Understanding the correlation between various pollutants and Cancer across geographical clusters in the U.S.

Atreya Naik and Morteza Sarmadi

SUMMARY

In the United States, chemical, water, and air pollution all severely undermine marine and human life. Inhaling these pollutants may affect vital organs, including the heart, brain, and liver because of their harmful properties. Many prior studies have shown that pollutants can increase the risk of life-threatening disease in humans. This study, performed at a state/regional granularity, explored whether different types of pollutants increase Cancer incidence rate. The endeavor of this project was to determine the strength of the correlation between the prevalence of certain types of Cancer and environmental factors across specific regions in the United States using the Pearson Correlation Metric as a mathematical tool across three years (2018-2020). 36 different types of Cancer (e.g. Cervix, Hodgkin Lymphoma, Leukemia) and 12 pollutant metrics (e.g. Ozone, Lead, Total toxic pounds in Majors) were analyzed using publicly available data, cleaned and processed through a python script. Our research has shown that certain clusters in the United States showcase a strong correlation between different environmental pollutants and certain types of Cancers. For example, in the Southwest region, there is a strong correlation between Days PM2.5 (air pollutant) and Hodgkin Lymphoma (Cancer type). These findings can potentially help prevent premature deaths caused by Cancer and help explain why certain clusters in the United States have abnormally high Cancer-incidence rates compared to other clusters.

INTRODUCTION

Cancer is the second leading cause of death in the United States, exceeded only by heart disease (1). From 2016-2020, 8,491,642 new cases of Cancer were reported, out of which 2,998,331 people died of Cancer (2). Additionally, the Cancer incidence rates for adolescents/adults aged 15-39 increased an average of 0.9% each year between 2014 and 2018 (3). Premature deaths from Cancer are around 2% across all premature age groups (4).

Pollution is contamination of the indoor or outdoor environment by any chemical, physical, or biological agent that modifies the natural characteristics of the atmosphere (5). According to the American Lung Association, more than 119 million residents in the United States live in places with unhealthy amounts of pollution (6). Pollutant levels are continuing to rise to uncharted levels, highlighting that pollution is a key environmental issue and is a huge detriment to human, animal and marine life (**Figure 1, 7)**.

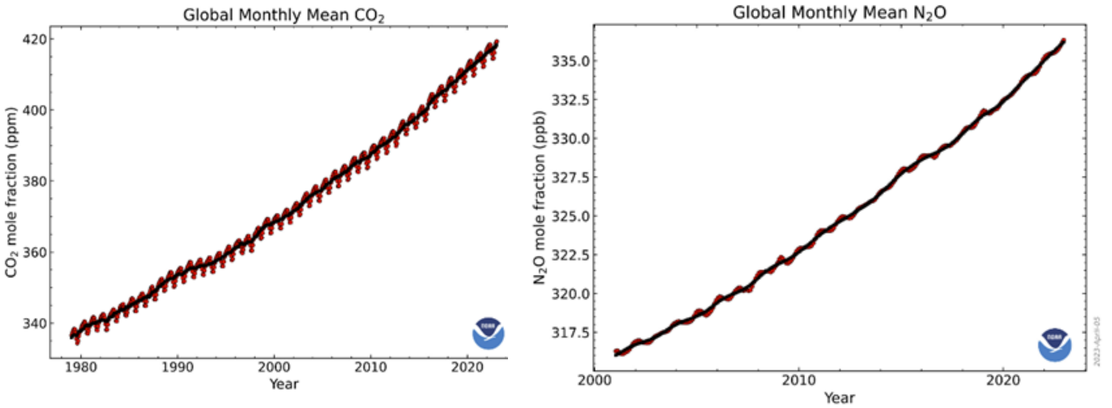

Figure 1. Increase in Carbon dioxide (left), and nitrous oxide (right) within the last 2 decades. Scatterplot showing the ppb of Carbon Dioxide (left) and Nitrous Oxide (right) in the atmosphere over the last 2-3 decades. Carbon Dioxide rose from 340 ppb to 420 ppb over 3 decades, while Nitrous Oxide rose from 317.5 ppb to 335 ppb over the last two decades **(7).**

Previous studies highlight a correlation between Cancer and geographical location. For example, residents living in Kentucky are 2.4 times as likely to be affected by Lung Cancer compared to Utah (8). The above correlation mentioned in literature, found a strong correlation between colorectal cancer and nitrate levels (9). Another study found that air pollution was associated with an 8% increase in Lung Cancer mortality (10). Scientists use statistics and other predictive models to find relationships between healthcare and environmental sciences (11). This study aimed to determine the correlation between various pollutant metrics and Cancers across certain regions of the United States using the Pearson Correlation Coefficient (gives R value) as a mathematical tool. R ranges from -1 to 1, with any R-value between 0.5 and 0.7 considered to signify strong correlation, and values between 0.7 and 1 are considered very strong.

