SC475 - Time Series Analysis Project

Kashyap Trivedi (202201191) and Zeel Boghara (202201201)
Time Series Analysis (SC475)
Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar, India
Email: 202201191@daiict.ac.in, 202201201@daiict.ac.in

Under the guidance of: Prof. Mukesh Tiwari

***Abstract*—Time series analysis is a powerful statistical technique used for understanding and forecasting data points collected over time. In this project, we have conducted a comprehensive time series analysis of an Air Quality dataset, specifically focusing on NO_2 concentration levels. The primary objective is to explore temporal patterns and forecast future values using methods such as moving averages, seasonal decomposition. Through visualization and statistical modeling, we analyze the underlying trend, seasonality, and randomness present in the data. This analysis provides insights into the behavior of NO_2 levels over time, which can be crucial for environmental monitoring and policy-making.**

I. Introduction

This dataset presents comprehensive NO_2(ug/m3) concentration data for India, spanning the years 2016 to 2023. Sourced from the Central Control Room for Air Quality Management, the data offers reliable and authoritative insights into NO_2(ug/m3) Concentration trends and conditions throughout the country. It serves as a valuable resource for analyzing environmental patterns and assessing the effectiveness of air quality management efforts over time. Time series analysis techniques are applied to extract meaningful insights and forecast future trends.

II. Interpolation For Null Values

The original dataset consisted of hourly NO_2 concentration readings. Approximately 6.5% of the data contained missing (NULL) values, which were handled using interpolation to ensure continuity. After filling the gaps, the data was resampled and aggregated to a daily frequency to facilitate trend and seasonality analysis.

Interpolation: Interpolation is a technique used to estimate and fill missing values within a dataset by using the known values surrounding the gaps. In this study, it was applied to maintain data continuity and preserve the overall trend and patterns before further analysis.

III. NO_2(ug/m3) Concentration Dataset Description

Title: Daily Average Of NO_2(ug/m3) Concentration (2016–2023)

Description:
This dataset contains hourly NO_2 concentrations in India (in ug/m3) from 2016 to 2023. The data exhibits an initial downward trend followed by a noticeable upward trend, accompanied by seasonal fluctuations—highlighting changes in NO_2 concentration levels over the years and recurring daily patterns.

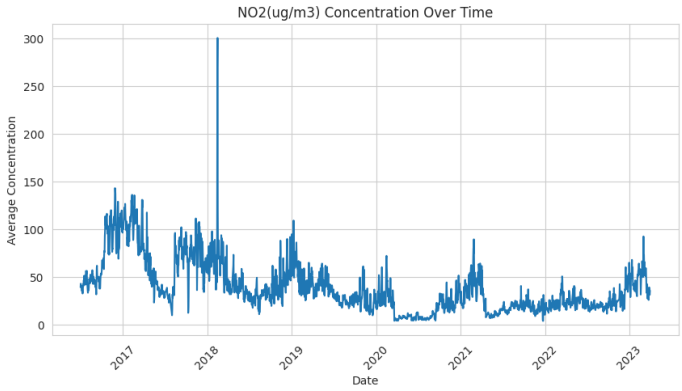

Fig. 1: Average Of NO_2(ug/m3) Concentration Over Time

Conclusion

The NO_2 concentration time series from 2016 to 2023 shows strong seasonal patterns and occasional sharp spikes, notably around 2018–2019. A significant drop is observed in 2020, likely due to COVID-19 lockdowns, followed by a gradual recovery. The recurring fluctuations indicate seasonality, making the data suitable for time series analysis and forecasting.

IV. Rolling Statistics for Stationarity Check

Before applying time series forecasting models, it is important to ensure that the data is stationary. A stationary series has constant mean and variance over time. In this section, we plot the 12-month Rolling Mean and Rolling Variance of the NO_2(ug/m3) Concentration dataset along with the original time series to visually inspect stationarity.

V. Observations from Time Series Plots

- The series does not exhibit a strong **upward or downward** trend over the full period, but shows significant short-term fluctuations.

