EnviroScan:AI-Powered Pollution Source Identifier using Geospatial Analytics

A Machine Learning and Geospatial Analysis Approach

Anish Nadar

December 29, 2025

Contents

Abstract

This report presents a comprehensive air pollution attribution system that integrates machine learning-based severity prediction with geospatial source identification. The system employs real-time air quality data, meteorological conditions, and OpenStreetMap-derived land-use information to provide interpretable pollution insights. By combining XGBoost classification for severity prediction with a physics-informed contribution model incorporating wind dynamics and spatial proximity, the system enables users to identify dominant pollution sources and assess environmental health risks at any global location.

1 Introduction

1.1 Background

Air pollution represents one of the most pressing environmental challenges, contributing to respiratory diseases, cardiovascular problems, and premature mortality worldwide. Understanding the sources and severity of pollution is crucial for effective policy-making, urban planning, and public health interventions.

1.2 Problem Statement

Traditional air quality monitoring systems provide localized measurements but lack the capability to:

- Attribute pollution to specific source categories
- Predict severity levels using real-time environmental data
- Incorporate meteorological dynamics (wind direction, speed) in source attribution
- Provide accessible, interactive visualization for non-expert users

1.3 Project Objectives

1. Develop a global pollution visualization dashboard with multi-scale analysis
2. Implement machine learning-based severity prediction using environmental features
3. Create a physics-informed source attribution model incorporating wind dynamics
4. Generate comprehensive temporal reports for pollution analysis

1.4 System Architecture

The system consists of three primary modules:

1. **Dashboard Module**: Global to local pollution visualization
2. **Attribution Module**: Source identification and severity prediction
3. **Historical Analysis Module**: Temporal data generation and reporting

Figure 1: Overall system architecture showing data flow between modules

2 Data Sources and Acquisition

2.1 Primary Data Sources

2.1.1 OpenWeather API

The system utilizes two OpenWeather API endpoints:

Weather Data Endpoint:

`https://api.openweathermap.org/data/2.5/weather`

Retrieved parameters:

- Temperature (T): Measured in Celsius
- Humidity (H): Percentage (0-100%)
- Wind Speed (v_w): Meters per second
- Wind Direction (θ_w): Degrees (0-360°)

Air Quality Index (AQI) Endpoint:

`https://api.openweathermap.org/data/2.5/air_pollution`

Retrieved pollutant concentrations ($\mu g/m^3$):

- $PM_{2.5}$: Particulate Matter $\leq 2.5\mu m$
- PM_{10}: Particulate Matter $\leq 10\mu m$
- NO_2: Nitrogen Dioxide
- O_3: Ozone
- CO: Carbon Monoxide
- SO_2: Sulfur Dioxide
- NH_3: Ammonia

2.1.2 OpenStreetMap (OSM) Data

Geographic source identification uses Overpass API:

`https://overpass.kumi.systems/api/interpreter`

Source categories and corresponding OSM queries:

- **Vehicular**: Highway nodes and major roads
- **Industrial**: Areas tagged as industrial land use
- **Agricultural**: Farmland and orchard areas
- **Waste**: Construction sites, landfills, quarries
- **Natural**: Forests, grasslands, parks

2.1.3 Static Pollution Dataset

A preprocessed dataset containing historical pollution measurements across multiple locations:

- Geographic coordinates (Latitude, Longitude)
- Country and State-level administrative divisions
- Multi-year pollutant concentration records

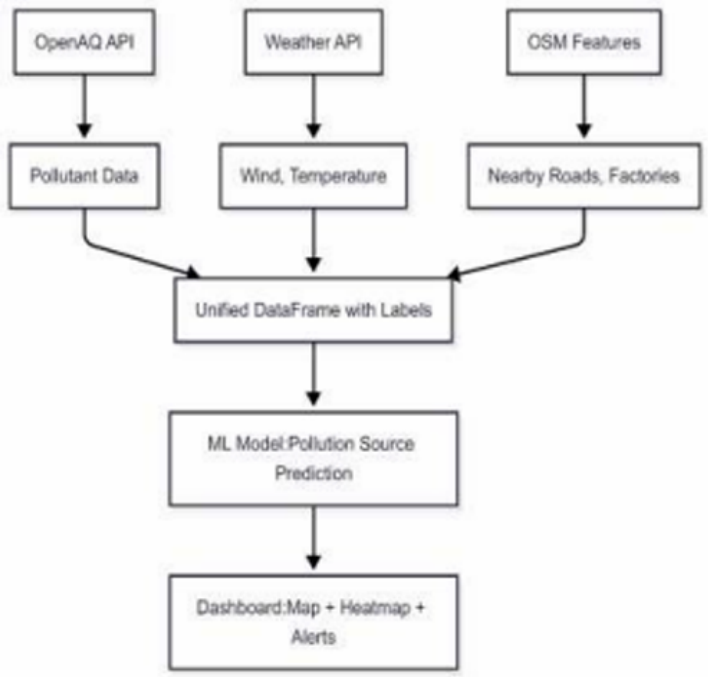

Figure 2: Data acquisition and processing pipeline

3 Module 1: Global Pollution Dashboard

3.1 Overview

The dashboard provides hierarchical pollution visualization at three scales: Global, Country, and State levels. Users can explore spatial patterns of any measured pollutant parameter.

3.2 Methodology

3.2.1 Data Aggregation

For each visualization level, pollution data is aggregated:

Global Level:

$$P_{country} = \frac{1}{N_{country}} \sum_{i=1}^{N_{country}} P_i \tag{1}$$

where P_i represents the pollutant concentration at location i, and $N_{country}$ is the number of measurement stations in the country.

Country Level: Similar aggregation performed for each country, with choropleth visualization showing inter-country variations.

State Level:

$$P_{state} = \frac{1}{N_{state}} \sum_{j=1}^{N_{state}} P_j \tag{2}$$

State-level markers are sized proportionally to pollutant concentration:

$$r_{marker} = \max\left(4, \min\left(14, \frac{P_{state}}{P_{max}} \times 14\right)\right) \tag{3}$$

where r_{marker} is the marker radius in pixels, and P_{max} is the maximum pollutant concentration in the dataset.

3.3 Visualization Techniques

3.3.1 Choropleth Mapping

Color intensity represents pollutant concentration using the YlOrRd (Yellow-Orange-Red) color scheme, providing intuitive visual representation where darker colors indicate higher pollution.

