Air Quality Index (AQI) Analysis and Forecasting

Introduction to Data Science - IT4142E

LECTURER

Assoc. Prof. Than Quang Khoat

TEAM

Nguyen Huy Hoang

Hoang Trung Hieu

Alexandre Guechtouli

Hanoi 2026

Agenda

- **Introduction**
  Background, Motivation, and Objectives
- **Data Description**
  Sources and ETL Pipeline
- **Exploratory Data Analysis**
  Univariate & Multivariate Analysis
- **Forecasting Methodology**
  Regression & Classification approaches
- **Results & Evaluation**
  Model performance comparison
- **Conclusion**
  Key Findings and Future Work

Background & Motivation

The Challenge

Air pollution is a critical environmental issue affecting public health and economic development, especially in rapidly urbanizing areas like Hanoi.

The Need

Real-time monitoring exists, but **forecasting** is crucial for proactive planning, protecting vulnerable populations, and regulatory action.

The Goal

To bridge the gap between historical raw environmental data and actionable future insights using advanced Data Science techniques.

Problem Statement

Core Problem

Predicting AQI based on historical time-series data is complex due to:

- Non-linear environmental data
- Seasonal and diurnal cycles
- Influence of traffic and meteorology

Analytical Approaches

1. Regression

Predicting the exact numerical value of AQI to track precise pollution trends.

2. Classification

Categorizing air quality into actionable health labels (e.g., "Unhealthy") for public safety.

Project Objectives

- **Data Pipeline Construction:** Automated ingestion, cleaning, and preprocessing from cloud/local storage.
- **Feature Engineering:** Extracting temporal (hour, day) and lag features to capture autoregressive properties.
- **Model Development:** Training and tuning Machine Learning (RF, XGBoost) and Deep Learning (LSTM) algorithms.
- **Comparative Analysis:** Empirically evaluating and comparing the performance of Regression vs. Classification approaches.

Future Applications

Mobile Apps

Push notifications for predicted "Unhealthy" levels.

Smart Cities

Public dashboards visualizing pollution trends.

Route Opt.

Suggesting "cleaner" paths for pedestrians/cyclists.

Weather Integration

Refining accuracy with real-time weather APIs.

Data Source: AQI.in

Platform Overview

Managed by Purelogic Labs, AQI.in aggregates data from:

- **Government Reference Stations:** High-precision instruments.
- **Low-cost Sensor Network:** Thousands of laser-scattering sensors filling geographical gaps.

Target Pollutants

PM2.5 & PM10

Particulate Matter

CO

Carbon Monoxide

NO2 & SO2

Nitrogen/Sulfur Dioxide

O3

Ozone

Data Scope & Crawl Pipeline

Locations: Major Vietnamese cities (Hanoi, Da Nang, Ho Chi Minh City, etc.)

1

Extraction

Python script using requests and BeautifulSoup to parse live HTML data.

2

Transformation

Regex cleaning to remove units and convert raw strings to numeric formats.

3

Storage

Batched insertion into a PostgreSQL database hosted on Neon.

