**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**

© The copyright for this project and all its associated products resides with Informatics Institute of Technology

# Table of Contents

[Table of Contents 2](#_Toc164446022)

[Introduction 3](#_Toc164446023)

[1. Data Preprocessing 4](#_Toc164446024)

[1.1 Clean and Prepare the Data 4](#_Toc164446025)

[1.1.1 Load the Data 4](#_Toc164446026)

[1.1.2 Explore Descriptive Statistics 4](#_Toc164446027)

[1.1.3 Handling Missing Values 5](#_Toc164446028)

[1.1.4 Handling Duplicates 6](#_Toc164446029)

[1.1.5 Handling Outliers 7](#_Toc164446030)

[2. Spatio-Temporal Analysis 13](#_Toc164446031)

[2.1 Analyze Trends Over Time 13](#_Toc164446032)

[2.1.1 Seasonal Variations 13](#_Toc164446033)

[2.1.2 Long-term Changes 14](#_Toc164446034)

[2.1.3 Trends Across Cities 15](#_Toc164446035)

[2.2 Changes in Gas Emissions due to the COVID-19 Lockdowns 22](#_Toc164446036)

[2.3 External Factors 24](#_Toc164446037)

[3. Machine Learning 29](#_Toc164446038)

[3.1 ARIMA 29](#_Toc164446039)

[3.2 SARIMA 33](#_Toc164446040)

[3.3 Model Performance 36](#_Toc164446041)

# 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.

# 1. Data Preprocessing

## Clean and Prepare the Data

### 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)

### 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|

+-------+--------------------+

### 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|

+------------+--------+------------+---------+

### 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)



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

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

### Handling Outliers

A graph of a distribution of hcho reading

Description automatically generatedDistribution 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

A graph of a distribution of hcho reading

Description automatically generated

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

Distribution of HCHO reading of second dataset (kan\_output.csv)

A graph of a distribution of hcho reading

Description automatically generated

Figure 3: Distribution of HCHO reading of second dataset.

Distribution of HCHO reading of second dataset after handling outliers

A graph of distribution of hcho reading

Description automatically generated

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

Distribution of HCHO reading of third dataset (mon\_kur\_jaf\_output.csv)

A diagram of a distribution of hcho reading

Description automatically generated

Figure 5: Distribution of HCHO reading of third dataset.

Distribution of HCHO reading of third dataset after handling outliers

A diagram of a distribution of hcho reading

Description automatically generated

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

Boxplot of HCHO reading of first dataset (col\_mat\_nuw\_output.csv)

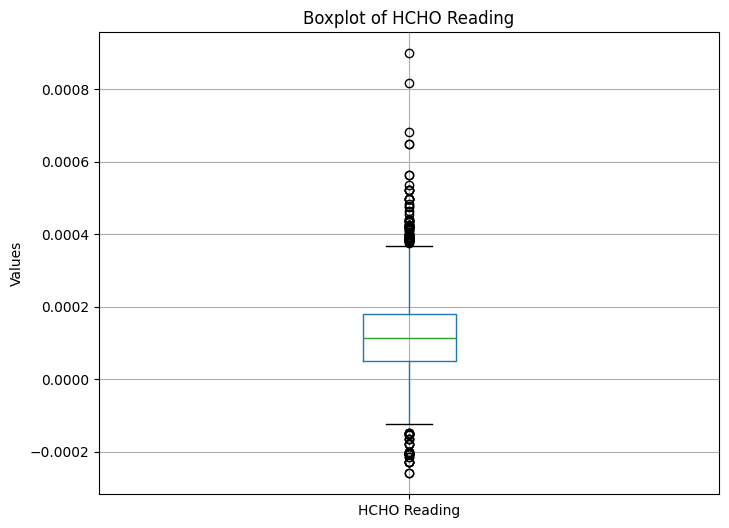


Figure 7: Boxplot of HCHO reading of first dataset.

Boxplot of HCHO reading of first dataset after handling outliers

A graph with a blue and green box

Description automatically generated

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

Boxplot of HCHO reading of second dataset (kan\_output.csv)

A graph with a box and a line

Description automatically generated with medium confidence

Figure 9: Boxplot of HCHO reading of second dataset.

