollution-Related Diseases Across California Counties

Tom Regpala

Department of Computer & Electrical Engineering & Computer Science

California State University, Bakersfield

Dr. Anjana Yatawara

August 19, 2024

Introduction

From vehicular exhaust to mass-combustion of fossil fuels, our global atmosphere is subject to an ever-growing injection of hazardous chemicals. Directly influenced by human industry, air pollutants such as carbon monoxide (CO), sulphur dioxide (SO_2), and particulate matter (PM_x) are increasingly melding into our atmospheric composition. These pollutants create adverse effects on public health, contributing to increased risk of respiratory and cardiovascular diseases. In the scope of Kern County, California, the impact of air pollution could manifest beyond human health. Characterized by an economy in agriculture and oil production, poor air quality from local petroleum industries could have lasting effects on crop yields and susceptibility to disease.

While many existing studies examine air pollution across different areas over time, there is a lack of emphasis on the socio-economic qualities of the examined regions. Lamont and Arvin, two cities within Kern County, are distinguished by their agricultural presence and a working-class, majority-Hispanic population. These communities are continuously exposed to a dusty, particle-polluted environment, creating considerable risk for respiratory disease. The degree of risk these people experience calls for an examination of disparities in pollution exposure between socio-economic groups.

Kern County's large oil industry presence and proximity to agriculture provides a unique microcosm of global pollution worthy of study. The main objective of this research is to investigate air pollution levels and trends of related diseases in Kern County while comparing it with other California regions. Specifically, concentration of major pollutants in the local atmosphere and their adverse health risks associated with them. Prevalence and predisposition to

diseases such as influenza-like illness (ILI), lung cancer, and stroke are anticipated to highlight differences between social strata.

Research into air pollution is inherently intertwined with public and environmental health, directly contributing to the wellbeing of the general populace. By contextualizing our study beyond computational analyses and into a socio-economic lens, unique insights to local policy could be derived and further contribute to public discourse. Kern County's distinctive mix of agriculture and industry naturally demands exploration into the atmospheric environment it creates. By establishing this foundational knowledge, future research could expand on the distribution of pollution-related diseases between racial and ethnic groups and how public policy could work to minimize morbidity.

This paper is organized as follows. The second section is a literature review with a focus on the health impacts of air pollution and how individual pollutants impact the human body. The third section outlines the collection process of various datasets used for analysis. The fourth section describes the methodology and statistical techniques used to derive meaning from the gathered data. The fifth section conducts an in-depth analysis of the examined data through numerous figures and visualizations. The sixth section discusses the results and limitations of the datasets collected and analyzed. The seventh section concludes this paper.

Literature Review

This literature review focuses on studies discussing the adverse effects of air pollution on human health, specifically stroke, lung cancer, and Influenza-like illness. Additionally, research on air pollution-related health impacts across racial, ethnic, and poverty groups is given particular attention, with emphasis on California public health.

While it is intuitive to draw a link between respiratory illness and air pollution, other health effects are not given as much attention. In his editorial from a medical journal, Michael Brauer discusses two studies on the effects of air pollution on stroke and anxiety (2015). On further study into Brauer's background, his research as a professor of public health at the University of British Columbia involves examining ethnic disparities in air pollution mortality across the United States. After a meta-analysis of "103 studies conducted in 28 countries and including 6.2 million events," the results of the first study (Shah et al., 2015) indicated a strong association between increases in particulate air pollutants and stroke incidence. Brauer highlights air pollution as a possible modifiable risk factor, meaning stroke hospitalizations and mortality can be managed by the change in air pollutant exposure. However, it is noted that further study is needed on the effects of pollution exposure on preconditions for stroke.

Malignant neoplasms of the lung, also known as lung cancer, remains at the top for cancer-related deaths in the United States. Unlike stroke, lung cancer has a known association with air pollution, which is examined by IAJ Kusumawardani (2023). In her study, she highlights four major mechanisms of lung cancer due to air pollution: Oxidative stress, inflammation, DNA damage, and epigenetic changes. Firstly, oxidative stress is caused by the presence of reactive oxygen species (ROS). More specifically, exposure to particulate matter of designation 2.5μm ($PM_{2.5}$) and smaller causes production of ROS. The resulting oxidative stress causes "nucleic acids, proteins, and fats to oxidize, facilitating inflammatory cell invasion" (Kusumawardani 2023). Next, the inflammatory response to pollutant exposure is studied and revealed to cause an increase in expression of cytokines. Cytokines, in the context of cancer spread, are known to trigger growth of malignant tumors. Additionally, DNA damage from air pollution manifests through reverse transcription of p53, a gene known known to suppress tumors. Specifically, $PM_{2.5}$ exposure stimulates the mutation of p53, causing cell death and cell growths.

