Visualizing Air Pollution and Related Mortality in the United States
Team 48 Final Report

Minh Anderson, Baris Kopruluoglu, Kayla Masozera, Joshua Perry, Yi-Chuen Wang

Introduction/Motivation

The United States Environmental Protection Agency (EPA) stated that “in 2022, about 66 million tons of pollution were emitted into the atmosphere in the United States” [17]. Despite the emergence of research and conversation revealing the impact of air pollution on the health of individuals and the environment, the threat continues to lurk in the overhead due to slow movement for change. The data is available but influential decision makers and citizens must be equipped with information that outlines a clear message which can be accomplished through effective visualization and data analysis.

Problem Definition

In this project, we created interactive maps of various forms of air pollution within each state and their associated mortality counts in certain counties. Although the EPA and the Centers for Disease Control and Prevention (CDC) collect thorough data concerning either air pollution or mortality, our approach layers both and examines their relationship. It offers insights into pollution hotspots and areas with heightened health risks due to the incorporation of geographic data. It also assists individuals in making informed decisions about where to reside or relocate, contributing to better public awareness and empowering local governments in their decision-making, taking into consideration environmental and public health effects. Furthermore, we run the data through analytical models to predict expected future air pollution and air pollution related death rates e.g., respiratory using various prediction models.

Literature Survey

A recent study investigated exposure disparities in air pollution by race/ethnicity and income on different time periods which will be a helpful resource to spot where these disparities exist [1]. However, it didn’t explore disparities at a fine grained geographic level which can be improved. It has been investigated how particulate matter (PM2.5), ozone and combination of high temperature and air pollution affects premature mortality and burden of death which we will use to address the substantial contribution of air pollution to deaths [2], [3], [8]. This can be extended to explore which different pollutants affect health more significantly on a wider and more diverse population.

We plan to survey some papers which discuss attributes of air pollution including wind patterns [7], motorcycle traffic [5] , and public transportation [9]. We plan to utilize these studies to find statistical inference between air pollution and other factors, e.g. number of factories in the city. These studies were based in small geographical areas, whereas we will try to apply their principles across the entire U.S.

Air pollution indices have been used to create awareness about health risks and mortality in different locations including Hong Kong [4], [6]. We will create our own pollution index for the U.S. based on studies that provide insight into the algorithms such as fuzzy c-means clustering [10] and give the user a more intuitive way to interpret air pollution distribution.
We also found articles that show visualization of air and environmental quality using real time pollution data on a macro and micro scale [11], [16] and cluster analysis of research relating pollution and health [12]. These will be useful in collecting data based on the trends found in air

pollution research and how we can approach forecasting air quality in the near future.

Some parametric machine learning algorithms have been utilized to predict air pollution including neural networks and deep learning in India, China and Iran [13], [14], [15]. Similarly, we apply some predictor based statistical learning methods to predict air pollution in the U.S. including decision tree regression and random forest regression.

Innovations

- We present the reader an interactive choropleth map of air pollutants with different features from 2000-2016 at the county level in the U.S.. Relevant ancillary details will be included in the tooltip.
- We present the reader an interactive choropleth map of the average nitrogen dioxide and some other air pollutants in a specific county over a ten year period and the death rate the year after to analyze their association.
- We present the reader prediction models on the annual nitrogen dioxide emission average in the U.S. with factors including population density, the number of registered vehicles, and the number of factories in that county.
- We present the reader prediction models for the number of air pollution-related deaths for the next year in a specific county based on the data for the past ten years.

Data Cleaning/Integration

The project goal requires multiple datasets including the U.S. Pollution 2000-2016 dataset that was accessed via Kaggle and used as the building point of our additional data. It contains missing values in features that will not be necessary for our analysis. It also includes repeating measurements that were resolved in the aggregation process. Features of the data include daily measurements of certain air pollutants including nitrogen dioxide, carbon monoxide at 204 different addresses in 139 counties in 47 states over the period from 2000 to 2016, totaling up to approximately 1.8 million data points. We used Spark for scalable computing.

Another dataset that we utilized is death causes which were scraped from the CDC website and grouped by year, state and county. Cause of death was, then, grouped into categories (e.g., respiratory, circulatory or cancer) to explore the depth of connection to air pollution.

To investigate the link between industrial pollution and public health, we accessed data from the Census Bureau website, offering information on the number of factories in various counties from 2000 to 2016.

