Final Research Paper

Isabel Berkeley, Justine Fretz, XinYuan Liang, & Yi Luo

May 2025

1 Motivation

Understanding environmental determinants of health is critical for identifying spatial and structural disparities in health outcomes across the United States. Air quality and weather conditions are two of the most well-documented environmental risk factors affecting cardiovascular and respiratory health, especially among vulnerable populations. However, these factors are often analyzed separately, and existing public health visualizations rarely offer high-resolution, county-level views over extended time periods. Our project aims to bridge this gap by building a harmonized, longitudinal dataset that merges air quality, weather, and health outcome data across all U.S. counties from 2000 to 2023. By doing so, we provide policymakers and researchers with tools to visualize and compare regional environmental burdens, assess long-term exposure trends, and link these patterns to chronic disease outcomes such as heart disease. Our final goal was to create accurate, interpretable spatial maps that summarize long-term exposure to key environmental stressors—namely PM2.5, ozone, temperature, humidity, and pressure—while accounting for gaps in raw data and maintaining consistency across counties and years.

2 Data Description

2.1 Air Quality Data

The air quality data used in this analysis was collected from the U.S. Environmental Protection Agency (EPA) using direct API integration and bulk downloads from the EPA AirData website. Specifically, we accessed daily monitoring data for particle matter 2.5 (PM2.5) and ozone (O_3) spanning from 2000 to 2023. To automate the data retrieval process, we wrote a script that looped through each year and pollutant, downloading ZIP files containing CSV records for all U.S. counties. These files were then programmatically unzipped and organized into a structured dataset. The raw files included daily concentration readings by monitoring site, which we later aggregated at the county level. This process required careful data cleaning, including standardizing FIPS codes, handling missing values, and ensuring consistent formatting across years. The result was

a harmonized dataset that allowed us to calculate long-term pollutant averages and build a comprehensive air quality index for each county.

2.2 Weather Data

The weather data used in this analysis was collected from the National Oceanic and Atmospheric Administration (NOAA). Much like the Air Quality dataset we used direct API integration and bulk downloads to compile the full dataset across the years 2000-2023. We used a very similar automated retrieval process to download and build the dataset. To remain consistent, we used the same data cleaning standards to ensure the merging process would be smooth. This helps to avoid issues when it comes to the modeling steps and ensuring we have less missing data.

2.3 Heart Disease/Demographic Data

To increase the effectiveness, precision, and professionalism of the county-level mortality mapping workflow using CDC WONDER data, a number of significant adjustments and enhancements were performed during this process. An important improvement was automating the procedure using a 'for' loop to run over all unique years in the dataset. This allowed for the automatic creation of maps from 2013 to 2023, whereas at first the study was restricted to a single year (2023). The map's subtitle was also changed to include the complete official source citation: "CDC WONDER (NCHS, US DHHS, CDC, CMF)," in favor of the generic "mortality_data.csv" reference. This increases transparency and acknowledges the government agencies that provided the data. Additionally, tryCatch() was used to handle error scenarios when spatial data downloads fail, particularly for years beyond the tigris::counties() function's supported range (usually 2013–2023). This ensures that the script runs continuously. To keep a steady, close-up view of the 48 continental states, a filter was also used to omit Alaska, Hawaii, and US territories. In order to produce distinct PNG files for every year, file naming was optimized, which enhanced output reproducibility and organization. Technical consistency was also considered, including fixing missing or suppressed data by setting na.rm = TRUE and visually coding NA regions as light gray, converting countycode to a character format to match GEOID in the shapefiles, and setting static map boundaries for uniform framing across maps. There are still a number of optimization options in spite of these advancements. Spatial shapefiles from a single year (2020, for example) can be repeated across all maps to minimize runtime and download dependencies. Using dynamic legends or superimposing state borders can improve visual clarity. For increased interactivity, technologies like leaflet or tmap might be used to let users explore data interactively. For performance, parallel processing could be utilized to create maps more quickly.