Kusumawardani's mechanisms of lung cancer all point to $PM_{2.5}$ exposure as one of the higher risks for lung cancer. She cites numerous studies that report an increase in lung cancer incidence alongside a 10-unit increase in $PM_{2.5}$ exposure. While significant incidence is noted in active and former smokers, non-smoker risk maintains the same trends with the other populations. The article ends with a stressed importance for further study in air pollution as a risk factor for lung cancer in non-smokers.

With $PM_{2.5}$ and PM_{10} being able to carry viruses and spread infection (Zhang et al., 2023), incidence of Influenza and other like illnesses (ILI) throughout season and climate changes are a growing field of study. Coupled with irritant gases like Ozone (O_3) and Sulfur dioxide (SO_2), which invade the respiratory tract through the respiratory mucosa, air pollutants create a susceptibility to viruses by the weakening of the immune system. Zhang's research suggests that seasonal changes in temperature lead to variation in air quality and concentration of pollutants, hinting at a complex association between Influenza, air pollution, and climate. Furthermore, using variables such as daily mean temperature, daily precipitation, and daily concentration of various pollutants, Zhang explored the relationships between pollutants and Influenza through a quasi-Poisson regression model. By applying the model to various seasonal periods, it was found that winter temperatures caused pollutants to create harsher irritations in the respiratory tract, coinciding with the CDC's recognized "flu season". The study concludes with speculation that the findings could assist in optimizing influenza warning systems developed by environmental protection agencies.

From how easily it is sourced and associated with severe health impacts, research on air pollution most often targets $PM_{2.5}$ for study (Nawaz et al., 2023). In Nawaz's research, the health impacts of air pollution exposure are measured by the deaths grouped by pollutant. Consistently, O_3 exposure and related deaths were significantly smaller than that of $PM_{2.5}$. With population

and pollutant concentration data from fourteen cities around the United States, an anthropogenic fraction was calculated for each city and was used to highlight a substantial attribution between anthropogenic emissions and health impacts from $PM_{2.5}$ and O_3; Essentially, pollutant-related deaths were mostly sourced to man-made emissions. Additionally, some cities reported that most of their pollutants are traced out of city bounds and policies targeting urban emissions are "more effective in reducing NO_2-related health impacts" (Nawaz 2023). Regional cooperation is encouraged for city policymaking to target $PM_{2.5}$, which could demand further research on the impacts of interregional policy on pollutant concentration.

Data Collection

Our research focuses on trends in air-pollution related diseases, namely influenza-like illness, lung cancer, and stroke. Similar to Nawaz's research, our study utilizes number of deaths as a metric for the health impacts of air pollution. Additionally, U.S. air quality index (AQI) data will be used for analyzing the relationship between air quality and its health risks.

CDC's Wide-ranging Online Data for Epidemiologic Research (WONDER) system is used for grabbing death statistics for various underlying causes. Through the database query tool, results can be grouped by state, county, and frequency. In our case, monthly total deaths from 1999 to 2020 for lung cancer, influenza-like illnesses, and stroke by California county are queried and exported into separate spreadsheets by cause of death. Utilizing RStudio, the three spreadsheets are combined by matching the death count data points by their shared year, month, and county.

For outdoor AQI data, the United States Environmental Protection Agency (EPA) provides daily air quality statistics for a specified pollutant, year, and geographical area through the daily data tool. Spreadsheets are generated by year, meaning one was generated for each year

from 1999 to 2020 for $PM_{2.5}$ over all California counties. The spreadsheet columns are trimmed down to display date, county, and AQI before being combined through RStudio in a similar method to the death dataset.

Lastly, California county population total estimates are gathered from the United States Census Bureau's database. The website generates spreadsheets on a decade basis; tables for the 1990s, 2000s, 2010s, and 2020s are gathered. Finally, the datasets are formatted into long form, and then combined through RStudio into a 1999-2020 population dataset.

