

## INTRODUCTION

Fine particulate matter (PM2.5) is a major air pollutant linked to respiratory and cardiovascular diseases. Understanding what drives its variation across U.S. cities is essential for effective environmental policy.

Using data from the U.S. Environmental Protection Agency (EPA) and complementary federal sources, this study examines how transportation, social, and environmental factors influence PM2.5 levels. Variables include traffic density, green space, education and unemployment rates, temperature, and wind speed.

Through multiple linear regression and random forest modeling, the analysis identifies the top predictors of PM2.5 and highlights how urban, socioeconomic, and climatic factors interact to shape local air quality.

## METHODS

### Exploratory Data Analysis (Figures 1–6):

- PM2.5 distributions (Figure 1) and yearly trends (Figure 2) were visualized to understand pollutant patterns over time. Scatterplots examined relationships with key predictors: traffic density (Figure 3), green space (Figure 4), industrial NOx (Figure 5), and SO2 emissions (Figure 6).

### Multiple Linear Regression (Table 1):

- Used to quantify the direction and magnitude of predictors' effects on PM2.5. Table 1 shows estimated regression coefficients.

### Random Forest Regression (Table 2):

- Captured nonlinear relationships and interactions among predictors. Table 2 displays feature importance rankings.

### Interaction Effects:

- Explored how green space moderates traffic's impact on PM2.5, informed by Figures 3 and 4.

### Model Evaluation:

- Performance assessed via  $R^2$ , RMSE (Root Mean Squared Error), and residual analysis for linear regression, with Random Forest interpreted through feature importance.

## CONCLUSIONS

**Traffic & PM2.5:** Higher traffic density strongly increases PM2.5 concentrations across cities.

**Wind & Green Space:** Greater wind speed and higher urban green coverage reduce PM2.5 levels.

**Socioeconomic Factors:** Higher labor force participation increases PM2.5, while higher education is slightly protective.

**Temperature:** Warmer cities show modestly higher PM2.5, likely due to stagnant air.

**Interaction Effect:** Green space buffers the impact of traffic on pollution.

**Model Performance:** Random Forest ( $R^2 = 0.50$ ) outperformed Linear Regression ( $R^2 = 0.33$ ), confirming key predictors and interactions.

# Chasing Clean Air: Uncovering the Key Drivers of PM2.5

## Azam Shahbaz - Graduating Spring 2029

Professor Bowen Eldridge

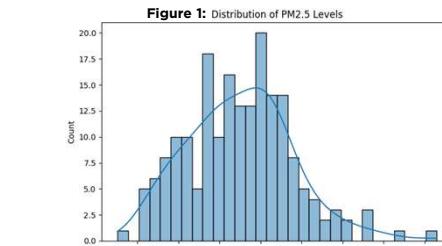


Figure 1: Distribution of PM2.5 Levels

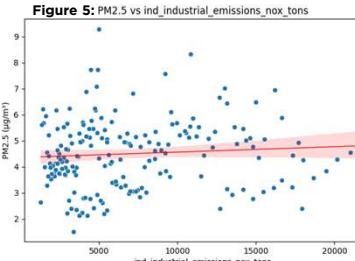


Figure 5: PM2.5 vs ind\_industrial\_emissions\_nox\_tons

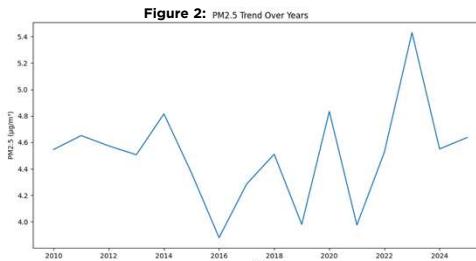


Figure 2: PM2.5 Trend Over Years

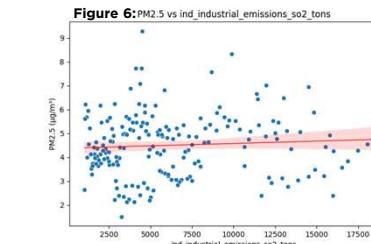


Figure 6: PM2.5 vs ind\_industrial\_emissions\_so2\_tons

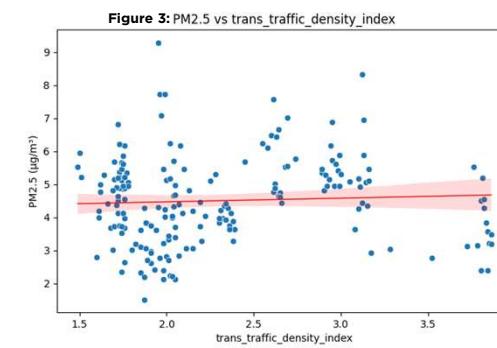


Figure 3: PM2.5 vs trans\_traffic\_density\_index

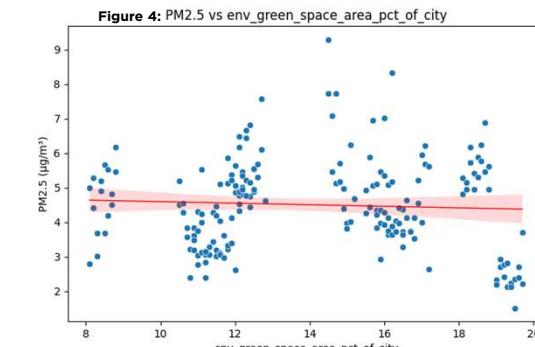


Figure 4: PM2.5 vs env\_green\_space\_area\_pct\_of\_city

Table 1: Linear Regression Performance:  
Train R<sup>2</sup>: 0.33  
Test R<sup>2</sup>: 0.33  
Train RMSE: 0.872  
Test RMSE: 1.030

