Understanding the Impact of Socioeconomic Status on Premature Deaths by Race

Saima Rahman
Colby College

Maximus Liu
University of Texas at Austin

Meris McElveen
University of North Carolina at Chapel Hill

Carnegie Mellon University
Statistics & Data Science

Research Question

How is the socioeconomic status of a county associated with the number of premature deaths of certain racial groups at the county level?

Background

We are motivated to explore how **race** and **geography** shape premature death, as each racial group faces unique leading causes and risk factors. Our research hopes to guide targeted interventions to **increase healthcare access** and **reduce preventable deaths**.

Methods

Models: Multiple linear regression (MLR), Random Forest (RF), XGBoost models

- To predict premature death (PD) rates across counties and assess feature importance by race.
- A final MLR was ran with selected features to provide interpretable results for policy recommendation.
- A model comparison of RF and XGBoost revealed XGBoost outperforms across all evaluation metrics (RMSE, R², cross-validation).
- XGBoost was used to extract the most influential variables and as the final predictive model.

Exploratory Data Analysis (EDA)

1. **Higher Years of Potential Life Lost rates in the southeast US and Alaska.**
Our focus shifted towards counties from this geographic region in order to provide a deeper analysis of how various socioeconomic factors impact racial communities in the south, many of which have historically been disadvantaged in healthcare.

2. **Rural counties consistently demonstrated significantly higher premature death rates compared to their urban counterparts.**
This led to the incorporation of an *Is Rural* variable, created using *% Rural* from the original dataset.

3. **Initial XGBoost & MLR model reveals top socioeconomic factors contributing to premature death.**
We found that top factors vary across all racial groups, but there are consistencies. Our final selection included: *% rural, food environment index, Primary physician rate, % limited access to healthy food, % some college, average PM2.5, % unemployed*, and *income ratio*. XGBoost also revealed some physical health factors associated with PD (e.g., firearm fatalities, alcohol-involved driving deaths).