Interactive Visualization Maps

We created different choropleth maps representing a distribution of different air pollutants county by county along with respiratory-based deaths. One example can be seen below in Figure 1. In this plot, we see the average NO_2 emission levels for each county for 10 years (2005-2014) and the number of deaths for each county in 2015. The size of the circles translates to mortality count while their color represents average NO_2 emissions. At first glance, there is a noticeable association between NO_2 level and respiratory deaths as the size of the circles increases, the color shifts to yellow along the scale. Similarly, we created choropleth maps that show the average emission level of SO_2, O_3 vs. related deaths in the same counties over the same time period. We found that NO_2 emission level has the strongest association with

respiratory deaths according to various resources that indicated that NO_2 is one of the most dangerous air pollutants to human health. (Fig.1)

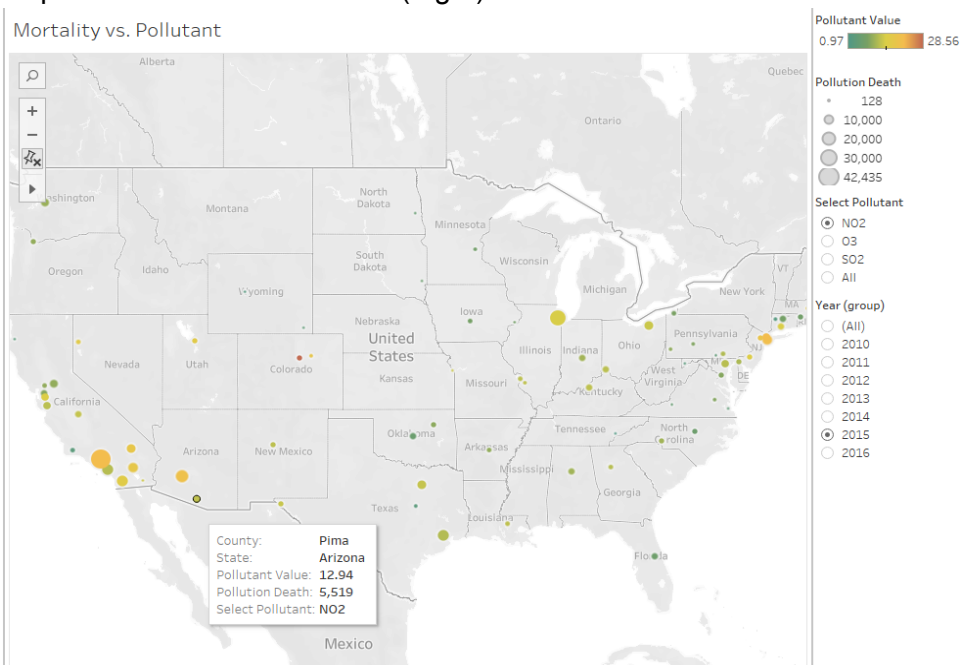

Fig.1 Pollution related death counts vs. three different pollutants by year.

At this juncture, we direct our attention to the state of California, aiming to scrutinize the interplay between the averaged pollutants and diverse pollution factors, to include number of factories and number of vehicles registered. We specify California because these counties have data for all of our predictors. Vehicle registration data for each state is highly varied in format and availability across states due to a lack of a single repository for county data across the country. A detailed examination of this relationship is facilitated through the visualization of correlation patterns on a heatmap. Upon selecting a specific value on the heatmap, a corresponding scatter plot provides additional insights about the nuanced interplay between these variables. (Fig.2 and Fig.3)

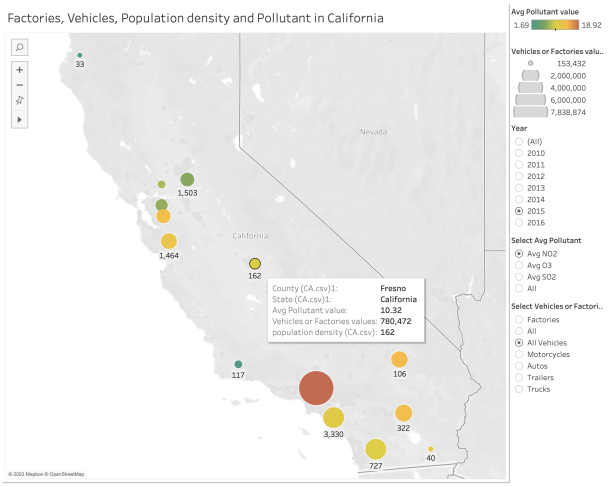

Fig.2 Average of three pollutants and various pollution factors by year in California

