International Journal of Engineering Trends and Technology
ISSN: 2231 – 5381 /doi:10.14445/22315381/IJETT-V69I5P227

Volume 69 Issue 5, 196-207, May 2021


# Data Preprocessing Techniques for Handling Time Series data for Environmental Science Studies

Ebin Antony[1], N S Sreekanth[2], R K Sunil Kumar [3], Nishanth T[4]

[1]*Department of Information Technology, Kannur University, Kannur, Kerala, India*

[2]*C-DAC Bangalore, #68 Electronics City Bangalore, Karnataka, India*

[3]*Department of Information Technology, Kannur University, Kannur, Kerala, India & Visiting Associate, Inter-University Centre for Astronomy and Astrophysics, Pune, India*

[4]*Department of Physics, Sree Krishna College Guruvayur, Affiliated to University of Calicut, Kerala, India*

[1]ebinantony21@gmail.com, [2]nssreekanth@gmail.com, [3]seuron74@gmail.com

***Abstract*** *- The present article discusses various preprocessing techniques suitable for dealing with time series data for environmental science-related studies. The errors or noises due to electronic sensor fault, fault in the communication channel, etc., are considered here. Such errors or glitches that occur during the data acquisition or transmission phases need to be eliminated before it fed to the forecasting or classification systems. Computationally simple and efficient techniques are discussed here so that they can even be adopted for a hard real-time system environment. While adopting these techniques, we may also end up with some of the real genuine values, which may consider as an outlier. A special indicator function, the moving Inter Quartile Range (MIQR) algorithm, is proposed to overcome such special cases.*

**Keywords —** *Time Series Analysis, Data Preprocessing, Moving Inter Quartile Range, Environmental Science, Data Science*

## I. INTRODUCTION

Time series data and its modeling are very popular among researchers who work in various domains like atmospheric science, financial sector, engineering science, etc. Any data or observations ( behavior); for a given subject at different time intervals, i.e., it may be equally spaced as in the case of metrics or unequally spaced as in the case of events, can be considered as time-series data. Data collected for health monitoring in intensive care units of a patient or collecting the environmental parameters like humidity, temperature, wind speed, presence of various elements in the atmosphere for weather prediction or air quality prediction [1, 2, 3] are some of the best examples for equally spaced time series data. Collecting the event-based logs and traces by control systems or server applications can be considered as unequally spaced time series data [4, 5, 6]. Such data collected from the real world can be used for modeling the system behavior, and it allows us to predict the future from the past. Weather modeling, reliability prediction in control systems, stock market prediction are some of the prominent applications which work based on time series modeling [7, 8, 9]. The quality of such prediction depends on the type of modeling techniques or algorithms used, and it also has a dependency on the preprocessing techniques adopted for cleaning the data collected from the real world [10, 11].

Any data collected through electronic sensors are prone to errors; this error may occur during data capturing, recording, or transmitting phases [12]. Especially the researchers in atmospheric science and environmental science heavily depend on electronic sensor-based data for their research work [13, 14. 15]. It may be difficult to notice the glitches or errors in the data collected over a while through electronic sensors deployed in the field. These errors or glitches may be in the form of missing values, noise or outliers, and seasonal variations. Unless these glitches are removed, this may contribute to unexpected variations in the results. Especially in a fully automated control system environment, interconnected with various sensors, to provide warning during adverse conditions, these faulty outlier values may generate false alarms. The usage of outlier/noise removal is very crucial for real-time systems which trigger certain action or warning by sensing the environmental parameters. The present article discusses various data preprocessing techniques and experimentally provides sufficient evidence for using some of the standard preprocessing practices while dealing with time-series data, especially in environmental science. The work carried out by Reshmi CT. et.al.[16], "Temporal Changes in Air Quality during a Festival Season in Kannur, India" is the basis of this study. The authors studied the difference in air quality during the fireworks period during the Vishu festival in the Kannur district of Kerala. The ambient concentration of PM10, $NO_2$, $O_3$, and NO were observed during the festival period for “four consecutive years in 2015, 2016, 2017, and 2018 in Kannur” [16].

In the proposed study, we only considered the concentration of surface $O_3$ to discuss some of the standard preprocessing techniques and also experimentally prove how to overcome any glitches/ errors in the reading due to error in the physical sensors. The study of the concentration of surface $O_3$ is very popular among environmental scientists, as the ozone concentration has lots of direct impact in agriculture [17,18,19,20,21], UV light protection, an aerobic process for the biological

treatment plant, etc. [22]. The most commonly available $O_3$ sensors work based on one of the following principles; they are absorption spectroscopy, photoacoustic, photo reductive, photostimulated, metal oxides, electrochemical methods. The study carried out by Michael David et al. discusses the specific performance-related problems, limitations of all these types of ozone sensors [23]. In the present experiment, the data was collected using an absorption spectroscopy-based, O342e UV Photometric Ozone Analyzer. The datasheet of the analyzer reports standard zero drift 1 ppb / 7days under the working temperature of range $0^oC$ to $35^oC$ [24]. The system may also behave unexpectedly beyond the specified operating temperature. Such conditions may cause to generate the values as outliers or noise. The present article discusses some of the standard practices used for data preprocessing to avoid outliers, noises, or missing values. A method is also proposed to avoid removing some of the genuine observations as outliers that may occur due to some specific phenomenon.

## II. LITERATURE REVIEW

Data cleaning and preprocessing is one of the interesting areas for data scientists. The preprocessing techniques they use vary concerning the problem under consideration. However, the error due to the data collected at the source should be deal with in the almost same way, irrespective of the problem statement. Kyriakidis (2009) et al. provides a comprehensive report on various techniques in data preprocessing, focuses on preparing higher data quality for data science applications[25]. This includes handling missing data, managing outliers, de-trending, and smoothing data. Kin Seng Lei (2010) et al. introduced a strategy for preprocessing, missing observational information using various imputation techniques for API (Air Pollution Index) forecast employing the Adaptive Neuro-Fuzzy Inference System (ANFIS). The forecast execution after data preprocessing is compared with the without preprocessing case, and the Root Mean Square Error (RMSE) shows the viability of the preprocessing techniques for API forecast against nine years of estimated data in Macau City [26].

Christophe Paoli (2010) et al. presents the usage of artificial neural networks (ANNs) in the sustainable strength space [27]. They used a Multi-Layer Perceptron (MLP) and a specially appointed time series preprocessing to build a strategy during the daily forecast of worldwide solar radiation on a horizontal surface. Advanced MLP presents comparable or better expectations to conventional and reference strategies like ARIMA methods, Bayesian hypotheses, Markov networks, and K-Nearest-neighbors. They tracked down that the proposed data preprocessing approach could fundamentally decrease forecasting errors contrasted with conventional forecasting strategies. S. Minu (2016) et al. review on estimating soil properties from hyperspectral air and satellite data, as well as preprocessing procedures used to overcome air attenuation and low signal-to-noise ratios, and to distinguish soil properties [28].

Juneja (2019) et al. reports various aspects to enhance the quality of original data. By exploring the processes for integration, noise filter, removal of bad data, and data transformation/normalization. In this study, they also proposed a preprocessing system to predict the nature of data in climate monitoring and prediction for identifying the dangerous atmospheric deviation boundaries and raises alerts for early warning to clients and researchers [29]. Khongsrabut (2019) et al. focused on finding the anomaly in the water utilization time series data. They tested two strategies: (1) Autoregressive Integrated Moving Average (ARIMA) modeling and (2) Median Absolute Deviation (MAD). A clustering strategy (K-Mean) is used as a pre-processing step to distinguish the correct parameters of the two defect detection strategies. The results showed that ARIMA and MAD performed well in the water use time-series data along with the specified parameter values obtained from K-Mean [30].

D. Andresic (2019) et al. two huge cosmic time series data were selected from the BRITE (BRIght Target Explorer) and project named Kepler K2 and explored feasible methods for searching for hidden periods and grouping them [31]. For these data collections were so huge that artificial neural networks must be used, that requires some data preprocessing. Therefore, the hidden periods of the galactic time series bring a concise outline of possible answers to search using absolute artificial neural networks or including extra traditional analytical methods, mainly from a data preprocessing and its visualization perspective.