Boxplot of HCHO reading of second dataset after handling outliers

A diagram of a box with a green and blue line

Description automatically generated

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

Boxplot of HCHO reading of third dataset (mon\_kur\_jaf\_output.csv)

A diagram of a box plot

Description automatically generated

Figure 11: Boxplot of HCHO reading of third dataset.

Boxplot of HCHO reading of third dataset after handling outliers

A graph with a blue and green line

Description automatically generated

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

# Spatio-Temporal Analysis

## Analyze Trends Over Time

### Seasonal Variations

A graph with lines and points

Description automatically generated

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.

Recognizing local events and activities that might impact these readings is crucial for accurate interpretation and ensuring that the measures adopted are well-suited to the specific challenges faced in each region or period.

### Long-term Changes

A graph of different colored lines

Description automatically generated

Figure 14: Average HCHO Reading by Year for Each City

The graph presenting average HCHO (formaldehyde) 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. Such insights can guide targeted actions to improve air quality and public health outcomes across different regions.

A graph showing different colored bars

Description automatically generated

Figure 15: Average HCHO Readings by Location

### Trends Across Cities

**Bibile, Monaragala**

A graph of different colored lines

Description automatically generated with medium confidence

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

This is about decomposing a time series signal from Bibile, Monaragala into three components: trend, seasonal, and residual.

* The trend component shows the long-term increasing or decreasing pattern in the data.
* The seasonal component shows the repetitive patterns within a year.
* The residual component shows the remaining fluctuations in the data that are not captured by the trend and seasonal components.

In the graph, the trend component is generally increasing, the seasonal component fluctuates around zero, and the residual component fluctuates around zero with no clear pattern.

**Colombo**

**A graph of different colored lines

Description automatically generated with medium confidence**

Figure 17: Colombo - Trend, Seasonal, Residual Component

The three graphs in the image compare the trend, seasonal, and residual components of a time series signal for HCHO readings in Colombo Proper.

* The trend component shows a generally increasing trend in HCHO readings over time.
* The seasonal component shows a cyclical pattern, with HCHO readings being higher in some months than others.
* The residual component shows the remaining fluctuations in the data that are not captured by the trend and seasonal components. There is no clear pattern to the residuals.

**Deniyaya, Matara**

**A graph of different colored lines

Description automatically generated**

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

Trend Component: This graph shows the long-term rise or fall in the climate variable. There seems to be a slightly increasing trend over time.

Seasonal Component: This graph depicts the cyclical pattern within a year. The values fluctuate around zero, with potentially higher values in specific months.

Residual Component: This graph shows the remaining fluctuations in the data after the trend and seasonal components are removed. There's no clear pattern in the residuals.

the trend component suggests a gradual increase in the climate variable, while the seasonal component indicates a yearly cycle.

**Jaffna**

**A group of graphs showing different colored lines

Description automatically generated**

Figure 19: Jaffna - Trend, Seasonal, Residual Component

Trend Component: The trend component shows a slight increasing trend in the climate variable over time.

Seasonal Component: The seasonal component shows a cyclical pattern, with the climate variable being higher in some months than others.

Residual Component: The residual component shows the remaining fluctuations in the data that are not captured by the trend and seasonal components. There is no clear pattern to the residuals.

**Kandy**

**A graph of different colored lines

Description automatically generated**

Figure 20: Kandy - Trend, Seasonal, Residual Component

Trend Component: This graph shows the long-term rise or fall in HCHO levels over time. It's difficult to discern a definitive trend from the image you provided.

Seasonal Component: This graph depicts the cyclical pattern within a year for HCHO levels. There might be periods with consistently higher or lower HCHO concentrations.

Residual Component: This graph shows the remaining fluctuations in the data after the trend and seasonal components are removed. There's no clear pattern in the residuals, representing unpredictable variations in HCHO levels.

