Smart Air Quality Forecasting in Jaipur: MLR and SARIMA Techniques for NO_2 Prediction

by

KHYATI SHARDA (2021UCE1564)
RAGHAV AGRAWAL (2021UCE1545)
SHREYA PANDEY (2021UCE1410)

Submitted

In partial fulfillment of the requirements for the degree of

Bachelor of Technology (Civil Engineering)

to the

DEPARTMENT OF CIVIL ENGINEERING
MALAVIYA NATIONAL INSTITUTE OF TECHNOLOGY JAIPUR

May 2025

CERTIFICATE

This is to certify that the project report entitled "**Smart Air Quality Forecasting in Jaipur: MLR and SARIMA Techniques for NO2 Prediction"** which is being submitted by **Raghav Agrawal (2021UCE1545), Khyati Sharda (2021UCE1564), Shreya Pandey (2021UCE1410)**, for the partial fulfillment of the requirements of the degree of **B.Tech. (Civil Engineering)** to the **Department of Civil Engineering, Malaviya National Institute of Technology Jaipur** has been carried out under our supervision and guidance.

(Prof. Sudhir Kumar)
Supervisor
Dept. of Civil Engg.
MNIT JAIPUR

(Dr. Ruchi Sharma)
Supervisor
Dept. of Civil Engg.
MNIT JAIPUR

Acknowledgements

We would like to express our heartfelt appreciation to all those who contributed to the successful completion of our major project and made this journey truly remarkable and memorable.

We are deeply grateful to our supervisors, **Prof. Sudhir Kumar** and **Dr. Ruchi Sharma**, for their constant support, invaluable guidance, and insightful feedback throughout the course of our work. Their mentorship served as a primary source of inspiration, helping us navigate challenges and remain motivated.

Our sincere thanks also go to **Mr. Chabi Kumar**, Ph.D. scholar, and **Mr. Saurav Joshi**, MTech Scholar at MNIT Jaipur, for their continuous encouragement and constructive suggestions. Their unwavering support and dedicated guidance played a significant role in shaping our academic efforts and fostering a mindset focused on growth and improvement.

To all those who played a part in this journey—directly or indirectly—we extend our deepest gratitude.

Shreya, Khyati & Raghav

Abstract

This study aims to forecast nitrogen dioxide (NO_2) levels in Jaipur using MLR and SARIMA models. With air pollution causing major public health risks in fast urbanizing locations such as Jaipur, the goal is to create accurate forecasting methods using environmental and meteorological data. Hourly air quality data from six CPCB monitoring sites (2018-2024) were preprocessed using season-specific mean imputation and strong outlier detection approaches, such as STL decomposition with Median Absolute Deviation. Pearson correlation analysis identified major predictors, including PM2.5, NH_3, CO, temperature, humidity, and wind speed. The MLR model using polynomial features combined with Ridge regularization showed a test R^2 of 0.6454. The SARIMA (1,1,1) x (1,0,0,24) model accurately represented temporal patterns, with a test R^2 of 0.5233 and RMSE of 8.22. Both models produced consistent results in both short-term (24-hour) and medium-term (7-day) predicting timeframes. Winter weather can limit pollution dispersion, leading to higher NO_2 concentrations. This study demonstrates the efficacy of combining statistical modeling with domain-specific environmental insights to improve air quality management and educate policy in expanding urban areas confronting rising pollution concerns.

Table of Contents

Particulars	Page No.
Certificate	i
Acknowledgements	ii
Abstract	iii
Table of contents	iv
List of figures	v
List of tables	vi
Abbreviations	vii
1. Introduction	**8**
1.1 Air Pollution and NO_2	8
1.2 NO_2 emissions in Jaipur	10
1.3 MLR and SARIMA techniques for NO_2 emission	13
2. Literature review	**13**
2.1 Effect of air pollution on the environment	13
2.2 Measurement of NO_2 and its prediction	15
2.3 Machine learning (MLR) and statistical models	16
2.4 Research gaps	18
2.5 Research objectives	19
3. Methodology	**20**
3.1 Site selection	20
3.2 Data preprocessing	22
3.2.1 Data collection	22
3.2.2 Data cleaning	22
3.2.3 Outlier processing	24
3.3 Pearson's correlation analysis	25
3.4 Model development	26
3.4.1 Multiple Linear Regression (MLR)	26
3.4.2 Regularization techniques for linear models	27
3.4.3 SARIMA	30
3.5 Model Evaluation	33
4. Results and discussion	**34**
4.1 Data cleaning & Merging	34
4.2 Seasonal data insights on air pollutants	35
4.3 Parameter Selection	47
4.3.1 Correlation Matrix of parameters	47
4.3.2 Featured Engineering	49
4.3.3 Log transformation of targeted value	50

4.4 Model development 51
4.4.1 Development of MLR 51
4.4.2 Development of SARIMA 53
4.4.3 Forecasting performance of the SARIMA model 55

5. Conclusions 57

6. References 58

List of Figures

Figure No.	Title	Page No.
1.1	NO_2 emission load of different sources in the city of Jaipur	12
1.2	Annual average NO_2 concentration for all CAAQM monitors in Jaipur, 2023	13
1.3	Number of days exceeding guideline and NAAQS for daily average NO_2 concentration of CAAQM monitoring stations	14
2.1	Workflow of Data-Driven Air Pollution Prediction Model	19
2.2	AQI Prediction Model Pipeline for Ahmedabad city	20
3.1	Geographical locations of all six sites	23
3.2	Schematic representation of MLR	28
3.3	Illustration of linear vs non-linear data patterns	29
4.1	Box plots for pollutants vs seasons	37
4.2	Bar plots of seasonal averages for pollutants	39
4.3	Box plots for weather parameters vs seasons	40
4.4	Bar plots of seasonal averages for weather parameters	41
4.5	Daily average pollution levels from 2018 to 2024	42
4.6	Pollutants vs relative humidity	43
4.7	Pollutants vs wind speed & wind direction	44
4.8	Pollutants vs atmospheric temperature	46
4.9	Pearson’s correlation matrix	47
4.10	Feature importance ranking for NO_2 prediction using F-score	50
4.11	NO_2 distribution before and after log transformation	51
4.12	Performance comparison of Regression model for NO_2 predictor	52
4.13	True vs Predicted values - Polynomial + Ridge	53
4.14	ACF and PACF Plots of Differenced Log-transformed series	54
4.15	Actual vs. predicted NO_2 concentration using SARIMA (1,1,1) x (1,0,0,24) model	55
4.16	NO_2 concentration forecast - next 24 hours	57
4.17	NO_2 concentration forecast - next 7 days	58

List of Tables

Table No.	Title	Page No.
4.1	Comparative evaluation of regression models for NO_2 prediction using R^2 and RMSE	39
4.2	Comparative Performance of SARIMA Model Variants for NO_2 Prediction	43
4.3	Forecasted NO_2 concentrations with confidence intervals and error analysis	72

Abbreviations

PM2.5	:	Particulate Matter 2.5
PM10	:	Particulate matter 10
CO	:	Carbon Monoxide
NO_2	:	Nitrogen dioxide
SO_2	:	Sulphur dioxide
O_3	:	Ozone
WS	:	Wind speed
SR	:	Solar radiation
BP	:	Barometric Pressure
RH	:	Relative Humidity
AT	:	Atmospheric Temperature
SARIMA	:	Seasonal Autoregressive Integrated Moving Average
MLR	:	Multiple Linear Regression
AR	:	Auto-regressive component
MA	:	Moving average component
ACF	:	Autocorrelation function
PACF	:	Partial Autocorrelation function
D	:	Seasonal Differencing in SARIMA
WHO	:	World Health Organization
IQR	:	Interquartile Range
CAAQM	:	Continuous Ambient Air Quality Monitoring System
ARIMA	:	Autoregressive Integrated Moving Average
AIC	:	Akaike Information Criterion
RMSE	:	Root mean squared error
MAE	:	Mean Absolute Error
MSE	:	Mean squared error
AQI	:	Air quality index
CPCB	:	Central Pollution Control Board
RSPCB	:	Rajasthan State Pollution Control Board
STL	:	Seasonal-Trend Decomposition
MAD	:	Median Absolute Deviation
BIC	:	Bayesian Information Criterion
R^2	:	Coefficient of determination

1. Introduction

1.1 Air pollution & NO_2

Air quality remains one of the most important environmental issues today. It affects not only humans but also plants and animals around the world. Factors such as higher living standards, economic development, population growth, and increased transportation all contribute to air pollution. The release of complex gases and particles in cities affects public health, agriculture, weather, and climate. These emissions are influenced by factors such as population density, energy consumption, industrial activity, and mode of transportation (Oji & Adamu, 2020). The acceleration of industry and urbanization has worsened air pollution, and the future of pollution control is not bright. Climate change is a major global issue that humanity is currently dealing with. India has proposed a carbon peak and carbon neutrality to solve the climate crisis. Additional study is required to better understand the intricate relationships between atmospheric contaminants and meteorological variables affecting air quality (Khedekar & Thakare, 2023).

More than 80% of urban residents are exposed to air pollution levels that exceed World Health Organization (WHO) limits, and approximately 98% of cities in low- and middle-income countries do not meet recommended air quality norms (Manju et al., 2018). In 2012, air pollution was responsible for over 7 million fatalities globally, with outdoor air pollution exposure accounting for 4.2 million of those deaths (WHO, 2014). Outdoor air pollution causes roughly 3.3 million premature deaths worldwide each year, and if not handled successfully, this amount might increase fourfold by 2050 (Lelieveld et al., 2015).

Vehicles using spark ignition engines emit emissions from the fuel system, engine crankcase, and exhaust. It’s exhaust contains the primary combustion products, carbon monoxide (CO) and water vapor (H_2O). The main pollutants released by gasoline-powered automobiles are lead (Pb), carbon monoxide (CO), hydrocarbons (HCs), and nitrogen oxides (NO_x). Road traffic is the main source of transportation for both passengers and products, particularly in Nigeria's urban areas. The average speed in certain parts of the city is comparatively low due to the high traffic congestion. As a

result, motor vehicles frequently produce significant amounts of CO, H_2S, NO_2, SO_2, and particulate matter (PM). Along with secondary pollutants, mobile air pollutants like lead (Pb), acetaldehyde, formaldehyde, and benzene can harm human health (Oji & Adamu, 2020).

Fossil fuel burning in motor vehicles and stationary sources (heating, power generation) is the main human-caused source of nitrogen oxide emissions into the atmosphere. Under normal circumstances, air oxidants like ozone quickly convert nitric oxide to nitrogen dioxide. A combination of solid, liquid, or solid plus liquid particles suspended in the atmosphere is known as particulate air pollution.

Temperature, humidity, wind speed, and other meteorological factors all play a part in determining the levels of pollutants. The concentration and dispersion of air contaminants are significantly influenced by meteorological factors:

- Temperature: High temperatures can exacerbate the formation of ground-level ozone, particularly in urban areas with high vehicular emissions. During heat waves, pollutants tend to remain trapped near the ground, leading to spikes in air pollution levels.
- Humidity: High humidity can influence the formation of secondary pollutants, such as ozone and fine particulate matter. Humidity levels also affect the removal of pollutants from the atmosphere via wet deposition (rainfall).
- Wind Direction and Speed: Wind is a major factor in the spread of pollutants. Pollutant concentrations in the air can decreased by dispersing and diluting them with the aid of strong winds. On the other hand, quiet winds could lead to the buildup of contaminants in stagnant air masses, which would raise their concentration.
- Rainfall: Through wet deposition, rainfall purges the air of gaseous contaminants and particulate matter, acting as a natural cleanser. Pollutants can, however, linger in the atmosphere and raise concentrations in places with little rainfall.
- Air Pressure: Variations in atmospheric pressure can affect the air's vertical mixing, which can trap contaminants close to the surface. Because low-pressure systems can limit the upward passage of air and contaminants, they are generally linked to poor air quality (Adedokun Taofeek, 2025).

1.2 NO_2 emissions in Jaipur

The Rajasthan State Pollution Control Board (RSPCB) evaluated the air quality of Jaipur by the monitoring stations' data. The results indicated an exceedance of the limits of carbon monoxide (CO), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), and particulate matter (PM_{10} and $PM_{2.5}$) in several parts of the city concerning the national ambient air quality standards. For instance, the average PM_{10} concentration value measured was 150 µg/m³, which is significantly higher than the suggested value of 100 µg/m³. Also, nitrogen dioxide was reported to be at about 50 µg/m³, which is over the accepted value of 40 µg/m³. The study also emphasized how industrial operations, construction dust, and vehicle emissions all contribute to Jaipur's air pollution (Prasad Sharma, n.d.).

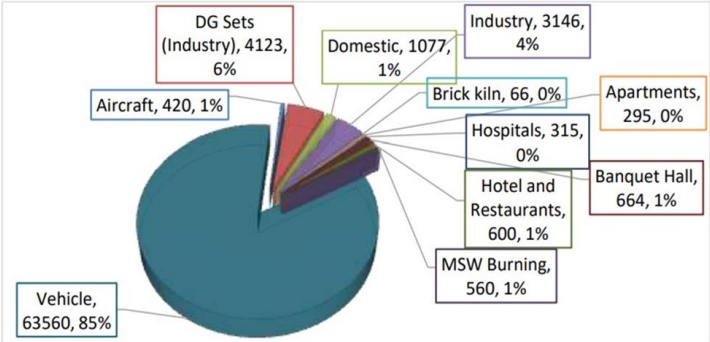

Figure 1.1: NO_2 emission load of different sources in the city of Jaipur (kg/d)

Annual NO_2 concentrations in 2023:

- All six of the air quality monitors under analysis exceeded the health-based WHO guideline, with the highest readings occurring at the Adarsh Nagar monitoring station.
- In 2023, daily NO_2 concentrations were higher than the WHO daily guideline at three stations for more than 60% of the year, and for 277 days at the Adarsh Nagar monitor.

