Journal of Environmental Protection, 2023, 14, 685-709
https://www.scirp.org/journal/jep
ISSN Online: 2152-2219
ISSN Print: 2152-2197

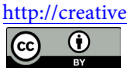
A Study of Detection of Outliers for Working and Non-Working Days Air Quality in Kolkata, India: A Case Study

Mohammad Ahmad*, Weihu Cheng, Zhao Xu, Abdul Kalam

Faculty of Science, Beijing University of Technology, Beijing, China
Email: *mahmad.or@emails.bjut.edu.cn

How to cite this paper: Ahmad, M., Cheng, W.H., Xu, Z. and Kalam, A. (2023) A Study of Detection of Outliers for Working and Non-Working Days Air Quality in Kolkata, India: A Case Study. *Journal of Environmental Protection*, **14**, 685-709.
https://doi.org/10.4236/jep.2023.148039

Received: July 30, 2023
Accepted: August 25, 2023
Published: August 28, 2023



Abstract

A variety of factors affect air quality, making it a difficult issue. The level of clean air in a certain area is referred to as air quality. It is challenging for conventional approaches to correctly discover aberrant values or outliers due to the significant fluctuation of this sort of data, which is influenced by Climate change and the environment. With accelerating industrial expansion and rising population density in Kolkata City, air pollution is continuously rising. This study involves two phases, in the first phase imputation of missing values and second detection of outliers using Statistical Process Control (SPC), and Functional Data Analysis (FDA), studies to achieve the efficacy of the outlier identification methodology proposed with working days and Nonworking days of the variables NO_2, SO_2, and O_3, which were used for a year in a row in Kolkata, India. The results show how the functional data approach outshines traditional outlier detection methods. The outcomes show that functional data analysis vibrates more than the other two approaches after imputation, and the suggested outlier detector is absolutely appropriate for the precise detection of outliers in highly variable data.

Keywords

Statistical Process Control, Functional Data Analysis, Fuzzy C Means, Outliers, Air Quality

1. Introduction

Modern humans are affected by anthropogenic sources of air pollution such as the burning of straw, coal, kerosene, and emissions from factories, cars, airplanes, and aerosol cans. Our environment is surrounded by a variety of dangerous contaminants every day, including CO, CO_2, PM, NO_2, SO_2, O_3, NH_3, and

Pb. The chemicals and particles that make up air pollution have an impact on plant, animal, and even human health. Numerous dangerous illnesses in people, including heart disease, pneumonia, bronchitis, and lung cancer, can be brought on by air pollution. Total outliers vary from the manual's points, which support the idea that they can be a substantial source of pollutants for the study of pollution of the air in cities. Additionally, data on irregular processes like emissions may be found in data that is not too high but differs from neighboring observations. The spatial-temporal study is carried out on the air quality in Kolkata using data from 2017 to 2020 [1]. For each of the sample locations, the spatial distribution was calculated using the radial basis function (RBF) approach. The model fit that estimates the spatial distribution of air pollutants was chosen using mean standardized error (MSE) and a root mean square standard error (RMSSE). In order to uncover obscure trends in the dataset and identify pollutants that have an immediate effect on the air quality index, an exploratory data analysis is conducted. Nearly all contaminants are seen to have greatly decreased in 2020. In order to analyze and predict air quality, the research endeavor takes a six-year period of pollutants data from 23 Indian cities. The resampling technique is used to study the data, and several machine learning techniques are used in forecasting the air quality. The results are compared with the available metrics [2]. To forecast PM2.5 concentrations for Kolkata during Covid 19 lockdown using MLR (multiple linear regression) and ANN (artificial neural network) models, as well as evaluation and precision of both approaches [3]. Rady A. E. S. *et al.* [4] conducted the comparative analysis of PM10 emission rates from controlled and uncontrolled cement silos in concrete batching facilities. Anyikwa S. *et al.* [5] discussed the study of monitoring and evaluation of air pollution at Ohaji/Egbema flow station and its environs via GPS in Ohaji Egbema lag, Imo state Nigeria and obtained by utilizing the air quality index to analyze whether the environment is contaminated and with which specific gas, it was determined that NO_x gases are the source of the pollution since their concentration is higher than the ambient temperature's usual value. SPC is a process that needs to be sufficiently adaptable to be effective around the objective or nominal dimensions of the quality attributes. By lowering variability, the control of statistical processes is an effective set of problem-solving techniques that is helpful for attaining process reliability and enhancing capacity. It is crucial to develop a normal variation for the process through ongoing process monitoring. If normal variation has any deviation, then disturbance has occurred; for the removal of disturbance, an adjustment has to be made. SPC is the technique of data collection, organization, analysis, and decision-making. If there is no disturbance or special cause in the data, then it is statistically controlled. The principle of SPC does decide the mean, Lower Control limit, and Upper Control Limit (LCL and UCL). If the values are falling, whether UCL or LCL, then the process is not under control.

In order to apply them to vector problems, FDA has been developed. The

FDA approach was inspired by the classical technique of data mining to cope with vectorial data treatment. The applications of the FDA have also been used for environmental research [6] [7], medical research [8], and manufacturing methods. This functional model provides two important features: first, the correlation of the data structure with time is taken into account, and second, the comparisons are made with the view of the global problem. The application compares functional depth principle curves, a metric that reflects within a group of curves the centrality of a given curve. The authors have used functional depth to solve different environmental problems [9]. Sancho *et al.* [10] used the FDA to classify the outliers for classic quality control in the communication of the report. A functional model analysis was defined by Martinez *et al.* [11] and used for the detection of outliers in samples of air quality. By generating a new functional sample, the model transforms the sample vectors to find functional outliers by adapting the principle of functional depth. In order to achieve an improved air quality management solution, the author compared the corresponding results with classical and functional approaches and obtained the most suitable methodology to evaluate the dataset. Majumdar *et al.* [12] projected emissions for the year 2030 in the Business-as-usual (BAU) case, thus taking into account the impacts of existing and proposed policies. The findings show that existing measures and policies are inadequate to significantly reduce the emissions of Delhi Metropolitan City (KMC) PM2.5 by 2030. There are three substitute setups considering different pollution control and non-technical city-specific control steps, laterally with related cost consequences, which were discussed by the author. Ahmad M. *et al.* [13] discussed the comparative study of outlier detection using classical technique and SPC for the Yamuna River. In Delhi, Haque and Singh [14] addressed Air Pollution and Human Health, where respondents with 85.1% of respiratory diseases exceeded 14.9% of waterborne diseases and included 60% of Acute Respiratory Infections (ARI), 7.8% of Chronic Obstructive Pulmonary Diseases (COPD), 1.2% of Upper Respiratory Tract Infections (URTI), and 12.7% influenza. While the level of pollution was reported as critical, only 39.3% of respondents thought that their health had been impacted by outdoor (air) pollution. Torres *et al.* [15] identified a solution that uses functional data analysis to discover outliers in urban gas emissions. Over time, the researchers considered gas emissions as curves, with outliers found by comparing curves rather than vectors. It was used for the identification of outliers in gas omissions in the city of Oviedo based on the principle of functional depth, and through the vector comparison, the results were contrasted with those generated by a conventional method. A tool for outlier detection in water quality parameters is given by Di Blasi *et al.* [16], using the variables conductivity, turbidity, and ammonium. The methods were based on the consideration of the various parameters as a vector, with components representing the concentration values. This novel technique considers measurements of water quality throughout time as continuous curves instead of individual points, *i.e.*, the analyzed dataset is

