Analysis of Air-Quality and Forecasting

ASSIGNMENT - II

Submitted in partial fulfillment of the requirements of
***MATH F342,** Applied Statistical Methods*
By

Dinesh Kumar 2022B4TS516P

Under the supervision of:
Dr. Sumanta Pasari

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE PILANI
November 2024

CONTENTS

Chapter 1

Introduction to Air Quality Index (AQI)

1.1 Overview of Air Quality

Air quality refers to the condition of the air within our surrounding environment. It is determined by measuring various pollutants such as particulate matter (PM2.5 and PM10), ozone (O3), nitrogen dioxide (NO2), sulfur dioxide (SO2), carbon monoxide (CO), and lead. Poor air quality poses significant risks to human health, ecosystems, and the overall climate.

1.2 Understanding the Air Quality Index (AQI)

AQI is a numerical scale used to communicate the quality of air and its impact on human health.
AQI is derived from the concentration of specific pollutants. For each pollutant, a sub-index is calculated and then aggregated.
PM2.5, PM10, CO, NO2, O3, and SO2 are primary contributors.
In India, AQI is defined by the Central Pollution Control Board (CPCB). International standards are set by organizations like the EPA and WHO.

1.3 Air Quality in Indian Cities (with Case Studies)

India, home to some of the world's most densely populated urban areas, is grappling with severe air pollution challenges. Rapid urbanization, unregulated industrial growth, vehicular emissions, and limited public awareness have exacerbated the situation. According to the State of Global Air Report (2022), over 1.67 million premature deaths in India were attributable to air pollution in 2021 alone, making it the second-largest risk factor for early mortality in the country.

Factors Contributing to Air Pollution

- **Industrial Emissions**: Manufacturing hubs emit large quantities of sulfur dioxide (SO_2) and nitrogen oxides (NO_x), which contribute to smog and acid rain.
- **Vehicular Pollution**: India's exponential growth in vehicle ownership has led to increasing emissions of carbon monoxide (CO), hydrocarbons, and particulate matter.
- **Construction Dust**: Urban construction activities release significant quantities of coarse particulate matter (PM10).
- **Burning of Biomass**: Traditional cooking methods and seasonal crop stubble burning significantly degrade air quality.

Case Studies: Jaipur and Kolkata

1. **Jaipur**: Known for its semi-arid climate, Jaipur faces unique pollution challenges such as high levels of suspended dust particles due to construction and vehicular movement. Despite being less industrialized compared to Kolkata, the city experiences frequent exceedances of permissible PM2.5 levels.
    - **Notable Sources**: Dust storms, vehicular emissions, and construction activities.
    - **Seasonality**: Winter months witness a spike in pollution due to temperature inversion.

2. **Kolkata**: As one of India's oldest metropolitan hubs, Kolkata's pollution stems from a dense population, outdated transportation systems, and industrial emissions. The city consistently records AQI levels in the "unhealthy" category.
    - **Notable Sources**: Diesel-based public transport, industrial smokestacks, and coal-based thermal power plants.
    - **Seasonality**: Monsoon improves air quality temporarily, but winter again leads to hazardous pollution levels.

1.4 Impact of Air Quality on Human Health & Environment

Health Impacts

Air pollution is a "silent killer," with both short-term and long-term health implications for millions of individuals globally. Pollutants such as PM2.5, PM10, SO_2, and NO_2 have direct links to the following:

- **Respiratory Diseases**: High levels of particulate matter aggravate asthma, bronchitis, and chronic obstructive pulmonary disease (COPD).
    - Example: A 2021 study in the *Lancet Respiratory Medicine* revealed that children living in polluted cities like Delhi are 1.7 times more likely to develop chronic respiratory issues.
- **Cardiovascular Issues**: Persistent exposure to air pollutants increases the risk of heart attacks, strokes, and arrhythmias.
    - WHO (2021) states that prolonged exposure to PM2.5 accounts for 24% of global cardiovascular deaths.
- **Neurodevelopmental and Cognitive Effects**: Emerging research links air pollution to developmental delays in children and higher risks of neurodegenerative diseases such as Alzheimer's in adults.
- **Premature Deaths**: A WHO report from 2022 highlights that air pollution contributes to 7 million premature deaths globally every year.

Environmental Impacts

- **Climate Change**: Pollutants like black carbon and methane contribute to global warming by trapping heat in the atmosphere.
- **Damage to Ecosystems**: Acid rain, caused by NO_x and SO_2, alters soil chemistry, damages forests, and disrupts aquatic ecosystems.
- **Reduced Agricultural Productivity**: Ozone pollution can impair crop yields by damaging leaf tissues, reducing food security in already vulnerable regions.
- **Visibility Reduction**: High levels of PM10 and PM2.5 can create dense smog, disrupting transportation and daily life.

Economic Costs

A study by TERI (The Energy and Resources Institute) estimated that air pollution costs India 8.5% of its GDP annually due to health expenses and lost productivity. This reinforces the urgent need for air quality monitoring and forecasting.

1.5 Significance of Air Quality Forecasting

Air quality forecasting is a critical tool for mitigating the harmful effects of air pollution. By predicting pollution levels in advance, authorities can take proactive measures to safeguard public health and the environment. Key reasons for forecasting include:

1. **Protecting Public Health**:
    - **Health Advisories**: Real-time and forecasted AQI levels allow governments to issue health advisories for sensitive groups, such as children, the elderly, and individuals with pre-existing conditions.
    - **School and Office Closures**: Forecasting helps authorities determine whether to restrict outdoor activities or declare temporary closures.
2. **Emergency Preparedness**:

- Advance forecasting allows urban areas to prepare for pollution spikes caused by events such as crop burning or industrial accidents.
- Example: Beijing's emergency measures during the 2008 Olympics successfully reduced pollution levels through predictive analysis.

3. **Policy Development**:
   - Long-term data trends from forecasting models assist in formulating air quality policies, such as vehicle emission standards, industrial zoning regulations, and cleaner energy initiatives.
4. **Sustainable Urban Planning**:
   - Forecasting aids in designing eco-friendly cities with better green spaces, optimized traffic systems, and pollution control mechanisms.
5. **International Collaboration**:
   - Forecasting systems are integral to transboundary air pollution agreements, ensuring nations can collectively combat shared challenges.

Applications of Air Quality Forecasting

- **Healthcare**: Hospitals can prepare for spikes in respiratory and cardiac emergencies.
- **Agriculture**: Farmers can adapt sowing and harvesting schedules based on forecasted ozone and particulate levels.
- **Tourism**: Forecasting ensures tourist destinations maintain optimal conditions, particularly in pollution-sensitive areas like the Himalayas.

Key Challenges in Air Quality Forecasting

- **Data Gaps**: Inadequate monitoring stations in rural and peri-urban areas reduce model accuracy.
- **Technological Limitations**: High computational requirements for models such as SARIMA or machine learning algorithms limit their widespread use.
- **Dynamic Nature of Pollution**: Meteorological changes, such as wind patterns and temperature inversions, add complexity to forecasting models.

Chapter 2

Data Pre-Processing

2.1 Shapiro-Wilk Test

The Shapiro-Wilk test is a statistical test used to assess whether a dataset follows a normal distribution. It is widely used in various fields like economics, biology, and data science to verify the assumption of normality, which is crucial for many statistical analyses.

The test checks the null hypothesis that the data comes from a normally distributed population.

- **Null Hypothesis (Ho)**: The data follows a normal distribution.
- **Alternative Hypothesis (Ha)**: The data does not follow a normal distribution.