Spatial Trends: Death rates generally appear higher in more rural, non-metropolitan areas, particularly in the southeastern U.S., Midwest, and some

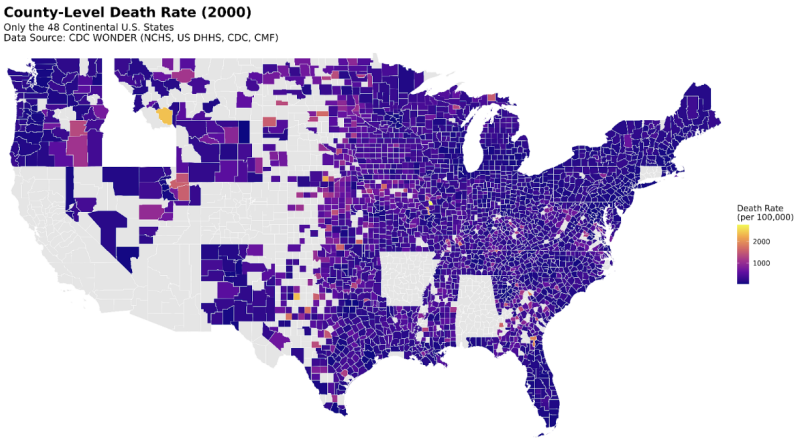

Figure 1: County Level Death Rate 2000

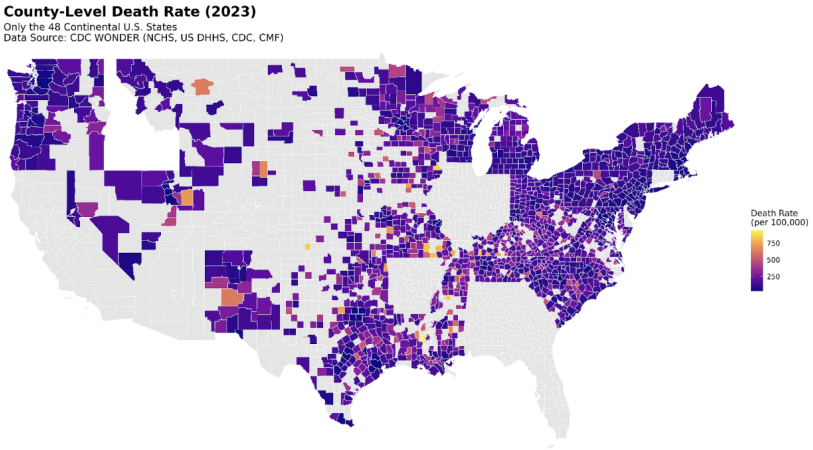

Figure 2: County Level Death Rate 2023

parts of the western states. Metropolitan and suburban counties exhibit relatively lower death rates consistently across the years. Temporal Trends: There is a visible intensification of death rates in rural counties from 2000 to 2023, reflecting potential issues like health care access, socioeconomic conditions, and aging populations. Regions like Appalachia and the Mississippi Delta exhibit persistently high mortality rates, indicating chronic public health and economic disparities.

3 Methodology

In this data processing workflow, we initially imported a dataset covering U.S. mortality from 2000 to 2023, specifically cleaning by removing entries lacking essential population data and excluding age groups younger than 20 to maintain analytical consistency. Each variable, including demographic details (county, year, age group, sex, and race) and mortality statistics, was systematically labeled to facilitate clear interpretation. We then calculated the total number of deaths and total population by county and year, providing aggregate metrics crucial for county-level analysis. Following this, we computed population proportions for distinct age-sex-race subgroups within each county-year, enabling a demographic-standardized evaluation of mortality. Next, age-sex-race-specific death rates (ASDR) per 100,000 population were derived, highlighting subgroup-specific mortality risks clearly. By weighting these ASDR values according to their population proportions, we generated standardized death rates reflective of each county's unique demographic structure. Finally, these standardized county-level death rates were aggregated, saved, and exported into multiple formats (Stata, Excel, and CSV) to ensure accessibility and usability for subsequent analysis and reporting.

