Open Access and Peer Reviewed Journal ISSN 2394-2231 https://www.ijctjournal.org/

RESEARCH ARTICLE OPEN ACCESS

An Efficient Data Preprocessing Framework for Time-Series Air Pollution Data

Dr Asha N[1], Dr Pushpalatha M, Dr Sowmya S B

[1] *Associate Professor, Department of Computer Science, Nrupathunga University (formerly Government Science College), Bangalore, India*

ORCID: 0000-0002-2615-5890

[2] *Associate Professor, Department of Computer Science, Government Women's College Hunsur, Karnataka, India*

[3]*Associate Professor, Department of Mathematics, Nrupathunga University (formerly Government Science College), Bangalore, India*

email: ashan.gsc@gmail.com

Abstract: Real-world datasets are often characterised by noise, missing values, inconsistencies, and anomalies, which can adversely affect the performance of data mining and machine learning algorithms. Therefore, data preprocessing is an essential step in ensuring data quality and improving the efficiency and accuracy of the analytical process. The time-series pollution dataset is initially preprocessed to ensure data quality and consistency before applying statistical or machine-learning techniques. The preprocessing begins with the identification and treatment of missing or invalid observations, represented as *NaN (Not a Number)* values. The paper emphasis on incorporating Appropriate imputation techniques to handle missing observations while preserving the temporal characteristics of the dataset. Also, outliers and anomalous observations are detected and treated using suitable statistical methods. Pollution data exhibit extreme observations due to sudden variations in pollutant concentrations, and their distribution may be left-skewed or right-skewed. Therefore, statistical measures such as the Interquartile Range (IQR) are employed to identify and appropriately handle abnormal observations. These preprocessing procedures help produce a reliable and consistent hourly time-series dataset, thereby improving the quality of subsequent analysis and the accuracy of pollution and AQI prediction models. The paper proposes an efficient Data preprocessing framework for time-series air pollution datasets.

Keywords— Data Preprocessing, Time series Data set, Not a Number(NaN), Interquartile range, Air pollution dataset.

I. INTRODUCTION

The annual and monthly average data was not sufficient to carry out future trends of air pollution data set. This gave an insight into the research work to collect time series data set which contains hourly-based data of pollutant concentration in 2018, and 2019 from 17 different monitoring stations, with 20 attributes and 18,673 instances. A **time-series air pollution dataset** consists of sequential observations of atmospheric pollutants recorded at specific and usually regular time intervals, such as hourly, daily, or monthly measurements. These datasets are valuable for understanding the temporal variation of air pollution and for developing predictive models for **Air Quality Index (AQI)** and individual pollutant concentrations. The behaviour of time-series data can generally be characterised as linear or nonlinear, depending on the nature of the relationships and temporal dependencies among the observed variables. Air pollution time-series data are often highly nonlinear, owing to complex interactions among multiple pollutants, meteorological conditions, seasonal variations, emission sources, and temporal patterns. Consequently, conventional linear statistical models may not always adequately capture the underlying relationships within the dataset. Therefore, various statistical, machine-learning, and nonlinear predictive models can be applied to air pollution time-series data to identify the patterns and relationships and to determine the model that provides the best predictive accuracy with minimum error [1]. Seasonal effects and trend variations are important characteristics of many time-series datasets, particularly in the analysis of air pollution. Monthly and quarterly pollution data may exhibit distinct seasonal patterns, reflecting recurring changes in pollutant concentrations over different periods of the year. In contrast, annual pollution data are more useful for identifying long-term trends and gradual changes in air-quality conditions. Seasonal variations in air pollution can be influenced by several factors, including weather and meteorological conditions, temperature, rainfall, wind patterns, human activities, holidays, industrial operations, and other periodic events. These factors can result in systematic fluctuations in pollutant concentrations over time. Therefore, identifying and appropriately modelling seasonality and trend components is an important aspect of time-series analysis and can contribute to improved accuracy in air pollution and AQI prediction

trend-based time-series data, as reliable predictions enable effective planning, resource allocation, risk management, and timely decision-making across various fields. In the context of air pollution forecasting, accurate prediction of pollutant concentrations and AQI can support environmental monitoring, public-health planning, and the implementation of appropriate pollution-control measures. Therefore, selecting appropriate forecasting models capable of effectively capturing trend, seasonality, temporal dependencies, and nonlinear patterns is essential for achieving reliable predictive performance [3].