viewed as a time-dependent function instead of a group of individual points at distinct periods in time. Based on functional depth, the approach was utilized to identify outliers in water quality samples in the Mino river basin (northwest Spain). Beevers *et al.* [17] addressed measurements of atmospheric emissions of NO_x, NO_2, ozone, species of particulate matter, and roadside vehicles, all leading to the improvement of inventories of road transport emissions in London and undertaking environmental observation to determine emerging projects. The actual environmental emission efficiency of Euro 6 and VI cars, the selective catalytic reduction, and the ultra-low emission zone in London have played a crucial role in achieving the European Union's environmental NO_2 limit values. In view of the growing health evidence of air pollution in cities, policymakers should seek to reduce levels of particulate matter towards guideline values centered on the World Health Organization's health. The approach of practical pattern recognition to the identification of fluvial valleys and topographic profiles of glaciers was proposed [18], using a functional method for the classification of vector machines. The author used the original findings and compared the anticipated functional form of support vector machines with general linear functional representations and vector versions: comprehensive linear models and support vector machines. The investigator noted the benefits of the proposed vector functional support machines and the benefits of using a functional rather than a vector approach. Several outlier detection methods were introduced [19] and differentiating between parametric and nonparametric procedures and univariate and multivariate techniques. When the outliers were presented, special care was taken to ensure the robustness of the estimators used. For data mining, outlier detection is based on distance measurements, clustering, and spatial methods. In order to determine process stability, Martinez *et al.* [20] addressed statistical process monitoring and control charts used for outlier detection methods. The author proposed the method of One-class Peeling (OCP), a scalable paradigm that incorporates statistical and machine learning techniques in multivariate data to identify multiple outliers. For statistical process control, the single-class peeling approach can be implemented, doesn't require covariance estimation, and is suitable for large-dimensional records with a large proportion of outliers. The theoretical assessment indicated that the OCP approach worked well in high dimensions and was more effective and stable than current methodologies in terms of computation. In the additional materials, examples of R commands and data sets determine the respective OCP distances and threshold availability. Garca-Nieto [21] observed, as a function of time, the foraging efficacies of aerosol elements for the removal device (congealing, heterogeneous nucleation, and gravitational subsidence) and examined the health effects of the aerosol earlier and later in the active processes mentioned above by associating the fractions of respirable dirt. The well-known scavenging equations, describing the aerosol PSDs in geography, manufacturing, and region, are functional with three atmospheric situations (pure, cloudy, and city). The investigator concluded

that respirable dust is hardly scavenged, and approximately 10% remains after 18 hours of gravitational subsiding associated with the primary amount of breathing air. In comparison to disintegration and moisture, the primary elimination mechanism of respirable aerosol was gravitational settling, which is nearly six times better than a rainout. The aerosol study conducted with 86 daily samples from July 2002 to July 2003 [22]. The numerous methods of data analytics that can be used to track and efficiently assess policies or interventions to minimize emissions of nitrogen oxide (NO_x) and identifying COVID-19 pandemic incidents of pollution and eliminating outliers [23] [24]. There are numerous approaches that can be used to find outliers, but no single approach is considered to be the best. The imputation method typically uses two different kinds of methods [25]. First, the statistical technique is used to anticipate missing data [26]. To fill in the missing numbers, specific statistical aspects like the mean and different indicators are used. Next, the lacking Machine learning algorithms are used to estimate values [27]. These methods, in particular, include fuzzy Support vector machine (SVM), random forest, and C-means (FCM) [28] [29] [30]. The essential process is to use many models to provide a number of potential values. The missing information is then best possible candidate values, as established by certain assessment criteria, in their place.

This study was conducted in two halves, one utilizing imputation for missing data using Fuzzy C mean method and the second, Statistical Process Control, and the other with a functional approach. Each strategy was explored, and the outcome is to be presented in terms of the most efficient method for detecting outliers in air pollution monitoring data in order to increase its capability of informing future measures to improve local air quality. This study is divided into several sections. Section 2 introduces the location of study, datasets, and methods. Section 3 presents the results that obtained using the proposed methodology. And finally, the section 4 concluded the results and discussion of this study.

2. Materials and Methods

2.1. Case Study—Air Quality in Kolkata India

Deprived air quality in India is now considered an imminent risk to public health as well as the country's ability to develop. The environment is significantly impacted by urban air pollution, especially in developing countries. Along with 10 other Indian cities that are even more impure, Kolkata, the third-largest city by population in India, is among the 25 cities with the highest pollution in the world. One of the Indian cities that requires involvement to ensure good air quality in the decades to come is Kolkata. The smog, manufacturing production, and other forms of pollution in the environment are affecting more places. One of the most undeveloped and dirty cities in the world is Kolkata, as stated [31]. Based on the research comparing data on air quality in four Indian cities, Kolkata has greater pollution levels compared to Mumbai and Chennai and is also

close to Delhi [32] and listed in polluted cities [33]. The winter season is when air pollution in Kolkata is at its worst, being more severe than throughout other seasons [34]. On the contrary, during peak hours, the worst traffic intersections triple the city's usual pollution levels [35]. The production of power, the transportation industry, soil and road dust, trash combustion, etc. are the main sources of pollution in India. The biggest contributor to air quality in Kolkata is transportation [36] [37], where the prevalence of badly managed automobiles, the usage of gasoline, and inadequate regulations have made this industry the dominant source of pollutants in the air [38] [39]. Three thermal power plants are also operational close to Kolkata, in addition to a number of small businesses that also affect our environment [40]. According to the West Bengal Pollution Control Board's research on the various causes of air pollution in Kolkata, automobiles are the main cause of air pollution (51.4%), followed by industry (24.5%) and dust particles (21.1%) [41]. Data on air pollution was collected hourly for the current study from the Central Pollution Control Board (CPCB) in India.