4 Results

4.1 Descriptive Statistics

Table 1: Summary Statistics of Death Rate by Urbanization Level

	Min	Max	Mean	SD
Large Central Metro	333.93	1674.55	718.89	226.00
Large Fringe Metro	328.34	24137.93	1131.21	735.41
Medium Metro	392.38	24193.55	1250.45	699.33
Micropolitan(Nonmetro)	285.88	27777.78	1951.02	921.40
NonCore(Nonmetro)	726.81	40000.00	3063.56	1906.81
Small Metro	581.06	28205.13	1534.43	937.29
Total	285.88	40000.00	1662.76	1277.08

The data reveals a striking urban-rural divide in standardized death rates

across U.S. counties, with mortality rates increasing consistently as areas become more rural. The large Central Metro areas exhibit the lowest average standardized mortality rate (718.89), while NonCore (Nonmetro) areas show the highest (3063.56), indicating that residents in the most rural counties face more than four times the mortality risk of those in dense urban centers. Furthermore, the standard deviation increases with rurality, reflecting greater variability and instability in health outcomes in less urbanized regions. Interestingly, even Small Metro and Micropolitan areas show elevated death rates, suggesting that rural health disadvantage extends beyond the most remote locations.

Table 2: Summary Statistics of Death Rate by Age Group

	Min	Max	Mean	SD
20-24 years	333.93	848.86	533.12	116.39
25-34 years	333.93	1318.65	659.27	183.07
35-44 years	328.34	2497.74	810.40	243.61
45-54 years	328.34	4203.15	1052.35	389.59
55-64 years	285.88	5925.93	1327.05	600.77
65-74 years	328.34	14925.37	1580.89	861.95
75-84 years	328.34	18181.82	1852.24	1267.67
85+ years	328.34	40000.00	2332.40	2105.55
Total	285.88	40000.00	1662.76	1277.08

The age-based mortality statistics show that the mortality rate steadily increases with age. The average standardized mortality rate for young adults aged 20-24 is the lowest, at 533.12, while the average standardized mortality rate for those aged 85 and over is the highest, at 2332.4, which is more than four times higher. The mortality rate increases with age, which is consistent with the general physiological decline among the elderly. In addition, the standard deviation (SD) also increases significantly with age, from 116.39 in the age group of 20-24 to 2105.55 in the age group of 85 and older. This indicates that not only are the mortality rates higher among the elderly, but also the differences and unpredictability of the mortality rate results are greater.

Table 3: Summary Statistics of Death Rate by Sex

	Min	Max	Mean	SD
Female	328.34	27906.98	1670.03	1387.84
Male	285.88	40000.00	1656.41	1171.74
Total	285.88	40000.00	1662.76	1277.08

The sex-based mortality statistics show that both males and females experience similar average standardized death rates, with females having a slightly higher mean (1670.03) than males (1656.42). However, a closer look reveals that the distribution of mortality is more varied among females, as indicated by a higher standard deviation (1387.84 compared to 1171.74).

Table 4: Summary Statistics of Death Rate by Race

	Min	Max	Mean	SD
Black or African American	333.93	25000.00	1165.42	813.51
White	285.88	40000.00	1772.54	1325.43
Other Races	333.93	6451.61	718.59	320.90
Total	285.88	40000.00	1662.76	1277.08

In terms of race-based mortality, the data reveals notable disparities across different racial groups. White individuals have the highest average standardized mortality rate at 1772.54, along with the highest standard deviation (1325.43). Black or African-American individuals, on the other hand, show a lower average standardized mortality rate (1165.42) but still face considerable variation (SD = 813.51). Other races have the lowest mean standardized mortality rate at 718.59 and the least variation (SD = 320.90), suggesting more consistent results.

