Leveraging California's Data for Nationwide Environmental Health Risk Prediction

Harper Baer, Samih Qureshi

Datasci 112 Winter 2024

Abstract

Due to historical inequities, there is a relationship between the demographics of a community and its environmental health risk. We analyzed these relationships on the scale of California to help build a machine learning model for the rest of the nation, creating a decision-making tool for environmental and public health policymakers. Our data exploration revealed correlations between race and pollution burden, and our ridge regression model pointed to hot spots of environmental health risk nationwide– specifically in the Cotton Belt and in isolated areas of the Southwest.

Research Question

How do population demographics determine the environmental health of a community?

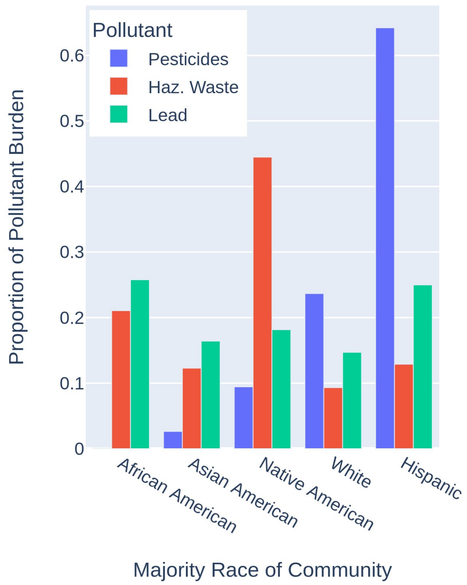
Environmentalism is intersectional; in order to advance work in climate solutions while centering equity, it is crucial to understand how environmental burdens have been unjustly shouldered by marginalized communities.

California is uniquely prolific in its data collection and analysis, exemplified by its CalEnviroScreen tool, which weighs demographic and pollution burden data to produce an overall health score for a community.

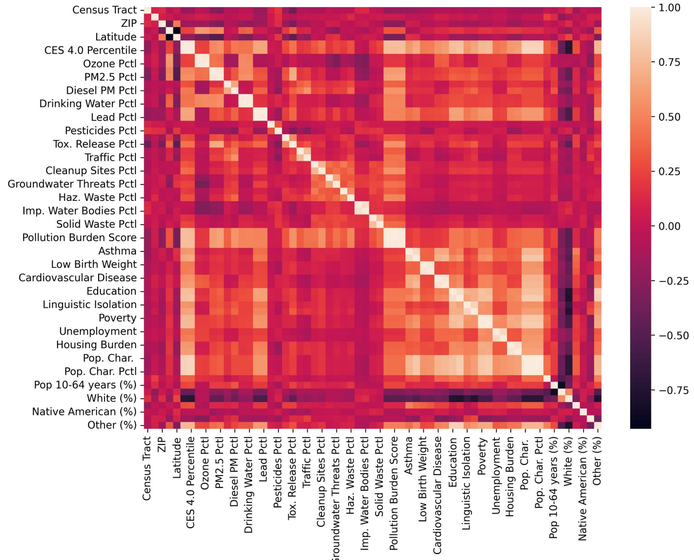
Data Collection

We merged the pollution burden and population demographic data provided by CalEnviroScreen on the key of census tract. For test data, we cleaned U.S. Census Bureau's nationwide demographic data, combining ethnicities into 5 groups, then merged it with census tract location data for geospatial visualization.

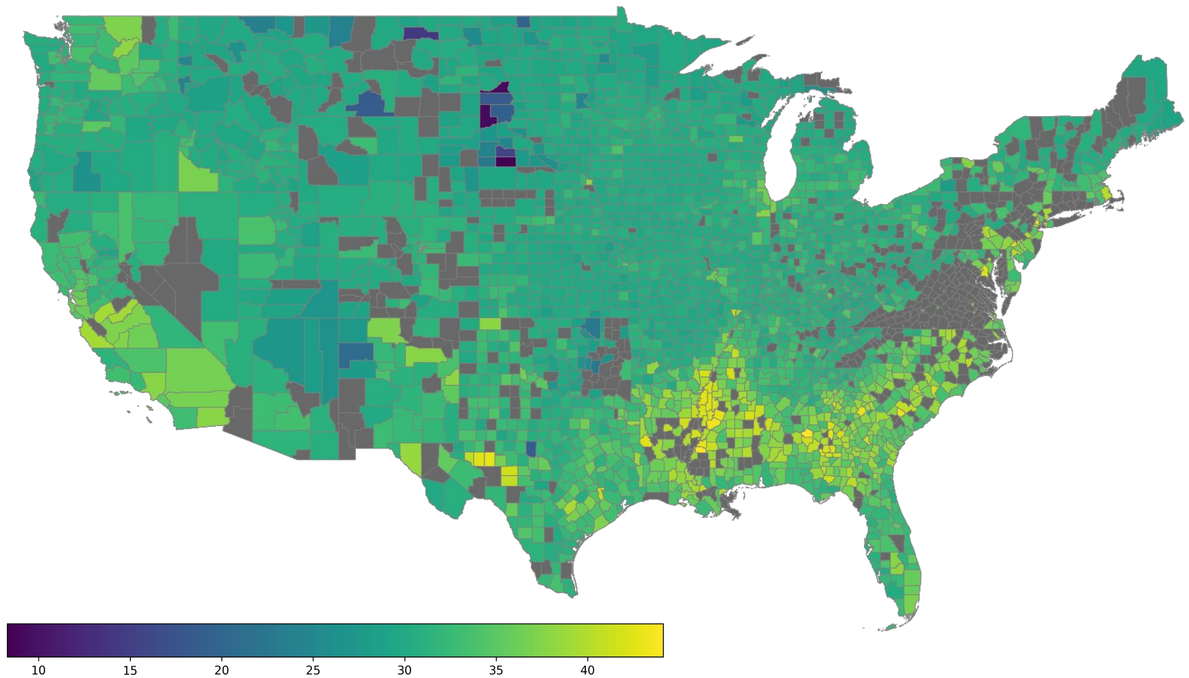
Pollution Burden vs. Race



Correlation Heatmap



Predicted Pollution Burden By County



Data Exploration and Analysis

Our correlation heatmap and bar chart highlight the relationships between race and pollutant levels within a community.

Certain pollutants showed varying correlations, such as pesticides clearly being most present in Hispanic (presumably agricultural) communities.

We then built a ridge regression model (with a test error of 10) using California demographic data as training data, with CalEnviroScreen's Pollution Burden Score as a label, and then tested it on the entire nation.

Conclusion

Our model points to the Cotton Belt and isolated areas in the Southwest as areas of particular national concern, though each state has its own areas of attention.

Data Sources

CalEnviroScreen 4.0 (2021)
U.S. Census Bureau (2020)

Stanford University