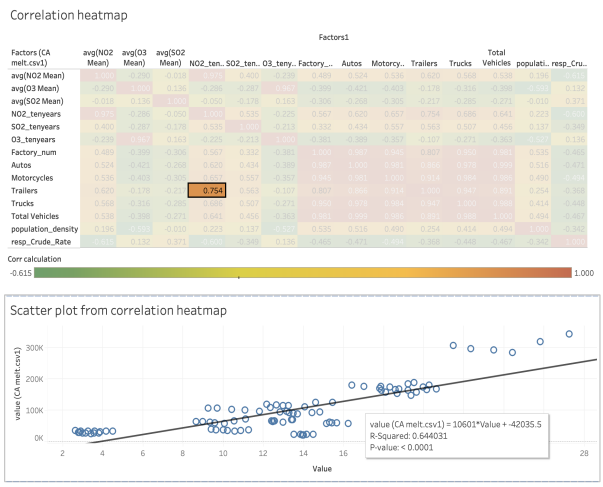

Fig.3 Correlation heatmap with related scatter plot

Proposed Methods

Intuition:
Air pollution can become an oft-forgotten concern when its effects are not presented clearly. Instead of separating mortality and pollution metrics, analyzing them head-on and providing an interpretable product to convey possible relationships and make predictions may amplify public-health recommendations of the EPA and CDC. This is what we hope to accomplish through our project.

Methods:
In the literature, most proposed methods are time series based methods which account for trends, seasonality and yield good accuracy. For example, we see that some air pollutants significantly increase over the summer months. However, to be able to take action to decrease air pollution and, more importantly, diseases and deaths related to air pollution, we prioritized factor-based methods. In addition, we account for the level of air pollutants over a long period and their effect at the end of that period. We propose to the reader two prediction approaches as follows.

I. Prediction of the average level of air pollutants (e.g. NO_2) based on some factors

We created prediction models for the annual NO_2 average based on the specific county. After the data integration process and grouping by state, county and year, our dataset resulted in 53,000 data points. We utilized correlation coefficient and *p*-value to determine what factors are most associated with NO_2 emission. First, we fit a few models without the number of registered vehicles due to the lack of available data for the time frame and locations we were analyzing. Then, we used the detailed vehicle registration information for California from 2000 to 2016 and fit models for data including the number of vehicles, in which we had about 1,100 rows. Here are the correlation coefficients for NO_2 response vs. our predictors:

	CO	O_3	SO_2	Population density	Number of factories	Number of autos	Number of trucks
Corr Coef	0.511820	-0.10586	0.166962	0.257804	0.469287	0.594719	0.656627

We checked the linearity and constant variance assumptions of multiple linear regression and ensured our regression models did not reveal any multicollinearity based on Variance Inflation Factor. We applied different regression models such as multiple linear, decision tree, and random forest to predict NO_2. In each model, we split our dataset as 80% training and 20% test. The results are described further in the Evaluation section. Random Forest and Decision Tree Regressions yielded better accuracy. We employed 10-fold cross-validation to choose the best fitted hyper-parameters for our decision tree and random forest regression models including number of trees, OOB score, max_depth, error criterion. Some parameters for these models can be seen in the chart.

II. Prediction of the mortality related to air pollution

Our second prediction approach is to predict pollution-related mortality (e.g. respiratory) counts in a certain county based on the average NO_2 and SO_2 emission level of the county over a ten-year period, the county's population and population density. Here are the correlation coefficients for mortality vs. our predictors:

	NO_2	SO_2	Population Density	Population
Corr Coef	0.492	0.115	0.171	0.983

We verified similar assumptions as we did before our first prediction experiment and, surprisingly, when we examined multicollinearity using VIF, we did not obtain any value that is greater than 5. We were expecting population density and population to show some collinearity. Hence, we kept both predictors. The *p*-values for each predictor showed that all predictors are significant at 90% level. The reason we did not include ozone is that it was not a significant predictor based on the *p*-value and it also had a negative correlation coefficient with mortality counts which does not align with the purpose of our experiment. Similar to that of the first experiment, we split our data into 80% of points for training and the remaining 20% became the test set.