Test Statistic:

The Shapiro-Wilk test statistic (W) measures how closely the data matches a normal distribution. It is calculated as:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Where:

- n: *Number of data points*
- $x_{(i)}$: *Ordered sample values (from smallest to largest).*
- $\bar{x}$: $Sample\ mean$
- a_i: $Precomputed\ constants\ derived\ from\ the\ expected\ values\ of$ normal order statistics and the $covariancematrixofthesestatistics$.

Interpretation:

- If the $W - value$ is close to 1, the data is likely normal.
- A small $W - value$ suggests deviations from normality.

p-value:

- The test generates a $p - value$ to determine statistical significance.
- If $p \leq \alpha$ (commonly α=0.05), reject Ho. This means the data is ***NOT*** normally distributed.
- If p > α, fail to reject Ho, indicating the data may be normally distributed.

Conclusion – Hourly AQI values are NOT normally distributed.

Advantages:

1. **Powerful for small samples:** The Shapiro-Wilk test is especially effective for small to moderate-sized datasets (typically $n \leq 2000$).
2. **Robust against non-normalities:** It reliably detects departures from normality like skewness and kurtosis.

Limitations:

1. **Sample size constraints:** The test may be less reliable for very large datasets (*n > 5000*), where even minor deviations from normality can produce significant results.
2. **Sensitive to outliers:** Extreme values can affect the test outcome.
3. **Interpretation dependency:** A significant result doesn't quantify *how* the data deviates from normality, only that it does.

2.2 Double Exponential Smoothing for Handling Missing Values in AQI Forecasting

Double Exponential Smoothing (DES) is a time series forecasting method that accounts for both level (current trend) and trend components in the data. It is well-suited for datasets with linear trends, making it ideal for time series data like Air Quality Index (AQI) values, which often exhibit seasonal patterns and temporal trends.

DES smooths the data using two equations:

$$\textbf{Level} \quad L_t = \alpha \cdot y_t + (1-\alpha) \cdot (L_{t-1} + T_{t-1})$$

$$\textbf{Trend} \quad T_t = \beta \cdot (L_t - L_{t-1}) + (1-\beta) \cdot T_{t-1}$$

Where:

- y_t: Observed value at time t,
- L_t: Smoothed value for the level at tt.
- T_t: Smoothed value for the trend at tt.
- α,β: Smoothing parameters (0 < α,β\alpha, \beta ≤ 1).

For missing values, the method interpolates based on the level and trend, ensuring continuity in the time series.

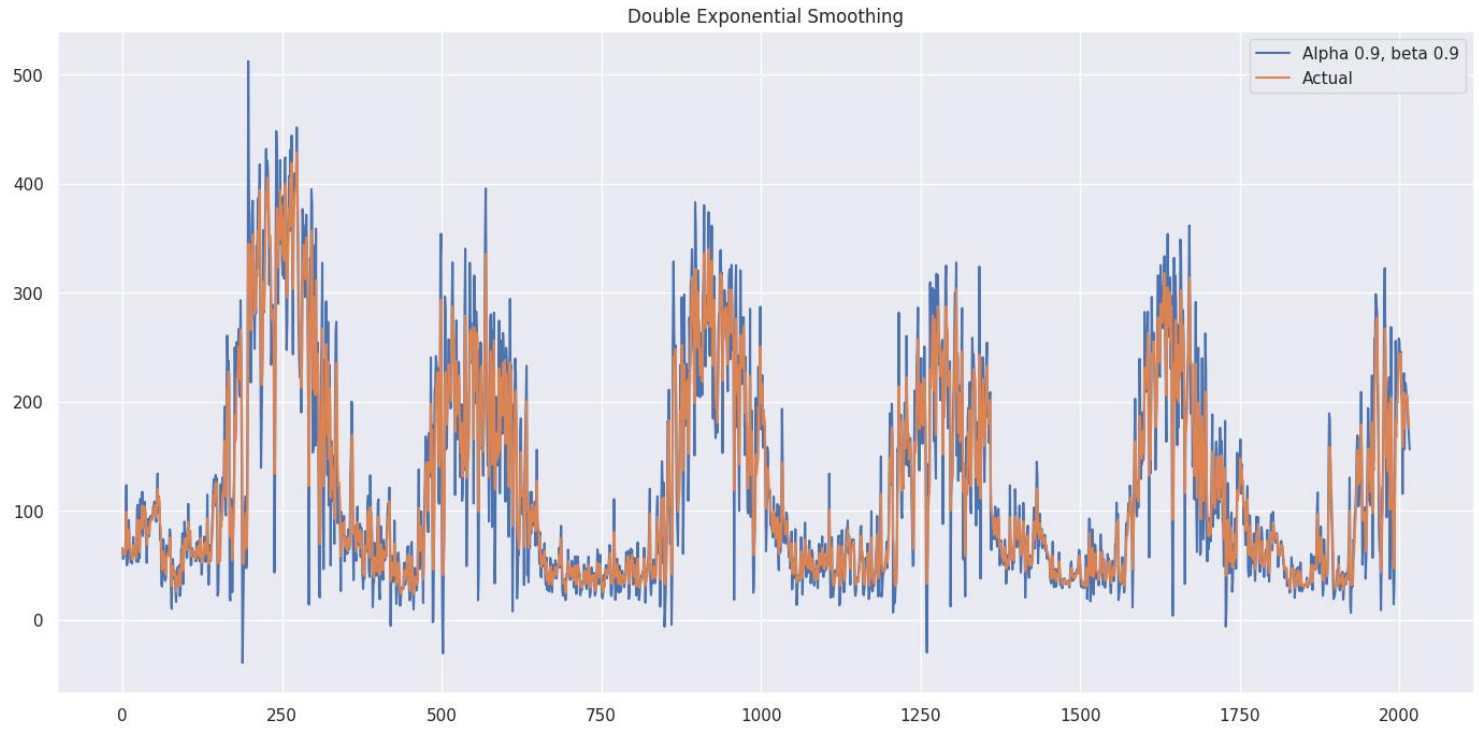

Figure 1 – Smoothed AQI values for Kolkata (daily)

