Global Air Quality Profiling: Mining Pollution Hotspots and Dominant Contaminant Sources

Sayuri Janbandhu, Vanshika Shah, Bhavana Talkute
Team SVB

***Abstract*—This paper presents a comprehensive data mining analysis of global air quality using a dataset of 23,463 cities. The primary objective was to move beyond traditional monitoring to identify the hidden drivers of PM2.5 pollution. We employed a multi-stage methodology involving K-Means Clustering for regime profiling, Association Rule Mining (Apriori) for pattern discovery, and Random Forest Regression for predictive modeling.**

Our results reveal a critical "Risk Asymmetry": cities with 'Unhealthy' Carbon Monoxide levels show a 100% conditional probability of reaching 'Hazardous' PM2.5 states, whereas cities with 'Unhealthy' traffic emissions (NO2) show a 0% probability of triggering the same hazardous threshold. Additionally, Association Rule Mining confirmed that High PM2.5 is strongly associated with specific geographic clusters in Southern Asia, rather than being a random global occurrence. Our predictive model achieved an accuracy (R^2) of 0.35, proving that chemical composition alone is insufficient for forecasting, highlighting the urgent need for meteorological data integration.

***Index Terms*—Data mining, K-Means Clustering, Association Rule Mining, PM2.5, Random Forest, Risk Probability, Anomaly Detection.**

I. Introduction

Air pollution is a leading global health risk, with PM2.5 (particulate matter less than 2.5 micrometers) posing the most severe threat to human respiratory systems. While traditional policy often focuses on vehicular traffic reduction, recent data suggests the sources are more complex and region-specific.

This project aims to apply advanced data mining techniques to a dataset spanning 170 countries to answer three key questions:

1) What are the distinct "pollution regimes" globally?
2) Is traffic (NO2) truly the main driver of PM2.5, or are we ignoring a bigger threat?
3) Can we quantify the risk of hazardous air quality based on specific precursor pollutants?

II. Methodology

A. Dataset Description

The analysis utilizes a global air quality dataset containing **23,463 records** representing cities across 170 countries. The data serves as a cross-sectional snapshot of air quality for the year 2024. Key features include AQI values for Carbon Monoxide (CO), Ozone (O_3), Nitrogen Dioxide (NO_2), and Particulate Matter (PM2.5).

B. Preprocessing

Data integrity was verified, with no missing values found in the pollutant columns.

- **Feature Engineering:** A new feature, `Dominant_Pollutant`, was created to categorize cities based on which pollutant had the maximum AQI value.
- **Normalization:** To ensure accurate distance calculations for clustering, all numeric pollutant features were scaled using `StandardScaler` to a mean of 0 and variance of 1.
- **Discretization:** Continuous AQI values were binned into categorical labels (Good, Moderate, Unhealthy, Hazardous) to facilitate Association Rule Mining.

C. Models Used

We applied a diverse set of Data Mining techniques:

- **K-Means Clustering:** To segment cities into clusters based on pollution severity ($k = 3$).
- **Association Rule Mining (Apriori):** To discover "If-Then" rules linking specific pollutants to overall air quality states.
- **Conditional Probability (Risk Analysis):** To calculate the probability $P(A|B)$ of Hazardous PM2.5 occurring given specific pollutant conditions.
- **Random Forest Regressor:** To determine Feature Importance and predict PM2.5 levels.
- **Local Outlier Factor (LOF):** To detect anomalies (e.g., wildfires).

III. Results

A. Global Pollution Profiling (Clustering)

Using the Elbow Method, the optimal number of clusters was determined to be $k = 3$. The K-Means algorithm identified three distinct air quality regimes:

- **Cluster 0 (Safe Zone):** Avg AQI $\approx$ 57. Low pollution across all metrics.
- **Cluster 2 (Moderate Zone):** Avg AQI $\approx$ 128. Typical urban pollution.
- **Cluster 1 (Hazardous Zone):** Avg AQI $\approx$ 199. Extremely high PM2.5 levels.

Geographic analysis of Cluster 1 reveals a "Hotspot" concentration in Southern and Eastern Asia, specifically India, China, and Pakistan.

