INFORMATICS INSTITUTE OF TECHNOLOGY

In Collaboration with

ROBERT GORDON UNIVERSITY ABERDEEN

Artificial Intelligence and Data Science

Module Leader: Mohamed Ayoob

CM2604 - Data Engineering

Assignment Type: Individual Coursework

Student Name: Vanuja Thihansith Sooriyaarachchi

IIT ID - 20222408

RGU ID - 2311130

Table of Contents

List of Figures

Introduction

This coursework provides a detailed analysis of formaldehyde (HCHO) concentrations in major cities across Sri Lanka, aiming to better understand air quality and its implications for public health and environmental health. The study focuses on identifying spatial and temporal variations in HCHO levels and explores potential emission sources.

The findings from this coursework aim to contribute valuable insights into air quality management, supporting efforts towards sustainable urban planning and public health improvement in Sri Lanka. By understanding and predicting HCHO levels, policymakers and environmental agencies can better implement effective pollution control strategies, thus enhancing the quality of life and environmental health in the region.

Dataset:

Link:
https://drive.google.com/drive/folders/1xzQ5pIEnaUN2DOyZTqYSJrxFMC8Unx73?usp=sharing

The provided dataset contains daily HCHO measurements from the Sentinel-5P satellite from the European Space Agency (ESA) for seven cities in Sri Lanka. The historical data is available from the 2019/1/1 to 2023/12/31 (YYYY-MM-DD). The cities are: 'Colombo Proper', 'Deniyaya, Matara', 'Nuwara Eliya Proper', 'Bibile, Monaragala', 'Kurunegala Proper', 'Jaffna Proper', and 'Kandy Proper'.

The project is managed using Git, fostering transparency and collaboration throughout the development phase. The source code is openly accessible to the public on GitHub at https://github.com/VanujaSooriyaarachchi/HCHO_Prediction.

Power BI Dashboard:

HCHO Level Dashboard

Future Forecasting Dashboard

1. Data Preprocessing

1.1 Clean and Prepare the Data

1.1.1 Load the Data

Load the col_mat_nuw_output.csv file.

```
data = spark.read.csv("D:/IIT/2 nd Year/2nd Sem/Data Engineering/Course Work/HCHO_Prediction/dataset/col_mat_nuw_output.csv", header=True, inferSchema=True)
```

Load the kan_output.csv file.

```
data_2 = spark.read.csv("D:/IIT/2 nd Year/2nd Sem/Data Engineering/Course Work/HCHO_Prediction/dataset/kan_output.csv", header=True, inferSchema=True)
```

Load the mon_kur_jaf_output.csv file.

```
data_3 = spark.read.csv("D:/IIT/2 nd Year/2nd Sem/Data Engineering/Course Work/HCHO_Prediction/dataset/mon_kur_jaf_output.csv", header=True, inferSchema=True)
```

1.1.2 Explore Descriptive Statistics

Summarize (mean, median, standard deviation) HCHO levels for each city on and across the entire dataset.

Describe the 'HCHO reading' column

```
data.select('HCHO reading').describe().show()
```

```
+-------+--------------------+
|summary|        HCHO reading|
+-------+--------------------+
|  count|                3058|
|   mean|1.200178195763001...|
| stddev|1.009287188756533...|
|    min|-2.59296176552668...|
|    max|8.997101837438971E-4|
+-------+--------------------+

+-------+--------------------+
|summary|        HCHO reading|
+-------+--------------------+
|  count|                1825|
|   mean|9.890951713730535E-5|
| stddev|9.651844491820422E-5|
|    min|-2.99702863135199...|
|    max|7.051621763962024E-4|
+-------+--------------------+

+-------+--------------------+
|summary|        HCHO reading|
+-------+--------------------+
|  count|                5477|
|   mean|1.192770268341103...|
| stddev|8.860002918894764E-5|
```

```
|    min|-3.52473024357239...|
|    max|5.837611392919413E-4|
+-------+--------------------+
```

1.1.3 Handling Missing Values

Check for null values in the DataFrame

This is obtained through the code below.

```
data.select([count(when(col(c).isNull(), c)).alias(c) for c in data.columns]).show()
```

Dataset 1

```
+------------+--------+------------+---------+
|HCHO reading|Location|Current Date|Next Date|
+------------+--------+------------+---------+
|        2419|       0|           0|        0|
+------------+--------+------------+---------+
```

Dataset 2

```
+------------+--------+------------+---------+
|HCHO reading|Location|Current Date|Next Date|
+------------+--------+------------+---------+
|         793|       0|           0|        0|
+------------+--------+------------+---------+
```

Dataset 3

```
+------------+--------+------------+---------+
|HCHO reading|Location|Current Date|Next Date|
+------------+--------+------------+---------+
|        1651|       0|           0|        0|
+------------+--------+------------+---------+
```

After removing null values

```
+------------+--------+------------+---------+
|HCHO Reading|Location|Current Date|Next Date|
+------------+--------+------------+---------+
|           0|       0|           0|        0|
+------------+--------+------------+---------+
```

1.1.4 Handling Duplicates

Calculate the length of DataFrame.

```
# Count the number of rows in the DataFrame
data_count = data.count()
# Show the length of the DataFrame
print("Length of DataFrame:", data_count)
```

```
Length of DataFrame: 5477
```

Drop duplicate values.

```
# Drop duplicates from the DataFrame
data_no_duplicates = data.dropDuplicates()
```

Calculate the length of DataFrame after dropping duplicate values.

```
# Count the number of rows in the DataFrame
data_count = data.count()
# Show the length of the DataFrame
print("Length of DataFrame:", data_count)
```

```
Length of DataFrame: 5477
```

Length of DataFrame(Before drop duplicates) = Length of DataFrame(After drop duplicates)

So, there are no duplicate values in each of the data sets.

1.1.4.1 Handling Outliers

Distribution of HCHO reading of first dataset (col_mat_nuw_output.csv)

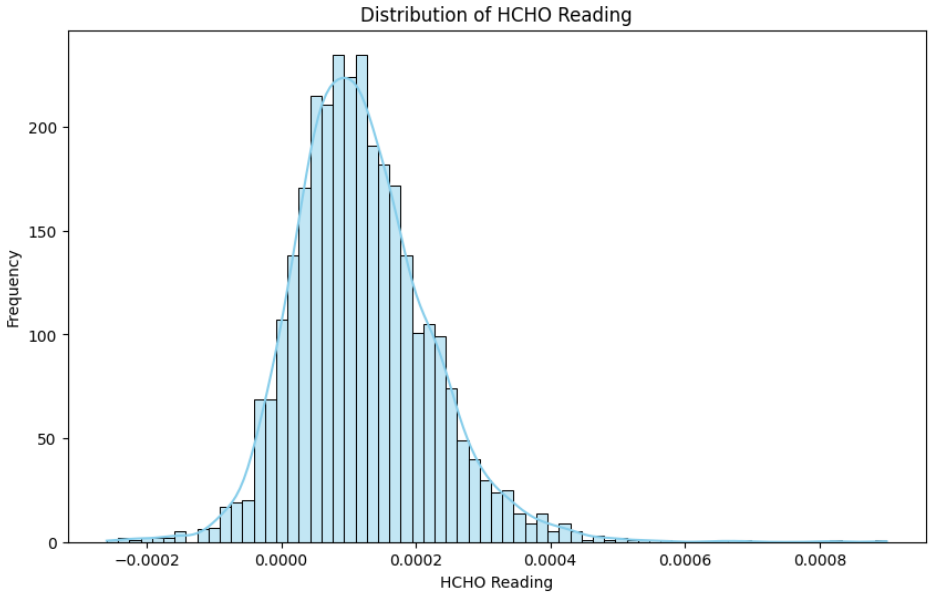


Figure 1: Distribution of HCHO reading of first dataset.

Distribution of HCHO reading of first dataset after handling outliers

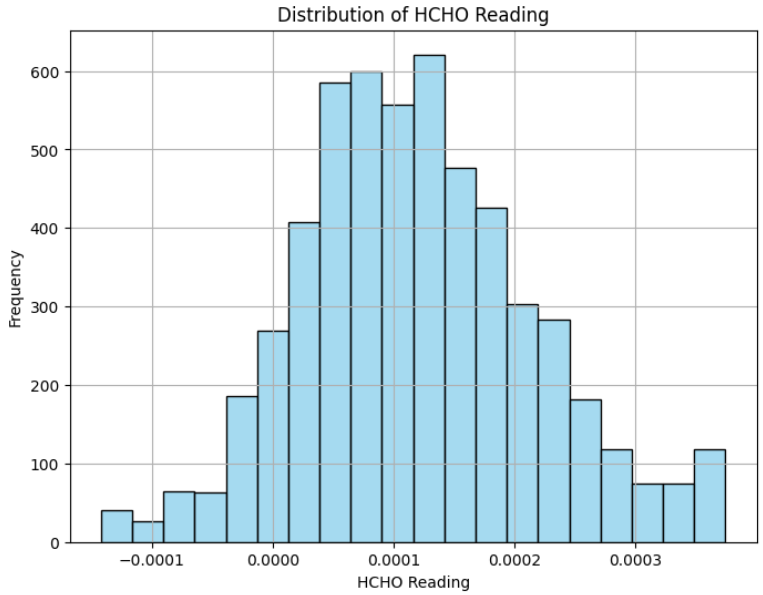


Figure 2: Distribution of HCHO reading of first dataset after handling outliers.

Distribution of HCHO reading of second dataset (kan_output.csv)

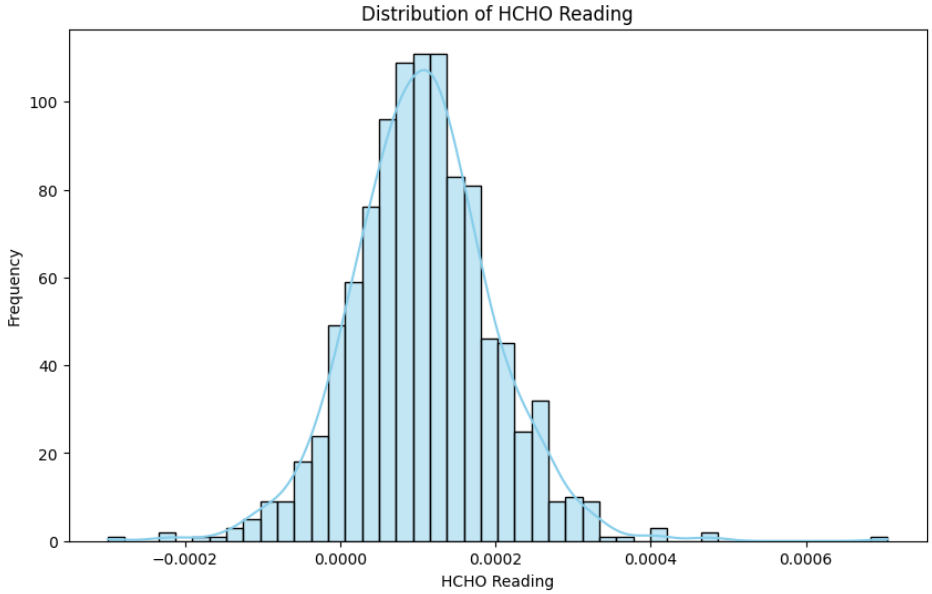


Figure 3: Distribution of HCHO reading of second dataset.

Distribution of HCHO reading of second dataset after handling outliers

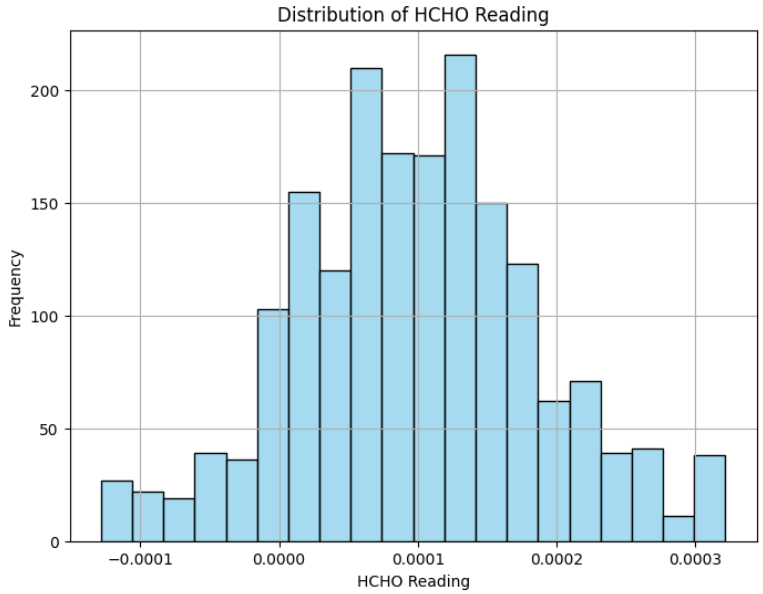


Figure 4: Distribution of HCHO reading of second dataset after handling outliers.

Distribution of HCHO reading of third dataset (mon_kur_jaf_output.csv)

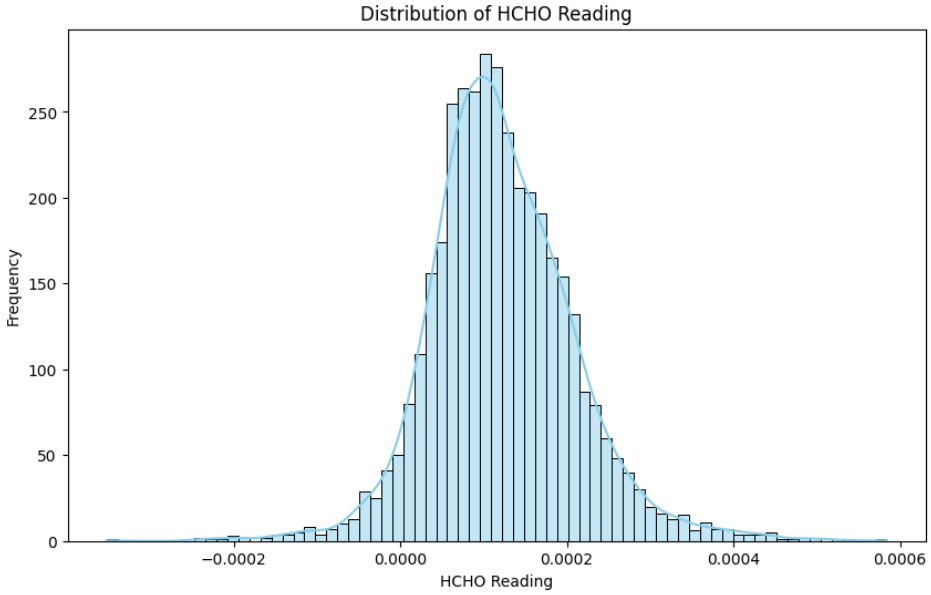


Figure 5: Distribution of HCHO reading of third dataset.