Crawled data

—34k data points

38 locations in Viet Nam

Collect from 14/11 to 21/12

Exploratory Data Analysis

Unveiling the Hidden Patterns of Pollution

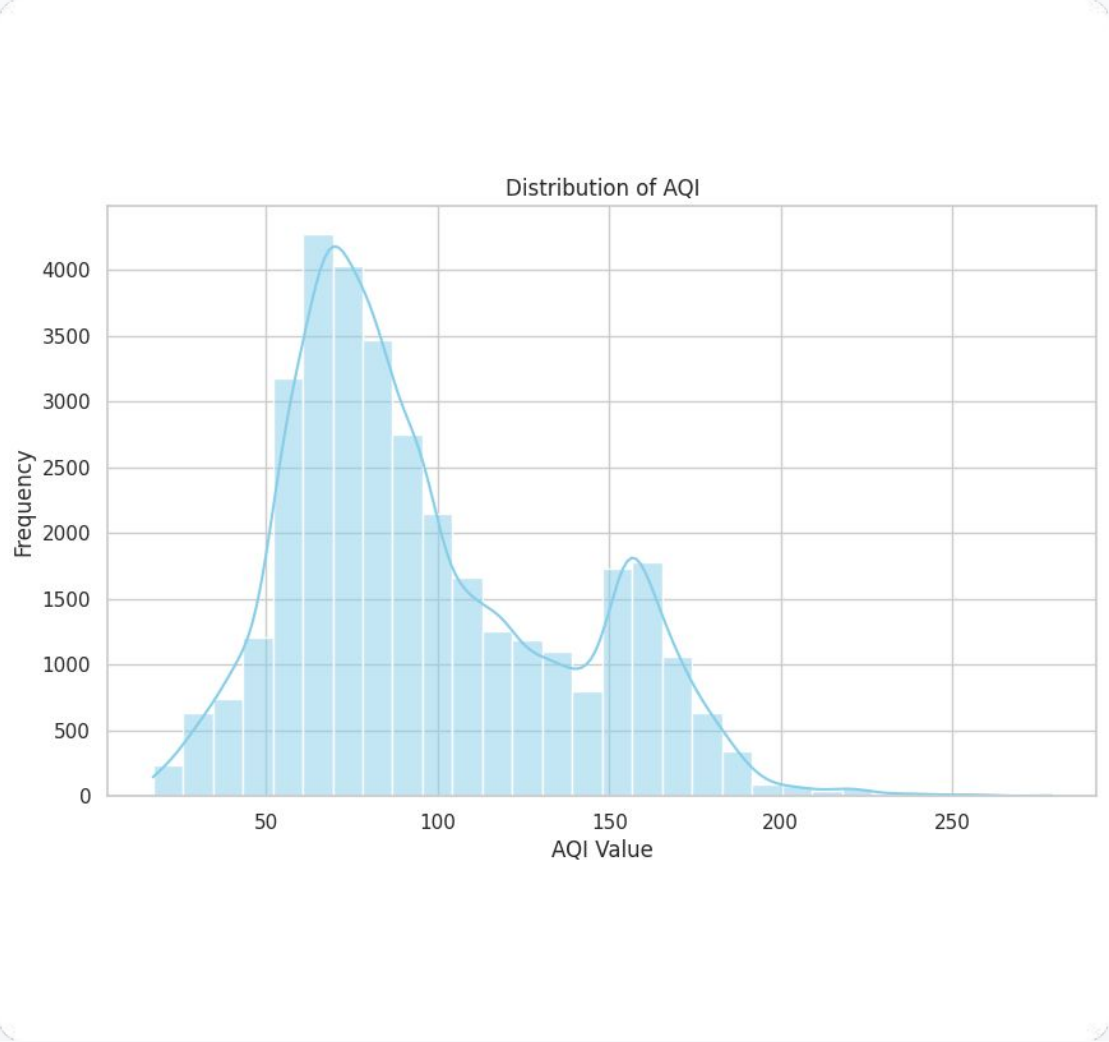
AQI Analysis: The Target

Bimodal Distribution

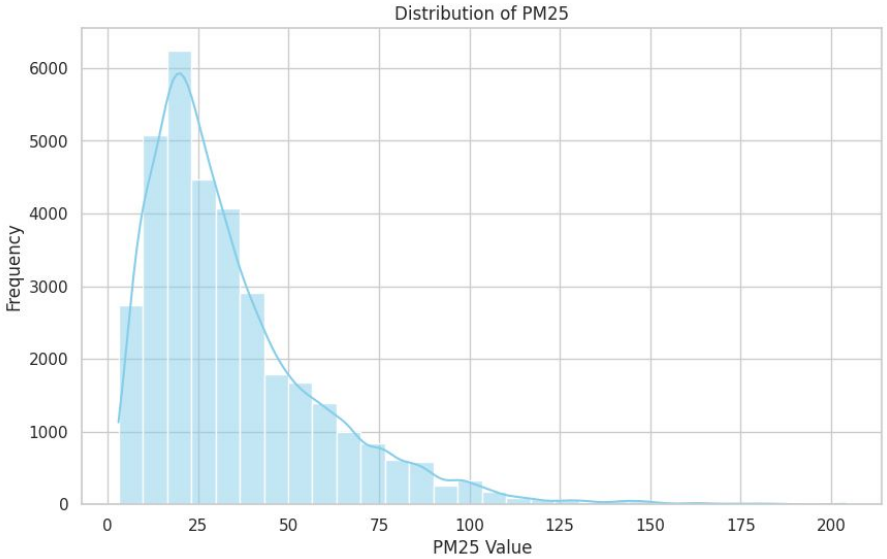
The AQI distribution reveals two distinct peaks:

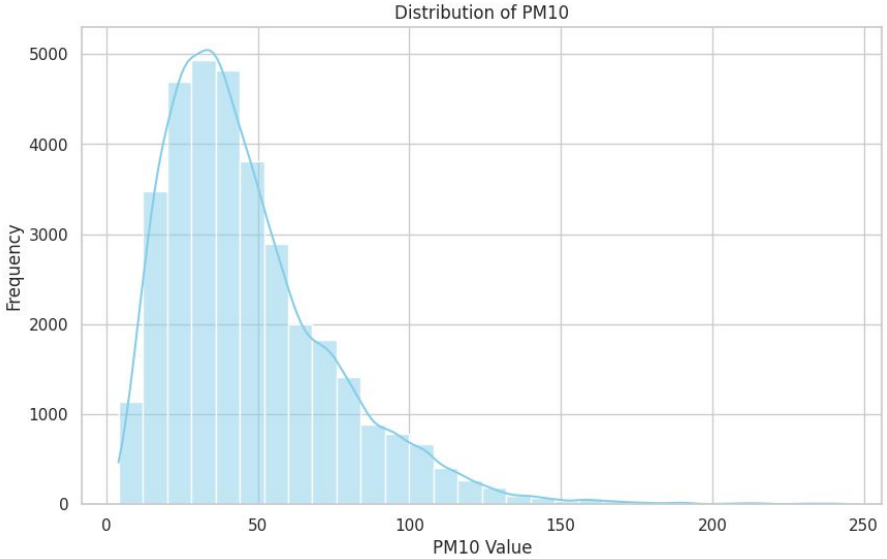
- **Primary Mode (~70):** "Moderate" air quality.
- **Secondary Mode (150-175):** "Unhealthy" recurring conditions.

Insight: A continuous cluster of outliers (210-260) indicates that "Very Unhealthy" events are persistent, not random.



Particulate Matter (PM2.5 & PM10)



