Visualizing Air Pollution and Mortality in the United States

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

Summary

We created interactive visualizations depicting air pollutants within each state and associated mortality counts in certain counties. We examined the relationship between the two with a different approach. It offers insights into pollution hotspots and areas with heightened health risks with geographic data and accounting for different air pollutants, factors over a long time period. This information assists individuals in making informed decisions about where to reside or relocate, contributing to better public awareness and empowering local governments in their policy making. Furthermore, we run the data through analytical models to predict expected future air pollution and air pollution related death counts e.g., respiratory using various prediction models.

Background

The United States Environmental Protection Agency stated that “in 2022, about 66 million tons of pollution were emitted into the atmosphere in the United States” [www.epa.gov/air-trends/air-quality-national-summary]. 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 analysis.

Data Aggregation

This project required multiple data sources. The base dataset was the U.S. Pollution 2000-2016 file accessed via Kaggle. Features include daily measurements of certain air pollutants including nitrogen dioxide and 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. Mortality data, including cause of death was scraped from the CDC website and grouped by year, state and county. To explore a connection with industrial pollution, the number of factories by county was pulled from the Census Bureau. All data was joined into a “master” file by year, state and county.

Methods & Evaluations

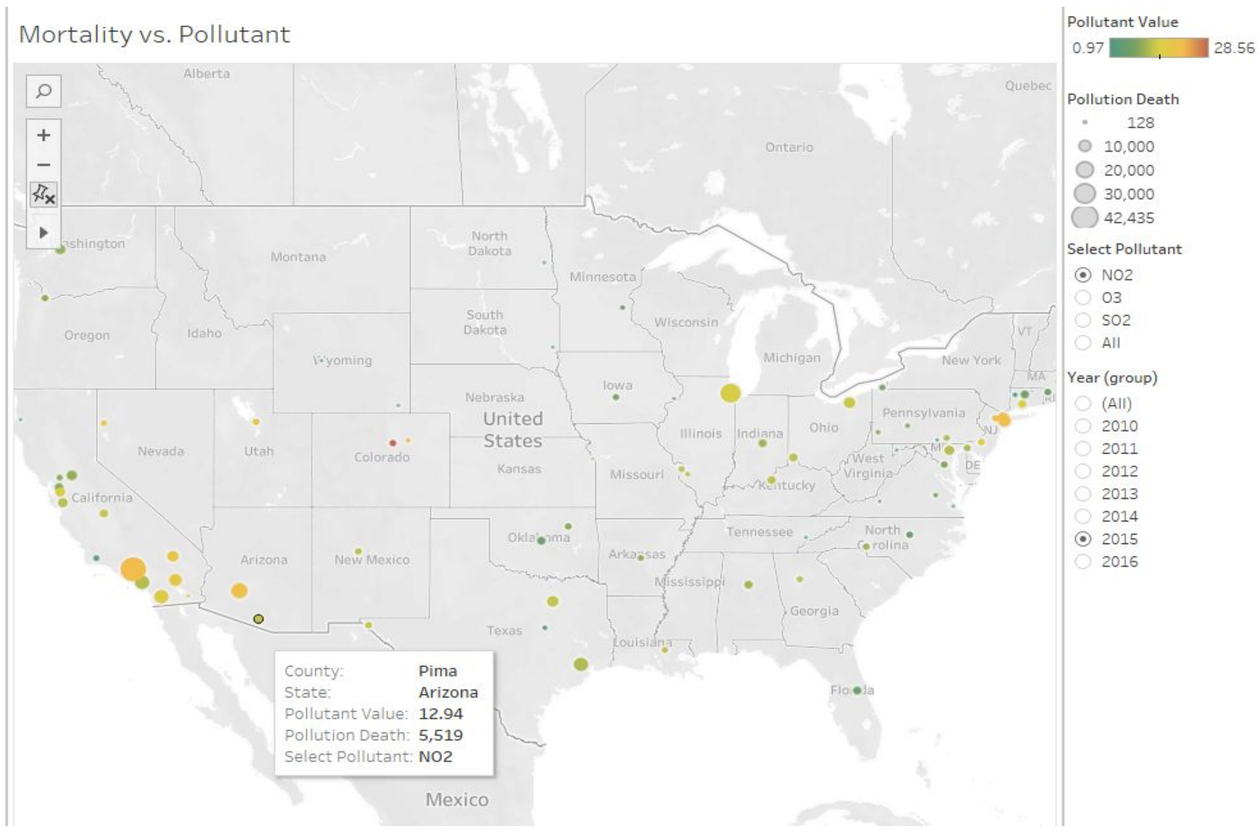
We created prediction models for the annual NO2 average and also for respiratory related death counts. We prioritized factor-based methods rather than time series models, because we wanted to be able answer the question “what contributes to air pollution?” rather than “when does it happen most?”. We did thorough analysis to determine what predictors to use in each model as seen in the correlation heatmap. Due to limited availability of county-specific vehicle registration data, we focused on California for models that include number of vehicles as a predictor.