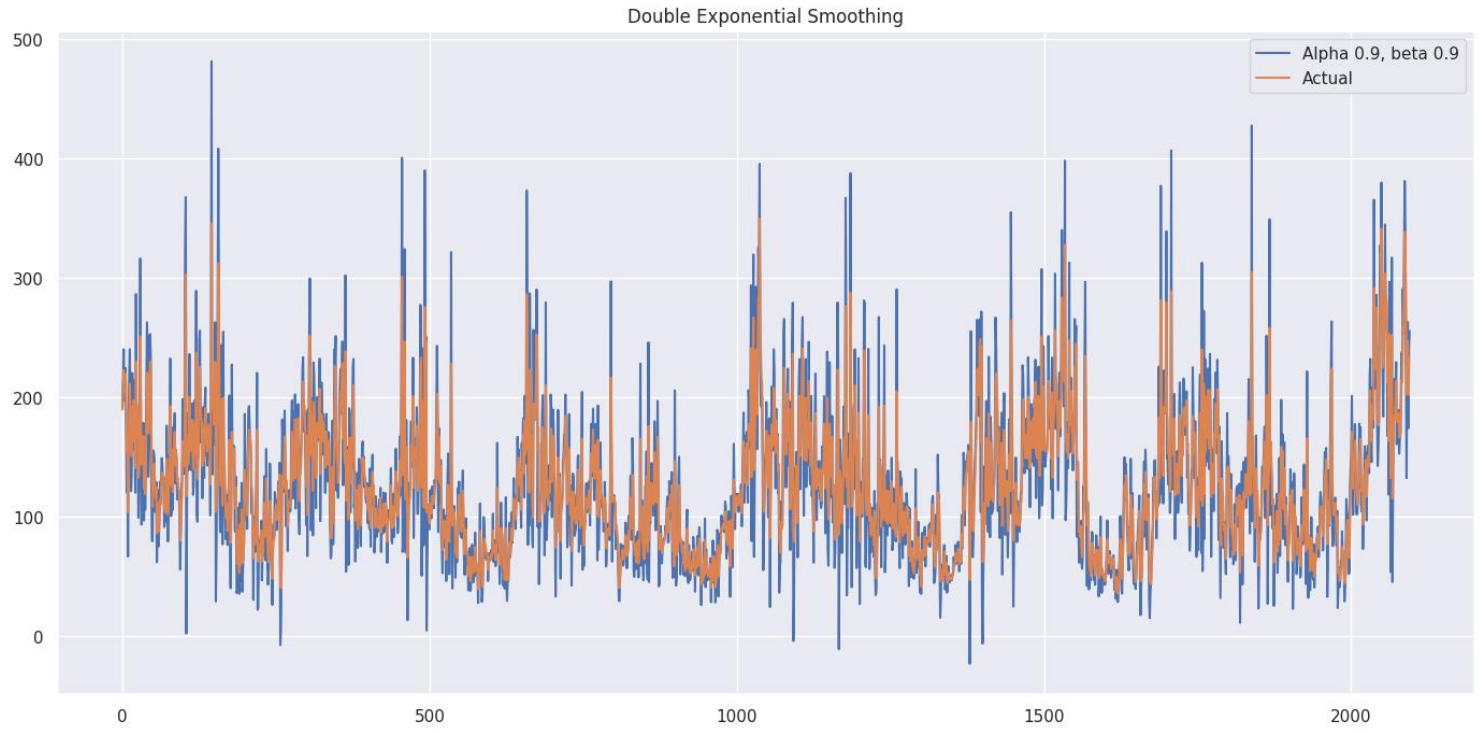

Figure 2 – Smoothed AQI values for Jaipur (daily)

Justification for Use

1. **Handles Trends Effectively:** AQI values often show trends influenced by temporal factors like weather or emissions. DES captures these dynamics, enabling realistic imputation of missing data.
2. **Avoids Overfitting:** Unlike more complex methods, DES provides a balanced approach to filling gaps without introducing unnecessary model complexity.
3. **Maintains Time Series Structure:** By interpolating missing values in alignment with the observed trends, DES preserves the overall time series integrity, crucial for accurate forecasting.
4. **Computationally Efficient:** DES is quick to implement and works well for moderately sized AQI datasets, making it practical for real-time or large-scale systems.

Using DES for filling missing AQI values ensures the continuity of data trends and supports reliable forecasting, aligning with the goal of maintaining high predictive accuracy while being methodologically sound.

2.3 ADF test

Augmented Dickey-Fuller test is used to check the stationarity. We performed the test with the hypotheses are as follows:

H_0: The series has a unit root
H_a: The series has no unit root

The test statistic for the ADF test is -

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \cdots + \delta_k \Delta y_{t-k} + \epsilon_t$$

where,

- Δy_t ***is the first difference of the series,***

- t *is a time trend* ($if included$)
- y_{t-1} *is the lagged level of the series*,
- Δy_{t-i} *are lagged differences*,
- k *is the number of lags used to account for autocorrelation*,
- γ *is the coefficient tested for significance*.

Note - ADF test is not a conclusive measure of stationarity of data hence we moved to ACF and PACF plots.

2.4 ACF and PACF Plots: Understanding and Applications

Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots are essential tools in time series analysis. They help understand the dependencies between observations and guide model selection, particularly for autoregressive integrated moving average (ARIMA) models.

Autocorrelation Function (ACF):

The ACF measures the correlation between a time series and its lagged values over different time lags.

- **Mathematical Representation:** For a time series $\{X_t\}$, the autocorrelation at lag k is

$$r_k = \frac{\mathrm{Cov}(X_t, X_{t-k})}{\sqrt{\mathrm{Var}(X_t) \cdot \mathrm{Var}(X_{t-k})}}$$

- **Purpose:**
  - Quantifies how current values depend on past values.
  - Indicates the presence of patterns, such as seasonality or persistence in trends.
- **ACF Plot:** A graphical representation of the autocorrelation coefficients at different lags. Peaks beyond a confidence interval (usually 95%) indicate statistically significant autocorrelations.

Partial Autocorrelation Function (PACF):

The PACF isolates the correlation between a time series and its lagged values, removing the influence of intermediate lags.

- **Conceptual Difference:** While ACF includes all indirect correlations (e.g., lag 3 includes effects of lags 1 and 2), PACF removes these indirect effects, showing the *direct* relationship at each lag.
- **Mathematical Representation:** The partial autocorrelation at lag k is the correlation between X_t and X_{t-k} after controlling for the effects of intermediate lags $1,2,\ldots,k-1$.
- **PACF Plot:** Similar to the ACF plot but represents partial autocorrelations. Peaks outside the confidence interval denote significant direct correlations at specific lags.

How They Are Used:

❖ **Model Identification (ARIMA):**

ACF and PACF plots are crucial for determining the order of autoregressive (AR) and moving average (MA) components in ARIMA models:

- **AR Models (p):**
  - ACF: Gradually declines.
  - PACF: Significant cutoff after lag *p*.
- **MA Models (q):**
  - ACF: Significant cutoff after lag *q*.
  - PACF: Gradually declines.
- **Mixed ARMA Models:** Both plots may show a combination of declining and cutoff patterns.

❖ **Identifying Seasonality:**

ACF plots reveal recurring peaks at multiples of seasonal lags, indicating periodic patterns in the data.