Predictor	Coefficient
14 trans_traffic_density_index	2.49954e+00
16 env_annual_avg_wind_speed_m/s	1.771402e-01
7 soc_labor_force_participation_rate_pct	5.34113e-01
5 soc_education_hs_degree_plus_pct	4.90466e-01
6 soc_education_bachelors_degree_plus_pct	-4.63915e-01
18 env_green_space_area_pct_of_city	4.39253e-01
19 env_annual_avg_temperature_c	2.903748e-01
8 soc_unemployment_rate_pct	2.283124e-01
0 soc_poverty_rate_pct	2.109482e-01
4 soc_total_population	-7.014842e-02
17 env_annual_total_precipitation_mm	-1.762493e-03
2 soc_population_density	1.262896e-03
12 ind_industrial_emissions_vocs_tons	-1.036266e-03
10 ind_industrial_emissions_nox_tons	5.156711e-04
11 ind_industrial_emissions_so2_tons	3.981475e-04
9 ind_manufacturing_establishments	-1.943964e-04
13 trans_annual_vmt_million_miles	-2.702784e-05
3 soc_median_household_income_usd	-1.962302e-05

Table 2: Random Forest Performance:  
Test R<sup>2</sup>: 0.500  
Test RMSE: 0.890

Predictor	Importance
18 env_green_space_area_pct_of_city	0.195167
5 soc_education_hs_degree_plus_pct	0.111174
8 soc_unemployment_rate_pct	0.095361
7 soc_labor_force_participation_rate_pct	0.082357
17 env_annual_total_precipitation_mm	0.066078
2 soc_population_density	0.053568
13 trans_annual_vmt_million_miles	0.044535
0 soc_total_population	0.033968
3 soc_median_household_income_usd	0.033879
1 soc_total_population	0.031967
16 env_annual_avg_wind_speed_m/s	0.031706
14 trans_traffic_density_index	0.029829
9 ind_manufacturing_establishments	0.026768
10 ind_industrial_emissions_nox_tons	0.022226
11 ind_industrial_emissions_so2_tons	0.020784
4 soc_poverty_rate_pct	0.020735
6 soc_education_bachelors_degree_plus_pct	0.020520
12 ind_industrial_emissions_vocs_tons	0.020400



## RESULTS

### Predictors of PM2.5 Concentration

- **Model Performance:** The Multiple Linear Regression achieved a test  $R^2$  of 0.33 with RMSE = 1.03, while the Random Forest model achieved a test  $R^2$  of 0.50 with RMSE = 0.89, indicating strong predictive ability.
- **Traffic Density:** The traffic density index emerged as the strongest predictor of PM2.5 in both models, with higher traffic density consistently linked to increased PM2.5 levels.
- **Wind Speed:** Average annual wind speed was positively associated with PM2.5 in the linear regression model (coefficient = 1.77), but Random Forest ranked it as the second most important predictor, highlighting its role in pollutant dispersion.
- **Labor Force Participation:** Higher labor force participation rates were associated with elevated PM2.5, suggesting links to commuting and industrial activity.
- **Education:** Percent of population with a high school degree was positively associated with PM2.5, while bachelor's degree attainment was negatively associated, indicating complex socioeconomic interactions.
- **Green Space & Temperature:** Green space area and annual average temperature were moderate contributors, with more green space linked to lower PM2.5 and warmer temperatures associated with higher PM2.5.

### Interaction: Traffic Density x Green Space

- The interaction term between traffic density and green space had a negative coefficient (-0.188), indicating that higher green space mitigates the effect of traffic density on PM2.5 concentrations. Cities with more green coverage experienced weaker traffic-related air pollution increases.
- **Model Diagnostics:**
  - Linear regression residuals showed no extreme outliers, supporting a generally linear relationship between predictors and PM2.5. Q-Q plots indicated slight deviations from normality, but overall fit was reasonable.
  - Random Forest predictions showed low RMSE and high feature importance stability, confirming robustness of the top predictors.

### The top five predictors ranked by combined importance:

1. Traffic Density Index
2. Average Annual Wind Speed
3. Labor Force Participation Rate
4. High School Education (%)
5. Bachelor's Degree (%)

### Future Actions:

- **Electrify Transportation** – Expand electric transit, biking infrastructure, and carpool incentives to reduce traffic-related emissions.
- **Expand Green Space** – Increase urban tree cover and parkland to absorb pollutants and buffer high-traffic areas.
- **Regulate Industrial Emissions** – Tighten NOx and SO<sub>2</sub> controls and encourage cleaner manufacturing processes.
- **Adopt Renewable Energy** – Transition cities and industries toward wind and solar to reduce fossil fuel dependence.
- **Promote Air Quality Policy** – Support data-driven environmental legislation and community awareness programs.

