Predicting Air Pollution Levels in India Using Multivariate Regression Analysis

Abstract

Air pollution remains one of India's most pressing environmental challenges, significantly impacting public health, the economy, and climate change. This project aims to predict air pollution levels in India by developing a multivariate regression model using various predictors such as location, season, weather data, temperature, and other relevant environmental and socio-economic factors. By leveraging data mining techniques and analyzing existing literature, the study seeks to identify key determinants of air pollution and provide insights that could inform policy decisions and mitigation strategies.

Contents

1 Introduction **2**

2 Data Sources **2**

3 Initial Data Analysis **3**

3.1 Data Preprocessing and Feature Selection 3

3.2 Time Series Analysis 3

3.3 Correlation Analysis 4

4 Literature Review **5**

1 Introduction

Air pollution in India has reached alarming levels, with several cities consistently ranking among the most polluted globally. This pollution is generated by various chemical pollutants, such as carbon monoxide (CO) and particulate matter (PM_{10} and $PM_{2.5}$). The adverse effects of air pollution extend beyond environmental degradation, affecting human health, economic productivity, and contributing to climate change. According to the World Health Organization, air pollution is a leading cause of premature deaths worldwide, with India accounting for a significant share [1].

Understanding the factors contributing to air pollution is crucial for developing effective mitigation strategies. This project aims to build a predictive model for air pollution levels by analyzing data from five major Indian cities: Bengaluru (Bangalore), Delhi, Hyderabad, Jaipur and Mumbai. At this stage, the most suitable modeling techniques and the specific dependent variables have yet to be determined. It is not yet clear whether the developed model will predict pollution levels on a daily or hourly basis. This decision will depend on factors such as data availability, complexity, and the desired balance between model accuracy and manageability, making it a key consideration for the study.

2 Data Sources

This project utilizes two primary sources of data to analyze air pollution levels across five major Indian cities:

- **Air Quality Data in India**: Available at Kaggle. This dataset provides hourly measurements of various air pollutants and particulate matter. The data is collected from multiple weather stations located within each of the five major cities. Recognizing these larger cities may have several monitoring stations to capture spatial variability in air quality, we aggregate the pollutant levels by averaging the measurements from all stations within a city for each hour. This averaging process ensures that the data represents the overall air quality of the city rather than isolated monitoring points.

- **Historical Weather Data for Indian Cities**: Available at Kaggle. This dataset includes hourly weather-related features for the same set of cities, encompassing over 20 variables such as precipitation (mm), wind speed, temperature, humidity, and other meteorological parameters.

Merging these two datasets is a strategic choice for several reasons. Firstly, it allows for a comprehensive analysis of the relationship between air pollution levels and various weather conditions. By integrating pollutant concentrations with corresponding meteorological data, we can better understand how factors like temperature, wind speed, and precipitation influence air quality. This holistic approach enhances the model's ability to capture the multifaceted nature of air pollution dynamics.

Furthermore, we extended the merged dataset by incorporating precise geographical coordinates for each of the five cities analyzed. Using the Google Maps API, we retrieved the exact latitude and longitude for each city, adding these as additional predictors in our model. Including geographic coordinates is expected to account for spatial dependencies and regional differences in pollution patterns, thereby potentially improving the model's predictive accuracy.

In summary, the final dataset comprises hourly air quality measurements, detailed historical weather data, and geographical information for the mentioned Indian cities. This integrated dataset provides a robust foundation for developing a predictive model aimed at forecasting daily air pollution levels, balancing data comprehensiveness with manageability.

3 Initial Data Analysis

3.1 Data Preprocessing and Feature Selection

Given the different sources of our data, we needed to filter and merge only the cities present in both datasets. Fortunately, all the cities selected for this project—Bengaluru, Delhi, Hyderabad, Jaipur, and Mumbai—are among the top ten most populous cities in India, including the top two.