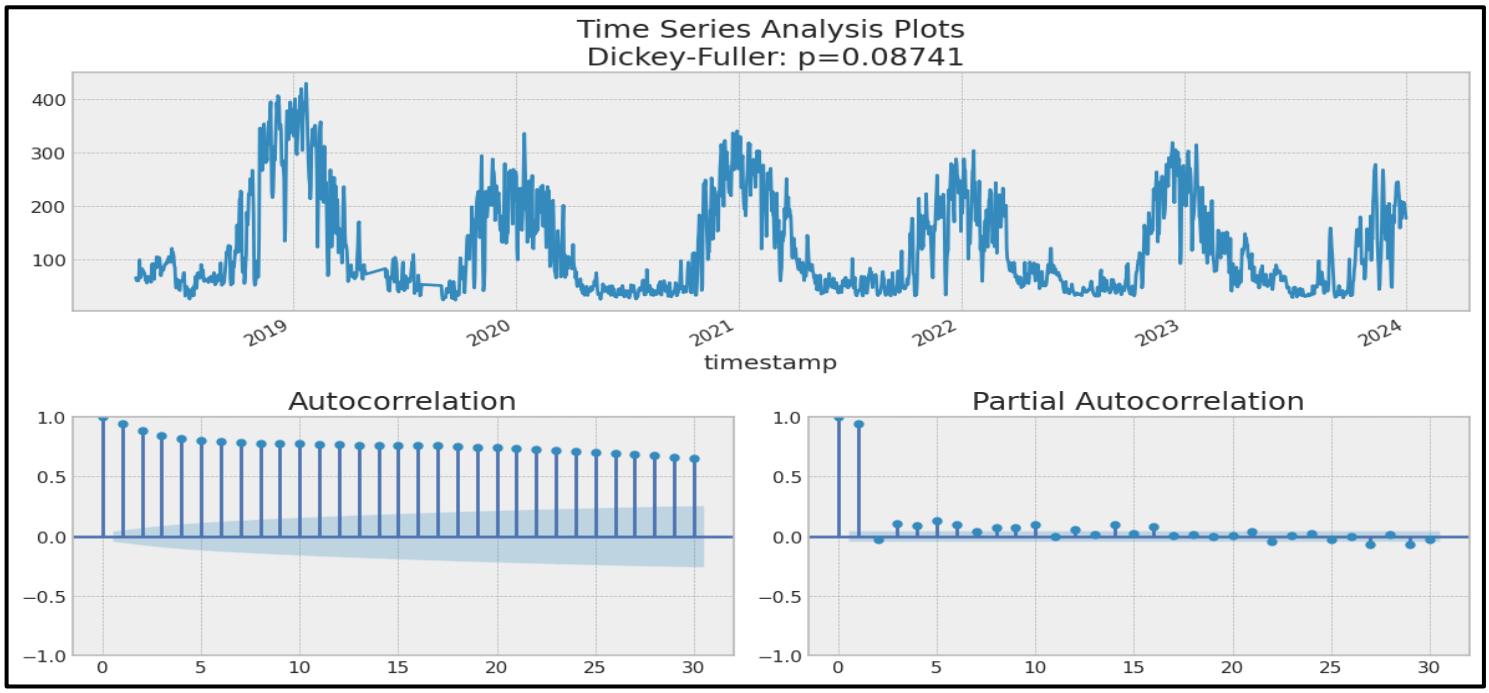

Figure 3 – ACF and PACF plots to check stationarity of data & determining model parameters for Kolkata (daily)

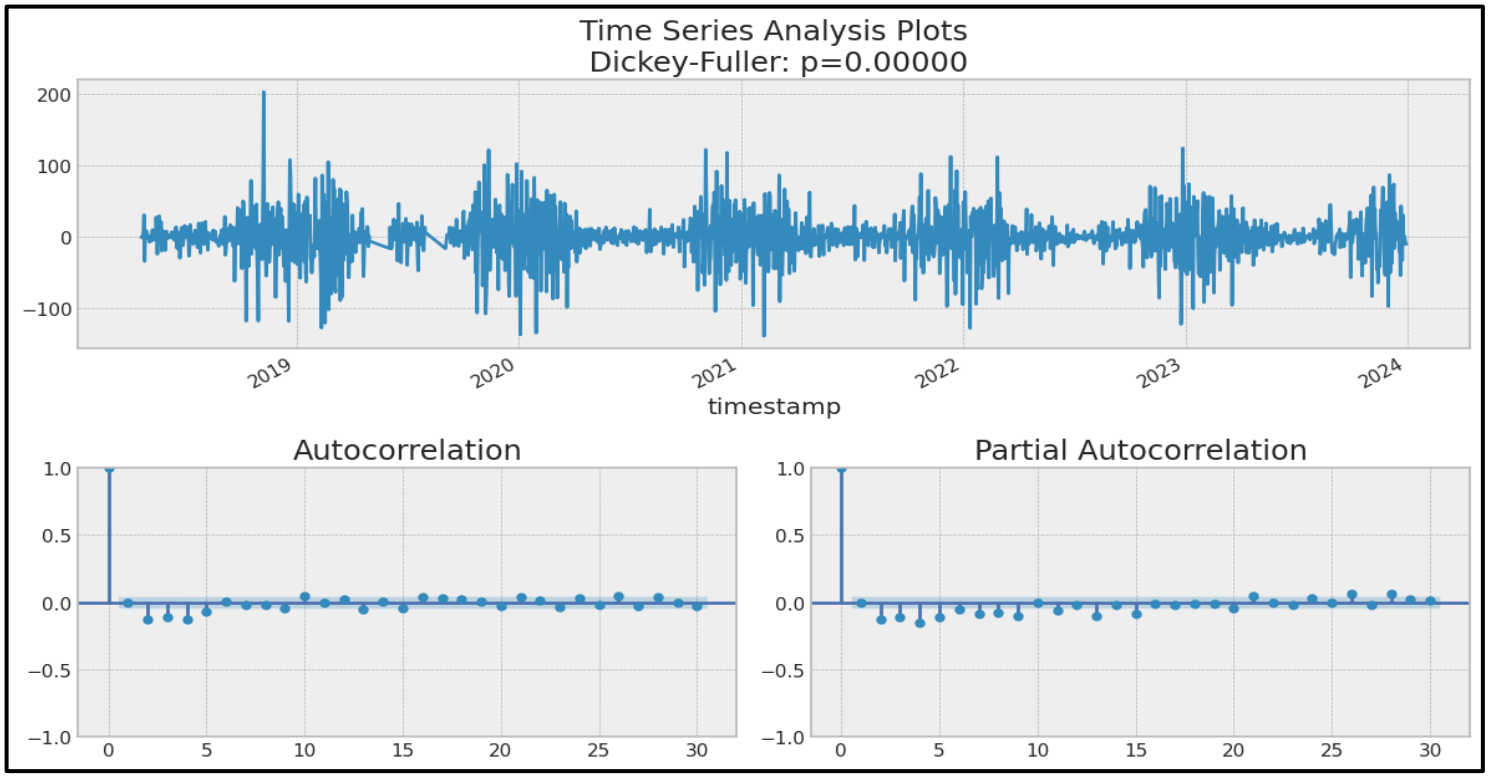

Figure 4 – ACF and PACF plots after differencing (d=1) for Kolkata (daily)