Fig. 2: Rolling Mean and Rolling Variance (Window=12) over the Original NO_2(ug/m3) Concentration Series.

- Sharp spikes are observed, particularly around **late 2017 and mid-2018**, indicating possible pollution events or data anomalies.
- There is some cyclical or seasonal behavior, although less pronounced than in datasets like air travel.

A. Rolling Mean and Rolling Variance

- The **12-month rolling mean (red dashed line)** captures the underlying trend by smoothing the short-term noise, making it easier to observe persistent shifts in NO_2 concentration over time.
- The **rolling variance (orange dashed line)** shows fluctuations over time, with peaks aligning with extreme events or volatile periods (notably around 2017–2018).
- The variance remains elevated during periods of instability, suggesting that the volatility of NO_2 levels is not constant.

B. Conclusion

From the rolling plots:

- Both the rolling mean and rolling variance change over time, implying that the series is non-stationary.
- This non-stationarity suggests that techniques like differencing or transformation might be necessary before applying forecasting models.
- The volatility and episodic spikes indicate the influence of external environmental factors or pollution episodes.

VI. Year-wise Monthly Trends

A grid of compact subplots was generated, each representing Monthly NO_2(ug/m3) Concentration Trends for Each Year (2016–2023)

- The yearly plots reflect distinct **seasonal variations** in NO_2 concentration levels, with noticeable fluctuations observed across months.
- A significant spike in early 2018 suggests a potential **pollution event or data anomaly**, which stands out from other years.
- From 2020 onwards, NO_2 levels appear relatively **lower and more stable**, possibly influenced by lockdowns or environmental regulations during the COVID-19 pandemic.
- The consistency in values during 2021 to 2023 indicates a phase of **controlled emissions or reduced urban activity**.

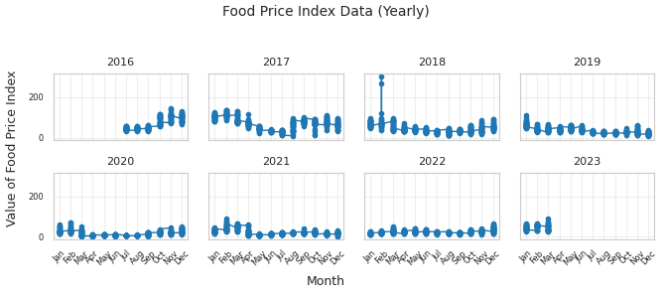

Fig. 3: Monthly NO_2(ug/m3) Concentration Trends for Each Year (2016–2023)

Conclusion: The data shows both periodic seasonal patterns and distinct anomalies in NO_2 concentration. Early years exhibit higher variability, while recent years suggest improved air quality and emission control.

VII. Rolling Variance vs. Rolling Mean

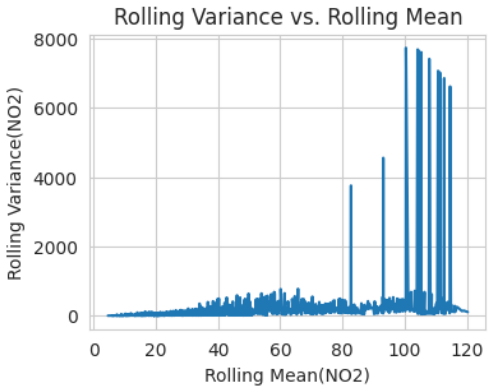

Fig. 4: Relationship Between Rolling Mean and Rolling Variance

The plot reveals a non-linear relationship between the rolling mean and rolling variance, indicating variability in the data. Therefore, a transformation of the original series is recommended to stabilize variance before further modeling.

Conclusion: The direct relationship between rolling mean and variance is strong evidence of **non-stationarity**.

VIII. Rolling Variances for Various Transformations

This set of plots visualizes the rolling variance for different transformations of the passenger series, each with a 12-month rolling window.