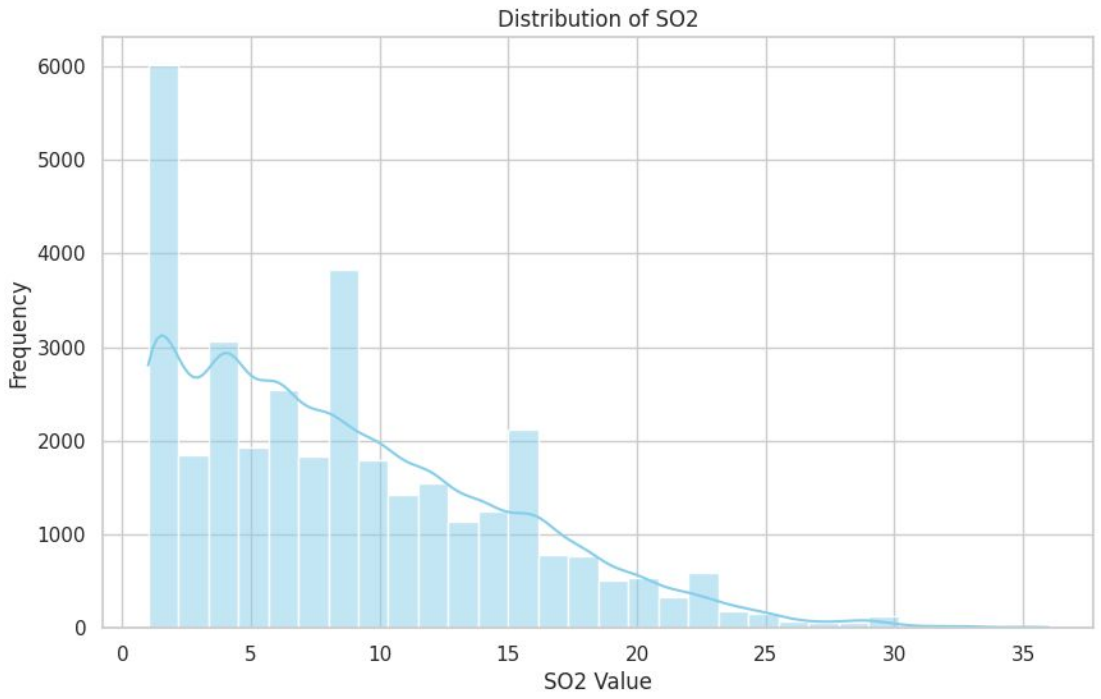


Distribution Characteristics

- **PM2.5:** Log-Normal distribution with a heavy right tail. Mode at 15-25 µg/m³.
- **PM10:** Similar right-skew but broader baseline.

Key Finding: Unlike AQI, PM2.5 is *unimodal*. This suggests that while PM is a major driver, it is not the sole cause of the secondary "Unhealthy" AQI peak.

Sulfur Dioxide (SO2)



Distribution Characteristics

- **Baseline & Concentration**: The data has a very low baseline, with the highest density of points occurring between 1 and 2.

- **Secondary Peaks:** Unlike unimodal distributions, this features distinct secondary "humps" at approximately 4, 8, and 15.

Key Finding: The specific secondary peaks (4, 8, 15) likely correspond to the operational cycles of distinct pollution sources, such as industrial discharge schedules or heavy shipping traffic.

CO Analysis: Sensor Saturation

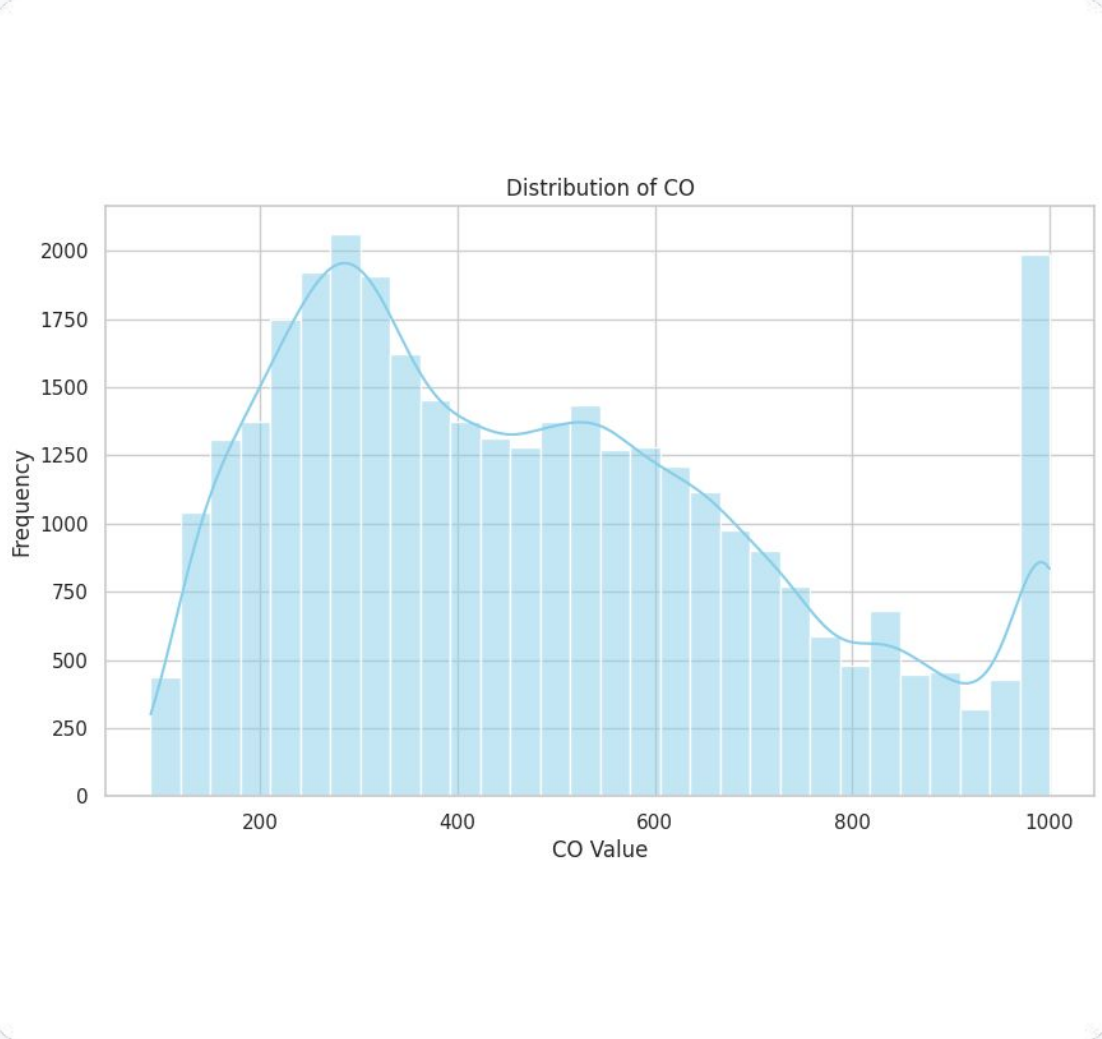
Data Quality Issue

Carbon Monoxide (CO) data exhibits a complex, multimodal distribution.

