

# Predicting Air Pollution Levels in Five Major Indian Cities

Aleksandr Jan Smoliakov    Danial Yntykbay    Davide Giuseppe Griffon

Data Science Study Programme  
Faculty of Mathematics and Informatics

2024

# Table of Contents

- 1 Introduction
- 2 Data Analysis
- 3 Methodology
- 4 Conclusion

# Project Overview

- Background:
  - Air pollution ranks among the most pressing global health threats
  - India faces some of the highest pollution levels globally
  - Driven by rapid urbanization and economic growth
- Study Objectives:
  - Analyze pollution and weather data from five major Indian cities:
    - Bengaluru, Delhi, Hyderabad, Jaipur, Mumbai
  - Develop predictive models using multivariate linear regression
  - Focus on understanding key meteorological and temporal predictors

# Literature Review

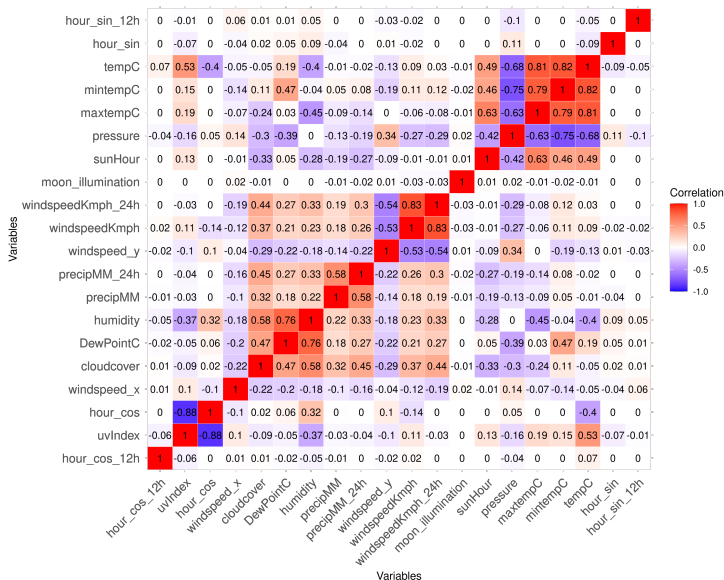
- Extensive research available:
  - Widespread interest among scientists and data analysts
  - Numerous studies on air pollution
  - Many also are focused on India due to severe air quality issues
- Research gap:
  - No existing studies examining these five specific cities together
  - Unique combination of air quality and weather datasets
- Most existing approaches:
  - Deep learning / neural network models
  - Complex machine learning algorithms
- Our approach:
  - Focus on interpretable linear regression
  - Emphasis on feature engineering
  - Independent models for each city and pollutant

- Air Quality Data in India (2015-2020):
  - Hourly pollution measurements
  - Coverage: 27 major Indian cities
  - Over 700,000 records
- Historical Weather Data (2006-2019):
  - Over 20 meteorological variables
  - Coverage: 8 Indian cities
  - Over 700,000 records
- Combined Dataset:
  - Time period: January 2015 to December 2019
  - Five cities with complete data coverage:
    - Bengaluru, Delhi, Hyderabad, Jaipur, Mumbai

# Preliminary Analysis

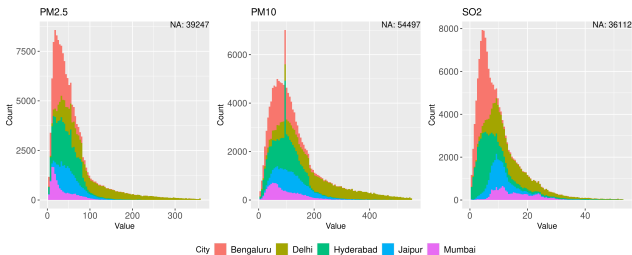
- Correlation between pollutants:
  - Most pollutants positively correlated with each other
  - O<sub>3</sub> shows distinct pattern from other pollutants
- Weather correlations:
  - Humidity: negative correlation with most pollutants
  - Wind speed: negative correlation, aids pollutant dispersion
  - Temperature: positive correlation with O<sub>3</sub>
  - UV index: positive correlation with O<sub>3</sub>
- Temporal patterns:
  - Strong seasonal trends observed
  - Higher pollution in autumn and winter

# Preliminary Analysis



# Original Data Distribution

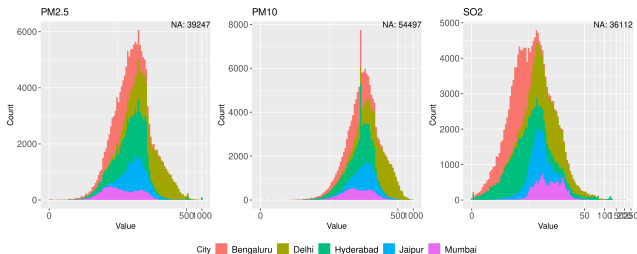
- Initial analysis revealed strongly right-skewed distributions:
  - High frequency of lower values
  - Long tail extending toward higher concentrations
  - Pattern consistent across all pollutants and cities





# After Log Transformation

- Applied logarithmic transformation to address skewness:
  - Added constant of 1 to handle zero values
  - Resulted in more normal-like distributions
  - Data better suited for linear regression
  - Transformation applied consistently across all pollutants



- Temporal Features:
  - Hour of day (cyclic encoding)
  - Day of week
  - Month of year
- Weather-Related Features:
  - Wind components (X and Y axes)
  - 24-hour cumulative precipitation
  - 24-hour cumulative wind speed
  - Normalized sunrise times
- All continuous features were scaled

# Model Development

- Independent linear regression models for each:
  - City
  - Response variable
- Data splitting:
  - Training: 2015-2018
  - Testing: 2019
- Feature selection using VIF analysis
- Removed features with  $VIF > 4$ :
  - minTempC
  - maxTempC
  - DewPointC

# Model Performance by Pollutant

Response	$r$	$R^2$	RMSE
PM2.5	0.728	0.538	0.437
PM10	0.694	0.492	0.435
O3	0.660	0.443	0.576
NO <sub>x</sub>	0.430	0.214	0.604
NH3	0.338	0.161	0.420
CO	0.311	0.132	0.307
SO2	0.273	0.102	0.388

# Model Performance by City

City	$r$	$R^2$	RMSE
Delhi	0.615	0.403	0.412
Hyderabad	0.598	0.390	0.390
Bengaluru	0.427	0.246	0.481
Jaipur	0.411	0.198	0.472
Mumbai	0.061	0.007	0.649

# Key Findings

- Best predictions for:
  - $\text{PM}_{2.5}$  ( $R^2 = 0.538$ )
  - $\text{PM}_{10}$  ( $R^2 = 0.492$ )
  - $\text{O}_3$  ( $R^2 = 0.443$ )
- Best performing cities:
  - Delhi ( $R^2 = 0.403$ )
  - Hyderabad ( $R^2 = 0.390$ )
- Key predictors:
  - Month of year (seasonal patterns)
  - Humidity (negative correlation)
  - Wind speed and UV index

# Conclusions

- Models show moderate predictive power
- Performance varies significantly across:
  - Pollutants
  - Cities
- Temporal patterns are strong predictors
- Meteorological variables show consistent influence
- Linear models capture main relationships

- Explore more sophisticated approaches:
  - Generalized Additive Models (GAMs)
  - Machine learning techniques
- Additional predictors:
  - Satellite data
  - Traffic information
  - Industrial activity metrics
  - Special events data
- Extend to more cities and longer time periods



Thank you for your attention