2.2. Analysis Methodology

To assess information about the environment, numerous kinds of devices are accessible, such as respiratory detectors. The expected value of the sample position and taking into account classical analysis, patterns, and differences between neighboring stations may be used to identify particular data values that are not usual. The pattern analysis in R-programming [42] was an illustration of the expert structure of data and validation of an environmental parameter. The results were simply statistically examined for traditional interpretation. In order to extract conclusions, the suggested approach requires using a huge amount of data that already exists, with some incomplete findings. The amount of data that is saved in systems nowadays requires the use of machine-learning tools. The methodology of research discussed here is oriented towards the discovery of information in databases (Knowledge Discovery Database) (KDD) [43]. This provides a full data extraction procedure and yields an accessible technique for preparing data and examining the obtained results. In order to provide information and aid in conclusion creation, the KDD provides a clear and collective method of observing the design and model parameters that are practical for outlier identification, future prediction, and/or classification. We apply some steps to this research: Imputation for handling missing values, classical analysis, SPC, and FDA.

2.3. Imputation

Incomplete datasets pose a challenge in data preprocessing, making machine learning algorithms ineffective for training models. Various data imputation approaches have been proposed to predict appropriate values using different algorithms. Accurate estimation of imputation methods is crucial for completing

missing values, especially in air quality. The FCM method generates satisfactory estimation results for multiple-dimensional datasets, but clustering results are sensitive to membership degree and cluster centroids. Huang J. *et al.* [44] pointed out the fuzzy C means method for data imputation. The clustering methods are often used in data imputation when dealing with missing values. K-means is a centroid-based algorithm that partitions data into k clusters, while fuzzy K-means assigns membership degrees to each data point for soft clustering.

2.4. Statistical Process Control

Traditional statistical analysis seeks to analyze the observed frequency distribution, which yields the absolute frequency of occurrence of each of the potential results of a discrete class [23]. If there are just a finite number of various outcomes (a discrete example), if the distribution function is utilized in the situation of an indefinitely frequent and randomly trustworthy calculation and each result is different, the outcome of relative frequency will not be very enlightening. This returns all values of the absolute frequency of occurrence that are less than x in this example [45]. By applying SPC to monitor the system, it is possible to identify the outliers. But in conditions where the points do not reach the defined limit, the analysis focuses on substantially low and high measurements. To study individual observation, the techniques can be used to study individual or average maps. You should split the dataset into logical subgroups [46]. The ability to cluster variance and simply identify variability in the presence of unique causes makes it crucial for the creation of rational groupings. For instance, when a measurement occurs again in the same way, it changes due to laboratory or analytical error, unless it is impractical to use rational subgroups. The method of gathering the data is the logical subgroup. The data collected shows that some of them display intrinsic variation, which is the common cause of variation; we can ignore this. The approach that makes it possible to identify special-cause variation, which may have an unfavorable impact on the subgroups when it is avoided, In addition, if the mechanism is excessively violated, the limit of the control chart that establishes the border to be defined is established based on variability within each subgroup. Therefore, only subgroups that replicate the common cause variance of the process should be gathered [25]. When the data have some missing observations, the data will have been imputed, and if normality has been established, then the data will be correctly structured. If we reject the null hypothesis, the data can be normalized in two different ways. The first is to utilize customized procedures for non-normal distributions to turn them into normal distributions or to modify data to normalize the data set [47]. A classical process analysis can be divided into two stages: the first, when a test is performed to remove normality and unusual measurement from the outcomes, and the second, when the pattern is examined and circumstances outside of control are experienced. The average, UCL, and LCL are specified at the first

level. The average is defined precisely by the control model and signifies the objective point. Then, we set the confidence interval as the standard deviation of the process [48]. The Shewart control chart is the most commonly used for SPC due to its high success rate in spotting significant modifications in a process. The control chart is more accurately described as a monitoring system for graphical statistical processes. While the underlying distribution of processes is understood, it is usually built into a traditional control chart to monitor process parameters. Although the most current data is used in these figures, the slight or gradual improvement in the procedure is not proven. Different criteria have been devised by different writers to identify specific deviations [49] [50] and to add to the basic rules. Using these extra criteria [23] makes Shewart's control charts more vigilant and leads to a significant capacity for detecting a non-random sample. Once the data is structured and normality test has done and the data is not normal then use Box-cox transformation [51]. The transformation is as follows:

$$X_J^{(\lambda)} = \begin{cases} \dfrac{X_J^{(\lambda)} - 1}{\lambda}, & \text{if } \lambda \neq 0 \\ \log\left(X_J\right), & \text{if } \lambda = 0 \end{cases}$$

where λ denotes the maximizes the profile likelihood of the data X_J.

The average run length (ARL) is the most commonly used and simple mode to measure the capacity of a control chart with supplementary run rules. In the control charts, the run rule is used before the warn alarm indication when the process is not controlled. The important thing to do if this happens is to identify it as soon as possible. On the other hand, it would be reasonable to have a few false alarms when the mechanism is statistically under control. This term is defined specifically as an error of α (type I) and an error of β (type II). The technique of sensitivity is also defined and is highly linked to the number of outliers. It must be considered that the potential to identify out-of-control techniques is high; for this reason, there will be a lot of points that fall outside [52].

2.5. Functional Data Analysis

FDA is a collection of methods for studying curves and functions to analyze data across time [53]. Begin by converting vector samples into functional samples. The beginning points, which come from the study's generated discrete values, are used to create the curves. Smoothing is the process of transforming vector points into a continuous function over time. This data composition is valuable in the research of air pollution since it takes all of the values from the day as a single unit. As a result, a day with NO_2, SO_2, and O_3 values of varying variability may have an average identical to the other days, and the vectorial approach detects the outliers. These days would be identified as possible outliers by the functional analysis. For outlier detection in these types of investigations, functional techniques have always been shown to be superior.

Let $x(t_f)$ represent the initial observations, $t_f \in R$ signifies the time steps, and p represents the number of observations ($f = 1, 2, \cdots, p$). The individual value of the function $x(t) \in x \subset F$, where F is a functional space, can be observed. The functional space $F = span(\phi_1, \phi_2, \cdots, \phi_p)$ is used to estimate $x(t)$, where ϕ_g is the set of basis functions ($g = 1, 2, \cdots, n_b$) and p is the number of basis functions necessary to generate a functional sample. In statistics, there are various types of bases, but the Fourier basis is the most commonly employed. Furthermore, for periodic data like the ones we have in our study, the Fourier basis is the best option [25].

$$\min_{x \in F} \sum_{f=1}^{p} \left(z_f - x(t_f) \right)^2 + \lambda \Gamma(x) \tag{1}$$

$z_f = x(t_f) + \epsilon_f$ Where x is the observing point at t_f, ϵ_f is the random noise with zero mean, λ is the level of regulization and Γ is penalized, operator.

$$x(t) = \sum_{g=1}^{p} c_g \phi_g(t) \tag{2}$$

where $\{c_g\}_{g=1}^{p}$ is the coefficient that multiplied the basis function. We can write the problem of smoothing as:

$$\min_c \left\{ (z - \phi_c)^{\mathrm{T}} (z - \phi_c) + \lambda c^{\mathrm{T}} R c \right\} \tag{3}$$