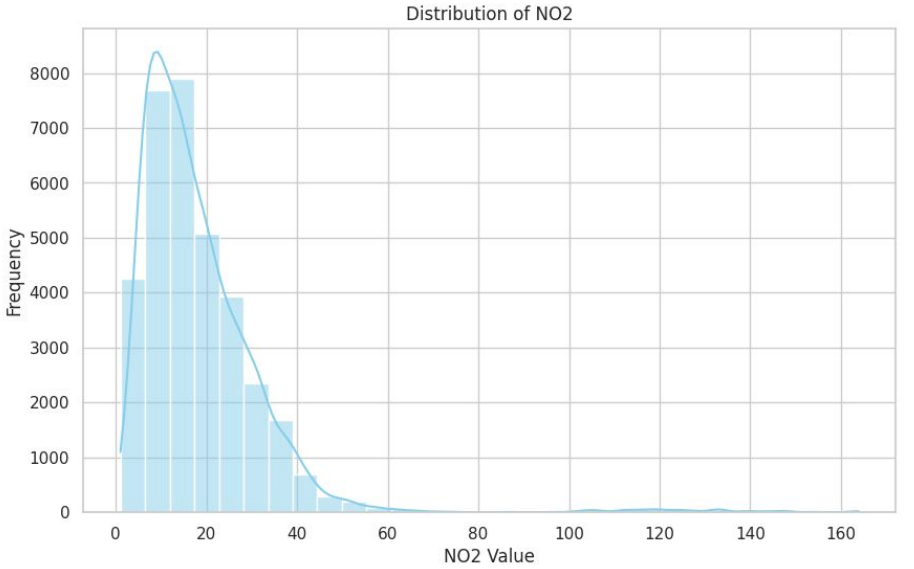
Critical Finding: A massive, isolated spike exactly at 1000 units.

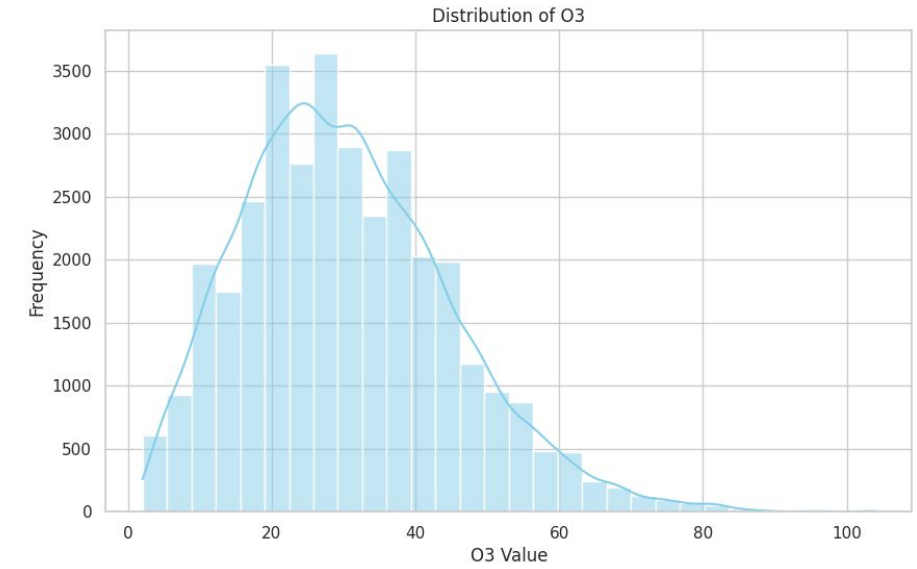
Sensor Saturation (Capping)

The sensor's max limit is 1000. True values above this are masked, creating an artificial "wall" in the data.



Nitrogen Dioxide & Ozone (NO2 & O3)





NO2 (Traffic Indicator)

Extreme right skew. A significant gap between 60-100 separates normal traffic from "bursty" pollution plumes.

O3 (Ozone)

the distribution is right-skewed, it lacks the extreme, heavy tails seen in other pollutants.