Over the last five years:

- Trends in NO_2 concentrations from ground level and satellite monitors are generally increasing
- Two of three ground-level monitors with sufficient data have a worsening trend, and

• Satellite observations of NO_2 in the atmosphere suggest pollution across the city as a whole is worsening.

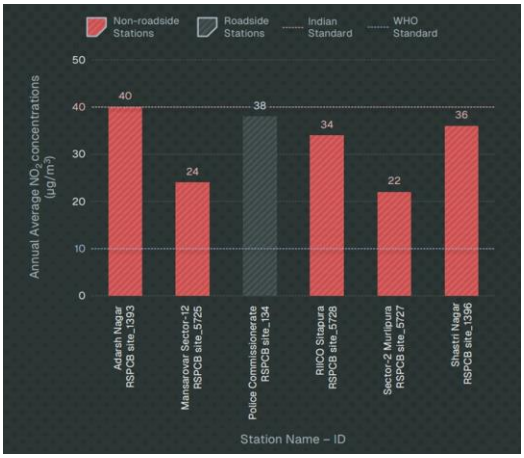

Figure 1.2: Annual average NO_2 concentration for all CAAQM monitors in Jaipur, 2023

Health risks from prolonged exposure to elevated levels of nitrogen dioxide (NO_2) are particularly troubling for young children aged 0-6. This age group represents roughly 12% of Jaipur's population. As Achakulwisut et al. (2019) estimated, in 2015, roughly 2,430 cases of pediatric asthma in Jaipur could be attributed to NO_2 pollution. Data from Continuous Ambient Air Quality Monitoring (CAAQM) stations indicate rising levels of NO_2 over the past five years. Two out of three stations showed significant increases. Moreover, satellite measurements of atmospheric NO_2 over Jaipur from 2019 to 2023 have corroborated NO_2 pollution's intensifying trend across the city (Abbas et al., n.d.).

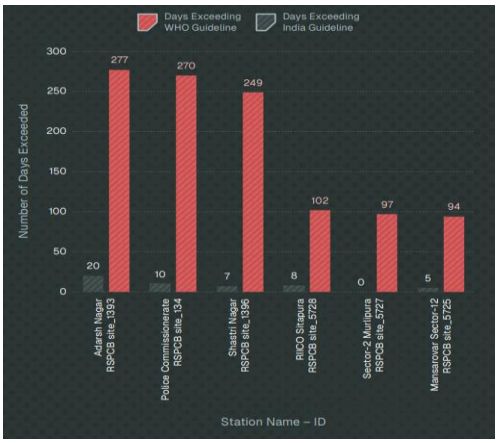

Figure 1.3: Number of days exceeding guideline and NAAQS for daily average NO_2 concentration of CAAQM monitoring stations in Jaipur

1.3 MLR and SARIMA techniques for NO_2 emission

Time series modeling is an essential tool for analyzing the temporal behavior of air pollutants and understanding their seasonal and diurnal patterns. By modeling historical data on pollutant levels and meteorological factors, time series techniques can help forecast future concentrations, enabling early warnings of high pollution events. The ability to predict pollutant levels is crucial for informing public health advisories, guiding regulatory actions, and optimizing traffic management to reduce pollution.

When it comes to air quality forecasting, the AIC, a popular statistical measure that balances model fit and complexity, can be used to compare various time series models, like ARIMA and SARIMA, in terms of their ability to predict pollutant concentrations while minimizing overfitting. A lower AIC value indicates a more efficient model, making it an invaluable tool for choosing the best model to forecast pollutants influenced by meteorological factors.

SARIMA expands the ARIMA model by consolidating regularity, which is regularly shown in discussions of contamination information due to components like climate designs, activity cycles, and regular agricultural activities. SARIMA models incorporate extra regular terms (p, d, q) to account for occasional changes. These regular components permit SARIMA to handle regular variations in toxins such as $PM_{2.5}$ and NO_2, which may display higher concentrations amid certain seasons (e.g., winter exhaust cloud or rural burning seasons). This demonstration makes a difference in progressing the exactness of long-term estimating by capturing both the drift and regular cycles in toxin information. (Taofeek Adedokun, 2025)

2. Literature review

2.1 Effect of Air pollution on the environment

The criteria for the discussion of pollutants are necessary because of their huge impact on public health, requiring critical analysis. Meteorological conditions have a vital role in determining the concentration of pollutants in the air. It is important to know the relationship between meteorological parameters and pollutants to counteract their future impacts, particularly in tropical areas (Putri Nilam, 2024).

Population growth, urbanization, and industrialization have raised energy demand and pollutant discharge rapidly, making pollution a critical urban problem that is associated with health hazards and social injustice. Meteorological conditions also directly affect pollutant concentrations and persistence in the air. This research examines the correlation between prime air pollutants and atmospheric conditions (Shahriar & Parisa, 2023).

Conversation about contamination is a notable matter for environmental and well-being concerns in developing nations such as Iran. Engine motor vehicles are the cause of eighty percent of the pollution that occurs in Iran. Delhi and Tehran are two Asian countries that produce the most notable conversations regarding contamination levels. As a consequence, some arrangements are laid out to deal with discourse of pollution, such as the purification and standardization of vehicle fuel, improvement of motors with efficient fuel consumption, inconvenience of punishment for firms that pollute the environment, and application of a moo charge strategy in green innovation, among other activities that are suggested (Ghorani et al., 2016).

Normally, six main air pollutants are extensively researched and compared. These pollutants may find their way into the air through both natural processes—e.g., sandstorms and forest fires—and man-made processes such as industrial operations, wood, straw, and fossil fuel burning, and motor vehicle emissions (Amanollahi et al., 2013; Ke et al., 2019; Lavin et al., 2012).

$PM_{2.5}$, PM_{10}, O_3, NO_2, SO_2, and CO were the predicted air pollutants in this research. The pollutants are known to produce a range of health issues upon longer-term exposure. The complex source lists needed by the chemical models utilized to predict

air pollution make them difficult to utilize. Even though machine learning models have also been employed, they are of no use due to their reliance on meteorological features, since modeling meteorology involves the utilization of other models and facilities. When fossil fuels containing nitrogen and sulfur are burned excessively, harmful air pollutants called NO_2 and SO_2 are produced (Salonen et al., 2019; Al-Naimi et al., 2015; Chen et al., 2007; Liu et al., 2016).

Shen et al. (2020) point out the deleterious effects of nitrogen dioxide (NO_2) and sulfur dioxide (SO_2). Several authors have pointed out these gases as extreme respiratory irritants that play a part in causing asthma and respiratory inflammation (Salonen, Salthammer & Morawska, 2019; Al-Naimi, Balakrishnan & Goktepe, 2015; Chen et al., 2007; Liu et al., 2016). In addition, NO_2 and SO_2 easily react with atmospheric chemicals, resulting in the production of secondary pollutants like acid rain and fine particulate matter ($PM_{2.5}$ and PM_{10}) (Salonen et al., 2019; Al-Naimi et al., 2015; Chen et al., 2007; Liu et al., 2016). Worth mentioning is that NO_2 has an important role in ozone (O_3) formation and photochemical smog generation through several chemical processes (Al-Naimi et al., 2015; Salonen et al., 2019).

One way to represent periodic patterns in time series data is through the use of cyclic encoding. Cyclic encoding can assist models such as SARIMA in better understanding seasonal fluctuations and trends by encoding the cyclical behavior of the data. This is particularly useful for data that reflects consistent, repeat trends over pre-specified intervals, e.g., daily traffic volumes or monthly sales data. Regular cycles that occur at repeated intervals, such as daily, weekly, or yearly cycles, are referred to as seasonal patterns. Cyclic encoding assists SARIMA models in identifying and forecasting these repeating trends by encoding these patterns (Devianto et al., 2024).

2.2 Measurement of NO_2 and its prediction

By 2017, the People's Republic of China Ministry of Environmental Protection (MEP) had already set up 1,497 environmental monitoring stations across the country to assess air quality. The National Urban Air Quality Real-time Release Platform (NUAQRRP) provides information on the Air Quality Index (AQI) and hourly average concentrations of major air pollutants such as NO_2, $PM_{2.5}$, PM_{10}, sulfur dioxide (SO_2), ozone (O_3), and carbon monoxide (CO). This research employs ground-based NO_2 concentration data

from 2018 to 2020. The analysis synchronizes TROPOMI satellite overpass times with matching average ground-level NO_2 measurements between 13:00 and 14:00 local time (LT) (Liu et al., 2007; Wang & Zhao, 2017). Statistical filtering was used to ensure data quality by excluding values greater than three times the standard deviation. Ground-level NO_2 values between 1.0 μg/m³ and 300 μg/m³ were then chosen for model development.

2.3 Machine learning (MLR) and statistical models

Two-step estimation of ground-level NO_2 concentrations using satellite data has already been accomplished with the aid of machine learning techniques. To identify the global relationship between ground-measured NO_2 and influencing variables, the initial step is to establish a regression application model (Chen et al., 2019; de Hoogh et al., 2019). To train and test the model, the sample data are divided into training and test datasets. Parameter optimization is subsequently utilized to identify the optimal regression model. Using the regression model and inputting the appropriate data for an application analysis to calculate the estimated outcome is the second process.

The forecasting framework employed a sophisticated ensemble learning method called stacked generalization (SG) that was designed to maximize the system's ability to represent nonlinear relationships and reduce errors in generalization. SG, originally reported by Wolpert (1992), works by aggregating several models to minimize the bias of individual models. Particularly, the predictions made by each of the base learners are stacked and utilized as input for a meta-learner that is trained with cross-validation. In this research, the meta-learner is either a Ridge regression or a Support Vector Regression (SVR) model, whereas the base learners are combinations of five sub-models: Multiple Linear Regression (MLR), Multi-Layer Perceptron (MLP), Random Forest (RF), Gradient Boosted Decision Trees (GBDT), and SVR. This makes a total of 52 SG models. In light of predictive performance and computational efficiency, no other models were chosen as meta-learners (Ke et al., 2022).

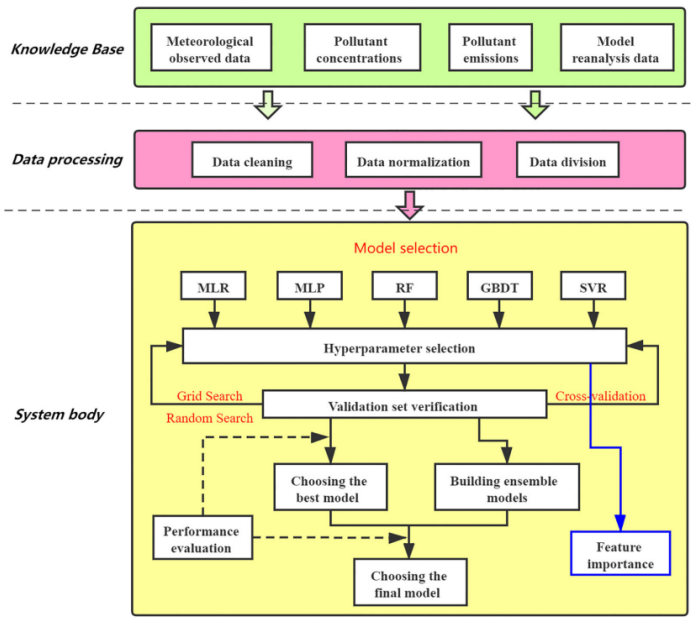

Figure 2.1: Workflow of Data-Driven Air Pollution Prediction Model

To analyze the performance of the model, several statistical measures were computed, such as Pearson's Correlation Coefficient, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and coverage. Pearson's Correlation Coefficient was employed to investigate the degree of overfitting or underfitting of the model. A value of about 0.5 is ideal, as this indicates that the model is neither overfitting nor properly capturing the overall trend. As defined by Shen et al. (2020), MSE is the mean of the squared error between predicted and observed values, and RMSE is the square root of the MSE. MAE is the mean absolute error between predicted and observed values. One of the primary differences between RMSE and MAE is that RMSE weights larger errors more than MAE (Shen et al., 2020).

2.4 Research Gaps

Existing research focuses primarily on metro cities (Delhi, Mumbai, and Kolkata), leaving growing urban areas such as Jaipur unexplored despite rising urbanization and pollution levels. Advanced feature selection and preprocessing approaches (e.g., cyclical encoding, seasonal decomposition) that are suited to Jaipur's specific climate and data patterns are rarely used. There is a lack of studies employing machine learning to estimate NO_2 levels in mid-tier cities using meteorological data. In Jaipur, there is a

lack of forecasting models for NO_2, underlining the necessity for comprehensive health-impact research employing strong modeling methodologies.

2.5 Research Objectives

Based on the identified research gaps, the major objective of our study is-

"Smart Air Quality Forecasting using MLR and SARIMA Techniques in Jaipur City"

And to achieve this, the sub-objectives are

- To collect and preprocess air quality (NO_2) and meteorological data from CPCB monitoring stations in Jaipur, addressing missing values from RSPCB and seasonal trends for robust analysis.
- To analyze correlations between pollutants and weather parameters using Pearson's correlation analysis, identifying key predictors for modeling.
- To develop multiple linear regression (MLR) and seasonal SARIMA models for evaluating meteorological impacts and forecasting NO_2 levels, incorporating both temporal and seasonal patterns.
- To evaluate model performance using metrics (R^2, RMSE, MAE) and determine the optimal approach for NO_2 prediction in urban settings.

Overall, this project introduces sophisticated parameter selection techniques to NO_2 prediction in Jaipur, addressing a significant gap in regional air quality modeling that has previously focused only on major metros.