4.2 Maps

To generate the final maps of county-level air quality and weather across the contiguous United States, we compiled and processed EPA monitoring data from 2000 to 2023. The primary goal was to create a normalized, interpretable index of air quality and weather conditions at the county level, using the most consistently reported and environmentally relevant pollutants and weather metrics available. After evaluating the data completeness and consistency across pollutants, we selected ground-level ozone (O_3) and fine particulate matter (PM2.5) as our two core indicators. Both variables are critical components of the Air Quality Index (AQI) used by the EPA and are strongly linked to respiratory and cardiovascular health risks. In contrast, we chose to exclude lead from the final map due to substantial missingness, only a limited number of counties had reliable lead measurements over the 23-year period, which would have skewed spatial representation and reduced overall data coverage.

For weather we decided that the most relevant indicators that are most closely associated with cardiovascular health were temperature, relative humidity and atmospheric pressure as extreme heat/cold as well as humidity are among the most burdensome weather conditions in terms of cardiovascular systems. We calculated county-level averages of O_3 and PM2.5, rescaled them to a [0, 1] range, and combined them into a composite air quality index using the row-wise mean. We used the same method to create a weather quality index using temperature, atmospheric pressure, and relative humidity. This approach ensured that a county could still be included in the index if it had data for at least one variable, maximizing coverage while maintaining interpretability. Below are the final visualization maps for these indexes across counties, using a perceptually uniform color scale, and excludes Alaska and Hawaii for spatial clarity. Gray regions represent counties with no valid data for and variables within the index.

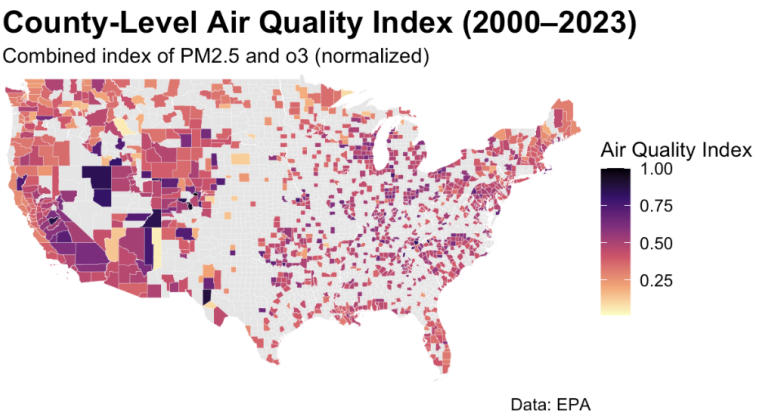

Figure 3: County Level Air Quality Index

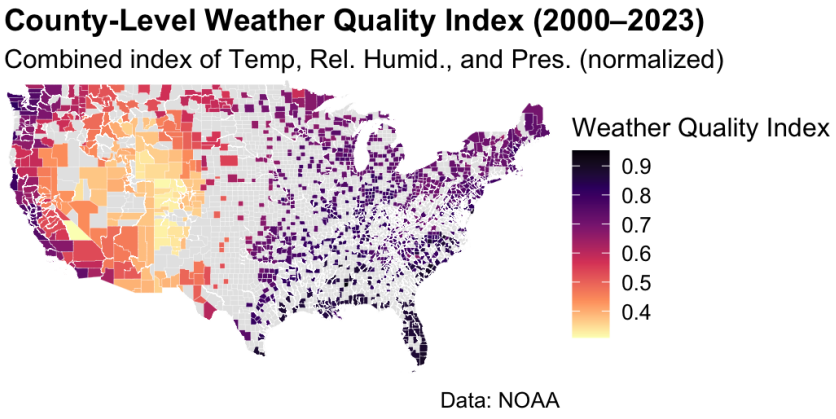

Figure 4: County Level Weather Quality Index

4.3 Regression Results

The regression study identifies key factors that influence mortality in U.S. counties between 2000 and 2023, with a focus on meteorological, demographic, and air pollution factors. While the effects of coarse particulate matter (PM10) are inconsistent and need more research, fine particulate matter (PM2.5) is consistently found to be a significant factor related with higher mortality, highlighting its detrimental impact on respiratory and cardiovascular health. Significantly higher death rates are correlated with cooler temperatures, indicating the significance of healthcare treatments and support during colder months. Given their demographics, older age groups are at far higher risk of mortality, which emphasizes the necessity of aging population-focused preventive measures and specialized treatment. Furthermore, the findings show that male mortality is higher, which calls for targeted health education and preventative care initiatives. The observed racial disparities, which include differences in death rates among various racial groups, underscore the intricacy of socioeconomic and healthcare access concerns.