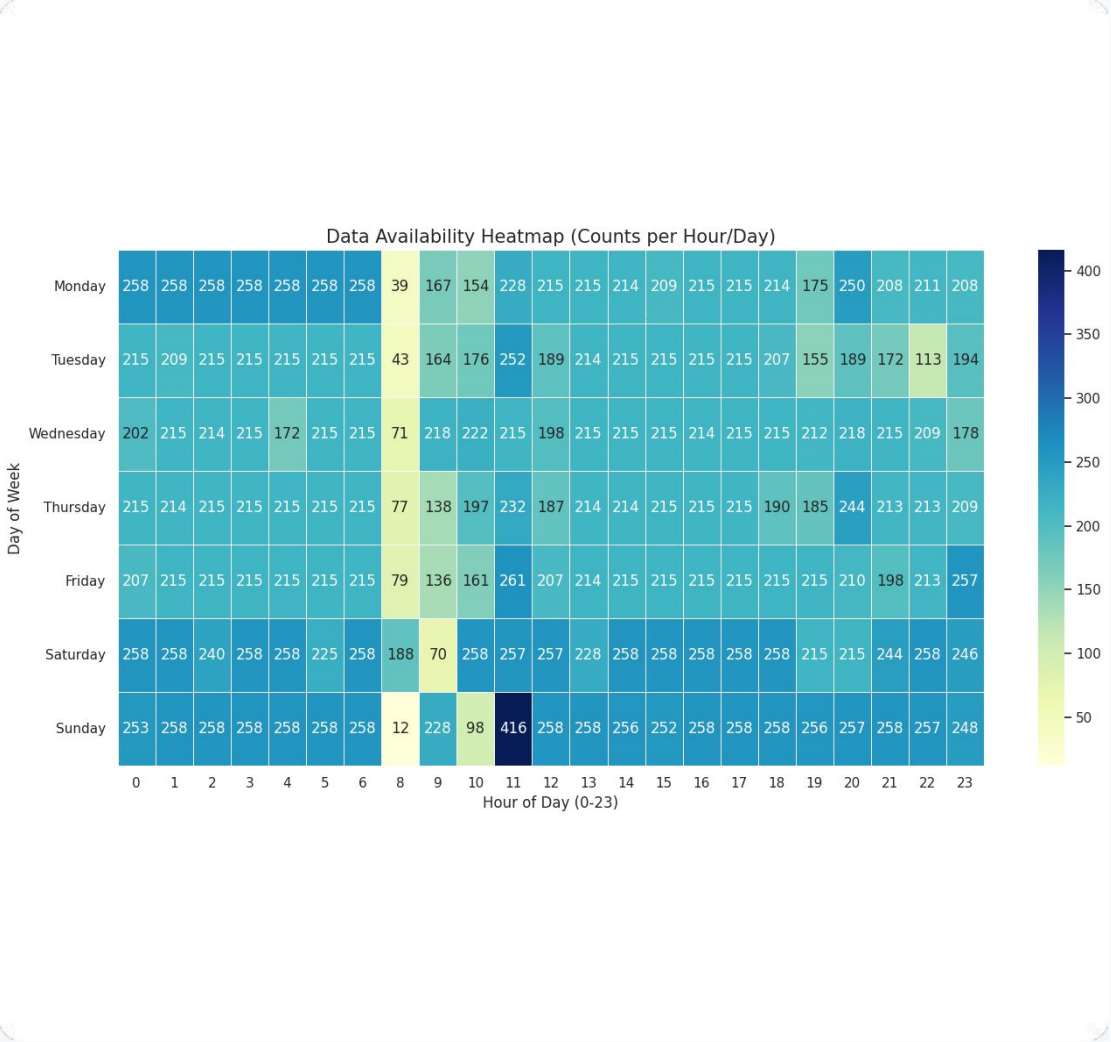
Data Availability & Quality Control

Audit Findings

- **Morning Gap (07:00-10:00):** Systematic missing data due to server congestion.
- **Sunday Spike:** Abnormally high records at 11:00 AM (retry mechanism).

Imputation Strategy

Used **Linear Time Interpolation** (limit 3h) to fill gaps while preserving trends, rather than simple mean imputation.



Multivariate Correlations

Key Insights

- **Dominant Driver:** PM2.5 and AQI have a near-perfect linear correlation.
- **Multicollinearity:** PM2.5 and PM10 provide redundant information.
- **Weak Gas Correlation:** High AQI events can occur even when NO2 or SO2 levels are low.

Geographic Pollution Fingerprinting

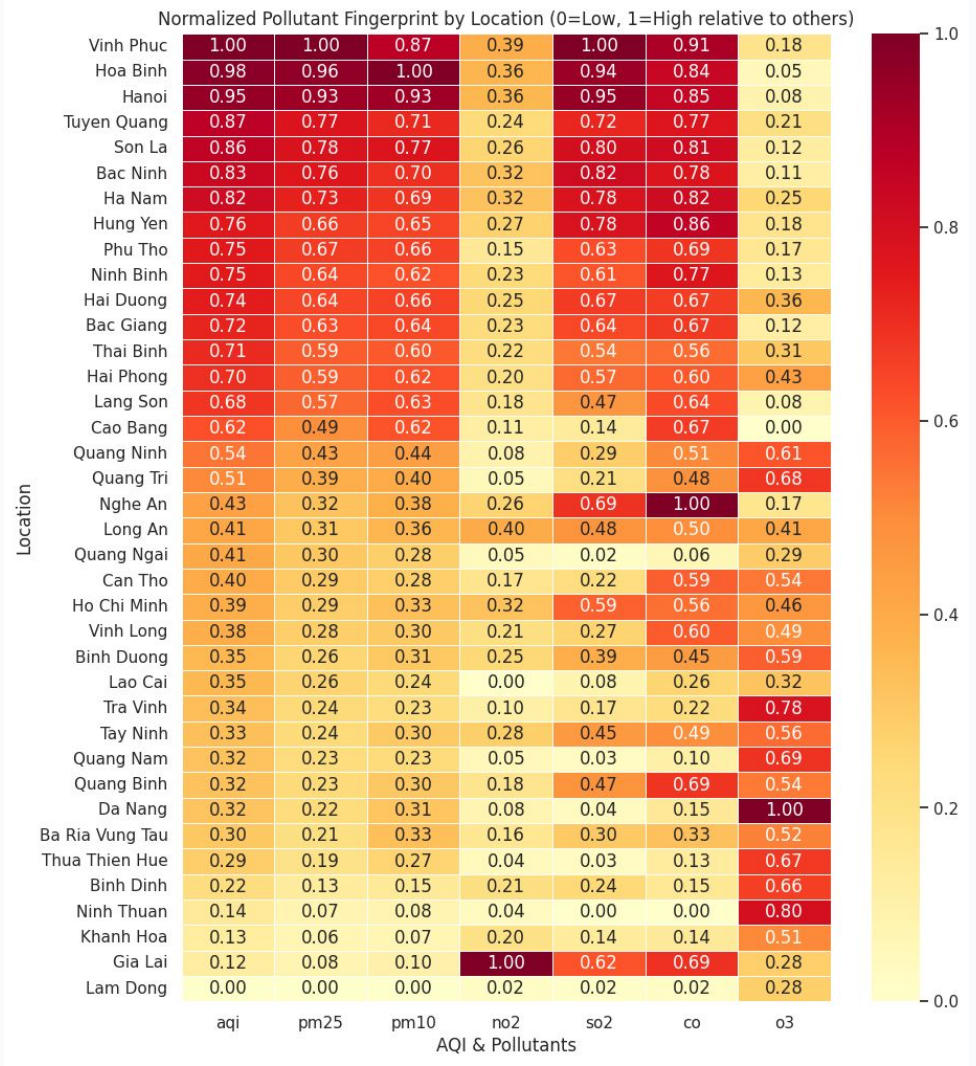
Pollution Clusters

Northern Cluster (Hanoi): Scores near 1.00 for Particulates (PM). Driven by dust accumulation.

Central & Coastal Cluster (Da Nang): Low PM scores but maximum Ozone (O3) scores.

Conclusion

The North battles dust; the Coast battles photochemical smog.