3. Methodology

3.1 Site Selection

Rajasthan, the most extensive state in India in terms of area, accounts for about 10.4% of the country's total land area. It is geographically situated in the western region of the country, stretching between latitudes 23°30′ to 30°11′ N and longitudes 69°29′ to 78°17′ E (Barupal et al., 2022). Jaipur, the capital of Rajasthan, is situated around 260 kilometers southwest of New Delhi on a sandy, triangular arid plain in northern India. According to Jaipur Nagar Nigam (2024), the city is situated between 26°46′ N to 27°01′ N latitude and 75°37′ E to 76°57′ E longitude. Jaipur receives approximately 650 mm of average annual rainfall, which takes place mainly during July to September months of monsoon. Wind speeds in the city typically range from 2.5 to 10.0 km/h, with summer months witnessing the highest speed (6–10 km/h) and a notable reduction during winter (Dadhich et al., 2018). Six monitoring sites were chosen for the study: Site-1 (26°54′9″ N, 75°50′12″ E), Site-2 (26°54′45″ N, 75°47′12″ E), Site-3 (26°56′29″ N, 75°47′50″ E), Site-4 (26.7751° N, 75.8514° E), Site-5 (26.9776° N, 75.7639° E), and Site-6 (26.8505° N, 75.7628° E). Jaipur, with a population of 467 square kilometers, is the most populous city in Rajasthan and is almost surrounded by hills, apart from the south. The population is expected to be around 4.33 million by 2024 (Rajasthan Population Census, 2011). The city has experienced a sustained rise in air pollution as a result of increased industrial activity and increasing residential and commercial areas, which has made air quality management more difficult. Based on these concerns, Jaipur was chosen for the current study. Hourly air quality data were obtained from six monitoring stations: Shastri Nagar (Site 1), Police Commissionerate Office (Site 2), Adarsh Nagar (Site 3), Mansarovar (Site 4), Sitapura (Site 5), and Murlipura (Site 6). Sites 1–3 had longitudinal data from the years 2018 to 2024, and Sites 4–6 provided data from the years 2023–2024. The main dataset was obtained from the Central Pollution Control Board (CPCB), and missing values were imputed using data from the Rajasthan State Pollution Control Board (RSPCB).

The city's geographic coordinates and the locations of the three sites chosen for this investigation are displayed in the figure.

Figure 3.1: Geographical locations of all six sites

3.2 Data Preprocessing

3.2.1 Data Collection

The data for this study were obtained from the Central Pollution Control Board (CPCB) and the Rajasthan State Pollution Control Board (RSPCB) for six monitoring sites in Jaipur. The dataset contains a variety of environmental factors, including air temperature (AT), relative humidity (RH), particulate matter (PM2.5 and PM10), and other meteorological and pollutant variables. The time range and data availability varied per site, with three sites having more than 61,368 records and the other three having roughly 17,545 records apiece. All data were collected at regular intervals, although there were gaps and anomalies caused by sensor failures or transmission issues. These raw datasets formed the basis for future cleaning, imputation, and analytic procedures.

3.2.2 Data cleaning

The dataset shows the number of null entries and the percentage of missing data. The missing percentage was calculated by multiplying the number of null instances by the total number of data points for each attribute. These values were counted using the Python computer language, and the data was shown using heat maps. Identifying usable information from enormous datasets is extremely tough and complex. In many real-world data sets, there was always a requirement for dataset visualization to check out

specific information rather than doing it manually or using another less efficient method. Data visualization is a valuable tool for increasing the ability to understand and engage with large, complicated datasets (Sadiku et al., 2016). It involves presenting facts graphically or pictorially, which makes it much easier to interpret. These tools can be divided into three categories: spreadsheets, programming libraries, and data visualization software (Keim et al., 1996). Spreadsheets commonly use scatter plots, bar charts, and line graphs; programming libraries include matplotlib. These tools include advanced visualization capabilities, such as heatmaps, network diagrams, and interactive dashboards (Ramloll et al., 2004; Srivastav, 2023). To visualize missing data in our dataset, we used heat maps to develop a suitable strategy for dealing with the missing data. The initial parameters investigated to find the missing values were NO_2, NH_3, SO_2, CO, O_3, PM_{10}, $PM_{2.5}$, benzene, toluene, ethylene, MP xylene, oxyline, TC, RH, AP, and sunlight. All of these characteristics were plotted using the Matplotlib and Seaborn libraries. The datasets from the six sites were evaluated for null or missing values using Python programming, and missing data percentages ranged from 15 to 25%. The 2018 dataset contained the majority of the missing values. As a result, null data statistics were calculated and applied to the remaining data preparation steps. Imputation methods entail filling in missing data to produce a complete data matrix that can be evaluated using standard procedures. There are several ways for dealing with missing data, such as eliminating all incomplete data points, filling in missing values with the most prevalent ones, or replacing null entries with the mean, median, or other relevant statistics. (Zhang 2012). Single imputation methods include replacing a missing data element with a single value without giving a model for the partially missing data. In this study, missing data points were imputed with mean values. The mean substitution approach replaces missing information with the variable's average score.

The mean value, often known as the arithmetic mean, is a fundamental concept in statistical analysis. The mean is a statistical metric that represents the arithmetic average of a dataset. It is widely used to summarize and describe the central tendency of distributions. A recent study looked at several components of the average value, including outliers that can have a significant impact on the mean, perhaps leading to inaccurate interpretations. This characteristic is referred to as sensitivity to outliers.

Scientists have proposed powerful alternatives, such as the trimmed mean and the Winsorized mean, that are less affected by outliers (Rousseeuw & Hubert, 2011).

Three sites contain 61,368 entries, while others have 17,544. Sorting by timestamp with date-time object. Jaipur time series data is cleaned using the seasonal mean approach.

3.2.3 Outlier Processing/Missing Value Handling

The dataset for Jaipur city contains a large number of missing values that need to be addressed before analysis. Simply eliminating rows with missing values would have led to severe data loss and imprecise results (Emmanuel et al., 2021). To solve this issue, we used a variety of imputation algorithms adapted to the data features. We initially determined the amount of missing data in each column. We excluded columns with missing values greater than 30% from the dataset because they were deemed unreliable for meaningful analysis.

For the remaining columns with acceptable levels of missing data, we used mean imputation to fill the gaps. Rather than utilizing the global mean, which would have ignored seasonal changes, we employed a season-specific mean substitution strategy. This approach substituted null values with mean values obtained only from the same season, retaining the data's natural seasonal variability. For the Ambient Temperature (AT) variable, which had 2,928 missing values, we used linear interpolation to estimate the missing entries based on nearby data points, as this method was more appropriate for the continuous nature of temperature data.

To handle outliers that may skew modeling findings, we developed a multi-method identification technique based on time series features. Our major method was Seasonal-Trend Decomposition (STL), which decomposed each variable into seasonal, trend, and residual components during a 24-hour period, with 13 seasonal components. Outliers were detected when normalized residuals derived using Median Absolute Deviation (MAD) surpassed a 3-sigma threshold. When STL could not be applied owing to data restrictions, we used two fallback methods: a rolling Interquartile Range (IQR) technique employing a one-week frame with values outside $Q_1 - \left(\frac{3}{1.35}\right)$*IQR or $Q_3 + \left(\frac{3}{1.35}\right)$*IQR highlighted as outliers; and a rolling Z-score method that found points

deviating more than 3 standard deviations from the rolling mean. Instead of removing outliers, we replaced them with NaN values to keep the time series structure while eliminating their influence. To confirm our approach, we documented the percentage of outliers removed from each variable and compared descriptive statistics before and after processing. This methodical process secured the dataset's integrity and reliability while preserving the temporal continuity required for time series analysis.

3.3 Pearson's Correlation Analysis

Understanding the complexity of data involves choosing the right input variables, which can be done through various statistical approaches. Within this research, r value was employed to identify input parameters for the prediction of NO2. The r value is a measure of statistical significance that is used to find the relationship between two variables on the same interval or ratio scale. It measures how strong two variables are. The r value can be obtained by using Equation 2 (Mao et al., 2021):

$$r = \frac{Cov(X,Y)}{\sigma x \, \sigma y}$$

Cov (X.Y) represents the covariance of two variables, while σ_x σ_y represent the standard deviation of two variables, X and Y.

The values of the coefficients range between +1 and -1, where +1 signifies a perfect positive correlation, -1 signifies a perfect negative correlation, and 0 signifies that there is no correlation whatsoever. Degree of correlation:

- Perfect: If the value is nearly 1, the correlation is referred to as perfect, meaning that when one variable increases, the other will also increase (if positive) or decrease (if negative).
- High correlation is defined as a coefficient value between 0.50 and 1.
- Moderate: A correlation with a value between 0.30 and 0.49 is considered moderate.
- Low: A correlation is termed low degree if its value is less than +.29 and more than -.29.
- A value of zero indicates that there is no correlation.

To show the relationship between the various meteorological parameters, like temperature, rainfall, wind speed, UV index, and the pollutants considered.

3.4 Model Development

3.4.1 Multiple Linear Regression (MLR)

According to Chen et al. (2019), multiple linear regression (MLR) is a reliable and comprehensible statistical method for determining the connection between ground-level NO_2 concentrations and possible explanatory factors. The cleaned datasets now offer trustworthy ground-truth values that are necessary for creating precise MLR models for the city of Jaipur.

The resulting cleaned NO_2 datasets offer a strong basis for precise air quality prediction in the various urban settings of Jaipur. The data now accurately depicts the dynamics of NO_2 concentrations at each monitoring point after anomalies were effectively eliminated while maintaining natural temporal patterns.

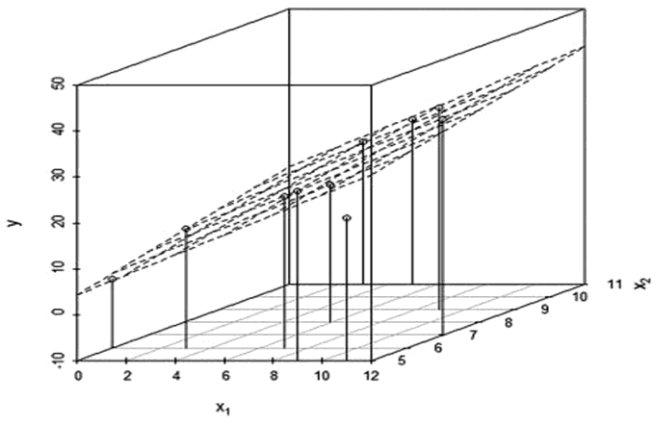

Figure 3.2: Schematic representation of MLR

The equation for multiple linear regression is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$$

- Y is the dependent variable.
- $X_1, X_2, \cdots X_n$ = independent variables(predictor)
- β_0 = Intercept term

- $\beta_1,\beta_2,\cdots\beta_n$ = coefficients of predictor variables
- ε = Error term (residuals)

The algorithm is intended to find the equation of the best-fitting line that can be used to predict values using independent variables. The regression model learns a function from the data (which includes known X and Y values) and uses it to predict Y values for unknown X.

3.4.2 Regularization Techniques for Linear Models

1. Baseline linear regression

Simple linear regression indicates the relationship between a dependent variable (input) and an independent variable (output). Mostly, this type of regression explains the following:

- The relationship strength between the given variables

Example: Correlation between temperatures and increased levels of pollution.

- The value of the dependent variable is based on the value of the independent variable.

Example: The value of the pollution level at a given temperature.

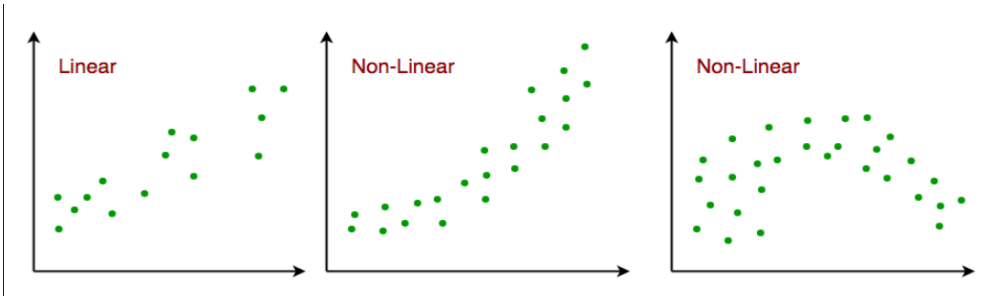

Figure 3.3: Illustration of Linear vs Non-Linear Data Patterns

2. Ridge Regression (L_2 Regularization)

Ridge regression is a linear regression method that adds a regularization factor to the conventional least squares goal. Its main objective is to penalize high coefficient values in the model in order to prevent overfitting. This is especially effective when features are highly correlated, as ridge regression manages multicollinearity and decreases model complexity. It employs an L_2 penalty, which decreases the regression

coefficients towards zero without setting them exactly to zero, retaining all predictors in the model while controlling their influence.

The objective function after applying ridge regression is:

$$\min\{\|\mathbf{y}-\mathbf{X}\mathbf{w}\|^2+\lambda\|\mathbf{w}\|^2\}$$

- $\|\mathbf{y}-\mathbf{X}\mathbf{w}\|^2$ is the sum of squared errors (to measure fit),
- $\|\mathbf{w}\|^2 = \sum_{j=1}^{p} w_j^2$ is the L_2 norm, penalizing large coefficient values.

3. Lasso Regression (L_1 Regularization)

Lasso regression is a regularization method for linear regression that lessens overfitting by adding a penalty term to the model's objective function. Its primary objective is to achieve equilibrium between forecast accuracy and model simplicity. Using L_1 regularization, Lasso efficiently performs feature selection by reducing the coefficients of less significant features to precisely zero. This simplifies the model by removing extraneous variables while still providing decent predicted performance.

The objective function after applying lasso regression is:

$$\min\{\|\mathbf{y}-\mathbf{X}\mathbf{w}\|^2+\lambda\|\mathbf{w}\|_1\}$$

- $\|\mathbf{y}-\mathbf{X}\mathbf{w}\|^2$ is the residual sum of squares (measures model fit),
- $\|\mathbf{w}\|_1=\sum_{j=1}^{p} |\mathbf{w}_j|$ is the L_1 norm, which penalizes the absolute values of the coefficients.

4. Elastic Net Regression

Elastic Net Regression is a regularization technique that applies both L_1 (Lasso) and L_2 (Ridge) penalties to a linear regression model. This hybrid strategy makes use of the advantages of both approaches: Ridge's multicollinearity handling and Lasso's feature selection. Because it promotes group selection of associated characteristics rather than selecting one at random, Elastic Net works especially well with datasets that contain a large number of correlated predictors. To improve model performance and generalization, it offers a more flexible and reliable method by balancing the effects of L_1 and L_2 regularization.