Table 5: Determinants of Mortality: The Influence of Air Pollution, Weather, and Demographics Across U.S. Counties (2000–2023)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
PM2.5	27.08***	-38.36***	-38.36	-36.89***	-34.50***	-32.55***	-32.55
	(41.88)	(-40.90)	(-2.20)	(-39.69)	(-37.04)	(-35.32)	(-2.01)
PM10	-6.058***	-1.293***	-1.293	-1.030***	-1.703***	-1.449***	-1.449
	(-21.78)	(-4.49)	(-0.40)	(-3.61)	(-5.97)	(-5.14)	(-0.49)
Temperature	-35.38***	-29.45***	-29.45*	-28.34***	-26.55***	-25.06***	-25.06*
	(-46.15)	(-37.79)	(-2.66)	(-36.71)	(-34.32)	(-32.76)	(-2.68)
25–34 years				23.84		66.27	66.27*
				(0.66)		(1.85)	(2.70)
35–44 years				58.40		112.7**	112.7
				(1.66)		(3.25)	(2.29)
45–54 years				122.3***		184.5***	184.5*
				(3.50)		(5.33)	(2.61)
55–64 years				166.0***		234.1***	234.1*
				(4.75)		(6.77)	(2.87)
65–74 years				197.3***		268.5***	268.5*
				(5.64)		(7.77)	(2.99)
75–84 years				229.8***		301.3***	301.3*
				(6.58)		(8.72)	(2.98)
85+ years				267.1***		334.2***	334.2*
				(7.63)		(9.65)	(2.88)
Male					6.753**	13.08***	13.08
					(2.75)	(5.38)	(2.40)
White					111.9***	116.6***	116.6*
					(37.83)	(39.85)	(2.79)
Other Races					-83.58***	-100.4***	-100.4*
					(-14.73)	(-17.88)	(-2.63)
Constant	1623.2***	2796.1***	2796.1**	2550.8***	2611.3***	2277.3***	2277.3**
	(88.09)	(80.64)	(5.04)	(51.63)	(75.56)	(46.37)	(4.69)
R-squared	0.228	0.346	0.346	0.359	0.360	0.375	0.375
Year × State Interaction	No	Yes	Yes	Yes	Yes	Yes	Yes
Clustered Urbanization	No	No	Yes	No	No	No	Yes

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

5 Discussion & Limitations

As demonstrated in our results section, we have found significant geographic and demographic trends in the mortality rates across U.S. counties from 2000 to 2023. We found notable increases in risks for people of older populations and

various racial and ethnic groups. Additionally, we found a consistent association between elevated PM2.5 levels and increased mortality. These findings carry substantial economic implications especially when considering the rising costs associated with treating cardiovascular diseases that are exacerbated by such environmental factors.

5.1 Economic Significance

As we found, poor air quality can lead to an increased risk for cardiovascular illness which drives up health care expenditures. The American Heart Association estimates that the annual inflation adjusted health costs of cardiovascular conditions in the U.S. are expected to quadruple from $393 billion in 2020 to 1.49 trillion in 2050. Due to worsening weather and air quality factors such as PM2.5 being linked to geography, certain areas may experience higher costs of treating cardiovascular disease as a result of prolonged exposure to poor conditions. This can also cause economic strain on public health systems and private insurance due to a higher prevalence of these illnesses. It is also important to consider the link between demographic disparity and economic vulnerability. As shown, racial and ethnic minorities are at a higher risk for cardiovascular disease. In some cases, these groups may also be more likely to experience economic inequality widening the gap and placing a large burden on these vulnerable groups.