$z = (z_1, \cdots, z_p)^{\mathrm{T}}$, the expansion of vector coefficient $c = (c_1, \cdots, c_p)^{\mathrm{T}}$, a (p, n_b)-matrix ϕ whose elements are $\phi_{fg} = \phi_g(t_f)$; and a (p, n_b)-matrix R whose elements are:

$$R_{gl} = \left\langle D^2 \phi_g, D^2 \phi_l \right\rangle_{L_2(T)} = \int_T D^2 \phi_g, D^2 \phi_l \mathrm{d}t \tag{4}$$

The problem can be solved with

$$c = (\phi' \phi + \lambda R)^{-1} \phi z \tag{5}$$

The functional data allows us to determine whether or not different time intervals, such as days, weeks, or months, are higher than the mean feature and how far they differ. It also enables the removal of outliers that aren't real but are caused by system failure. The notion of depth allows you to sort a collection of data in Euclidian space by how close it is to the sample core. In multivariate analysis, the concept of depth emerged and was generated to calculate a point centrality among a cloud of them. This idea started to be incorporated into practical data analysis over the course of the year. In this region, the centrality of a certain curve xi is defined by depth, and the center of the sample is the mean curve. The two depth measurements, Fraiman-Muniz depth (FMD) and H-model depth (HMD) [25], are most usual in the sense of functional data.

Through the estimation of depths, it is also possible to classify outliers with a practical approach. In this case, it will take into account elements that have different behavioral designs than the rest. Instead of summarizing the curve observations into a single point, such as the average, the definition of depth makes it possible to deal with observations identified at a given interval in curve types.

The depth technique is used for the identification of outliers and significance: there will be a low depth of an element that is distant from the sample. Thus, practical outliers are the curves with the least depth.

Firstly, the $F_{n,t}\left(x_e(t)\right)$ is the cumulative empirical distribution function of the values of the curves $\left\{x_e(t)\right\},(e=1,2,\cdots,n)$ in a certain time $t\in[a,b]$ it is contemplated. It can be defined as:

$$F_{n,t}\left(x_e(t)\right)=\frac{1}{n}\sum_{g=1}^{n} I\left(x_g(t)\le x_e(t)\right) \tag{6}$$

where $I(.)$ is an indicator function, next, the FMD for curve x_i is calculated as:

$$\mathrm{FMD}_n\left(x_e(t)\right)=\int_a^b D_n\left(x_e(t)\right)\mathrm{d}t \tag{7}$$

where $t\in[a,b]$. The functional mode in HMD, on the other hand, is the element or curve that is most densely surrounded by the other curves in the dataset. HMD is written as:

$$\mathrm{HMD}_n\left(x_e,h\right)=\sum_{g=1}^{n} K\left(\frac{\left\|x_e-x_g\right\|}{h}\right) \tag{8}$$

In a functional space, with a kernel function $K:R^{+}\to R^{+}$, a bandwidth parameter h and $\|\cdot\|$ as the norm. In a vast majority of cases, it is norm L_2, expressed as:

$$\left\|x_e(t)-x_f(t)\right\|=\left(\int\left(x_e(t)-x_f(t)\right)^2\mathrm{d}t\right)^{1/2} \tag{9}$$

There are also a number of parameters for the kernel functions $K(\cdot)$. The truncated Gaussian kernel is a popular one, can be expressed as:

$$K(t)=\frac{2}{\sqrt{2\pi}}\exp\left(-\frac{t^2}{2}\right),\ t>0 \tag{10}$$

A functional sample set may have elements that, although not containing error, exhibit characteristics that are distinct from the rest of the set. Instead of only comparing the mean values over the measurement time interval, the depth measurements mentioned above allow sets of observations over time fitted to curves to be compared in order to find outliers in functional samples. An outlier in a functional sample will therefore have significantly less depth because depth and outlier are opposite terms. In order to find functional outliers, the deepest curves are sought after. The value of bandwidth h was chosen as the 15th percentile of the empirical distribution using the HMD to create the outlier selection criterion $\left\{\left\|x_e(t)-x_f(t)\right\|^2,e,f=1,2,\cdots,n\right\}$. The cut-off C was chosen so that around 1% of accurate observations were incorrectly classified as outliers (type I error) [54]:

$$P_r\left(\mathrm{HMD}_n\left(x_e(t)\right)\right)<c=0.01,\ \ e=1,2,\cdots,n. \tag{11}$$

Unfortunately, the distribution of the selected functional depth is unknown, necessitating an estimate of C. For the purposes of this study, we selected a

method based on bootstrapping [55] [56] [57] [58] the curves of the original set with a probability proportional to depth out of the several approaches to estimate this value. As an overview, the bootstrapping strategy is as stated:

1) By using sampling with replacement, a new sample is taken from the previous sample (each element is replaced after extraction so it can be chosen again). In addition, order 10 has been chosen for resampling.

2) The populational parameter of interest is estimated using this new sample as a basis to generate a statistic.

3) Repeat the steps overhead a significant number of times.

4) Finally, Determine the empirical statistical distribution.

3. Result and Discussion

Two stages were involved in gathering and analyzing the results. Each database variable for air quality was analyzed statistically and in the first phase is to impute the data with FCM method. The second phase proceeded with the building of the chart with reasonable monthly subgroups in order to test its correctness and offer a wide overview of the trends and mean values of each variable. At the same time, the monthly data were studied through functional data analysis with functional depth. The hourly data collected from the CPCB for one year (2019) As a result, our investigation focuses on the research of daily hourly measurements of NO_2, SO_2, and O_3 in two cases: working and nonworking. There are 298 working days and 67 nonworking days, including Sunday, public holidays such as Republic Day, Independence Day, and Diwali etc. Following that, the results obtained using the recommended approaches for air pollution analysis in Kolkata's urban area are provided individually.

3.1. Working Case

The Second and third phase of NO_2, SO_2 and O_3 analysis includes the $\bar{X}$ charts with hourly rational groups for 298 days, as well as functional data analysis of hourly data with depth. During learning stage, non-normality measurements were observed, requiring Box-Cox transformation to verify normality assumption using monthly concentration averages. The Box-cox transformation is applied to the data set. The $\bar{X}$ Charts shows the annual cycle for working days. It is observed that there are 248 outliers for NO_2, 231 outliers for SO_2, and 136 outliers for O_3 as shown in **Figures 1-3**. The lowest concentration of NO_2 values takes place March to September and highest concentration in between January, February and October to December. For SO_2 the lowest concentration takes place January to March and November, December and the highest concentration in between April to October. And for O_3 the lowest concentration in between January, February and September to November and the highest concentration in between March to August and December. The results obtained with functional depth method are shown in **Figures 4-9**. There are 76, 63, and 22 outliers for NO_2, SO_2, and O_3 concentration and are shown in **Figure 5**, **Figure 7**, and **Figure 9**.