- **Square Root Transformation**: The rolling variance of the square root-transformed passenger data is shown in purple and dark orange.

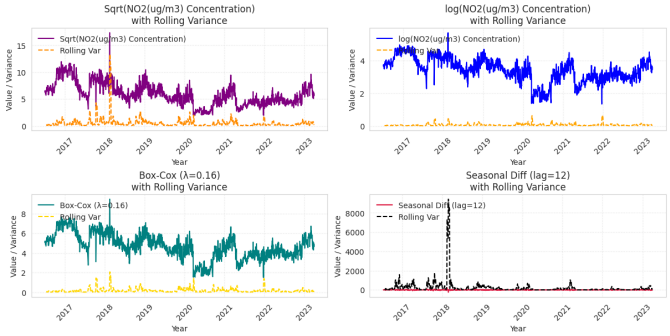

Fig. 5: Rolling Variance for Various Transformations

- **Logarithmic Transformation**: The rolling variance of the log-transformed passenger data is plotted in blue and orange.
- **Box-Cox Transformation**: The rolling variance after applying the Box-Cox transformation with $\lambda \approx$ fitted_lambda is displayed in teal and gold.
- **Seasonal Differencing**: The rolling variance for the seasonal difference (lag=12) is represented in crimson and black.

Conclusion: After taking log transformation, variance becomes almost constant

IX. Analysis of Time Series Transformations

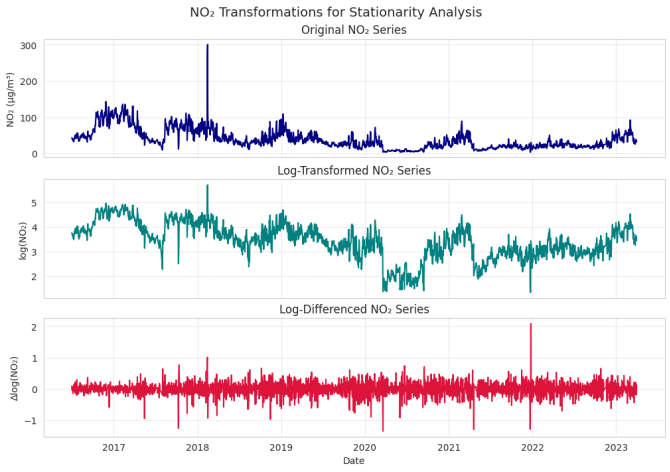

Fig. 6: Data Transformations Plots

- **Original NO_2 Series (Top)**:
  - Shows clear long-term variation with periods of volatility, indicating non-stationary behavior.
  - Variance appears to increase during certain time intervals, suggesting variance of the data changes over time .
- **Log-Transformed NO_2 Series (Middle)**:
  - Transformation compresses extreme values and stabilizes variance to some extent.
  - Trend remains visible, but fluctuations are more controlled compared to the original series.
- **Log-Differenced NO_2 Series (Bottom)**:
  - Removes both trend and changing variance, making the series more stationary.
  - Residual fluctuations are more consistent and suitable for ARIMA(1,1,1) modeling.

ADF Test Results: To evaluate stationarity, the Augmented Dickey-Fuller (ADF) test was conducted on the transformed NO_2 series:

ADF Test Results on Log-Transformed NO_2:

- **Before Differencing:**

$$\text{ADF Statistic} = -2.7695$$
$$\text{p-value} = 0.0628$$

Interpretation: The p-value is greater than 0.05, so we fail to reject the null hypothesis. The series is likely non-stationary.

- **After Differencing:**

$$\text{ADF Statistic} = -20.2411$$
$$\text{p-value} = 0.0000$$

Interpretation: The p-value is less than 0.05, and the ADF statistic is far below critical values. This confirms that the differenced log-transformed series is stationary.