Distribution of HCHO reading of third dataset after handling outliers

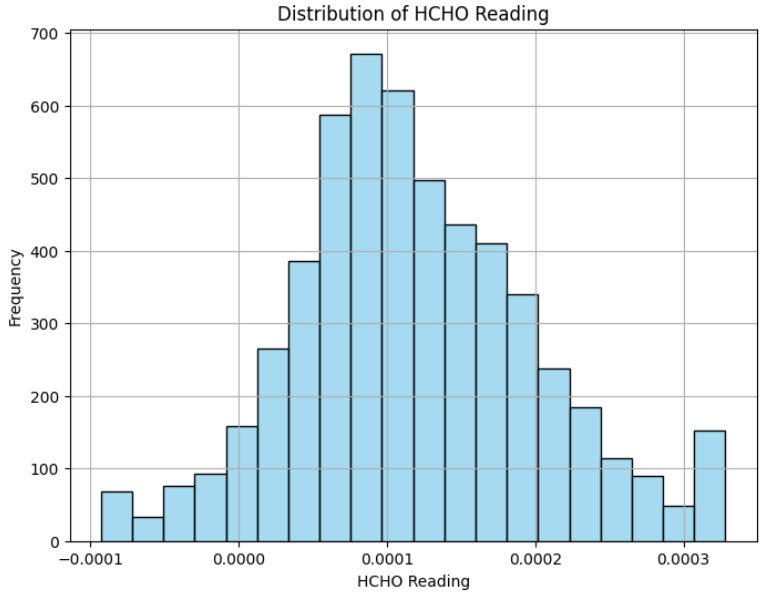


Figure 6: Distribution of HCHO reading of third dataset after handling outliers.

Boxplot of HCHO reading of first dataset (col_mat_nuw_output.csv)

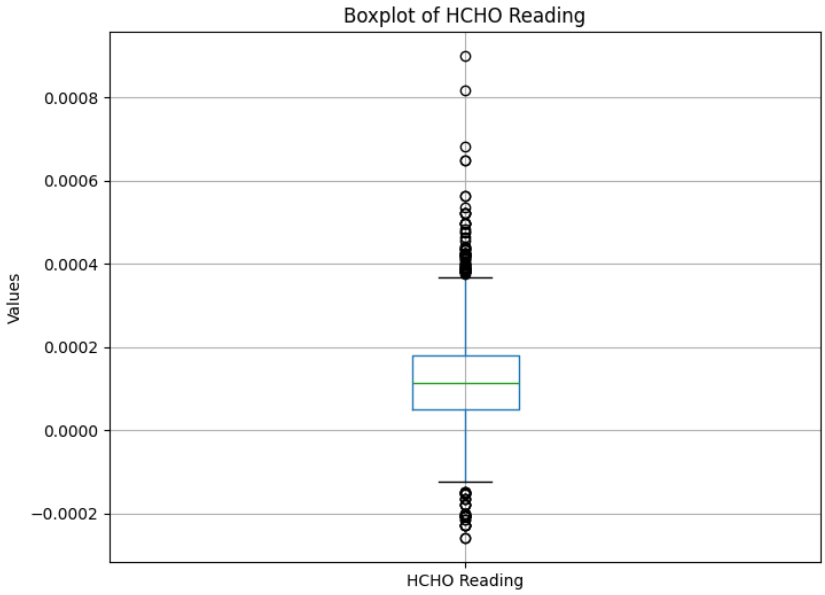


Figure 7: Boxplot of HCHO reading of first dataset.

Boxplot of HCHO reading of first dataset after handling outliers

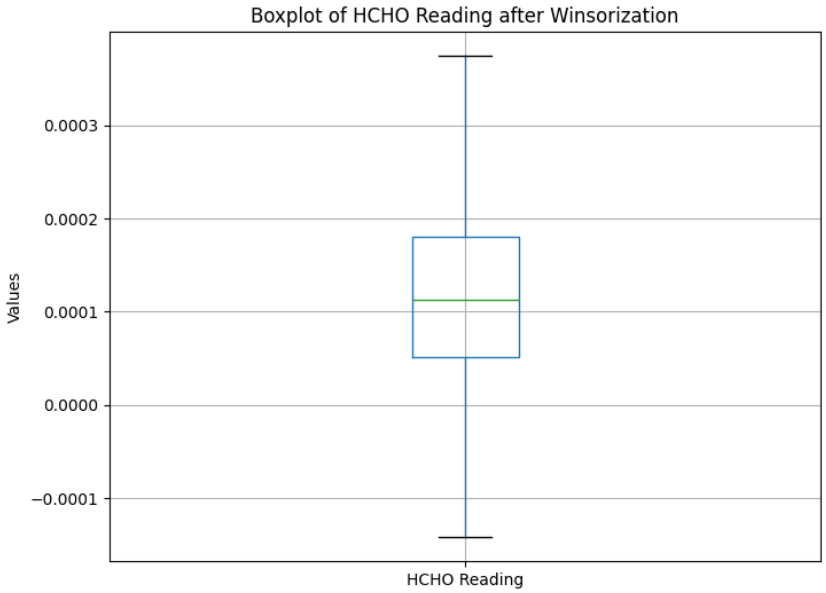


Figure 8: Boxplot of HCHO reading of first dataset after handling outliers.

Boxplot of HCHO reading of second dataset (kan_output.csv)

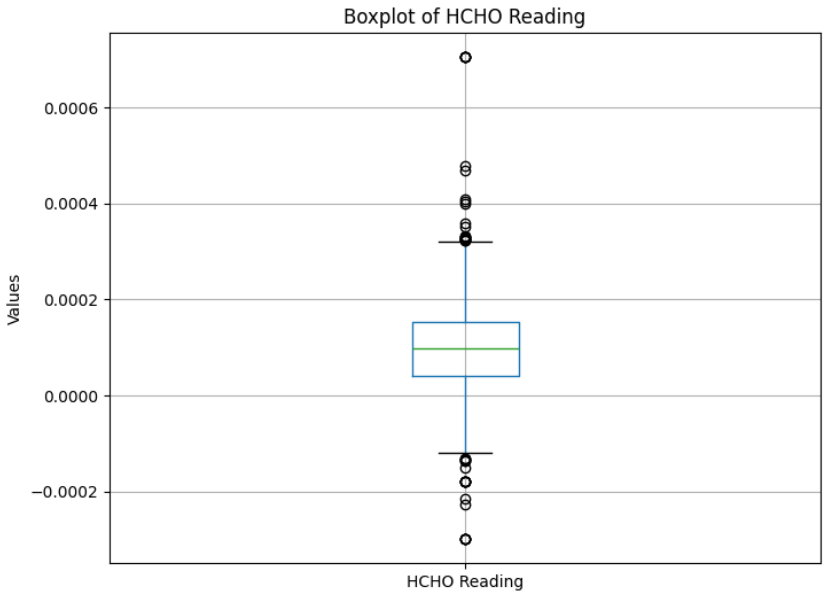


Figure 9: Boxplot of HCHO reading of second dataset.

Boxplot of HCHO reading of second dataset after handling outliers

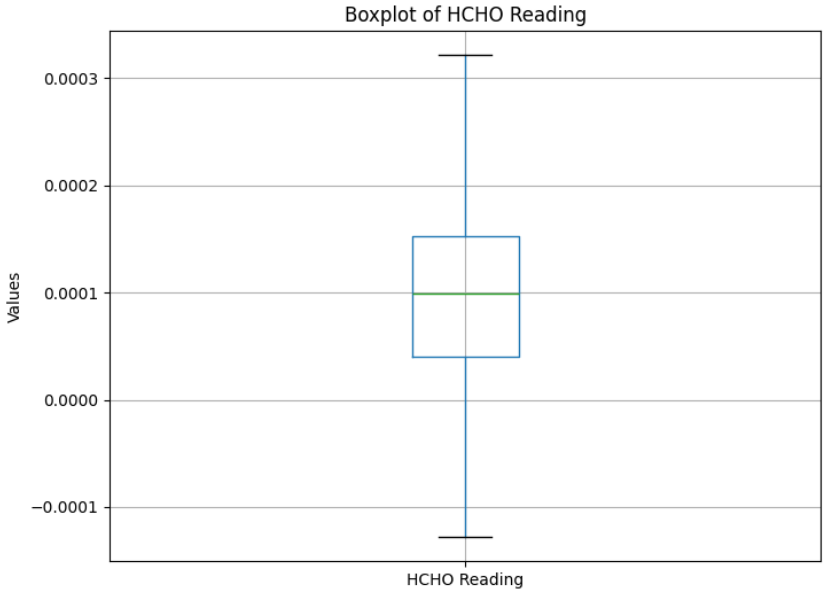


Figure 10: Boxplot of HCHO reading of second dataset after handling outliers.

Boxplot of HCHO reading of third dataset (mon_kur_jaf_output.csv)

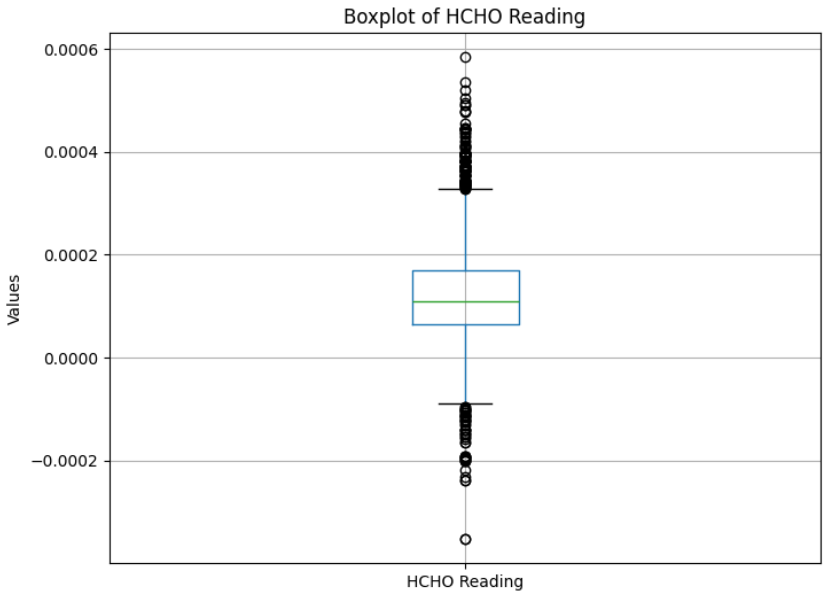


Figure 11: Boxplot of HCHO reading of third dataset.

Boxplot of HCHO reading of third dataset after handling outliers

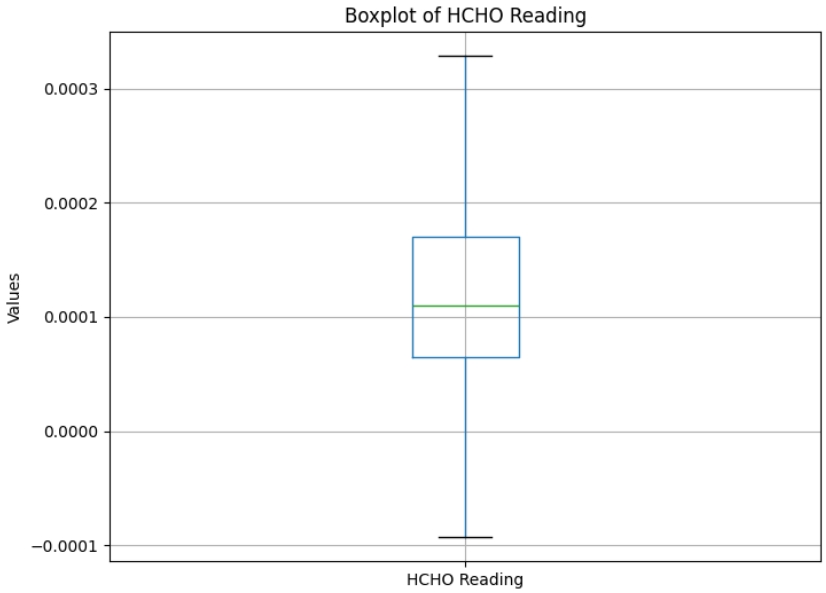


Figure 12: Boxplot of HCHO reading of third dataset after handling outliers.

2. Spatio-Temporal Analysis

2.1 Analyze Trends Over Time

2.1.1 Seasonal Variations

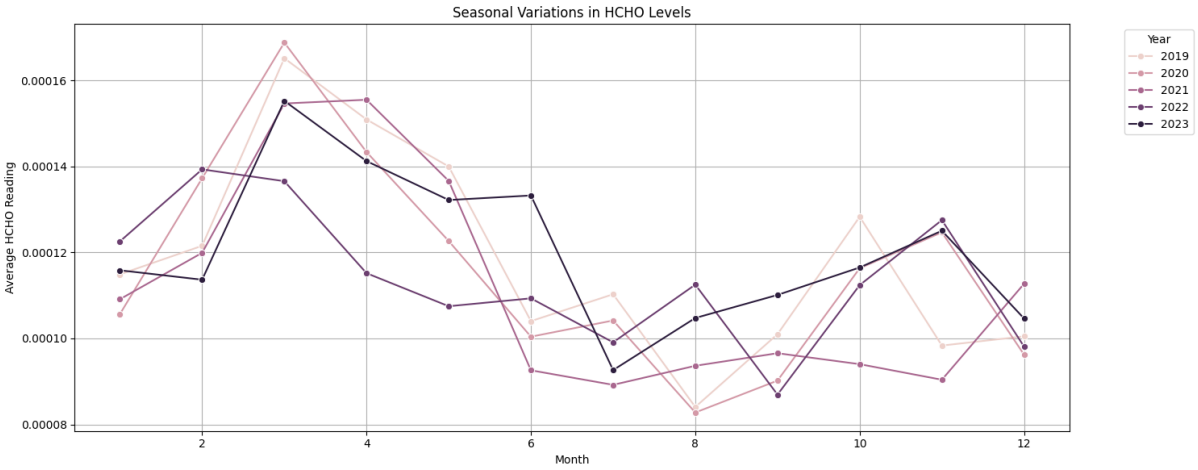


Figure 13:Seasonal Variations in HCHO Level

The graph illustrating the seasonal variations in HCHO (formaldehyde) levels over several years, from 2019 to 2023, offers a visual representation of how these levels fluctuate over time. Each line on the graph represents a different year, helping to identify trends and changes across the period.