RESULTS

The aim of this study was to understand the correlation between various pollutants and various types of Cancer across different regions of the United States. Plotting the different environmental pollutant variables changes across the eight regions reveals that there is definite variability of these pollutants and Cancer count across these regions. For example, in **Figure 2**, the West coast region had the most days with PM2.5 and the least days with Ozone detected in 2019, while the mid-Atlantic region had more days with NO2. Similarly, the New England region had the highest count of Cancer per 10000 people (a.k.a. normalized count).

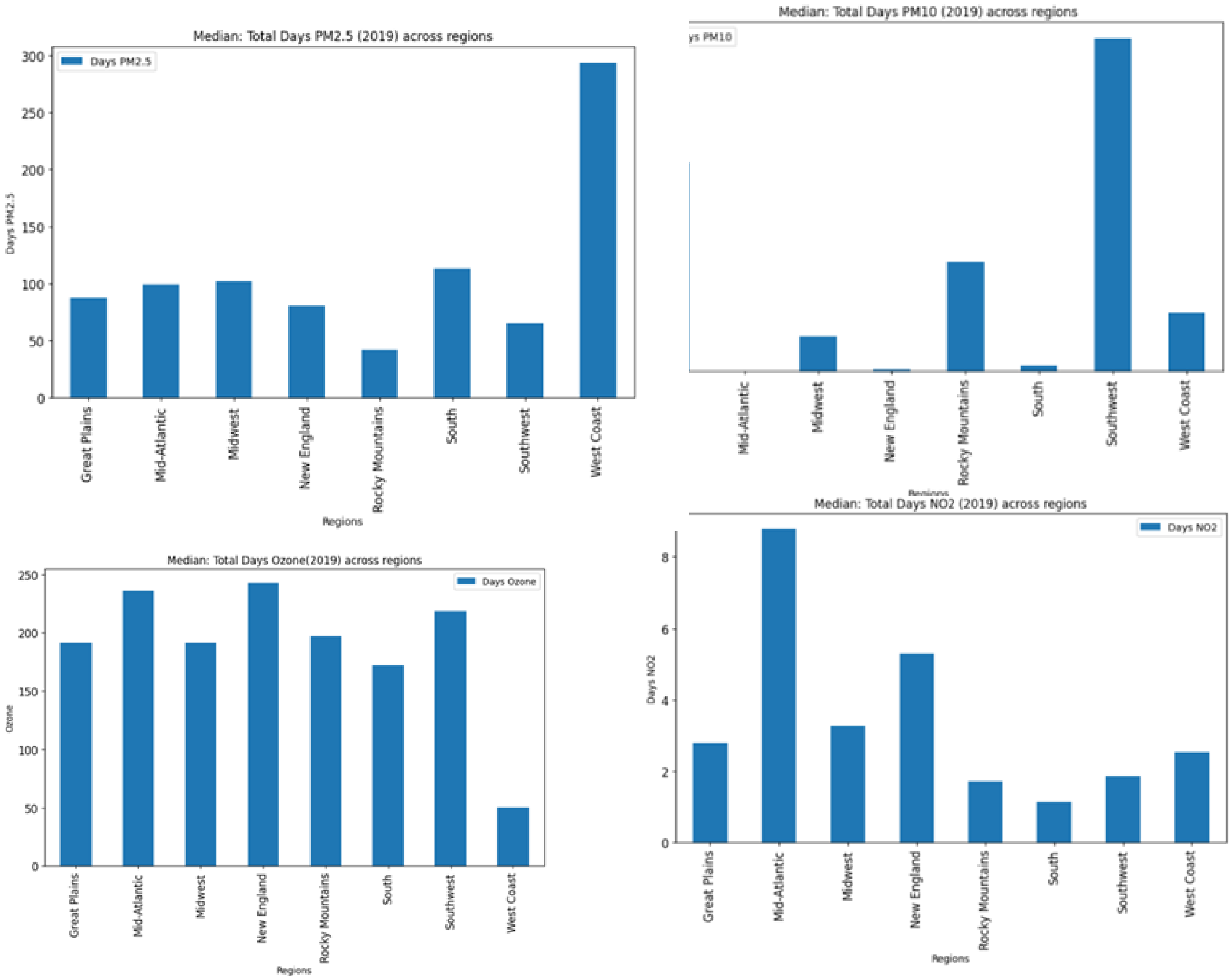

Figure 2: EDA: Region variability across different air-pollutants and Cancer. Bar graphs showing median cancer counts across regions/states in that region. Python script was developed and pre-processed to determine these values. (Determined via code - 22)

We analyzed 36 different types of Cancer, 12 pollutant metrics (**Table1**) across eight regions over 3 years (2018-2020). A custom python script was developed to clean, pre-process publicly available raw data and determine the correlation using the Pearson Correlation Metric as a mathematical tool. The findings revealed strong correlations (R > 0.7) between specific air, water, and chemical pollutants and various Cancers in certain regions, aligning with some prior research in the United States. For example, if you look at **Figure 3**, in the Midwest, between 2018 and 2020 we observed a strong correlation between air pollution and Melanomas of the Skin with the average R value = 0.88. In the Mid-Atlantic, we observe a strong correlation between chemical pollution and Colon & Rectum Cancer **(Figure 4)**. In the

Great Plains, we observed a strong correlation between water pollution and Kidney & Renal Pelvis Cancer **(Figure 5)**. See **Table 2** for more details on the correlation findings across the different regions.