The objective function after applying elastic net regression is:

$$\min\{\|y-Xw\|^2+\lambda[\alpha\|w\|_1+(1-\alpha)\|w\|^2]\}$$

- $\|y-Xw\|^2$ is the residual sum of squares.
- $\|w\|_1 = \sum_{j=1}^{p} |w_j|$ is the L_1 norm (Lasso),
- $\|w\|^2 = \sum_{j=1}^{p} w_j^2$ is the L_2 norm (Ridge).

4. Polynomial + Ridge Regression

Polynomial Regression with Ridge Regularization goes beyond linear regression by using polynomial features (such as squared terms and interaction terms) to capture non-linear correlations between variables. While polynomial transformations boost model flexibility, they also introduce a high number of features, which can result in overfitting. Ridge Regression (L_2 regularization) is used to address this, penalizing big coefficients and assisting in the control of model complexity. The best regularization strength (alpha) determined from prior cross-validation is reused in this strategy, although cross-validation is not repeated after polynomial transformation. This method allows the model to learn complex patterns while yet maintaining generalization performance.

The objective function after applying polynomial + ridge regression is:

$$\min\{\|y-\Phi(X)w\|^2+\lambda\|w\|^2\}$$

- The first term is the sum of squared errors (data fitting term).
- The second term is the L_2 regularization penalty.
- λ controls the strength of the regularization.

3.4.3 SARIMA (Seasonal Autoregressive Integrated Moving Average)

SARIMA models are sophisticated, advanced statistical approaches built primarily for analyzing and forecasting time series data with both trend and seasonal characteristics. SARIMA offers a complete technique for modeling complicated temporal dynamics in environmental data by expanding the standard ARIMA methodology to include seasonal components. SARIMA provides an analytical method that is especially useful

for monitoring air pollution, because concentrations are greatly impacted by cyclical elements like daily traffic patterns, weekly cycles of human activity, and seasonal weather fluctuations. This statistical model can capture both long-term trends and repeating patterns in pollutant concentrations in urban areas, making it particularly useful for forecasting future air quality conditions and analyzing atmospheric pollutants' temporal behavior. The sections below explain the mathematical structure, component interpretation, and implementation methods of the SARIMA model for NO_2 concentration data.

The SARIMA model is denoted as SARIMA (p, d, q) × (P, D, Q) s, where:

■ Basic Components:

- **AR (Autoregressive) Term (p)**: Captures the relationship between a current observation and its previous values.
  Example: If pollution levels were high yesterday, they are likely to be high today as well.
- **MA (Moving Average) Term (q)**: Models the relationship between an observation and past forecast errors.
  Example: If previous forecasts underestimated pollution, the model adjusts accordingly in future predictions
- **Differencing (d)**: Removes underlying trends by analyzing differences between consecutive observations, enhancing stationarity.
  Example: Helps stabilize the mean of a time series by eliminating trend components.

■ Seasonal Components:

- **Seasonal AR Term (P)**: Accounts for correlations between observations separated by seasonal lags
  Example: Pollution at 8 AM today may resemble pollution at 8 AM yesterday.
- **Seasonal MA Term (Q)**: Models the effect of past seasonal errors on current observations.

  Example: If recent mornings have shown unexpected spikes in pollution, the model adapts accordingly.

- **Seasonal Differencing (D)**: Removes recurring seasonal patterns by analyzing changes between seasonal periods.
  Example: Eliminates repeated daily or monthly cycles to better capture underlying trends.

The mathematical representation of the SARIMA model is expressed as:

$$\phi_p(B)\,\Phi_P(B^S)(1-B)^d(1-B^S)^D Z_t = \theta_q(B)\,\Theta_Q(B^S)a_t$$

Where:

1. Ordinary Autoregressive Component: $\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$
2. Ordinary Moving Average Component: $\theta_q(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q$
3. Seasonal Autoregressive Component: $\Phi_P(B^S) = 1 - \Phi_1 B^S - \Phi_2 B^{2S} - \cdots - \Phi_P B^{PS}$
4. Seasonal Moving Average Component: $\Theta_Q(B^S) = 1 + \Theta_1 B^S + \Theta_2 B^{2S} + \cdots + \Theta_Q B^{QS}$

Where B is the backshift operator, and Z_t represents the time series after appropriate transformation (Box et al., 2008; Dubey et al., 2021).

Model Selection: AIC and BIC

When dealing with SARIMA models, it is critical to pick the appropriate configuration of parameters (p, d, q) x (P, D, Q). Two crucial metrics aid in this choosing process:

Akaike Information Criterion (AIC)

The Akaike Information Criterion (AIC) is a statistical metric applied to model comparison and selection. It measures the balance between a model's goodness of fit and complexity, penalizing models with higher parameters to avoid overfitting. AIC is highly useful for model comparison across different models, including ARIMA, SARIMA, ARIMAX, and hybrid models. Lower values of AIC indicate a more efficient model since it fits better with fewer parameters. For ARIMA and SARIMA models, AIC is obtained by initially running the model on the data and then computing the likelihood of the model predictions.

Bayesian Information Criterion (BIC)

The Bayesian Information Criterion (BIC) is a crucial model selection metric that is comparable to the AIC but places a greater focus on simplicity. BIC successfully selects the most parsimonious model by imposing a higher penalty for model complexity, particularly as sample size increases. While AIC prioritizes predictive performance, BIC selects the model that is most likely to reflect the genuine underlying data generation process. In practice, models with lower BIC values are preferable because they provide a better mix of goodness-of-fit and model simplicity.

3.5 Model Evaluation

In this study, the R^2 Score, Mean Squared Error, and Root Mean Squared Error (RMSE) metrics were used. These validation measures are used to verify whether regression models are correct or inaccurate.

Mean absolute error (MAE)

Mean absolute error (MAE) is a metric used to quantify mistakes between paired data (Sammut, Claude Webb, 2010). The error increases with the MAE. The difference between the actual value (Yi) and the predicted value ($\hat{Y}i$) is known as the error, and n is the total number of observations.

$$\text{MAE} = \frac{\sum_{i=1}^{n} |Yi - \hat{Y}i|}{n}$$

Root mean square error (RMSE)

RMSE is a measure to indicate how closely a regression line approximates the data points (Nevitt and Hancock, 2000). Another way to define RMSE is as the standard deviation of residuals. In a regression using variables observed over T periods, the RMSE of values forecast at time t of the dependent variable Yt is calculated using the square root of the mean of the squares of the errors.

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{T} (\widehat{Yt} - Yt)^2}{T}}$$

R^2 score

The R^2 Score (Chicco et al., 2021) measures the percentage of the variation in a scenario where the independent variables set the dependent variable. The coefficient of determination (R^2 Score) gives more accurate and informative scores without interpretational challenges. The coefficient of determination (R^2 or r^2) measures the variance percentage in the dependent variable that is predictable from the independent variable and is read "R squared". The formula of R^2 is presented as

$$R^2 = 1 - \frac{sum\ squared\ regression\ (SSR)}{total\ sum\ of\ squares\ (SST)}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(Yi - \hat{Y}i)^2}{\sum_{i=1}^{m}(Yi - \hat{Y})^2}$$

Where $\hat{Y}_i$ is the predicted i^{th} value, Y_i is the actual i^{th} value, and $\hat{Y}$ is the mean of actual values, and m is the total number of observations.

4. Results and discussions

4.1 Data Cleaning & Merging

All six monitoring stations (Adarsh Nagar, Sitapura, Murlipura, Mansarovar, Police Commissionerate, and Shastri Nagar) had their NO_2 datasets thoroughly cleaned to guarantee uniformity and preparedness for time series modeling. When used inside the seasonal portions of each year, seasonal mean imputation was quite successful in filling in the missing numbers. Despite the 30% imputation threshold we established, almost all missing records could be recovered using seasonal division, preserving the data's temporal structure.

Depending on the station-specific data distribution, outliers were treated using a combination of STL decomposition (using LOESS), IQR, and Z-score approaches. About 15–20% of outliers per station were eliminated by this multi-step procedure, which mostly consisted of anomalous spikes or sensor faults. A strong basis for precise forecasting was provided by the final cleaned datasets, which nevertheless included significant daily patterns and seasonal trends.

During the data cleaning step, we discovered that the merged dataset consisted of 61,368 records. However, we discovered that approximately 2,928 items had missing values for air temperature, accounting for nearly 4.7% of the dataset. Linear interpolation was used to fill this gap because temperature is a crucial meteorological component affecting pollution levels. This strategy allowed us to estimate and fill in missing temperature values using surrounding data points, ensuring the dataset's continuity and dependability for subsequent modeling and analysis.

4.2 Seasonal Data Insights on Air Pollutants

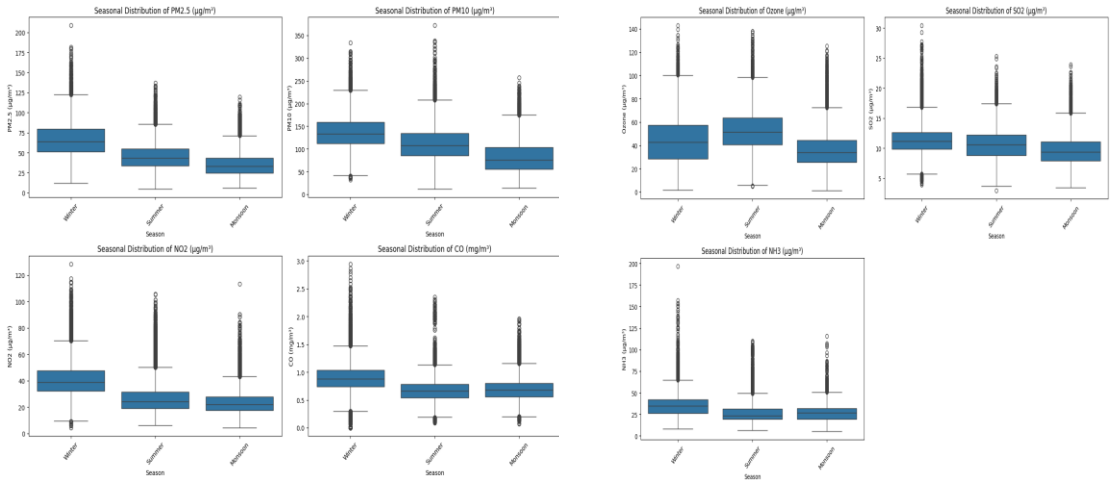

Figure 4.1: Box plots for pollutants vs seasons

We see distinct seasonal behaviors across the seven key pollutants:

- Winter Peak: Pollutants like $PM_{2.5}$, PM_{10}, NO_2, and CO consistently show higher mean concentrations during the winter months (likely December to February in Jaipur). The box plots also suggest a wider spread and potentially higher extreme values during this season for some of these pollutants. This indicates that air quality tends to be poorer in the winter.
- Summer Peak: Ozone exhibits an opposite trend, with higher concentrations and a higher median during the summer months (likely April to June in Jaipur).
- Relatively Consistent Distribution: SO_2 appears to have a more stable distribution across the seasons, with median levels and the spread of data being somewhat similar in Winter, Summer, and Monsoon.
- Elevated in Winter and Monsoon: NH_3 shows a tendency for slightly higher median concentrations and a wider spread in both the Winter and Monsoon seasons compared to Summer.

Bar Plot Analysis of Seasonal Averages in Air Quality Data:

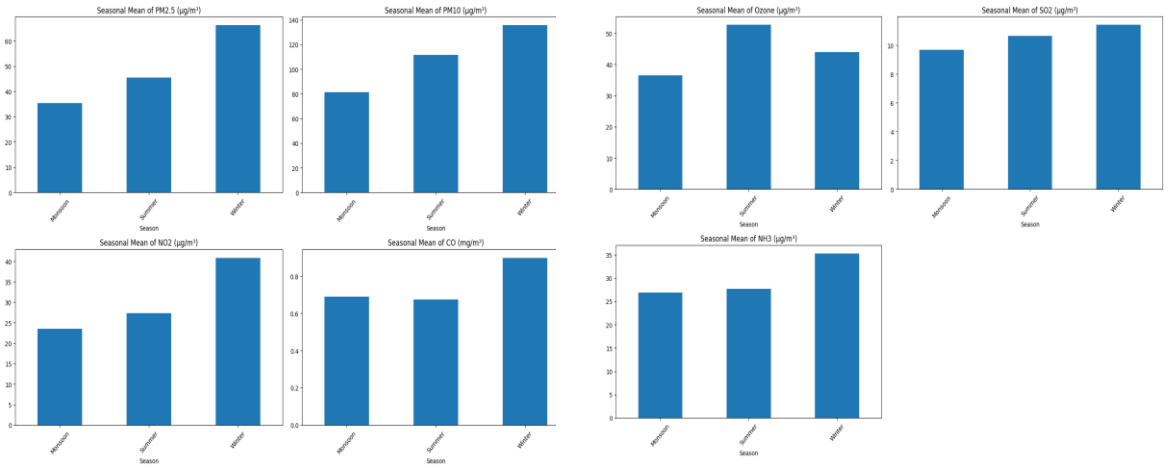

Figure 4.2: Bar plots of seasonal averages for pollutants

These trends underline the need for season-specific air quality management strategies, especially during the winter months when public health risks are elevated due to higher pollution levels

- Winter emerges as the most polluted season across all major pollutants (PM2.5, PM10, NO_2, and CO), showing significantly higher concentrations compared to summer and monsoon. This is likely due to a combination of lower wind speeds, temperature inversions, and increased emissions from heating sources.
- NO_2 concentrations also follow a similar seasonal pattern, being lowest during the monsoon, slightly higher in summer, and peaking in winter. This indicates that traffic and industrial emissions, which are consistent sources of NO_2, have a more pronounced impact when meteorological conditions in winter trap pollutants near the ground.
- The seasonal variation suggests a strong influence of meteorological factors on pollutant dispersion and accumulation. Reduced atmospheric mixing in winter allows pollutants to accumulate, while higher temperatures and precipitation in other seasons aid in their dispersion or removal.