Seasonal Trends:

The graph suggests a recurring annual pattern in HCHO levels, which may be influenced by seasonal factors such as temperature fluctuations, vegetation cycles, or human activities like heating and agriculture. These seasonal impacts are typical as they affect the photochemical processes and emission sources contributing to HCHO levels.

Annual Comparison:

2019: This year shows fluctuating levels with a significant peak in the middle of the year, likely corresponding to the summer months when increased sunlight enhances photochemical reactions, leading to higher HCHO production.

2020: Starting higher than 2019, the levels see a notable dip, possibly reflecting the decreased human and industrial activities during the COVID-19 lockdowns, affecting emissions and atmospheric conditions.

2021: This year continues the trend of variability observed in 2020, with peaks not reaching the heights of 2019, possibly indicating the prolonged impact of the pandemic on activities contributing to HCHO levels.

2022: Shows continued variability with a general reduction in the highest levels compared to 2019, suggesting either effective regulatory measures or lasting changes in emission sources.

2023: With only partial data available, the early months of 2023 indicate lower levels than previous years, which might suggest ongoing adaptations in industrial practices or environmental policies.

Interannual Variability:

The year-to-year changes highlight that, aside from seasonal effects, factors such as policy changes, economic conditions, or long-term environmental shifts are likely influencing HCHO levels. This variability underscores the importance of continuous monitoring and analysis to adapt and fine-tune environmental and public health policies.

Peaks and Valleys:

Consistently, all years exhibit mid-year peaks possibly due to increased vegetation activity and industrial operations during warmer months. The valleys typically occur at the beginning and end of the year, which may be influenced by lower temperatures or different atmospheric conditions that reduce the formation or dispersion of HCHO.

Implications for Policy and Research:

The data from this graph is crucial for understanding the dynamics of air quality concerning formaldehyde levels. It aids policymakers, researchers, and the public in identifying the specific factors that may influence these levels, such as seasonal changes, policy effectiveness, and the impact of global events like the pandemic. This understanding is vital for developing targeted strategies to manage and reduce HCHO emissions effectively, ultimately improving air quality and public health outcomes.

2.1.2 Long-term Changes

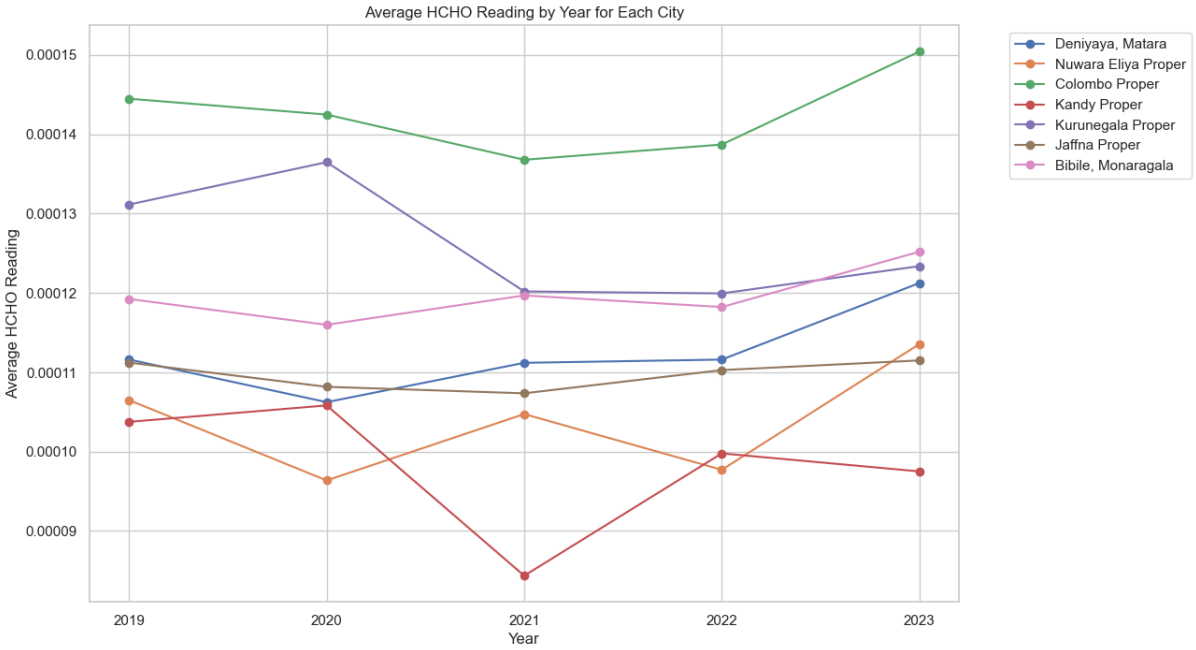


Figure 14: Average HCHO Reading by Year for Each City

The graph presenting average HCHO readings across seven different cities from 2019 to 2023 offers insightful data on air quality trends and their temporal dynamics. Each city, represented by a unique color on the graph, reveals specific patterns in HCHO concentrations over the years, allowing for both inter-city comparisons and the assessment of temporal trends.

Inter-City Comparison:

Dambulla, Matale: Shows a clear upward trend in HCHO readings over the five-year period, suggesting increasing formaldehyde levels which could be indicative of growing industrial or vehicular activity.

Colombo Proper: Indicates a decline from 2019 to 2020 with subsequent stabilization, possibly reflecting effective regulatory measures or changes in urban activities.

Nuwara Eliya Proper: Exhibits a decrease followed by an increase and another dip, indicating fluctuating HCHO levels that may be influenced by seasonal tourism activities and agricultural practices.

Kandy Proper: Displays minor fluctuations but maintains relative stability, suggesting consistent ambient conditions or effective air quality management.

Kurunegala Proper: Shows an initial increase followed by a decrease, reflecting possible variations in local industrial activities or changes in traffic patterns.

Jaffna Proper: This line shows a significant decrease initially, slight recovery, and then a steady decline, which could be related to specific local policies or economic factors impacting emissions.

Bibile, Monaragala: Demonstrates the most dramatic decrease initially, with a partial recovery followed by another decline, suggesting significant impacts from local environmental or policy changes.

Temporal Trends:

General Decrease in 2020: Most cities experienced a decline in HCHO readings in 2020, likely correlated with the global slowdown in industrial activities and transportation during the COVID-19 pandemic.

Varying Recovery Patterns: Post-2020, cities displayed different recovery patterns, with some returning to pre-pandemic levels while others remained at lower levels, indicating diverse local responses or adaptations.

Yearly Variability:

2019 to 2020: A widespread decrease across cities, reflecting the immediate impact of pandemic-related restrictions.

2020 to 2021: Shows mixed trends with some cities stabilizing or recovering, while others continued to decline, underscoring varied local conditions or measures.

2021 to 2022: Most cities either stabilized or experienced declines, suggesting a possible adaptation to new normal in urban activities and emissions.

2022 to 2023: Early data for 2023 suggests some cities may be experiencing rises in HCHO levels, possibly due to resumed activities or less stringent regulations.

City-Specific Observations:

Dambulla, Matale, and Nuwara Eliya Proper: Exhibit more significant year-on-year variability, possibly due to local factors such as seasonal agricultural burns, tourism, or changes in local industries.

Bibile, Monaragala: The sharp decrease and subsequent partial recovery could be tied to specific local events or interventions, warranting closer investigation to understand the causes and implications of such fluctuations.

This analysis is crucial for urban planning and environmental policymaking, as it helps identify cities with worsening air quality and those showing improvements. It also assists in understanding the impact of global and local events, like the COVID-19 pandemic, on urban air pollution.

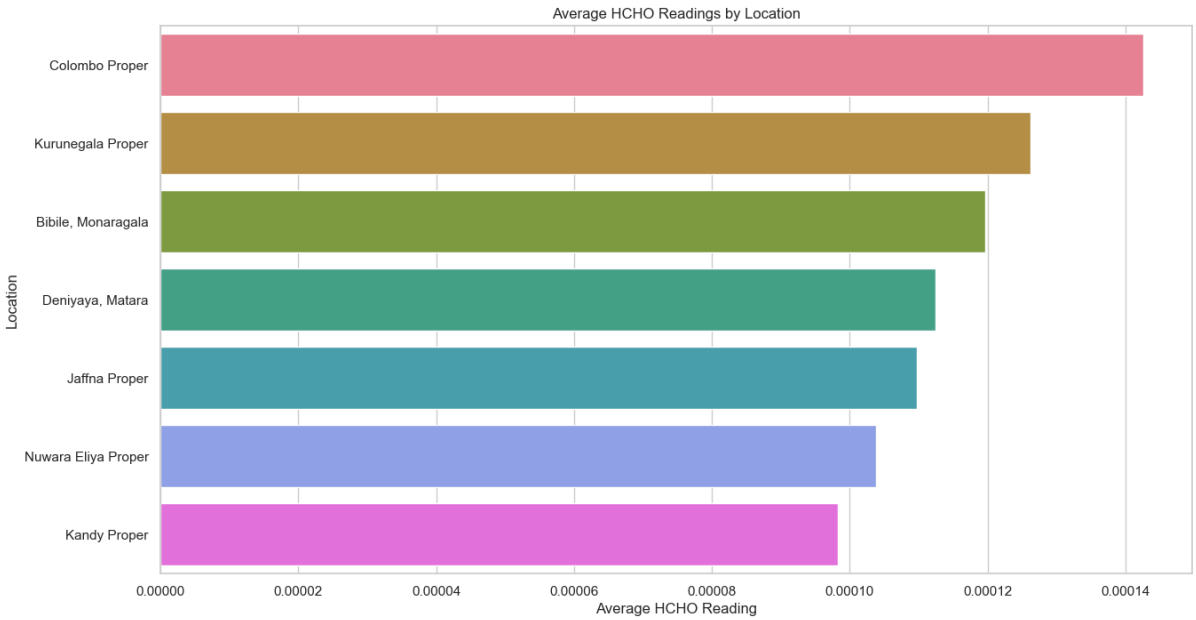


Figure 15: Average HCHO Readings by Location

2.1.3 Trends Across Cities

2.1.3.1 Bibile, Monaragala

Figure 16:Bibile, Monaragala - Trend, Seasonal, Residual Component

The decomposition of time series data into trend, seasonality, and residuals is a crucial method for understanding the underlying patterns and irregularities in environmental data, such as HCHO levels in Bibile, Monaragala.

Trend Component:

This component reflects the long-term progression or changes in the data, highlighting overall upward or downward movements over time.

If this graph shows a consistent upward trend, it might suggest an increase in HCHO levels, possibly due to escalating industrial activity or changes in environmental regulations. Conversely, a downward trend could indicate effective pollution control measures or a decrease in activities that produce HCHO.

Seasonal Component:

This graph captures the regular patterns or fluctuations that recur at consistent intervals within the data, typically aligned with seasonal cycles.

Regular spikes might correspond to specific seasons, reflecting seasonal influences such as temperature variations, agricultural burning, or seasonal industrial outputs that affect HCHO levels. Understanding these patterns helps predict times of higher or lower air quality.

Residual Component:

The residuals represent what remains in the data after accounting for the trend and seasonality. This component should ideally show randomness or 'noise' that is not explained by the other components.

If the residuals are centered around zero with no clear pattern, the model is likely to perform well. However, significant spikes or patterns in the residuals suggest that there are other factors influencing the data that the model has not captured, indicating a potential area for model refinement.

2.1.3.2 Colombo

Figure 17: Colombo - Trend, Seasonal, Residual Component

Trend Component: This plot shows the long-term movement in the HCHO data, smoothing out the short-term fluctuations to highlight the underlying trend in the dataset. It appears to have some periodicity, with peaks and troughs indicating possible cyclical behavior.

Seasonal Component: This graph illustrates the seasonal fluctuations in the HCHO levels. It shows a clear pattern that repeats over time. This might correspond to regular seasonal changes in the environment or human activity that affect HCHO levels.

Residual Component: This displays the residuals of the model that is, the difference between the actual data and the combined trend and seasonal components. Ideally, the residuals should

be random and centered around zero, suggesting that the model has captured all the significant information in the data. In this case, the residuals show some structure, indicating potential room for model improvement.

2.1.3.3 Deniyaya, Matara

Figure 18: Deniyaya, Matara - Trend, Seasonal, Residual Component

The analysis of the time series decomposition of HCHO readings in Deniyaya, Matara, like the previously discussed decomposition for Colombo Proper, provides a detailed look at how HCHO levels are influenced over time through trend, seasonality, and residuals.

Trend Component: The trend component in Deniyaya, Matara shows a generally increasing progression with periodic fluctuations. This could indicate an overall rise in HCHO levels over time, potentially driven by urban expansion, escalation in industrial activities, or other environmental changes.

Seasonal Component: There's a clear, repetitive seasonal pattern in the HCHO levels, suggesting significant seasonal impacts. Factors such as agricultural practices, weather conditions, and seasonal industrial or domestic activities might heavily influence these variations.

Residual Component: The residuals, which represent what the trend and seasonal adjustments haven't accounted for, should ideally look like white noise random with no pattern and centered around zero. The presence of some variability in the residuals suggests that the model might

be missing some explanatory variables or that there are non-linear interactions not captured by the current modeling approach.

Residual analysis can point to areas where the model may be improved by incorporating additional data or refining the model structure.

2.1.3.4 Jaffna

Figure 19: Jaffna - Trend, Seasonal, Residual Component

The time series decomposition of HCHO readings in Jaffna Proper provides a structured analysis into the dynamics affecting formaldehyde levels in the area. This breakdown into trend, seasonal, and residual components help identify underlying patterns and anomalies, offering valuable insights for environmental monitoring and policymaking.

Trend Component: The trend in Jaffna Proper shows slight fluctuations without a clear directional increase or decrease, suggesting a relatively stable HCHO level over time with intermittent periods of rise and fall.

This stability might indicate that while there are periodic increases or decreases in HCHO emissions, they tend to return to a baseline level. This could be due to effective regulation and steady state of industrial and vehicular emissions in the area.

Seasonal Component: The seasonal fluctuations are quite pronounced, indicating significant and consistent seasonal influences on HCHO levels.