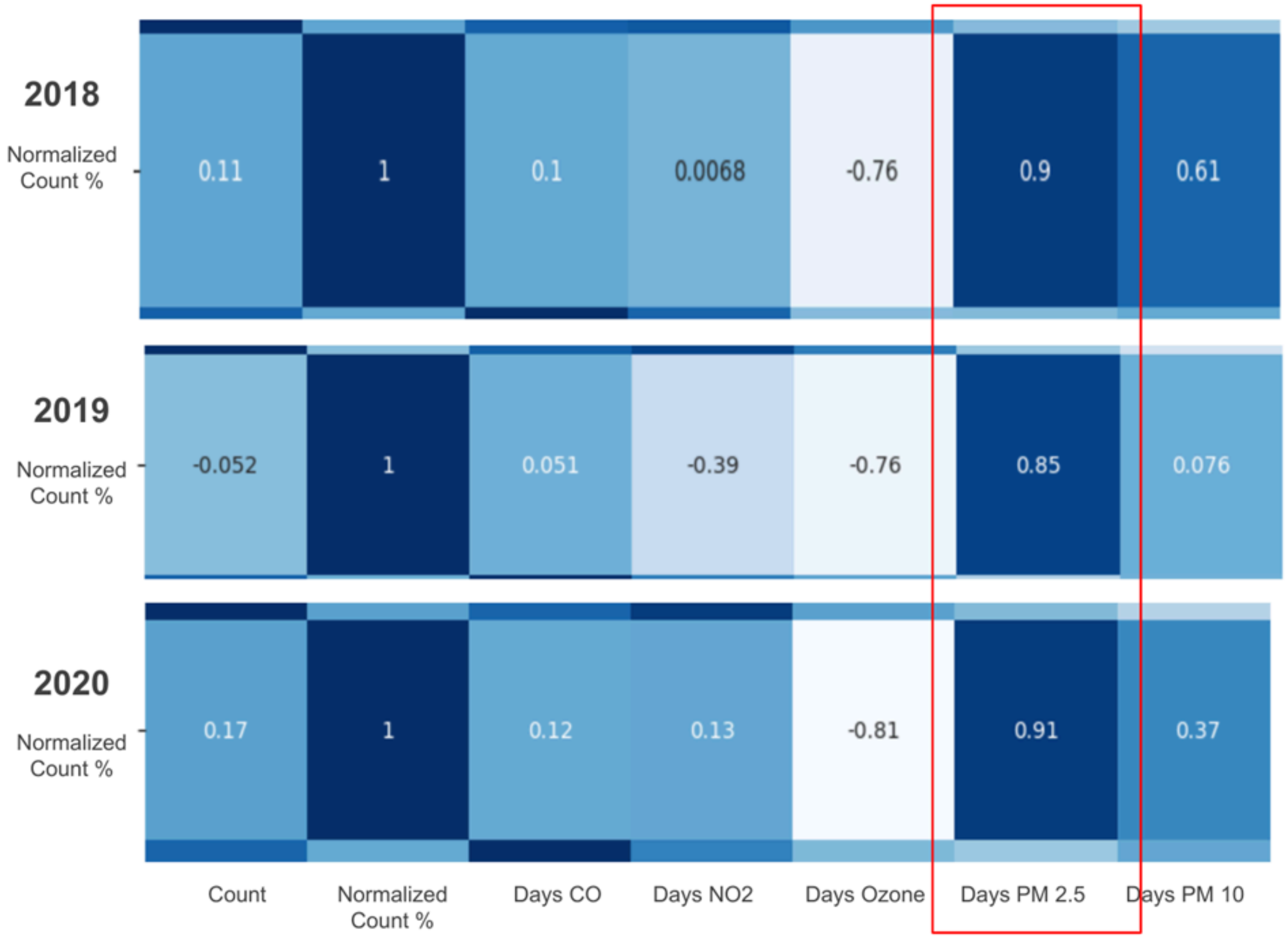

Figure 3: Strong correlation between Melanomas of Skin Cancer and air pollution (PM2.5). Correlation matrix between Melanomas of Skin Cancer and air pollution (PM2.5) in the Midwest across 2018-2020. A Python script was developed and pre-processed to determine such correlations. R=0.9 in 2018, R = 0.85 in 2019, R = 0.91 in 2020. (Determined by code - 22).