Results/Evaluation

I. Can we predict the average level of NO_2 based on some factors?

The results will be discussed in two parts within this subsection. The first chart displays the results without the number of registered vehicles and the second chart displays results with vehicle data included. The predictor, number of vehicles, significantly improves decision tree and random forest models' accuracy so to expand this model to other territories within the United States, vehicle registration data for all counties would require intensive data scraping due to lack of availability and uniformity of data collection between states. We were able to get registration information from California for the years in which our investigation was conducted. Those results can be seen in the second chart.

Our selected performance metrics, Mean Squared Error and Mean Absolute Error, inform the reader of errors in the test dataset predictions. Our aim is to find the model that predicts NO_2 emission with lowest of these errors. The results are in terms of parts per billion (ppb), a measure unit for air pollutants. Here are the results without number of vehicles as a predictor:

	Linear Reg. (R^2 =0.654)	**Random Forest Reg. (200 trees, max_depth: None)**	**Decision Tree Reg. (max_depth: 5)**
MSE	11.794	6.167	11.987
MAE	2.639	1.826	2.696

Random forest regression resulted in the least errors in both metrics. Hence, we propose the random forest regression model with the indicated predictors in the previous section to predict annual average NO_2 level for a certain county. 75% of counties we examined throughout the U.S. have annual NO_2 averages more than 9 ppb. Hence, our model provides an estimation within 20% error for most counties. Secondly, the results below include the number of vehicles as a predictor, i.e. for the state of California. For 85% of the counties in California, our error margin is less than 20%.

While these results are not the best, they are arguably acceptable and can be improved upon further as more predictors are introduced.

	Linear Reg. (R^2 =0.726)	**Random Forest Reg. (100 trees, max_depth: None)**	**Decision Tree Reg. (max_depth:None)**
MSE	8.285	3.03	2.991
MAE	2.347	1.243	1.296

II. Can we predict respiratory deaths based on air pollution?

	Linear Reg.	**Random Forest Reg. (100 trees, max_depth: None)**	**Decision Tree Reg. (max_depth: None)**
MSE	28,605.57	6,503.42	11,959.87
MAE	115.29	49.9	57.851
MAPE	41.75%	11.96%	10.32%

Since MSE and MAE become more difficult to interpret as the scale increases, we included MAPE (Mean Absolute Percentage Error) to provide additional clarity about accuracy results. To predict respiratory death toll in a certain county and achieve the lowest MAPE, we propose decision tree regression. This prediction model was more successful than the first model we fit to predict NO_2 level.

Conclusion/Discussion

In this study, we created interactive maps to inform the reader about various types and facets of air pollution and provide insight concerning the impact of air pollution on respiratory-based deaths. We examined the effect of air pollution over an extended time period (ten years) on respiratory-based deaths. As additional data becomes available, this methodology can and should be extended to longer time periods to more closely investigate the association between air pollution and respiratory deaths. Based on our observations, nitrogen dioxide has the largest impact on air pollution related deaths.

Moreover, we focused on answering the question “what is contributing to air pollution?” rather than “when/where is the air being polluted?”. In future work, it would be worth exploring seasonality and additional time-based trends in tandem with our factor-based approach. Our project serves as a starting point for local governments and agencies to use information to decrease actions and decisions that increase air pollution and affect public health quality. Examples include writing legislation that limits the number of factories per county and optimizing infrastructure to encourage public transport use.

We acknowledge that our prediction models have room to improve significantly with the availability of recent data and inclusion of additional features. Both air pollution and death count are difficult to predict as the uncertainty in these measurements is inherent, therefore, we are satisfied with the results as a starting ground for further development.
All team members have contributed a similar amount of effort.