Box plot analysis of the Seasonal mean of weather parameters

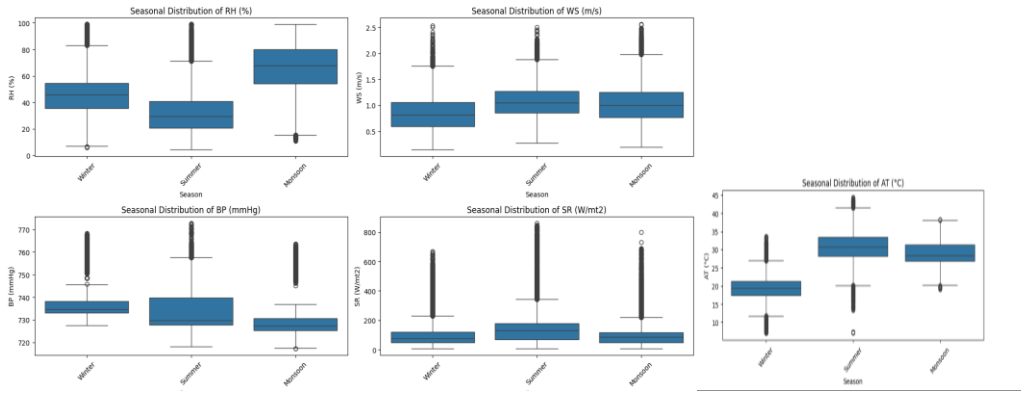

Figure 4.3: Box plots for weather parameters vs seasons

- Lower wind speeds, especially in winter, contribute to higher NO_2 levels due to limited dispersion. When wind movement is weak, NO_2 emitted from vehicles and industries tends to accumulate in the lower atmosphere, increasing overall concentration.
- High atmospheric pressure and low temperatures in winter create stable atmospheric conditions that trap NO_2 near the surface. The boxplots indicate that winter exhibits both higher pressure and lower temperatures compared to other seasons, aligning with periods of peak NO_2 concentrations.
- Reduced solar radiation in winter slows down the breakdown of NO_2, since sunlight is essential for photochemical reactions that convert NO_2 into other compounds. This results in prolonged atmospheric presence of NO_2, compounding pollution during the colder months.

Bar plots of the seasonal mean of weather parameters

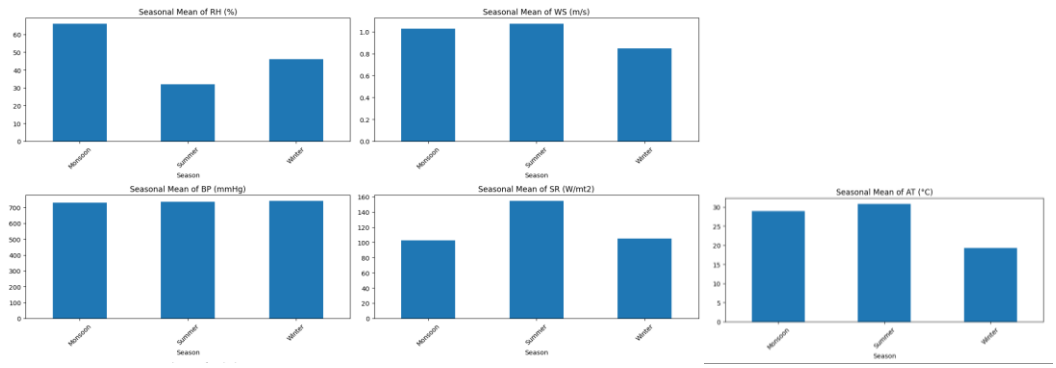

Figure 4.4: Bar plots of seasonal averages for weather parameters

- NO_2 levels show a clear seasonal pattern, with concentrations peaking during the winter season. This suggests a strong seasonal influence of weather on pollutant behavior.
- Collectively, meteorological parameters in winter—such as lower temperature, weaker winds, and atmospheric stability—create unfavorable dispersion conditions, causing NO_2 to accumulate more compared to other seasons.
- These patterns highlight the crucial role of weather conditions in regulating NO_2 levels, suggesting that even without significant changes in emissions, pollution can intensify simply due to seasonal meteorological variations.

Daily average pollutant levels

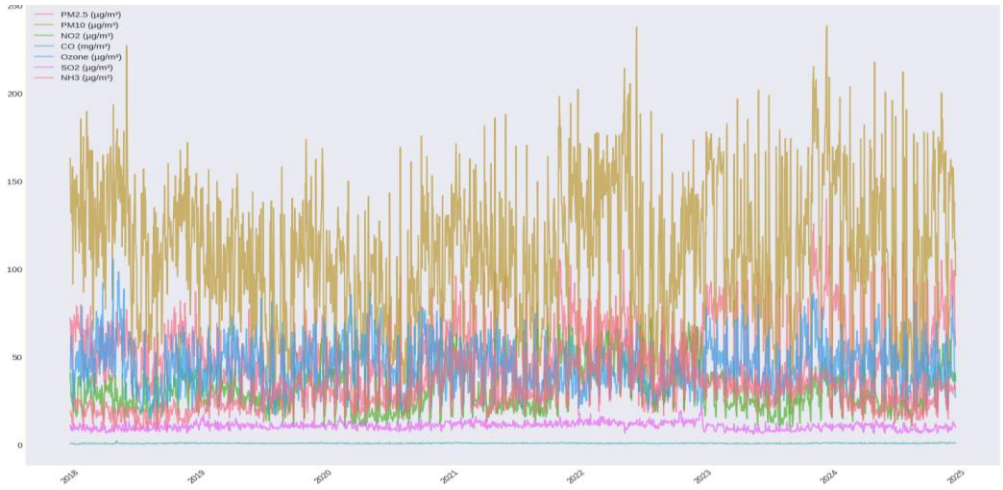

Figure 4.5: Daily average pollutant levels from 2018-24

This graph displays the daily average levels of seven key air pollutants in Jaipur, Rajasthan, India, from the beginning of 2018 to 2024. Each colored line represents a different pollutant, and the y-axis shows the concentration levels in their respective units (µg/m³ for $PM_{2.5}$, PM_{10}, NO_2, Ozone, SO_2, NH_3; and mg/m³ for CO). The x-axis represents the time progression over the years.

Pollutants vs. Relative humidity graph

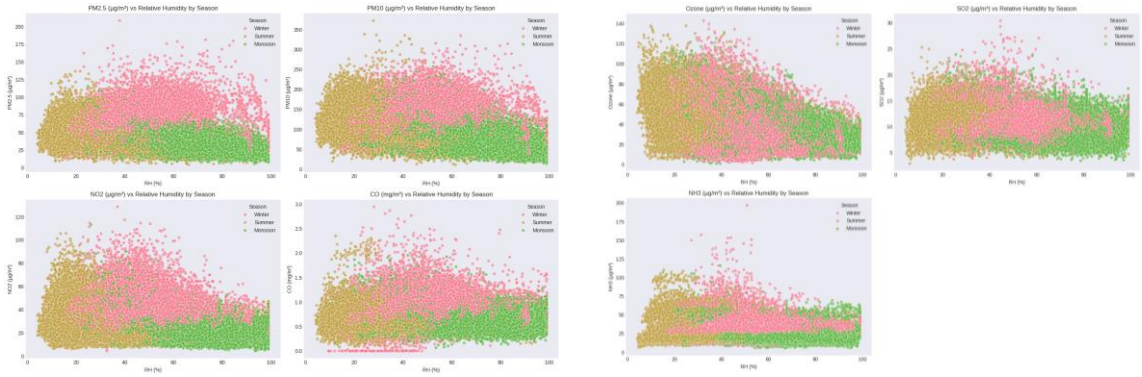

Figure 4.6: Pollutants vs relative humidity

This figure contains seven scatter plots, each examining the relationship between a specific air pollutant and relative humidity, RH (%), with the data points colored according to the season (Winter, Summer, Monsoon).

These scatter plots reveal how relative humidity is associated with different pollutant levels across the seasons in Jaipur:

- Particulate Matter ($PM_{2.5}$ and PM_{10}), NO_2, and CO: Tend to have higher concentrations at lower to mid-range relative humidity, particularly during the winter. Higher humidity, especially during the monsoon, is generally associated with lower levels of these pollutants, likely due to wet deposition and reduced atmospheric stability.
- Ozone: Shows an inverse relationship with humidity, with the highest concentrations occurring during the summer at lower to mid-range humidity, conditions that favor its photochemical formation. Higher humidity during the monsoon is associated with lower ozone levels.
- SO_2 and NH_3: Also tend to show lower concentrations at higher relative humidity, especially during the monsoon, suggesting removal processes related to moisture in the air.

Pollutants vs. Wind Direction & Wind Speed graph

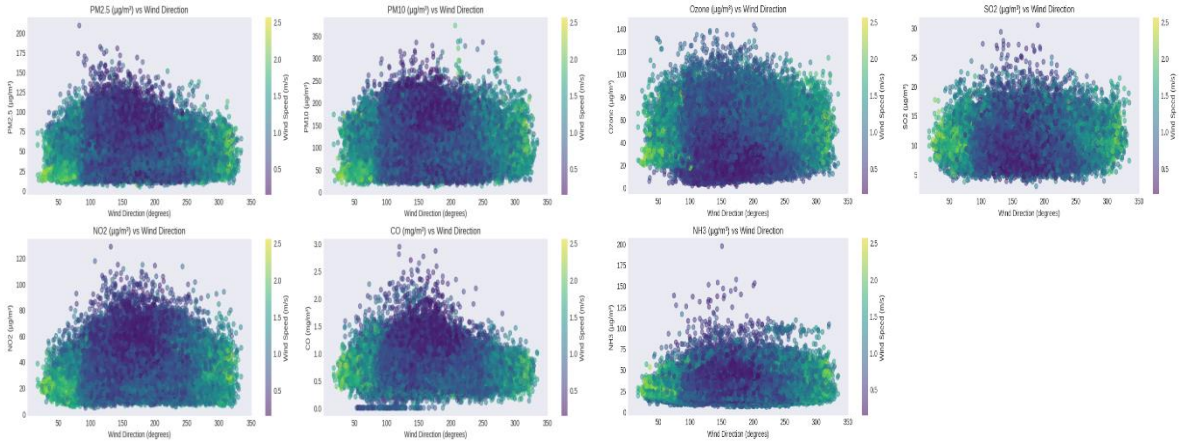

Figure 4.7: Pollutants vs wind speed & wind direction

This figure presents seven density scatter plots, each examining the relationship between a specific air pollutant and wind direction (in degrees, where 0/360 is North, 90 is East, 180 is South, and 270 is West). The color intensity of the points indicates the frequency of occurrence of those specific combinations of pollutant concentration and wind direction, with the color scale on the right representing the wind speed (in m/s) associated with those data points.

The scatter plots reveal the following patterns in the relationship between air pollutants, wind direction, and wind speed:

- $PM_{2.5}$ and PM_{10}: Both particulate matter pollutants show relatively high concentrations when wind directions are between 0°–90° and 270°–360°, particularly at lower wind speeds (dark-colored points). This indicates that stagnant or low-movement air contributes to particulate accumulation.
- NO_2 and CO: These gaseous pollutants also show elevated levels at lower wind speeds, especially when wind is coming from the eastern and southwestern sectors. This suggests possible emission sources (e.g., traffic or industrial zones) located in those directions.
- Ozone (O_3): Ozone concentrations appear more evenly distributed across wind directions, though higher concentrations are seen at moderate wind speeds, possibly due to secondary formation during sunlight exposure rather than direct emission sources.

- SO_2: Sulfur dioxide shows higher values when wind is coming from 40°–100°, again under lower wind speed conditions. This may indicate localized industrial emissions or combustion sources in that wind sector.
- NH_3: Ammonia concentrations are highest when wind direction is between 250°–350°, with most high concentrations associated with calm wind conditions. This may point to agricultural activity or waste treatment plants as potential sources.

Pollutants vs. Atmospheric Temperature graph

Pollutants vs. Relative humidity graph

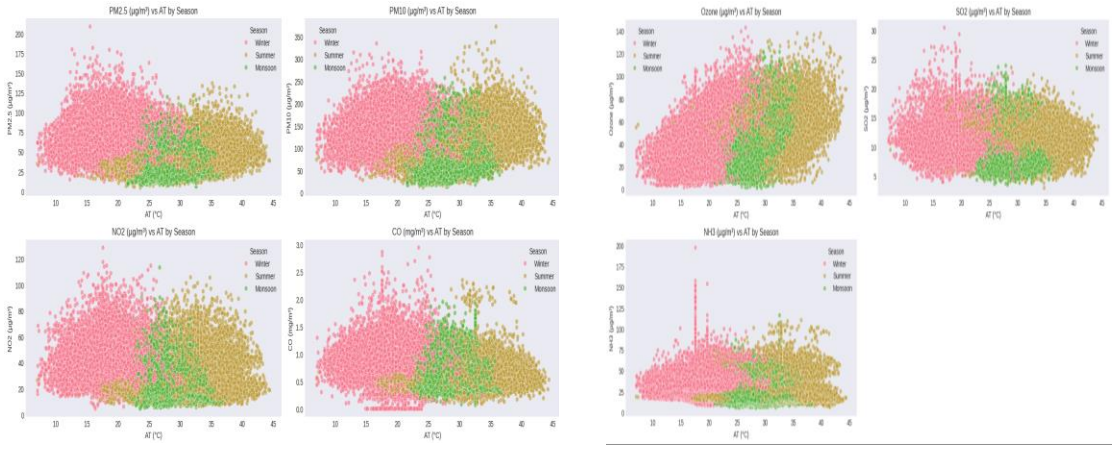

Figure 4.8: Pollutants vs Atmospheric Temperature

The plots show how concentrations of various air pollutants vary with air temperature (AT) across Winter, Summer, and Monsoon seasons:

- $PM_{2.5}$, PM_{10}, NO_2, CO, SO_2, and NH_3 show higher concentrations in winter, especially at lower temperatures (<20°C), and decrease with rising temperature.
- Ozone (O_3) increases with temperature, peaking in summer and monsoon due to enhanced photochemical activity.
- Pollution levels are lowest in summer and monsoon, likely due to better dispersion and rainfall.
- Winter season consistently shows the worst air quality across most pollutants.