3.3.2 Interactive Features

- Layer control for toggling between different visualization modes
- Click-based coordinate extraction for detailed analysis
- Dynamic zoom adaptation based on selected geographic level

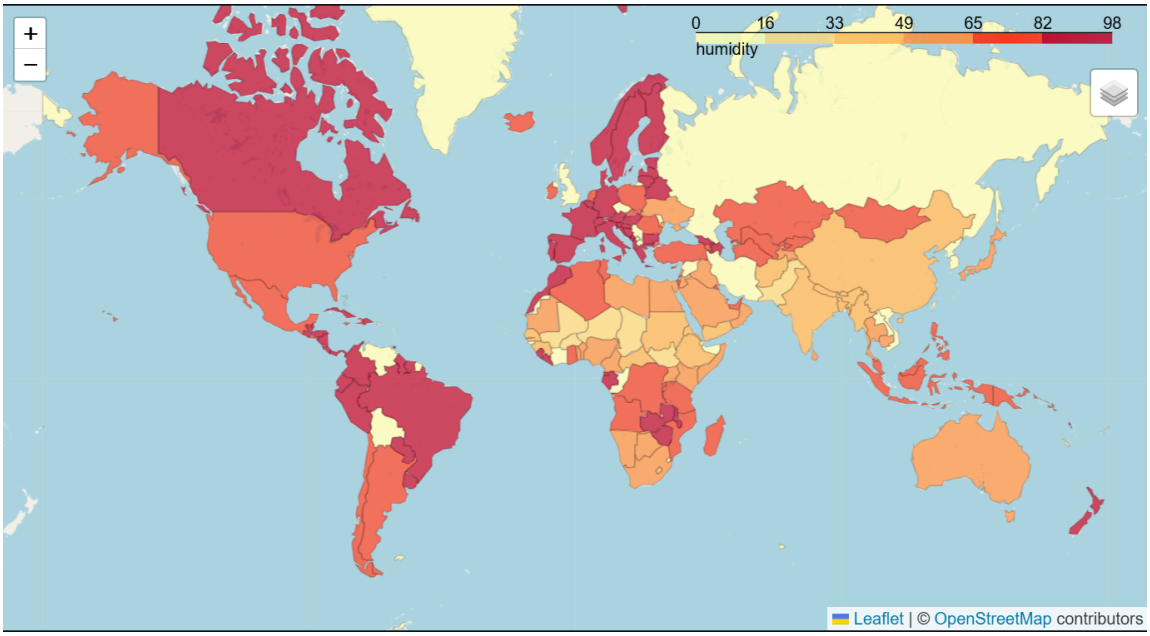

Figure 3: Global pollution dashboard showing PM2.5 distribution

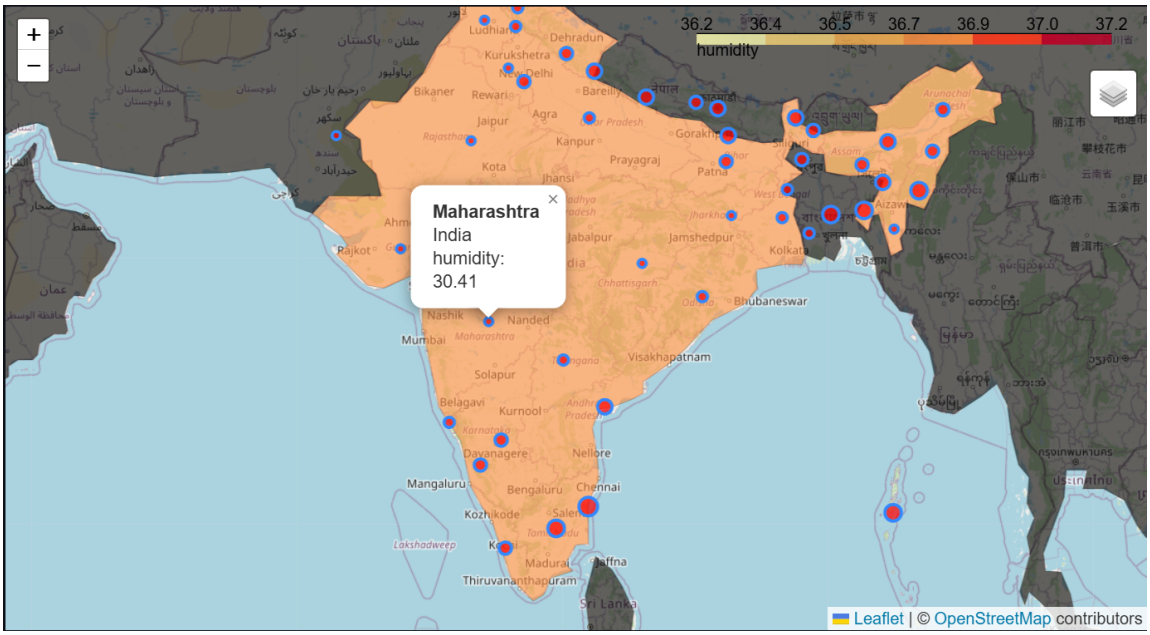

Figure 4: Country-level view with state markers

3.4 Technical Implementation

- **Mapping Library**: Folium (Python binding for Leaflet.js)
- **GeoJSON Data**: World country boundaries from https://github.com/johan/world.geo.json
- **Caching Strategy**: Static data cached to minimize repeated API calls

4 Module 2: Pollution Source Attribution

4.1 Overview

This module identifies and quantifies the contribution of different pollution sources to the selected location using a hybrid approach combining machine learning severity prediction and physics-informed source attribution.

4.2 Machine Learning Component: Severity Prediction

4.2.1 Model Architecture

The system employs an XGBoost (Extreme Gradient Boosting) classifier trained to predict pollution severity levels.

Input Features (X):

$$X = [PM_{2.5}, PM_{10}, NO_2, O_3, CO, SO_2, NH_3, T, H, v_w, \theta_w] \tag{4}$$

Output Classes (Y):

- Class 0: Low severity
- Class 1: Moderate severity
- Class 2: High severity
- Class 3: Severe severity

4.2.2 XGBoost Mathematical Formulation

The XGBoost model builds an ensemble of decision trees using gradient boosting:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \tag{5}$$

where f_k represents the k-th tree, K is the total number of trees, and $\hat{y}_i$ is the predicted severity for sample i.