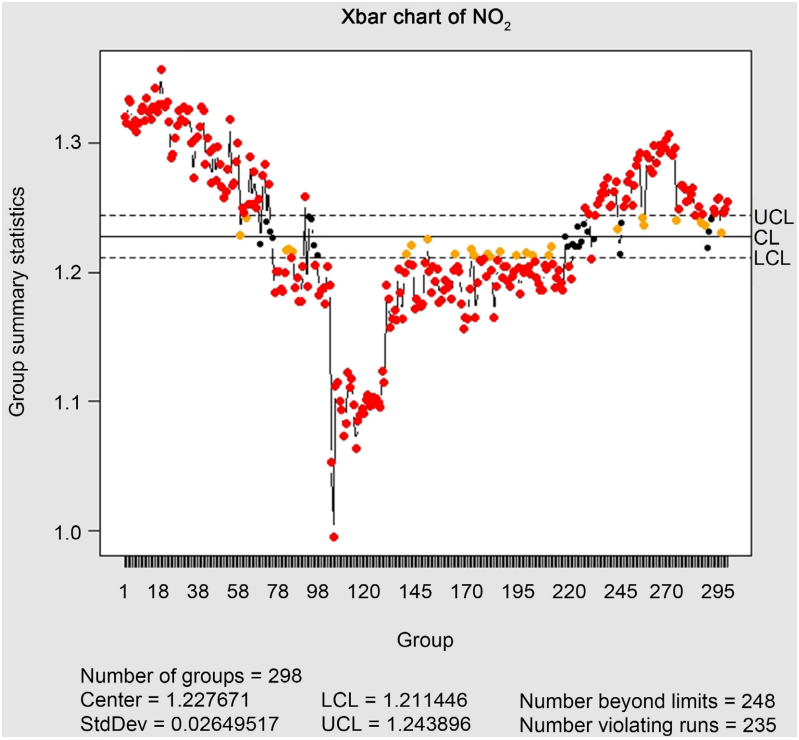


Figure 1. $\bar{X}$ Chart of NO_2 (Working Days).

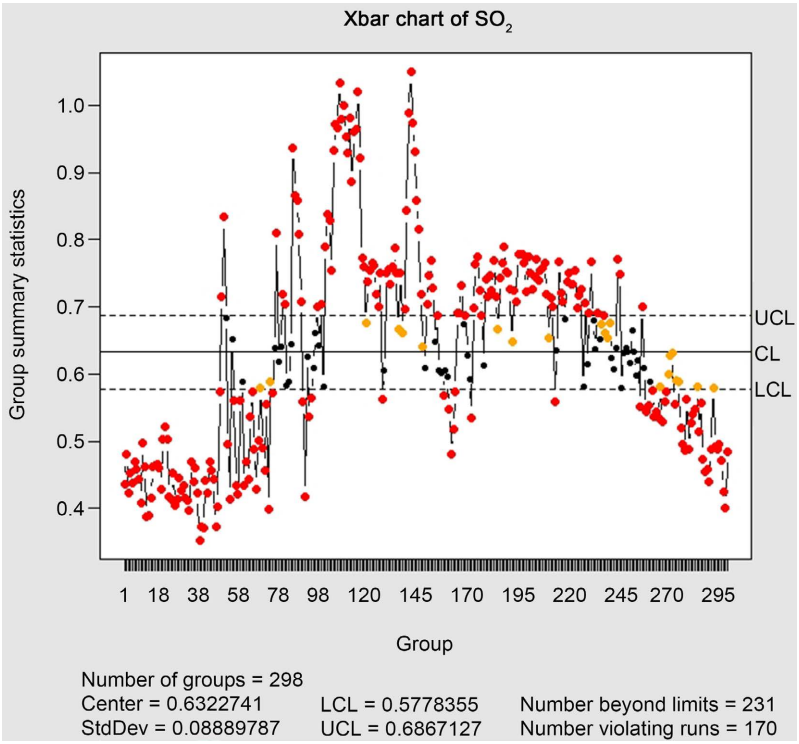


Figure 2. $\bar{X}$ Chart of SO_2 (Working Days).

3.2. Non Working Case

In this case the variable NO_2, SO_2 and O_3 analysis includes the $\bar{X}$ charts with

hourly rational groups for 67 days, as well as functional data analysis of hourly data with depth. The $\bar{X}$ Charts shows the annual cycle for working days. There are 53 outliers for NO_2, 52 outliers for SO_2, and 30 outliers for O_3 as shown in Figures 10-12. The results obtained with functional depth method are shown in Figures 13-18. There are 76, 63, and 22 outliers for NO_2, SO_2, and O_3 concentration and are shown in Figure 14, Figure 16, and Figure 18.

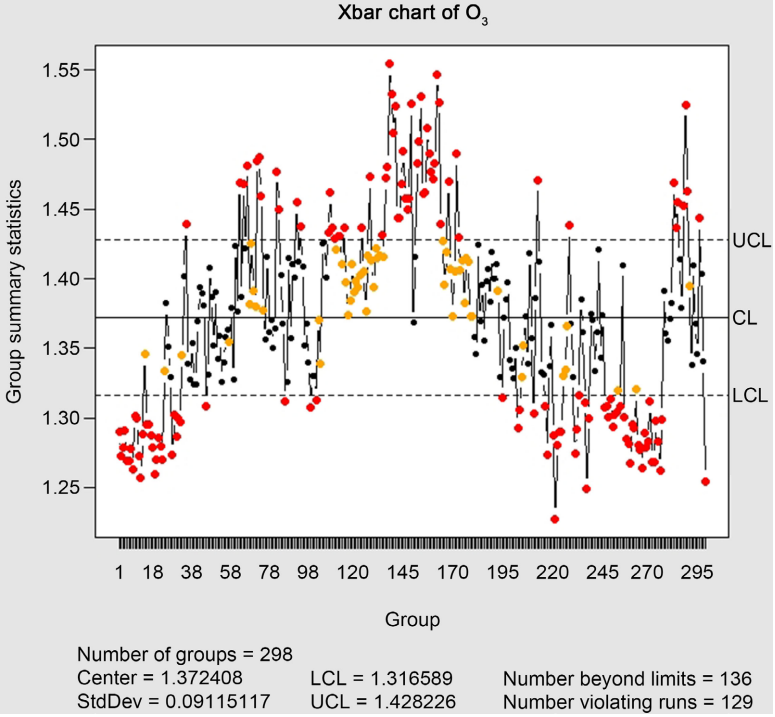


Figure 3. $\bar{X}$ Chart of O_3 (Working Days).