X. QQ-Plot Analysis

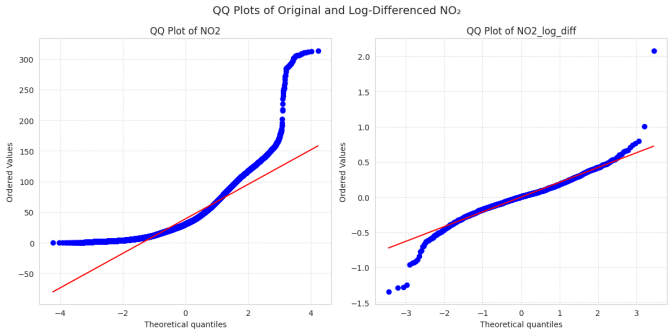

Fig. 7: QQ-plot of NO_2 And Log Difference of NO_2

Right Plot (Log-Differenced NO_2 Series): After applying a logarithmic transformation followed by differencing, the data points align more closely with the theoretical quantile line. This suggests that the transformed series approximately follows a normal distribution, satisfying one of the key assumptions for ARIMA modeling.

XI. ACF PACF Analysis

1. ACF and PACF of Original NO_2 Series (Top Row):

- **ACF:** The autocorrelation function shows a very slow exponential decay, suggesting the presence of strong trend and non-stationarity in the original NO_2 data.
- **PACF:** The partial autocorrelation function has a significant spike at lag 1 and then cuts off more quickly. This

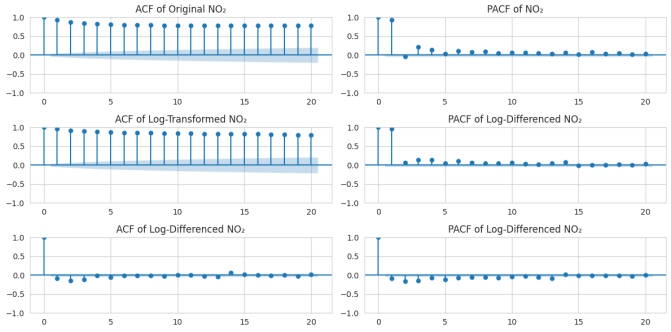

Fig. 8: ACF And PACF Plot

indicates that while some short-term dependency exists, the series is still non-stationary.

- **Conclusion:** The original series is non-stationary and not suitable for ARIMA(1,1,1) modeling in its raw form.

2. ACF and PACF of Log-Transformed NO_2 Series (Middle Row):

- **ACF:** The ACF still shows strong autocorrelation across many lags, though the pattern is slightly more dampened than in the original series.
- **PACF:** Significant spikes in the first few lags are still visible, suggesting persistent autocorrelation and continued non-stationarity.
- **Conclusion:** Log transformation alone is insufficient to achieve stationarity; differencing is still required.

3. ACF and PACF of Log-Differenced NO_2 Series (Bottom Row):

- **ACF:** ACF cuts off quickly after lag 1, which is characteristic of a moving average (MA) process.
- **PACF:** PACF also drops sharply after lag 1, indicating a possible autoregressive (AR) component at lag 1.
- **Conclusion:** The log-differenced series appears to be stationary, as the autocorrelations are much weaker and short-lived. This transformation prepares the data well for ARMA (1,1) modeling.

XII. DECOMPOSITION OF LOG(NO_2)

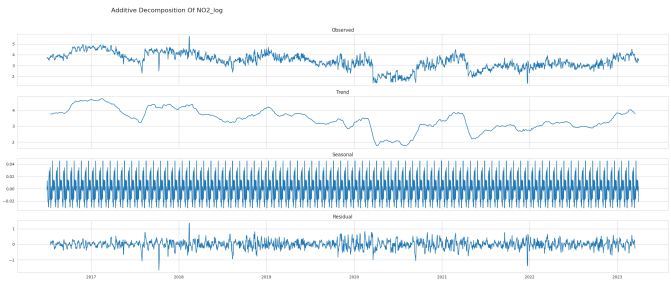

Fig. 9: Trend , Seasonal Residual Decompose

- Any time series consist of the components **trend** , **seasonality** and **Residual**.