Worsening weather and air quality conditions can also have a large impact on the tourism industry. Our regression analysis indicates that there is a significant relationship between cooler temperatures and higher mortality. This has economic implications not only for increased winter healthcare demands but also for winter tourism, thus requiring adequate infrastructure and healthcare preparedness. Extreme weather conditions such as heat or cold or high humidity may deter people especially in vulnerable populations from visiting places geared towards outdoor recreation. While cities may experience worse air quality conditions due to industrial air pollution which can also have a negative impact on susceptible travelers. These considerations can negatively impact revenue, job markets, and tax revenue in areas with unfavorable conditions.

5.2 Policy Implications

The Clean Air Act sets standards for air quality to project public health. Stricter emission standards for industries and vehicles, increased monitoring of PM2.5 levels and policies to promote cleaner energy alternatives are crucial. Given the observed relationship between PM2.5 and higher mortality rates, it is worth advocating for stricter emission standards for key sources of this pollutant among others. These standards can be implemented across numerous industries including Industrial, Automotive, Agriculture, and Construction. We would also suggest that more resources be allocated to increasing the number of monitoring stations to improve the network and better monitor rural areas. Through advancements in technology, lower cost sensors and more comprehensive satellite

data can provide real time air-quality information allowing policy makers to act quicker and constantly update models to motivate new policy.

There are also areas where our results can be leveraged to improve public health policy, as our research underscores the direct link between air and weather conditions and health outcomes. These policies could include more timely public health advisories to notify especially vulnerable populations of periods of poor air quality. Additionally, there could be more awareness surrounding environmental and health risks of air pollution and educate people on mitigation practices and protective measures. A successful policy to address concerns of air pollution connected to cardiovascular disease would require intervention from various government agencies from the Department of Motor Vehicles to healthcare agencies. The worsening air quality conditions also highlights the need for more sustainable practices when it comes to renewable energy and sustainable land use.

5.3 Limitations

While this study offers a robust analysis of environmental health trends across U.S. counties from 2000 to 2023, several limitations should be acknowledged. First, although we compiled data from authoritative sources such as the EPA, NOAA, and CDC WONDER, these datasets still suffer from inconsistent spatial and temporal coverage. For example, lead data from the EPA was excluded due to substantial missingness, with reliable measurements reported in only a small fraction of counties. Even among the selected pollutants (PM2.5 and ozone), certain rural or low-population areas had sparse or irregular monitoring, which could bias spatial patterns despite our use of imputation and index-based smoothing.

Additionally, weather and pollution variables were averaged over long time periods, which may obscure seasonal trends, short-term spikes, or event-driven anomalies (e.g., wildfires, heatwaves). The air and weather quality indexes rely on normalized values, which improve comparability but may downplay absolute exposure levels in high-risk areas. Our mapping pipeline also excluded Alaska, Hawaii, and U.S. territories for consistency in spatial scale, which limits the national generalizability of our findings. Lastly, while we identified strong associations between environmental conditions and mortality, our analysis is observational and cannot establish causal relationships. Unmeasured confounding factors—such as access to healthcare, comorbidities, or local economic conditions, may also influence the outcomes. Future work could enhance this framework by incorporating more granular exposure metrics, exploring temporal variation, and integrating health outcome models that explicitly account for time lags and regional policy differences.

5.4 Conclusion

This project brings together long-term air quality, weather, and health outcome data to produce a comprehensive, county-level view of environmental conditions

across the contiguous United States from 2000 to 2023. By harmonizing data from multiple federal sources and constructing interpretable index measures for pollution and weather stressors, we offer a scalable framework for visualizing and analyzing spatial disparities in environmental health risks. Our findings highlight consistent regional patterns, particularly in rural and southeastern counties, where exposure to environmental stressors and mortality rates remain disproportionately high. These maps and indexes provide a foundation for future modeling, policy evaluation, and public health decision-making aimed at reducing place-based environmental health inequities.