The following code snippet creates a mapping between the city names used in the air quality data and those in the weather data, and loads the air quality data:

```r
# Create a mapping between city names in air quality data and weather data
city_name_map <- data.frame(
    air_quality = c("Bengaluru", "Delhi", "Hyderabad", "Jaipur", "Mumbai"),
    weather = c("bengaluru", "delhi", "hyderabad", "jaipur", "bombay")
)

# Load the air quality data
data_air_quality <- read_csv(
    file.path(INPUT_DIR, "air_quality", "city_hour.csv"),
    show_col_types = FALSE
)
```

Observations:

We identified 12 numeric columns that could serve as response variables: $PM_{2.5}$, PM_{10}, NO, NO_2, NOx, NH_3, CO, SO_2, O_3, Benzene, Toluene, and Xylene. The **Air Quality Index (AQI)** is a standardized measure that quantifies the overall air quality based on the concentrations of these pollutants. Since the AQI is derived from these pollutants, we chose not to use AQI itself as a response variable.

All these numeric columns have up to 55% missing values. Therefore, we decided to focus on the pollutants with the most complete data: $PM_{2.5}$, PM_{10}, NOx, NH_3, CO, SO_2, and O_3.

As an initial analysis, we plotted the selected pollutants, using different colors to represent each city. We observed that some cities have more complete records of both pollution and weather data than others. This disparity is illustrated in Figures 1 and 2.

Notably, data availability is consistent across both datasets: for example, Bengaluru consistently has more data, while Mumbai has less. Based on this preliminary analysis, we decided to focus on the three cities with the most data: Bengaluru, Delhi, and Hyderabad.

3.2 Time Series Analysis

We analyzed whether the time of year affects pollution levels. Although we only present the graph for O_3 pollution, all pollutants exhibit similar behavior: pollution levels tend to be lower during the summer months, from July to September.

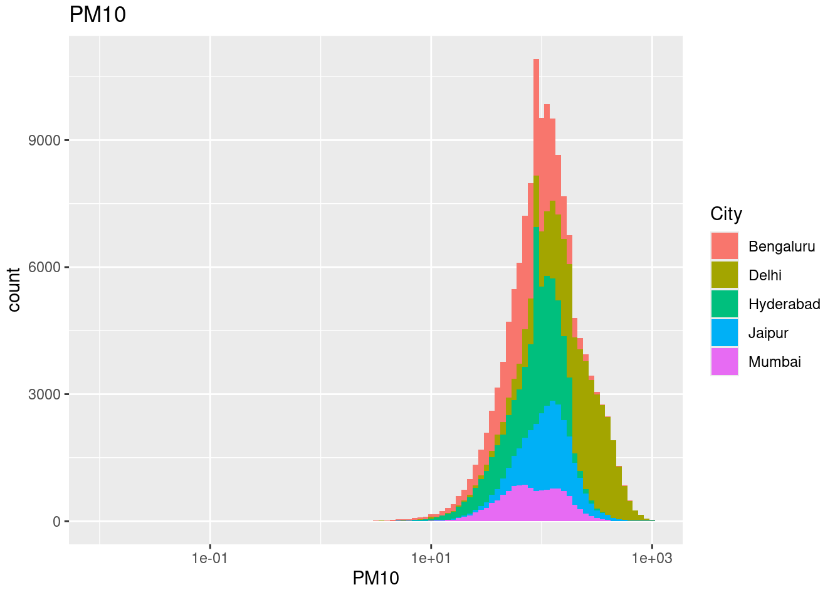

Figure 1: PM_{10} levels by city

3.3 Correlation Analysis

We conducted a correlation analysis to examine the relationships among the weather variables and between the features and response variables.

First, we calculated the correlation matrix for the weather variables:

```
# Calculate the correlation matrix
correlation_matrix <- cor(
    data_merged[, weather_vars],
    use = "pairwise.complete.obs"
)

# Remove rows and columns with all NA values
correlation_matrix <- correlation_matrix[
    !apply(is.na(correlation_matrix), 1, all),
    !apply(is.na(correlation_matrix), 2, all)
]

# Plot the correlation matrix
corrplot(
    correlation_matrix,
    method = "color",
    tl.col = "black",
    order = "hclust"
)
```

Observations:

The correlation matrix (Figure 4) shows that some weather variables are highly correlated with each other. To avoid multicollinearity, we decided to remove variables such as *FeelsLikeC*, *HeatIndexC*, *WindChillC*, *minTempC*, and *maxTempC*, which are highly correlated with *tempC* and are derived from combinations of temperature, humidity, and wind speed. Similarly, *humidity* is correlated with *DewPointC*, and certain UV index variables are highly correlated.