## III. DATA PREPROCESSING

Preprocessing is the first step in data science and machine learning for classification and information retrieval problems. The raw data collected from the real world will go through the preprocessing techniques before it processes with any machine learning or data mining algorithms [32]. There is three kinds of data preprocessing methods: data cleaning, data transformation, and data reduction [33]. The data cleaning addresses the missing data and noise or outlier removal. Data transformation addresses normalizing the data, feature selection, domain-level transformation, etc. Data reduction deals with dimensionality reduction, attribute subset selection, etc. Among the three techniques, data transformation and data reduction are problem-dependent processes, but data cleaning is a must step to be followed across any data science problem. This paper discusses various popular preprocessing techniques used to handle the time series data. The Ozone $O_3$ concentration collected during four consecutive days, 13 April to 16 April of the year 2015 to 2018, is considered for this study.

Reshmi C.T et al., as reported in their work, collected data from “Kannur university campus (KUC) (11.9° N, 75.4° E), situated 15 km north from Kannur. The observational site is situated 1 km away from the National Highway (NH 17) and 6 km away from the Arabian Sea

and is surrounded by a densely populated residential area" [16]. Variations in concentrations of surface $O_3$ from 13 April to 16 April for the years 2015, 2016, 2017, and 2018 are considered for the proposed study. In addition to existing noise, random noises were introduced in the collected data to test the performance of various preprocessing techniques, and its performance is studied and reported as part of this work. The sample data set collected every thirty minutes duration of surface $O_3$ concentration of the year 2015 is shown below.

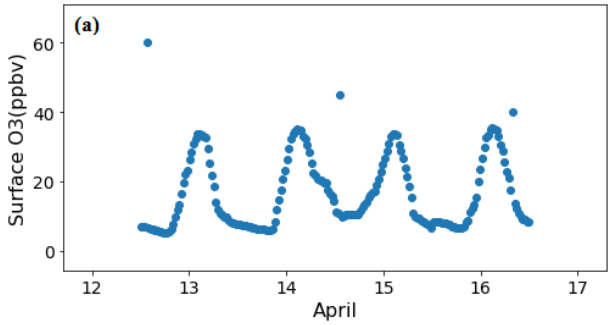
$O_3$ Sample Data Set for 2015:-

| Time (IST) | Aril 13 | Apr-14 | Apr-15 | Apr-16 |
|---|---|---|---|---|
| 0.5 | 7.16 | 7.65 | 11.22 | 8.54 |
| 1 | 6.99 | 7.62 | 10.82 | 8.48 |
| 1.5 | 6.88 | 7.5 | 10.22 | 8.35 |
| 2 | 6.69 | 7.39 | 10.02 | 8.3 |
| 2.5 | 6.56 | 7.3 | 10.2 | 8.23 |
| | ..... | | | |
| | ..... | | | |
| 22 | 9.22 | 19.7 | 8.34 | 9.77 |
| 22.5 | 8.45 | 17.5 | 8 | 9.27 |
| 23 | 8.19 | 16.4 | 7.68 | 9.01 |
| 23.5 | 7.96 | 15.8 | 7.26 | 8.54 |
| 24 | 7.8 | 14.3 | 6.73 | 8.27 |

This paper discusses the three important data cleaning aspects; noise or outlier removal, missing value handling, and smoothing. Fig. 1 (a) shows the surface $O_3$ with outlier and Fig. 1 (b) shows the surface $O_3$ with a missing value.

### *A. Detect and Remove the Outliers or Noise*

An outlier or a noise is an observation that differs or deviates from the overall pattern in a given data set [34]. The outlier may happen due to a fault in the data acquisition system, i.e., sensor node failure, error in the communication channel, etc. Such kind outliers will generally be in the form of one or more glitches for a shorter interval, or it can even be continuous if the sensing device or communication channel is faulty. Distinguishing such exceptions from the real data and figuring out how to manage them are on various artifacts. Understanding the information, data and their source, information on the logical area that addresses, and the scope of values expected by the related boundaries are some of the parameters considered for noise removal [35].



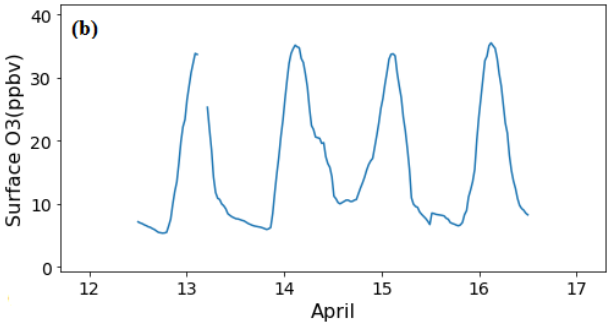




**Fig. 1 Surface $O_3$ (a) with outlier (b) with missing value**

In any forecasting application using time series data, it's essential to detect the presence of such outliers or noise and rectify it; otherwise, it can lead to wrong logical ends or forecasts. In this article, discuss two prominent techniques data scientist uses to detect such outliers or noise in the data under consideration. They are visual techniques and mathematical techniques.

### *B. Discover outliers with visual*

Simple visual techniques are employed to detect the glitches, i.e., outliers or noise in the given data. Instances of data under consideration can be plotted in specific graphical representation, and anomalies can be brought out. This article discusses two such methods, box plot and scatters plot.

#### *a) Box Plot*

Boxplot is a standard way of displaying data distribution based on five numbers summary ("minimum," first quartile (Q1), median (Q2), third quartile (Q3), and "maximum"). The data range between Q1 and Q3 is identified as Inter Quartile Range (IQR). Box plot will also help us to understand the general behavior of data, whether it is symmetrical or how tightly data is grouped or skewed [36, 37].

Box plot is one of the altogether less factual diagram strategies that decide exceptions. There may be one anomaly or numerous exceptions inside an informational index, which happenstance both underneath or more the base and most prominent data regards. The box plot gives outliers or obscure results by increasing the lesser and greater data values to a maximum of 1.5 times the inter-quartile range. Any information results that come outside of the greatest and least values known as outlier are not difficult to decide on a box plot chart.

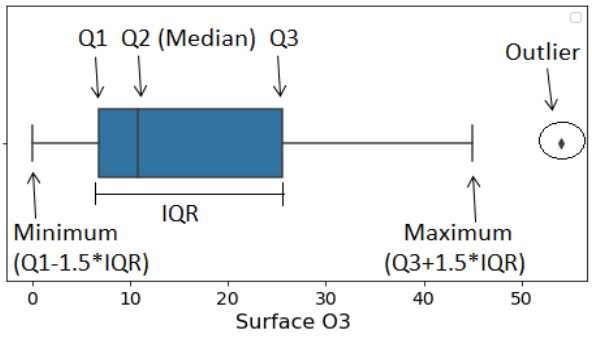



**Fig. 2 Outlier detection using Box plot**

Fig. 2 shows the outlier detection for surface $O_3$ using a box plot. The figure shows that one point above Q3+(1.5*IQR) as outliers because they do not include the other measurements. Here we analyzed the uni-variant outlier. We used the daily variation of surface $O_3$ to test the outlier.

***b) Scatter Plot***

A scatter graph is the most effortless method of the diagrammatic portrayal of bivariate information. One variable addresses along the X-axis, and the other variable addressed along the Y-axis. In time-series data, for the present study X-axis represent the time interval, and the Y-axis represents $O_3$ concentration at each time interval. The pair of points plots on the two-dimensional chart forms a scatter graph. The bearing of the progression of points shows the kind of relationship between the two given factors.

When there is a regression line in the scatter plot, we can recognize the outlier easily. Point or points far away from the regression line is generally considered as an outlier for the scatter plot. The outlier for a scatter and box plot is different from each other [38]. Fig. 3 shows the outlier detection using a scatter plot.

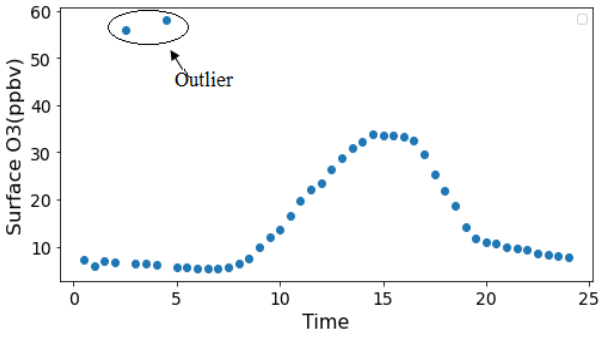
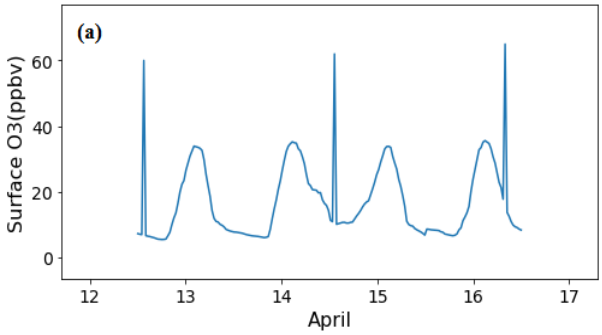



**Fig. 3 Outlier detection using scatter plot for $O_3$**

***C. Discover outliers with a mathematical function***

***a) Z – Score***

Z-Score, also called standard score, is an important method used for outlier detection. Z-score helps to understand how far the data value from the mean [39].



