Predicting Air Pollution Levels in Five Major Indian Cities

A. J. Smoliakov D. Yntykbay D. G. Griffon

Data Science Study Programme
Faculty of Mathematics and Informatics

2024–12–16

Project Overview

- Background
  - Air pollution among most pressing global health threats
  - Particularly severe in India
- Study Objectives
  - Analyze pollution data in five major Indian cities
  - Develop predictive models for main air pollutants
  - Uncover key meteorological and temporal predictors

Literature Review

- Numerous studies on air pollution, including in India
- Research gap
  - No studies examining these five specific cities together
  - Unique combination of air quality and weather datasets
  - Most studies focus on time series models
- Our approach
  - Focus on interpretable linear regression
  - Emphasis on feature engineering
  - Independent models for each city and pollutant

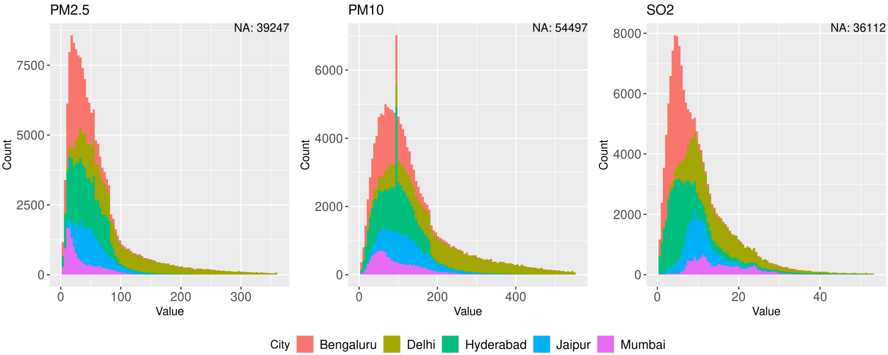
Data Sources

- Air Quality Data in India (2015-2020)
  - Hourly data
  - 27 major Indian cities
  - Seven pollutants: $PM_{2.5}$, PM_{10}, NO_2, SO_2, CO, O_3, NH_3
- Historical Weather Data (2006-2019)
  - Hourly data
  - 8 major Indian cities
  - $>$20 meteorological variables
- Combined Dataset
  - Intersection of the two datasets
  - Time period: January 2015 to December 2019
  - 5 cities: Bengaluru, Delhi, Hyderabad, Jaipur, Mumbai

Raw Outcome Variable Distribution

Right-skewed distributions

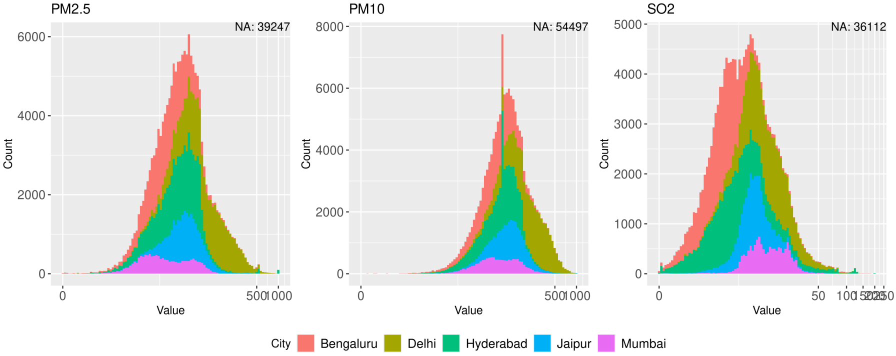
- Pattern consistent across all pollutants and cities



After Log Transformation

Applied $\log(1 + x)$ transformation to all outcome variables

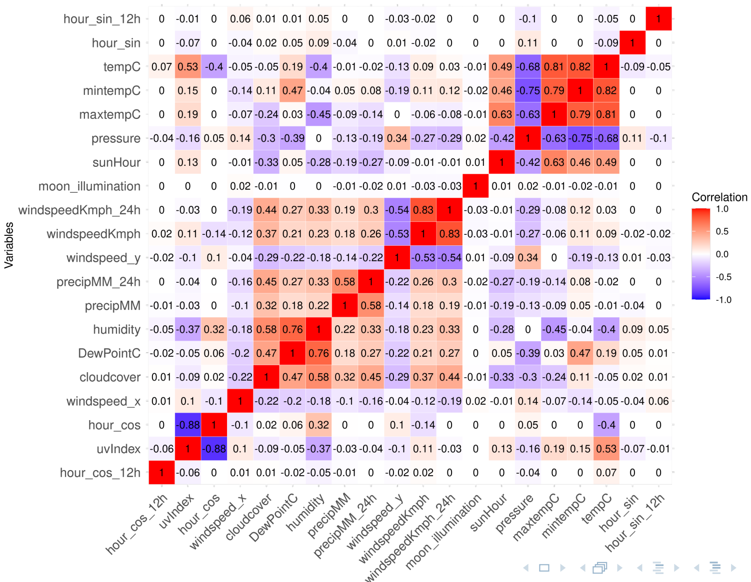
- Resulted in more normal-like distributions



Feature Engineering

- Temporal Features
  - Hour of the day
    - Cosine/Sine features to capture cyclic (12 and 24-hour) patterns
  - Day of the week (categorical)
  - Month of the year (categorical)
- Weather-Related Features
  - Wind components (X and Y axes)
  - 24-hour cumulative precipitation and wind speed
- All continuous features were scaled

Preliminary Analysis



Preliminary Analysis

- Weather correlations
    - Humidity: negative with most pollutants
    - Wind speed: negative with most pollutants
    - Temperature, UV index: positive with O_3
- Significantly higher pollution in autumn and winter
- Pollution correlations
    - Most pollutants positively correlated with each other
    - O_3 shows distinct pattern from other pollutants

Model Development

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

Model Performance (Mean per City/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
NOx	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

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

* Mumbai had substantial missing data, only 2/7 pollutants were modeled

Conclusions

- Models show moderate predictive power
  - R^2 between 0.006 and 0.672, mean = 0.292
  - Varies significantly across pollutants and cities
- Key predictors
  - Month of the year
  - Humidity (negative link)
  - Temperature (positive link)
  - $\cos(\textit{hour of day})$
  - Precipitation over 24 hours

Room for Improvement

- Explore more sophisticated approaches
  - Non-linear models
  - Time series models (e.g. LSTM)
- 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