II. RELATED WORK

The experimental research considers $PM_{2.5}$, PM_{10}, NO, NO_2, NOx, O_3, SO_2 and CO. Importantly, the authors emphasise that preprocessing should be reproducible and auditable, which could strengthen the proposed "efficient data preprocessing framework."[4]. The outlier/anomaly detections are Error-based outliers - instrument failure, sensor drift, calibration problems, etc. Event-based extreme values: genuine sudden changes in pollution caused by environmental conditions. The distinction is particularly important for the research as the research should not automatically remove every extreme pollution value. A genuine pollution episode may be an important observation rather than an error [5]. The dataset includes $PM_{2.5}$, PM_{10}, NO_2, SO_2, CO, O_3 and NH_3.The preprocessing includes data cleaning, missing-value treatment, timestamp standardisation, and AQI generation [6]. The paper explicitly describes preprocessing of $PM_{2.5}$, PM_{10}, SO_2, NO_2, CO, O_3 and meteorological variables. The preprocessing includes: Outlier testing → missing-value filling → data normalisation. The paper uses a box-plot/IQR-based approach for outlier detection, making it particularly useful for supporting the statistical methodology [7].

III. METHODOLOGY

The Time series data collected has been used for the following phases during the research work,

Phase 1: Preparation of Time series data
Phase 2: Predict future trends of pollutant concentration
Phase 3: Impacts of various pollutants and Meteorological parameters on PM_{10} and $PM_{2.5}$

A. Preparation of Time series data

The hourly time-series air pollution dataset is initially subjected to a systematic preprocessing procedure to improve its quality, consistency, and suitability for subsequent statistical and machine-learning analysis. The preprocessing involves identifying and treating missing or invalid values, particularly NaN (Not a Number) observations, as well as measurements may contain extreme values resulting from sudden changes in pollutant concentrations, it is necessary to distinguish genuine pollution events from erroneous observations. Furthermore, pollution variables may exhibit left-skewed or right-skewed distributions, indicating deviations from a symmetric distribution. Therefore, appropriate statistical methodologies are applied to the hourly time-series dataset to examine its distributional characteristics and identify potential outliers. In particular, the Interquartile Range (IQR) method can be employed for robust outlier detection. The preprocessing of missing values and anomalous observations ensures that the resulting dataset retains its temporal structure while providing a reliable and consistent foundation for subsequent air pollution and AQI prediction [8].

B: Predict future trends of pollutant concentration

After preprocessing, the hourly time-series air pollution dataset is used to identify and predict future patterns in pollutant concentrations. The dataset contains multiple pollutant variables that may influence overall air quality, with PM10 and PM2.5 being important indicators of particulate pollution. Therefore, the influence of other pollutant parameters, including NO_2, SO_2, CO, O_3, and NH_3, is examined to determine their contribution to variations in PM10, PM2.5, and overall AQI. The temporal relationships and dependencies among these pollutant concentrations are analysed to identify significant patterns and influential variables for future prediction. The resulting relationships are subsequently utilized by appropriate statistical and machine-learning models to improve the accuracy and reliability of air pollution and AQI forecasting.

C: Impacts of various pollutants and Meteorological parameters on PM_{10} and $PM_{2.5}$

The influence of meteorological parameters on PM10 and PM2.5 concentrations is also examined over the corresponding time period to understand their temporal relationships and contribution to variations in particulate matter. Parameters such as temperature, relative humidity, wind speed, and rainfall can influence the dispersion, accumulation, and transformation of air pollutants. To improve the reliability of future predictions, the hourly time-series dataset is decomposed into three fundamental components: trend, seasonality, and residual. The trend component represents the long-term movement in pollutant concentrations, the seasonal component captures recurring temporal patterns, and the residual component represents irregular fluctuations that cannot be explained by the trend and seasonal components. This decomposition facilitates a better understanding of the underlying temporal structure of PM10

more accurate air-pollution forecasting models.

Methods used in Data Pre-processing

The Data Pre-processing of the Pollution data set is carried out with different methods during every phase of the modules. The reason behind Data Pre-processing is to

1. Replacing the Missing values
2. Removal of Outliers

A. Replacing Missing Values

The pollution data set archived may contain missing values. The missing values may be blank, dashes, or may be NAN (Not a Number values) which will cause noise in the data set. These noises are generated due to human error, hardware failure, or loss of data transfer. These noises have to be removed using mean-median computation. This method is implemented in Python in different phases of the modules during replacing the missing values of the pollution data set.