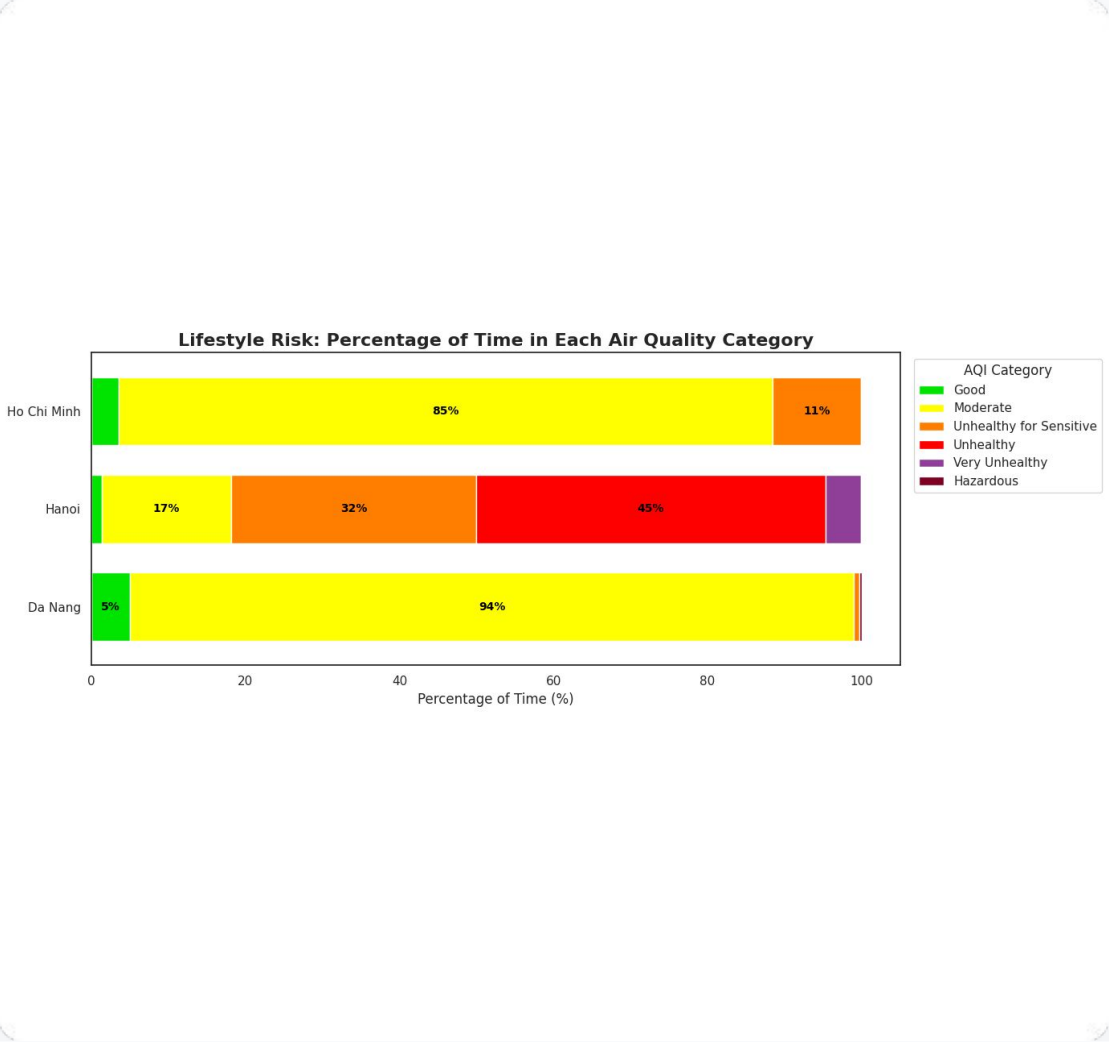
Lifestyle Risk: The North-South Divide

Unsafe Air Frequency

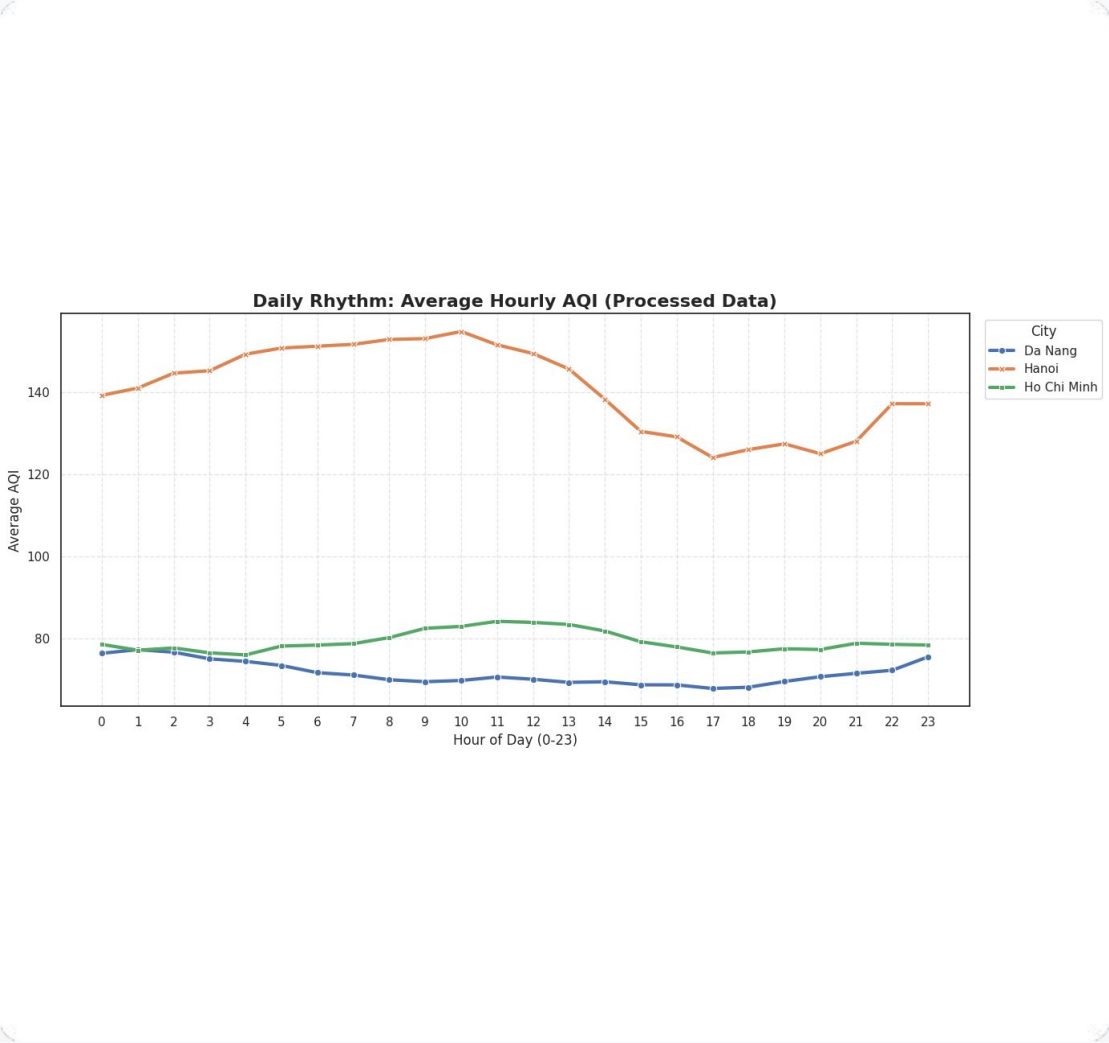
We analyzed the percentage of time spent in unsafe AQI categories.