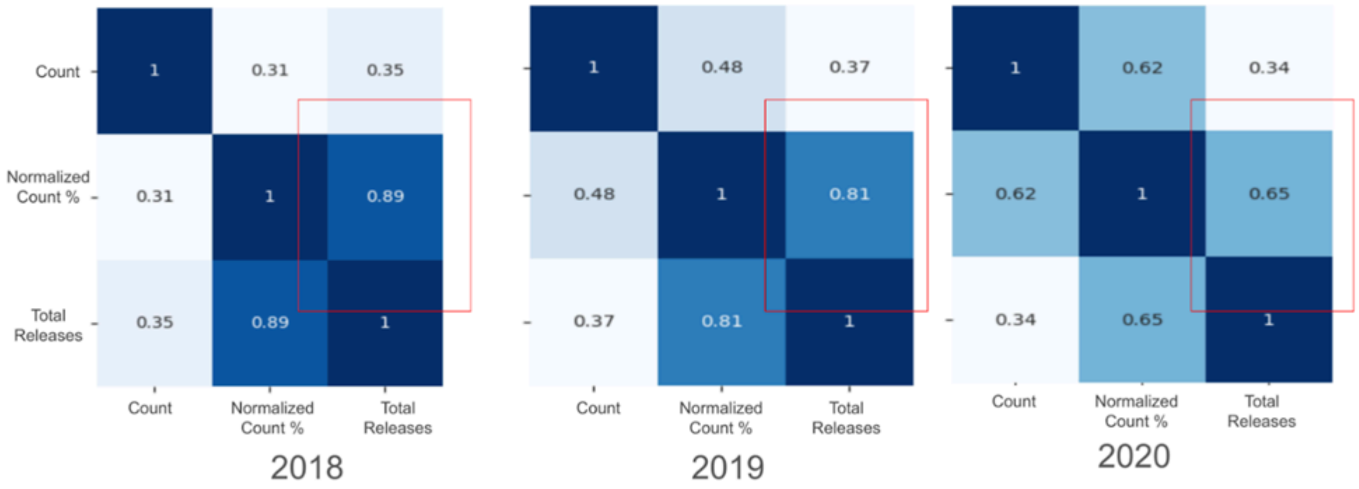

Figure 4: Strong correlation between Colon and Rectum Cancer and chemical pollution (Benzene). Correlation matrix between Kidney and Renal Pelvis Cancer and chemical pollution (Benzene) in the Mid-Atlantic across 2018-2020. A Python script was developed and pre-processed to determine such correlations. R=0.89 in 2018, R = 0.81 in 2019, R = 0.65 in 2020. 2020 (Determined by code - 22)

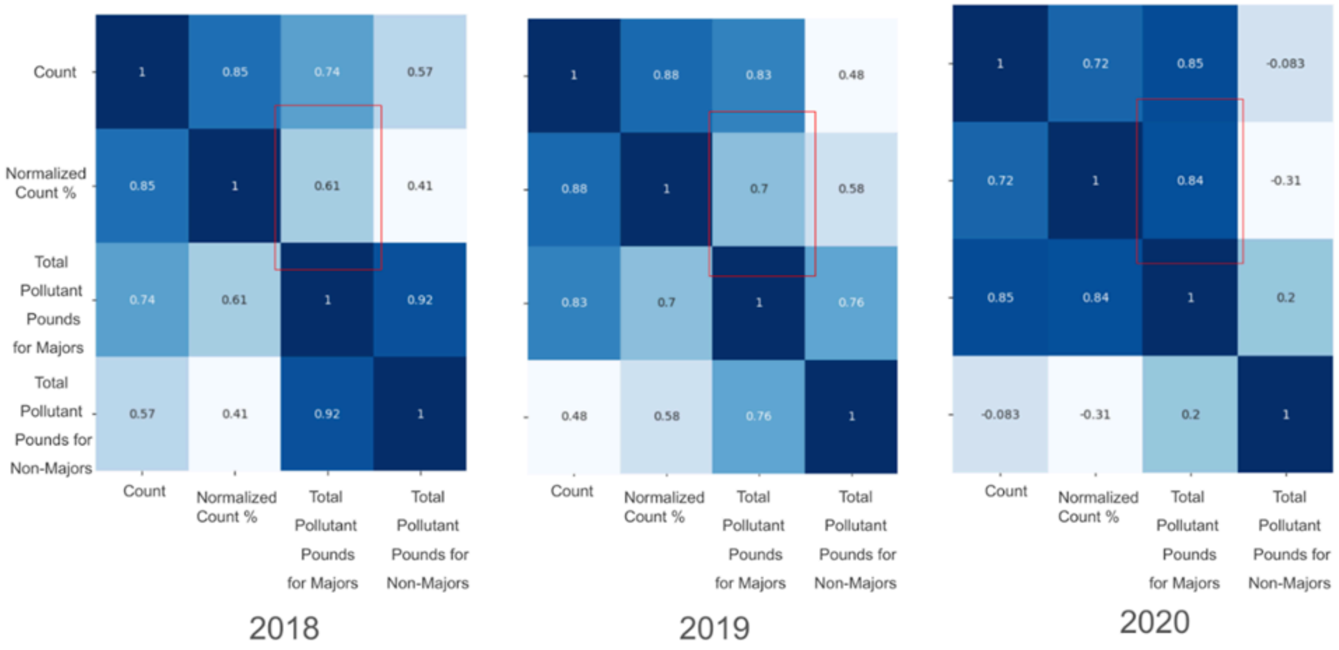

Figure 5: Strong correlation between Kidney and Renal Pelvis Cancer and total water pollutant pounds. Correlation matrix between Kidney and Renal Pelvis Cancer and total water pollutant pounds in

the Great Plains across 2018-2020. A Python script was developed and pre-processed to determine such correlations. R=0.41 in 2018, R = 0.7 in 2019, R = 0.84 in 2020. (Determined by code - 22)