Next, we analyzed the correlation between features and response variables:

```
feature_vars <- c(
    "sunHour", "uvIndex", "moon_illumination",
    "DewPointC", "WindGustKmph", "cloudcover", "humidity",
    "precipMM", "pressure", "tempC", "visibility", "winddirDegree", "windspeedKmph"
```

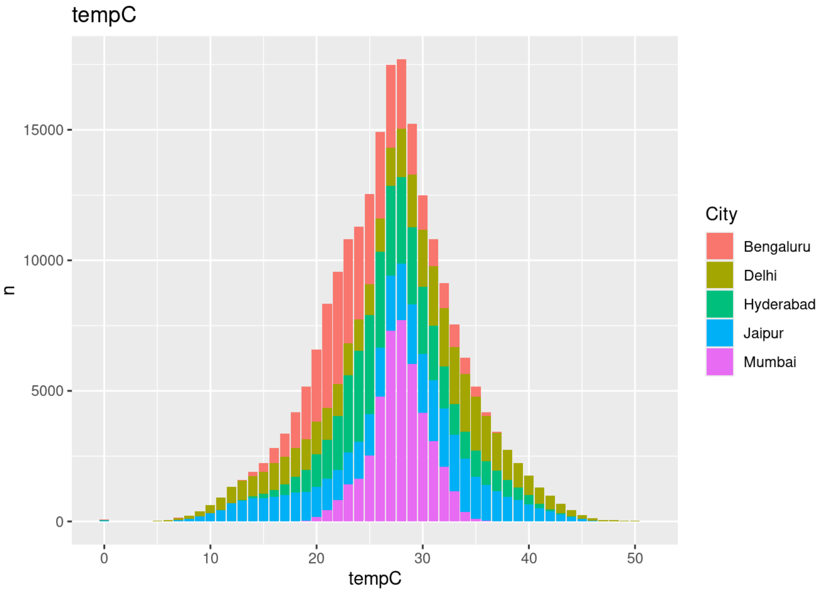

Figure 2: Temperature readings by city

```
)

# Calculate the correlation matrix
correlation_matrix <- cor(
    data_merged[, response_vars],
    data_merged[, feature_vars],
    use = "pairwise.complete.obs"
)

# Plot the correlation matrix
corrplot(
    correlation_matrix,
    method = "color",
    tl.col = "black"
)
```

Observations:

The correlation matrix (Figure 5) indicates that the weather variables are weakly correlated with the response variables. However, since we only used a linear correlation measure, there may be non-linear relationships that are not captured in this analysis.