4.3 Parameter Selection

4.3.1 Correlation Matrix of Parameters

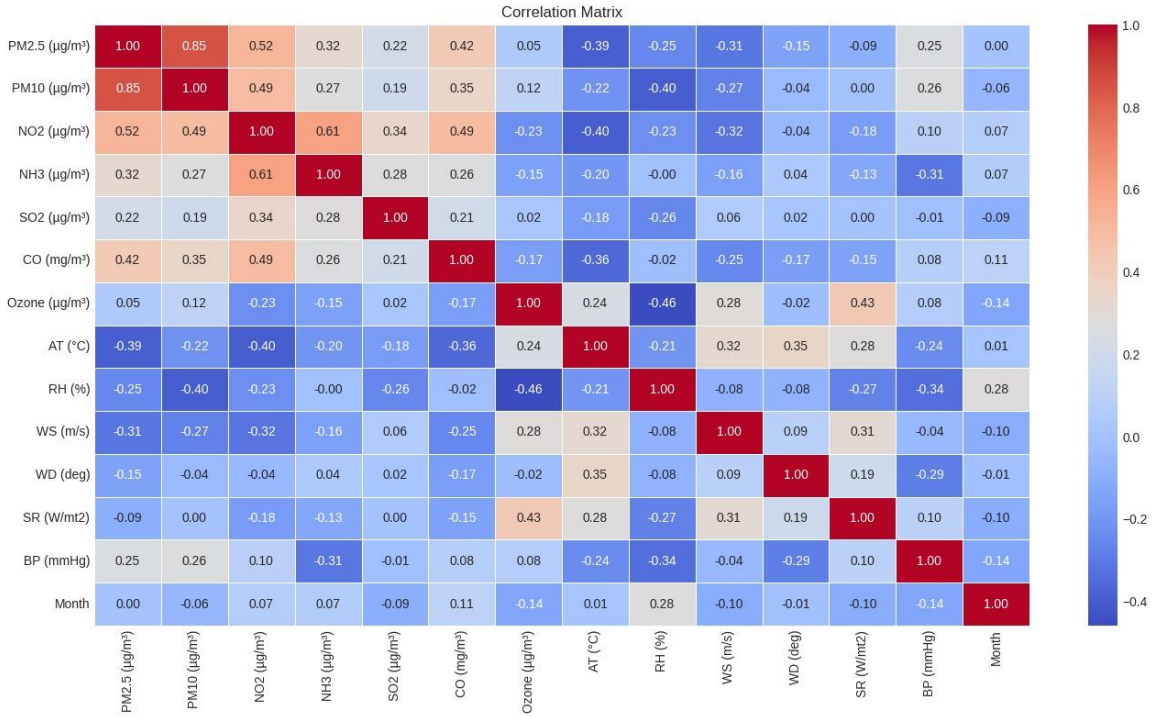

Figure 4.9: Pearson's correlation matrix

This is our correlation matrix, which visually represents the pairwise linear relationships between different variables. The variables included are various air pollutants ($PM_{2.5}$, PM_{10}, NO_2, NH_3, SO_2, CO, Ozone) and meteorological parameters (Atmospheric Temperature (AT), Relative Humidity (RH), Wind Speed (WS), Wind Direction (WD), Solar Radiation (SR), Barometric Pressure (BP)).

Key Findings and Relationships:

NO_2 - Pollutant Correlations:

- **$PM_{2.5}$ (0.52):** Shows a moderate positive correlation with $PM_{2.5}$. This suggests that days with higher levels of fine particulate matter tend to also have higher levels of nitrogen dioxide. This could be due to shared combustion sources like vehicular traffic or industrial activities that release both pollutants.

- **PM_{10} (0.49):** Similar to PM2.5, NO_2 exhibits a moderate positive correlation with PM_{10}. This further supports the idea of common emission sources contributing to both particulate matter and NO_2 pollution.
- **NH_3 (0.61):** Displays a moderate positive correlation with Ammonia. This might indicate some shared sources, such as agricultural activities or industrial processes that release both gases, or potential atmospheric chemical interactions.
- **SO_2 (0.34):** Shows a weak positive correlation with Sulfur Dioxide. This could point to some overlap in industrial or combustion sources that emit both these gaseous pollutants.
- **CO (0.49):** Has a moderate positive correlation with Carbon Monoxide. This is likely because both NO_2 and CO are primary pollutants released from incomplete combustion processes, particularly from vehicles.
- **Ozone (-0.23):** Exhibits a weak negative correlation with Ozone. This is a common observation in urban environments. NO is emitted along with NO_2 from combustion sources. NO can react with ozone, reducing its concentration ($NO + O_3 \rightarrow NO_2 + O_2$). Therefore, higher NO_2 (often associated with higher NO emissions) can lead to lower ozone levels near sources. However, further downwind, NO_2 can participate in photochemical reactions that contribute to ozone formation.

NO_2 - Meteorological Parameter Correlations:

Atmospheric Temperature (AT) (-0.40): Shows a weak negative correlation. This suggests that higher temperatures might be associated with slightly lower NO_2 levels, possibly due to increased atmospheric mixing and dispersion or changes in emission patterns.

Relative Humidity (RH) (-0.23): Displays a weak negative correlation. Higher humidity might lead to some removal of NO_2 through deposition processes or influence its atmospheric chemistry.

Wind Speed (WS) (-0.23): Exhibits a weak negative correlation. Higher wind speeds generally promote the dispersion of pollutants, leading to lower local concentrations of NO_2

Wind Direction (WD) (-0.04): Shows a very weak negative correlation, close to zero. This reinforces the idea from the wind rose plots that the relationship between wind direction andNO_2 concentration is likely more complex and source-location dependent than a simple linear correlation can capture.

Solar Radiation (SR) (-0.13): Displays a very weak negative correlation. While NO_2 is involved in photochemical reactions leading to ozone formation, the direct linear relationship with solar radiation at ground level might be weak.

Barometric Pressure (BP) (0.10): Shows a very weak positive correlation, suggesting a minimal linear relationship with NO_2 levels.

4.3.2 Featured Engineering

To improve the model's ability to capture temporal trends in NO_2 concentrations, we applied various cyclic encodings and time-based transformations. These were designed to reflect the periodic nature of environmental data and enhance both explanatory and forecasting accuracy.

We engineered multiple cyclical time features, such as month_cos, month_sin, hour_cos, hour_sin, day_sin, and day_cos, using sine and cosine transformations. These helps represent continuous time cycles like daily, weekly, and monthly patterns. Among these, the month_cos = cos ($2\pi \times$ month / 12) transformation proved most influential in the feature importance analysis.

While many cyclic features were considered, month_cos emerged as a dominant temporal predictor, exhibiting an F-score of approximately 8,000 in our analysis, indicating a significant relationship with NO_2 levels.

Model Specific Feature Use -

- For the MLR model: despite month_cos showing high importance, it was excluded to avoid multicollinearity with other features. Instead, we focused on pollutant and meteorological parameters with strong independent correlations to NO_2, ensuring reliable coefficient estimates.

➢ For the SARIMA model: month_cos was retained and emphasized as it directly supports the model's seasonal component. Including this feature helped the SARIMA model capture annual variation patterns and improve forecasting accuracy.

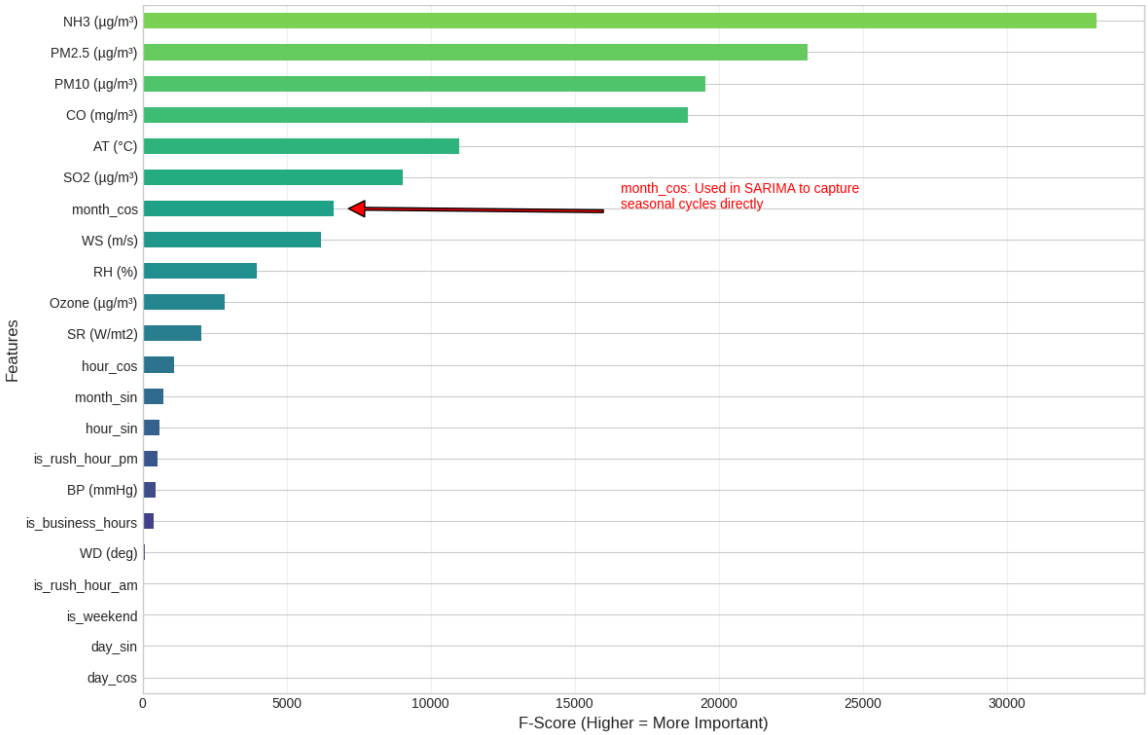

Figure 4.10: Feature importance ranking for NO_2 prediction using F-score

Parameters Selected for MLR Model:

PM2.5, SO_2, NH_3, CO, AT (Atmospheric temperature), RH (Relative Humidity), WS (Wind Speed), SR (Solar Radiation).

Parameters Selected for the SARIMA Model:

PM2.5, SO_2, NH_3, CO, AT (Atmospheric temperature), RH (Relative Humidity), WS (Wind Speed), month_cos.

4.3.3 Log Transformation of Targeted Value

To improve model performance and satisfy statistical assumptions, a log transformation was applied to the NO_2 concentration values. The original data exhibited a positively

skewed distribution with signs of heteroscedasticity, both of which can adversely affect models like Multiple Linear Regression (MLR) and SARIMA that assume normally distributed residuals and constant variance.

After the transformation, the distribution became approximately normal, the variance was stabilized, and the influence of extreme values was reduced. This resulted in a nearly symmetrical, bell-shaped distribution, effectively normalizing the data. These improvements enhanced model interpretability and robustness.

Since the model outputs are now on a logarithmic scale, final predictions were exponentiated to revert them back to the original unit (μg/m³) for practical interpretation.

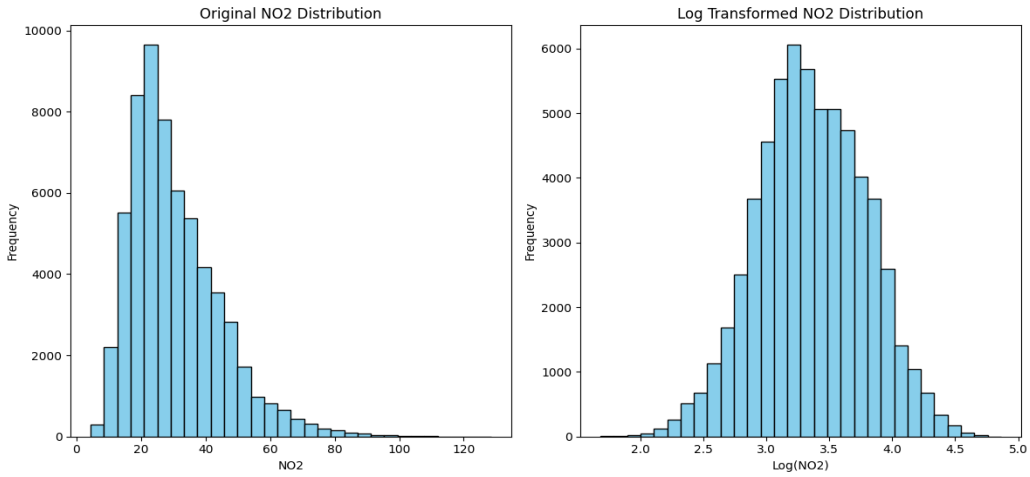

Figure 4.11: NO_2 distribution before and after log transformation

4.4 Model Development

4.4.1 Development of MLR

The model development began with Baseline Linear Regression, which demonstrated low predictive accuracy due to its failure to capture non-linear patterns (Test R^2 = 0.6005, RMSE = 0.2707). In order to manage multicollinearity and increase generalization, regularized regression techniques like Ridge, Lasso, and Elastic Net were applied to the model utilizing grid search cross-validation. Ridge Regression

was shown to be the most stable of these, and α = 10 was found to be the ideal regularization strength.