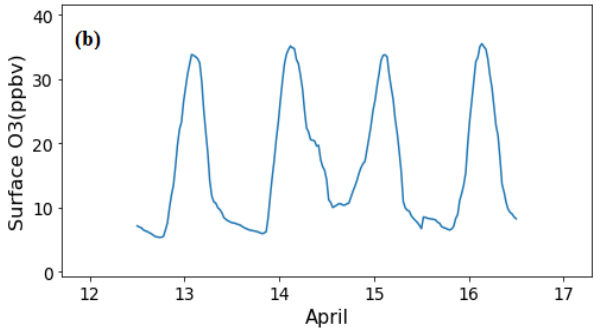




**Fig. 4 Surface $O_3$ (a) before removing outlier, (b) after removing outlier using z-score**

More specifically, the Z-score tells you how many standard deviations are there from a data point to the average. The Z-score is calculated as follows:

$$Z = \frac{x - \mu}{\sigma}$$

The standard score represents the Z, x depicts the observed value, $\mu$ indicates the sample's mean, and $\sigma$ shows the sample's standard deviation. Z-Score outliers are defined as, If the z-score of a data point is greater than 3, i.e., $|z| > 3$, then it is an outlier. Else it is a part of data. Fig. 4 (a) shows the surface $O_3$ before removing the outlier, and fig. 4 (b) shows the $O_3$ after removing the outlier using a z-score.

***b) IQR Score***

The Inter Quartile Range (IQR) is a statistical measure of the difference between $75^{th}$ and $25^{th}$ percent. It is represented by the formula $IQR = Q3 - Q1$ concerning the box plot discussed earlier [Figure 2]. The general rule of thumb for detecting outliers using IQR is any data point beyond Q1-1.5* IQR, i.e., minimum and Q3+1.5*IQR, i.e., maximum are considered as outliers, and they will be removed. Fig. 5 shows the surface $O_3$ after removing the outlier with Inter Quartile Range (IQR) [40].

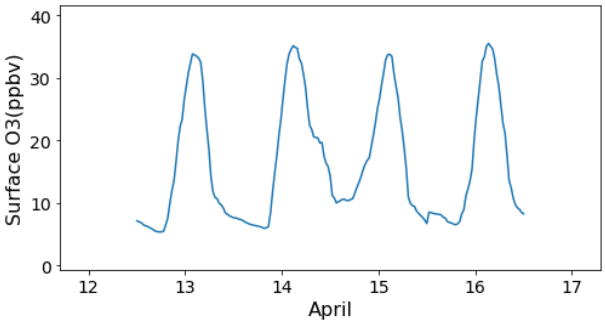



**Fig. 5 Surface $O_3$ after removing outlier using IQR (Resultant of Fig. 4 (a))**

### *D. Handling Missing Value*

Handling missing values is one of the important aspects of data preprocessing. During the data acquisition process, or transmission process from the sensor to the information processing node, it is quite often that we may encounter missing data problems. When the amount of data is limited, fixing the lost value problem by removing the corresponding data index can lead to the loss of important information, and it also results in a deficit of information available for training [41]. This article discusses two important missing value handling techniques, linear interpolation and KNN (K-Nearest Neighbors).

#### *a) Linear Interpolation*

Linear interpolation is an imputation procedure that accepts a linear relationship between data points and uses non-missing values from neighboring data points to register a missing data point [42]. The linear interpolation formula is:

$$y = y_1 + \frac{(x - x_1)(y_2 - y_1)}{x_2 - x_1}$$

Where $x_1$ and $y_1$ are the first coordinates, $x_2$ and $y_2$ are the second coordinates, x is the point to perform the interpolation, and y is the interpolated value. Fig. 6 (a) shows the surface $O_3$ data with missing values, and fig. 6 (b) shows the $O_3$ data after handling missing values using linear interpolation.

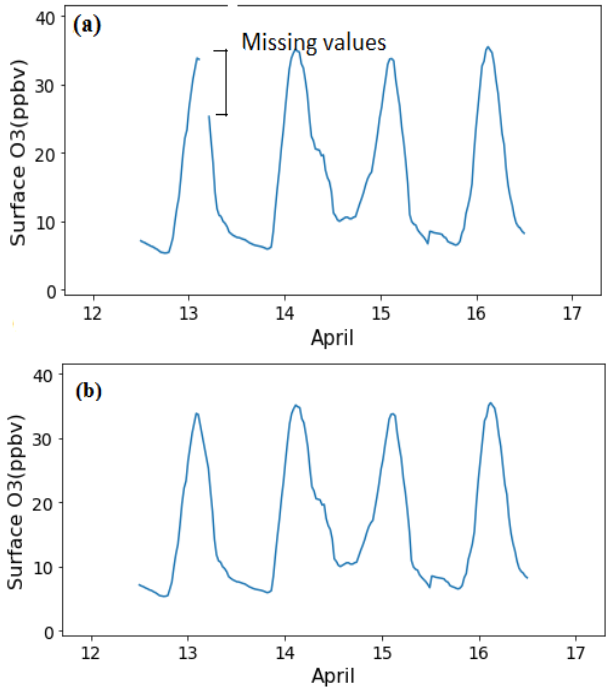



**Fig. 6 Surface $O_3$ (a) with missing values (b) after handling missing value using linear interpolation**

#### *b) KNN (K - Nearest Neighbors)*

Using the KNN method, a loss value calculated by the majority among its closest neighbors, K is the mean value of the "nearest neighbors," which is calculated as forecasting of a mathematical estimation of value, known as the "majority / mean rule" [43]. Fig. 7 represents the $O_3$ data after handling the missing value using KNN techniques. The idea of the KNN method is to identify the 'K' samples in the dataset as having the same or near space. We use these 'K' samples to determine the value of missing data points. The lost values of each sample are calculated using the average value of the 'K' neighbors found in the dataset.

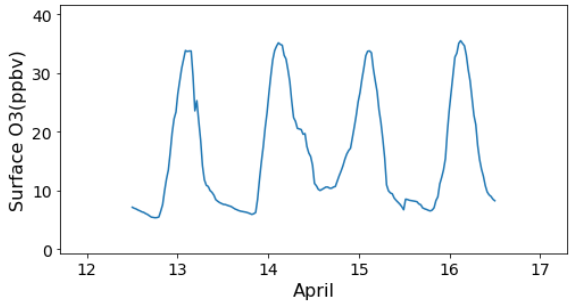



**Fig. 7 Surface $O_3$ data after handling missing value using KNN Imputation. (Resultant of figure 6 (a))**

### *E. Smoothing*

Data smoothing eliminates noise and concentrates on genuine patterns and trends. They distributed by a clear picture of the nature of the time series considered. In some cases, chronological variability is substantial, and they do not provide indicators of a trend or accuracy, which are parts of high significance for understanding the learning cycle. Smoothing eliminates periodicity and reveals delayed fluctuations within the data set [44]. Simple Moving Average (SMA) smoothing is the most common smoothing technique. The formula for SMA is:

$$SMA = \frac{A_1 + A_2 + \cdots + A_n}{n}$$

Where $A_n$ is the value at period $n$, and $n$ is the total number of periods.

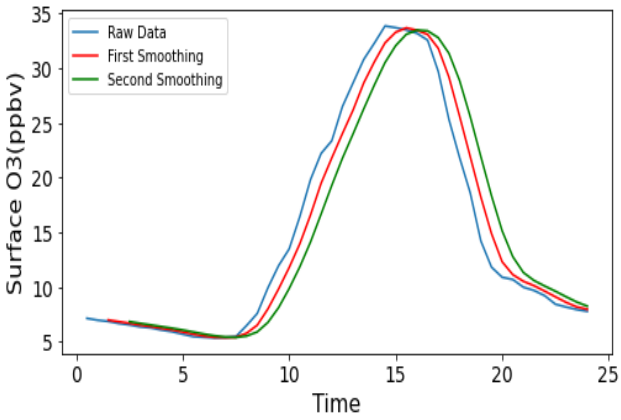



**Fig. 8 Daily variation of smoothed $O_3$ data**

Fig. 8 shows the daily variation of smoothed surface $O_3$ data. The blue line shows original data, and the red line shows the results of the first smoothing of original data, and the green line indicates second smoothing.