B. Removal of Outliers

An outlier is defined as a data point that deviates significantly from the norm [9]. Outliers are cases that do not fit the same model as the rest of the data and appear to be from a different probability distribution [10]. As a result, an outlier includes both incorrect and surprisingly correct data. Outlier detection is a procedure that selects k samples that are significantly different, exceptional, or inconsistent with the remaining data [11,12]. Global outliers, contextual (or conditional) outliers, and collective outliers are the three types of outliers. If a data object deviates significantly from the rest of the data set, it is considered a global outlier. A contextual outlier is an object that deviates significantly from a specific context of the object. A collective outlier is formed when a subset of data objects deviates significantly from the entire data set [13].

There are three different categories of outlier detection methods. They are

1. Statistical Methods
2. Clustering Methods
3. Classification Methods

The Data Pre-processing to remove outliers is carried out using Statistical Methods such as,

1. Interquartile range
2. Gaussian Distribution

A. **Interquartile range**

If the data points are found to be continuously distributed, the Interquartile range (IQR) is a technique that assists to find outliers in the data [14,15]. It is given by the equation

IQR= Q3-Q1

Where,

Q1 is the first quartile

The interquartile range (IQR) can be used to divide the points by 25% and to create a boxplot, which is a simple graphical representation of the interquartile range. In this method, the pollution dataset is divided into quartiles and divided into four equal parts. Q1, Q2, and Q3 are the first, second, and third quartiles, respectively [11 12].

Algorithm 1.1 to detect an interquartile range

(1) Median is used to calculate quartiles recursively.
(2) If 2n is the even number of datapoint entries Then,
Q1 is the first quartile = smallest data point entry median
Q3 is the third quartile = largest data point entry median,
(3) If 2n + 1 is the odd number of entries Then,
Q1 is the first quartile = smallest data point entry median
Q3 is the third quartile = largest entry median,
(4) Q2 is the second quartile of an ordinary median.

Gaussian Distribution

The data points in the hourly-based pollution dataset contain extremely high or low data points. “Since the data points need to have symmetrical data to forecast future values, Gaussian Distribution also called Normal Distribution is applied to make the data points symmetrical” [16].

The graph of the Normal Distribution is categorized by two parameters

1. Mean and average which is the maximum of the graph and about which the graph is always symmetric [16].
2. Standard Deviation determines the dispersion away from the mean. A small Deviation produces a steep graph whereas a large deviation produces a flat graph. [16].

The hourly data, based on Time series data as shown in Figure 1.1. does not have discrete or distinct values. It is not possible to forecast the data when plotting all the pollutant concentrations with respect to time series; the data in the graph is either left or right-skewed. So, to scale the data to a given range, a normal distribution is used. The Data Pre-processing for Time series was very essential as it included NAN values. The NAN values were replaced with the nearest data points using the interpolation method. It was also necessary to study the influence of pollutant concentrations

also the influence of meteorological parameters on PM_{10} and $PM_{2.5}$.