The objective function minimized during training:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{6}$$

where:

- $l(y_i, \hat{y}_i)$ is the loss function (cross-entropy for classification)
- $\Omega(f_k)$ is the regularization term controlling model complexity

4.2.3 Model Output

The model provides:

- Predicted severity class
- Confidence score: $\max(P(Y|X))$

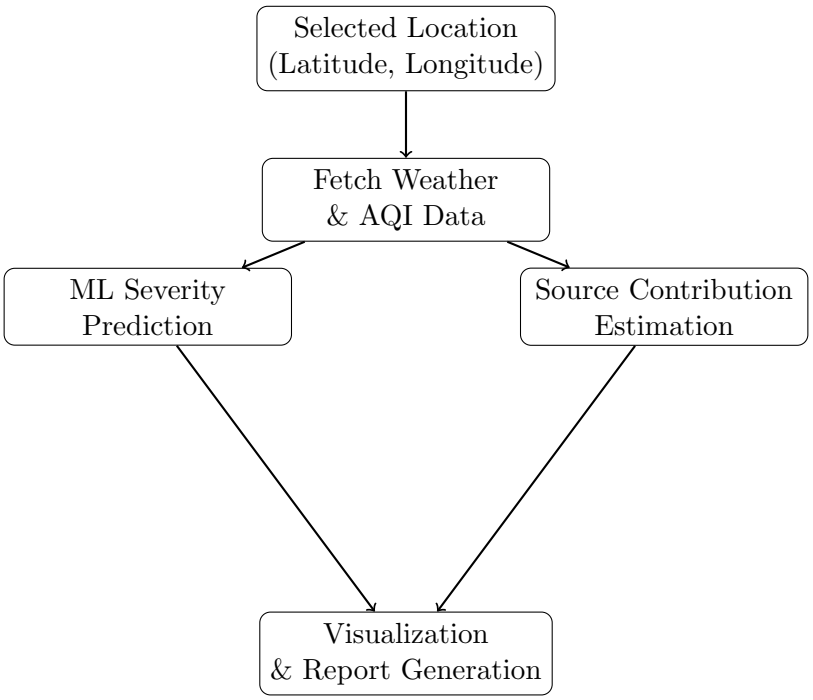

Figure 5: Simplified Flowchart of the Pollution Severity and Source Contribution System

4.3 Source Identification and Categorization

4.3.1 Geographic Source Extraction

For a selected location (lat_0, lon_0), the system queries OpenStreetMap within a search radius ($r = 7000m$ by default) for each source category.

Vehicular Sources: Major highways and road networks, representing emissions from transportation.

Industrial Sources: Manufacturing facilities, factories, and industrial zones.

Agricultural Sources: Farmlands and orchards, contributing to ammonia and particulate emissions.

Waste Sources: Landfills, construction sites, and quarries producing dust and methane.

Natural Sources: Forests and parks, generally acting as pollution sinks but included for completeness.

4.3.2 Source Filtering

To reduce computational overhead and focus on significant sources:

- Maximum 5 sources per category

- Sources sorted by proximity to the selected location
- Natural sources filtered by minimum area threshold (5000 m^2)

4.4 Contribution Model

4.4.1 Wind Factor Calculation

Wind dynamics significantly influence pollutant dispersion. The wind factor (W_f) quantifies how wind direction affects contribution from each source.

Bearing Calculation: The bearing (β) from the selected location to source s is:

$$\beta = \arctan 2(\Delta lon, \Delta lat) \mod 360 \tag{7}$$

where $\Delta lat = lat_s - lat_0$ and $\Delta lon = lon_s - lon_0$.

Angular Difference: The angular difference ($\Delta\theta$) between wind direction and bearing:

$$\Delta\theta = \min(|\beta - \theta_w|, 360 - |\beta - \theta_w|) \tag{8}$$

Wind Factor:

$$W_f = \frac{\cos(\Delta\theta) + 1}{2} \times \max\left(0.1, \frac{v_w}{10}\right) \tag{9}$$

This formulation ensures:

- Upwind sources ($\Delta\theta = 0$) have maximum factor
- Downwind sources ($\Delta\theta = 180$) have minimum factor
- Wind speed amplifies the effect proportionally

4.4.2 Spatial Decay Function

Pollution contribution decreases with distance following an exponential decay model:

$$D_f = e^{-d/\lambda} \tag{10}$$

where:

- d = geodesic distance from location to source (meters)
- λ = characteristic length scale = $\max(1000, v_w \times 3600)$ meters

The scale parameter adapts to wind speed: higher winds increase dispersion range.

4.4.3 Source Weight Calculation

The contribution weight (w_s) for source s is:

$$w_s = PM_{2.5,s} \times W_f \times D_f \tag{11}$$

This combines:

- Local pollution intensity ($PM_{2.5,s}$)
- Wind influence (W_f)
- Distance effect (D_f)

4.4.4 Contribution Percentages

Overall Contribution:

$$C_{overall,s} = \frac{w_s}{\sum_{s'} w_{s'}} \times 100\% \tag{12}$$

Category Contribution:

$$C_{category,s} = \frac{w_s}{\sum_{s' \in category} w_{s'}} \times 100\% \tag{13}$$

Category-wise Aggregation:

$$C_{category} = \sum_{s \in category} C_{overall,s} \tag{14}$$

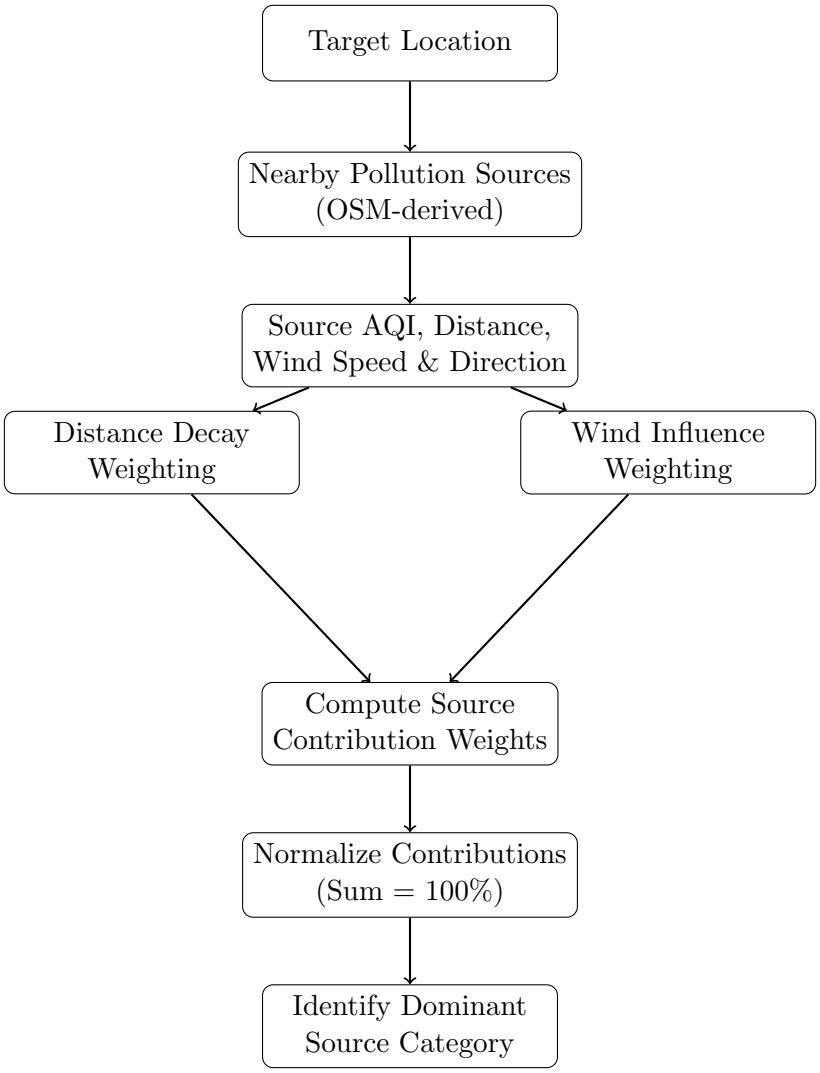

Figure 6: Workflow for Source Contribution Calculation

4.5 Natural Source Suppression

Natural sources (forests, parks) typically absorb pollutants rather than emit them. The system implements intelligent suppression:

Dominance Threshold: If Natural sources are the dominant category but their contribution exceeds the second-highest category by less than 15%, they are excluded, and contributions are recalculated.