### *F. Considering the Unexpected Variations from the Normal Behaviors*

In some cases, especially in environmental-related studies, may have surprise observations that will have a considerable deviation from normal observations. This may be due to a specific phenomenon that occurred during the day, or seasonal variations, or any other similar instances. We have to ensure that such deviations from the normally observed patterns should not be eliminated by considering that as noise or outlier. In this article, discuss one such instance with a specific example observed by an environmental scientist. As reported by Reshmi CT et al.,[16] in their work, there is a deviation in the observed surface $O_3$ concentration in Vishu days from the normal days. Fig. 9 shows the line graph, where surface $O_3$ concentrations are deviated from the normal day's observation pattern during the Vishu festival day because of the fireworks.

These observations studied and reported that a 100% increase in the "$NO_2$ photolysis rate during the fireworks episode would lead to a 100% increase in surface ozone production"[16]. This is the best example to demonstrate to have a separate indicator function to account for such unexpected observations due to some specific phenomenon. Modeling such unseen variations is always challenging, but it is quite possible to model such variations if they are seasonal. Considering the observations of 4 years during the specific days, i.e., Vishu festival days, an additional indicator function can be fit to consider such variations. A continual learning-based system is always recommended for considering such genuine deviations from normal behavior. If such observations are beyond the threshold defined for outlier removal or noise removal, separate discounting parameters need to be considered as part of such systems to ensure reliability. In this particular case, on the Vishu festival day, the $O_3$ concentration level is shifted from the observations on a normal day; but it should not be eliminated as an outlier or noise. As stated earlier, outlier/ noise readings will be generally observed as a glitch for very shorter intervals, especially in time series data, unless the sensing device is faulty.

But these types of observations generally last for a considerable amount of time intervals. Recognizing such specific variations requires a domain-level understanding of the problem under consideration. A sufficient amount of data under that specific season should be treated separately to model such kinds of special occasions. The seasoning parameters should be derived for handling and treating such cases as special cases. Collecting data under some of the specific occasions in environmental studies have lots of importance. Hence defining the indicator functions as part of the system for dealing with such instances are very important.

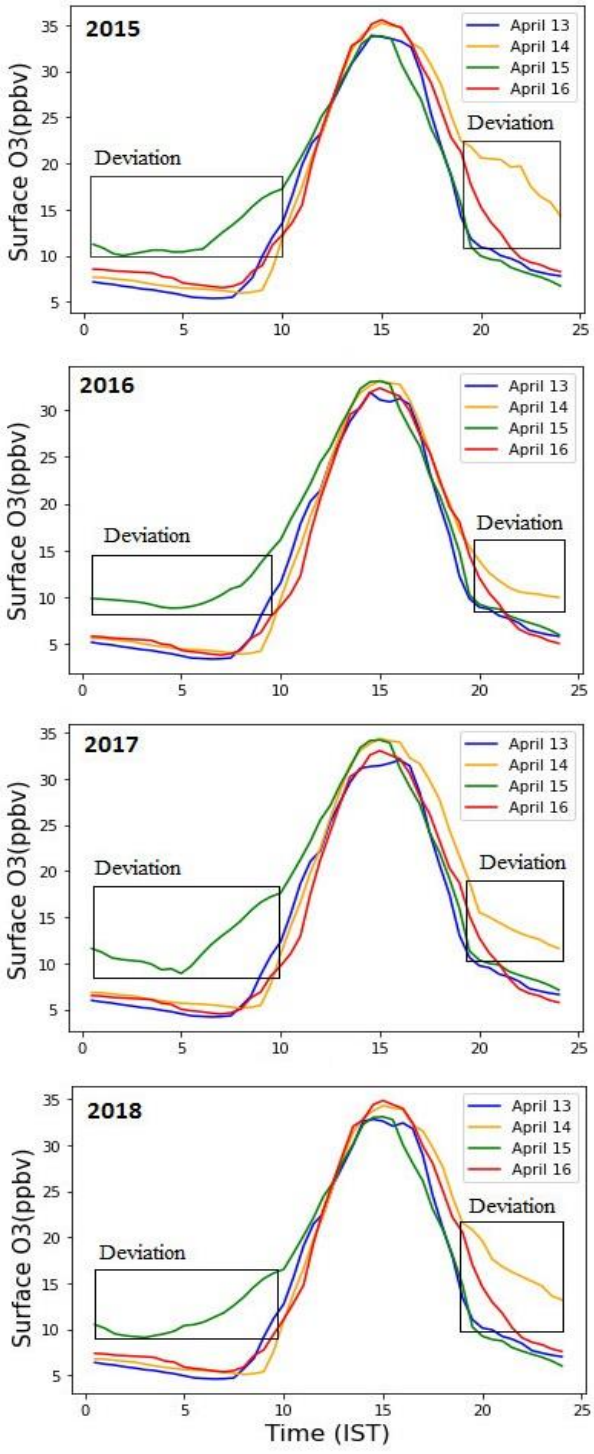



**Fig. 9 $O_3$ data during fireworks**

In the present case, propose a moving Inter Quartile Range (MIQR) algorithm to resolve such unexpected, genuine variations in the observations. In the MIQR method instated of considering the global IQR values, we locally set the IQR limit during every instance of the observations. Since the data under consideration is time-series data, generally, it is observed that there can be an accepted range of values that can be considered as probable observation in the succeeding intervals. MIQR algorithm is defined as follows.

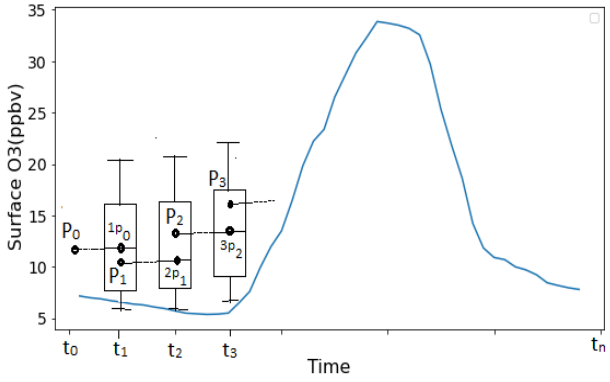



**Fig. 10 Moving Inter Quartile Range**

In Fig.10, The blue line in the line graph indicates the concentration of $O_3$ on non-Vishu day. On Vishu festival day at the time $t_0$, $P_0$ be the observed value which is away from the normal observation (mean). At this point, consider an IQR box where the current value $P_0$ will be considered as a projected mean value corresponding to $t_1$, and it is denoted as $^{1p}_0$ indicate that the projected mean at interval $t_1$ based on the observation at $t_0$. An IQR box will be constructed by keeping the point $^{1p}_0$ be the mean point at $t_1$. The upper and lower range is being set based on the global observation. At $t_1$, the actual observation is $P_1$, which is pure within the IQR box defined at $t_1$. Similarly, the new projected mean value at $t_2$ is denoted as $^{2p}_1$, i.e., the IQR mean value at time interval $t_2$ based on the observation at $t_1$. This process will be repeated for the rest of the observation. Hence the observed sequences $P_1$, $P_2$, $P_3$…. will be considered as genuine observations even though it is far from the normal observation. This ensures that only the abruptness in the reading will only be considered as outliers/noise. If any such values appear as part of real observation that will be neglected by the system, and it may be considered as a missing value. Linear interpolation or KNN Imputation method can be used to handle such missing value problems, as discussed earlier.

## IV. Results and Discussions

The experiments are conducted using the data form collected by Reshmi C.T et al.,[16]. Only Surface Ozone concentrations, i.e., $O_3$ concentrations, are considered for explaining the data preprocessing and smoothening concepts. In addition to existing noise, random noises are added to the signal using pseudo-random numbers and whose effects are studied. The outlier removal, missing value problems, and data smoothening are considered in this experiment to demonstrate the preprocessing requirement in environmental studies.