These patterns likely correlate with specific local activities such as agricultural cycles that involve burning or pesticide use, seasonal variations in traffic, or changes in industrial production schedules. Recognizing these patterns helps predict times of higher pollution levels, crucial for health advisories and environmental management.

Residual Component: The residuals exhibit some volatility, suggesting the presence of irregular influences on HCHO levels.

This indicates that there are additional factors or random events impacting HCHO levels that are not accounted for by the seasonal or trend components. These could include unexpected industrial emissions, environmental incidents like fires, or shifts in regulatory enforcement.

2.1.3.5 Kandy

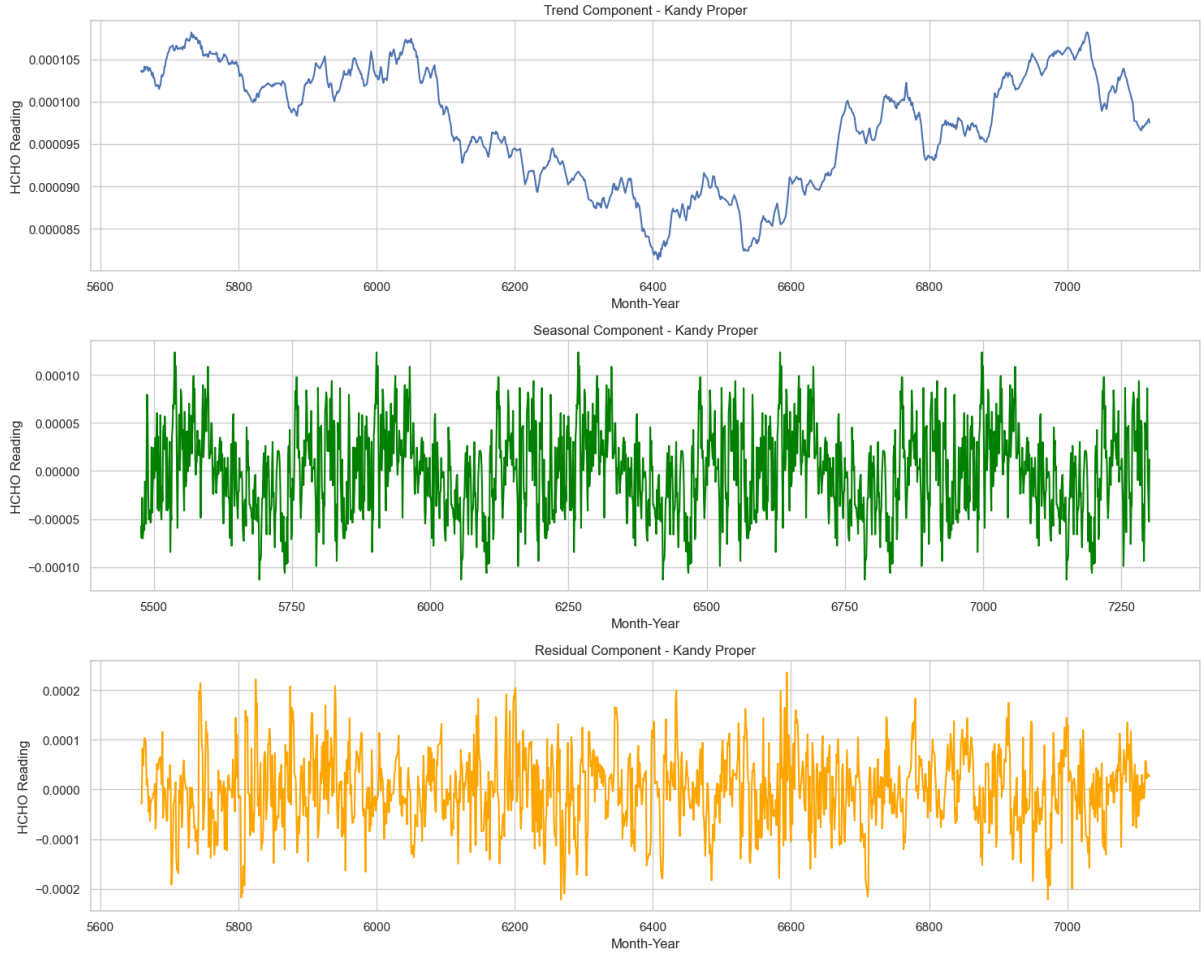


Figure 20: Kandy - Trend, Seasonal, Residual Component

The time series decomposition for Kandy Proper, which separates data into trend, seasonal, and residual components, offers crucial insights into the dynamics of HCHO levels in the area. Understanding these components helps in identifying patterns and irregularities that are essential for developing effective environmental policies and forecasting models.

Trend Component: The trend component in Kandy Proper displays an undulating pattern, characterized by fluctuations without a distinct long-term upward or downward trend.

This suggests that while HCHO levels in Kandy Proper vary over time, there is no consistent increase or decrease. These fluctuations might reflect transient changes in environmental conditions or emissions without long-term impact on the baseline HCHO levels.

Seasonal Component: There is a clear seasonal pattern, indicating regular fluctuations at consistent intervals.

Such regularity likely stems from seasonal variations in activities that affect HCHO emissions, such as agricultural burning during harvest seasons, increased industrial activity at certain times of the year, or seasonal variations in traffic flow. Identifying these patterns is vital for predicting periods of high or low air quality, which can inform seasonal public health advisories and environmental regulations.

Residual Component: The residuals show considerable variability, suggesting influences on HCHO levels that are not captured by the seasonal or trend components.

This variability could be due to random or unexpected events like industrial accidents, fires, or sudden changes in local regulations. The presence of significant residuals indicates that while the model accounts for much of the variability in HCHO levels, there are still factors at play that it does not capture.

2.1.3.6 Kurunegala

Figure 21: Kurunegala - Trend, Seasonal, Residual Component

The time series decomposition for Kurunegala Proper, broken down into trend, seasonal, and residual components, provides an insightful view into the dynamics of HCHO levels in this specific area. This analysis helps to identify long-term movements, cyclical patterns, and unexplained variations, which are crucial for effective environmental management and policy formulation.

Trend Component: The trend component in Kurunegala Proper indicates a slight downward trajectory, suggesting a gradual decrease in HCHO levels over the observed period.

This decreasing trend might be indicative of successful environmental policies, improved industrial practices, or a reduction in activities that produce HCHO emissions. It could also reflect broader environmental shifts that reduce pollutant levels.

Seasonal Component: There is clear seasonality in the HCHO levels, with consistent patterns of peaks and troughs occurring at regular intervals throughout the year.

This component suggests that HCHO levels in Kurunegala Proper are significantly influenced by seasonal factors, which could include agricultural burning cycles, seasonal industrial outputs, or climatic conditions that affect how HCHO is formed or dispersed in the atmosphere.

Residual Component: The residuals show irregularity, capturing fluctuations in the data that are not explained by the trend or seasonal components.

These irregularities might be due to random environmental events such as unexpected industrial emissions, accidental releases, or measurement errors. The presence of significant residuals indicates that there are additional factors influencing HCHO levels that the model has not accounted for.

2.1.3.7 Nuwara Eliya

Figure 22: Nuwara Eliya - Trend, Seasonal, Residual Component

The time series decomposition for Nuwara Eliya Proper, like those of other cities, offers crucial insights into the patterns and influences on HCHO levels through its trend, seasonal, and residual components.

Trend Component: The trend component shows a gradual upward movement in HCHO levels over the period analyzed.

This upward trend suggests a long-term increase in activities or environmental changes contributing to higher HCHO levels. Factors could include increased industrial activity, changes in land use, or growing vehicular traffic.

Seasonal Component: There is clear seasonality in HCHO levels, with regular peaks and troughs.

The seasonal fluctuations could be linked to specific local activities. Understanding these patterns allows for predictive measures and strategic planning to manage peak HCHO periods effectively.

Residual Component: The residuals show some volatility, indicating that there are still irregular influences on HCHO levels after accounting for trend and seasonality.

2.2 Changes in Gas Emissions due to the COVID-19 Lockdowns

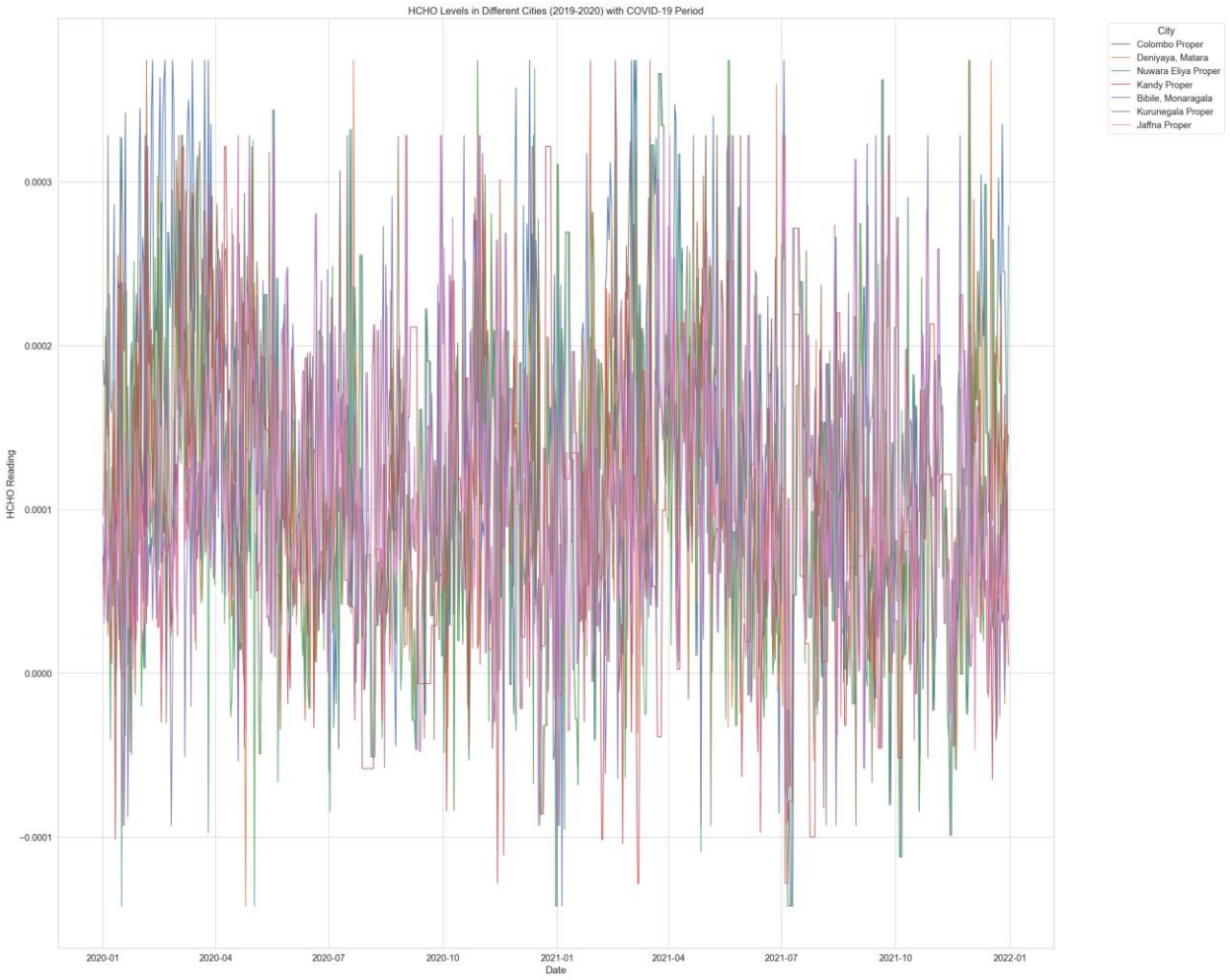


Figure 23: HCHO Levels in Different Cites with Covid-19 Period

This graph presents HCHO (formaldehyde) levels in all the cities from 2019-2020, including a highlighted section that likely indicates the COVID-19 lockdown period.

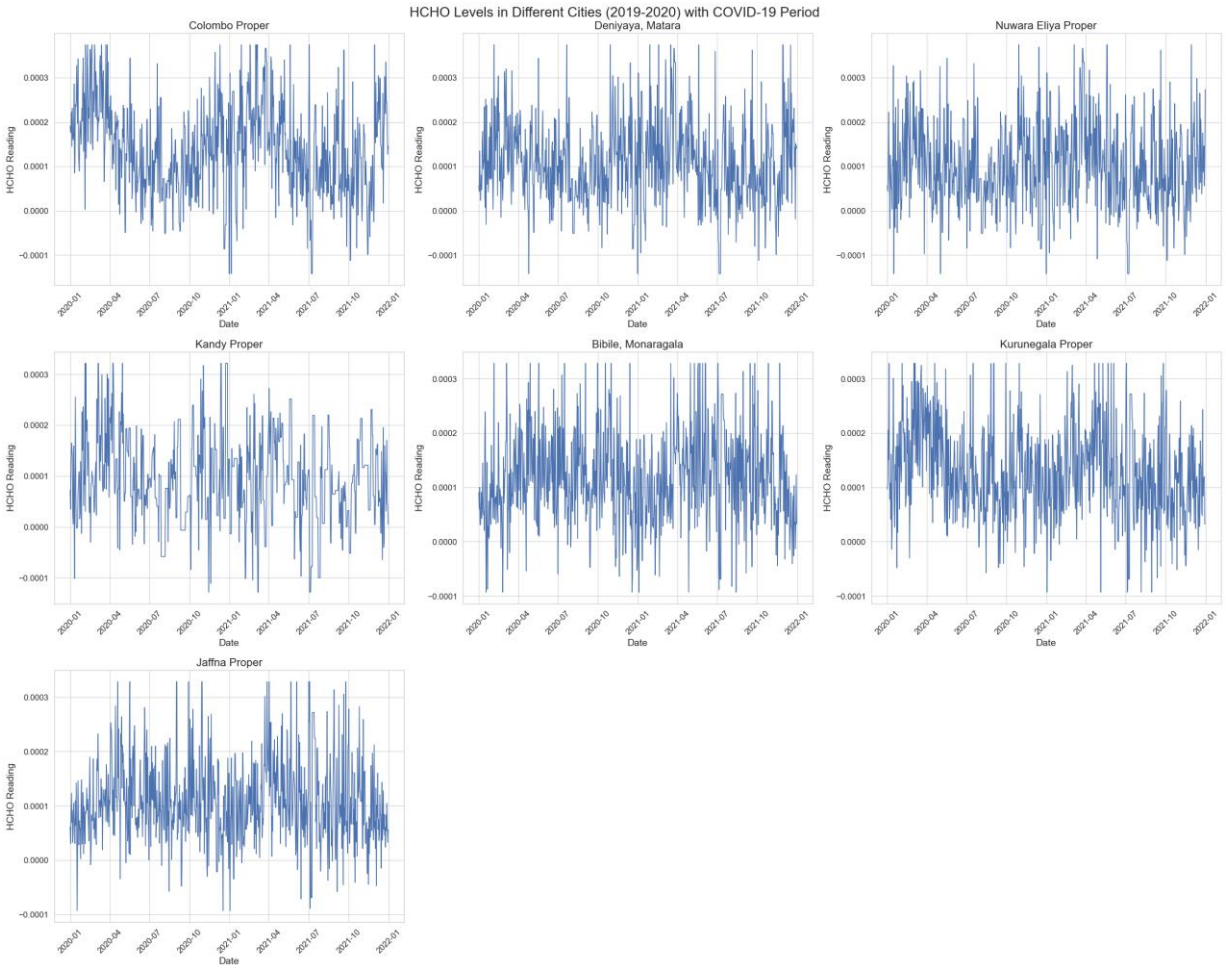


Figure 24: HCHO Levels in Different Cites with Covid-19 Period (One graph for each city)

This graph presents HCHO (formaldehyde) levels in various cities from 2019-2020, including a highlighted section that likely indicates the COVID-19 lockdown period.