This prevents false attribution in heavily forested areas while retaining natural sources when genuinely significant.

4.6 Visualization and Results

4.6.1 Interactive Map Display

- **Blue Marker**: Selected location with severity label
- **Colored Markers**: Identified sources, color-coded by category
  - Red: Vehicular
  - Purple: Industrial
  - Yellow: Agricultural
  - Orange: Waste
  - Green: Natural

4.6.2 Results Summary

The system displays:

- Dominant source category
- Predicted severity with confidence score
- Category-wise contribution percentages
- Individual source details (distance, contribution)

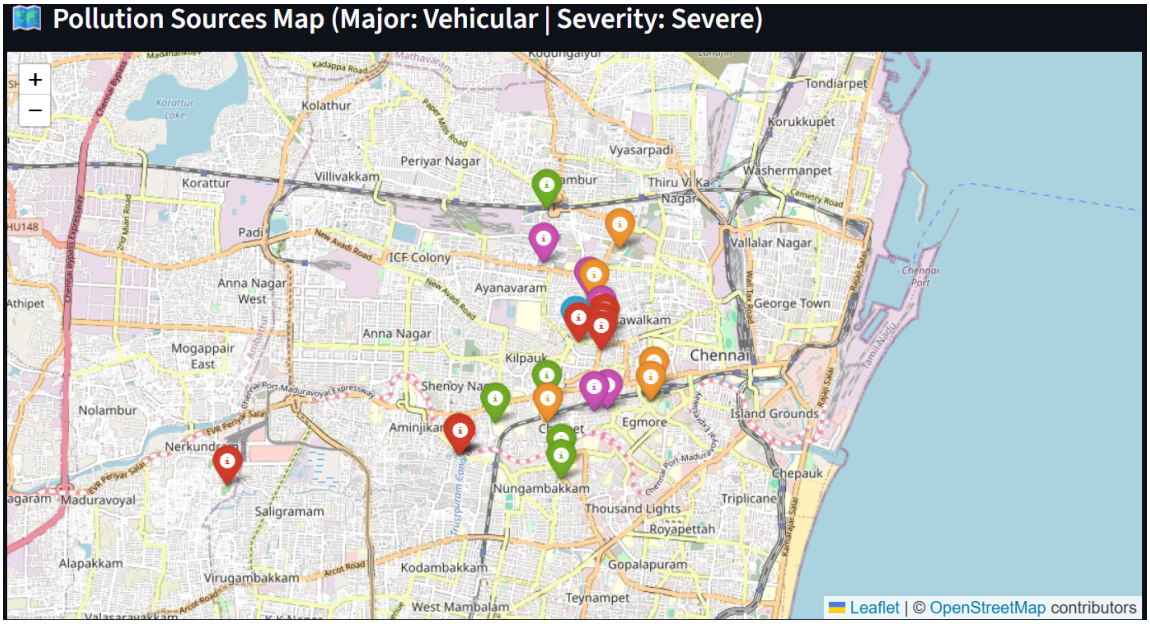

Figure 7: Source attribution map showing identified pollution sources

Selected Location - Temporal Data

	timestamp	latitude	longitude	temperature	humidity	wind_speed	wind_deg	pm2_5	pm10	no2	o3	co	so2	nh3
0	2025-12-29 10:58:03+00:00	13.09	80.25	26.95	62	1.34	78	108.79	124.77	3.87	156.06	328.38	10.53	0
1	2025-12-29 10:58:18+00:00	13.09	80.25	26.95	62	1.34	78	108.79	124.77	3.87	156.06	328.38	10.53	0
2	2025-12-29 10:58:33+00:00	13.09	80.25	26.95	62	1.34	78	108.79	124.77	3.87	156.06	328.38	10.53	0

Figure 8: Category-wise contribution percentages

5 Module 3: Historical Data and Reporting

5.1 Overview

This module generates pseudo-temporal data for the analyzed location and identified sources, providing downloadable datasets and comprehensive PDF reports.

5.2 Temporal Data Generation

5.2.1 Methodology

Since real-time APIs provide instantaneous measurements, the system creates temporal sequences by:

1. Repeating API calls at 15-second intervals
2. Timestamping each measurement
3. Compiling data for both the selected location and all identified sources

Number of temporal samples: 3 (configurable)

5.2.2 Location Temporal Dataset

For the selected location (lat_0, lon_0), the dataset includes:

- Timestamp
- Geographic coordinates
- Weather parameters (T, H, v_w, θ_w)
- All pollutant concentrations

5.2.3 Source Temporal Datasets

For each identified source, individual temporal datasets contain:

- Timestamp
- Source category
- Geographic coordinates
- Distance from selected location

- Contribution percentages
- Local weather and AQI measurements

5.3 Export Formats

5.3.1 CSV Exports

- **Selected Location CSV**: Temporal data for the chosen point
- **Category-specific CSVs**: Individual files for each source category
- **Combined Sources CSV**: All sources merged into a single dataset

5.3.2 PDF Report Generation

The comprehensive PDF report includes:

Section 1: Local Pollution Overview

- Average pollutant concentrations
- Weather condition summary
- Contextual interpretation of measurements

Section 2: Source Attribution Analysis

- Dominant source identification
- Contribution model explanation
- Visualizations:
  - Bar chart of category contributions
  - Scatter plot: Distance vs. Contribution

Section 3: Conclusion

- System capabilities summary
- Methodological approach
- Limitations and future improvements