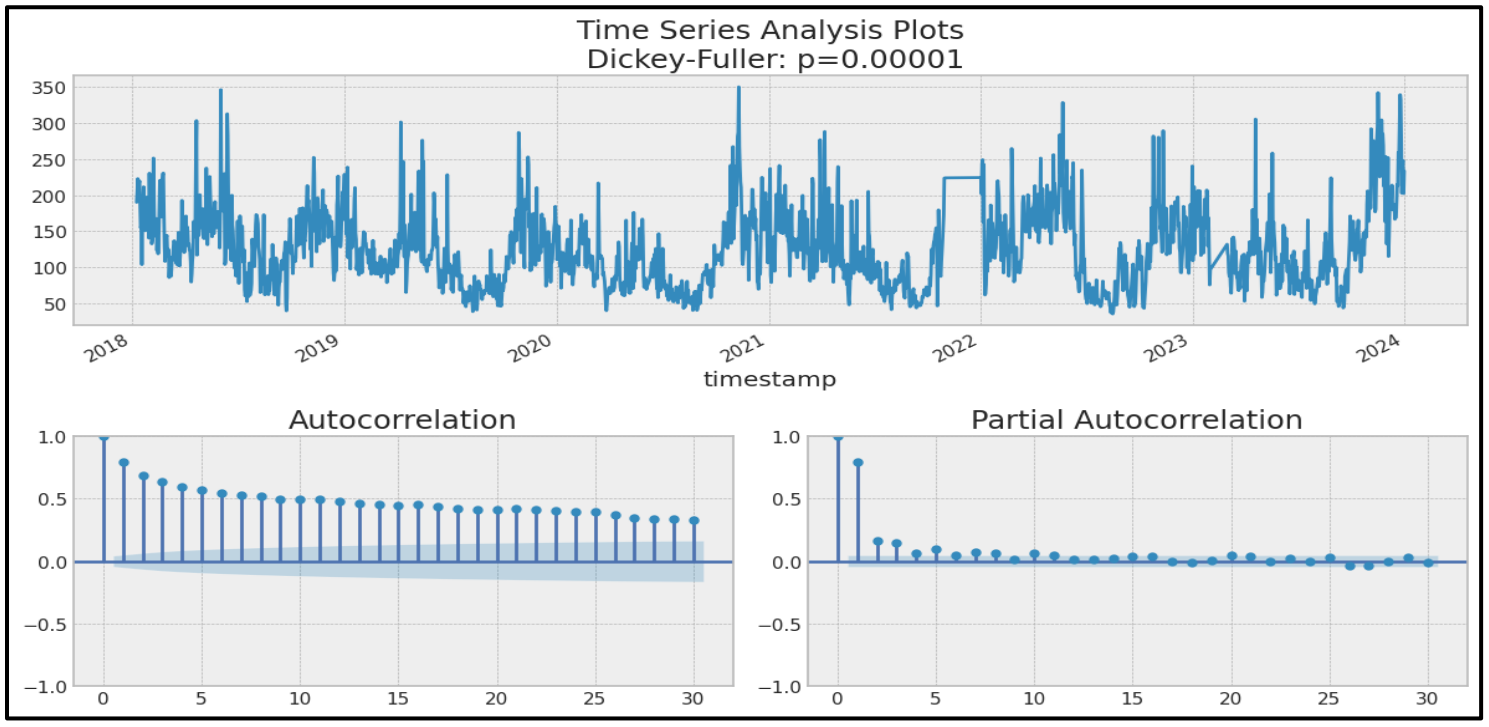

Figure 5 – ACF and PACF plots to check stationarity of data & determining model parameters for Jaipur (daily)

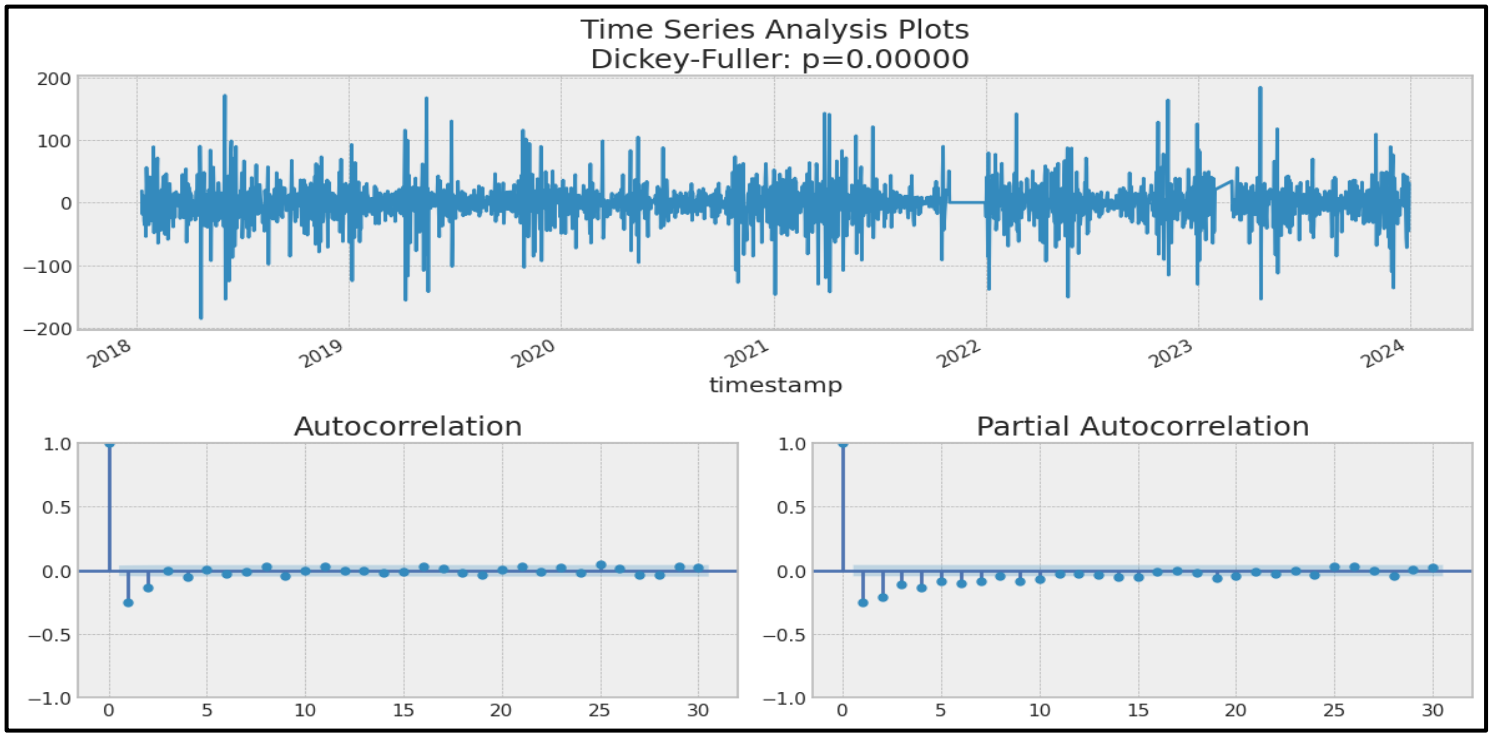

Figure 6 – ACF and PACF plots after differencing (d=1) for Jaipur (daily)

Conclusion – ACF plots show almost exponential decay proving that the data is not stationary hence differencing is used.

Chapter 3

ARMA, ARIMA, SARIMA Analysis of AQI data

3.1 ARMA

Autoregressive Moving Average or ARMA, as the name suggests, is a combination of the previous two models, Auto Regression and Moving Averages. ARMA can be used to describe a stationary time series in terms of two component polynomials, one of which corresponds to AR and the other to MA. The model has two parameters, *p* and *q*, for Auto-regression and Moving Averages respectively. ARMA (*p*,*q*) model is as follows –

$$\theta_p(B)x_t = \phi_q(B)w_t$$

where θ_p is a polynomial of order *p* and φ_q is a polynomial of order *q*.

3.2 ARIMA

ARIMA Autoregressive Integrated Moving Average is the full form of ARIMA. here we notice an addition of I or 'Integrated' term in ARMA; This is a measure of how many non-seasonal differences are needed to achieve stationarity. So, it converts the non stationary model into a stationary one by using differencing. Due to this, ARIMA can be applied on non-stationary models as well, unlike ARMA.

$$(1 - B)^d x_t$$

Here *B* denotes the backshift operator. The white noise is introduced due to differencing by order *d*. The ARIMA(*p*,*d*,*q*) essentially performs ARMA(*p*,*q*) on data that has been integrated by order *d*, which is the meaning of the appended 'I' - for integration. This model can be represented compactly as.

$$\theta_p(B)(1 - B)^d x_t = \phi_q(B)w_t$$

where θ_p and φ_q are polynomials of order *p* and *q* respectively.

3.3 SARIMA

ARIMA models are also capable of modelling a wide range of seasonal data. A Seasonal ARIMA model is formed by including additional seasonal terms in the ARIMA models.

$$\text{ARIMA} \quad \underbrace{(p,d,q)}_{\text{Non-seasonal part of the model}} \quad \underbrace{(P,D,Q)_S}_{\text{Seasonal part of of the model}}$$

The seasonal part of the model consists of terms that are similar to the non-seasonal components of the model, but involve backshifts of the seasonal period. The additional seasonal terms are simply multiplied by the non-seasonal terms. This can be expressed mathematically as:

$$\Theta_P(B^s)\theta_p(B)(1-B^s)^D(1-B)^d x_t = \Phi_Q(B^s)\phi_q(B)w_t$$

where $\Theta_P, \theta_p, \Phi_Q, \varphi_q$ are polynomials of orders *P*, *p*, *Q* and *q* respectively.

