KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING
School of Data Science and Analytics

Chasing Clean Air: Uncovering the Key Drivers of PM2.5
Azam Shahbaz – Graduating Spring 2029

Professor Bowen Eldridge

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.

This study uses data from the U.S. Environmental Protection Agency (EPA) and complementary federal sources for 12 major U.S. cities spanning 2010–2025. The dataset includes yearly averages of PM2.5 concentration along with key transportation, environmental, and socioeconomic variables.

Predictors include traffic density, green space area, industrial emissions (NOx and SO_2), education and unemployment rates, temperature, and wind speed.

Using multiple linear regression and random forest modeling, the analysis identifies the top predictors of PM2.5 and explores how urban activity, social factors, and climate conditions interact to influence 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.

Figure 1:

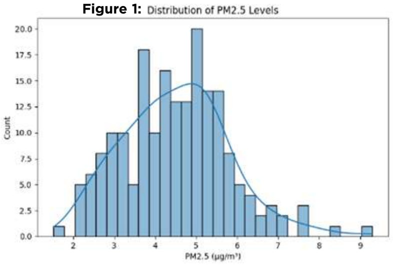


Figure 2:

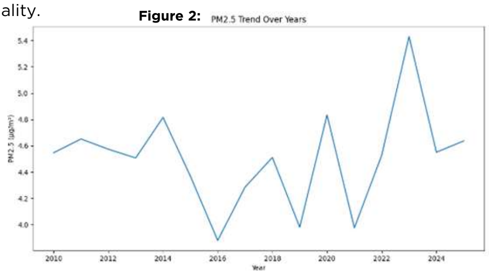


Figure 3: PM2.5 vs trans_traffic_density_index

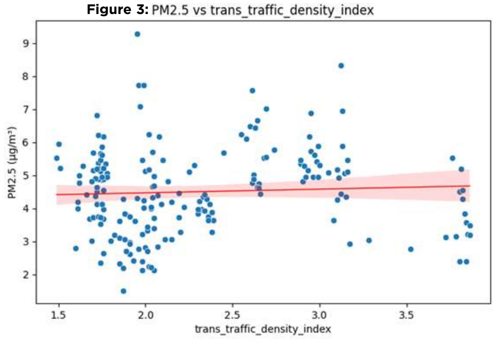


Figure 4: PM2.5 vs env_green_space_area_pct_of_city

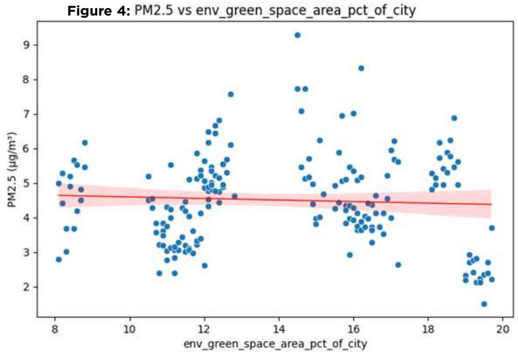


Figure 5: PM2.5 vs ind_industrial_emissions_nox_tons

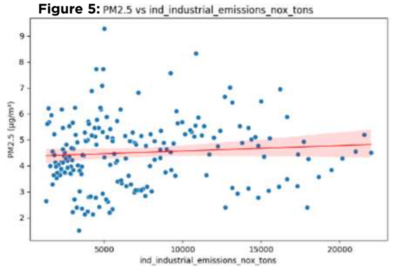


Figure 6: PM2.5 vs ind_industrial_emissions_so2_tons

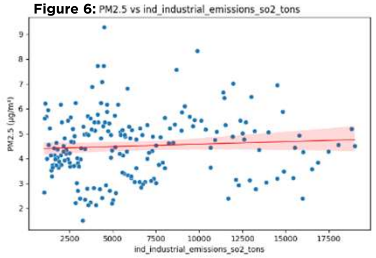
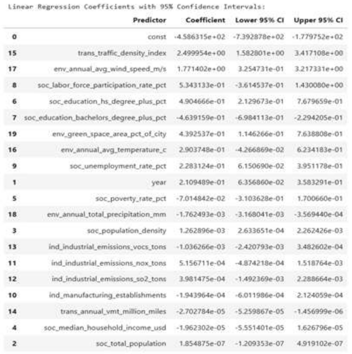