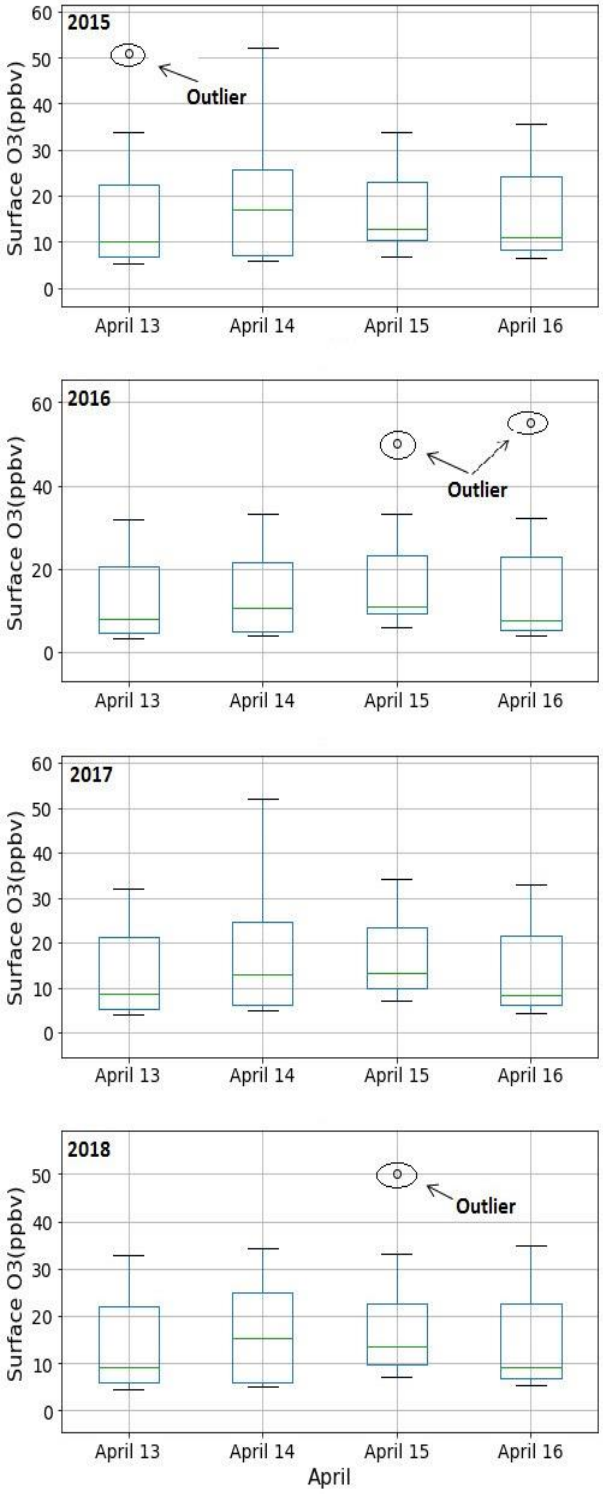



**Fig. 11 Box plot for detecting outlier in four years**

Fig. 11 shows the Box plot for detecting the outliers for four consecutive years, 2015 to 2018.Fig. 12 shows the Scatter plot for detecting the outliers for four consecutive years, 2015 to 2018.

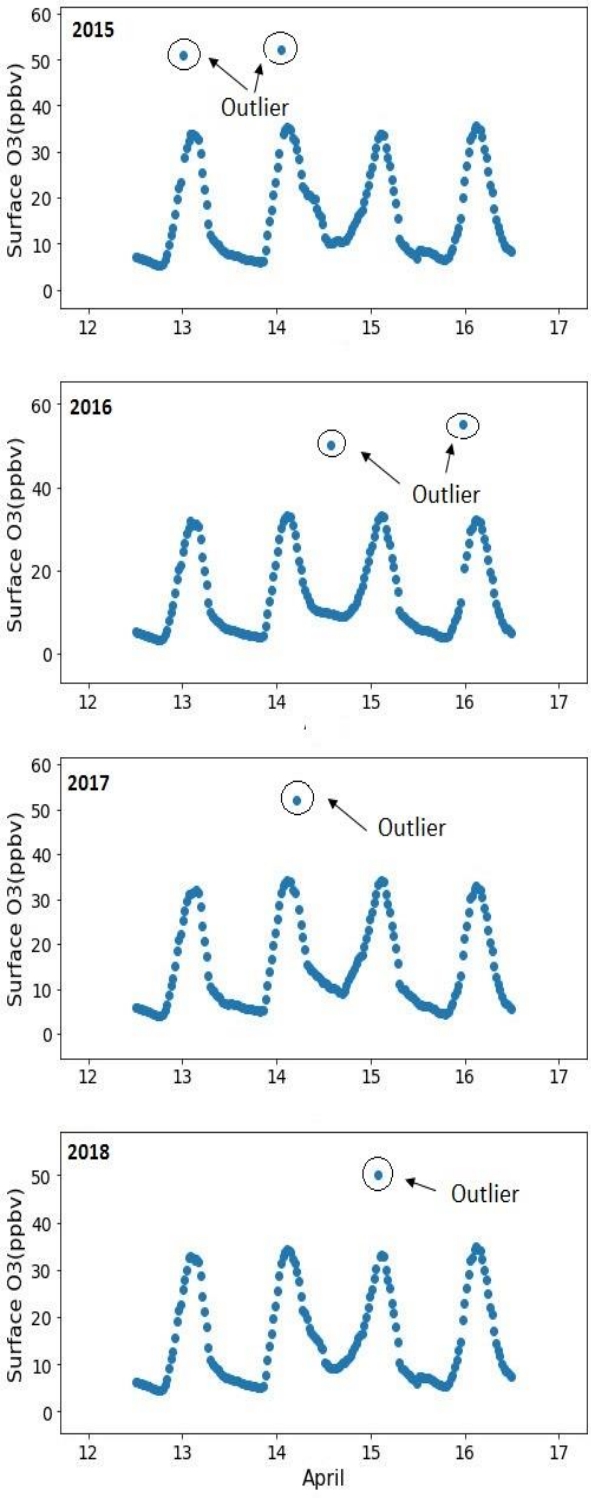



**Fig. 12 Scattered plot for detecting outlier in four years**

.

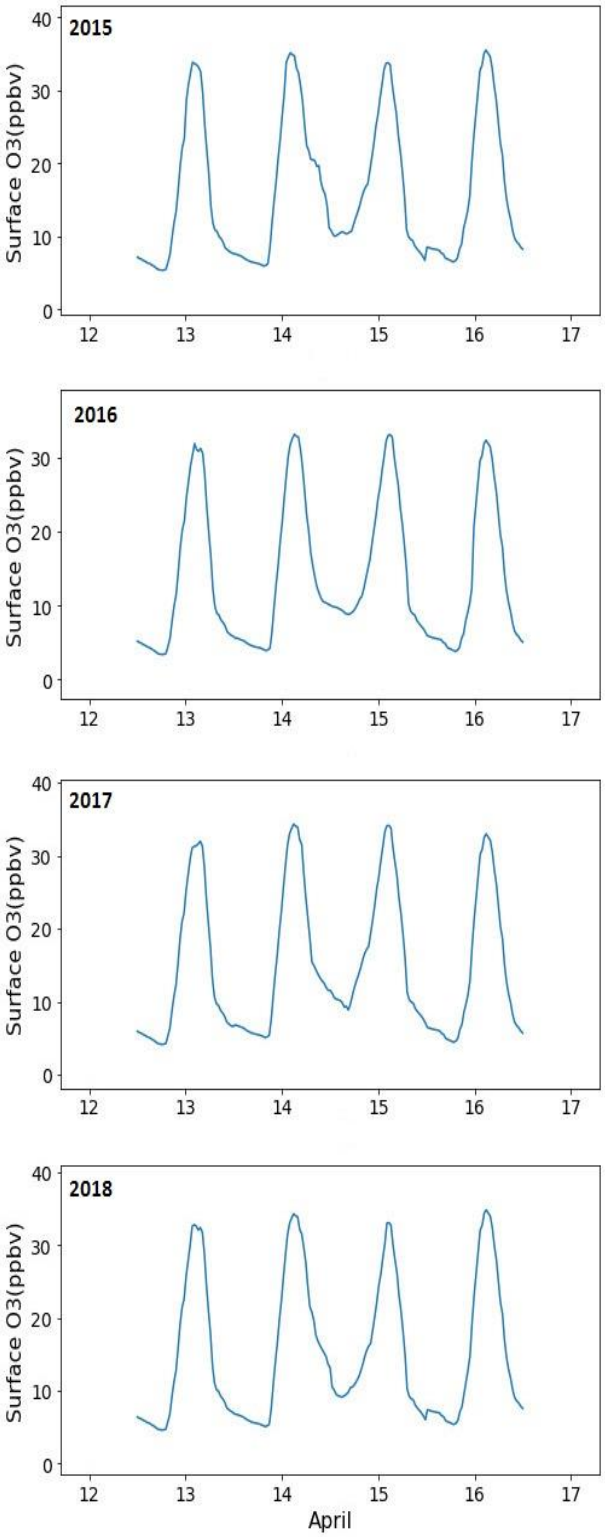



**Fig. 13 Outlier removed using Z Score**

The resultant of outlier removal methods Z-score and IQR for the four consecutive years are shown in fig. 13 and fig. 14 – resultant of fig 11 and 12.