**Kurunegala**

**A graph of different colored lines

Description automatically generated with medium confidence**

Figure 21: Kurunegala - Trend, Seasonal, Residual Component

**Nuwara Eliya**

**A graph of different colored lines

Description automatically generated with medium confidence**

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

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

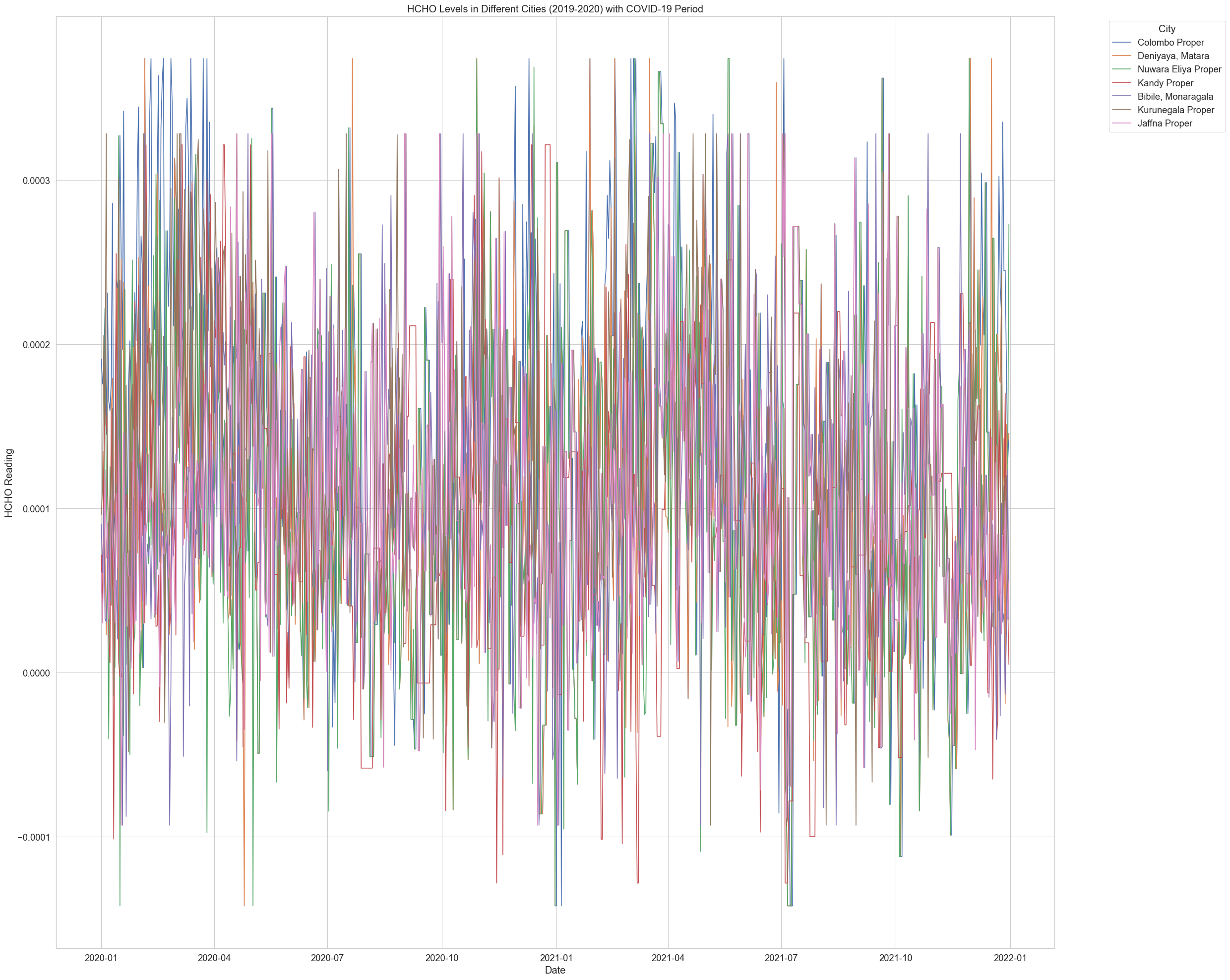


Figure 23: HCHO Levels in Different Cites with Covid-19 Period

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.