Methodology

After importing and formatting datasets for AQI, ILI deaths, stroke deaths, lung cancer deaths, and population by California county, RStudio is used to apply numerous statistical techniques for meaningful analysis. For the initial exploration of data, yearly averages are calculated from the monthly health condition and daily AQI datasets. Next, A death rate per 100,000 population for each condition is calculated by dividing each year's average death count by the population of the corresponding year, then multiplying by 100,000. The yearly averages and death rates are centralized under a single data object.

With the data collected under one large dataset, both trend and regression analysis are applied to study the changes in air-pollution related illnesses over time as well as modeling the relationship between air quality and health condition death rates. Over the 1999-2020 period, death rates for each health condition for selected counties are graphed. Next, correlation matrices are calculated and plotted for AQI and the health conditions in three environments: Kern, selected counties, and the entire CA state.

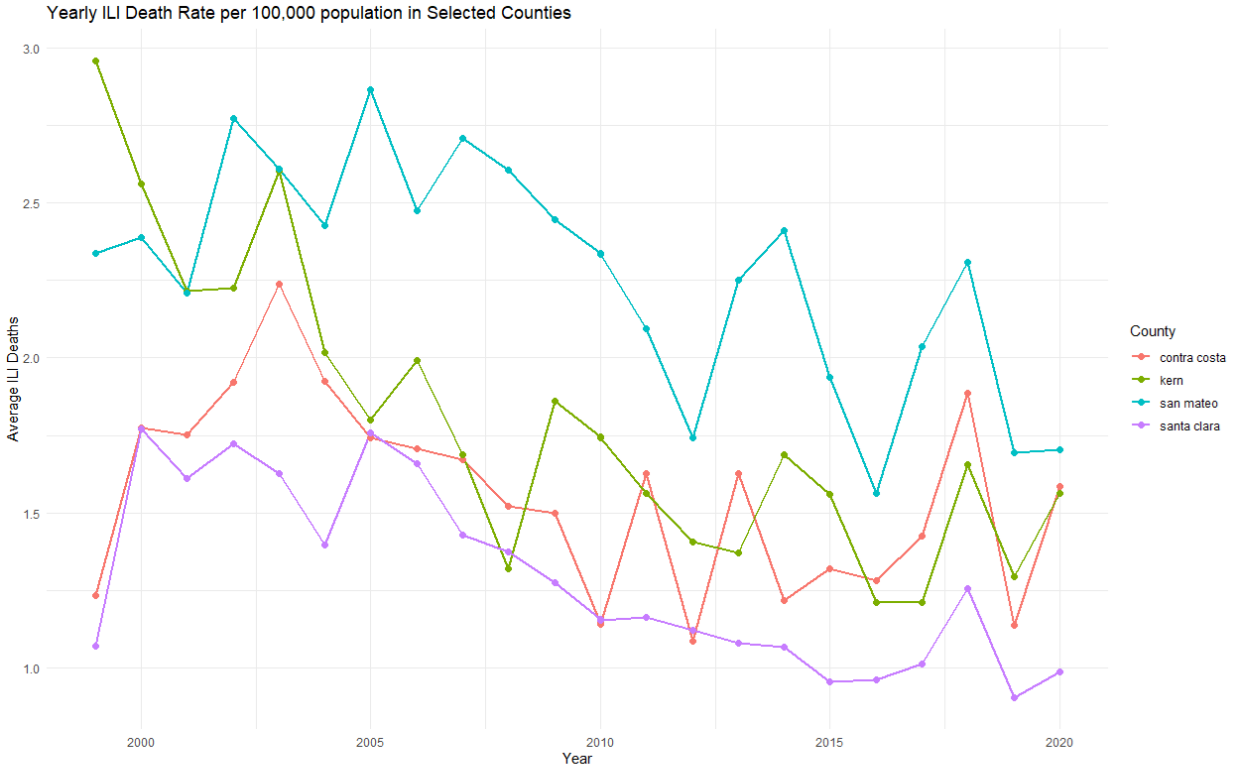
Lastly, the data undergoes geospatial analysis to generate dynamic heatmaps for AQI and death rates for ILI, stroke, and lung cancer. Utilizing RStudio, shapefiles for CA counties are