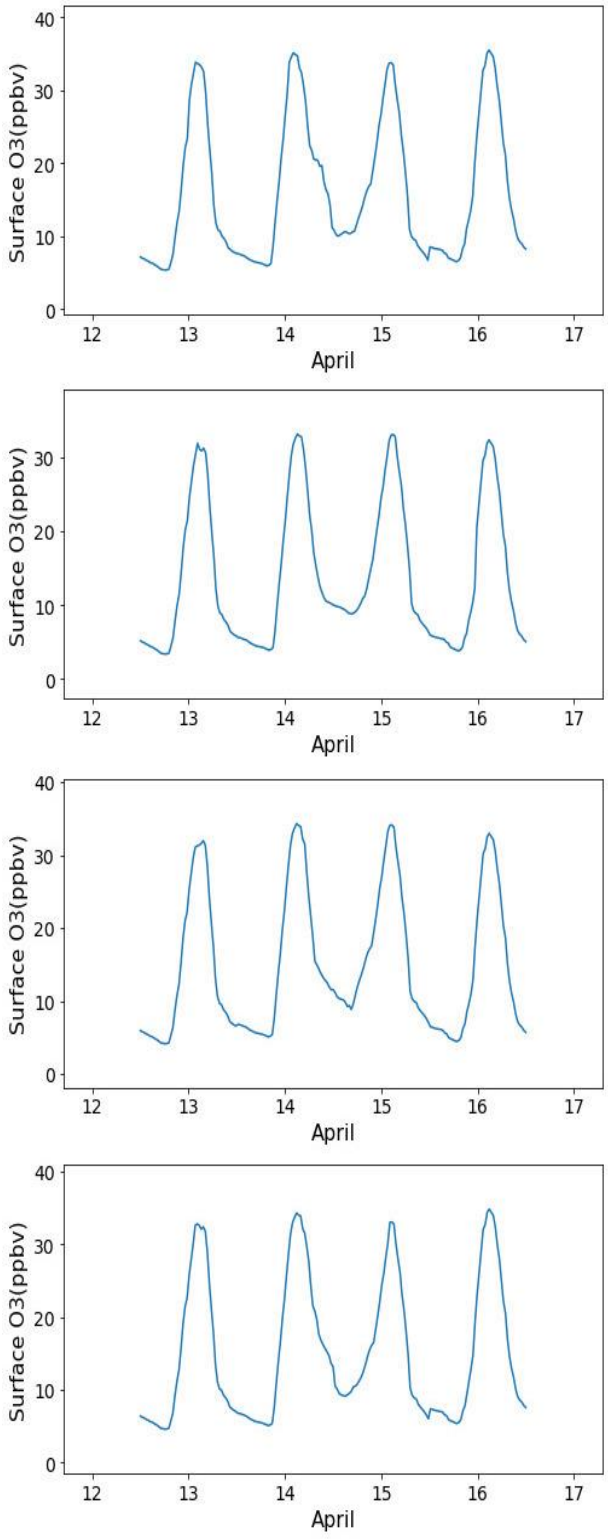



**Fig. 14 Outlier removed using IQR**

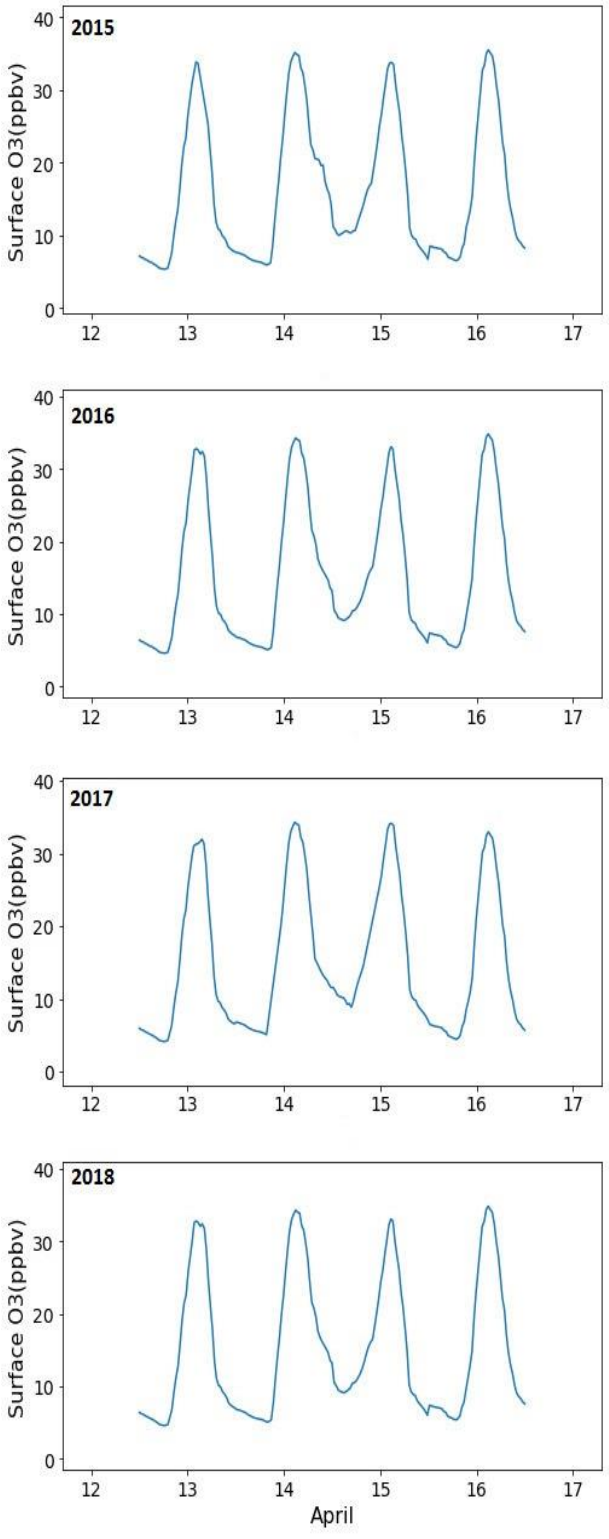



**Fig. 15 Handling missing value using linear interpolation**

The resultant of missing value handling methods, linear interpolation, and KNN Imputations algorithms for the four consecutive years are given in fig. 15 and fig. 16.