Time	CO mg/m³	Ozone µg/m³	NO µg/m³	NO2 µg/m³	NOx µg/m³	NH3 µg/m³	SO2 µg/m³	PM2.5 µg/m³	PM10 µg/m³	BEN µg/m³	TOL µg/m³	m,p-XYL	o-XYL µg/m³	Et-BEN	AT °C	RH %	WS m/s	WD deg.	BP mm
01-01-2018 01:00	0.92	3.7	19.5	46.3	65.8	43.4	3.7	57	104	3.4	9.8	5.6	0.5	0.1	-	73	0.8	190	706
01-01-2018 02:00	0.80	8.4	12.6	38.9	51.6	44.3	2.9	43	77	3.2	14.8	4.8	0.1	0.1	29.0	82	1.4	221	706
01-01-2018 03:00	0.69	10.8	6.3	35.4	41.6	48.9	2.6	40	69	2.4	10.0	3.6	0.1	0.1	29.0	88	1.2	217	705
01-01-2018 04:00	0.57	7.5	6.9	30.5	37.3	33.1	4.2	40	72	2.0	24.6	3.2	0.2	0.1	29.0	90	0.7	191	705
01-01-2018 05:00	0.69	4.1	26.7	25.2	52.0	33.8	3.9	47	69	1.8	0.1	3.1	0.1	0.0	29.0	91	0.6	183	705
01-01-2018 06:00	1.03	0.2	85.4	36.5	122.1	29.2	3.9	64	116	1.8	8.9	3.0	0.0	0.1	29.0	91	0.4	190	705
01-01-2018 07:00	1.26	0.6	127.9	37.2	165.5	27.8	5.8	75	125	2.2	0.2	3.7	0.0	0.1	29.0	90	0.5	207	705
01-01-2018 08:00	1.60	1.6	158.8	38.0	197.3	32.6	7.1	83	151	2.8	0.2	4.9	0.1	0.1	29.0	84	0.6	218	706
01-01-2018 09:00	1.14	10.4	62.1	52.5	114.7	54.2	4.5	69	126	3.3	0.2	5.0	0.0	0.1	29.0	72	1.0	230	707
01-01-2018 10:00	1.03	13.5	28.2	42.9	71.2	47.9	3.7	46	117	3.0	14.5	4.3	0.1	0.2	29.0	62	1.2	243	707
01-01-2018 11:00	1.03	18.0	18.3	43.4	61.8	28.9	6.3	38	103	2.7	9.6	4.0	0.0	0.2	29.0	52	1.6	169	707
01-01-2018 12:00	1.03	18.2	21.7	54.2	75.9	28.3	6.3	38	95	2.5	17.4	4.7	0.0	0.1	29.0	47	1.7	192	706
01-01-2018 13:00	1.03	31.2	8.1	49.1	57.2	31.5	4.5	44	85	2.3	4.7	5.1	0.9	0.1	29.0	44	1.9	256	705
01-01-2018 14:00	0.92	45.3	4.8	41.0	45.8	34.3	2.9	43	92	2.2	8.9	4.5	0.0	0.1	29.0	42	1.4	222	704
01-01-2018 15:00	1.14	44.1	5.2	54.2	59.3	32.9	2.6	53	111	2.2	8.5	3.6	0.3	0.1	29.0	40	1.3	157	703
01-01-2018 16:00	1.14	41.4	5.2	56.4	61.6	37.7	3.1	44	104	2.3	0.2	3.7	1.5	0.1	29.0	40	1.5	220	702
01-01-2018 17:00	1.03	39.6	6.7	49.1	55.8	37.3	4.2	43	110	2.2	10.9	3.6	0.8	0.3	29.0	41	1.4	252	702

Figure 1.1 : Time series dataset of Air pollution data

Visualize the Time series data with more clarity by decomposing it into a trend, seasonal and residual.

1. Level - shows the average values in the series of hourly-based pollution dataset
2. Trend - shows increasing and dipping values in the series
3. Seasonal - shows the repeating patterns in the season
4. Residual - shows the data points varying randomly in the series

So, all the pollutants and meteorological parameters are plotted with respect to a timeline to examine how all the pollutant concentrations and meteorological parameters behave and change with time. The following Algorithm is an *approach to* Normalize the Time Series dataset during Data Pre-processing.

Algorithm 1.2 for Adaptive Normalization approach for Time series data set

Step 1: Loading the data file Silk board hourly data from the excel file format

Step 2: Remove all special characters in the dataset by NAN Values (Not a Number)

main_file = pd. read excel (path+file, sheet_name='Table1', nan_values='-')

Step 3: Finding the Header Names and counting the number of Headers in the excel file

column name = [var for var in main_file. columns]

print ('Total number of Header: ',

Step 4: Check if any column has NAN or null values

check null = [var for var in Data. Columns if Data[var]. isna(). sum()]

Step 5: Replacing the NAN values to their two nearest data points using the Linear Interpolation method and Rechecking for any nan values.

Step 6: Checking if any Outlier data exists using Statistical Models

Step 7: Observing the influence of other pollutants on PM_{10} and $PM_{2.5}$

Step 8: Variation of PM_{10} and $PM_{2.5}$ with respect to pollutants and metrological parameters with a timeline

Step 9: Visualize the Time series data with more clarity by decomposing it into the trend, seasonal and residual.

IV. *RESULTS AND DISCUSSION*

Step 1 : Determines loading a sample file in excel format. The hourly-based data for Silk Board Station is shown in Figure 1.1.

Step 2 : Remove all NAN values in the data series.

Step 3 : Count the number of attributes and a number of instances in each column.