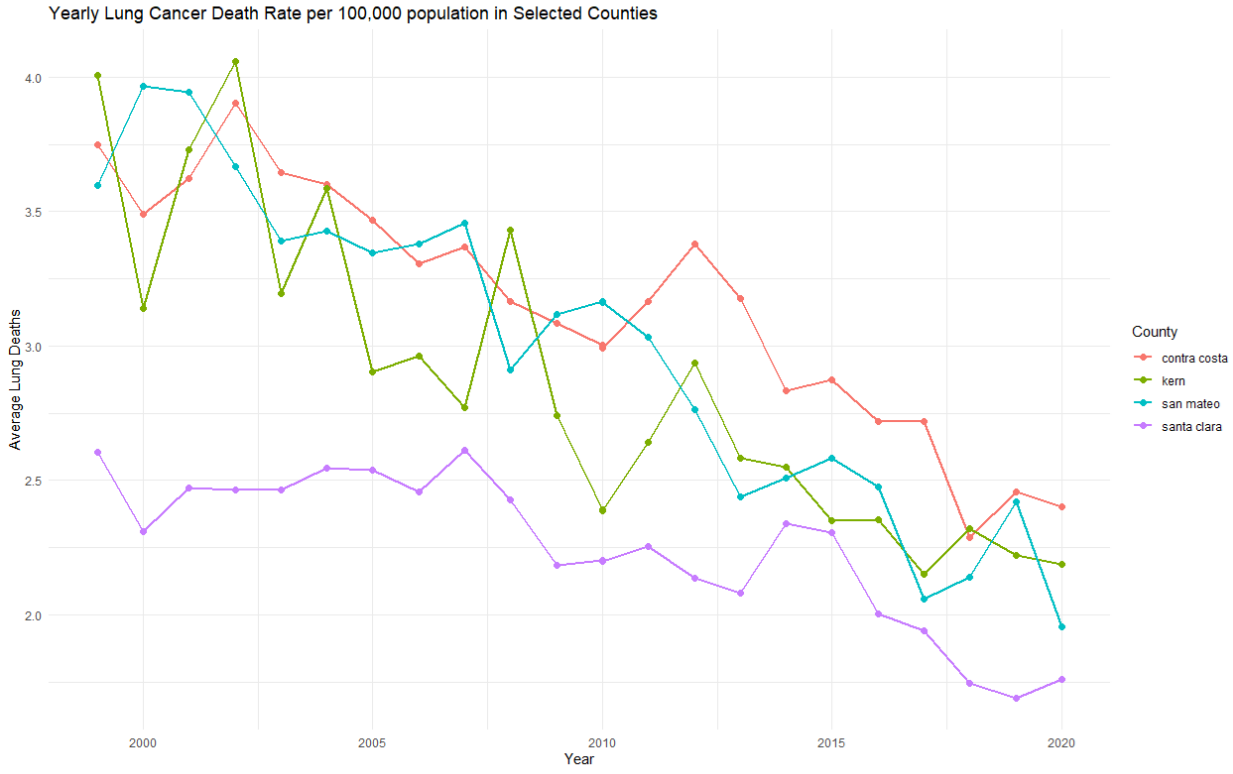
generated and matched with our combined dataset to create heatmaps for each year and for each variable. These heatmaps are combined into GIFs from external web sources to create an animated overview of how AQI and the health conditions change over time across California.