Table1: Pollutants considered for analysis in this study

Pollutant	Factors considered
Air	Days CO, Days NO2, Days Ozone, Days PM2.5, Days PM10
Water	Total Pollutant Pounds (lb./yr) for Majors, Total Pollutant Pounds (lb./yr) for non-Majors
Chemical	Lead, Sulfuric Acid, Chromium compounds, Benzene, Polycyclic aromatic compounds

Table2: Correlation Cancer Type Vs Region Vs Pollutant type (Determined by code - 22)

Region	States	Types of Cancer	Cause	Pollutant Type	Avg Correlation (2018-2020)

Great Plains	Kansas, Nebraska, North Dakota, Oklahoma, South Dakota	Kidney and Renal Pelvis	Chemical	Benzene	0.9
		Kidney and Renal Pelvis	Water	Pollutant Pounds Maj	0.7
		Stomach	Chemical	Lead	0.7
Midwest	Illinois, Indiana, Iowa, Michigan, Minnesota, Missouri, Ohio, Wisconsin	Melanomas of the Skin	Air	PM2.5	0.9
Mid-Atlantic	Delaware, District of Columbia, Maryland, New Jersey, New York, Pennsylvania	Colon and Rectum	Chemical	Benzene	0.8
		Larynx	Chemical	Benzene	0.7
		Colon and Rectum	Chemical	Chromium compounds	0.7
		Colon and Rectum	Water	Pollutant Pounds Maj	0.6
		Brain and Other Nervous System	Chemical	Benzene	0.6
		Ovary	Water	Pollutant Pounds Maj	0.5
		Brain and Other Nervous System	Chemical	Lead	0.5

		Brain and Other Nervous System	Chemical	Sulphuric Acid	0.5
New England	Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, Vermont	Myeloma	Air	CO	0.9
		Kidney and Pelvis	Chemical	Sulphuric Acid	0.7
		Stomach	Chemical	Chromium Compounds	0.7
Rocky Mountains	Colorado, Idaho, Montana, Utah, Wyoming	Leukemias	Air	PM2.5	0.9
West Coast	California, Oregon, Washington	Cervix	Air	PM2.5	0.9
		Brain and Other Nervous System	Air	PM2.5	0.9
		Melanomas of the Skin	Air	PM2.5	0.9
Southwest	Arizona, Nevada, New Mexico, Texas	Esophagus	Air	PM10	0.9
		Hodgkin Lymphoma	Air	PM2.5	0.8
South	Alabama, Arkansas, Florida, Georgia, Kentucky, Louisiana, Mississippi,	Hodgkin Lymphoma	Air	Ozone	0.6

	North Carolina, South Carolina, Tennessee, Virginia, West Virginia	Leukemias	Air	Ozone	0.6

For the analysis performed across the three years, the correlations of Cancer and pollution factors stayed very consistent across the different regions of the United States. For example, the correlation between Days PM2.5 and Hodgkin Lymphoma were very strong across the years in the Southwest region 2018 (R = 0.8), 2019 (R=0.7), and 2020 (R=0.7), meaning that in the Southwest region, PM2.5 caused Hodgkin Lymphoma. Similarly, for water pollutant Total Pollutant Pounds (lb./yr) for Majors in the Great plains, there was a strong correlation to Kidney/Renal/Pelvis Cancer across the years 2018 (R=0.6), 2019 (R=0.7) and 2020 (R=0.8). There was strong correlation across the years between Brain/Nervous system Cancer and chemical pollutants like Sulphuric Acid, Lead, Benzene in the Mid-Atlantic, eg: for Benzene 2018 (R=0.6), 2019 (R=0.7), 2020 (R=0.6).

DISCUSSION

Certain clusters in the United States are more susceptible to harmful diseases such as Cancer. Looking at our results, we see that within these clusters there is a strong correlation between different environmental (water, air and chemical) pollutants and certain types of Cancers (eg: In the Southwest region, there is a strong correlation between the air pollutant PM2.5 and Hodgkin Lymphoma) (**Figure 6**).