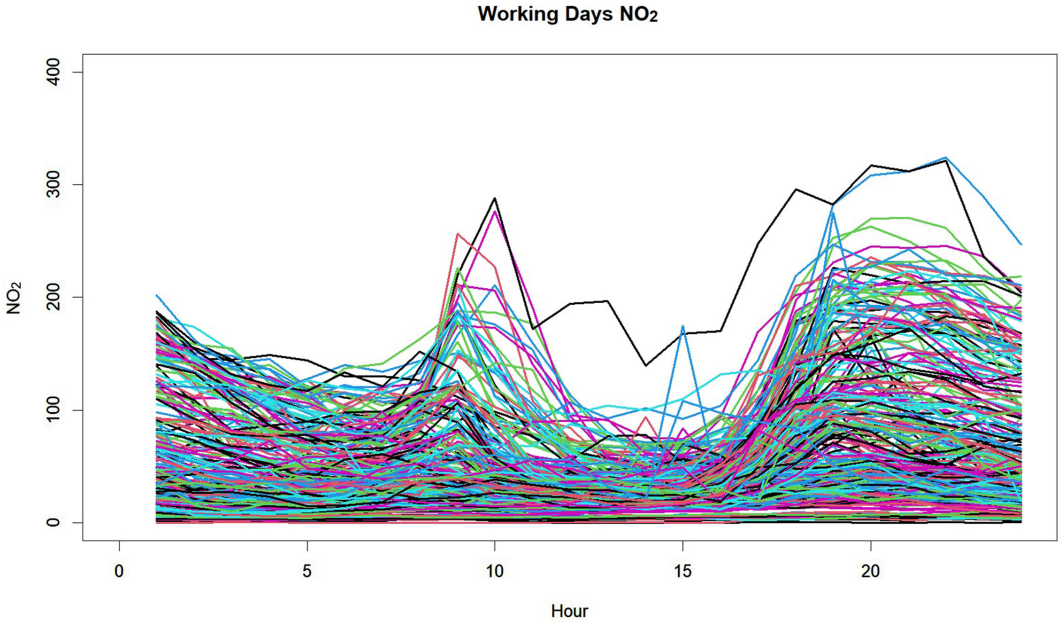


Figure 4. Functional representation of NO_2 concentration (Working Days).

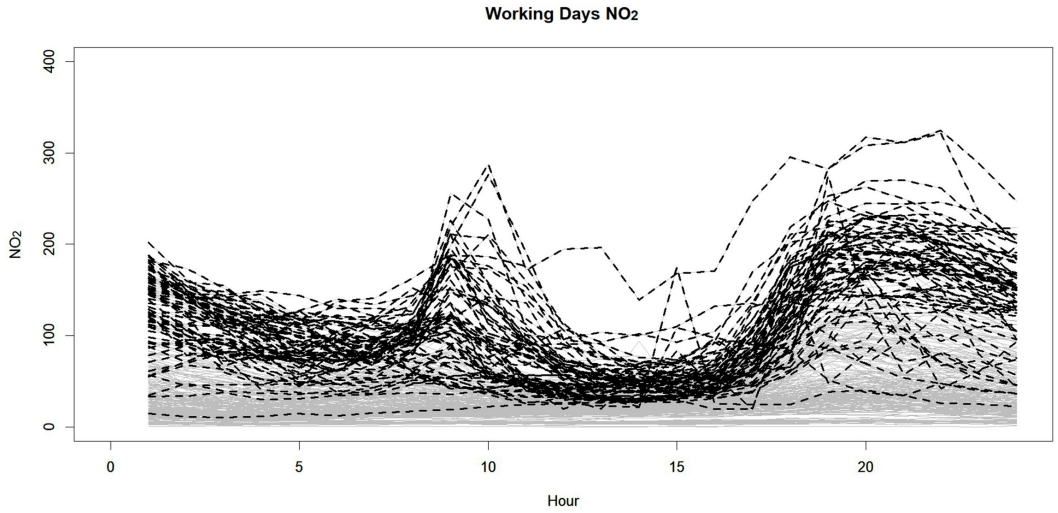


Figure 5. Functional outlier representation of NO_2 (Working Days).

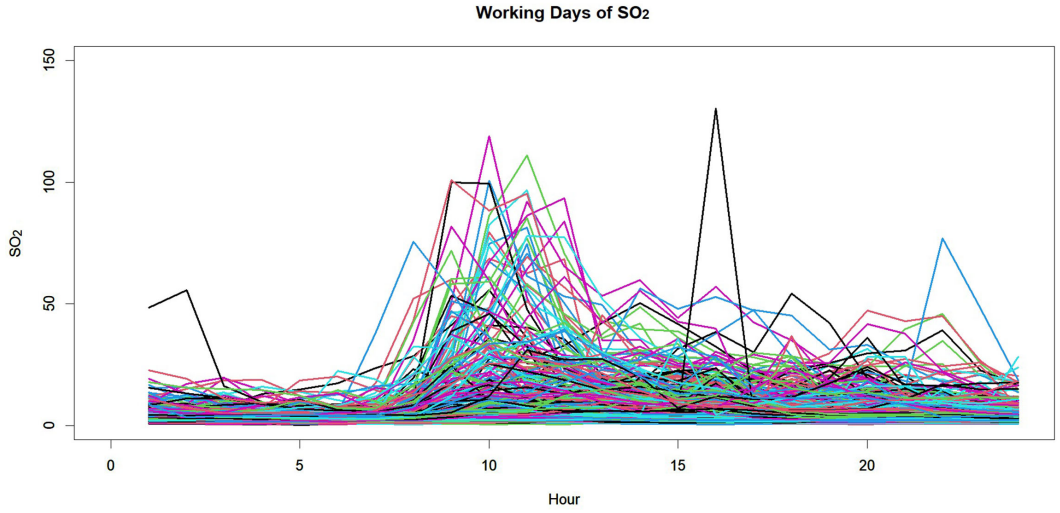


Figure 6. Functional representation of SO_2 concentration (Working Days).