The seasonality in SARIMA appears as an exponent in this model and it essentially denotes the number of observations per year. It can be noticed that daily seasonality of 365 is highly computationally expensive.

We fit SARIMA (0,1,3)x(2,1,0) [12] following the results of this journal paper as finding optimal parameters for such large data with annual seasonality is beyond our computational capacity.

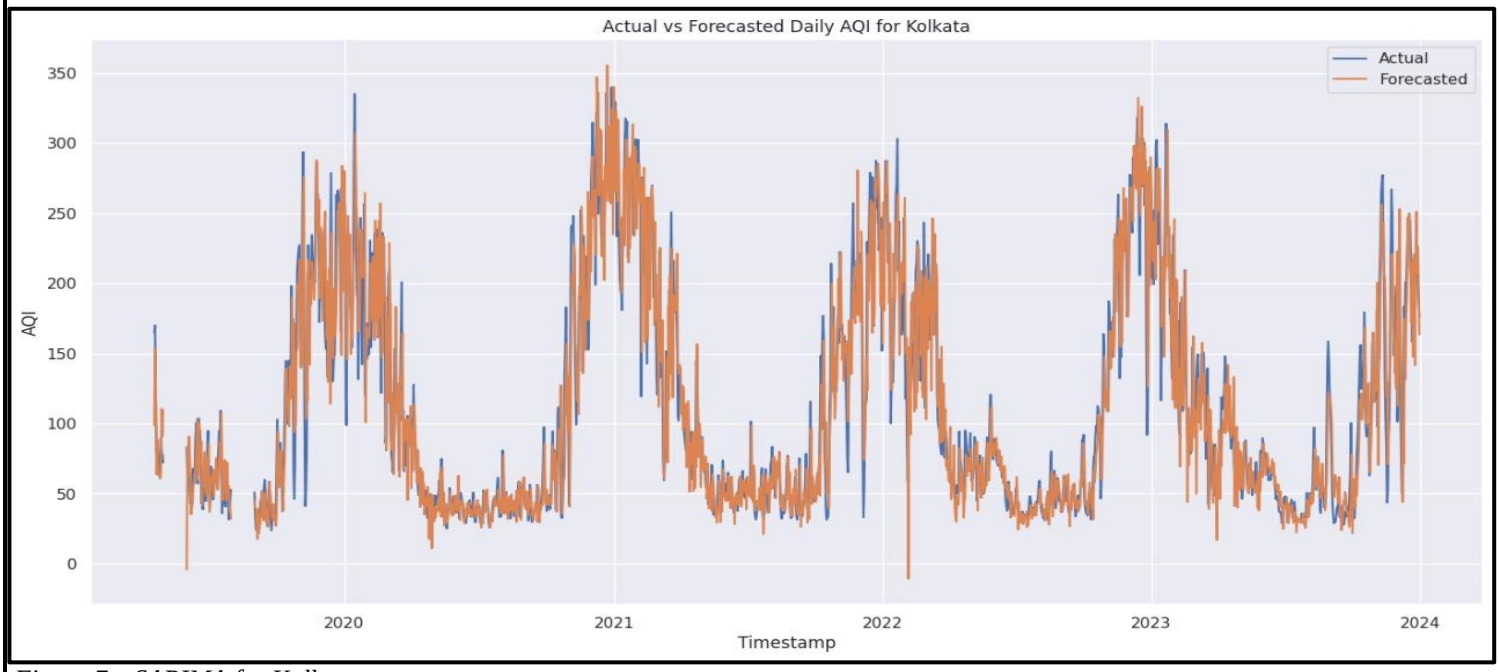

Figure 7 – SARIMA for Kolkata

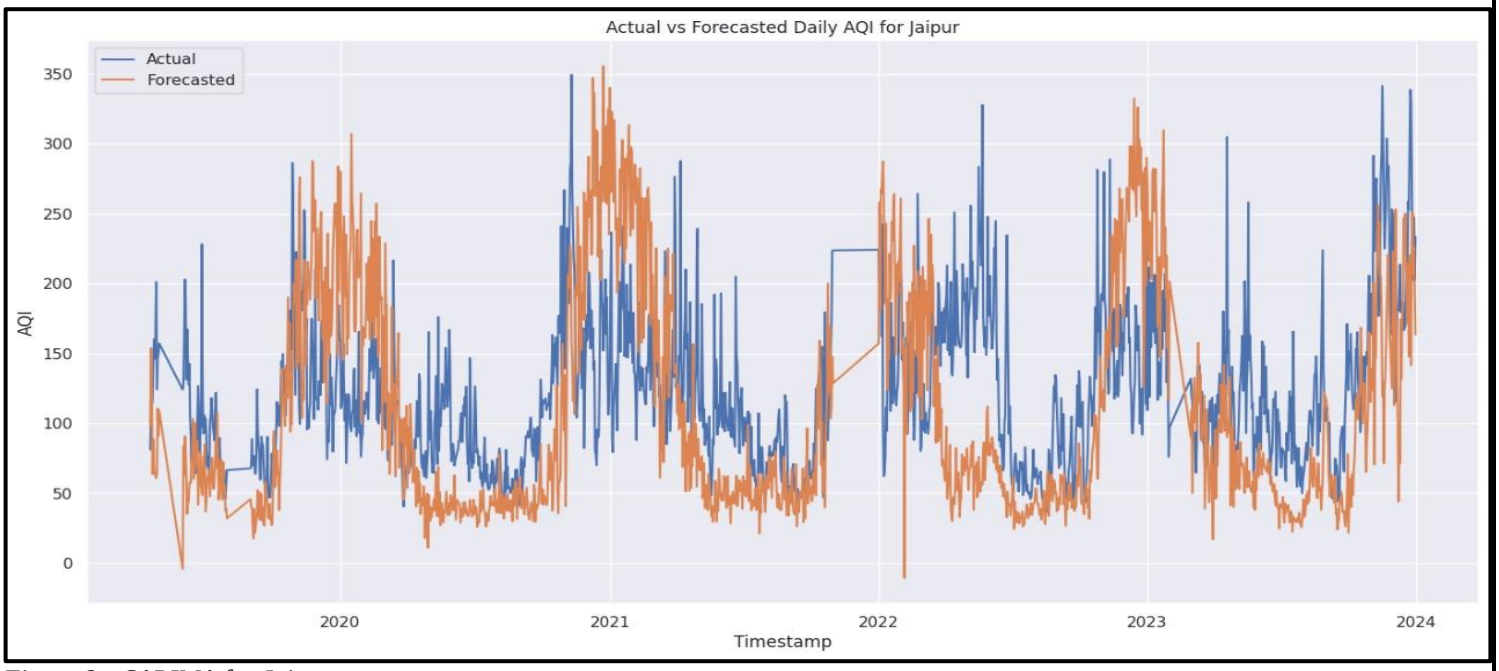

Figure 8 – SARIMA for Jaipur

Chapter 4

Time Series Model Validation

4.1 Metrics and Techniques

Time series models are evaluated and validated using various metrics and statistical tests to ensure accuracy, robustness, and suitability for the underlying data. Key metrics and tests include:

- **Sum of Squared Errors (SSE):**
  Measures the cumulative squared differences between observed and predicted values. Lower SSE indicates a model that fits the data better.

- **Root Mean Squared Error (RMSE):**
  A scale-dependent measure that provides the square r

- **Akaike Information Criterion (AIC):**
  Balances goodness of fit with model complexity. Lower AIC indicates a better model but penalizes overfitting.

  $$AIC=2k-2ln(L)$$

  where k is the number of parameters, and L is the likelihood of the model.