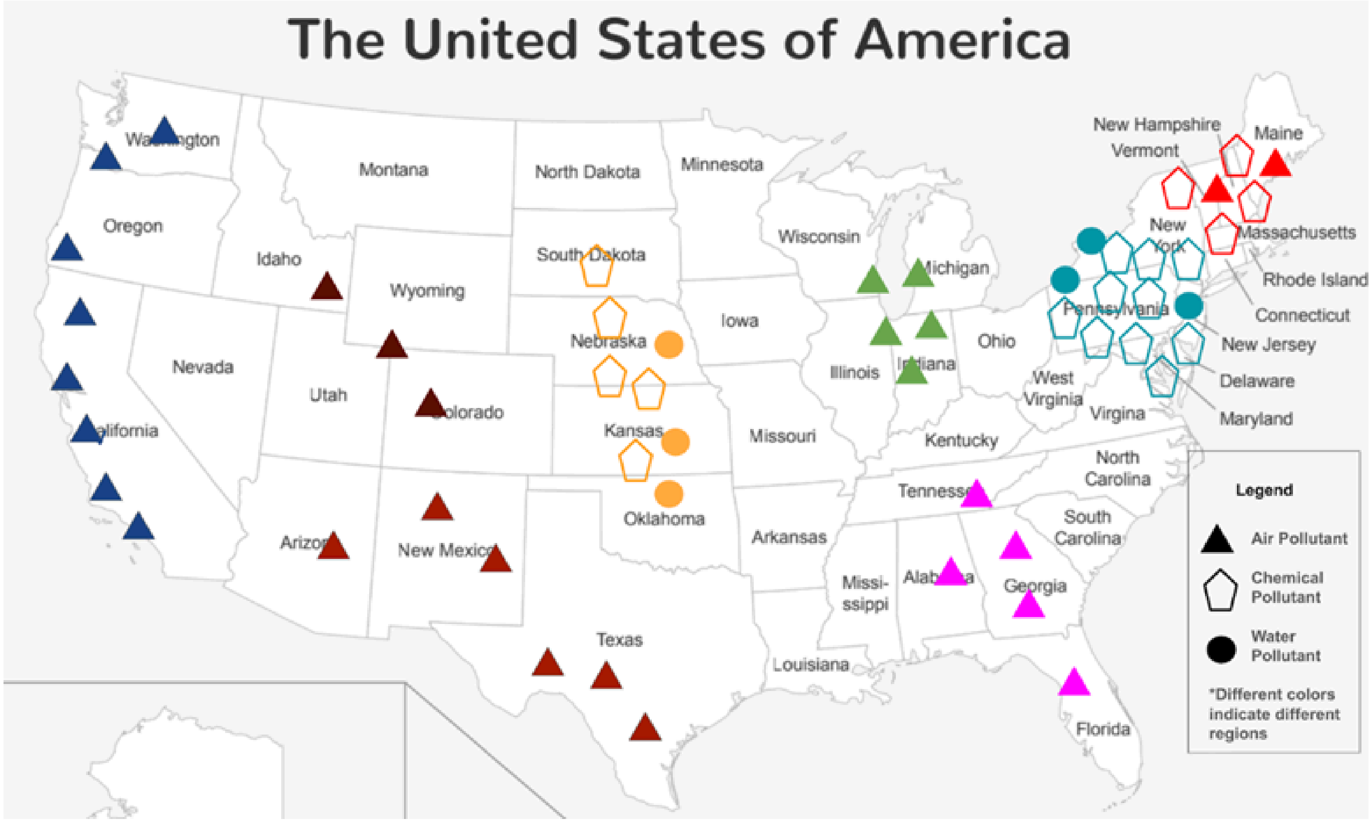

Figure 6: Strong cancer correlations with pollutants across the different regions. Map showing distribution of the number of environmental factors that had a strong correlation to various Cancers in that region. A Python Script was pre-processed to determine correlations in regions.

Although it is unclear why certain air, water, and chemical pollutants have extremely high correlations to specific cancer incidence rates in specific regions and have much lower correlations in other regions, many studies have proposed broader explanations towards why these pollutants have a positive correlation toward cancer incidence rate (12 - 17). One example related to water pollution: according to the National Cancer Institute, women who were exposed to higher nitrate levels, had increased risks of ovarian cancer (12). We also found this specific trend across the years 2018-2020 when observing the correlation between ovary cancer to total water pollutant pounds in certain regions. In general, when people drink water contaminated with nitrates or arsenic, it commonly leads to the endogenous formation of N-nitroso compounds, which are potent animal concentrations (13). One example related to chemical pollution: according to a 2021 study concluded that exposure to chemical pesticides increases the risk of brain cancer (14). We also found that chemical pollutants like Benzene, Lead and Sulfuric Acid had a strong correlation to Brain Cancer in the Mid-Atlantic across the years 2018-2020. In general, chemical pollution has a strong link to cancer because these chemical carcinogens cause cancer by changing the DNA found in a person's somatic cells (15). One example related to air pollution: according to a recent environmental study air pollution has a strong link to cervical cancer (16). We also found an extremely strong correlation between Cervix Cancer and the air pollutants, PM 2.5 in the west coast region across the years 2018-2020. In general, air pollutants have a strong correlation to cancer because tiny air pollutant particles may build up in the lungs and damage the DNA in somatic body cells, which can affect mitosis, how cells divide (17).