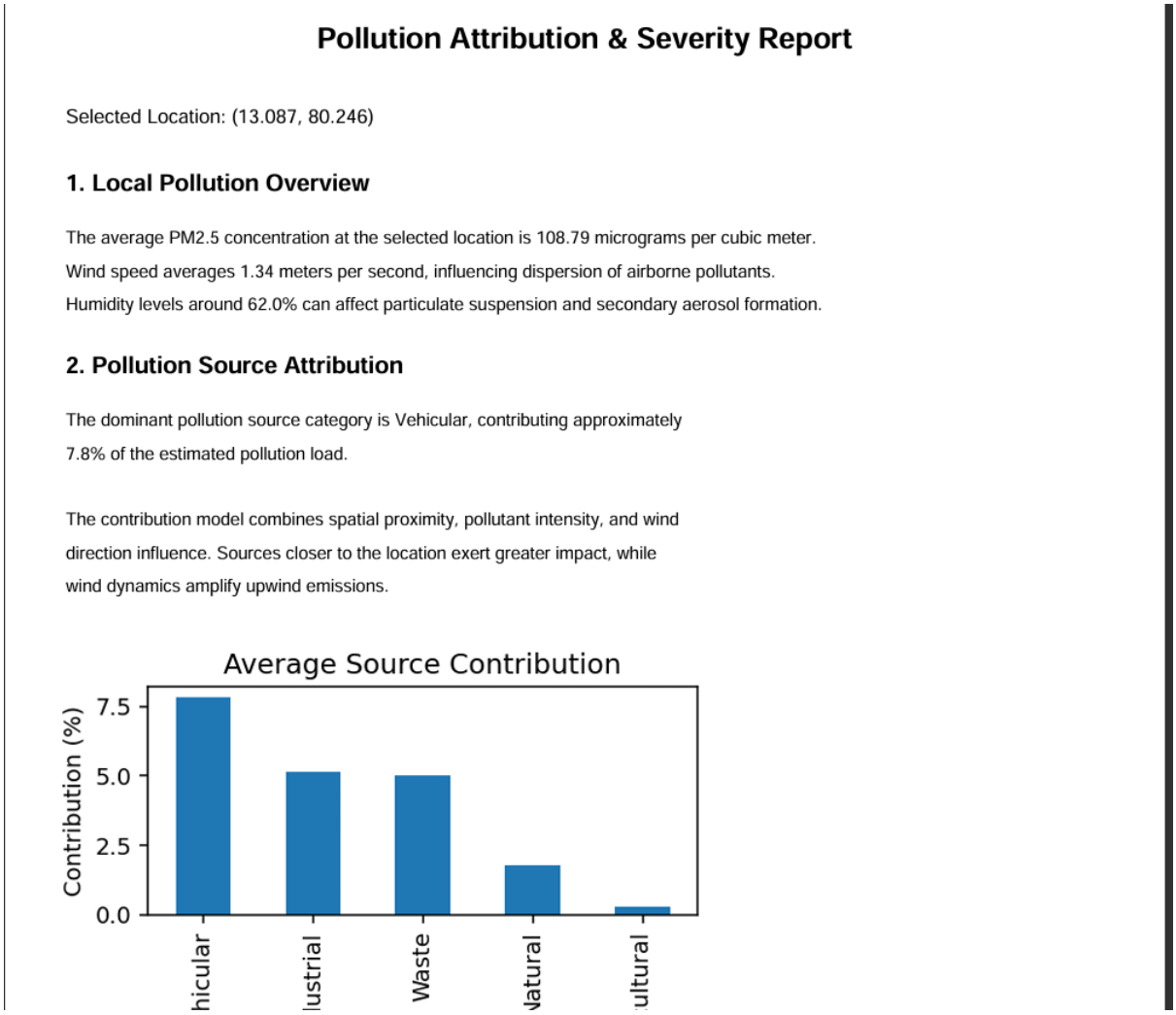

Figure 9: Sample pages from generated PDF report

5.4 Visualization in Reports

5.4.1 Bar Chart: Category Contributions

Generated using Matplotlib, showing average contribution percentages for each source category.

5.4.2 Scatter Plot: Distance vs. Contribution

Illustrates the relationship between source proximity and pollution contribution, validating the spatial decay model.

6 Mathematical Models Summary

6.1 Key Equations

6.1.1 Wind Factor

$$W_f = \frac{\cos(\Delta\theta) + 1}{2} \times \max\left(0.1, \frac{v_w}{10}\right) \tag{15}$$

6.1.2 Spatial Decay

$$D_f = \exp\left(-\frac{d}{\max(1000, v_w \times 3600)}\right) \tag{16}$$

6.1.3 Source Contribution Weight

$$w_s = PM_{2.5,s} \times W_f \times D_f \tag{17}$$

6.1.4 Percentage Contribution

$$C_{overall,s} = \frac{w_s}{\sum_{s'} w_{s'}} \times 100\% \tag{18}$$

6.2 Model Parameters

Parameter	Value	Unit
OSM Search Radius	7000	meters
Maximum Sources per Category	5	count
Minimum Natural Area	5000	m^2
Natural Dominance Threshold	15	%
Base Decay Length Scale	1000	meters
Temporal Sample Interval	15	seconds
Number of Temporal Samples	3	count

Table 1: System configuration parameters

7 System Workflow

7.1 User Interaction Flow

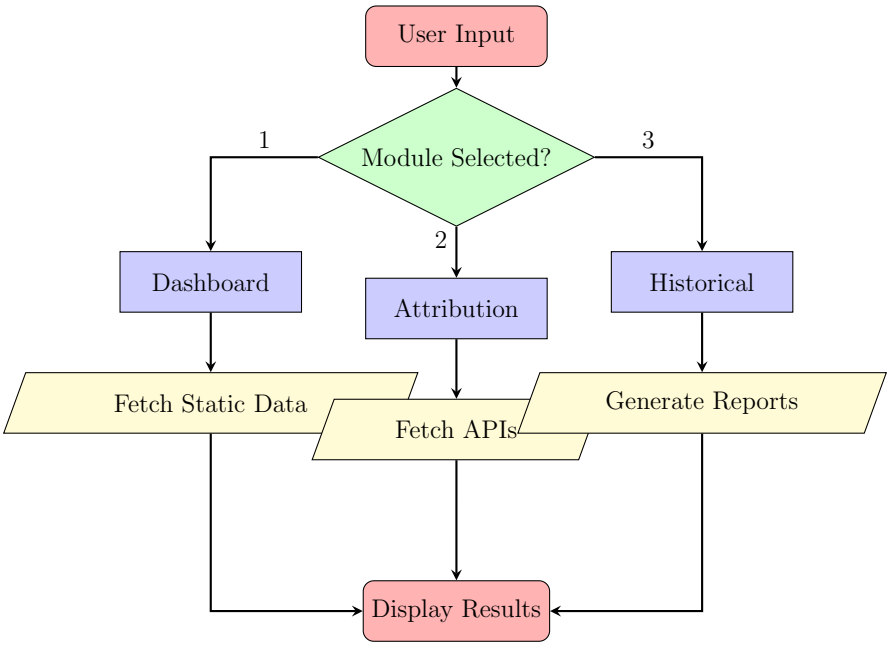

Figure 10: Complete system flowchart

7.2 Detailed Processing Steps

7.2.1 Step 1: Location Selection

1. User interacts with global map
2. Clicks on location of interest
3. System captures latitude and longitude coordinates