A. Choice of Decomposition Model

There are two primary types of time series decomposition:

1) **Additive Decomposition**
2) **Multiplicative Decomposition**

In additive decomposition, the components of a time series are assumed to combine linearly:

$$Y_t = T_t + S_t + R_t$$

In our case, the trend component is not proportional to the seasonal component, and the seasonal variation remains relatively constant over time. Hence, **additive decomposition** is more suitable for analyzing the log-transformed NO_2 data.

B. Additive Decomposition of Log-Transformed NO_2

Figure 9 presents the additive decomposition of the log-transformed NO_2 time series into its primary components: *Observed*, *Trend*, *Seasonal*, and *Residual*.

- **Observed:** The log-transformed NO_2 values exhibit notable fluctuations, indicating the presence of both long-term trends and recurring seasonal patterns.
- **Trend:** The trend component reveals underlying long-term movement in the data. A clear rise and fall is observed around 2018 and 2020–2021, followed by a gradual increase thereafter. These changes could be associated with environmental policies, weather influences, or socio-economic disruptions (e.g., lockdowns).
- **Seasonal:** A distinct and consistent seasonal pattern is evident. The repetition suggests a regular cycle—likely daily or weekly—possibly related to traffic patterns or industrial activity. This justifies incorporating seasonality in the modeling approach.
- **Residual:** The residuals capture random variations that remain after removing trend and seasonality. These values fluctuate around zero with a few spikes, indicating occasional anomalies or shocks. The residuals appear stationary, which supports further time series modeling.

XIII. DETRENDING OF LOG(NO_2)

Detrending Methods:

- Differencing the log-transformed series.
- Moving average of the log-transformed series.
- Using In-Built Function
- Regression Method
- **Detrended (First Difference of log(NO_2))** :
  - After applying first-order differencing to the time series, the trend component is effectively removed, resulting in a stationary series suitable for further time series modeling.
- **Detrended (log(NO_2) - Moving Average)** :
  - Moving average stabilizes the fluctuations, and the variance shows less volatility after removing the trend using a moving average.

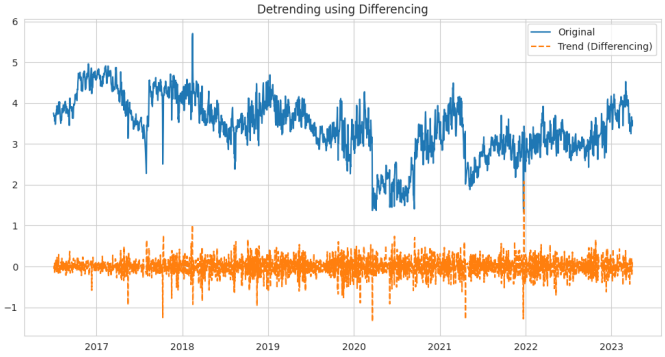

Fig. 10: Detrended (First Difference of log(NO2))

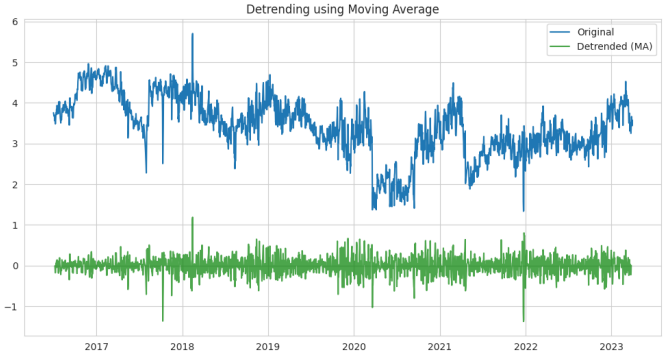

Fig. 11: Detrended (log(NO2) - Moving Average)

- **Detrended ($\log(NO_2)$) - Using In-Built Function)** :
  - To better understand and isolate the components of the NO_2 time series, the built-in `seasonal_decompose` function from the `statsmodels` library was employed using an additive model. The decomposition separates the time series into trend, seasonal, and residual components.

