Data Mining (2024) Submitted ; Published

# Predicting Air Pollution Levels in Five Major Indian Cities

**Aleksandr Jan Smoliakov** ALEKSANDR.SMOLIAKOV@MIF.STUD.VU.LT
**Danial Yntykbay** DANIAL.YNTYKBAY@MIF.STUD.VU.LT
**Davide Giuseppe Griffon** DAVIDE.GRIFFON@MIF.STUD.VU.LT
*Data Science study programme*
*Faculty of Mathematics and Informatics*

**Advisor:** Jurgita Markevičiūtė

## Abstract

Air pollution ranks among the most pressing global health threats, causing an estimated 3 to 9 million deaths annually. India, with its rapid urbanization and economic growth, faces some of the highest pollution levels globally. The country's air pollution stems from a combination of natural and anthropogenic sources, emitting harmful pollutants such as carbon monoxide (CO) and particulate matter ($PM_{10}$ and $PM_{2.5}$).

This project aims to develop a predictive model for air pollution levels using a multivariate regression approach. The model will incorporate a wide array of predictors, including geographic location, seasonal variations, meteorological data and temperature. By evaluating various modeling techniques—starting with linear regression and potentially expanding to more complex methods such as Random Forest—and analyzing data correlations, the study seeks to identify the most effective methods for accurately forecasting urban air pollution levels.

This topic was chosen due to its relevance in addressing the escalating air quality issues in India and the availability of extensive historical weather and air quality data. Previous studies on air quality prediction often focus on pollutants like $PM_{2.5}$ and $PM_{10}$. Here, we broaden the scope by analyzing seven different pollutants across five major Indian cities—Bengaluru (Bangalore), Delhi, Hyderabad, Jaipur, and Mumbai—assessing their seasonal trends and interdependencies, and examining a wider set of meteorological and temporal variables. The study also enhances interpretability by comparing feature importance across cities, offering a clearer understanding of regional differences.

This study will demonstrate the practical application of data mining techniques using real-world environmental data, serving as a foundation for further exploration in this field.

## 1. Literature Review

Air pollution has been extensively studied worldwide due to its significant impacts on public health, ecosystems, and the economy. Numerous studies have demonstrated the adverse health effects of air pollutants, including respiratory and cardiovascular diseases, leading to increased morbidity and mortality rates. Particularly, countries like India and China have received considerable attention because of their severe air quality issues, which are exacerbated by rapid industrialization, urbanization, and population growth.

During our literature review, we encountered a vast number of articles related to our research topic. The widespread interest among scientists and data analysts facilitated the collection of numerous sources. As our research progressed, we found that many studies have attempted to develop statistical and artificial intelligence models to predict air pollution levels, utilizing both global datasets and data specific to particular regions. Researchers have employed a variety of techniques, including multivariate regression, neural networks, and machine learning algorithms, to forecast pollutant concentrations based on diverse predictors such as meteorological conditions, emission sources, and socio-economic factors. These models have been applied at various regional scales, providing valuable insights for environmental management and policy-making. Despite this extensive body of research providing a solid foundation for our study, we have not found any existing studies that have examined the five major Indian cities using the same datasets we are using in this project.

To effectively manage and synthesize the extensive body of literature, we have chosen not to include general studies on air pollution in India, as they do not directly contribute to the development of our predictive model. Instead, we have focused our review on two specific categories of research that are more pertinent to our objectives: "Causal and Correlational Studies on Urban Air Pollution" and "Predictive Models". The first category delves into the factors influencing air pollution levels in urban areas, providing valuable insights that inform the selection of variables and the structural framework of our model. The second category encompasses studies that have developed predictive models for air pollution, offering methodologies and approaches that we can build upon to enhance the accuracy and reliability of our own model. By concentrating on these two groups, we aim to leverage existing knowledge effectively and advance our research in a meaningful way.

### 1.1 Causal and Correlational Studies on Urban Air Pollution

Understanding the dynamics of urban air pollution is essential for developing effective strategies to enhance air quality and protect public health. Several studies have focused on analyzing the correlations and underlying causes of air pollution in metropolitan environments, rather than constructing predictive models.