- **Bayesian Information Criterion (BIC):**
  Similar to AIC but with a stronger penalty for complexity, favoring simpler models for large datasets.

  $$BIC=kln(n)-2ln(L)$$

- **Kolmogorov-Smirnov (KS) Test:**
  The KS test assesses whether the residuals (differences between observed and predicted values) follow a specified distribution, typically a normal distribution for time series models. It compares the cumulative distribution function (CDF) of the residuals to the CDF of the assumed distribution.

4.2 Kolmogorov-Smirnov (KS) Test in Detail

The KS test is a *non-parametric statistical test* used to determine if a sample conforms to a specified distribution.

- **Null Hypothesis (*Ho*):** The sample data comes from the specified distribution.
- **Alternative Hypothesis (*Ha*):** The sample data does not come from the specified distribution.

Steps of the KS Test:

1. Compute the empirical cumulative distribution function (ECDF) of the sample data.
2. Compute the CDF of the reference distribution (e.g., normal or uniform distribution).
3. Calculate the maximum absolute difference between the ECDF and the reference CDF:

$$D = max|F_n(x) - F(x)|$$

   where $F_n(x)$ is the ECDF and $F(x)$ is the reference CDF.

4. Compare the D-statistic to the critical value for a given significance level (α).

Interpretation:

- If D is greater than the critical value, reject H0H_0H0, indicating the data does not follow the assumed distribution.
- A p-value can also be computed; a small p-value (e.g., p<0.05) implies rejection of *Ho*

Application in Time Series:

- Ensures that residuals are white noise, i.e., they have no structure or predictable pattern.
- Validates model assumptions about error distribution, essential for forecasting accuracy.

By combining these metrics and tests, analysts can thoroughly evaluate the performance and assumptions of their time series models.

4.3 Tests Results on AQI data

Table 1 – Metrics for SARIMA Model (Jaipur)

AIC	BIC	MAPE	KS Test
17234	**17239.5**	**21.92**	**Fail to reject *Ho***

Table 2 – Metrics for SARIMA Model (Kolkata)

AIC	BIC	MAPE	KS Test
57081.8	57088.5	21.8496	Fail to reject *Ho*

Chapter 5

LSTM for AQI forecasting

Long Short-Term Memory (LSTM) networks are a sophisticated type of recurrent neural network (RNN) designed to address the challenges of learning from sequential data, particularly the vanishing gradient problem that affects traditional RNNs. Introduced by Sepp Hochreiter and Jürgen Schmidhuber in 1997, LSTMs have become essential in various applications involving time series data, natural language processing, and more.

5.1 Unique Features of LSTMs

Memory Cells and Gates: LSTMs incorporate memory cells that can retain information over long periods. This is achieved through three types of gates:

- Input Gate: Controls the addition of new information to the memory cell.
- Forget Gate: Determines which information to discard from the memory cell.
- Output Gate: Regulates what information is output from the memory cell.

This architecture allows LSTMs to selectively remember or forget information, making them adept at capturing long-range dependencies within data sequences.

Handling Long-Term Dependencies: Unlike traditional RNNs, LSTMs can effectively manage sequences with long-term dependencies, enabling them to learn patterns that span thousands of time steps. This capability is critical for tasks where earlier inputs significantly influence current outputs.

5.2 Applications in Time Series Analysis

LSTMs are particularly advantageous for time series data due to their ability to model complex temporal patterns. They excel in:

- Financial Forecasting: Predicting stock prices based on historical data.
- Weather Prediction: Analysing past weather patterns to forecast future conditions.
- Anomaly Detection: Identifying unusual patterns in time series data, crucial for industries like finance and healthcare.

Their robustness in managing noise and irregularities makes LSTMs a preferred choice for many real-world applications involving sequential data.

5.3 Visual Representation

The following diagram illustrates the structure of an LSTM unit, highlighting its key components:

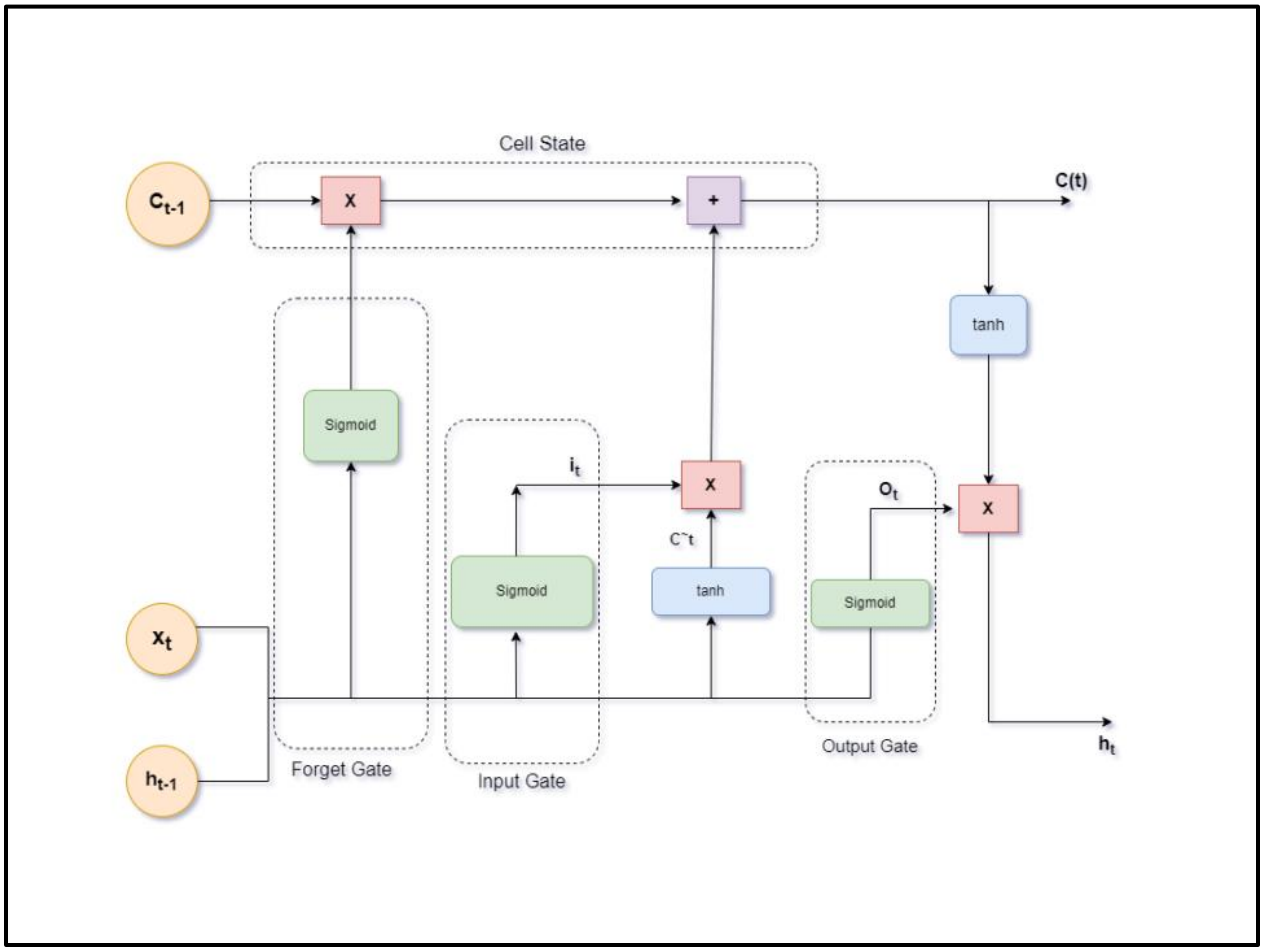

Figure 9 – Visual Depiction of LSTM Architecture

This image shows how the input, forget, and output gates interact with the memory cell to manage the flow of information effectively. The design enables LSTMs to maintain relevant information over extended periods while discarding unnecessary details, which is essential for accurate predictions in time-dependent tasks.