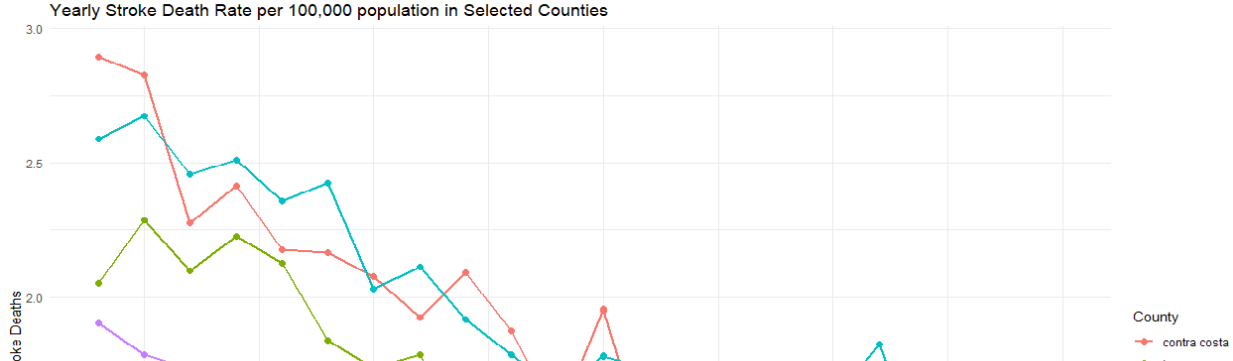
Data Analysis

For intuitive and readable plots, the amount of examined counties are limited to four: Kern, San Mateo, Santa Clara, and Contra Costa. These counties are specifically chosen to compare Kern with wealthier counties by median family household income. Some counties are excluded from assessment due to an extended period of missing air monitoring data, which could be attributed to low population or lack of reports during earlier periods of time. We showcase the trends in various air pollution related health impacts for the selected counties through three graphs (Figure 1).

Figure 1

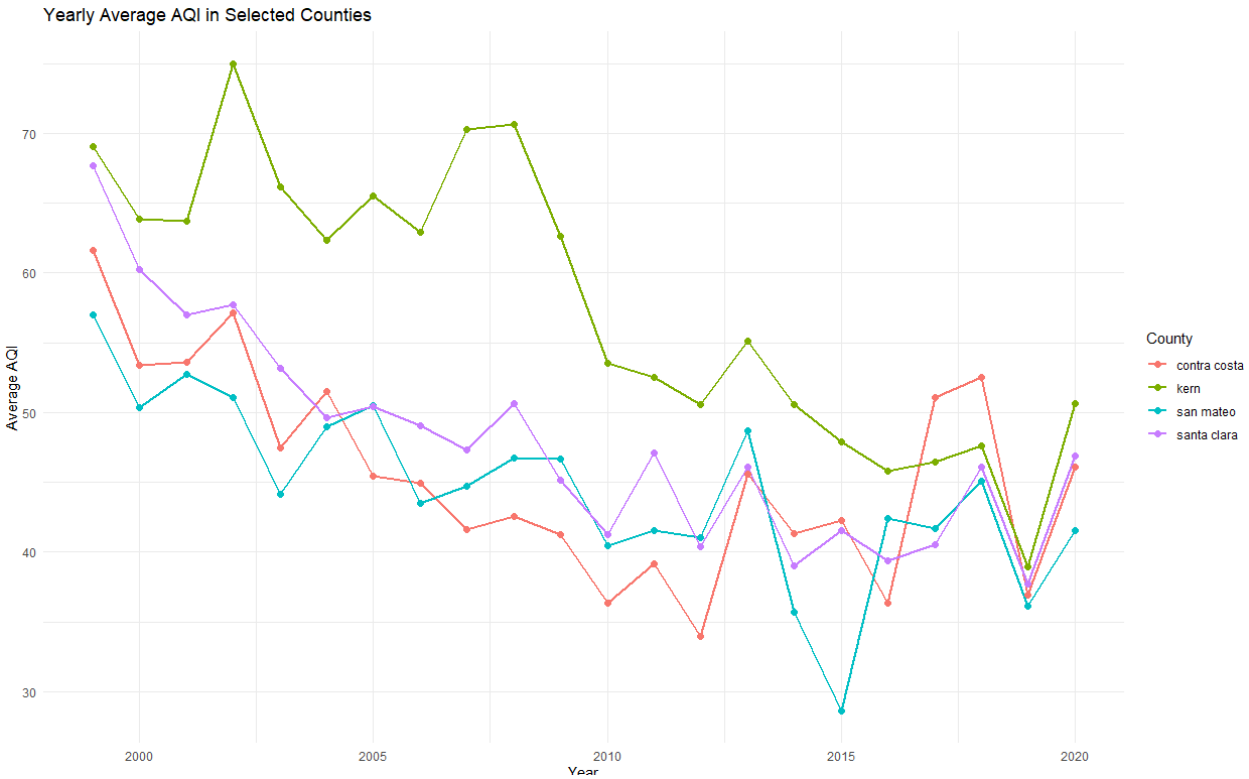

Yearly ILI Death Rate per 100,000 population in Selected Counties
Average ILI Deaths
3.0
2.5
2.0
1.5
1.0
2000
2005
2010
2015
2020
Year
County
contra costa
kern
san mateo
santa clara


