Practicum I Final Report

Public Health & Environmental Inequality Analysis Using Multi-Agency Federal Data (2018–2023)

Student: Salman Shareef

Program: Master of Science in Applied Data Science

Course: Practicum I

Institution: University of St. Thomas

Term: Fall 2025

Abstract

This practicum project investigates county-level relationships between air pollution exposure, respiratory health outcomes, and socioeconomic vulnerability by integrating data from three United States federal agencies. Annual fine particulate matter (PM2.5) data were obtained from the Environmental Protection Agency (EPA), asthma prevalence data were sourced from the Centers for Disease Control and Prevention (CDC), and socioeconomic indicators were extracted from the U.S. Census Bureau's American Community Survey (ACS). The analysis focuses on four major metropolitan counties—Los Angeles County (CA), Harris County (TX), Cook County (IL), and Maricopa County (AZ)—from 2018 to 2023.

Exploratory data analysis revealed substantial spatial and temporal variation in pollution exposure and socioeconomic conditions. A cross-sectional regression analysis conducted for 2023 suggested moderate associations between PM2.5, poverty, and asthma prevalence; however, statistical significance was not achieved due to the limited sample size. Despite this limitation, the study demonstrates the feasibility and policy relevance of integrating environmental, health, and socioeconomic datasets to support environmental justice and public health decision-making.

Introduction

Air pollution is a critical environmental determinant of respiratory health. Fine particulate matter (PM2.5), defined as airborne particles with diameters less than or equal to 2.5 micrometers, is particularly harmful because it penetrates deeply into the respiratory system and bloodstream. Numerous epidemiological studies associate long-term PM2.5

exposure with asthma development, exacerbation of chronic respiratory conditions, cardiovascular disease, and premature mortality.

Pollution exposure is not distributed evenly across populations. Lower-income communities and historically marginalized populations tend to experience disproportionately higher levels of pollution, a phenomenon commonly referred to as **environmental injustice**. These inequities compound existing health vulnerabilities and place unique burdens on already disadvantaged communities.

This practicum applies applied data science methods to construct a reproducible, policy-relevant data pipeline that integrates environmental, health, and socioeconomic data at the county level. The objectives of the study are:

1. To evaluate multi-year PM2.5 trends in major U.S. metropolitan counties.
2. To examine changes in median income and poverty over time.
3. To assess the cross-sectional relationship between PM2.5, poverty, and asthma prevalence in 2023.
4. To demonstrate a complete public health data integration and analysis workflow suitable for policy analytics.

Data Sources

Three independent federal datasets were used in this study.

2.1 Environmental Protection Agency (EPA)

- **Dataset:** Air Quality System (AQS)
- **Variable:** Annual average PM2.5 ($\mu g/m^3$)
- **Geographic Level:** County
- **Years:** 2018–2023
- **Access Method:** EPA AQS API

2.2 Centers for Disease Control and Prevention (CDC)

- **Dataset:** Asthma Prevalence Surveillance Data
- **Variable:** Adult asthma prevalence (%)
- **Geographic Level:** County
- **Year:** 2023
- **Access Method:** Public CDC data release

2.3 U.S. Census Bureau

- **Dataset:** American Community Survey (ACS), 5-Year Estimates
- **Variables:**
  - Median household income (USD)
  - Poverty rate (%)
- **Geographic Level:** County
- **Years:** 2018–2023
- **Access Method:** U.S. Census API

Study Area

The analysis focuses on four large, diverse metropolitan counties selected for their high population, economic diversity, and environmental policy relevance:

County	State	FIPS
Los Angeles County	CA	06037
Harris County	TX	48201
Cook County	IL	17031
Maricopa County	AZ	04013

These counties represent distinct geographic regions and environmental conditions within the United States.

Data Engineering & Integration

A reproducible data pipeline was developed using Python. Data were extracted via APIs or curated files, processed, and integrated using the following steps:

1. API-based extraction from EPA and Census servers
2. Manual ingestion of CDC asthma data
3. Standardization of Federal Information Processing Standards (FIPS) codes
4. Year alignment and type harmonization
5. Multi-source dataset merging
6. Duplicate record removal
7. Missing data diagnostics

The final master dataset contains the following variables:

- county_name

- state_fips
- county_fips
- year
- annual_pm25
- asthma_prevalence
- median_household_income
- poverty_percent

This structure supports both longitudinal environmental analyses and cross-sectional health analyses.

Exploratory Data Analysis

5.1 PM2.5 Trends (2018–2023)

PM2.5 levels varied across counties and across years. Los Angeles County consistently exhibited higher PM2.5 concentrations compared with Maricopa County. Harris and Cook Counties displayed intermediate pollution levels. A notable increase in PM2.5 occurred in 2020, likely reflecting wildfire activity and regional atmospheric conditions.

5.2 Socioeconomic Trends

Median household income increased steadily across all four counties between 2018 and 2023. Poverty rates declined during the same period, although meaningful differences persisted across counties. Maricopa County demonstrated the lowest poverty rate by 2023, while Harris County retained higher poverty levels relative to the other study areas.

5.3 Asthma & Pollution (2023)

Cross-sectional analysis of 2023 data revealed that counties with higher PM2.5 levels tended to exhibit higher asthma prevalence. Cook and Harris Counties showed the highest combined PM2.5 exposure and asthma burden, while Maricopa County demonstrated the lowest asthma prevalence.

5.4 Environmental Inequality

When comparing PM2.5 exposure to poverty rates, counties with higher poverty tended to exhibit elevated pollution exposure, reflecting patterns consistent with environmental justice concerns. However, the relationship varied by geographic and regional context.

Statistical Modeling

An ordinary least squares (OLS) regression was estimated for 2023:

Asthma Prevalence = $\beta_0 + \beta_1(\text{PM2.5}) + \beta_2(\text{Poverty})$

Model Results

- **R-squared:** 0.551
- **Number of observations:** 4
- **Predictors:** PM2.5, Poverty Rate
- **Statistical significance:** Not achieved

Interpretation

The model indicates that approximately 55% of the variation in asthma prevalence across counties in 2023 is explained by PM2.5 and poverty. However, neither predictor achieved statistical significance due to the extremely small sample size. These results are considered **exploratory and hypothesis-generating**, rather than confirmatory.

Policy Implications

The results highlight several important implications for environmental health policy:

- Pollution exposure remains uneven across major metropolitan regions.
- Socioeconomic vulnerability is closely linked to environmental risk.
- Integrated federal data systems are essential for identifying environmental justice disparities.
- Policymakers can use similar frameworks to prioritize regulatory and healthcare interventions in high-risk communities.

Limitations

- Asthma prevalence data were only available for a single year.
- The ecological design limits individual-level inference.
- Small sample size restricts statistical power.
- Within-county disparities are not captured.

Conclusion

This practicum demonstrates a complete public health data science workflow that integrates pollution exposure, health outcomes, and socioeconomic vulnerability using national-scale federal data. While statistical significance was constrained by sample size, the project establishes a replicable framework for environmental justice and health policy analytics. The results reinforce the importance of data-driven environmental regulation and targeted public health interventions.

Software & Tools

- Python
- Pandas
- NumPy
- Matplotlib
- Seaborn
- Statsmodels
- EPA AQS API
- U.S. Census Bureau API