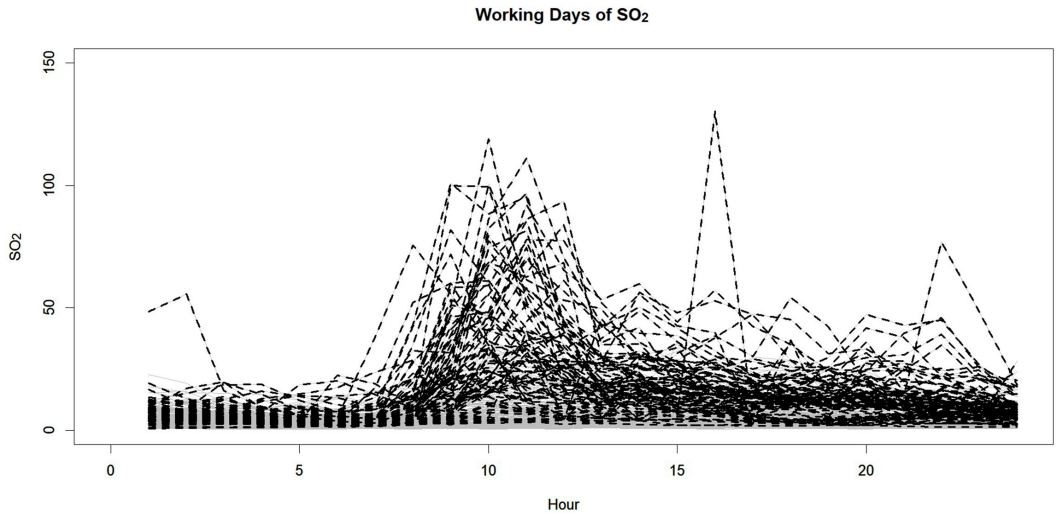


Figure 7. Functional outlier representation of SO_2 (Working Days).

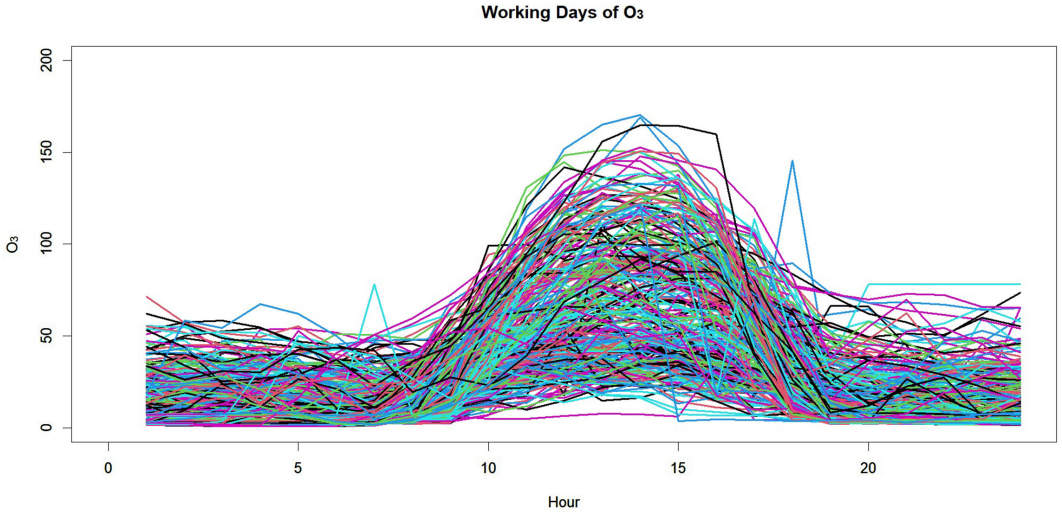


Figure 8. Functional representation of O_3 concentration (Working Days).

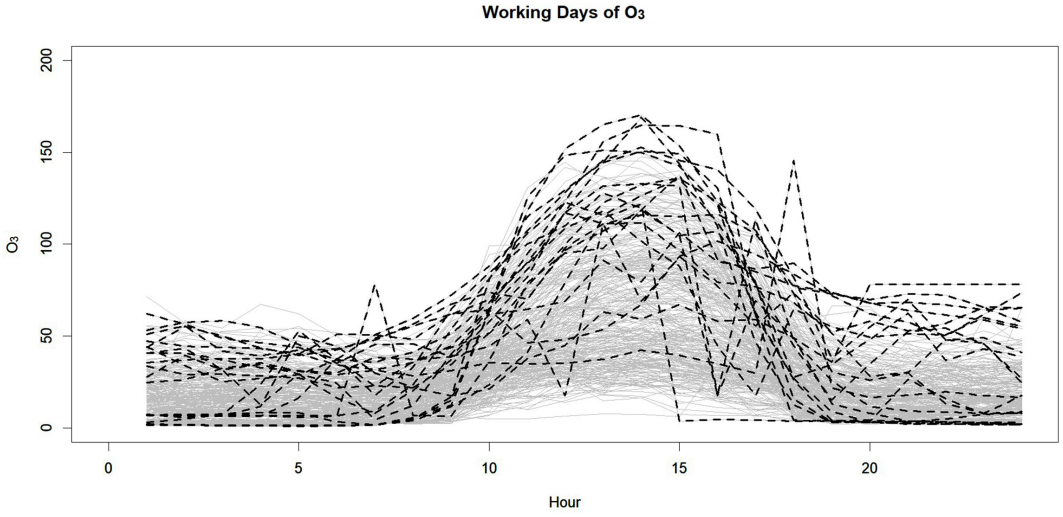


Figure 9. Functional outlier representation of O_3 (Working Days).

It was discovered during the phase of learning that the data sets identifying days as the rational subgroups of the chart for working and nonworking days are not under control due to non-normality. Despite the fact that after applying Box-cox transformation the chart was not under control and has a high degree of variation, it should be noted that after outlier detection is complete, the study is regarded as having accomplished its aim, and no back transformation is conducted. Shewart graphs for the control stage revealed the considerable variability of the pollutant's average. Figures 1-3 and Figures 10-12 shows control charts for working and non-working days of NO_2, SO_2, and O_3 recorded in Kolkata station. The red dot in the graph indicates a process that is not statistically controlled, whereas the black dots show a process that is statistically controlled. On the contrary, the visual interpretation of these graphs indicates that there is similar behavior in both working and nonworking (NO_2, SO_2, and O_3) cases, resulting in outcomes above the control limits.

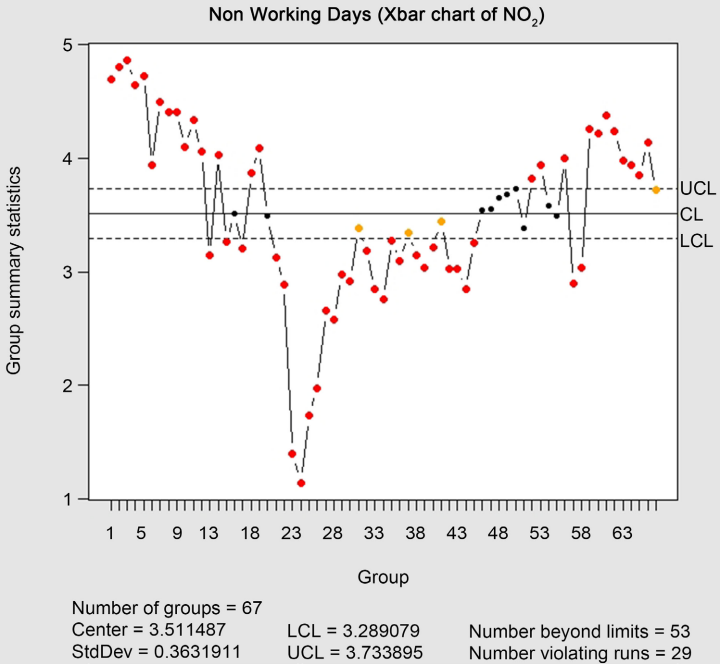


Figure 10. $\bar{X}$ Chart of NO_2 (Non-Working Days).