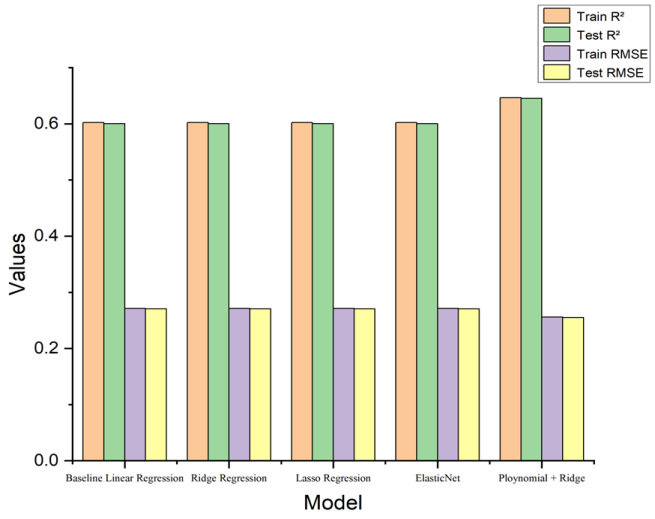

Figure 4.12: Performance comparison of the regression model for the predictor

Then, using the determined α value, Polynomial Features and Ridge Regression were coupled. This increased the model's capacity to capture non-linear correlations and yielded the best performance (Test R^2 = 0.6454, RMSE = 0.2551). This strategy was the most reliable because it successfully balanced regularization and model complexity.

Model	Train R^2	Test R^2	Train RMSE	Test RMSE
Baseline Linear Regression	0.6025	0.6005	0.2714	0.2707
Ridge Regression	0.6025	0.6005	0.2714	0.2707
Lasso Regression	0.6025	0.6005	0.2714	0.2707
Elastic Net	0.6025	0.6005	0.2714	0.2707
Polynomial + Ridge Regression	**0.6467**	**0.6454**	**0.2558**	**0.2551**

Table 4.1: Comparative Evaluation of Regression Models for NO_2 Prediction Using R^2 and RMSE

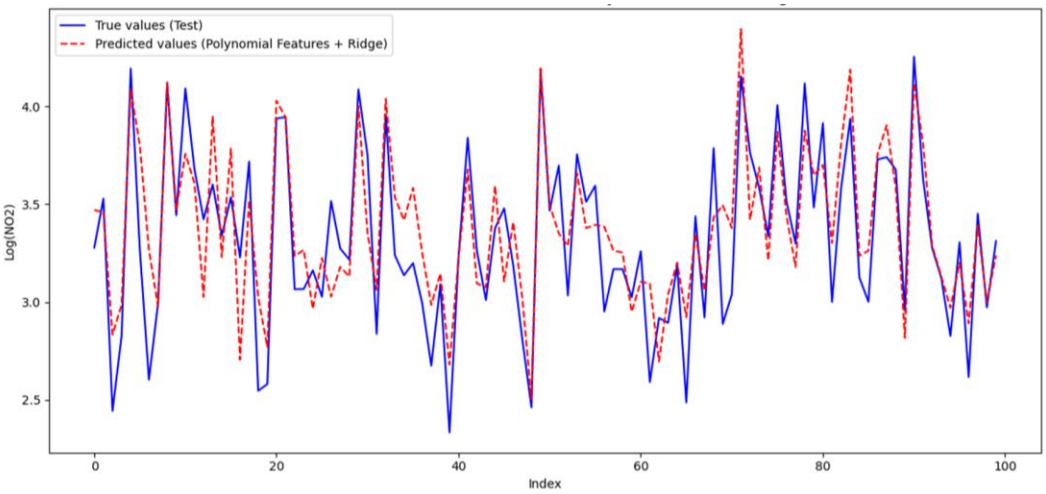

Figure 4.13: True vs Predicted values - Polynomial + Ridge

The graph compares actual (blue solid line) and predicted (red dashed line) log-transformed NO_2 concentrations using the Polynomial Features + Ridge Regression model. The close tracking between lines demonstrates the model's effectiveness at capturing non-linear patterns in the data, validating its strong performance metrics (R^2 = 0.6454, RMSE = 0.2551) and confirming the value of combining polynomial features with ridge regularization ($\alpha = 10$).

4.4.2 Development of SARIMA

To address variance instability in the original NO_2 time series, a natural logarithmic transformation (log_1p) was applied. Subsequent Augmented Dickey-Fuller (ADF) and KPSS tests indicated that, while the transformation stabilized variance, the series remained non-stationary. Consequently, first-order differencing (d=1) was employed to induce stationarity by removing the underlying trend. Analysis of the autocorrelation (ACF) and partial autocorrelation (PACF) functions of the differenced, log-transformed series revealed significant autocorrelation at a lag of 24 hours, indicative of a pronounced daily periodicity. To account for this recurring pattern, seasonal differencing with a lag of 24 (s=24) was therefore applied, effectively capturing the diurnal cycles characteristic of NO_2 concentration fluctuations.

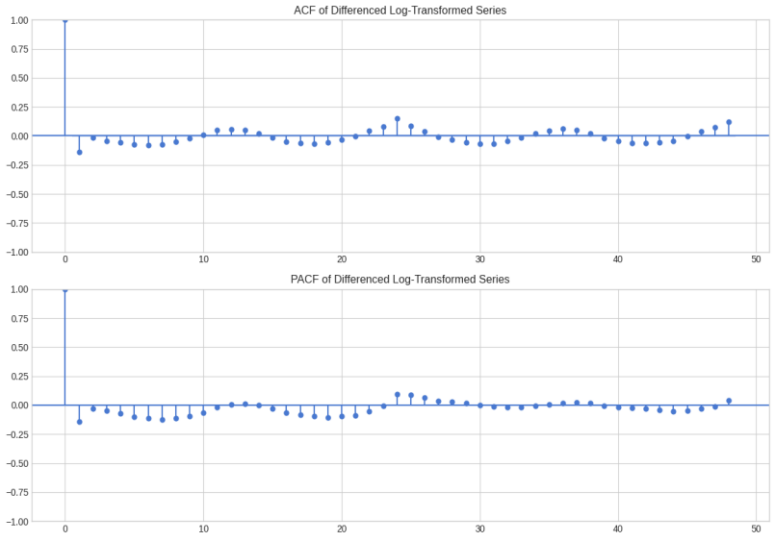

Figure 4.14: ACF and PACF Plots of Differenced Log-transformed NO_2 series

The initial SARIMA model with configuration (1, 1, 1) × (0, 0, 0, 24) served as a baseline for evaluating subsequent improvements. This baseline model demonstrated limited predictive capability with a test RMSE of 8.30 and a low coefficient of determination (R^2) of 0.35. The modest performance suggested inadequate capture of the underlying temporal patterns, particularly the seasonal component of the NO_2 concentration data.

A systematic enhancement approach was then implemented by progressively adjusting both seasonal and non-seasonal parameters. The first significant improvement came from incorporating a seasonal autoregressive component, resulting in the SARIMA (1, 1, 1) × (1, 0, 0, 24) model. This modification substantially enhanced predictive performance, reducing the RMSE to 8.22 and increasing the R^2 to 0.52, while simultaneously lowering the AIC to 0.9968. The marked improvement confirmed that incorporating seasonal memory helped the model better represent the 24-hour cyclical pattern inherent in urban NO_2 concentration dynamics.

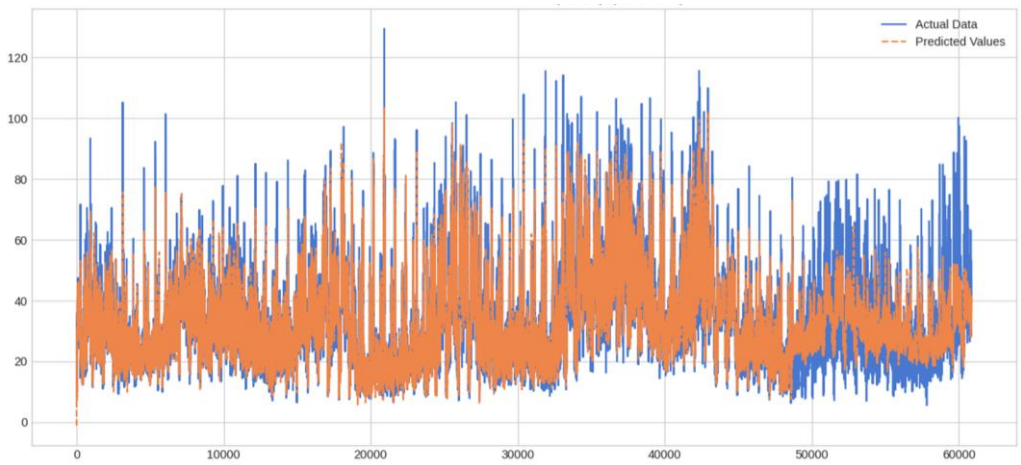

Figure 4.15: Actual vs Predicted NO_2 Concentration Using SARIMA (1, 1, 1) x (1, 0, 0, 24) Model

An alternative configuration was tested by substituting the seasonal autoregressive term with a seasonal moving average component, yielding SARIMA (1, 1, 1) × (0, 0, 1, 24). This variant also performed reasonably well with a test R^2 of 0.52 and RMSE of 8.24, though slightly less effectively than the model with the seasonal autoregressive parameter. The comparable performance of these two seasonal variants underscores the importance of capturing cyclical patterns in the data, regardless of the specific mechanism employed.

Attempts to increase model sophistication through higher-order non-seasonal components yielded diminishing returns. The SARIMA (2, 1, 1) × (0, 0, 0, 24) configuration showed marginal improvement in training fit but exhibited poorer generalization to unseen data, with an elevated RMSE of 8.47. Similarly, increasing the non-seasonal moving average order to produce SARIMA (1, 1, 2) × (0, 0, 0, 24) offered no substantive benefits, as evidenced by the persistently high RMSE of 8.46 and deteriorated R^2 values.

Following a comprehensive evaluation across multiple parameter configurations, the SARIMA (1, 1, 1) × (1, 0, 0, 24) emerged as the optimal model. This specification effectively balances computational complexity with forecasting accuracy, while appropriately capturing both the trend component and the diurnal cycle in NO_2 concentrations. The model incorporates first-order non-seasonal autoregressive and

moving average components, first-order differencing for trend removal, and a first-order seasonal autoregressive component with 24-hour periodicity.

Model Specification	AIC/BIC	Train R^2	Test R^2	Test RMSE	Note
SARIMA(1,1,1)x(0,0,0,24)	1.0023	0.8692	0.3510	8.3035	Baseline Model
SARIMA(1,1,1)x(1,0,0,24)	0.9968	0.8374	0.5233	8.2271	Added seasonal AR component
SARIMA(1,1,1)x(0,0,1,24)	0.9966	0.8372	0.5217	8.2409	Added seasonal MA component
SARIMA(2,1,1)x(0,0,0,24)	0.9966	0.8384	0.4946	8.4713	Increased non-seasonal AR order
SARIMA(1,1,2)x(0,0,0,24)	0.9966	0.8384	0.4948	8.4691	Increased non-seasonal MA order

Table 4.2: Comparative Performance of SARIMA Model Variants for NO_2 Prediction

4.4.3 Forecasting Performance of the SARIMA Model

Following the model development phase, the finalized SARIMA (1,1,1) × (1,0,0,24) model was applied to generate forecasts of NO_2 concentrations over two critical time horizons: 24 hours (short-term) and 7 days (medium-term). These forecasts were based on the cleaned and log-transformed dataset, effectively capturing both short-term fluctuations and the strong daily cyclicity characteristic of urban air pollution patterns.

24-hour forecast for best model (2025-01-01)

The 24-hour forecast demonstrated excellent predictive accuracy with minimal error margins. As shown in Table 1, the model successfully captured the distinct diurnal cycle of NO_2 concentrations at key points throughout the day:

Time	NO_2 Forecast ($\mu g/m^3$)	Lower CI	Upper CI	Actual Values	Absolute Percentage Error
00:00	39.72	29.34	53.78	38	4.52
06:00	35.66	22.75	55.91	36	0.94

12:00	37.07	23.10	59.50	39	4.95
18:00	42.43	26.11	68.96	46	7.76
22:00	46.24	28.25	75.69	51	9.33

Table 4.3: Forecasted NO_2 Concentrations with Confidence Intervals and Error Analysis

The model's forecasting accuracy is particularly notable, with absolute percentage errors ranging from as low as 0.94% during early morning hours to a maximum of 9.33% during the evening peak pollution period. This level of precision confirms the model's ability to anticipate NO_2 concentration patterns throughout different times of the day, including both low concentration periods and evening peaks.

The complete hourly forecast revealed important temporal patterns:

- Mean forecast: 38.44 µg/m³
- Minimum forecast: 33.71 µg/m³ (occurring at 03:00)
- Maximum forecast: 46.24 µg/m³ (occurring at 22:00)

The confidence intervals widened slightly during peak hours, reflecting the greater inherent variability in NO_2 concentrations during periods of high human activity. Even with this increased uncertainty during peak periods, the model maintained strong predictive performance.

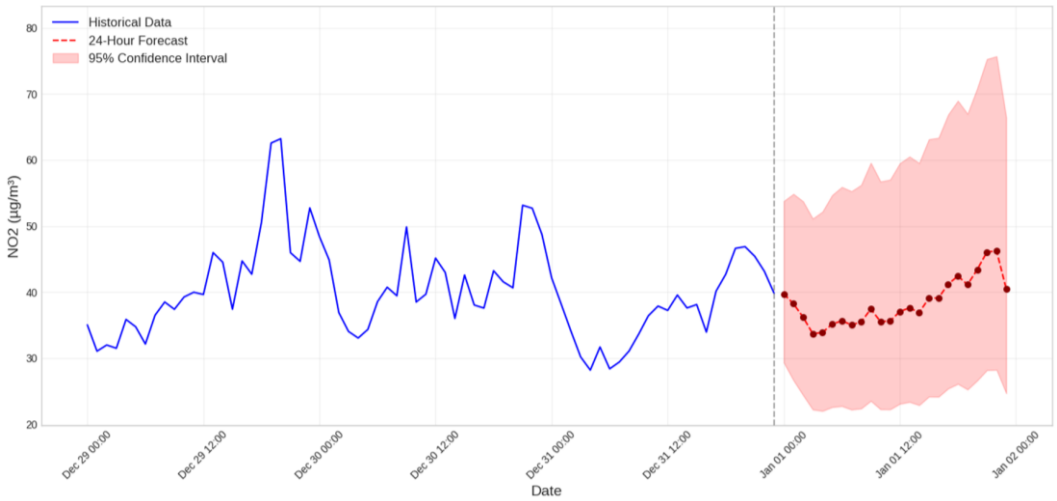

Figure 4.16: NO_2 concentration Forecast - Next 24 hours

One Week (7-days) Forecast

While the 7-day forecast maintained a good representation of the general trend in NO_2 concentrations, a widening 95% confidence interval was observed as the prediction horizon extended. This progressive increase in uncertainty is an expected characteristic of time series forecasting, reflecting the compounded effects of inherent noise and external unmodeled variables over time.