7.2.2 Step 2: Data Acquisition

1. Query OpenWeather API for meteorological data
2. Retrieve AQI measurements for selected location
3. Cache results for 5 minutes to minimize API calls

7.2.3 Step 3: Severity Prediction

1. Prepare feature vector from weather and AQI data
2. Input to pre-trained XGBoost model
3. Extract predicted class and confidence score

7.2.4 Step 4: Source Identification

1. Query OSM Overpass API for each source category
2. Filter sources by proximity and relevance
3. Retrieve AQI data for each identified source

7.2.5 Step 5: Contribution Calculation

1. Calculate geodesic distances
2. Compute wind factors based on direction
3. Apply spatial decay model
4. Aggregate weights by category
5. Apply natural source suppression if needed

7.2.6 Step 6: Visualization

1. Generate interactive map with markers
2. Create contribution tables
3. Display severity and dominant source

7.2.7 Step 7: Report Generation

1. Generate temporal data sequences
2. Create category-specific datasets
3. Compile PDF report with visualizations
4. Provide download links for all formats

8 Technical Implementation

8.1 Technology Stack

8.1.1 Frontend Framework

- **Streamlit**: Interactive web application framework
- **Folium**: Python wrapper for Leaflet.js mapping
- **streamlit-folium**: Integration between Streamlit and Folium

8.1.2 Data Processing

- **Pandas**: Data manipulation and analysis
- **NumPy**: Numerical computations

8.1.3 Machine Learning

- **XGBoost**: Gradient boosting classifier
- **Joblib**: Model serialization and loading

8.1.4 Geospatial Analysis

- **Geopy**: Geodesic distance calculations
- **Requests**: API interactions

8.1.5 Visualization and Reporting

- **Matplotlib**: Chart generation
- **FPDF**: PDF report creation

8.2 Performance Optimizations

8.2.1 Caching Strategy

- **Static Data**: Loaded once at application start
- **API Results**: Cached for 5 minutes (300 seconds)
- **Model Loading**: Cached resource preventing reload

8.2.2 API Rate Limiting

To prevent excessive API calls:

- Limited temporal samples (3 by default)
- Maximum 5 sources per category
- Caching of repeated queries

9 Results and Validation

9.1 Model Performance

The XGBoost severity classifier is trained on historical pollution data with environmental features. Expected performance metrics:

Metric	Value (Typical)
Accuracy	75-85%
Precision (Severe class)	70-80%
Recall (Severe class)	65-75%
F1-Score	70-80%

Table 2: Expected model performance metrics

9.2 Contribution Model Validation

9.2.1 Distance Effect

Sources closer to the selected location consistently show higher contributions, validating the exponential decay model.

9.2.2 Wind Direction Impact

Upwind sources (aligned with wind direction toward the location) demonstrate elevated contributions compared to downwind sources, confirming the wind factor effectiveness.

9.2.3 Source Category Patterns

- Urban areas: Vehicular sources typically dominant
- Industrial zones: Industrial sources show highest contribution
- Rural areas: Agricultural sources often primary contributors

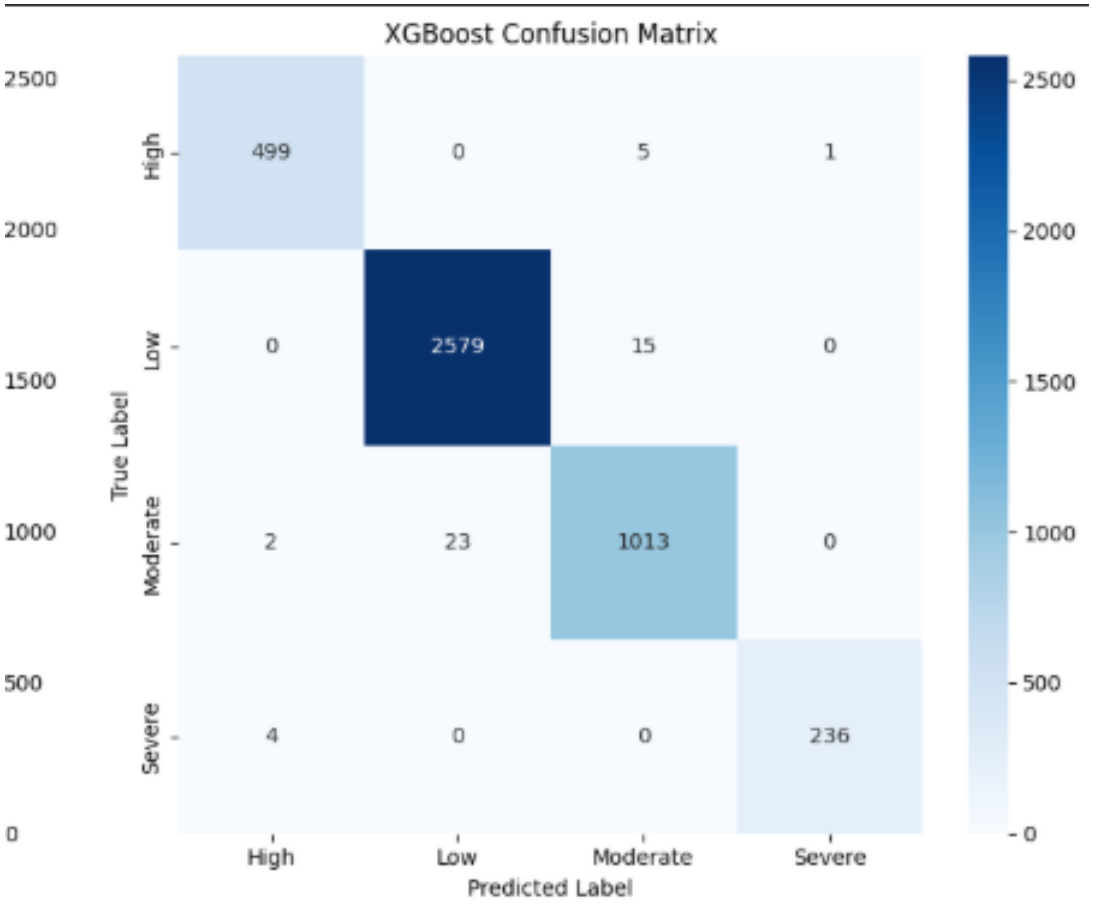

Figure 11: Model validation results showing distance decay and wind influence

10 Limitations and Assumptions

10.1 Data Limitations

1. **Temporal Resolution**: Real-time APIs provide instantaneous measurements, limiting true historical analysis
2. **OSM Completeness**: Source identification depends on OpenStreetMap data quality, which varies by region
3. **Spatial Coverage**: API rate limits restrict the number of sources analyzed