References

1. Jiawen Liu, Lara P. Clark, Matthew J. Bechle, Anjum Hajat, Sun-Young Kim, Allen L. Robinson, Lianne Sheppard, Adam A. Szpiro, and Julian D. Marshall (2021). Disparities in Air Pollution Exposure in the United States by Race/Ethnicity and Income, 1990–2010. Environmental Health Perspectives, 129:12 https://doi.org/10.1289/EHP8584
2. Dedoussi, I.C., Eastham, S.D., Monier, E. et al. Premature mortality related to United States cross-state air pollution. Nature 578, 261–265 (2020). https://doi.org/10.1038/s41586-020-1983-8
3. Bowe B, Xie Y, Yan Y, Al-Aly Z. Burden of Cause-Specific Mortality Associated With PM2.5 Air Pollution in the United States. JAMA Netw Open. 2019 Nov 1;2(11):e1915834. doi: 10.1001/jamanetworkopen.2019.15834. PMID: 31747037; PMCID: PMC6902821. https://pubmed.ncbi.nlm.nih.gov/31747037/
4. Li, L. et al., (2015). Can the Air Pollution Index be used to communicate the health risks of air pollution? Environmental Pollution, 205, 153-169 https://www.sciencedirect.com/science/article/abs/pii/S0269749115002729?via%3Dihub
5. Kumar, N. et al., (2023). Traffic-Related Air Pollution and Associated Human Health Risk. Macromolecular Symposia, 407, 2100486. https://doi.org/10.1002/masy.202100486
6. Thach, T. et al., (2018). A novel method to construct an air quality index based on air pollution profiles. International Journal of Hygiene and Environmental Health, 221, 17-26 https://pubmed.ncbi.nlm.nih.gov/28988894/
7. Anderson, M.L. As the Wind Blows: The Effects of Long-Term Exposure to Air Pollution on Mortality. Journal of the European Economic Association, Volume 18, Issue 4, August 2020, Pages 1886–1927, https://academic.oup.com/jeea/article/18/4/1886/5580747
8. Willers SM, Jonker MF, Klok L, Keuken MP, Odink J, van den Elshout S, Sabel CE, Mackenbach JP, Burdorf A. High resolution exposure modelling of heat and air pollution and the impact on mortality. Environ Int. 2016 Apr-May;89-90:102-9.https://pubmed.ncbi.nlm.nih.gov/26826367/ Epub 2016 Jan 28. PMID: 26826367.
9. Triguero-Mas M, Martínez-Solanas È, Barrera-Gómez J, Agis D, Pérez N, Reche C, Alastuey A, Querol X, Pérez K, Basagaña X. Public Transport Strikes and Their Relationships With Air Pollution, Mortality, and Hospital Admissions. Am J Epidemiol. 2020 Feb 28;189(2): 116-119. https://academic.oup.com/aje/article/189/2/116/5578513?login=false PMID: 31566673.
10. R. K. Grace, K. Aishvarya S., B. Monisha and A. Kaarthik, "Analysis and Visualization of Air Quality Using Real Time Pollutant Data," 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2020, pp. 34-38, doi: 10.1109/ICACCS48705.2020.9074283. https://ieeexplore.ieee.org/document/9074283
11. Chen, P. Visualization of real-time monitoring datagraphic of urban environmental quality. J Image Video Proc. 2019, 42 (2019). https://doi.org/10.1186/s13640-019-0443-6
12. Liu D, Cheng K, Huang K, Ding H, Xu T, Chen Z, Sun Y. Visualization and Analysis of Air Pollution and Human Health Based on Cluster Analysis: A Bibliometric Review from

2001 to 2021. Int J Environ Res Public Health. 2022 Oct 5;19(19):12723. Doi: 10.3390/ijerph191912723. PMID: 36232020; PMCID: PMC9566718. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9566718/

13. Kumar K, Pande BP. Air pollution prediction with machine learning: a case study of Indian cities. Int J Environ Sci Technol (Tehran). 2023;20(5):5333-5348. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9107909/ Epub 2022 May 15. PMID: 35603096; PMCID: PMC9107909.
14. Xu R, Wang D, Li J, Wan H, Shen S, Guo X. A Hybrid Deep Learning Model for Air Quality Prediction Based on the Time–Frequency Domain Relationship. Atmosphere. 2023; 14(2):405. https://doi.org/10.3390/atmos14020405
15. Delavar MR, Gholami A, Shiran GR, Rashidi Y, Nakhaeizadeh GR, Fedra K, Hatefi Afshar S. A Novel Method for Improving Air Pollution Prediction Based on Machine Learning Approaches: A Case Study Applied to the Capital City of Tehran. ISPRS International Journal of Geo-Information. 2019; 8(2):99. https://doi.org/10.3390/ijgi8020099
16. Enigella, Sumanth Reddy, and Hamid Shahnasser. "Real time air quality monitoring." 2018 10th International Conference on Knowledge and Smart Technology (KST). IEEE, 2018. https://ieeexplore.ieee.org/abstract/document/8426102
17. Air Quality - National Summary | US EPA, www.epa.gov/air-trends/air-quality-national-summary. Accessed 2 Nov. 2023.