Yearly Lung Cancer Death Rate per 100,000 population in Selected Counties
Average Lung Deaths
4.0
3.5
3.0
2.5
2.0
2000
2005
2010
2015
2020
Year
County
contra costa
kern
san mateo
santa clara


Yearly Stroke Death Rate per 100,000 population in Selected Counties
3.0
2.5
2.0
County
contra costa

Santa Clara, the wealthiest county, maintains the lowest death rates for ILI, stroke, and lung cancer across every year. The other three counties, however, hold similar death rates across all health conditions except ILI; San Mateo reports considerably higher rates for ILI across all years. For all studied health conditions, death rates seem to decrease over the observed years. A plot for yearly average AQI in the selected counties is created for analysis alongside these graphs (Figure 2).

Figure 2



For the AQI plot, disparities between the wealthy counties and Kern County are observed; Kern holds substantially higher AQI than the other three counties until 2017, where all counties start to converge. For every county, the 2020 average AQI is significantly lower than the starting 1999 average AQI, implying a reduction in air pollutant concentration. To better

observe the relationship between air pollution and related health conditions, correlation matrices are visualized (Figure 3).

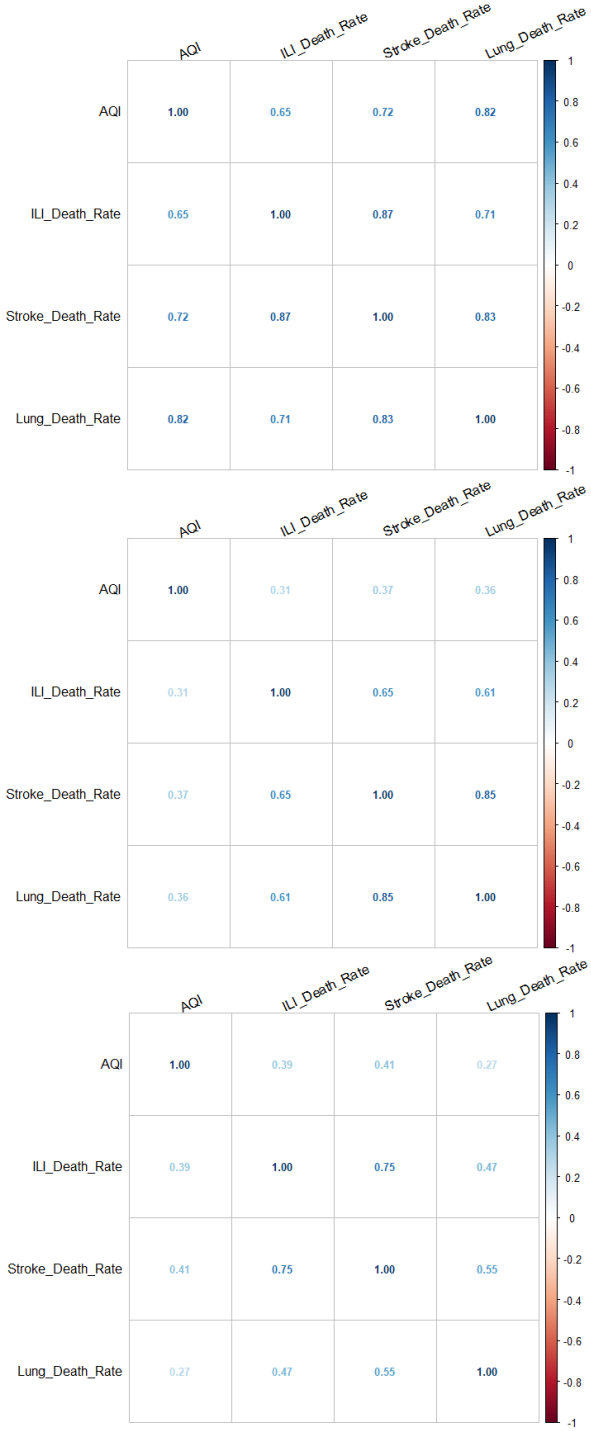


Figure 3. Plot of the correlation matrix between AQI and the death rates for ILI, Stroke, and Lung Cancer in Kern County (A), Across selected counties (B), and across the state of California (C).