The header names are: 20

['Time', 'CO\nmg/m³', 'Ozone\nµg/m³', 'NO\nµg/m³', 'NO2\nµg/m³', 'NOx\nµg/m³', 'NH3\nµg/m³', 'SO2\nµg/m³', 'PM2.5\nµg/m³', 'PM10\nµg/m³', 'BEN\nµg/m³', 'TOL\nµg/m³', 'm, p-XYL\n µg/m³', 'o-XYL\nµg/m³', 'Et-BEN\nµg/m³', 'AT\n°C', 'RH\n%', 'WS\nm/s', 'WD\ndeg.', 'BP\nmmHg']

Figure 1.2: Header names in Silk Board Hourly based time series data

```
CO\nmg/m³                910
Ozone\nµg/m³             590
NO\nµg/m³                675
NO2\nµg/m³               676
NOx\nµg/m³               676
NH3\nµg/m³               659
SO2\nµg/m³               624
PM2.5\nµg/m³             632
PM10\nµg/m³              660
BEN\nµg/m³               556
TOL\nµg/m³               556
m,p-XYL\n   µg/m³        557
o-XYL\nµg/m³             557
Et-BEN\nµg/m³            561
AT\n°C                   554
RH\n%                    814
WS\nm/s                  541
WD\ndeg.                 542
BP\nmmHg                1541
dtype: int64
```

Figure 1.3: Number of NAN Instances in each Column of the Silk Board Hourly based time Step 6: series data

Step 4: Checking any column as NAN values

```
Time                     0
CO\nmg/m³                0
Ozone\nµg/m³             0
NO\nµg/m³                0
NO2\nµg/m³               0
NOx\nµg/m³               0
NH3\nµg/m³               0
SO2\nµg/m³               0
PM2.5\nµg/m³             0
PM10\nµg/m³              0
BEN\nµg/m³               0
TOL\nµg/m³               0
m,p-XYL\n   µg/m³        0
o-XYL\nµg/m³             0
Et-BEN\nµg/m³            0
AT\n°C                   0
RH\n%                    0
WS\nm/s                  0
WD\ndeg.                 0
BP\nmmHg                 0
dtype: int64
```

Figure 1.4: Checking any column having NAN Values

Step 5: Fill the NAN values with the two nearest data points using Linear Interpolation for all columns.

mean	1.003793	32.186991	28.363638	34.648789	63.084582
std	0.696149	26.865981	43.555145	19.798322	53.335804
min	0.011449	0.196195	0.085856	0.940247	2.609700
25%	0.572462	11.771670	3.699976	19.933230	28.135300
50%	0.858693	26.290070	10.670650	30.652040	47.742200
75%	1.259416	41.200860	33.974370	44.379640	81.787250
max	10.659240	236.414400	625.398100	166.987800	756.521000

Figure 1.5: A sample of data showing all NAN values filled with the two nearest data points using Linear Interpolation

Checking for Outliers using Skewed and Gaussian Distribution for all pollutant Concentrations

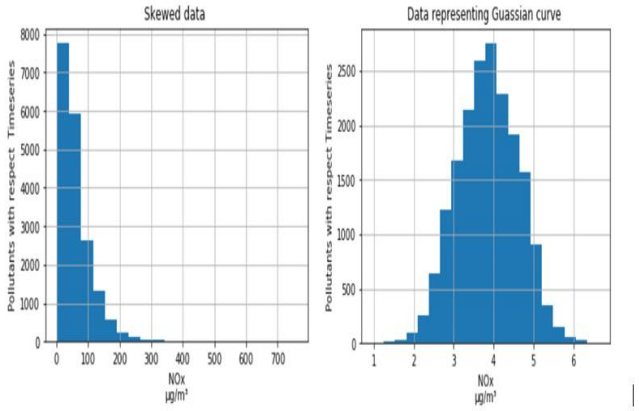

Figure 1.6: Skewed and Gaussian Representation of NOx Pollutant Concentration

Step 7: The above Figure 1.6 is the influence of pollutant concentrations of SO_2, NO, NH_3, Ozone, CO, and NOx on PM_{10} and $PM_{2.5}$.