For instance, [Khedekar, Sneha and Thakare, Sunil, 2023] conducted a six-year analysis in Pune, India, assessing the correlations between pollutants and meteorological factors. The study revealed that most pollutants were positively correlated with each other and with temperature, except for $O_3$, which had a negative correlation. Wind speed showed a strong negative correlation with pollutant levels, emphasizing its role in pollutant dispersion.

Building on similar themes, [Diya et al., 2024] investigated air pollution across various urban hotspots in Chennai, India. This research assessed hourly concentrations of pollutants such as $PM_{10}$, $PM_{2.5}$, $SO_2$, $NO_2$, and CO across key areas—industrial, traffic, commercial,

and residential zones—over the course of 2022. A key methodological approach employed in this study is the Coefficient of Divergence (COD), which quantifies spatial variations in pollutant concentrations among the different hotspots. One of the significant findings of the Chennai study is the impact of wind on pollution dispersion. When wind speeds are low (0–3 m/s), CO levels tend to be higher, indicating that pollutants are not dispersing effectively and are accumulating near their sources. Conversely, when the wind blows from the south and southeast at moderate speeds (2–6 m/s), the concentrations of $PM_{2.5}$ and $PM_{10}$ increase. This suggests that pollutants from nearby industries are being transported toward the monitoring stations, highlighting the crucial role of meteorological conditions in air quality.

In another study, [Suthar et al., 2024] aimed to identify seasonal patterns and understand how meteorological factors influence pollutant levels in a different Indian city. The research included a correlation analysis between air pollutants and meteorological parameters—wind speed (WS), wind direction (WD), relative humidity (RH), and solar radiation (SR). Over three consecutive years, the analysis revealed that WD, WS, and RH generally had a negative correlation with all measured air pollutants. Calm wind conditions inhibit the dispersion of pollutants, resulting in higher concentrations near the ground, underscoring the importance of WS and WD in the dispersion and transport of air pollutants.

Expanding this line of research to European cities, [Rowland, 2024] examined the relationship between meteorological parameters and the concentrations of $NO_2$, $O_3$, $PM_{10}$, and $PM_{2.5}$ in Krakow, Paris, and Milan during 2021. The study found that $NO_2$, $PM_{10}$, and $PM_{2.5}$ concentrations were higher during winter and lower during summer, exhibiting negative correlations with temperature, while $O_3$ showed the opposite trend. Wind speed was inversely related to particulate matter and $NO_2$ levels but positively correlated with $O_3$ concentrations. These findings highlight the influence of meteorological conditions on pollutant levels and the occurrence of the "Ozone weekend effect" in these cities.

Speaking of temporal trends, [Bozhkova et al., 2020] conducted research in urban areas of Belarus. Seasonal patterns revealed higher pollution in autumn and winter, with increased dispersion of pollutants and ozone formation in spring and summer. The study observed daily pollution peaks occurring in the morning and evening, driven by human activities and affected by wind and atmospheric stability. The reduced dispersion efficiency during these periods, combined with higher emission intensities, contributes to these peaks.

These studies collectively underscore the significant impact of meteorological and temporal factors on urban air pollution across diverse geographic regions. The consistent observations of pollutant behavior in relation to temperature, wind speed, and other meteorological parameters highlight the necessity of incorporating environmental conditions into air quality management and policy-making.

### 1.2 Predictive Models

Predictive modeling plays a crucial role in understanding and forecasting air pollution levels, which is essential for public health planning and environmental management. Various studies have employed different statistical and machine learning approaches to predict concentrations of air pollutants and the Air Quality Index (AQI), a standardized measure that indicates the overall air quality and its potential impact on human health.

Singh et al. [Singh et al., 2012] investigated both linear and nonlinear methods for forecasting urban air quality, aiming to improve prediction accuracy in complex urban environments. The study examined the effectiveness of different modeling approaches for predicting concentrations of common urban pollutants such as $PM_{10}$, NO, CO, and $O_3$. Specifically, they applied linear models like multiple linear regression and nonlinear models including Artificial Neural Networks (ANNs) to compare their performance in capturing pollution patterns. The findings indicated that nonlinear models, particularly ANNs, provided better prediction accuracy than linear models, highlighting the importance of nonlinear approaches in modeling air pollution in urban settings.