4 Literature Review

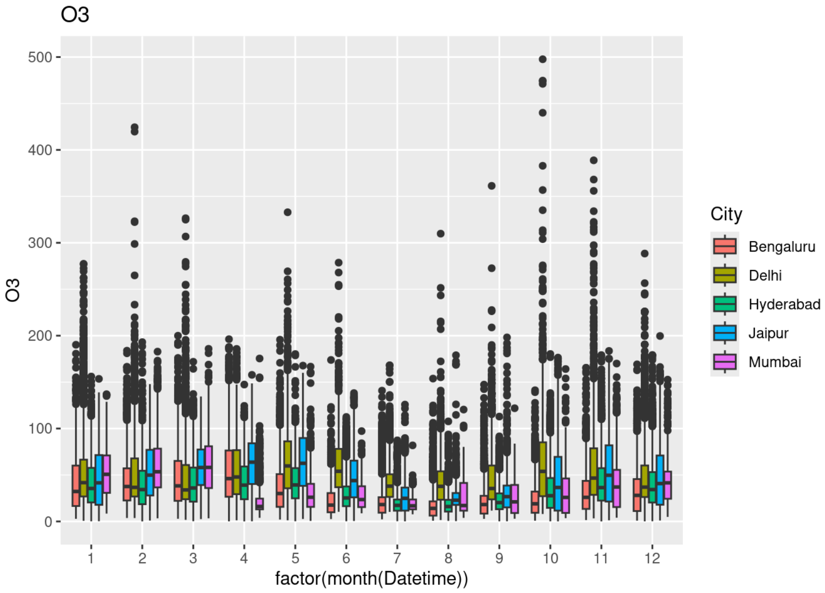

Figure 3: O_3 levels by month

References

[1] S. Dey and [Other Authors]. "Ambient Air Pollution and Daily Mortality in Ten Cities of India: A Causal Modelling Study". In: *The Lancet Planetary Health* 4.7 (2020), e287–e298. DOI: 10.1016/S2542-5196(24)00114-1. URL: https://www.thelancet.com/journals/lanplh/article/PIIS2542-5196(24)00114-1/fulltext.

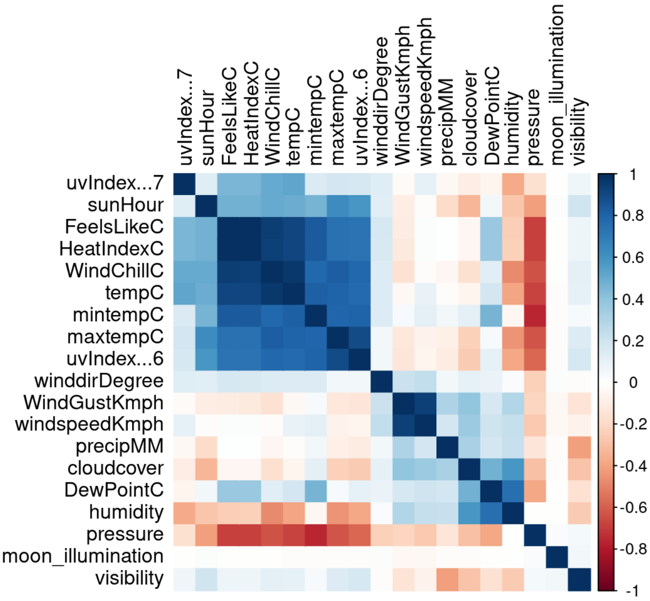

Figure 4: Correlation matrix of weather variables

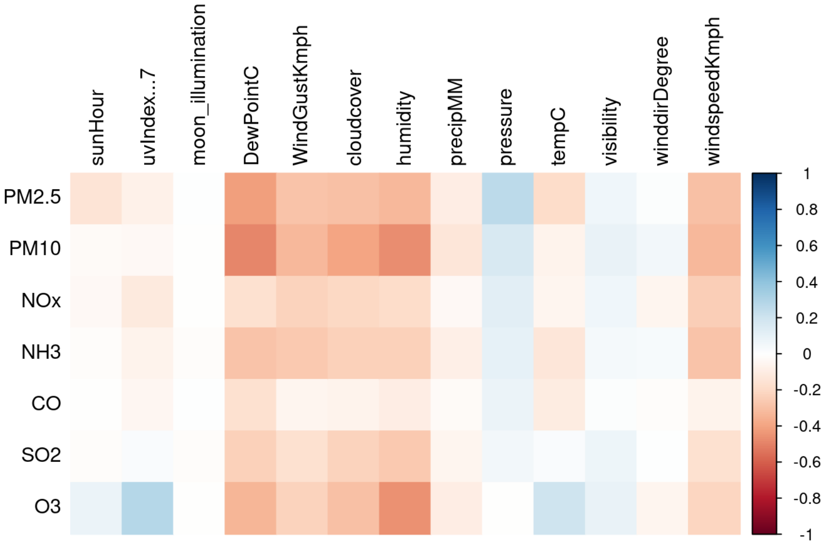

Figure 5: Correlation between features and response variables