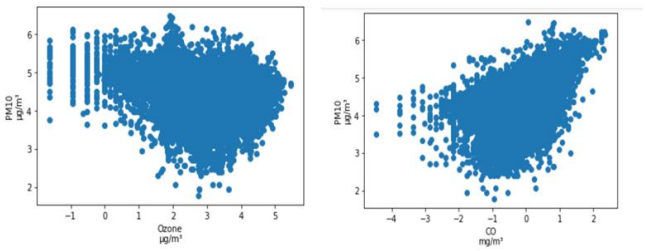

Figure 1.7: Influence of CO and Ozone on PM_{10}

Step 8: Variation of PM_{10} and $PM_{2.5}$ with respect to pollutants and metrological parameters with a timeline

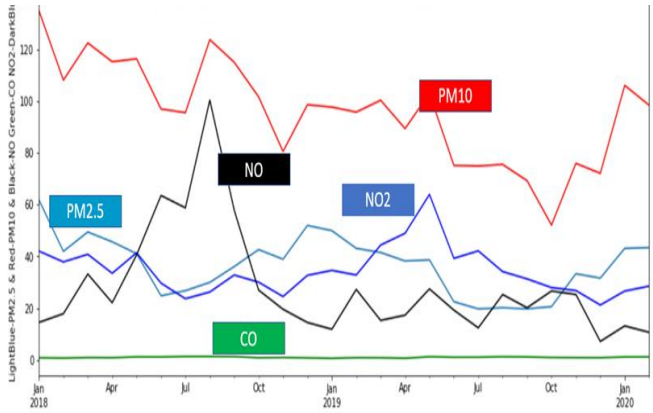

Figure 1.8 Plotting the pollutants concentration of $PM_{2.5}$ and PM_{10} with Pollutants NO, CO, and NO2 with respect to the timeline

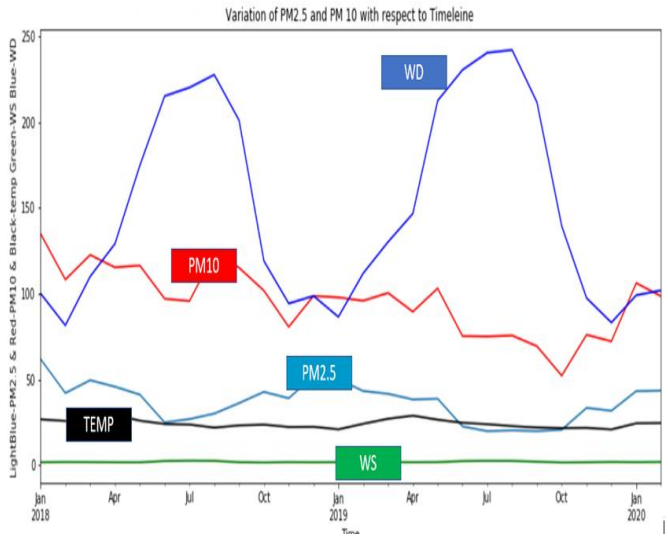

Figure 1.9: Plotting the pollutants concentration of $PM_{2.5}$ and PM_{10} with Meteorological Parameters Temp, WS, and WD with respect to the timeline

Step 9: Visualizing the Time series data with more clarity by decomposing into the trend, seasonal and residual.

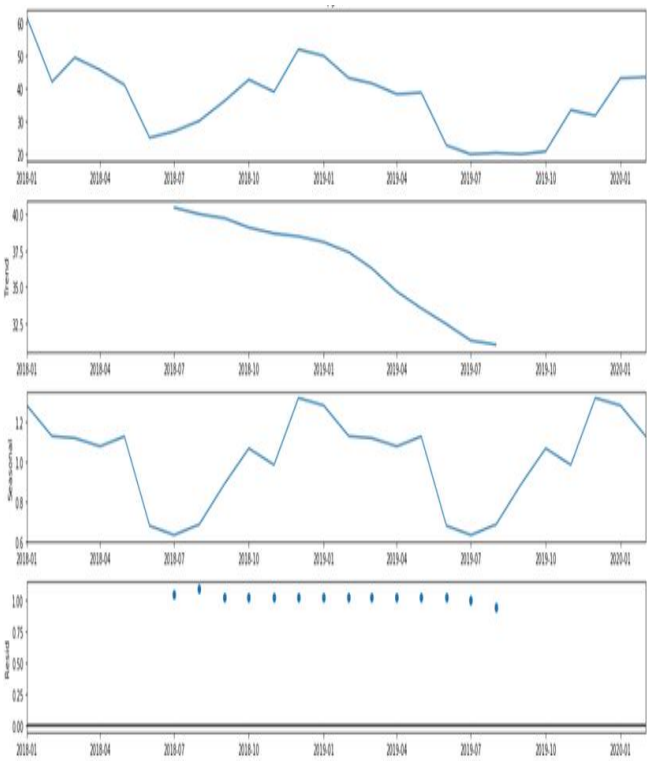

Figure 1.10: Trend, Seasonal and Residual Analysis of $PM_{2.5}$

The Actual graph represents the average values of $PM_{2.5}$ in the series. The Trend Analysis in $PM_{2.5}$ Pollutant Concentration depicts that it slightly starts increasing in July month and start dipping during January and April and again increases in July. The Seasonal Analysis of $PM_{2.5}$ Pollutant Concentration depicts the variation of the concentration during every month.The Residue Analysis in $PM_{2.5}$ represents the noise that is a random variation in the series.