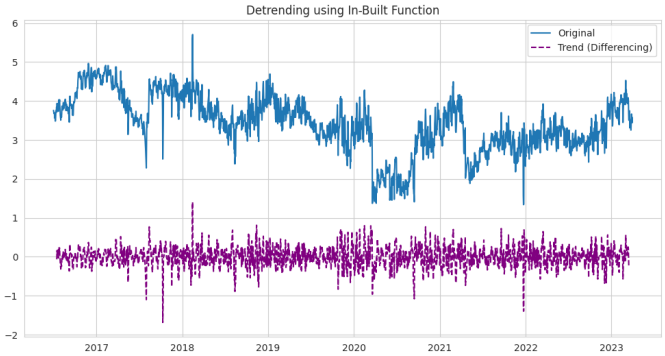

Fig. 12: Detrended (log(NO2) - Decomposition Trend)

- **Detrended ($\log(NO_2)$) - Regression Method)** :
  - In this method, an in-built polynomial regression function of degree 4 is applied,as is has minimum MSE to model the underlying trend in the original NO_2 data.
  - The fitted polynomial curve (red dashed line) represents the estimated trend, which is subtracted from the original data (blue line) to obtain the detrended series (green line).
  - This technique is especially useful when the trend is non-linear and can be well-approximated by a polynomial of a suitable degree.

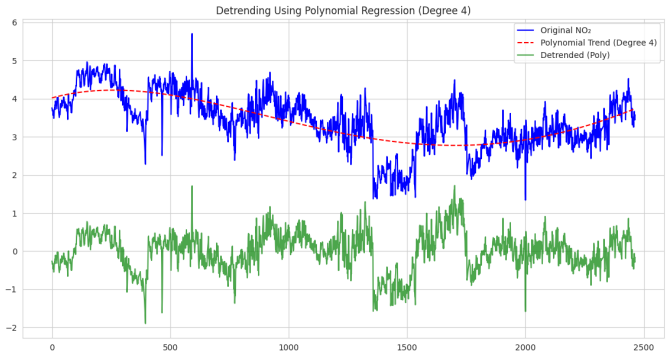

Fig. 13: Detrended ($\log(NO_2)$) -Regression Method)

Conclusion: The rolling mean and variance of the detrended data help assess the effect of various detrending methods. The results indicate stabilization of the trend, with reduced volatility in the moving averages after the trend removal process.

XIV. Deseasonalized Of $\log(NO_2)$

Deseasonalized Methods:

- Local Trend Method.
- Moving average of the log-transformed series.
- Lag Operator Method

- **Deseasonalized (Local Trend Method)** :
  - The Local Trend Method is a smoothing-based technique that separates the time series into trend, seasonal, and residual components.
  - It identifies and removes seasonal effects by estimating local trends using moving averages or locally weighted regression (LOESS).
  - This method adapts to short-term fluctuations, making it suitable for time series with non-constant seasonal patterns.
  - As shown in the graph, once the seasonal component is removed, the resulting series (in orange) captures the underlying trend and irregular variations without periodic seasonality.

- **Deseasonalized ($\log(NO_2)$) - Moving Average)** :
  - The **original series** (shown in blue) includes both trend and seasonal variations.

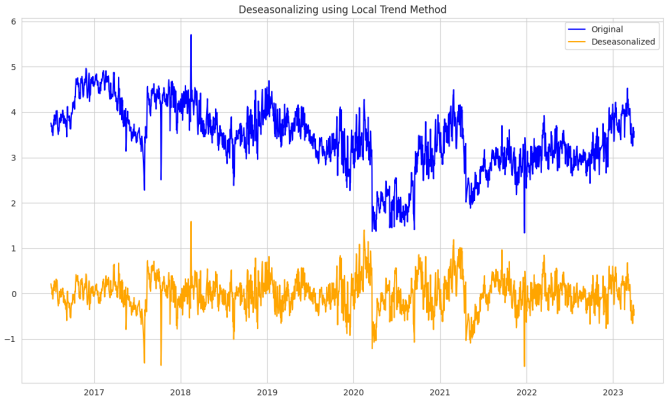

Fig. 14: Deseasonalized (Local Trend Method)