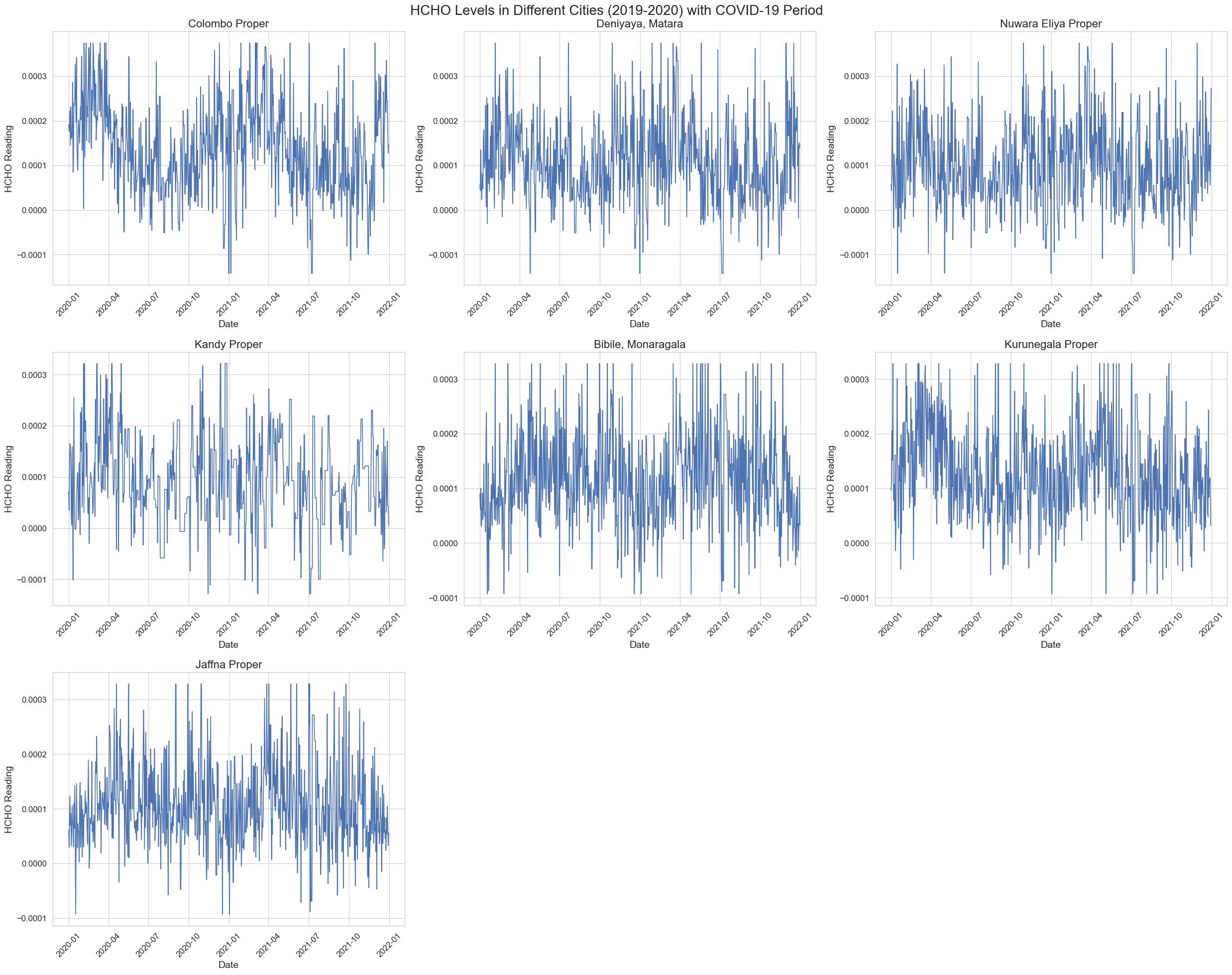


Figure 24: HCHO Levels in Different Cites with Covid-19 Period (One graph for each city)

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.

In general, HCHO levels are influenced by various factors, including industrial activities, traffic emissions, and possibly natural sources. The lockdown may have impacted these activities differently in each city. However, environmental factors such as temperature, humidity, and wind patterns can also affect HCHO levels and should be considered when analyzing these patterns. Moreover, the consistency of data collection and any local events (such as fires or industrial incidents) can also impact the data.

## External Factors

A graph showing a blue line

Description automatically generated

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.

**A graph showing a blue line

Description automatically generated**

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.

**A graph with a line going up

Description automatically generated**

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.

**A graph showing a graph

Description automatically generated with medium confidence**

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.

**A graph showing a graph

Description automatically generated with medium confidence**

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.

**A graph of a graph

Description automatically generated**

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.