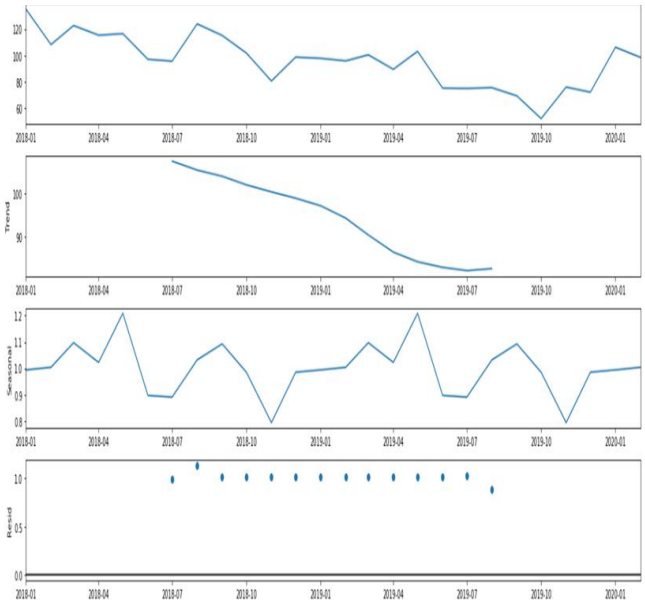

Figure 1.11: Trend, Seasonal and Residual Analysis of PM_{10}

The Actual graph represents the average values of PM_{10} in the series. The Trend Analysis in PM_{10} Pollutant Concentration depicts that it starts increasing slightly in July and starts dipping during the months of January and April and again increases in the month of July. The Seasonal Analysis of PM_{10} Pollutant Concentration depicts the variation of the concentration during every month. The Residue Analysis in PM_{10} represents the noise, which is a random variation in the series.

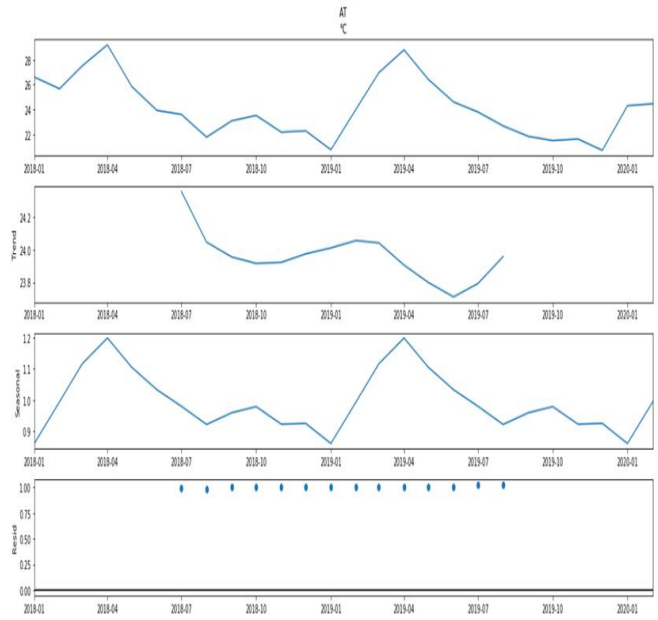

Figure 1.12: Trend, Seasonal and Residual Analysis of AT (Atmosphere)

series. The Trend Analysis in AT Pollutant Concentration depicts that it slightly starts decreasing from July month October to January and slightly increases from October to January to March and again dips. The Seasonal Analysis of AT Pollutant Concentration depicts the variation of the concentration every month. The Residue Analysis in AT represents the noise, which is a random variation in the series.

V. CONCLUSION

The proposed framework employs statistical data preprocessing techniques to improve the quality, consistency, and reliability of air pollution datasets before predictive modelling. Although the framework is primarily designed for time-series air pollution data, the underlying preprocessing procedures can be adapted and applied to other structured datasets with similar data-quality issues, such as missing values, outliers, inconsistencies, and non-normal distributions. By systematically identifying and treating these data-quality problems, the proposed framework can provide a more reliable input dataset for statistical and machine-learning models, thereby contributing to improved predictive accuracy and reduced prediction error across different application domains.