- **Hanoi:** Air is unsafe **>77%** of the time.
- **Da Nang / HCMC:** "Moderate" air quality **85-94%** of the time.

Reality: Pollution in Hanoi is a daily crisis; in the South, it is merely a background factor.



Temporal Dynamics: Daily Rhythm



Hanoi

High pollution at night. Drops 12:00-17:00 due to "Afternoon Ventilation" breaking the inversion layer.

HCMC

Gentle morning rise, plateaus during day. Driven by human activity rather than meteorological trapping.

Da Nang

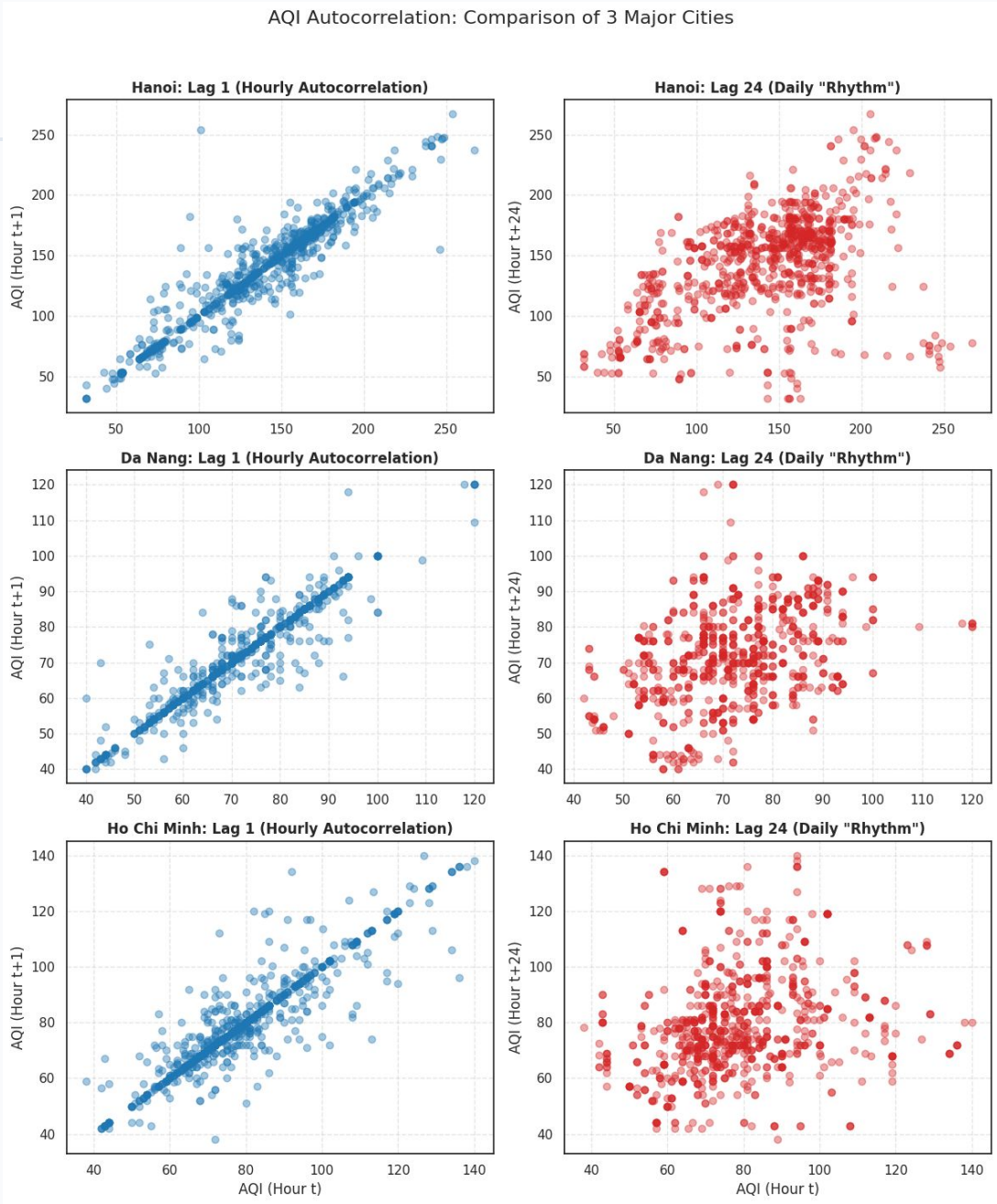
Exhibits a relatively stable and low AQI profile throughout the day, gradually decreasing to reach its minimum around 17:00 before slightly rising again in the evening.