As we increase in geographical scope, the correlation between each variable diminishes. Kern showcases a strong positive relationship between AQI and death rate for all health conditions. Expanding to the selected counties weakens this relationship, but slightly strengthens when applied across California. These changes could be attributed to the inclusion of sparsely populated and isolated counties. In counties with low population, data on any condition's monthly deaths are often low enough to be suppressed or have zero deaths over significant periods of time. The calculation and introduction of the death rate assists in

normalizing the data across California by weighing it over the population, which improves correlation relationships considerably.

For geospatial data, AQI and death rate heatmaps of California were created for years 1999-2020. Additionally, the heatmaps were combined to create four animated images that intuitively showcase the changes in each variable over the observed years (Figure 4).

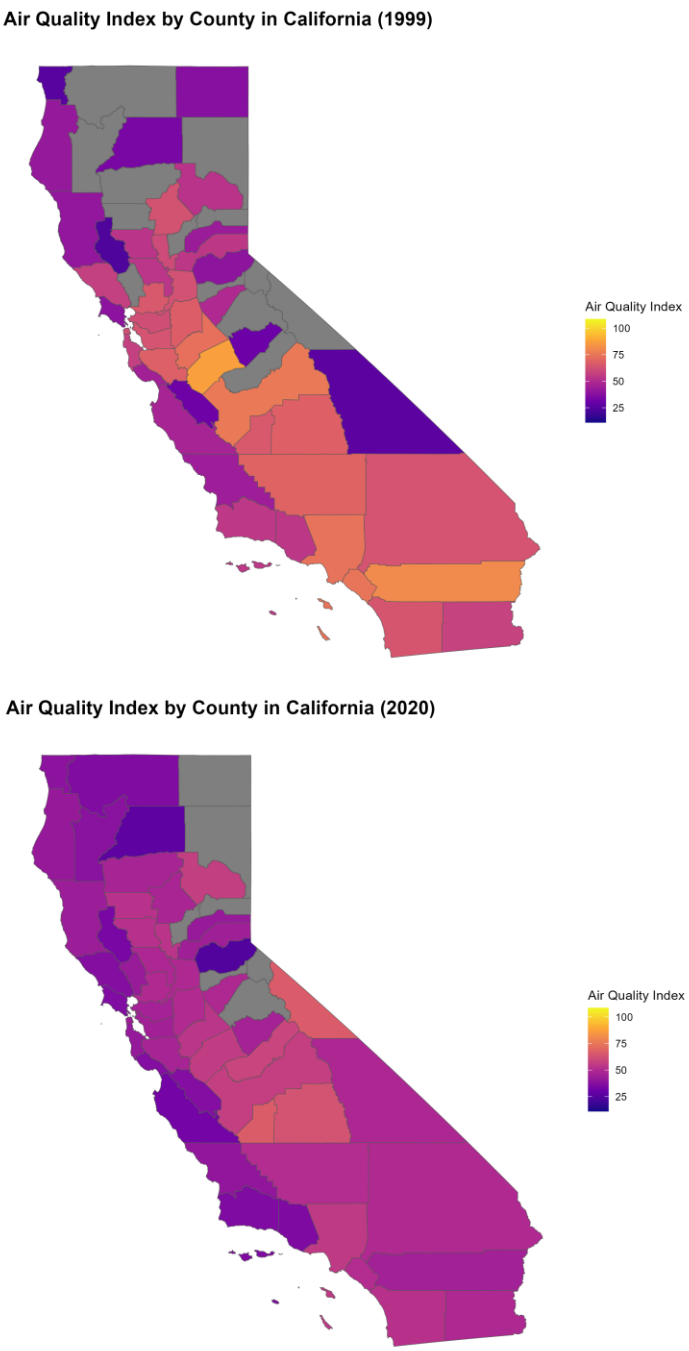


Figure 4. Comparison of 1999 and 2020 heatmaps for Air Quality Index by County in California. Similar heatmaps are created for ILI, stroke, and lung cancer death rates and converted to animated GIF format.

Across all heatmaps, counties along Northern California develop usable data as more monitoring stations become available. This is made apparent through the difference in number of "gray counties" with no data between 1999 and 2020 maps. The geospatial data for AQI visualizes an improvement in air quality over time across California. Similarly, heatmaps for ILI, stroke, and lung cancer death rates "cooled" over time, meaning air pollution and related health impacts experienced a reduction across California from 1999-2020.