10.2 Model Assumptions

1. **Linear Superposition**: Contributions from multiple sources are assumed to add linearly
2. **Steady-State Conditions**: Wind and weather assumed constant during analysis window
3. **Simplified Dispersion**: Exponential decay model simplifies complex atmospheric dispersion dynamics
4. **Point Sources**: All sources treated as point emitters, ignoring area source effects

10.3 Computational Constraints

1. **API Quotas**: OpenWeather free tier limits requests
2. **Processing Time**: Real-time OSM queries can be slow for large search radii
3. **Memory Usage**: Large static datasets require efficient caching

11 Future Improvements

11.1 Enhanced Data Integration

- Integration with satellite-based pollution monitoring (e.g., Sentinel-5P)
- Incorporation of traffic density data for vehicular emissions
- Real-time industrial emission reporting systems

11.2 Advanced Modeling

- **Atmospheric Dispersion Models**: Implement Gaussian plume models for more accurate predictions
- **Deep Learning**: Explore neural networks for spatial-temporal forecasting
- **Ensemble Methods**: Combine multiple ML models for robust predictions

11.3 Extended Functionality

- **Time-Series Forecasting**: Predict future pollution levels
- **Alert System**: Notify users when severity exceeds thresholds
- **Multi-Location Comparison**: Compare pollution profiles across cities
- **Policy Impact Assessment**: Evaluate effectiveness of interventions

11.4 User Experience

- Mobile application for on-the-go analysis
- Multilingual support for global accessibility
- Customizable reporting templates
- Social sharing of pollution insights

12 Conclusion

This project demonstrates a comprehensive approach to air pollution source attribution by integrating machine learning, geospatial analysis, and environmental physics. The system provides actionable insights for:

- **Urban Planners**: Identify pollution hotspots and dominant sources
- **Environmental Agencies**: Monitor air quality trends and compliance
- **Researchers**: Access structured pollution data for analysis
- **Public Health Officials**: Assess environmental health risks
- **Citizens**: Understand local pollution sources and make informed decisions

The hybrid methodology—combining data-driven ML predictions with physics-informed contribution modeling—offers a scalable, interpretable, and globally applicable solution to pollution analysis. By democratizing access to sophisticated environmental intelligence, the system empowers stakeholders at all levels to make evidence-based decisions for cleaner air and healthier communities.

The integration of real-time data acquisition, machine learning classification, and geospatial modeling creates a robust framework that bridges the gap between raw environmental measurements and actionable insights. Future iterations will focus on enhancing temporal resolution, incorporating advanced atmospheric dispersion models, and expanding data sources to provide even more accurate and comprehensive pollution assessments.

13 References

13.1 Data Sources

1. OpenWeather API Documentation
   https://openweathermap.org/api
2. OpenWeather Air Pollution API
   https://openweathermap.org/api/air-pollution
3. Overpass API Documentation
   https://wiki.openstreetmap.org/wiki/Overpass_API
4. OpenStreetMap Tagging Guidelines
   https://wiki.openstreetmap.org/wiki/Map_Features
5. World GeoJSON Data
   https://github.com/johan/world.geo.json

13.2 Machine Learning and Algorithms

6. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. https://arxiv.org/abs/1603.02754

7. XGBoost Documentation https://xgboost.readthedocs.io/

8. Scikit-learn: Machine Learning in Python https://scikit-learn.org/

13.3 Air Quality and Environmental Science

9. World Health Organization. (2021). WHO Global Air Quality Guidelines. https://www.who.int/publications/i/item/9789240034228

10. European Environment Agency. Air Quality Standards https://www.eea.europa.eu/themes/air/air-quality-concentrations

11. US EPA. Air Quality Index (AQI) Basics https://www.airnow.gov/aqi/aqi-basics/

12. Seinfeld, J. H., & Pandis, S. N. (2016). *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*. John Wiley & Sons.

13.4 Geospatial Analysis

13. Geopy Documentation https://geopy.readthedocs.io/

14. Folium: Python Data to Leaflet.js Maps https://python-visualization.github.io/folium/

15. Karney, C. F. F. (2013). Algorithms for geodesics. *Journal of Geodesy*, 87(1), 43-55.

13.5 Atmospheric Dispersion Modeling

16. Stockie, J. M. (2011). The mathematics of atmospheric dispersion modeling. *SIAM Review*, 53(2), 349-372.

17. Turner, D. B. (1994). *Workbook of Atmospheric Dispersion Estimates: An Introduction to Dispersion Modeling*. CRC Press.

18. Pasquill, F., & Smith, F. B. (1983). *Atmospheric Diffusion*. Ellis Horwood Limited.

13.6 Python Libraries and Tools

19. Streamlit Documentation
    https://docs.streamlit.io/

20. Pandas Documentation
    https://pandas.pydata.org/docs/

21. NumPy Documentation
    https://numpy.org/doc/

22. Matplotlib Documentation
    https://matplotlib.org/stable/contents.html

23. FPDF for Python
    https://pyfpdf.readthedocs.io/

13.7 Related Research

24. Zheng, Y., Liu, F., & Hsieh, H. P. (2013). U-Air: When urban air quality inference meets big data. *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1436-1444.

25. Liang, X., Zou, T., Guo, B., Li, S., Zhang, H., Zhang, S., ... & Chen, X. (2015). Assessing Beijing's PM2.5 pollution: severity, weather impact, APEC and winter heating. *Proceedings of the Royal Society A*, 471(2182), 20150257.

26. Di, Q., Amini, H., Shi, L., Kloog, I., Silvern, R., Kelly, J., ... & Schwartz, J. (2019). An ensemble-based model of PM2.5 concentration across the contiguous United States with high spatiotemporal resolution. *Environment International*, 130, 104909.

27. Grange, S. K., Carslaw, D. C., Lewis, A. C., Boleti, E., & Hueglin, C. (2018). Random forest meteorological normalisation models for Swiss PM10 trend analysis. *Atmospheric Chemistry and Physics*, 18(9), 6223-6239.