Average Years of Potential Life Lost (YPLL) Rate by State

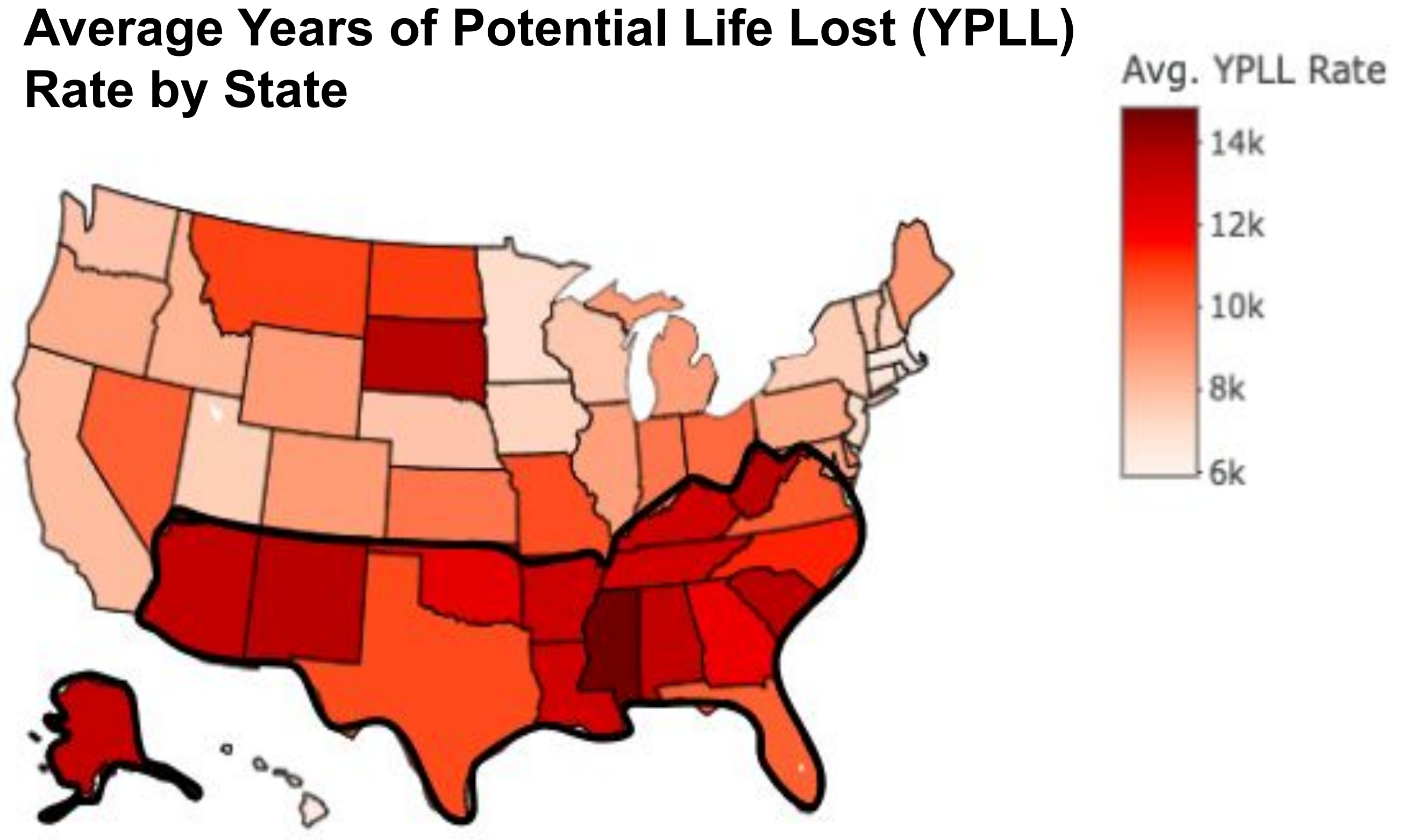

Figure 1. Choropleth map showing Premature Death rates across the United States.

Discussion

Rurality is a key predictor of premature death across all racial groups. Other influential factors include **health-related risks** (e.g., child mortality, firearm deaths) and **socioeconomic markers** (e.g., food access, education, income).

Limitations: Missing data across racial groups limited full national coverage.
Some models may still reflect **biases in data availability or sampling**.

Future work: Expanded analysis to **rural counties nationwide** and **regional subgroups** (e.g., Southeast, Alaska).

- Future work could further refine **models by region or subgroup** and investigate **policy interventions**.

Sources

- University of Wisconsin Population Health Institute. (2024). County Health Rankings & Roadmaps. https://www.countyhealthrankings.org/
- National Institute on Minority Health and Health Disparities. (2024). Health Disparities Data Portal. https://hdpulse.nimhd.nih.gov/