Colombo Proper: Shows significant variability with some periods of elevated HCHO levels. During the lockdown, there is a visible reduction in the upper range of the fluctuations, suggesting a decrease in HCHO emissions during that period.

Deniyaya, Matara: Also exhibits variability, but with fewer extreme spikes compared to Colombo. The lockdown period does not show a substantial change in pattern.

Nuwara Eliya Proper: Displays a relatively consistent range of variability. Like Deniyaya, the lockdown doesn't appear to have a pronounced effect on the HCHO levels.

Kandy Proper: The HCHO levels vary widely over time. The lockdown period doesn't demonstrate a significant deviation from the overall variability pattern.

Bibile, Monaragala: The data show a high degree of fluctuation, and it seems the lockdown period corresponds with a section of reduced variability, although not as marked as in Colombo.

Kurunegala Proper: The fluctuations are less extreme than in Colombo but still show a wide range of variability. The lockdown period here also does not indicate a clear shift in the pattern of HCHO levels.

Jaffna Proper: The variability is consistent with no evident change during the lockdown period.

Comparative Analysis:

Volatility: All cities show significant fluctuations in HCHO levels. Colombo stands out with higher peaks, suggesting periods of elevated HCHO concentration, possibly due to more industrial activities or traffic.

Lockdown Impact: The impact of the lockdown on HCHO levels is inconsistent across the cities. Colombo shows a reduction in variability, which might indicate an environmental response to reduced human and industrial activity.

Baseline Levels: Cities like Jaffna, Kandy, and Kurunegala exhibit a stable pattern throughout, with no apparent impact from the lockdown. This could imply that sources of HCHO in these areas were less affected by lockdown measures, or other factors may be influencing the levels.

Data Consistency: Bibile, Monaragala has a wide range but less pronounced spikes, suggesting different sources or dynamics of HCHO emissions.

2.3 External Factors

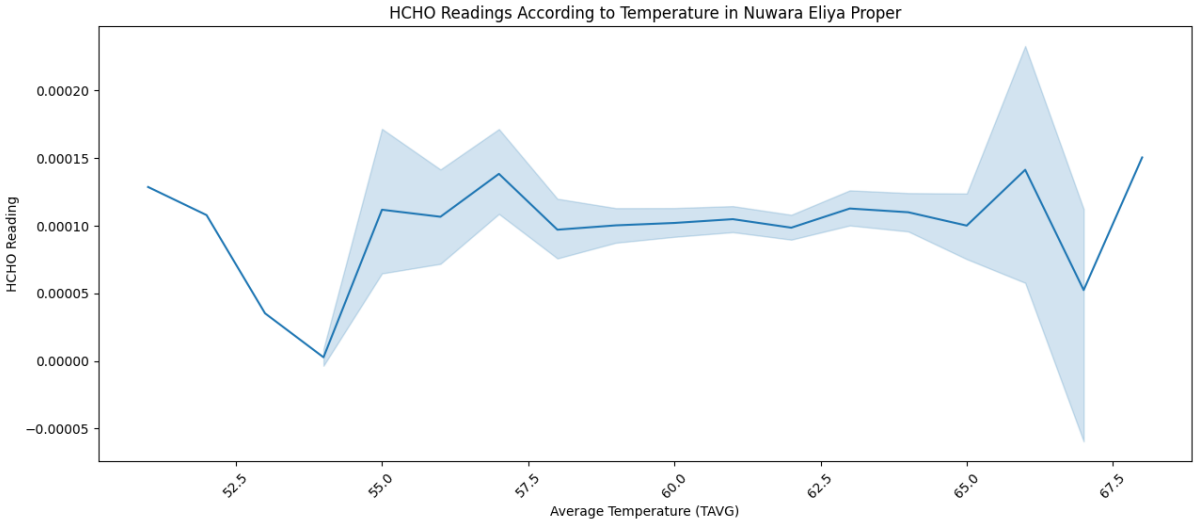


Figure 25: HCHO Readings According to Temp in Nuwara Eliya

This chart shows how the HCHO readings change as the average temperature increases or decreases in Nuwara Eliya.

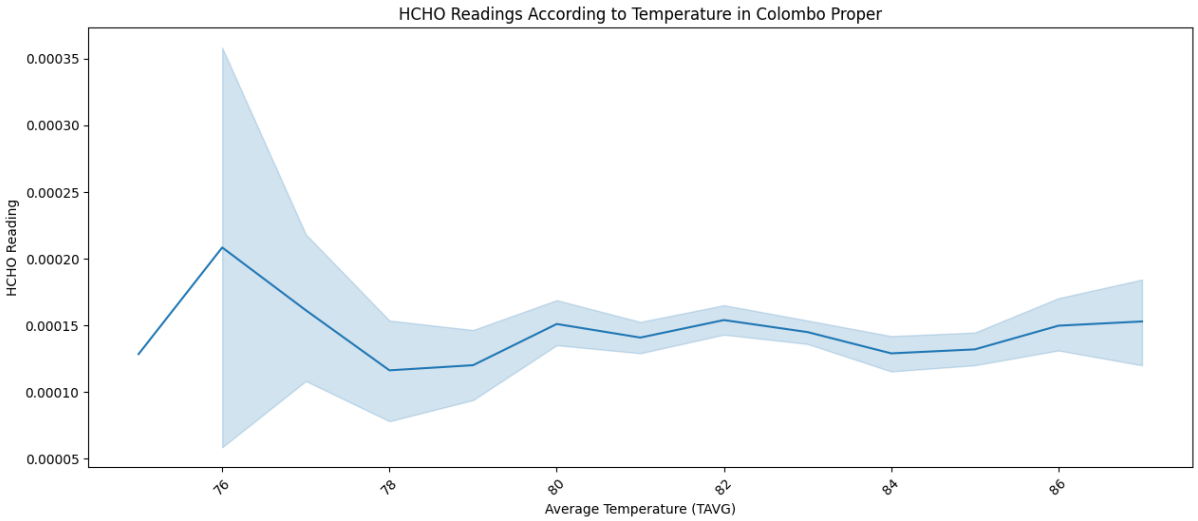


Figure 26: HCHO Reading According to Temp in Colombo

This chart shows how the HCHO readings change as the average temperature increases or decreases in Colombo.

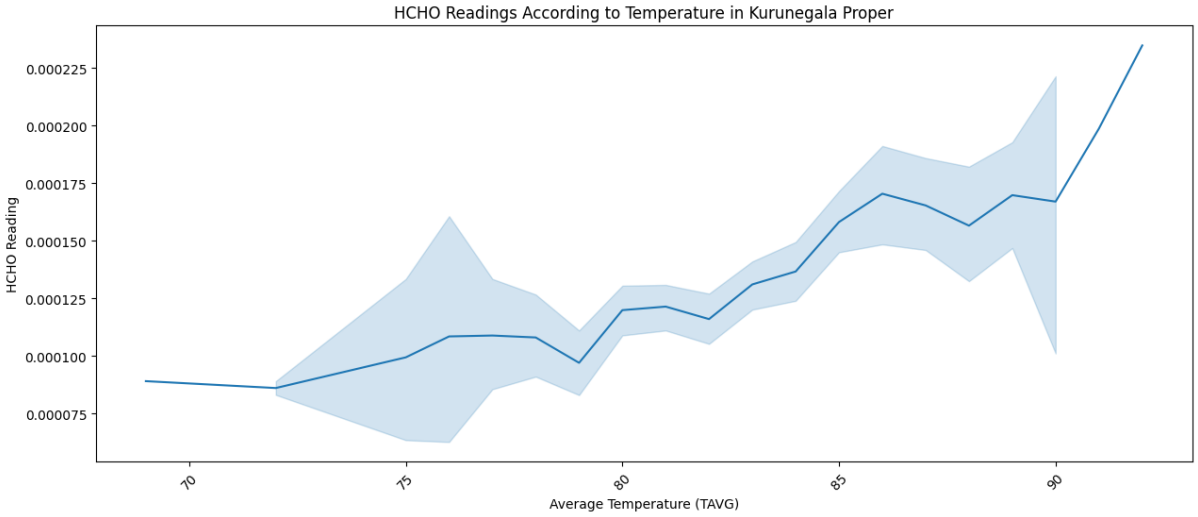


Figure 27: HCHO Reading According to Temp in Kurunegala

This chart shows how the HCHO readings change as the average temperature increases or decreases in Kurunegala.

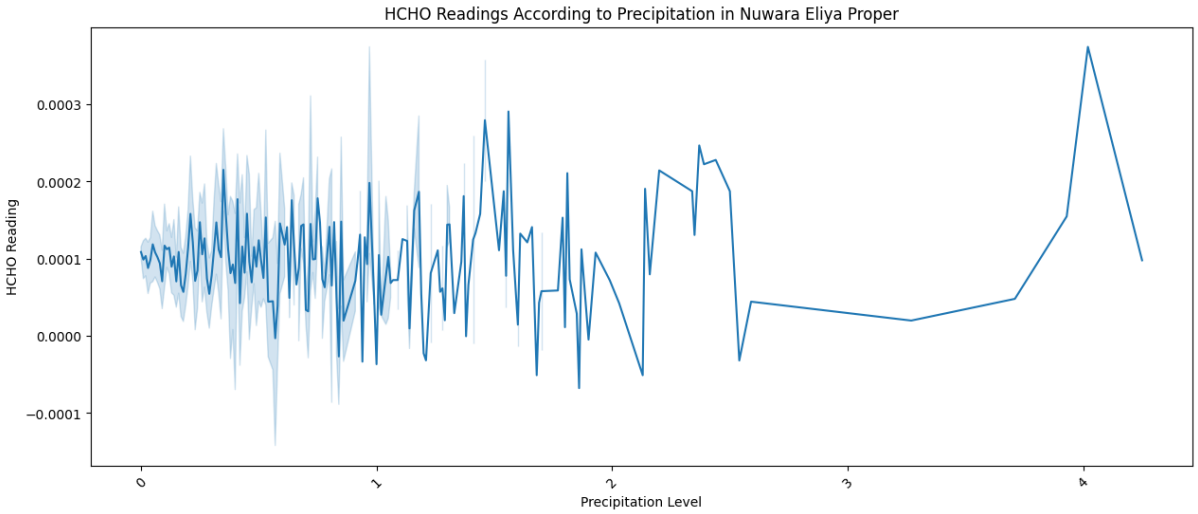


Figure 28: HCHO Readings According to Precipitation in Nuwara Eliya

This chart shows how the HCHO readings change as the average precipitation level increases or decreases in Nuwara Eliya.

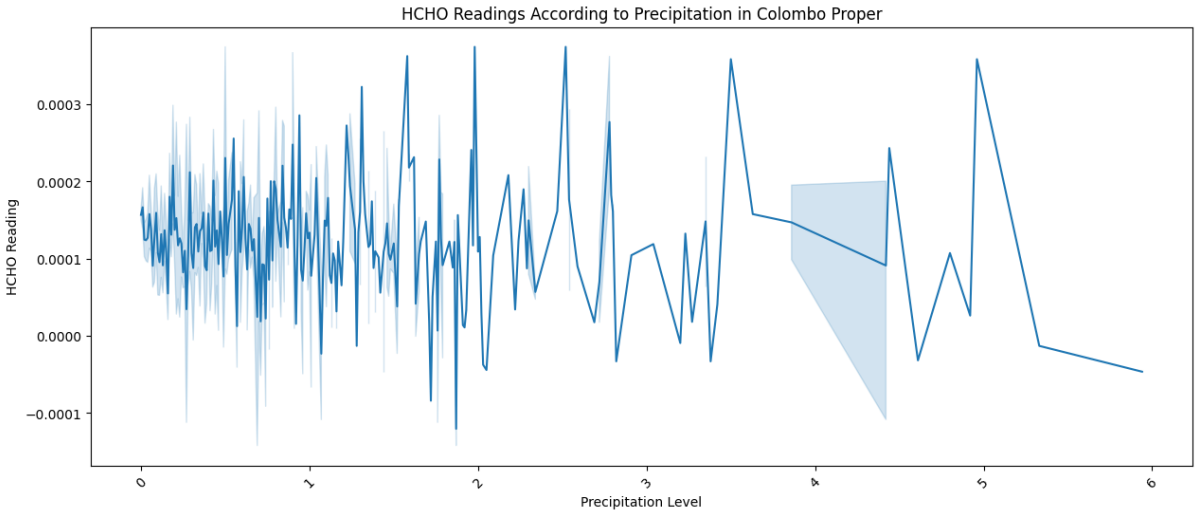


Figure 29: HCHO Reading According to Precipitation in Colombo

This chart shows how the HCHO readings change as the average precipitation level increases or decreases in Colombo.