# 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.

## ARIMA

**Monaragala**

A graph with a line drawn on it

Description automatically generated

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

**Colombo**

A graph with numbers and a number of data

Description automatically generated with medium confidence

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

**Matara**

**A graph with a number of data

Description automatically generated with medium confidence**

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

**Jaffna**

**A graph with a line drawn on it

Description automatically generated**

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

**Kandy**

**A graph with red and blue lines

Description automatically generated**

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

**Kurunegala**

**A graph with a red line

Description automatically generated**

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

**Nuwara Eliya**

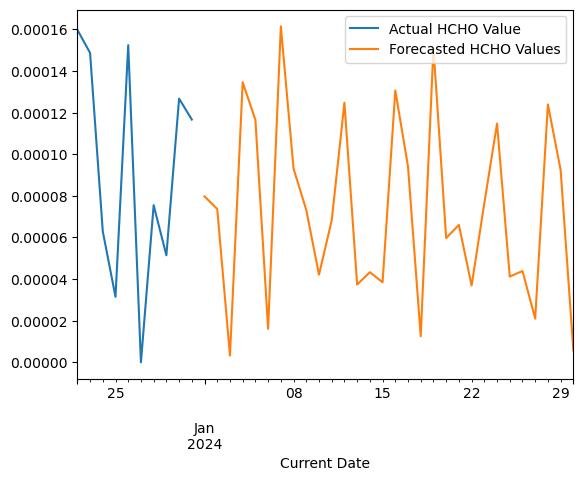
**A graph with a line drawn on it

Description automatically generated**

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

## SARIMAX

### Monaragala

****

A graph with a line and a dotted line

Description automatically generatedA graph with blue lines and a dotted line

Description automatically generatedFigure 40: Future forecast of Monaragala (SARIMAX)

Figure 38: Autocorrelation Function.

Figure 39: Partial Autocorrelational Function

### C**olombo**

Figure 41: Future forecast of Colombo (SARIMAX)

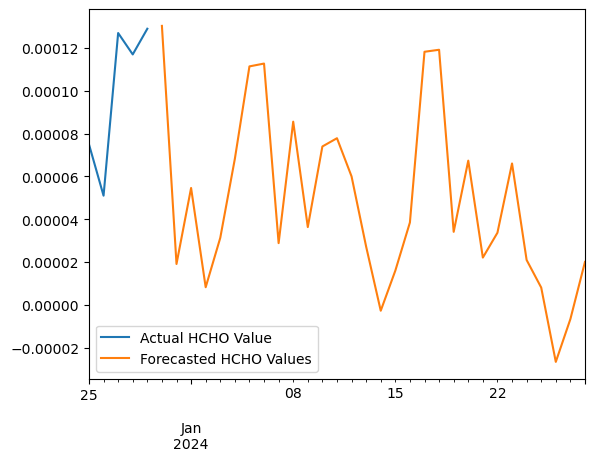
**A graph with a line

Description automatically generatedA graph with a line and a blue line

Description automatically generated**

Figure 42: Autocorrelation Function

Figure 43: Partial Autocorrelation Function

****

### Matara

Figure 44: Future forecast of Matara (SARIMAX)

### Jaffna

Figure 45: Future forecast of Jaffna (SARIMAX)

### Kandy

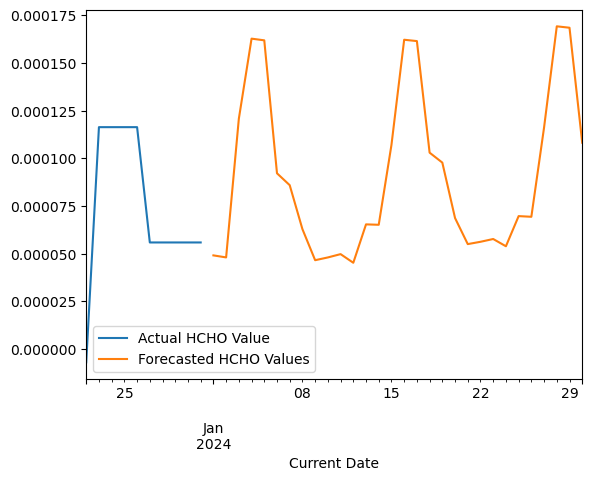
****

Figure 46: Future forecast of Kandy (SARIMAX).

### Kurunegala

Figure 47: Future forecast of Kurunegala (SARIMAX).

### Nuwara Eliya

Figure 48: Future forecast of Nuwara Eliya (SARIMAX)

## 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.