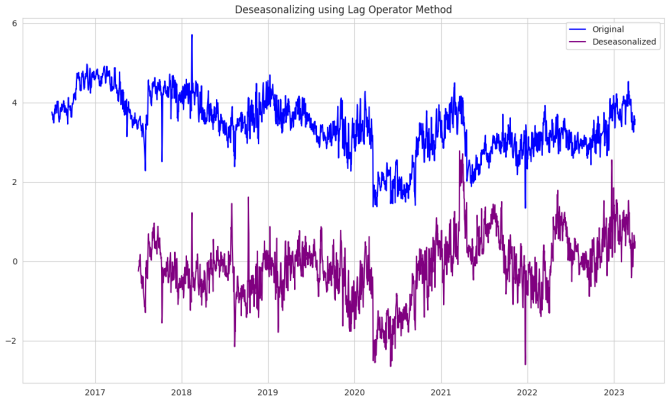

Fig. 16: Deseasonalized ($\log(NO_2)$ - Lag Operator Method)

- A **centered moving average** (orange dashed line) is calculated to estimate the seasonal component by smoothing out short-term fluctuations.
- The **deseasonalized series** (green line) is obtained by subtracting the estimated seasonal component from the original series:

$$Y_{\text{deseasonalized}} = Y_{\text{original}} - Y_{\text{seasonal}}$$

- This helps isolate the trend and irregular components, making further modeling or analysis more effective.

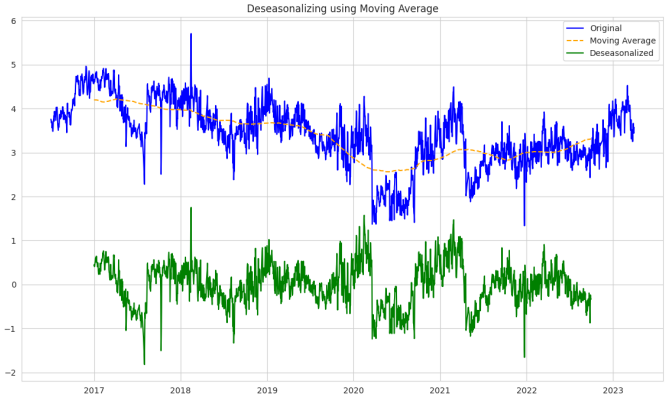

Fig. 15: Deseasonalized ($\log(NO_2)$ - Moving Average)

- **Deseasonalized ($\log(NO_2)$ - Lag Operator Method)** :
  - The **original series** (blue line) contains both seasonal and trend components.
  - We apply a lag difference, typically using a lag corresponding to the seasonal period (e.g., 12 for monthly data):

$$Y_t^{\text{deseasonalized}} = Y_t - Y_{t-s}$$

where s is the seasonal lag.
  - The resulting **deseasonalized series** (purple line) effectively removes the repeating seasonal structure, highlighting the trend and irregular components.

Conclusion: Deseasonalization is essential for analyzing the underlying patterns in a time series by removing seasonal effects. The Moving Average method smooths the data to extract seasonal components, while the Local Trend method isolates and removes both local trend and seasonality using smoothing filters. The Lag Operator method eliminates seasonality by differencing the series at a fixed seasonal interval. Each method has its strengths, with selection depending on the nature of the data and analysis goals. All approaches ultimately help in revealing the true structure of the series for better forecasting and modeling.

XV. References

References

[1] Peixeiro, Marco. Time Series Forecasting in Python. Simon and Schuster, 2022.
[2] Hyndman, R.J., Athanasopoulos, G. (2018). Forecasting: Principles and Practice.
[3] Hamilton, J.D. (1994). Time Series Analysis. Princeton University Press.