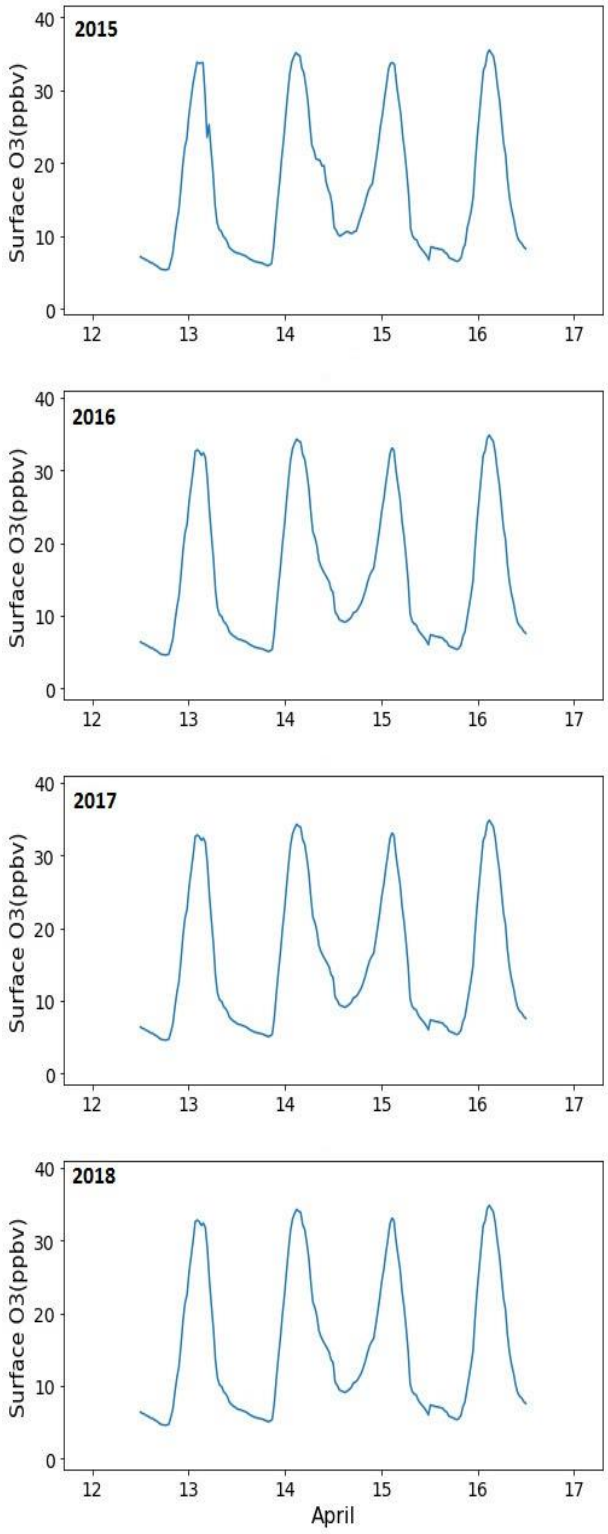



**Fig. 16 Handling missing value using KNNImputation**

The proposed moving IQR method to handle the special cases is also tested against all four years. The system effectively handled such deviations. The results are also validated by using some of the popular forecasting techniques Autoregressive Moving Average (ARMA), Autoregressive Integrated Moving Average (ARIMA), Seasonal Autoregressive Integrated Moving Average (SARIMA), etc. And the Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) are calculated for the noise added data and smoothed data. Here 75% of data is used for training, and 25% of data used for testing, i.e., forecasting. The results obtained are given in table 1. And bar chart for MSE and RMSE is shown in fig.17.

**TABLE 1**
**MSE and RMSE value for preprocessed data**

| | | ARMA | ARIMA | SARIMA |
|---|---|---|---|---|
| **Original data** | MSE | 24.483 | 22.951 | 23.012 |
| | RMSE | 4.948 | 4.79 | 4.797 |
| **Z-Score** | MSE | 13.025 | 11.327 | 13.971 |
| | RMSE | 3.609 | 3.365 | 3.737 |
| **IQR** | MSE | 13.025 | 11.327 | 13.971 |
| | RMSE | 3.609 | 3.365 | 3.737 |
| **Linear Interpolation** | MSE | 1.413 | 1.334 | 3.117 |
| | RMSE | 1.188 | 1.154 | 1.765 |
| **KNN Imputation** | MSE | 1.423 | 1.334 | 3.117 |
| | RMSE | 1.192 | 1.154 | 1.765 |

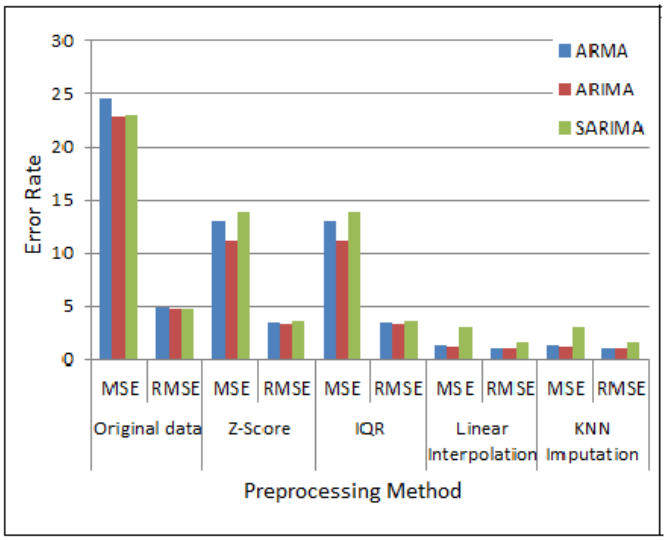



**Fig. 17 Bar chart for MSE and RMSE value**

## V. CONCLUSION

Real-world data tend to be incomplete, inconsistent, noisy, and missing. Data preprocessing is the major process in data science. Data preprocessing include data cleaning such as remove outlier or noisy data, handling of missing value using the standard techniques. The removal of genuine observations by considering them as noise or outlier is one of the issues discussed here. Proposed a moving interquartile range (MIQR) algorithm to resolve such unexpected, genuine variations in the observations. The current study emphasizes the importance of employing various preprocessing techniques to ensure the data quality collected through electronic sensors for prediction and forecasting in environmental science-related studies.

## REFERENCES


[1] Gocheva-Ilieva, Snezhana & Ivanov, A. & Voynikova, Desislava & Boyadzhiev, Doychin. (2013). Time series analysis and forecasting for air pollution in a small urban area: An SARIMA and factor analysis approach. Stochastic Environmental Research and Risk Assessment. 28. 1045-1060. 10.1007/s00477-013-0800-4.

[2] J. K. Sethi and M. Mittal, Analysis of Air Quality using Univariate and Multivariate Time Series Models 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, (2020) 823-827, doi: 10.1109/Confluence47617.2020.9058303.

[3] Wu, X.; Zhou, J.; Yu, H.; Liu, D.; Xie, K.; Chen, Y.; Hu, J.; Sun, H.; Xing, F. The Development of a Hybrid Wavelet-ARIMA-LSTM Model for Precipitation Amounts and Drought Analysis. Atmosphere 2021, 12, 74. https://doi.org/10.3390/atmos12010074

[4] P. M. T. Broersen and R. Bos, Estimating time-series models from irregularly spaced data, in IEEE Transactions on Instrumentation and Measurement, 55(4) (2006) 1124-1131, doi: 10.1109/TIM.2006.876389.

[5] Richard H. Jones, Time series analysis with unequally spaced data," Handbook of Statistics, Elsevier, 5(1985) 157-177, ISSN 0169-7161, ISBN 9780444876294, https://doi.org/10.1016/S0169-7161(85)05007-6.

[6] Jones, Richard H. Time Series Regression with Unequally Spaced Data. Journal of Applied Probability, 23(1986) 89–98. JSTOR, www.jstor.org/stable/3214345. Accessed 1 Apr. 2021.

[7] Shukla, A. & Garde, Yogesh & Jain, Ina. (2014). Forecast of weather parameters using time series data. Mausam. 65. 509-520.

[8] J. G. Harris and M. D. Skowronski, Automatic Speech Processing Methods for Bioacoustic Signal Analysis: A Case Study of Cross-Disciplinary Acoustic Research, 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, Toulouse, France, (2006)V-V, doi: 10.1109/ICASSP.2006.1661395.

[9] Sarwar, Umair & Muhammad, Masdi & Abdul Karim, Zainal Ambri. (2014). Time Series Method for Machine Performance Prediction Using Condition Monitoring Data. 10.13140/2.1.4520.3201.

[10] Kumar, Raghavendra & Kumar, Pardeep & Kumar, Yugal. (2020). Time Series Data Prediction using IoT and Machine Learning Technique. Procedia Computer Science. 167. 373-381. 10.1016/j.procs.2020.03.240.

[11] P. He, Y. Yuan, and G. Liu, Web Services Quality Prediction Based on Multivariate Time Series Analysis, 2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS), Beijing, China, (2018) 881-884, doi: 10.1109/ICSESS.2018.8663771.