B. Association Rule Mining

We applied the Apriori algorithm to find hidden relationships. Key rules discovered include:

- *Rule 1:* $\{\text{High PM2.5}\} \rightarrow \{\text{Hazardous AQI}\}$ (Confidence: 1.0).
- *Rule 2:* $\{\text{High CO}\} \rightarrow \{\text{High PM2.5}\}$ (Confidence: 0.92).
- *Rule 3:* $\{\text{High NO2}\} \rightarrow \{\text{Moderate PM2.5}\}$ (Confidence: 0.85).

Rule 2 and 3 notably suggest that High CO is a much stronger predictor of severe PM2.5 than High NO2.

C. Risk Probability Analysis

Beyond correlation, we calculated the conditional probability of a city reaching "Hazardous" PM2.5 levels ($AQI > 300$) given that a specific source pollutant was "Unhealthy".

TABLE I
PROBABILITY OF HAZARDOUS PM2.5 BY SOURCE

If Source is 'Unhealthy'...	Probability of Hazardous PM2.5
Carbon Monoxide (Burning)	**100.0%**
Ozone (Smog)	2.6%
Nitrogen Dioxide (Traffic)	**0.0%**

This indicates that while traffic (NO2) contributes to pollution, it almost never triggers "Hazardous" events on its own. In contrast, high CO is a definitive predictor of hazardous air quality.

D. Predictive Modeling & Feature Importance

A Random Forest Regressor was trained to predict PM2.5 AQI values. The Feature Importance analysis ranked the predictors as follows:

1) **Carbon Monoxide (CO):** Impact Score > 0.55 (Primary Predictor).
2) **Ozone (O_3):** Impact Score ≈ 0.35.
3) **Nitrogen Dioxide (NO_2):** Impact Score ≈ 0.10.

This confirms that CO levels are the most reliable indicator of PM2.5 spikes.

IV. DISCUSSION

A. The Dominance of PM2.5

Our analysis shows that PM2.5 is the **Dominant Pollutant** in 18,276 cities (78% of the dataset), compared to only 5,186 for Ozone and just 1 city for NO2. This proves that physical particulate matter is the primary global health threat.

B. The Neglected Factor: Non-Traffic Sources

We identified **19,959 cities** where PM2.5 levels are High, but NO2 levels are Low. Our Risk Analysis confirms that high NO2 carries a **0% risk** of triggering hazardous PM2.5 levels in this dataset. This suggests that policies focusing on "Odd-Even" car bans are targeting a pollutant (NO2) that is not responsible for the most dangerous toxicity spikes. The real neglected factor is **Biomass Burning and Solid Fuels**, indicated by the 100% risk correlation with CO.

C. Anomaly Detection

The LOF algorithm successfully identified critical outliers, such as Durango (USA) with an AQI of 500. Such anomalies indicate specific events (likely wildfires) rather than chronic urban pollution, demonstrating the model's utility for emergency response systems.

D. Limitations

The Random Forest model achieved an R^2 score of 0.35. This low predictive accuracy is a significant finding: it proves that chemical data alone explains only 35% of the variance in pollution. The missing 65% is attributed to meteorological data (Wind Speed, Humidity, Precipitation) which was absent from this dataset.

V. CONCLUSION

This project applied data mining to uncover the hidden dynamics of global air quality. We conclude that:

- **Combustion over Cars:** With a **100% conditional probability** of hazardous air, CO (burning) is the critical target for reduction, not NO2 (traffic).
- **Regional Hotspots:** Hazardous pollution is heavily concentrated in specific Asian regions, requiring targeted geopolitical interventions.
- **Sensor Upgrades:** Future monitoring networks must integrate meteorological sensors to improve prediction accuracy beyond the current 35% ceiling.

REFERENCES

APPENDIX

This appendix provides details regarding the project's GitHub repository and associated file structure for reproducibility and open-source collaboration.

A. Repository Link

The complete source code, datasets, trained models, and documentation are available at:

https://github.com/yureiblack/Global-Air-Quality

B. Repository Structure

The project repository follows the structure shown below:

C. Description of Key Components

- **data/**: Contains raw, processed, and external datasets.
- **notebooks/**: Jupyter Notebooks for EDA, preprocessing, model training, and evaluation.
- **src/**: Modular Python source code for reproducible experiments.
- **models/**: Trained models and intermediate outputs.
- **results/**: Final plots, tables, and exported evaluation artifacts.
- **requirements.txt**: Exact Python dependencies.
- **README.md**: Instructions for running and reproducing results.