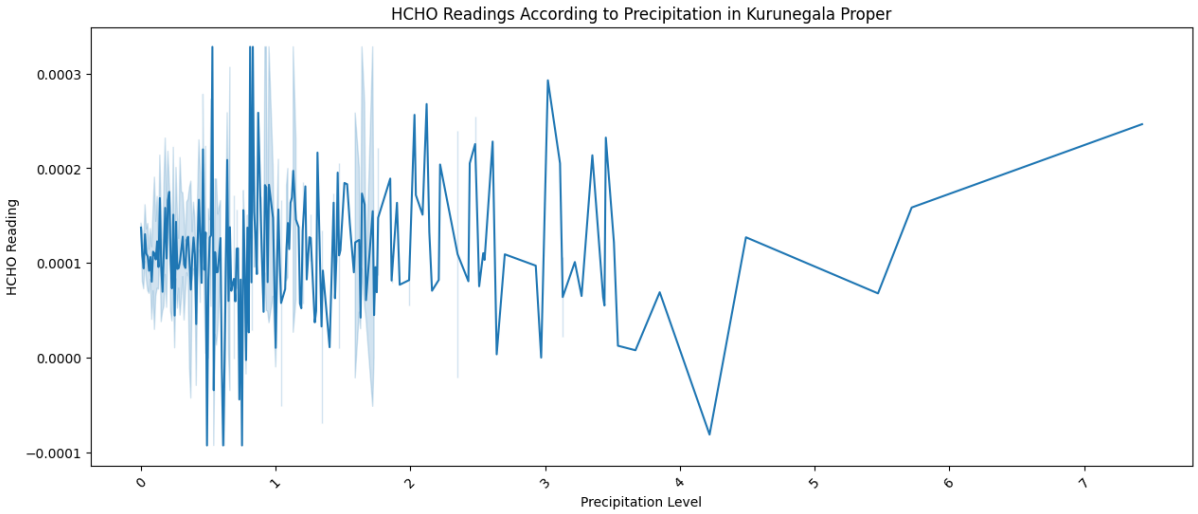


Figure 30: HCHO Reading to Precipitation in Kurunegala

This chart shows how the HCHO readings change as the average precipitation level increases or decreases in Kurunegala.

3. Machine Learning

To analyze the time series data for HCHO readings at a specific location, I utilized the SARIMAX model. This model extends the basic ARIMA model by incorporating seasonal patterns in the data, which makes it particularly effective for datasets that exhibit periodic fluctuations.

Overall, the SARIMAX model proved to be a powerful tool for time series analysis in this context, offering a more comprehensive approach than traditional ARIMA and providing valuable insights into the seasonal and non-seasonal trends of the HCHO readings.

3.1 ARIMA

Monaragala

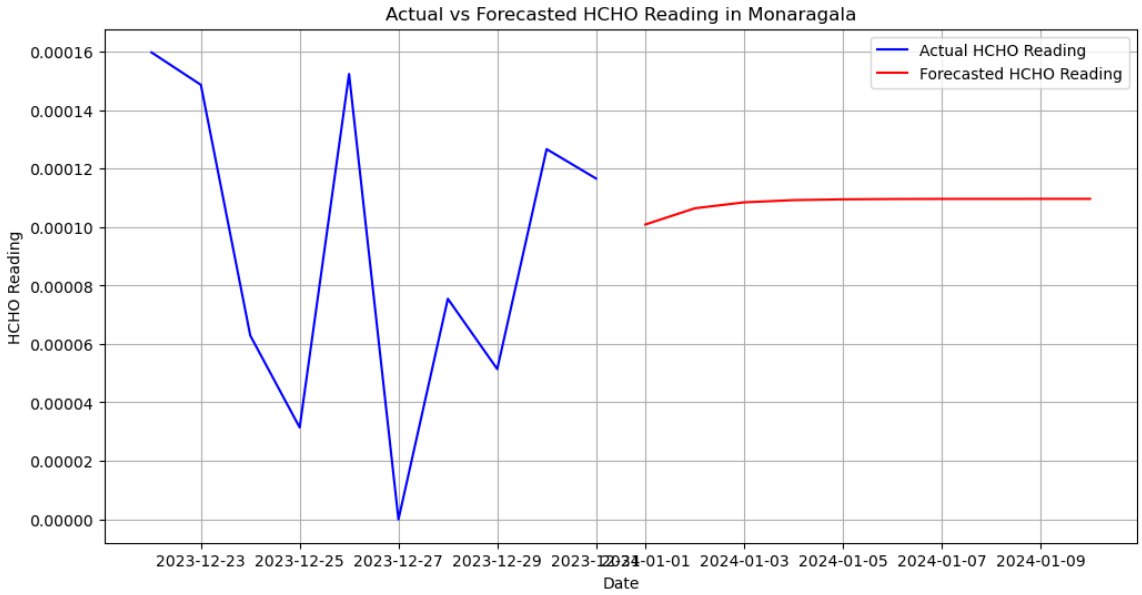


Figure 31: Actual vs Forecasted HCHO Reading in Monaragala (ARIMA)

Colombo

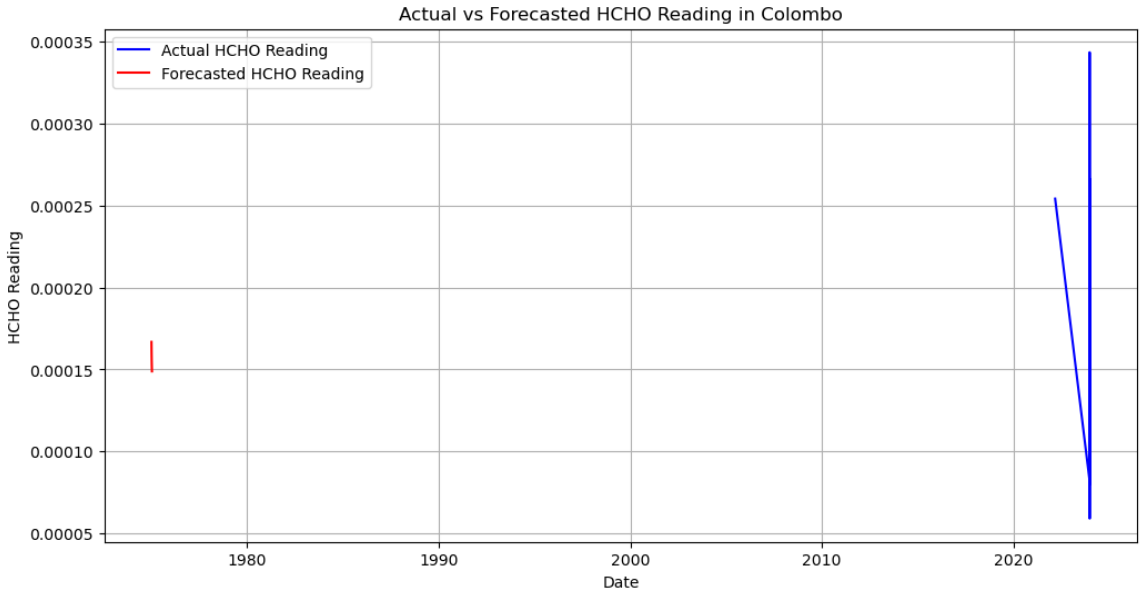


Figure 32: Actual vs Forecasted HCHO Reading in Colombo (ARIMA)

Matara

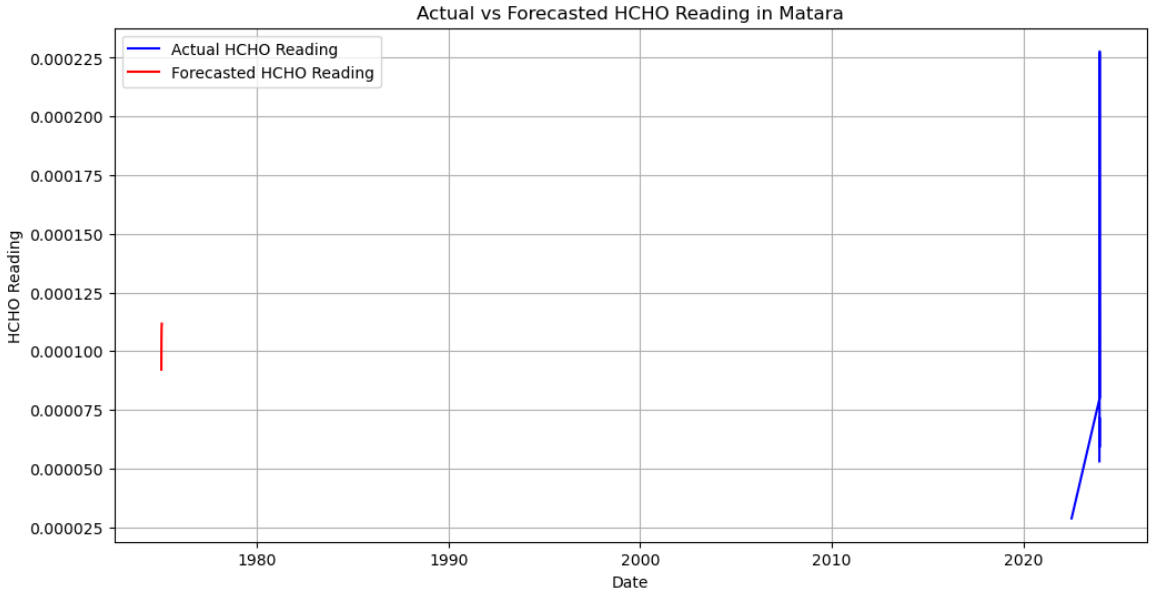


Figure 33: Actual vs Forecasted HCHO Reading in Matara (ARIMA)

Jaffna

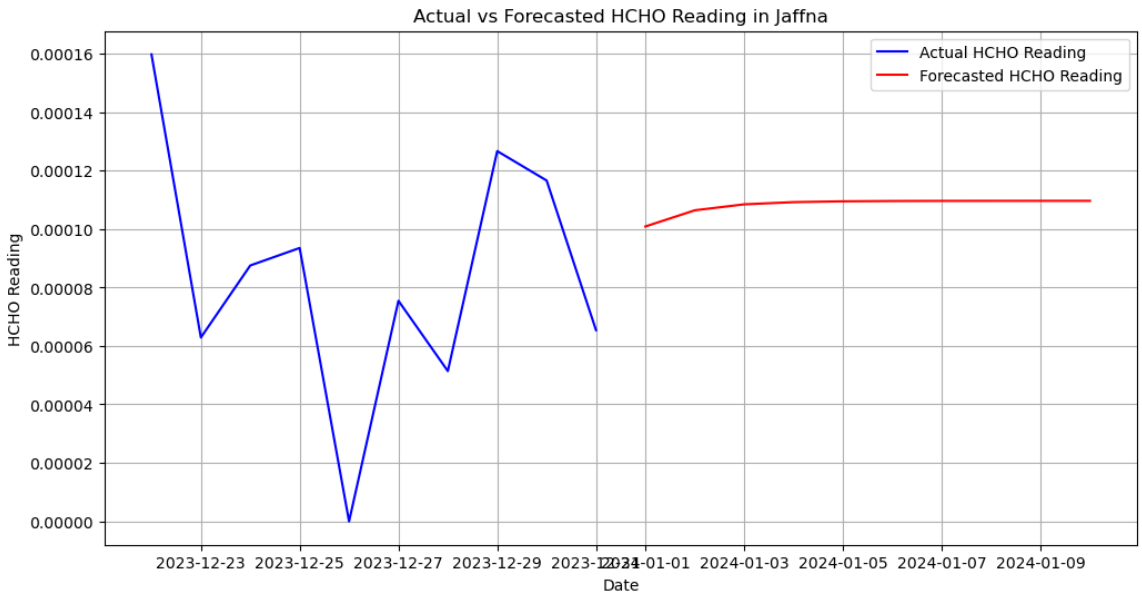


Figure 34: Actual vs Forecasted HCHO Reading in Jaffna (ARIMA)

Kandy

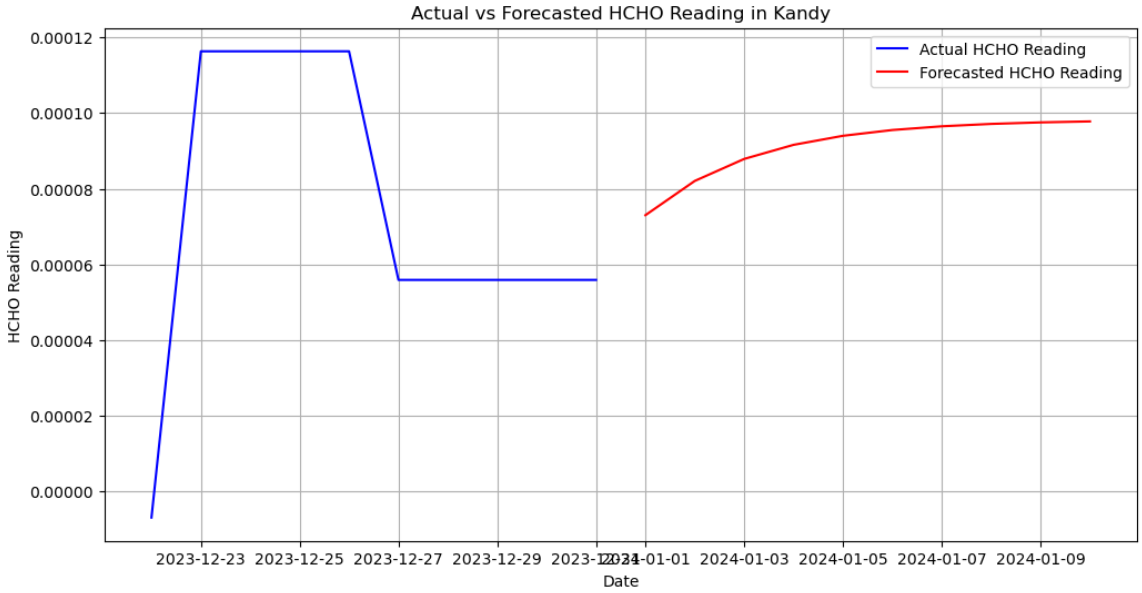


Figure 35: Actual vs Forecasted HCHO Reading in Kandy (ARIMA)