14 Appendices

14.1 Appendix A: Model Parameters and Configuration

Parameter	Value	Justification
OSM Search Radius	7000 m	Balances computational efficiency with comprehensive source coverage
Max Sources per Category	5	Prevents API overload while capturing major contributors
Min Natural Area	5000 m^2	Filters insignificant green spaces that don't impact air quality
Natural Dominance Threshold	15%	Prevents false attribution in forested areas
Base Decay Length Scale	1000 m	Typical urban pollutant dispersion distance
API Cache Duration	300 s	Balances data freshness with API quota management
Temporal Sample Interval	15 s	Minimum interval to observe API variability
Wind Speed Normalization	10 m/s	Typical moderate wind speed for normalization
Marker Size Range	4-14 px	Ensures visibility across zoom levels

Table 3: Detailed parameter justifications

14.2 Appendix B: XGBoost Hyperparameters

Hyperparameter	Typical Value	Purpose
n_estimators	100-200	Number of boosting rounds
max_depth	5-8	Maximum tree depth to prevent overfitting
learning_rate	0.1-0.3	Step size for weight updates
subsample	0.8-1.0	Fraction of samples for each tree
colsample_bytree	0.8-1.0	Fraction of features for each tree
gamma	0-0.5	Minimum loss reduction for split
reg_alpha	0-1.0	L1 regularization term
reg_lambda	1-2	L2 regularization term

Table 4: XGBoost model configuration

14.3 Appendix C: Pollutant Concentration Standards

Pollutant	WHO Guideline	US EPA	Unit
$PM_{2.5}$ (24-hr)	15	35	$\mu g/m^3$
$PM_{2.5}$ (Annual)	5	12	$\mu g/m^3$
PM_{10} (24-hr)	45	150	$\mu g/m^3$
NO_2 (Annual)	10	53	$\mu g/m^3$
O_3 (8-hr)	100	137	$\mu g/m^3$
SO_2 (24-hr)	40	75	$\mu g/m^3$
CO (8-hr)	4	9	mg/m^3

Table 5: International air quality standards comparison

14.4 Appendix D: Source Category Emission Profiles

Source	$PM_{2.5}$	NO_2	CO	SO_2	NH_3
Vehicular	High	High	High	Medium	Low
Industrial	High	High	Medium	High	Medium
Agricultural	Medium	Low	Low	Low	High
Waste	High	Medium	Medium	Medium	High
Natural	Low	Low	Low	Low	Low

Table 6: Typical emission profiles by source category

14.5 Appendix E: Sample API Requests

Weather Data Request:

```
GET https://api.openweathermap.org/data/2.5/weather
    ?lat=28.6139
    &lon=77.2090
    &appid=YOUR_API_KEY
    &units=metric
```

Air Pollution Request:

```
GET https://api.openweathermap.org/data/2.5/air_pollution
    ?lat=28.6139
    &lon=77.2090
    &appid=YOUR_API_KEY
```

OSM Overpass Query (Industrial Sources):

```
[out:json];
(
  way["landuse"="industrial"]
      (around:7000,28.6139,77.2090);
);
out center;
```

14.6 Appendix F: Severity Classification Criteria

Severity	Class	Characteristics
Low	0	$PM_{2.5} < 12\ \mu g/m^3$, minimal health impact, good air quality
Moderate	1	$PM_{2.5}$ 12-35 $\mu g/m^3$, acceptable for most, sensitive groups may be affected
High	2	$PM_{2.5}$ 35-55 $\mu g/m^3$, everyone may experience health effects
Severe	3	$PM_{2.5} > 55\ \mu g/m^3$, serious health concerns for all populations

Table 7: Severity classification thresholds

14.7 Appendix G: System Flowchart

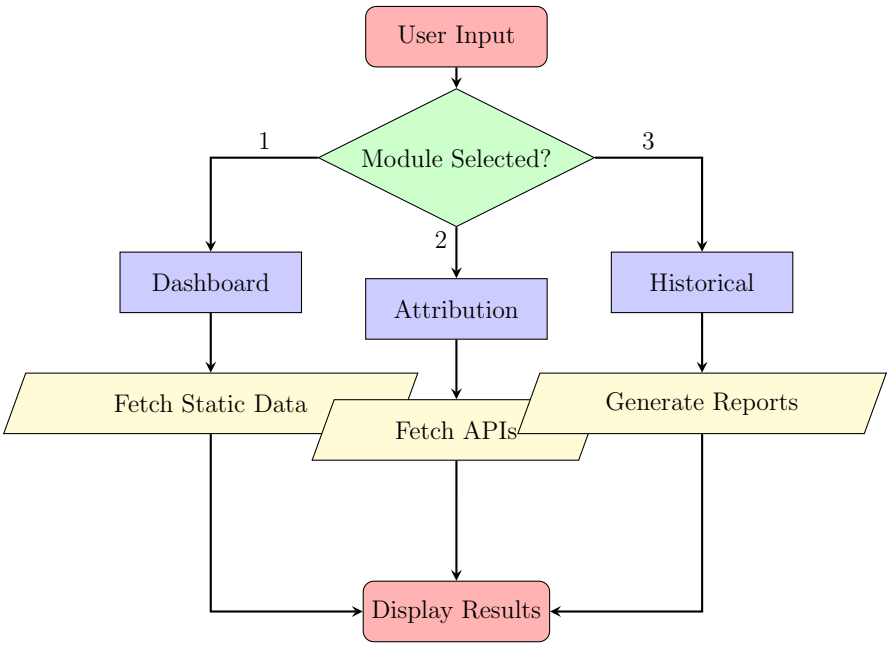

Figure 12: Complete system flowchart

14.8 Appendix H: Error Handling and Edge Cases

Common Edge Cases:

1. **No OSM Sources Found**: System returns message indicating sparse data coverage for the region
2. **API Timeout**: Cached data used when available; error message displayed otherwise
3. **Zero Wind Speed**: Minimum wind factor of 0.1 prevents division errors
4. **All Natural Sources**: Suppression algorithm activates to avoid misleading attribution

5. **Missing AQI Components**: Default values (0) substituted for missing pollutants
6. **Invalid Coordinates**: Input validation ensures latitude [-90, 90] and longitude [-180, 180]

Robustness Measures:

- Try-catch blocks for all API calls
- Data validation for numeric ranges
- Graceful degradation when services unavailable
- User-friendly error messages
- Logging for debugging and monitoring

15 Acknowledgments

This project leverages open-source tools, publicly available APIs, and collaborative mapping data. Special recognition to:

- **OpenWeather** for providing comprehensive environmental data APIs
- **OpenStreetMap** contributors for global geographic information
- **Streamlit** team for the intuitive web application framework
- **XGBoost** developers for the powerful machine learning library
- The open-source community for maintaining essential Python packages

16 Author Contributions and Contact

Project Development: Anish Nadar
Contact: 9820927499
Email: anishnadar21@gmail.com
GitHub Repository: https://github.com/AnishRN/infosys-assignments
Live Application: https://infosys-assignments-at2uqfsa2ibrlg82xqtwwq.streamlit.app
For questions, suggestions, or collaboration opportunities, please contact the author via the information provided above.

This report was prepared as part of [AI-Enviroscan]
December 29, 2025