Results and Discussion

The results from this study are subject to inconsistencies due to many different factors from our data analysis. These discrepancies manifest mainly through data availability. When a county is too low in population, suppression of measured population parameters is a major privacy concern; Suppression thresholds are established as a ratio, meaning greater data values fall short when the population is lower. Suppressed data points were treated as non-applicable data, meaning data on the number of deaths for the observed health conditions are underestimated. This carries over to the correlation and geospatial analysis, indicating the relationships between AQI and pollution-related health risks are likely stronger than they appear in the visualizations.

When brought to a countywide scale, Kern County showed significant positive correlation with air pollution and related health effects. This supports our expected outcome that greater exposure to poor air quality translates to sizeable health impacts. Compared to the other selected counties, Kern experienced significantly higher pollutant concentrations. The other three counties were similarly grouped in AQI magnitude and maintained relatively lower levels, meaning wealthier counties experienced better air quality.

Socio-economic disparities are further influenced by limitations in gathered data as demonstrated by examination of death rates by county. While Santa Clara consistently experienced lower death rates for all observed health conditions, San Mateo and Contra Costa were comparable to Kern. Notably, yearly changes in ILI death rate, Lung Cancer death rate, and annual average AQI had greater variances for Kern County. This is exacerbated by the data having not undergone bias analysis. Overall, Santa Clara's low death rates coupled with Kern's high variance and pollution exposure hints at the existence of socio-economic disparities, but consideration of more factors is demanded.

Conclusion

Exposure to air pollution is a significant component to the onset of public health risk. Over two decades of study in Kern County and the state of California, a gradual improvement of air quality is generally associated with lower death rates for lung cancer, stroke, and influenza-like illness. The considerably higher concentration of air pollutants in Kern compared to wealthier counties provides unique insight for future studies. Further research should demand consideration of other metrics for health impacts, such as incidence and hospitalizations data. California policymakers should consider monitoring stations on a county-by-county level to provide accurate data on concentrations of individual pollutants such as $PM_{2.5}$, O_3, and NO_2. Additionally, research on the concentrations of these pollutants over time across each county should be used to provide more information on the sourcing of emissions. Should these changes be implemented into future study, it would be easier to mitigate public health burden by targeting sources of anthropogenic emissions.

References

Brauer, M. (2015). Air pollution, stroke, and anxiety. *BMJ*, h1510. https://doi.org/10.1136/bmj.h1510

Centers for Disease Control and Prevention**.** (n.d.). *CDC WONDER: Data request for D76* [Data set]. U.S. Department of Health & Human Services. Retrieved July 15, 2024, from https://wonder.cdc.gov/controller/datarequest/D76

Kusumawardani, I. A. J. D., Indraswari, P. G., & Komalasari, N. L. G. Y. (2023). Air pollution and lung cancer. *Jurnal Respirasi*, *9*(2), 150–158. https://doi.org/10.20473/jr.v9-I.2.2023.150-158

Nawaz, M. O., Henze, D. K., Anenberg, S. C., Ahn, D. Y., Goldberg, D. L., Tessum, C. W., & Chafe, Z. A. (2023). Sources of air pollution-related health impacts and benefits of radially applied transportation policies in 14 US cities. *Frontiers in Sustainable Cities*, *5*, 1102493. https://doi.org/10.3389/frsc.2023.1102493

Shah, A. S. V., Lee, K. K., McAllister, D. A., Hunter, A., Nair, H., Whiteley, W., Langrish, J. P., Newby, D. E., & Mills, N. L. (2015). Short term exposure to air pollution and stroke: Systematic review and meta-analysis. *BMJ (Clinical Research Ed.)*, *350*, h1295. https://doi.org/10.1136/bmj.h1295

U.S. Census Bureau**.** (n.d.). *Data tables*. U.S. Department of Commerce. Retrieved August 4, 2024, from https://www.census.gov/data/tables.html

U.S. Environmental Protection Agency**.** (n.d.). *Air data: AQI plot*. Retrieved July 15, 2024, from https://www.epa.gov/outdoor-air-quality-data/air-data-aqi-plot

Zhang, R., Li, Y., Bi, P., Wu, S., Peng, Z., Meng, Y., Wang, Y., Wang, S., Huang, Y., Liang, J., & Wu, J. (2023). Seasonal associations between air pollutants and influenza in 10 cities of southern China. *International Journal of Hygiene and Environmental Health*, *252*, 114200. https://doi.org/10.1016/j.ijheh.2023.114200