There were many challenges faced in this case study. In the raw data, there were many erroneous rows, columns, and problems with joining the cancer and environmental datasets which required lots of data cleaning and processing. This study was performed at a state granularity rather than at a city granularity because there was limited information in publicly available datasets on both cancer and environmental metrics for cities and counties. Additionally, for the cancer dataset chosen, there wasn't enough publicly available data on certain types of Cancer such as Mouth Cancer and Eye Cancer which limited the scope of the study.

This field holds significant potential for innovation and expansion. Validating our hypothesis by examining correlations in other countries could be a valuable next step. Further, the project scope could be broadened by incorporating additional diseases, cancer types, and environmental metrics. Conversely, narrowing the focus to a single pollutant and cancer type would enable a more comprehensive analysis, incorporating factors like race, gender, and income to better explain the Pearson correlation coefficient. As a next step, conducting the study at a city level could enhance accuracy by concentrating environmental factors and cancer types to a specific area.

MATERIALS AND METHODS

In order to determine if there was a correlation between pollutants metrics and cancer deaths, we first needed to obtain a reliable dataset for cancer deaths and these pollutants. We chose epa.gov for generating our environmental dataset, and cdc.gov for our cancer dataset, as they had the most comprehensive data across many years. Epa.gov has raw data for air pollution (18), water pollution (19), and chemical pollution (20) from 1980 to 2023 and considers over 200 different air pollution metrics including NOx, CO2, PB, Ozone, PM10, Lead, Toxic Waste in Oceans, etc. at a state granularity. CDC has cancer datasets (2) from 1980 to 2020 and counts the total number of cancer deaths as well as the population at a state granularity. The correlation between chemical, water, and air pollution in geographical clusters and cancer mortality rate was investigated. To pre-process, clean and analyze the data, we used a custom developed python script. After normalizing and merging the cancer and environmental data we used the Pearson Correlation Coefficient to determine the strength of the correlation between the prevalence of cancer incidences and air pollutants. Finally, we drew conclusions on the correlation between environmental factors and disease-related indicators after performing the analysis for 3 years (2018-2020).

The pandas, csv, seaborn and matplotlib python libraries were used with Kaggle as the platform to develop code (**22**) and plot all the graphs for this research analysis. A total number of 36 Cancer types, across eight regions in the USA, with 5 air pollutants, 2 water-pollutants and 5 chemical pollutants were analyzed (**Table 1**). A total number of 36 x 8 x (5+2+5) = 3456 Correlation Coefficient Matrices were created and 1.5+ Million rows of raw data processed into valid data frames. The flowchart in **Figure 7** describes the data procedures and data cleaning in more detail. We tried to find raw cancer and environmental data using Google Scholar, Perplexity.AI, and Google Search. However, we chose epa.gov for environmental data and cdc.gov for cancer data because these sources had the most comprehensive data across many years at a state granularity. Then, the raw data was imported into Kaggle and was extensively cleaned to convert the raw unusable data into usable datasets. This included removing unnecessary rows and columns, updating certain column data names, removing incomprehensible and erroneous values in certain rows such as squiggly signs, operators etc. to prepare the dataset for correlation study. In the pollution data frame, we selected the necessary columns (air pollutants/metrics) and grouped the data by each state. Then, we created the cancer dataset. We selected the following columns: cancer count, population, and area (for a Cancer type). To normalize the cancer incidence rate, we created an additional column to account for the normalized count, calculated as count divided by population x 10000. This new column accounts for an unbiased percent of people with cancer in each state and population distribution. After obtaining these new data frames, we clustered states into eight geographic clusters, the Great Plains, New England, Southwest, Rocky

Mountains, West Coast, Midwest, East, and Mid-Atlantic. Hawaii and Alaska were excluded from the study because there would only be one state in those regional levels leading to invalid correlation results. Then, we mapped the environmental and cancer data into the different clusters. Then, we determined the strength of the correlation between the prevalence of cancer deaths and pollutants across these eight clusters using the Pearson Correlation Coefficient.