[12] Dybko, A. Errors in Chemical Sensor Measurements. Sensors 2001, 1, 29-37. https://doi.org/10.3390/s10100029

[13] Rogulski, M.; Badyda, A. Investigation of Low-Cost and Optical Particulate Matter Sensors for Ambient Monitoring. Atmosphere 2020, 11, 1040. https://doi.org/10.3390/atmos11101040

[14] He, X.; Xu, X.; Zheng, Z. Optimal Band Analysis of a Space-Based Multispectral Sensor for Urban Air Pollutant Detection. Atmosphere 2019, 10, 631. https://doi.org/10.3390/atmos10100631

[15] Woodall, G.M.; Hoover, M.D.; Williams, R.; Benedict, K.; Harper, M.; Soo, J.-C.; Jarabek, A.M.; Stewart, M.J.; Brown, J.S.; Hulla, J.E.; Caudill, M.; Clements, A.L.; Kaufman, A.; Parker, A.J.; Keating, M.; Balshaw, D.; Garrahan, K.; Burton, L.; Batka, S.; Limaye, V.S.; Hakkinen, P.J.; Thompson, B. Interpreting Mobile and Handheld Air Sensor Readings in Relation to Air Quality Standards and Health Effect Reference Values: Tackling the Challenges. Atmosphere 2017, 8, 182. https://doi.org/10.3390/atmos8100182

[16] CT, Resmi & T, Dr. Nishanth & Kumar, Satheesh & M, Balachandramohan & Valsaraj, Kalliat. (2019). Temporal Changes in Air Quality during a Festival Season in Kannur, India. Atmosphere. 10. 137. 10.3390/atmos10030137.

[17] Jian, F., D. S. Jayas, and N. D. White. 2013. Can Ozone be a New Control Strategy for Pests of Stored Grain? Agricultural Research. 1–8.

[18] Horvitz, S. and M. Cantalejo. 2012. Application of Ozone for the Postharvest Treatment of Fruits and Vegetables. Critical Reviews in Food Science and Nutrition.

[19] Palou, L., C. H. Crisosto, J. L. Smilanick, J. E. Adaskaveg, and J. P. Zoffoli. 2002. Effects of Continuous 0.3 Ppm Ozone Exposure on Decay Development and Physiological Responses of Peaches and Table Grapes in Cold Storage. Postharvest Biology And Technology. 24: 39–48.

[20] Kim J. G., A. E. Yousef, and G. W. Chism. 1999. Use of Ozone to Inactivate Microorganisms on Lettuce. Journal of Food Safety. 19: 17–34.

[21] Karaca, H. and Y. S. Velioglu. 2007. Ozone Applications in Fruit and Vegetable Processing. Food Reviews International. 23: 91–106.

[22] Muz, M., M. Ak, O. Komesli, and C. Gökçay. 2012. An Ozone Assisted Process for Treatment of EDC's in Biological Sludge. Chemical Engineering Journal.

[23] Michael David,Mohd Haniff Ibrahim, Sevia Mahdaliza Idrus, Asrul Izam Azmi, Nor Hafizah Ngajikin, Tay Ching En Marcus, Maslina Yaacob, Mohd Rashidi Salim, Azian AbdulAziz, "Progress in Ozone Sensors Performance: A Review", Jurnal Teknologi 73(6) (2015) 23-29.

[24] http://norditech.com.au/wp-content/uploads/2019/09/O342e_Eng_17.03.pdf [ Technical Manual for O342e] (Last accessed 20-April-2021]

[25] Kyriakidis, Ioannis & Karatzas, Kostas & Papadourakis, Giorgos. (2009). Using Preprocessing Techniques in Air Quality forecasting with Artificial Neural Networks. Environmental Science and Engineering (Subseries: Environmental Science). 357-372. 10.1007/978-3-540-88351-7_27.

[26] Kin Seng Lei and Feng Wan, Pre-processing for missing data: A hybrid approach to air pollution prediction in Macau, 2010 IEEE International Conference on Automation and Logistics, Hong Kong, China, (2010) 418-422. doi: 10.1109/ICAL.2010.5585320

[27] Christophe Paoli, Cyril Voyant, Marc Muselli, Marie-Laure Nivet, Forecasting of preprocessed daily solar radiation time series using neural networks, Solar Energy, Volume 84, Issue 12(2010) 2146-2160, ISSN 0038-092X,https://doi.org/10.1016/j.solener.2010.08.011.

[28] S. Minu, Amba Shetty & Binny Gopal | Lachezar Hristov Filchev (Reviewing Editor) (2016) Review of preprocessing techniques used in soil property prediction from hyperspectral data, Cogent Geoscience, 2:1, DOI: 10.1080/23312041.2016.1145878

[29] A. Juneja and N. N. Das, Big Data Quality Framework: Pre-Processing Data in Weather Monitoring Application, 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Faridabad, India, 2019, pp. 559-563. doi: 10.1109/COMITCon.2019.8862267

[30] Khongsrabut and K. Waiyamai, Outliers Detection in Time Series Data: Case study: Provincial Waterworks Authority, 2019 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT-NCON), Nan, Thailand, (2019) 234-238.
doi: 10.1109/ECTI-NCON.2019.8692257

[31] D. Andrešič, P. Šaloun and B. Suchánová, Large Astronomical Time Series Pre-processing and Visualization for Classification using Artificial Neural Networks, 2019 IEEE 15th International Scientific Conference on Informatics, Poprad, Slovakia, (2019) 000311-000316. doi: 10.1109/Informatics47936.2019.9119283

[32] A. Famili, Wei-Min Shen, Richard Weber, Evangelos Simoudis, Data preprocessing and intelligent data analysis, Intelligent Data Analysis, 1(1–4) (1997)3-23, ISSN 1088-467X,https://doi.org/10.1016/S1088-467X(98)00007-9.

[33] A. Asok, Generalized approach to linear data transformation, 2016 International Conference on Data Science and Engineering (ICDSE), Cochin, India, (2016) 1-6, doi: 10.1109/ICDSE.2016.7823937.

[34] Z. Guan, T. Ji, X. Qian, Y. Ma, and X. Hong, A Survey on Big Data Pre-processing, 2017 5th Intl Conf on Applied Computing and Information Technology/4th Intl Conf on Computational Science/Intelligence and Applied Informatics/2nd Intl Conf on Big Data, Cloud Computing, Data Science (ACIT-CSII-BCD),

Hamamatsu, Japan, 2017, pp. 241-247, doi: 10.1109/ACIT-CSII-BCD.2017.49.

[35] Van Zoest, V.M., Stein, A. & Hoek, G. Outlier Detection in Urban Air Quality Sensor Networks. Water Air Soil Pollut 229, 111 (2018). https://doi.org/10.1007/s11270-018-3756-7

[36] Hird, Jennifer & Mcdermid, Greg. (2009). Noise reduction of NDVI time series: An empirical comparison of selected techniques. Remote Sensing of Environment. 113. 248-258. 10.1016/j.rse.2008.09.003.

[37] Michael Galarnyk. (2018, Sep 12), Understanding Boxplots, towards data science, https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

[38] Jajo, Nethal & Matawie, K.. (2009). Outlier Detection using Modified Boxplot. International Journal of Ecology and Development. 13. 116-122.

[39] Kaliyaperumal, Senthamarai & Kuppusamy, Manoj. (2015). Outlier detection in multivariate data. Applied Mathematical Sciences. 9. 2317-2324. 10.12988/ams.2015.53213.

[40] H P, Vinutha & Poornima, B. & Sagar, B.. (2018). Detection of Outliers Using Interquartile Range Technique from Intrusion Dataset. 10.1007/978-981-10-7563-6_53.

[41] Aggarwal, Vaibhav & Gupta, Vaibhav & Singh, Prayag & Sharma, Kiran & Sharma, Neetu. (2019). Detection of Spatial Outlier by Using Improved Z-Score Test. 788-790. 10.1109/ICOEI.2019.8862582.

[42] Pratama, Irfan & Permanasari, Adhistya & Ardiyanto, Igi & Indrayani, Rini. (2016). A review of missing values handling methods on time-series data. 1-6. 10.1109/ICITSI.2016.7858189.

[43] Abdullah, Mohd Mustafa Al Bakri. (2014). Filling Missing Data Using Interpolation Methods: Study on the Effect of Fitting Distribution. Key Engineering Materials. 594-595. 889-895. 10.4028/www.scientific.net/KEM.594-595.889.

[44] Raudys, Aistis & PABARŠKAITĖ, Židrina. (2018). Optimizing the smoothness and accuracy of moving average for stock price data. Technological and Economic Development of Economy. 24. 984-1003. 10.3846/20294913.2016.1216906.

[45] Pan, Ruilin & Yang, Tingsheng & Cao, Jianhua & Lu, Ke & Zhang, Zhanchao. (2015). Missing data imputation by K nearest neighbors based on grey relational structure and mutual information. Applied Intelligence. 43. 10.1007/s10489-015-0666-x.