The assessment for each model is done using 80% training 20% test split and computing mean squared error, mean absolute error and mean absolute percentage error for the test dataset. The table below shows the results for the models that predict respiratory death counts. Since the range of the response variable is very wide, MAPE is a more reliable metric, thus, we propose decision tree regression to predict respiratory death counts using average NO2, SO2 over the past 10 years period, population and population density of a certain county.

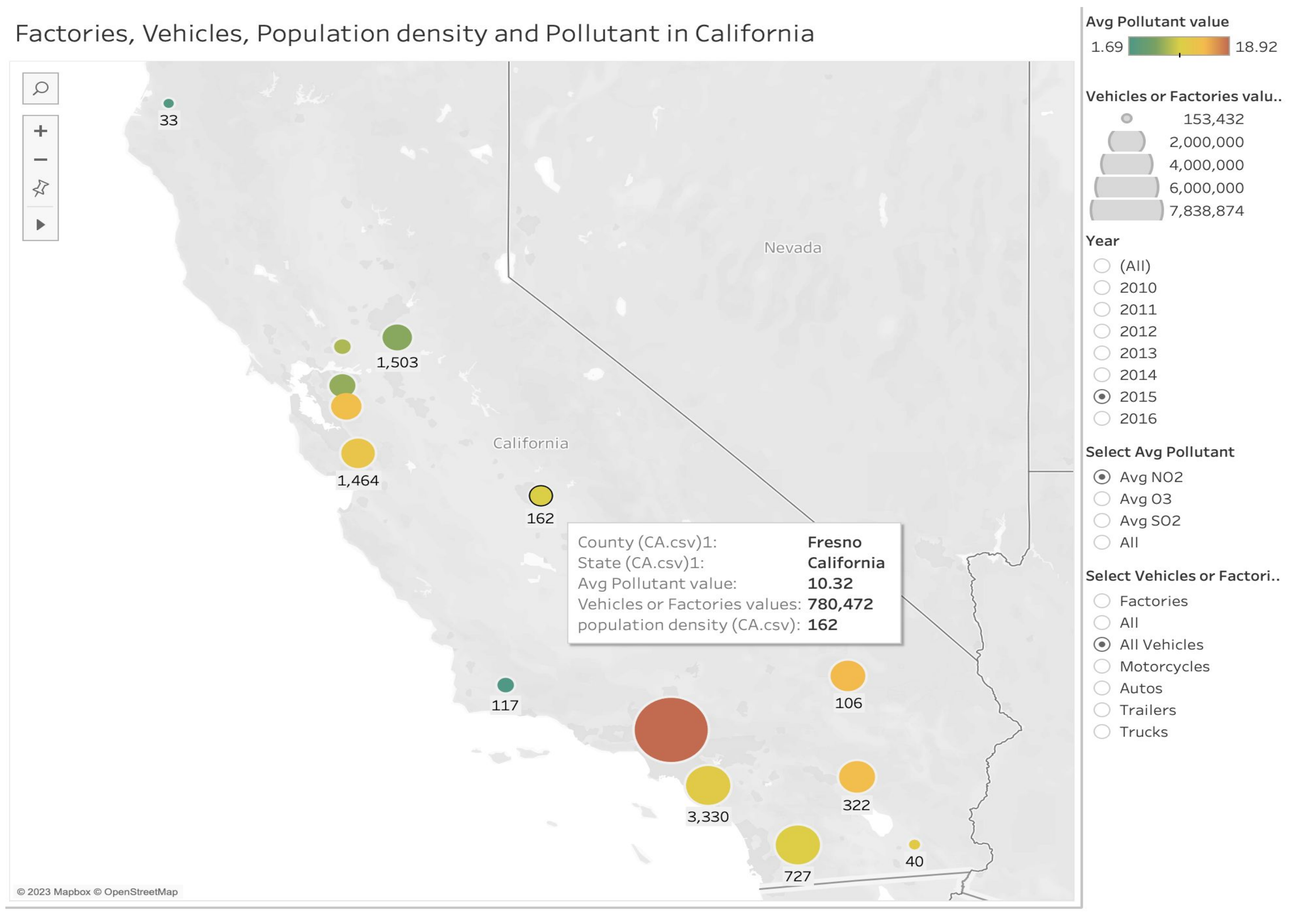
	Linear Reg.	Random Forest Reg. (100 trees, max_depth: None)	Decision Tree Reg. (max_depth: None)
MSE	28,605.57	6,503.42	11,959.87
MAE	115.29	49.9	57.851
MAPE	41.75%	11.96%	10.32%

In conclusion, our study sheds light on critical aspects of air pollution, emphasizing the long-term impact on respiratory deaths. While our prediction models performed acceptably, they have room for improvement with more recent data and additional features.

Pollution Level and Respiratory Deaths



Average Pollution by Number of Vehicles / Factories



Factor Based Model Correlation Coefficients