Premature death rate predictions vary by race

Predicted vs Actual Premature Death Rate by Racial Group

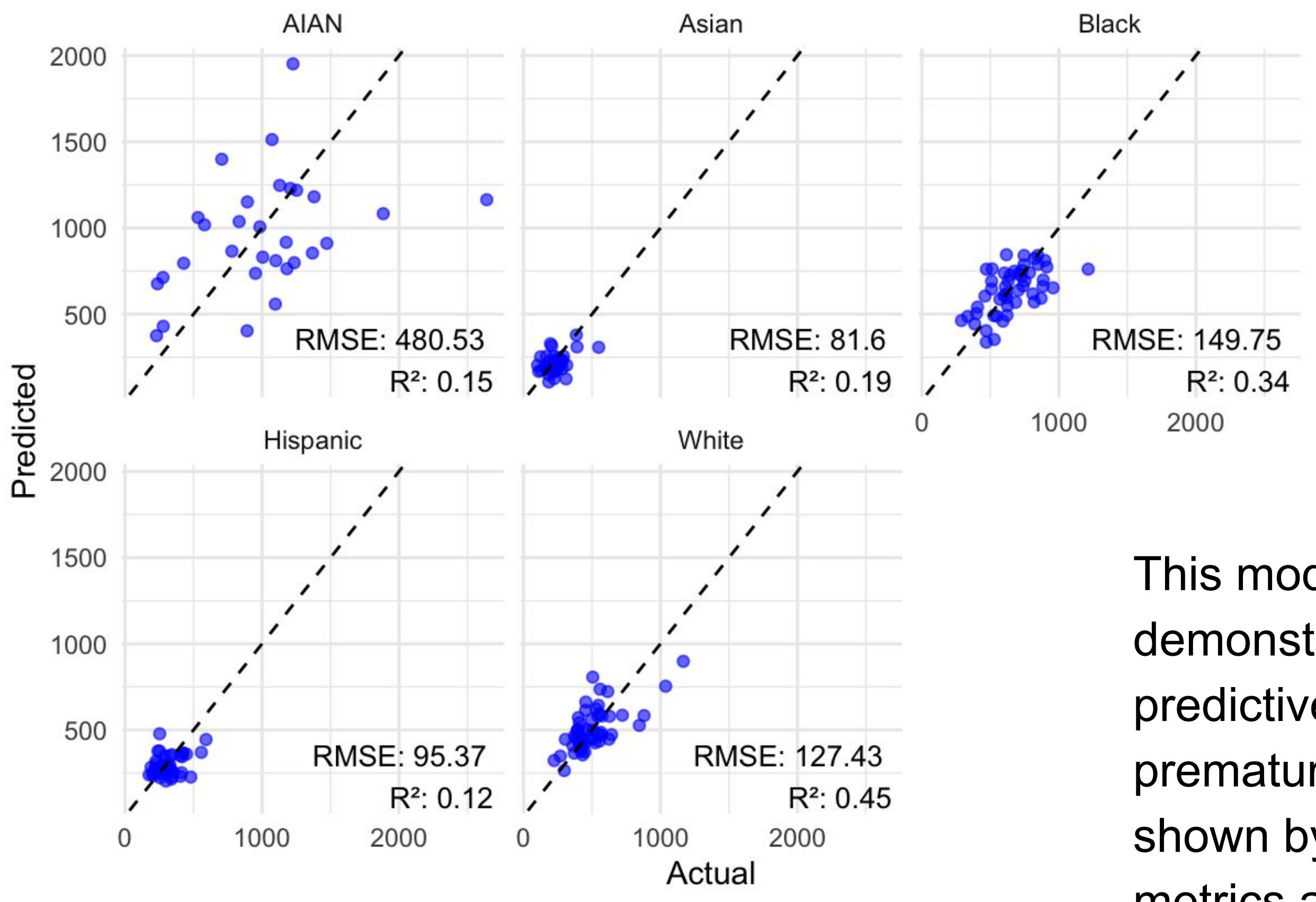

Figure 2. Final XGBoost model performance using cross-selected features across all racial groups.

This model demonstrates strong predictive power for premature death, as shown by low error metrics and tight clustering around the identity line.

Rurality highly affects premature death rate

Predicted Years of Potential Life Lost vs Actual in Rural Southeast US & Alaska

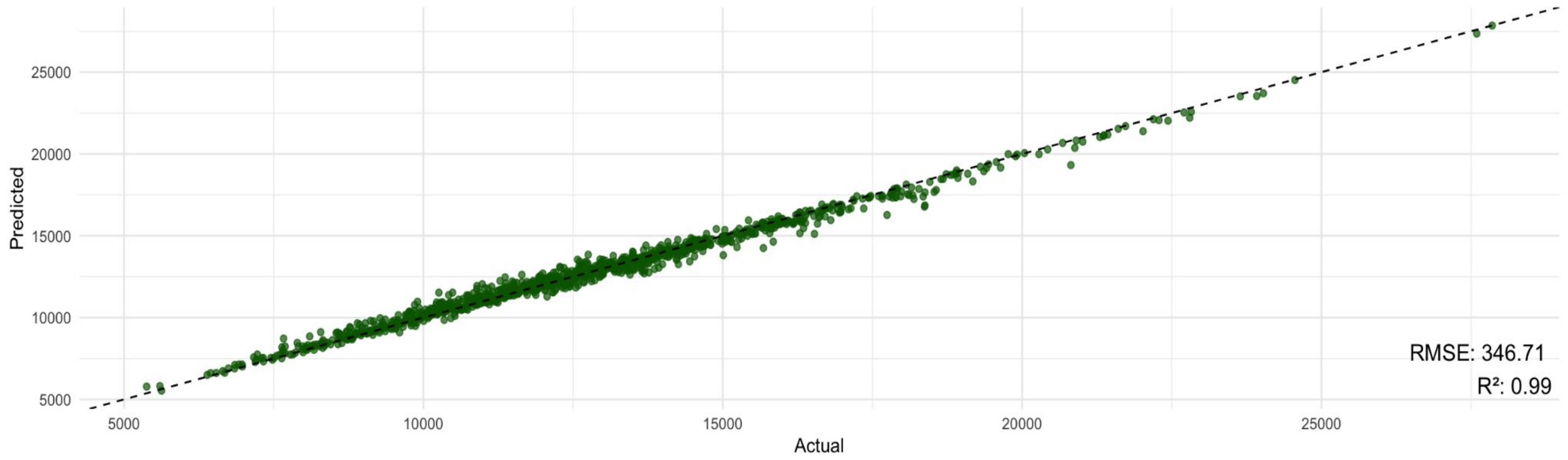

Figure 3. Improved model fit using revised rural-specific features.