VI. ACKNOWLEDGMENTS

*"This article is derived from the author's **Dr Asha N** PhD thesis titled 'Predictive Analysis, Forecasting, and Control of Air Pollution Using Data Analytic Techniques' submitted to Mother Teresa Women's University and available on Shodhganga.",**but not published elsewhere.***

VII. REFERENCES

[1] Wang, L., Zeng, Y., & Chen, T," Back propagation neural network with adaptive differential evolution algorithm for time series forecasting", Expert Systems with Applications, Volume 42, pp.855–863, http://doi.org/10.1016/j.eswa.2014.08.018.

[2] Hylleberg, S, "Modelling seasonality", (1st edition.) Oxford: Oxford University Press.

[3] Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., … Winkler, R, "The accuracy of extrapolation (time series) methods: Results of a forecasting competition", Journal of Forecasting, Volume 1(2), pp. 111–153. DOI:10.1002/for.3980010202.

[4] Fernández Palomares, N., Álvarez de Prado, L., Menéndez García, L. A., Fernández López, D., Buján, S., & Bernardo Sánchez, A. (2026). "A Reproducible QA/QC, Imputation and Robust-Series Workflow for Air-Quality Monitoring Time Series", Applied Sciences, 16*(7), 3396*. https://doi.org/10.3390/app16073396

[5] Khatri P, Shakya KS, Kumar P. "A probabilistic framework for identifying anomalies in urban air quality

Oct;31(49):59534-59570. doi: 10.1007/s11356-024-35006-x. Epub 2024 Oct 2. PMID: 39358655.

[6] Jayaraman, S., T, N., S, A. *et al.* "Enhancing urban air quality prediction using time-based-spatial forecasting framework". *Sci Rep* **15**, 4139 (2025). https://doi.org/10.1038/s41598-024-83248-z

[7] Ma Z, Luo W, Jiang J, Wang B, Ma Z, et al. (2023) "Spatial and temporal characteristics analysis and prediction model of PM2.5 concentration based on SpatioTemporal-Informer model", PLOS ONE 18(6): e0287423. https://doi.org/10.1371/journal.pone.0287423

[8] B.S. Freeman, G. Taylor, B. Gharabaghi, J. "Thé Forecasting air quality time series using Deep learning", J, Air Waste Manage. Association. 68 (8) (2018) 866–886.

[9] C. C. Aggarwal and P. S. Yu, "Outlier detection for high dimensional data", ACM SIGMOD International Conference on Management of data (ACM Sigmod 2001), ACM, June 2001, pp. 37- 46, DOI: l0.1145/375663.375668.

[10] X. Zhu and X. Wu, "Class noise vs. attribute noise: a quantitative study of their impacts", Artificial Intelligence Review, Volume. 22, no 3, November 2004, pp. 177-210, DOI: http://doi.org/10.1 007/s 10462-004-0751-8

[11] S. Chen, W. Wang, and H. van Zuylen, "A comparison of outlier detection algorithms for ITS data", Expert Systems with Applications, Volume. 37, no 2, March 2010, pp. 1169-1178, DOI: http://doi.org/10. 10 16/j. eswa. 2009. 06. 008.

[12] Antonio Javier talon Ballesteros, Jose Cristobal Riquelme Santos, "Deleting or Keeping Outliers for Classifier Training?", Institute of Electrical and Electronics Engineers (IEEE),2015, DOI:c10.1109/nabic.2014.6921892

[13] Kaur, K., Garg, A, "Comparative study of outlier detection algorithms", IJCA, Volume 147(9), (2016).

[14] J. Brownlee, "Time series forecasting as supervised learning", Machine Learning", Mastery,2020, https://machinelearningmastery.com/time-seriesforecasting-supervised-learning/

[15] H. P. Vinutha, B. Poornima and B. M. Sagar "Detection of Outliers Using Interquartile Range Technique from Intrusion Dataset", Information and Decision Sciences, Advances in Intelligent Systems and Computing 701, 2018, DOI:10.1007/978-981-10-7563-6_53

[16] Mia Hubert and Stephan Van der Veeken "Outlier detection of skewed data", Article in Journal of Chemometrics, March 2008, DOI: 10.1002/cem.1123