In summary, LSTMs represent a significant advancement in neural network architecture, providing robust solutions for sequence prediction problems across various domains.

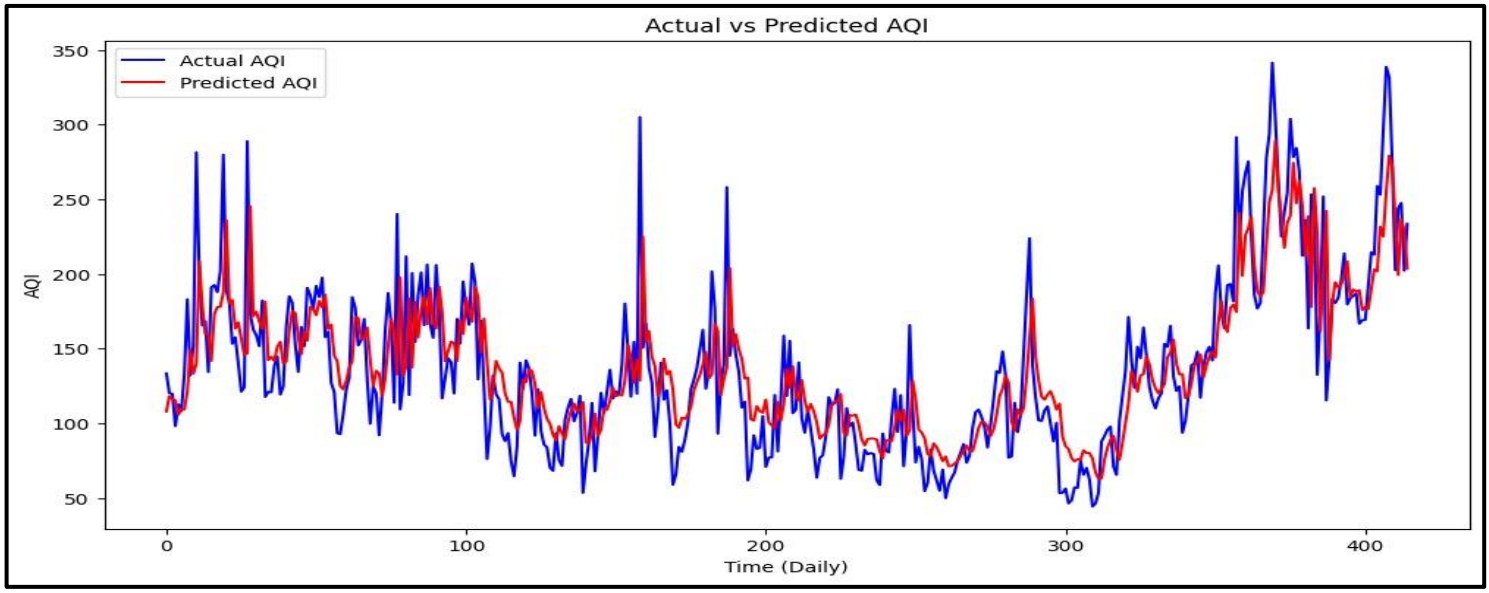

Figure 10 – LSTM forecasts for Jaipur

Code for Reproducibility of results

SARIMA - https://colab.research.google.com/drive/1waKwTbVD5dRyeJguV07EGrmq2onXhPLW?usp=sharing
LSTM - https://colab.research.google.com/drive/10geGQK03BC8cgL-xrPewjvdaVbBkGqiT?usp=sharing

Conclusion

In this study, we have conducted a comprehensive analysis of air quality forecasting methodologies applied to AQI data for the cities of Jaipur and Kolkata. By leveraging statistical and machine learning models, we aimed to understand the underlying patterns, assess data properties, and develop robust forecasting solutions for AQI prediction.

The analysis began with data preprocessing, where missing values were effectively handled using Double Exponential Smoothing (DES), ensuring continuity and preserving temporal trends. Statistical tests, including the Shapiro-Wilk and Augmented Dickey-Fuller (ADF) tests, confirmed that the hourly AQI data exhibited non-normality and non-stationarity, necessitating advanced modelling techniques. Autocorrelation and Partial Autocorrelation Function (ACF and PACF) plots further provided insights into the lag dependencies and guided the selection of appropriate ARIMA-based models.

The deployment of ARMA, ARIMA, and SARIMA models showcased their efficacy in capturing seasonal and non-seasonal patterns in AQI data. For Kolkata and Jaipur, SARIMA was instrumental in addressing the seasonal variations observed in AQI levels, with model parameters optimized for computational feasibility and accuracy. Metrics such as Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Mean Absolute Percentage Error (MAPE) validated the robustness of the SARIMA models. Both cities demonstrated a strong alignment between predicted and observed AQI values, confirming the suitability of these models for medium-term forecasting.

To enhance the accuracy and adaptability of the forecasting system, Long Short-Term Memory (LSTM) networks were introduced. The memory cell and gating mechanisms in LSTM allowed the model to learn complex temporal dependencies, outperforming traditional methods in scenarios with high variability and noise. The results demonstrated that LSTMs could capture intricate temporal dynamics and provide accurate short-term AQI forecasts, making them an ideal choice for modern air quality prediction systems.

Key findings from the study include:

1. **Seasonality and Trend Dependence**: AQI levels in Jaipur and Kolkata exhibit significant seasonal variations, influenced by meteorological conditions and localized emission sources.
2. **Model Efficiency**: SARIMA models are effective for structured seasonal forecasting, while LSTM models excel in capturing nonlinear patterns and providing flexible predictions.
3. **Statistical Validation**: Rigorous statistical testing, including the Kolmogorov-Smirnov test, confirmed that model residuals followed acceptable distributions, ensuring reliable forecasts.

This study underscores the critical role of data preprocessing, model selection, and validation in developing accurate AQI forecasting frameworks. The integration of traditional time-series models with advanced neural networks like LSTM presents a scalable and adaptive solution for air quality prediction. Future work can focus on incorporating real-time meteorological data, multi-source emission inventories, and external socioeconomic factors to further enhance forecasting accuracy.

In conclusion, this research provides a robust foundation for air quality management strategies, empowering policymakers with actionable insights to mitigate pollution and protect public health. The methodologies demonstrated here can be extended to other cities and regions, contributing to a sustainable and data-driven approach to environmental planning.

References

- Peixeiro, M. (2023, August 1). The complete guide to Time Series models. Built In. https://builtin.com/data-science/time-series-model

- Statsmodels. (n.d.). GitHub - statsmodels/statsmodels: Statsmodels: statistical modeling and econometrics in Python. GitHub. https://github.com/statsmodels/statsmodels

- How to Develop LSTM Models for Time Series Forecasting. (2020). Jason Brownlee. Retrieved November 22, 2024, from https://machinelearningmastery.com/how-to-develop-lstm-models-for-time-series-forecasting/

- Maltare, N. N., & Vahora, S. (2023). Air Quality Index prediction using machine learning for Ahmedabad city. Digital Chemical Engineering, 7, 100093. https://doi.org/10.1016/j.dche.2023.100093

- Pant, A., Joshi, R. C., Sharma, S., & Pant, K. (2023). Predictive Modeling for Forecasting Air Quality Index (AQI) using Time series analysis. Avicenna Journal of Environmental Health Engineering, 10(1), 38–43. https://doi.org/10.34172/ajehe.2023.5376