Kurunegala

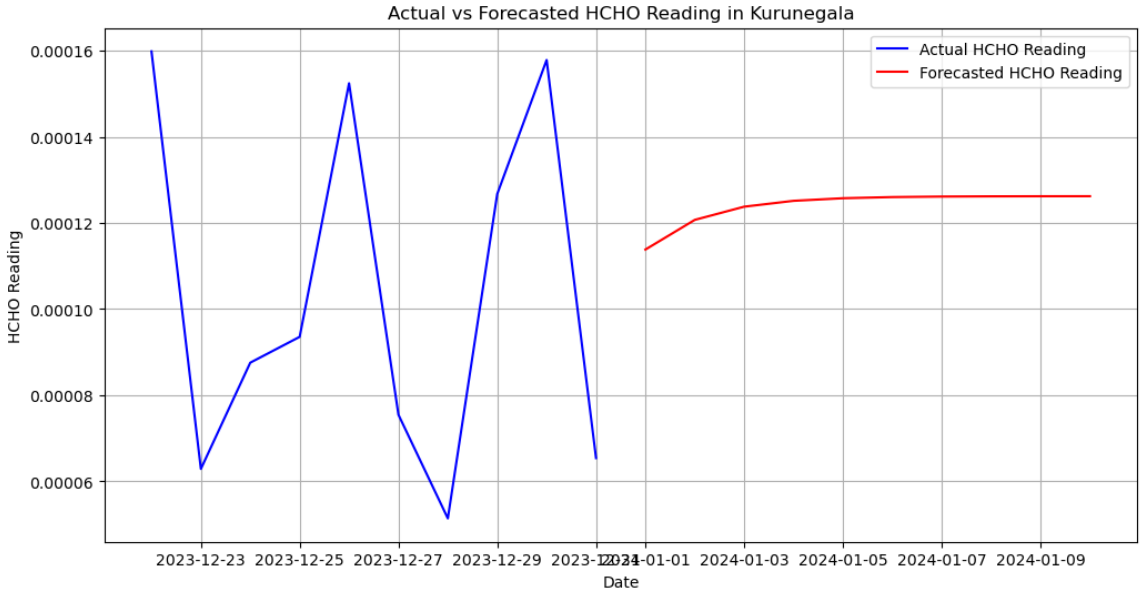


Figure 36: Actual vs Forecasted HCHO Reading in Kurunegala (ARIMA)

Nuwara Eliya

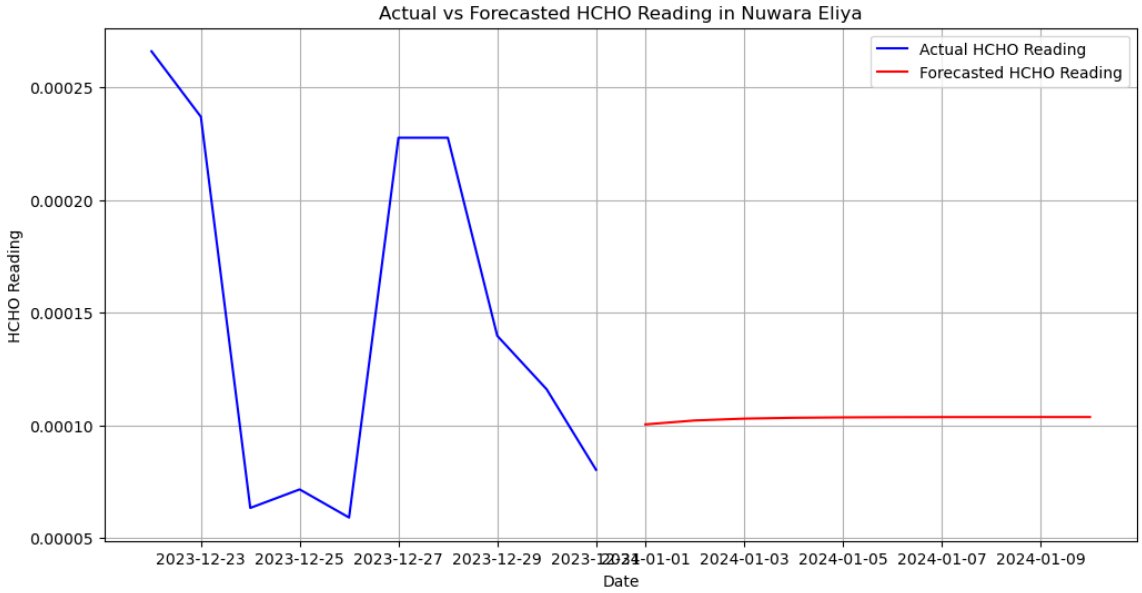


Figure 37: Actual vs Forecasted HCHO Reading in Nuwara Eliya (ARIMA)

3.2 SARIMAX

These graphs offer a valuable comparison between actual and forecasted HCHO levels, which is crucial for assessing the performance of a predictive model.

3.2.1 Monaragala

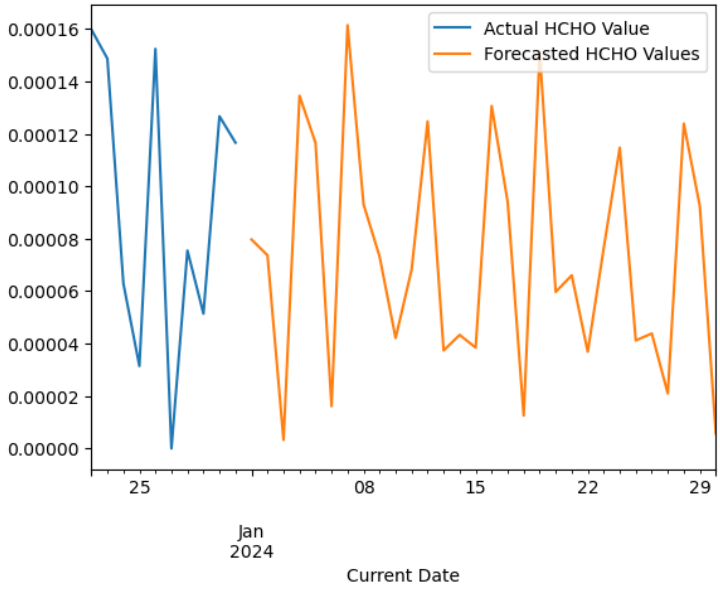


Figure 38: Future forecast of Monaragala (SARIMAX)

In the initial stages, the actual and forecasted values seem closely matched, indicating that the model may be effective under stable conditions.

However, as time progresses, the divergence between the two lines increases, suggesting that the model struggles with longer-term predictions.

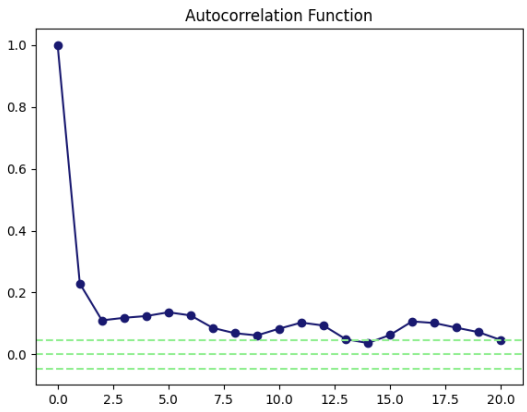


Figure 39: Autocorrelation Function.

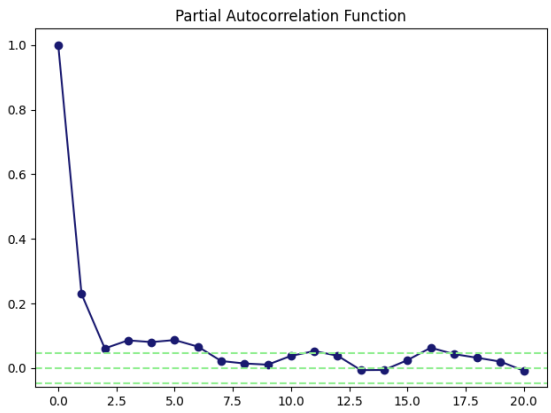


Figure 40: Partial Autocorrelational Function

The autocorrelation function and partial autocorrelation function graphs for each city here are pretty much the same. For that reason, only two cities have been graphed.

3.2.2 Colombo

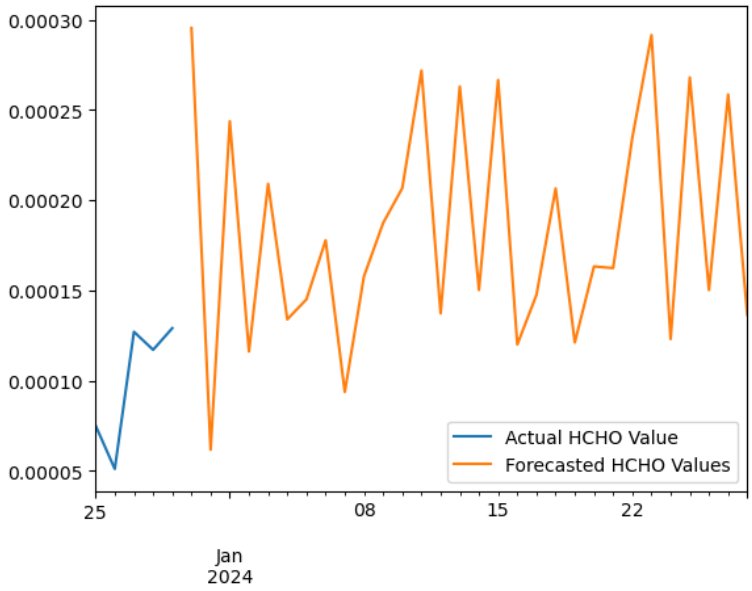


Figure 41: Future forecast of Colombo (SARIMAX)

While the model appears to be effective in making short-term predictions when the conditions are stable, it struggles with longer-term predictions, leading to a divergence between the actual and forecasted values.

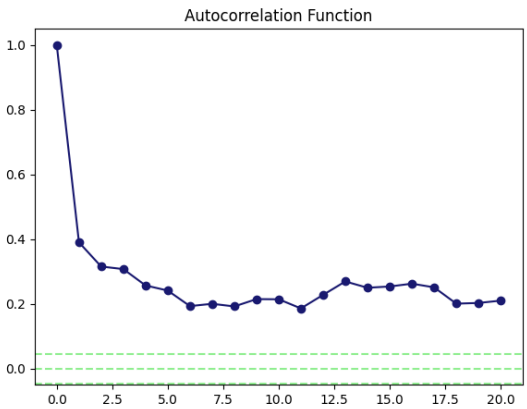


Figure 42: Autocorrelation Function

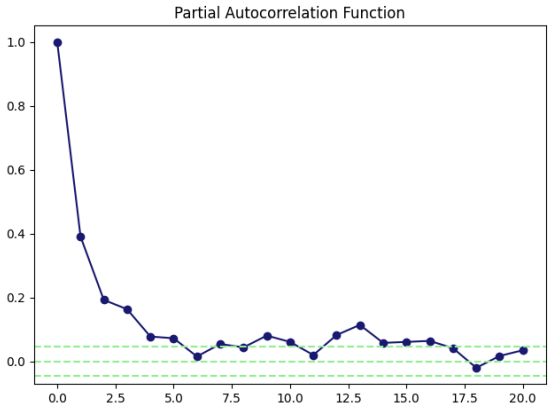


Figure 43: Partial Autocorrelation Function

3.2.3 Matara

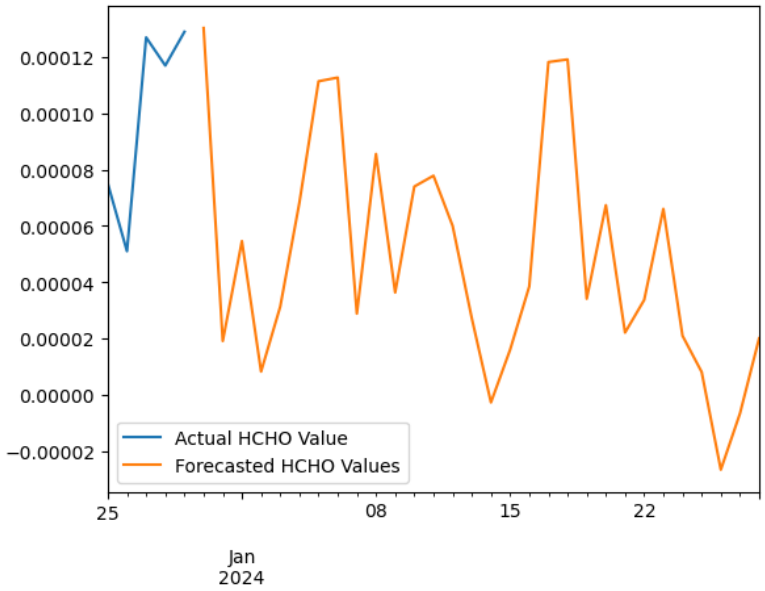


Figure 44: Future forecast of Matara (SARIMAX)

In the initial stages, the actual HCHO value and the forecasted values are closely matched, indicating that the model may be effective in making short-term predictions when the conditions are relatively stable.

The model appears to be effective in making short-term predictions when the conditions are stable, but it struggles with longer-term predictions, leading to a divergence between the actual and forecasted values.

3.2.4 Jaffna

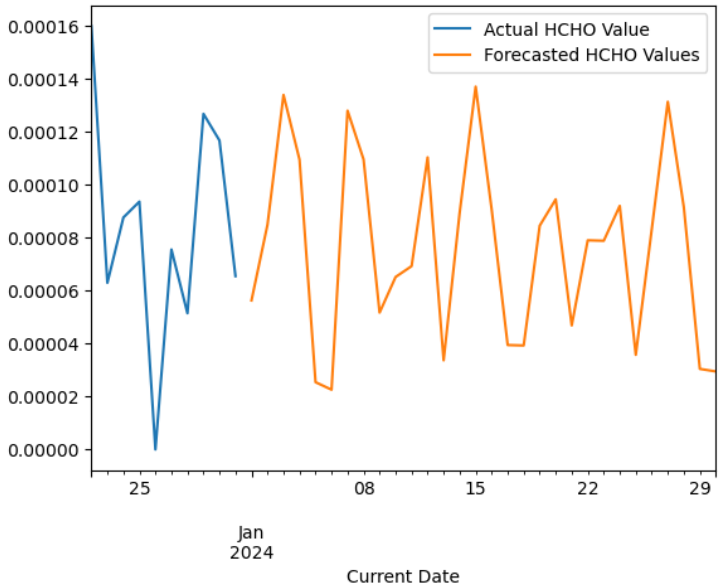


Figure 45: Future forecast of Jaffna (SARIMAX)

At the beginning of the graph, the actual and forecasted lines are closely aligned, indicating a good fit. However, as we move closer to the actual value, the forecasted line starts to deviate from the actual line, indicating a decrease in the model's accuracy.