Sanjeev et al. [Sanjeev, 2021] developed predictive models for air quality using machine learning algorithms, focusing on Artificial Neural Networks (ANN), Support Vector Machines (SVM), and Random Forests (RF). Their study aimed to identify the most efficient algorithm for air quality prediction. The models were evaluated based on accuracy scores, with the RF-based model achieving the highest accuracy.

Kothandaraman et al. [Kothandaraman et al., 2022] focused on predicting PM2.5 pollutant levels by employing a variety of machine learning algorithms, including linear regression, Random Forest, K-Nearest Neighbors, Ridge and Lasso regression, XGBoost, and AdaBoost. Their study utilized historical PM2.5 data and relevant meteorological features such as temperature, humidity, wind speed, and precipitation collected from monitoring stations in Anand Vihar, Delhi, over the period from January 2014 to December 2019. By evaluating the performance of these models through statistical metrics like Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared ($R^2$), they found that ensemble methods such as XGBoost and Random Forest outperformed other algorithms in terms of predictive accuracy. These results highlight the effectiveness of advanced machine learning techniques in modeling air pollution and the critical role of incorporating meteorological data.

Kumar et al. [Kumar and Pande, 2023] addressed the challenge of predicting the AQI by analyzing air pollution data from 23 Indian cities over six years. They carried out extensive data preprocessing, which involved handling missing values, correcting outliers, normalizing data, selecting features, and applying logarithmic transformations to fix skewed data. Their exploratory data analysis showed a significant decrease in pollution levels in 2020, likely due to COVID-19 lockdowns. To fix data imbalance, they used the Synthetic Minority Over-sampling Technique (SMOTE). They performed machine learning-based AQI predictions using various models, both with and without SMOTE resampling, and compared the results. The models were assessed using standard metrics like accuracy, precision, recall, F1-score, and error metrics (MAE, RMSE, RMSLE, $R^2$). The XGBoost model performed the best, achieving the highest accuracy in both training and testing phases, while the SVM model had the lowest accuracy. The Random Forest model also did well, especially when SMOTE was applied. The study emphasizes the effectiveness of ensemble learning methods in AQI prediction and suggests that future research could explore deep learning techniques to improve accuracy further.

Roy et al. [Roy et al., 2024] conducted a study in the densely populated northern Indian states of Delhi, Haryana, and Uttar Pradesh, analyzing PM2.5 concentrations in relation to meteorological factors such as temperature, precipitation, surface pressure, and wind. They employed Ordinary Least Squares (OLS) regression and Geographically Weighted

Regression (GWR) to explore the relationships between PM2.5 levels and environmental parameters across different seasons and locations. The OLS model identified significant predictors with $R^2$ values of 0.93 for summer and 0.94 for winter, while GWR accounted for spatial variability, enhancing model performance and highlighting the importance of geographical factors in air pollution modeling. However, our study does not utilize geographical data and cannot replicate the GWR analysis by Roy et al. Instead, we focus on the overall relationships between PM2.5 concentrations and meteorological factors without considering spatial variability.

Building on the significance of feature engineering and advanced modeling techniques, [Naz et al., 2024] emphasized the crucial role of feature engineering in time series prediction of air pollutants. They introduced a two-stage feature engineering and selection process that combines correlation-based selection with Variational Mode Decomposition (VMD). By developing and categorizing 22 new features into meteorological, temporal, statistical, and air pollutant types, their approach customizes optimal feature sets for each of the five major air pollutants. This customization enhances model performance by 1–5% compared to traditional lag-based methods and further improves accuracy by 3–13% when integrating VMD features. The optimized feature selection allows for simpler forecasting models with significant improvements in RMSE, MAE, and $R^2$ scores.

These studies demonstrate the effectiveness of various machine learning and statistical methods in predicting air pollution levels and AQI. Nonlinear models and ensemble learning techniques like Random Forest and XGBoost have shown high accuracy in forecasting pollutant concentrations and AQI. The incorporation of meteorological and temporal data significantly enhances model performance. For our study, which does not employ deep learning methods, these findings suggest that ensemble methods and regression techniques—especially those accounting for spatial variability—can serve as effective alternatives for accurate air quality prediction. Incorporating meteorological factors and addressing data imbalances may further improve prediction accuracy without the need for deep learning models.

## 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. [Soneji, 2020]

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.