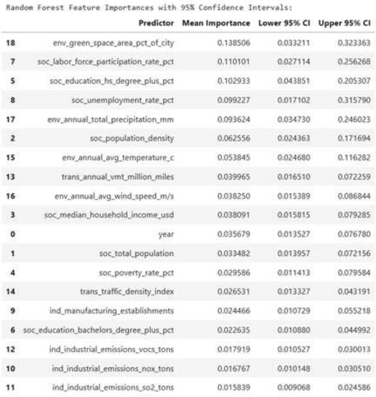
Table 1:



Linear Regression Coefficients with 95% Confidence Intervals:

	Predictor	Coefficient	Lower 95% CI	Upper 95% CI
0	const	-4.586315e+02	-7.392878e+02	-1.779752e+02
15	trans_traffic_density_index	2.499954e+00	1.582801e+00	3.417108e+00
17	env_annual_avg_wind_speed_m/s	1.771402e+00	3.254731e-01	3.217331e+00
8	soc_labor_force_participation_rate_pct	5.343133e-01	-3.614537e-01	1.430080e+00
6	soc_education_hs_degree_plus_pct	4.904666e-01	2.129673e-01	7.679659e-01
7	soc_education_bachelors_degree_plus_pct	-4.639159e-01	-6.984113e-01	-2.294205e-01
19	env_green_space_area_pct_of_city	4.392537e-01	1.146266e-01	7.638808e-01
16	env_annual_avg_temperature_c	2.903748e-01	-4.266869e-02	6.234183e-01
9	soc_unemployment_rate_pct	2.283124e-01	6.150690e-02	3.951178e-01
1	year	2.109489e-01	6.356860e-02	3.583291e-01
5	soc_poverty_rate_pct	-7.014842e-02	-3.103628e-01	1.700660e-01
18	env_annual_total_precipitation_mm	-1.762493e-03	-3.168041e-03	-3.569440e-04
3	soc_population_density	1.262896e-03	2.633651e-04	2.262426e-03
13	ind_industrial_emissions_vocs_tons	-1.036266e-03	-2.420793e-03	3.482602e-04
11	ind_industrial_emissions_nox_tons	5.156711e-04	-4.874218e-04	1.518764e-03
12	ind_industrial_emissions_so2_tons	3.981475e-04	-1.492369e-03	2.288664e-03
10	ind_manufacturing_establishments	-1.943964e-04	-6.011986e-04	2.124059e-04
14	trans_annual_vmt_million_miles	-2.702784e-05	-5.259867e-05	-1.456999e-06
4	soc_median_household_income_usd	-1.962302e-05	-5.551401e-05	1.626796e-05
2	soc_total_population	1.854875e-07	-1.209353e-07	4.919102e-07

Table 2:



Random Forest Feature Importances with 95% Confidence Intervals:

	Predictor	Mean Importance	Lower 95% CI	Upper 95% CI
18	env_green_space_area_pct_of_city	0.138506	0.033211	0.323363
7	soc_labor_force_participation_rate_pct	0.110101	0.027114	0.256268
5	soc_education_hs_degree_plus_pct	0.102933	0.043851	0.205307
8	soc_unemployment_rate_pct	0.099227	0.017102	0.315790
17	env_annual_total_precipitation_mm	0.093624	0.034730	0.246023
2	soc_population_density	0.062556	0.024363	0.171694
15	env_annual_avg_temperature_c	0.053845	0.024680	0.116282
13	trans_annual_vmt_million_miles	0.039965	0.016510	0.072259
16	env_annual_avg_wind_speed_m/s	0.038250	0.015389	0.086844
3	soc_median_household_income_usd	0.038091	0.015815	0.079285
0	year	0.035679	0.013527	0.076780
1	soc_total_population	0.033482	0.013957	0.072156
4	soc_poverty_rate_pct	0.029586	0.011413	0.079584
14	trans_traffic_density_index	0.026531	0.013327	0.043191
9	ind_manufacturing_establishments	0.024466	0.010729	0.055218
6	soc_education_bachelors_degree_plus_pct	0.022635	0.010880	0.044992
12	ind_industrial_emissions_vocs_tons	0.017919	0.010527	0.030013
10	ind_industrial_emissions_nox_tons	0.016767	0.010148	0.030510
11	ind_industrial_emissions_so2_tons	0.015839	0.009068	0.024586

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 × 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_2 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.

LinkedIn

GitHub