3.2.5 Kandy

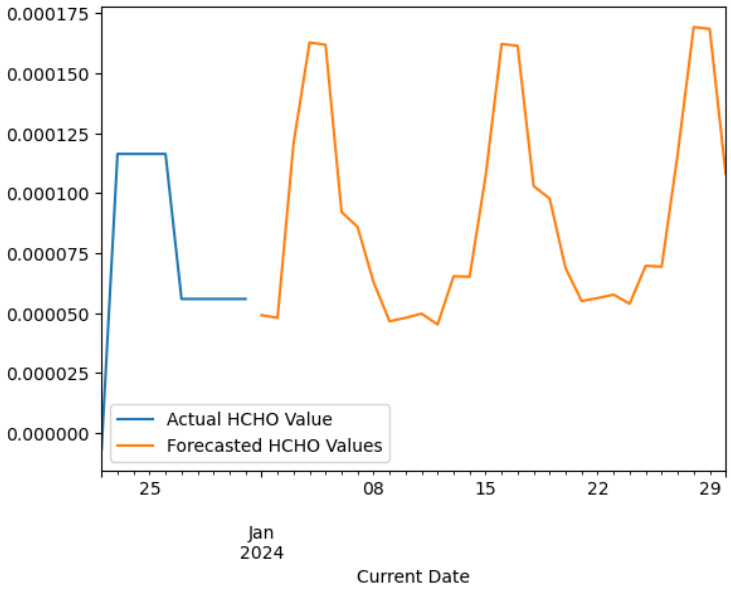


Figure 46: Future forecast of Kandy (SARIMAX).

Significant volatility in the actual values with a less volatile pattern in the forecasted values.

The model captures the general cyclical behavior but fails to match the amplitude of fluctuations, suggesting that it may average out the extreme values or is not sensitive enough to sudden changes in conditions.

3.2.6 Kurunegala

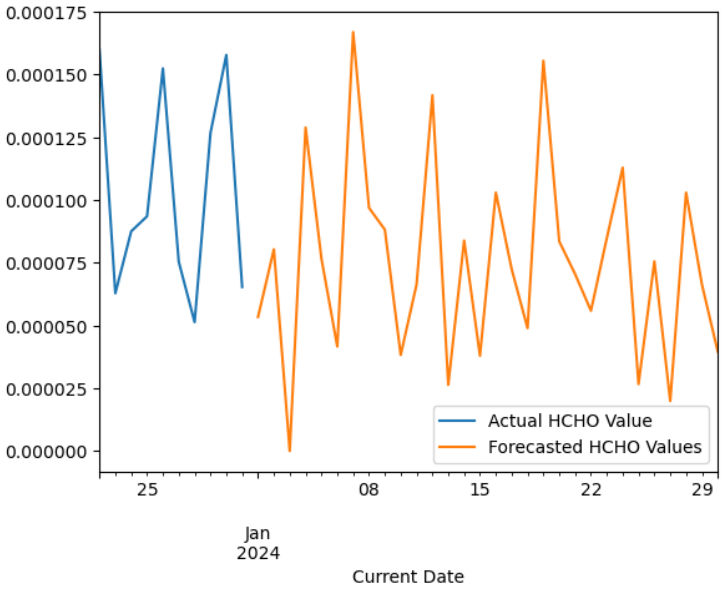


Figure 47: Future forecast of Kurunegala (SARIMAX).

Like the above, with smoother forecasted values and more pronounced actual fluctuations.

Consistency in the model's behavior across different datasets suggests a systematic smoothing effect, which could be due to the model's parameters or the nature of the input data it handles best.

3.2.7 Nuwara Eliya

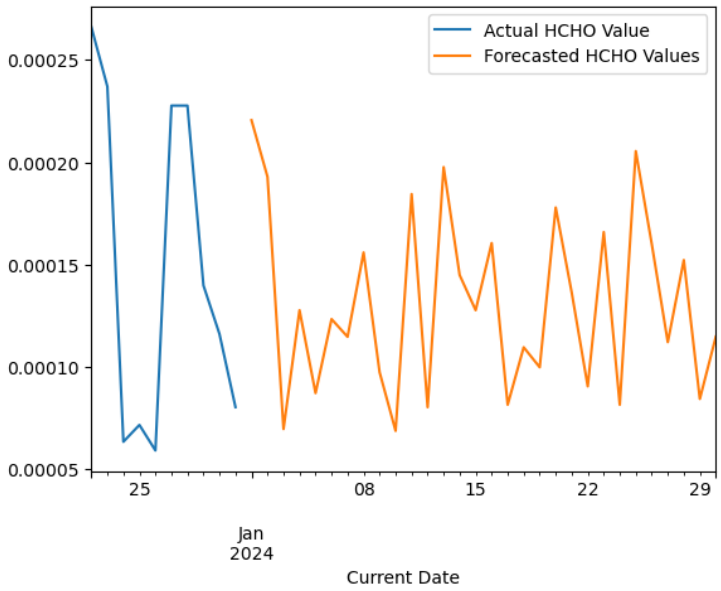


Figure 48: Future forecast of Nuwara Eliya (SARIMAX)

The forecast values are very smooth compared to the actual values, which show distinct weekly peaks.

The lack of responsiveness to weekly peaks suggests the model may not be incorporating data that captures weekly cyclic events, or it may not be adequately tuned to pick up such frequent fluctuations.

3.3 Model Performance

RMSE, MSE, and MAE are three common metrics used to evaluate the performance of regression models. They measure the difference between predicted values and actual values in a dataset.

- **RMSE (Root Mean Squared Error):** RMSE is the square root of the mean of the squared differences between predicted and actual values. It provides a measure of the average magnitude of prediction errors, giving more weight to larger errors due to the squaring.
- **MSE (Mean Squared Error):** MSE is the mean of the squared differences between predicted and actual values. It measures the average of the squares of the errors, providing a sense of the overall error in the model's predictions.
- **MAE (Mean Absolute Error):** MAE is the mean of the absolute differences between predicted and actual values. It provides a measure of the average magnitude of errors in a model's predictions.

Monaragala

```
RMSE: 0.00024749820518925764
MSE: 6.125536157190389e-08
MAE: 0.021318398004242033
```

Figure 49: Model Performance - Monaragala

Colombo

```
RMSE: 0.0009819849403105976
MSE: 9.64294422996808e-07
MAE: 0.08448734962574013
```

Figure 50: Model Performance - Colombo

Matara

```
RMSE: 0.0002263454564536852
MSE: 5.1232265657227105e-08
MAE: 0.019523924968189013
```

Figure 52: Model Performance - Matara

Jaffna

```
RMSE: 0.0001329102922970058
MSE: 1.7665145798475523e-08
MAE: 0.010857137509125417
```

Figure 51: Model Performance - Jaffna

Kandy

```
RMSE: 0.0011757030887444655
MSE: 1.3822777528832763e-06
MAE: 0.10391261424229134
```

Figure 53: Model Performance - Kandy

Kurunegala

```
RMSE: 0.0003598435626847625
MSE: 1.2948738960566258e-07
MAE: 0.031121378335040913
```

Figure 54: Model Performance - Kurunegala

Nuwara Eliya

```
RMSE: 0.00016415846548662472
MSE: 2.6948001790923362e-08
MAE: 0.013785933383821818
```

Figure 55: Model Performance - Nuwara Eliya

In each city, the RMSE, MSE, and MAE values are all very low, indicating that the model is highly accurate. The predictions closely match the actual data, suggesting that the model is well-suited for making predictions.

3.4 Limitations

1. **Volatility Clustering:** ARIMA and SARIMAX are not well-suited for data exhibiting volatility clustering, where periods of high volatility are followed by similar periods, as these models assume a consistent variance across the entire series.
2. **Data Frequency and Non-stationary Shocks:** These models require data with a fixed frequency and evenly spaced intervals. Irregularly spaced data points can complicate the model fitting process.
3. They generally assume that shocks (unexpected changes) are temporary, and that the series will revert to its mean, which may not be appropriate for data where shocks can cause a permanent change in the series level.
4. **Computational Complexity:** The inclusion of seasonal components and exogenous variables in SARIMAX increases the computational burden, particularly with larger datasets, which might be a concern for real-time processing or when handling massive datasets.
5. **Predictive Accuracy:** The accuracy of ARIMA and SARIMAX models may degrade over time or beyond the historical data range used for model fitting. Frequent re-estimation of the model may be necessary to maintain accuracy.

3.5 Potential Improvements

Deep Learning Approaches: Long Short-Term Memory Networks: For datasets with long-term dependencies or where the data sequence itself carries significant predictive power, deep learning models like LSTMs can often outperform traditional statistical models in capturing and forecasting such dynamics.

Volatility Modeling: Integration with GARCH Models: For time series data with apparent volatility clustering, such as financial markets, combining ARIMA/SARIMA with Generalized Autoregressive Conditional Heteroskedasticity (GARCH) models can effectively model and predict the level of volatility.

Error Correction: Systematic Error Adjustments: Incorporating error correction mechanisms into the forecasting model can adjust for any consistent bias or systematic errors identified in the forecasts, thus refining the accuracy of predictions.

Hybrid Models: Combination with Non-linear Models: Integrating ARIMA or SARIMA with machine learning algorithms that can capture non-linear patterns (e.g., random forests, neural

networks) can provide a more robust forecast by combining linear historical trend analysis with the ability to model complex interactions within the data.

4. Further Enhancements

1. Advanced Data Preprocessing

Objective: Utilize sophisticated techniques for managing missing data and outliers, ensuring data quality and consistency.

Impact: Cleaner, more reliable data leads to more accurate modeling and analysis, reducing the likelihood of skewed results due to poor data quality.

2. Machine Learning Enhancements

Objective: Implement advanced machine learning and deep learning techniques, such as CNNs for spatial data analysis and LSTMs for time-series forecasting.

Impact: These models can capture complex patterns and relationships in the data, enhancing predictive accuracy and providing deeper insights into temporal and spatial trends.

3. Real-time Analysis and Forecasting

Objective: Develop systems for real-time monitoring and forecasting of HCHO levels using live data integration.

Impact: Enables immediate response to elevated pollution levels and helps in proactive management of air quality.

4. Integration with Public Health Data

Objective: Analyze correlations between HCHO levels and health outcomes by integrating environmental data with public health records.

Impact: Supports the development of health advisories and targeted public health interventions.

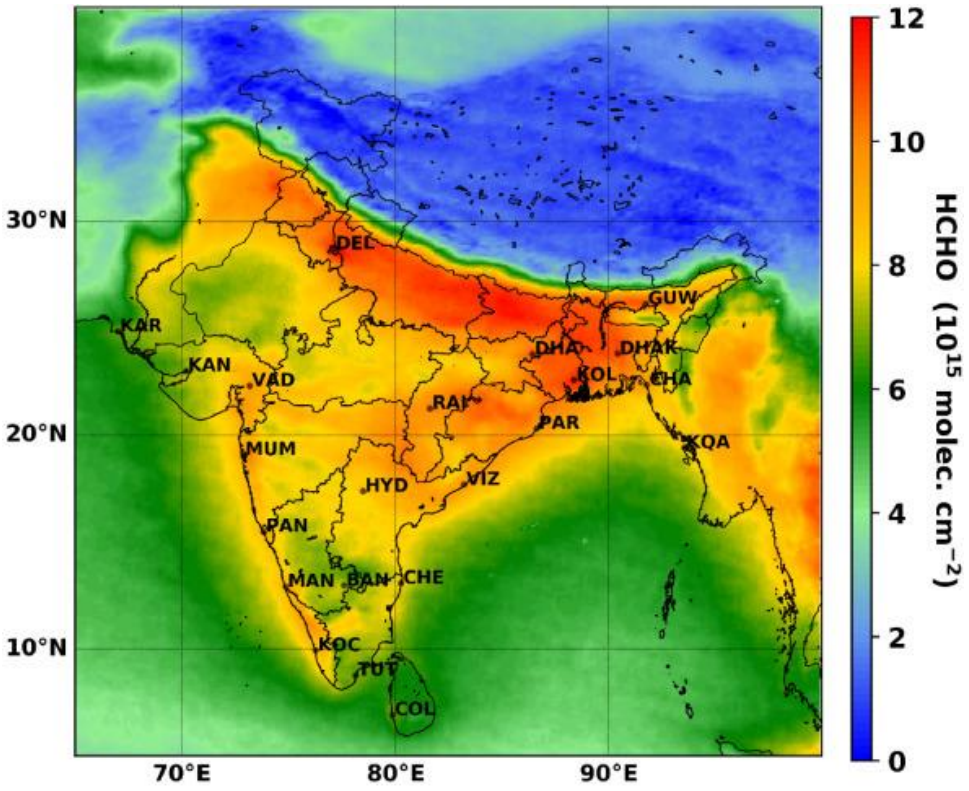
5. Similar Studies

Investigation of long–term trends and major sources of atmospheric HCHO over India

Link: https://www.sciencedirect.com/science/article/pii/S2667010022000373

This project has studied about the spread of HCHO level in India. The below map given here roughly shows the area where HCHO is spread in India. It also shows the HCHO expansion of Sri Lanka.

Figure 56: Map of HCHO level over India.


30°N
20°N
10°N
70°E
80°E
90°E
DEL
GUW
KAR
KAN
VAD
DHA
DHAK
KOL
CHA
RAI
PAR
MUM
HYD
VIZ
KOA
PAN
MAN
BAN
CHE
KOC
TUT
COL
12
10
8
6
4
2
0
HCHO (10^15 molec. cm^-2)

6. APPENDIX – References

GeeksforGeeks. (2020). How To Concatenate Two or More Pandas DataFrames? [online] Available at: https://www.geeksforgeeks.org/how-to-concatenate-two-or-more-pandas-dataframes/.

profile.w3schools.com. (n.d.). W3Schools. [online] Available at: https://profile.w3schools.com/log-in?redirect_url=https%3A%2F%2Fwww.w3schools.com%2Fpython%2Fpython_ml_preprocessing.asp [Accessed 20 Apr. 2024].

www.youtube.com. (n.d.). Forecasting in Power BI. [online] Available at: https://youtu.be/zl8F3e9SO2c?si=R2mBvsaNywAoImCA [Accessed 20 Apr. 2024].

GeeksforGeeks. (2023). Complete Guide To SARIMAX in Python. [online] Available at: https://www.geeksforgeeks.org/complete-guide-to-sarimax-in-python/.

kaggle.com. (n.d.). ARIMA Model for Time Series Forecasting. [online] Available at: https://www.kaggle.com/code/prashant111/arima-model-for-time-series-forecasting.