Autocorrelation Analysis

Inertia vs. Seasonality

- **Lag 1 (Hourly):** Strong linear relationship. Pollution is highly inertial (changes gradually).
- **Lag 24 (Daily):** Weak, "fuzzy" correlation.

Implication: Simple daily persistence models fail because weather disruptions (wind/rain) break the daily rhythm.



AQI Forecasting

Modeling Strategy

Phased Approach

Phase A (Baseline): Using raw data.

Phase B (Enhanced): Using domain-specific feature engineering.

Data Splitting

Location-Stratified Temporal Split

Train (80%) / Test (20%) per location. Prevents data leakage and ensures geographic generalization.

Feature Engineering (Phase B)

- **Cyclical Time:** Transformed hours into geometric coordinates ($\sin, \cos$) to smoothly connect 23:00 to 00:00.
- **Inertia (Lag):** Included AQI_{t-1}, AQI_{t-2}, AQI_{t-3}, etc.
- **Rolling Statistics:** 3-hour and 6-hour rolling means to smooth noise.
- **Dimensionality Reduction:** Dropped PM10 due to high correlation.
- **Target Transformation:** Log-transformation to normalize error distribution.

Regression Results

Model	**RMSE** (Lower is better)		**MAE** (Lower is better)		R^2 **Score** (Higher is better)	
	Phase A	**Phase B**	**Phase A**	**Phase B**	**Phase A**	**Phase B**
Linear Regression	8.16	6.41	6.01	4.81	0.951	0.971
Random Forest	2.06	1.27	0.95	**0.56**	**0.997**	**0.999**
XGBoost	**1.87**	**1.25**	**0.93**	0.67	**0.997**	**0.999**
LSTM	19.62	27.08	11.03	16.06	0.719	0.498

Classification Task

The Goal

Categorize air quality into labels: “Good”, “Moderate”, “Unhealthy for Sensitive Groups”, “Unhealthy”, “Very Unhealthy”, “Hazardous”

The Challenge

Class Imbalance & Covariate Shift: Southern cities have almost zero "Very Unhealthy / Hazardous" instances.

The Solution (Phase B)

SMOTE

Synthetic Minority Over-sampling Technique

Synthesized minority classes for training to prevent bias towards "Moderate" air.

Why not Accuracy?

The Bias Problem

Accuracy would bias toward the majority class ("Moderate"), ignoring dangerous pollution events.

Macro F1-Score

Harmonic mean of Precision and Recall.

Treats all health categories with **equal importance**, penalizing misclassification of rare "Hazardous" events.

Classification Results (Macro F1-Score)

Model	Phase A (Imbalanced)	Phase B (SMOTE + Enhanced)
Logistic Regression	0.852	0.965
Random Forest	**0.979**	**0.984**
XGBoost	0.963	**0.970**
LSTM	0.536	0.684

Key Findings

- **Regional Disparity:** A distinct "North vs. South" divide. Hanoi lifestyle risk is 77% vs. clean air in the South.
- **Pollution Driver:** PM2.5 is the primary driver, allowing dimensionality reduction.
- **Model Performance:** ensemble tree-based models (Random Forest and XGBoost) significantly outperformed the LSTM model
- **Feature Engineering:** Cyclical encoding and rolling statistics were more decisive than model complexity.

Limitations

Sensor Saturation

CO sensors cap at 1000 units, masking the true variance of extreme pollution events.

Missing Meteorology

Weak Lag-24 correlation suggests weather (wind/rain) drives changes, but this data was missing. Models cannot predict dispersal caused by sudden weather shifts.

Future Recommendations

Weather API

Integrate wind/humidity data to forecast pollution clearance.

Early Warning

Deploy classification model as a backend for mobile alerts.

Route Opt.

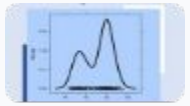
Integrate into navigation apps for "clean" route suggestions.

Thank You!

Students: Nguyen Huy Hoang, Hoang Trung Hieu, Alexandre Guechtouli

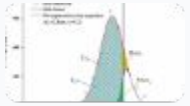
Class: IT4142E - 161306

Image Sources



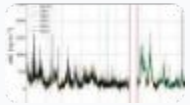
https://campuspi.com/wp-content/uploads/2025/03/article-30-01.jpg

Source: campuspi.com



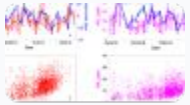
https://aaqr.org/images/article_images/2019/10/0103_fig1.png

Source: aaqr.org



https://ar.copernicus.org/articles/3/293/2025/ar-3-293-2025-avatar-web.png

Source: ar.copernicus.org



https://aaqr.org/images/article_images/2020/feature/20-05-0193.png

Source: aaqr.org

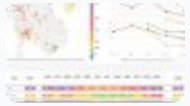
x-raw-image:///14ae383087413eb0cdf82a62c12e8b54c53b0f8089429c0b295f60cde2ff3843

Source: www.columbia.edu

https://www.mdpi.com/atmosphere/atmosphere-15-00896/article_deploy/html/images/atmosphere-15-00896-g002-550.jpg

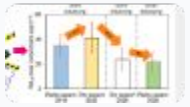
Source: www.mdpi.com

Image Sources



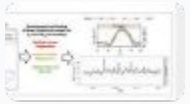
https://cdn.breathesafeair.com/wp-content/uploads/2024/01/Vietnam-Air-Pollution.png?strip=all

Source: breathesafeair.com



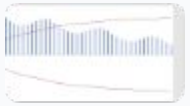
https://aaqr.org/images/article_images/2023/feature/22-09-0312.png

Source: aaqr.org



https://aaqr.org/images/article_images/2021/feature/20-07-0471.png

Source: aaqr.org



https://statisticsbyjim.com/wp-content/uploads/2021/05/CO2_data_ACF.png

Source: statisticsbyjim.com