An additional XGBoost model focused on all rural (≥30%) US counties revealed 12 new features: *# Not Proficient in English, % Enrolled in Free or Reduced Lunch, Segregation Index, # Unemployed, # Uninsured Children, % Children in Poverty, # Some College, Presence of Water Violation, % Excessive Drinking, Income Ratio, % Limited Access to Healthy Foods, % Uninsured.*

An Environmental Perspective

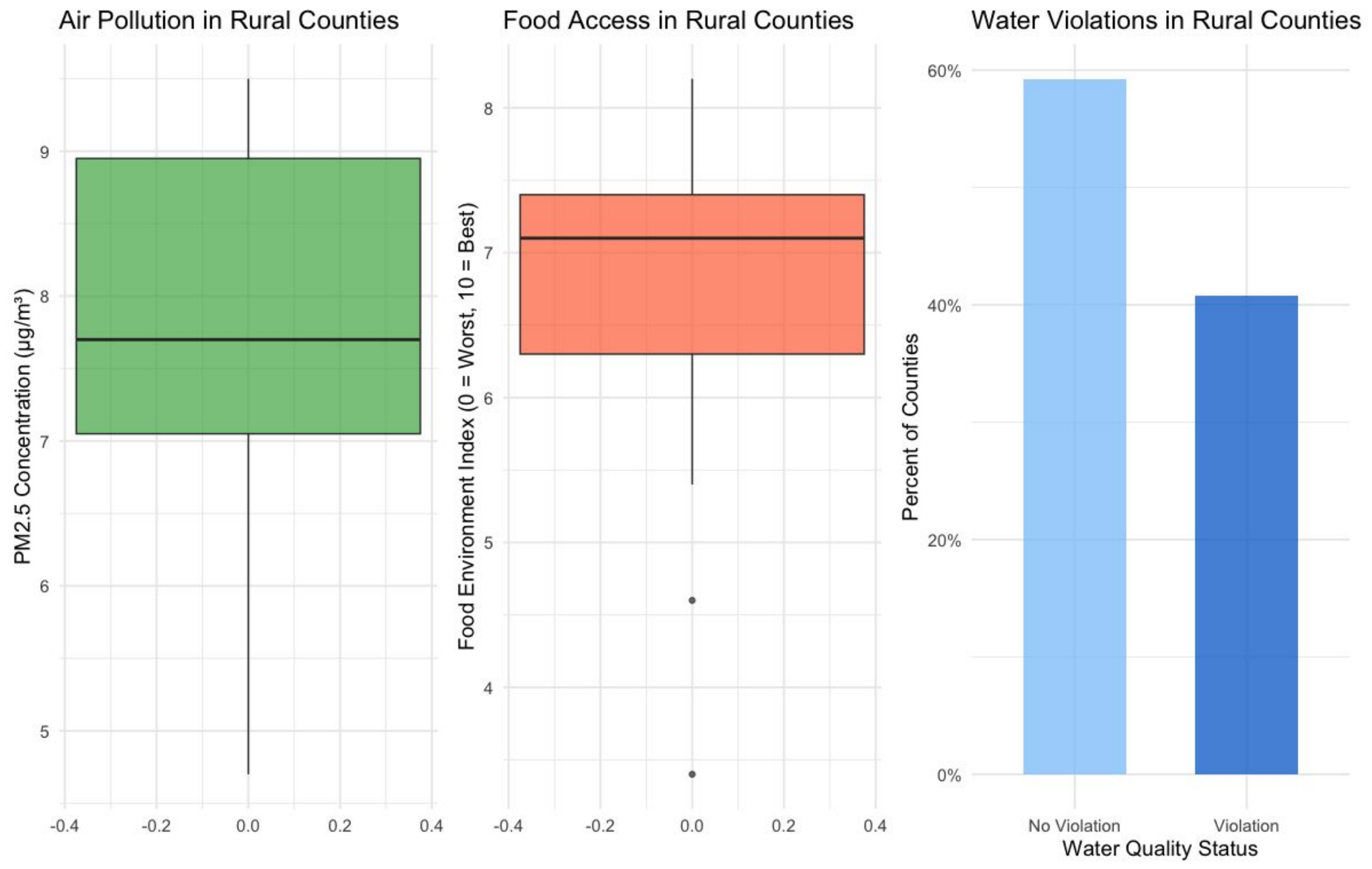

Figure 4. Environmental Health Challenges in Rural U.S. Counties

- Rural counties experience significant air pollution levels. This could be due to industrial sites, agriculture-related burning, or proximity to highways.
- The national average for Food Environmental Index is 7. For these rural counties, the plot is left skewed - 50% of the data is below the national average.
- Rural areas may have less monitoring, fewer resources, or less infrastructure to even detect/report violations.