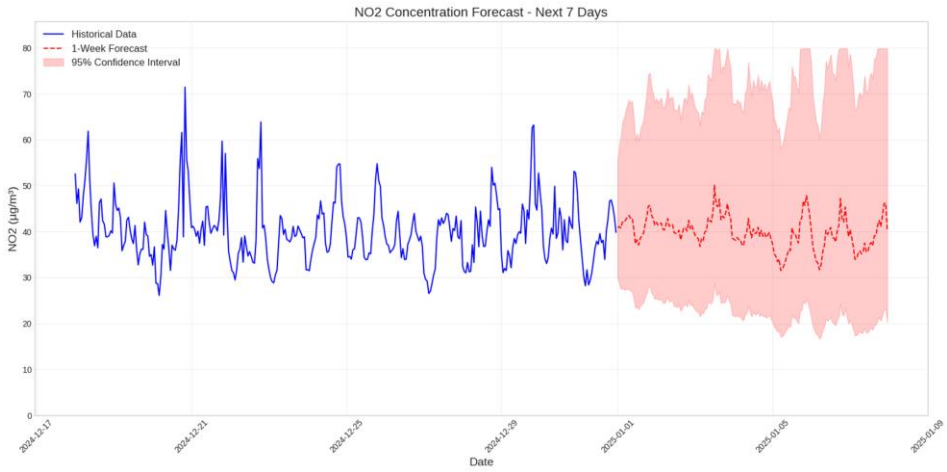

Figure 4.17: NO_2 concentration forecast - next 7 days

5. Conclusion

This study used 61,368 data points to create and assess statistical and time-series models for forecasting NO_2 concentrations in Jaipur. NO_2 levels were found to be negatively connected with temperature, humidity, precipitation, and wind speed, but favorably correlated with atmospheric pressure, showing the significant influence of climatic elements on air pollution levels.

The Multiple Linear Regression (MLR) model, augmented with polynomial features, efficiently caught non-linear correlations between NO_2 and its predictors, which were not properly recognized by Pearson correlation analysis alone. The Ridge Regression model using polynomial features outperformed all other regression models examined, with a test R^2 of 0.6454 and a Root Mean Square Error (RMSE) of 0.2551, indicating its robustness and predictive accuracy.

SARIMA models were used to analyze NO_2 concentrations on a daily and seasonal basis. The SARIMA $(1,1,1) \times (1,0,0,24)$ configuration scored best, with a test R^2 of 0.5233 and an RMSE of 8.2271. This model successfully caught daily cyclic trends and displayed great short-term forecasting ability, with a mean absolute error of 5.5% for 24-hour predictions. Even medium-term forecasts (up to 7 days) remained accurate despite wider confidence ranges.

The findings emphasize the need to include both meteorological and temporal characteristics when modeling air pollution. The hybrid modeling technique (MLR for multivariate correlation and SARIMA for time-series forecasting) effectively captured the complicated behavior of NO_2 levels. This methodology can serve as the cornerstone for environmental monitoring and pollution control methods in cities like Jaipur.

6. References

Abbas, A., Chanchal, A. K., Garnaik, S., Farrow, A., Glushkov, I., Wu, C., & Albertson-Kwok, J. (n.d.). *NO_2 POLLUTION AND HEALTH RISKS IN SEVEN MAJOR INDIAN CITIES This report was made possible through the collaborative efforts of*.

al Yammahi, A., & Aung, Z. (2023). Forecasting the concentration of NO_2 using statistical and machine learning methods: A case study in the UAE. *Heliyon*, *9*(2). https://doi.org/10.1016/j.heliyon.2022.e12584

Anusasananan, P., Suwanarat, S., & Thongprasert, N. (2023). The effect of meteorological parameters on air particulate matter. *Journal of Physics: Conference Series*, *2653*(1). https://doi.org/10.1088/1742-6596/2653/1/012068

Article Parisa Kahrari, O., Khaledi, S., Keikhosravi, G., & Jalil Alavi, S. (2024). Analyzing the Relationship Between Meteorological Elements and Criteria Atmospheric Pollutants in Tabriz Using Statistical Modeling. *Environmental Sciences (Environ. Sci.)*, *22*(1), 69–90. https://doi.org/10.48308/envs.2023.1348

Chen, P., Niu, A., Liu, D., Jiang, W., & Ma, B. (2018). Time Series Forecasting of Temperatures using SARIMA: An Example from Nanjing. *IOP Conference Series: Materials Science and Engineering*, *394*(5). https://doi.org/10.1088/1757-899X/394/5/052024

Chi, Y., Fan, M., Zhao, C., Yang, Y., Fan, H., Yang, X., Yang, J., & Tao, J. (2022). Machine learning-based estimation of ground-level NO_2 concentrations over China. *Science of the Total Environment*, *807*. https://doi.org/10.1016/j.scitotenv.2021.150721

Choudhary, S. S., & Saini, R. (2023). *Trend Analysis Using Meteorological Data and Non-parametric Statistical Tests: A Case Study of Jodhpur, Rajasthan, India*. https://doi.org/10.21203/rs.3.rs-2735325/v1

Dabral, P. P., & Murry, M. Z. (2017). Modeling and Forecasting of Rainfall Time Series Using SARIMA. *Environmental Processes*, *4*(2), 399–419. https://doi.org/10.1007/s40710-017-0226-y

Darmawan, G., Kurniasih, N., & Ahmar, A. S. (2018). Accuracy of Periodogram Analysis for Identification of Multiplicative SARIMA Models. *Journal of Physics: Conference Series*, *1028*(1). https://doi.org/10.1088/1742-6596/1028/1/012243

Devasekhar, M. v, & Natarajan, P. (n.d.). Prediction of Air Quality and Pollution using Statistical Methods and Machine Learning Techniques. In *IJACSA) International Journal of Advanced Computer Science and Applications* (Vol. 14, Issue 4). www.ijacsa.thesai.org

Gocheva-Ilieva, S. G., Ivanov, A. V., Voynikova, D. S., & Boyadzhiev, D. T. (2014). Time series analysis and forecasting for air pollution in a small urban area: An SARIMA and factor analysis approach. *Stochastic Environmental Research and Risk Assessment*, *28*(4), 1045–1060. https://doi.org/10.1007/s00477-013-0800-4

González-Enrique, J., Turias, I. J., Ruiz-Aguilar, J. J., Moscoso-López, J. A., & Franco, L. (2019). Spatial and meteorological relevance in NO 2 estimations: a case study in the Bay of Algeciras (Spain). *Stochastic Environmental Research and Risk Assessment*, *33*(3), 801–815. https://doi.org/10.1007/s00477-018-01644-0

Goyal, S., & Sharma, R. (2023). Prediction of the concentrations of PM2.5 and NOx using machine learning-based models. *Materials Today: Proceedings*. https://doi.org/10.1016/j.matpr.2023.07.121

Gupta, A., Srivastava, A. K., Singh Bisht, D., & Jhamaria, C. (n.d.). *Long-term assessment of near-surface air pollutants at Jaipur: Source identifications and their association with surface meteorology Long-term assessment of near-surface air pollutants at Jaipur: Source identifications and 2 their association with surface meteorology*. https://ssrn.com/abstract=4897841

Hesterberg, T. W., Bunn, W. B., McClellan, R. O., Hamade, A. K., Long, C. M., & Valberg, P. A. (2009). Critical review of the human data on short-term nitrogen dioxide (NO_2) exposures: Evidence for NO_2 no-effect levels. *Critical Reviews in Toxicology*, *39*(9), 743–781. https://doi.org/10.3109/10408440903294945

Hoang, T. T., Nguyen, H., Le, M. H., Ho, M. L., Tran, A. H., Nguyen, D. M. T., & Tran, N. A. (2024). The Application of Time Series Models Considering Seasonality

in Monthly Electricity Production Forecasting. *Proceedings of 2024 7th International Conference on Green Technology and Sustainable Development, GTSD 2024*, 200–206. https://doi.org/10.1109/GTSD62346.2024.10675225

Kayes, I., Shahriar, S. A., Hasan, K., Akhter, M., Kabir, M. M., & Salam, M. A. (2019). The relationships between meteorological parameters and air pollutants in an urban environment. *Global Journal of Environmental Science and Management*, *5*(3), 265–278. https://doi.org/10.22034/gjesm.2019.03.01

Ke, H., Gong, S., He, J., Zhang, L., Cui, B., Wang, Y., Mo, J., Zhou, Y., & Zhang, H. (2022). Development and application of an automated air quality forecasting system based on machine learning. *Science of the Total Environment*, *806*. https://doi.org/10.1016/j.scitotenv.2021.151204

Khedekar, S., & Thakare, S. (2023). Correlation analysis of atmospheric pollutants and meteorological factors using statistical tools in Pune, Maharashtra. *E3S Web of Conferences*, *391*. https://doi.org/10.1051/e3sconf/202339101190

Kumar, P. (2025). Exploring the influence of human activities and the COVID-19 lockdown on urban air quality in Rajasthan, India. *Theoretical and Applied Climatology*, *156*(4). https://doi.org/10.1007/s00704-025-05445-8

Kumari, S., & Muthulakshmi, P. (2024). SARIMA Model: An Efficient Machine Learning Technique for Weather Forecasting. *Procedia Computer Science*, *235*, 656–670. https://doi.org/10.1016/j.procs.2024.04.064

Lin, S., Lin, W., Hu, X., Wu, W., Mo, R., & Zhong, H. (n.d.-a). *CycleNet: Enhancing Time Series Forecasting through Modeling Periodic Patterns*. https://github.com/ACAT-SCUT/CycleNet.

Malhotra, M., Aulakh, I. K., Gupta, M., & Monika. (2023). Meteorological parameters affect pollutants in the air. *AIP Conference Proceedings*, *2768*. https://doi.org/10.1063/5.0148966

Maltare, N. N., & Vahora, S. (2023). Air Quality Index prediction using machine learning for Ahmedabad city. *Digital Chemical Engineering*, *7*. https://doi.org/10.1016/j.dche.2023.100093

Mohith, J., Kulshrestha, D., & Jothi, K. R. (2021). A Comprehensive Analysis of Machine Learning Methods for Air Pollution Forecasting. *2021 2nd International Conference on Innovative and Creative Information Technology, ICITech 2021*, 14–19. https://doi.org/10.1109/ICITech50181.2021.9590113

Oji, S., & Adamu, H. (2020a). ARTICLES INFORMATION Correlation between air pollutants concentration and meteorological factors on seasonal air quality variation. In *Journal of Air Pollution and Health* (Vol. 5, Issue 1). http://japh.tums.ac.ir

Oji, S., & Adamu, H. (2020b). Correlation between air pollutants concentration and meteorological factors on seasonal air quality variation. *Journal of Air Pollution and Health*, *5*(1), 11–32. https://doi.org/10.18502/japh.v5i1.2856

Prasad Sharma Associate Professor, B. (n.d.). Air Pollution in Jaipur Sources, Impacts, and Mitigation Strategies for a Healthier Future. *International Journal of Business and Management Invention (IJBMI)* ||, *9*, 2020–2072. https://doi.org/10.35629/8028-0901017278

Ruhela, M., Maheshwari, V., Ahamad, F., & Kamboj, V. (2022). Air quality assessment of Jaipur city Rajasthan after the COVID-19 lockdown. *Spatial Information Research*, *30*(5), 597–605. https://doi.org/10.1007/s41324-022-00456-3

Segovia, J. A., Toaquiza, J. F., Llanos, J. R., & Rivas, D. R. (2023). Meteorological Variables Forecasting System Using Machine Learning and Open-Source Software. *Electronics (Switzerland)*, *12*(4). https://doi.org/10.3390/electronics12041007

Shen, J., Valagolam, D., & McCalla, S. (2020). Prophet forecasting model: A machine learning approach to predict the concentration of air pollutants ($PM_{2.5}$, PM_{10}, O_3, NO_2, SO_2, CO) in Seoul, South Korea. *PeerJ*, *8*. https://doi.org/10.7717/peerj.9961

Theng, D., & Bhoyar, K. K. (2024). Feature selection techniques for machine learning: a survey of more than two decades of research. In *Knowledge and Information Systems* (Vol. 66, Issue 3, pp. 1575–1637). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/s10115-023-02010-5

Udristoiu, M. T., el Mghouchi, Y., & Yildizhan, H. (2023). Prediction, modelling, and forecasting of PM and AQI using hybrid machine learning. *Journal of Cleaner Production*, *421*. https://doi.org/10.1016/j.jclepro.2023.138496

Wikner, A., Pathak, J., Hunt, B. R., Szunyogh, I., Girvan, M., & Ott, E. (2021). Using data assimilation to train a hybrid forecast system that combines machine-learning and knowledge-based components. *Chaos*, *31*(5). https://doi.org/10.1063/5.0048050

Wong, P. Y., Su, H. J., Lee, H. Y., Chen, Y. C., Hsiao, Y. P., Huang, J. W., Teo, T. A., Wu, C. da, & Spengler, J. D. (2021). Using land-use machine learning models to estimate daily NO_2 concentration variations in Taiwan. *Journal of Cleaner Production*, *317*. https://doi.org/10.1016/j.jclepro.2021.128411