## References

Victoria V. Bozhkova, Alexander M. Liudchik, and Siarhey D. Umreika. Influence of meteorological conditions on urban air pollution. *Acta Geographica Silesiana*, 14(4):5–21, 2020. ISSN 1897-5100. URL https://ags.wnp.us.edu.pl/download/wydawnictwa/ags/tom_40_2.pdf.

M. Diya, Sudheer Kumar Kuppili, and S. M. Shiva Nagendra. Air quality in different urban hotspots in a metropolitan city in india and the environmental implication. *Environmental Monitoring and Assessment*, 196:1102, 2024. doi: 10.1007/s10661-024-13272-z. URL https://doi.org/10.1007/s10661-024-13272-z. Received: 28 May 2024 / Accepted: 16 October 2024.

Khedekar, Sneha and Thakare, Sunil. Correlation analysis of atmospheric pollutants and meteorological factors using statistical tools in pune, maharashtra. *E3S Web Conf.*, 391: 01190, 2023. doi: 10.1051/e3sconf/202339101190. URL https://doi.org/10.1051/e3sconf/202339101190.

D. Kothandaraman, N. Praveena, K. Varadarajkumar, B. Madhav Rao, Dharmesh Dhabliya, Shivaprasad Satla, and Worku Abera. Intelligent forecasting of air quality and pollution prediction using machine learning. *Adsorption Science & Technology*, 2022:5086622, 2022. doi: 10.1155/2022/5086622. URL https://doi.org/10.1155/2022/5086622.

K. Kumar and B. P. Pande. Air pollution prediction with machine learning: a case study of indian cities. *International Journal of Environmental Science and Technology*, 20 (5):5333–5348, 2023. ISSN 1735-2630. doi: 10.1007/s13762-022-04241-5. URL https://doi.org/10.1007/s13762-022-04241-5. Received: 2023/05/01.

F. Naz, M. Fahim, A. A. Cheema, N. T. Viet, T.-V. Cao, R. Hunter, and T. Q. Duong. Two-stage feature engineering to predict air pollutants in urban areas. *IEEE Access*, 12: 114073–114085, 2024. doi: 10.1109/ACCESS.2024.3443810.

Olawale Emmanuel Rowland. Comparative analysis of meteorological parameters and their relationship with $no_2$, $pm_{10}$, $pm_{2.5}$ and $o_3$ concentrations at selected urban air quality monitoring stations in krakow, paris, and milan. *Discover Environment*, 2(1):75, 2024. ISSN 2731-9431. doi: 10.1007/s44274-024-00060-2. URL https://doi.org/10.1007/s44274-024-00060-2.

S. Roy, C.M. Rao, and M. Abioui. Evaluation of non-stationary spatial relationship between meteorological-environmental parameters and pm2.5. *Advances in Space Research*, 73(8): 4106–4124, 2024. ISSN 0273-1177. doi: https://doi.org/10.1016/j.asr.2024.01.009. URL https://www.sciencedirect.com/science/article/pii/S0273117724000279.

D. Sanjeev. Implementation of machine learning algorithms for analysis and prediction of air quality. *International Journal of Engineering Research & Technology*, 10(3):533–538, 2021. doi: 10.17577/IJERTV10IS030323.

Kunwar P. Singh, Shikha Gupta, Atulesh Kumar, and Sheo Prasad Shukla. Linear and nonlinear modeling approaches for urban air quality prediction. *Science of The Total Environment*, 426:244–255, 2012. ISSN 0048-9697. doi: https://doi.org/10.1016/j.

scitotenv.2012.03.076. URL https://www.sciencedirect.com/science/article/pii/S0048969712004809.

Hitesh Soneji. Historical weather data for indian cities, 2020. URL https://www.kaggle.com/dsv/1129180.

Gourav Suthar, Rajat Prakash Singhal, Sumit Khandelwal, Nivedita Kaul, Vinod Parmar, and Abhay Pratap Singh. Annual and seasonal assessment of spatiotemporal variation in $pm_{2.5}$ and gaseous air pollutants in bengaluru, india. *Environment, Development and Sustainability*, 26:20629–20652, 2024. doi: 10.1007/s10668-023-03495-4. URL https://doi.org/10.1007/s10668-023-03495-4. Received: 17 August 2022 / Accepted: 9 June 2023 / Published online: 14 June 2023.