REFERENCES

1. “2022 Cancer Facts & Figures Cancer: Cancer Death Rate Drops.” 2022 Cancer Facts & Figures Cancer | Cancer Death Rate Drops | American Cancer Society, John Wiley & Sons, Inc., 12 Jan. 2022, www.cancer.org/research/acs-research-news/facts-and-figures-2022.html. Accessed 18 Aug. 2024.
2. “Cancer Data and Statistics.” Centers for Disease Control and Prevention, June 2024, www.cdc.gov/cancer/data/index.html. Accessed 18 Aug. 2024.
3. Annual Report to the Nation 2022 National Cancer Statistics.” SEER, Apr. 2023, seer.cancer.gov/report_to_nation/statistics.html. Accessed 18 Aug. 2024.
4. “Risk Factors: Age.” Risk Factors: Age - NCI, 5 Mar. 2021, www.cancer.gov/about-cancer/causes-prevention/risk/age. Accessed 18 Aug. 2024.
5. “Air Pollution.” World Health Organization, www.who.int/health-topics/air-pollution#tab=tab_1. Accessed 18 Aug. 2024.
6. Christensen, Jen. “A Quarter of Americans Live with Polluted Air, with People of Color and Those in Western States Disproportionately Affected, Report Says.” CNN, Cable News Network, 19 Apr. 2023, www.cnn.com/2023/04/19/health/state-of-the-air-2023/index.html. Accessed 18 Aug. 2024.
7. “Greenhouse Gases Continued to Increase Rapidly in 2022.” National Oceanic and Atmospheric Administration, www.noaa.gov/news-release/greenhouse-gases-continued-to-increase-rapidly-in-2022. Accessed 18 Aug. 2024.
8. “State of Lung Cancer: Key Findings.” State of Lung Cancer | Key Findings | American Lung Association, Simon & Schuster, 7 June 2024, www.lung.org/research/state-of-lung-cancer/key-findings. Accessed 18 Aug. 2024.
9. Ward, Mary, et al. “Drinking Water Nitrate and Human Health: An Updated Review.” International Journal of Environmental Research and Public Health, vol. 15, no. 7, 23 July 2018, p. 1557, doi:10.3390/ijerph15071557.

10. Pope III, C. Arden. “Lung Cancer, Cardiopulmonary Mortality, and Long-Term Exposure to Fine Particulate Air Pollution.” JAMA, vol. 287, no. 9, 6 Mar. 2002, p. 1132, doi:10.1001/jama.287.9.1132.
11. Sampaio, Nilo Antônio, et al. “Applications of Correlation Analysis in Environmental Problems.” Revista de Gestão Social e Ambiental, vol. 18, no. 3, 6 Mar. 2024, doi:10.24857/rgsa.v18n3-085.
12. “Water Contaminants and Cancer Risk: Arsenic, Disinfection Byproducts, and Nitrate.” Water Contaminants and Cancer Risk: Arsenic, Disinfection Byproducts, and Nitrate - NCI, dceg.cancer.gov/research/what-we-study/drinking-water-contaminants#:~:text=So%20far%2C%20they%20found%20that,and%20colorectal%20and%20pancreas%20cancers. Accessed 18 Aug. 2024.
13. “Water Contaminants and Cancer Risk: Arsenic, Disinfection Byproducts, and Nitrate.” Water Contaminants and Cancer Risk: Arsenic, Disinfection Byproducts, and Nitrate - NCI, dceg.cancer.gov/research/what-we-study/drinking-water-contaminants#:~:text=Rena%20Jones.-,Nitrate,which%20are%20potent%20animal%20carcinogens. Accessed 18 Aug. 2024.
14. Vienne-Jumeau, A., et al. “Environmental Risk Factors of Primary Brain Tumors: A Review.” Revue Neurologique, vol. 175, no. 10, Dec. 2019, pp. 664–678, doi:10.1016/j.neurol.2019.08.004.
15. “Determining If Something Is a Carcinogen.” American Cancer Society, www.cancer.org/cancer/risk-prevention/understanding-cancer-risk/determining-if-something-is-a-carcinogen.html#:~:text=Some%20carcinogens%20cause%20cancer%20by,that%20DNA%20changes%20will%20occur. Accessed 18 Aug. 2024.
16. Maisara, Siti, et al. “Air Pollution and Its Association with Cervical Cancer: A Scoping Review.” Medicine &amp; Health, vol. 18, no. 1, 16 June 2023, pp. 9–20, doi:10.17576/mh.2023.1801.03.
17. “How Can Air Pollution Cause Cancer?” Cancer Research UK, 28 June 2024, www.cancerresearchuk.org/about-cancer/causes-of-cancer/air-pollution-radon-gas-and-c

ancer/how-can-air-pollution-cause-cancer#:~:text=Because%20air%20pollution%20contains%20a,which%20can%20lead%20to%20cancer. Accessed 18 Aug. 2024.

18. “Air Data: Air Quality Data Collected at Outdoor Monitors Across the US.” EPA, Environmental Protection Agency, 30 Mar. 2024, www.epa.gov/outdoor-air-quality-data. Accessed 18 Aug. 2024.
19. “State Statistics.” EPA, Environmental Protection Agency, 30 Mar. 2024, echo.epa.gov/trends/loading-tool/get-data/state-statistics. Accessed 18 Aug. 2024.
20. “TRI Basic Data Files: Calendar Years 1987-Present.” EPA, Environmental Protection Agency, 30 Mar. 2024, www.epa.gov/toxics-release-inventory-tri-program/tri-basic-data-files-calendar-years-1987-present. Accessed 18 Aug. 2024.
21. Alexander, Kathy. Legends of America, www.legendsofamerica.com/ah-geosum/. Accessed 18 Aug. 2024.

Appendix

22. Link to code on github (homed in Kaggle):
https://github.com/noblepenguin43/cancer-environment-corelation