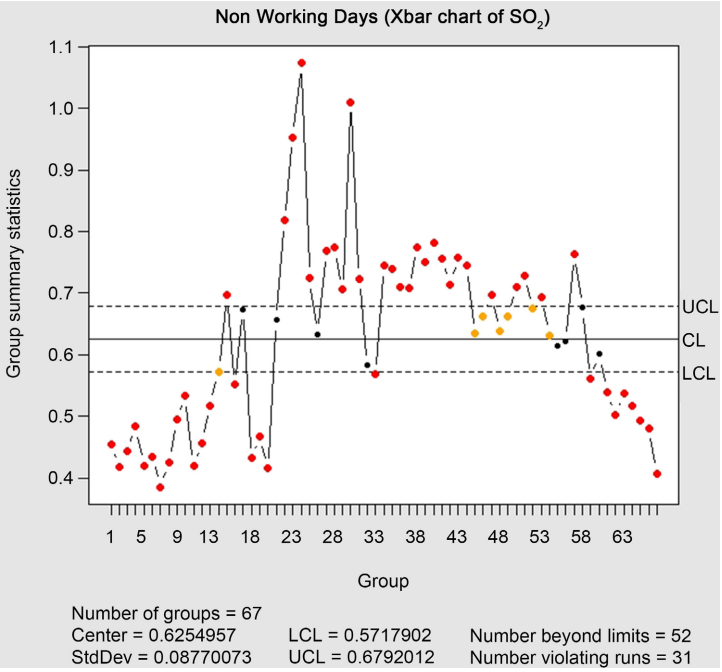


Figure 11. $\bar{X}$ Chart of SO_2 (Non-Working Days).

In the functional methodology technique, the initial step is to generate a sample curve based on discrete measurements taken each hour. The graph shows 298 functions for working and 67 functions for non-working derived from 24-hour data. If the data has been translated into functional form, *i.e.*, the curves with 24 points in a day, each of which takes into consideration the correlation between the NO_2, SO_2, and O_3 readings and may be examined for outliers, the

data can be analyzed. When the depths are taken into account, the functional analysis results let us discover days with aberrant functional points, even if there are no outliers. Despite the fact that the daily limit values were exceeded, the NO_2, O_3, and SO_2 absorption may demonstrate aberrant behavior throughout the course of a day. The functional technique, on the other hand, detects any variation from normal daily NO_2, O_3, and SO_2 emission behavior without depending on any distribution limits. This can be seen in **Figure 5**, **Figure 9**, **Figure 11**, **Figure 14**, **Figure 16**, **Figure 18**, which shows the functional outliers found in both cases for NO_2, SO_2, and O_3. Outliers are indicated by black lines in **Figure 5**, **Figure 9**, **Figure 11**, **Figure 14**, **Figure 16**, **Figure 18**.

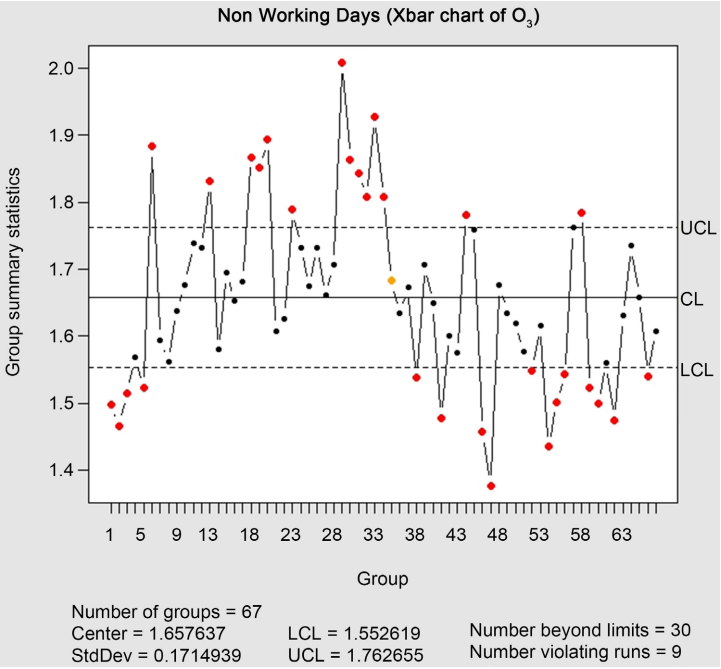


Figure 12. $\overline{X}$ Chart of O_3 (Non-Working Days).

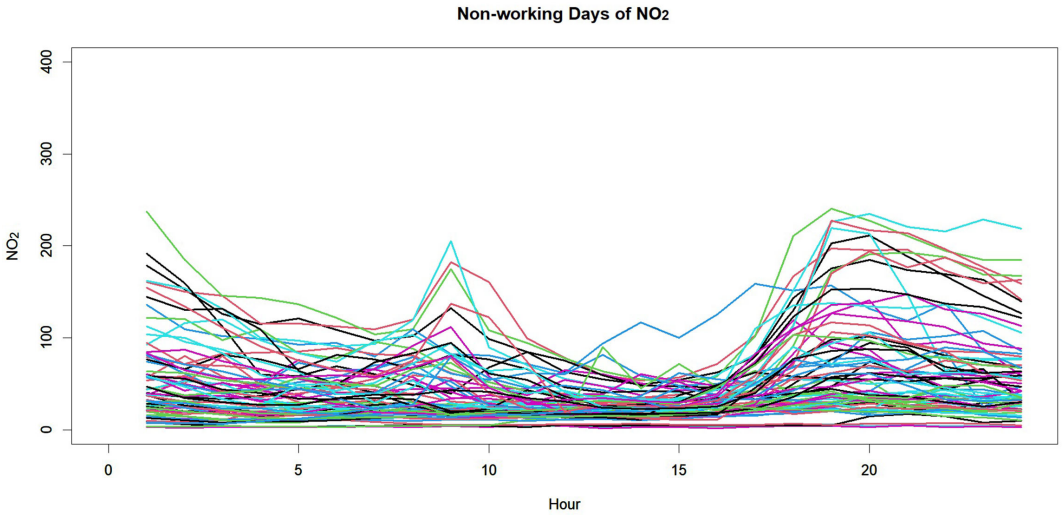


Figure 13. Functional representation of NO_2 concentration (Non-Working Days).

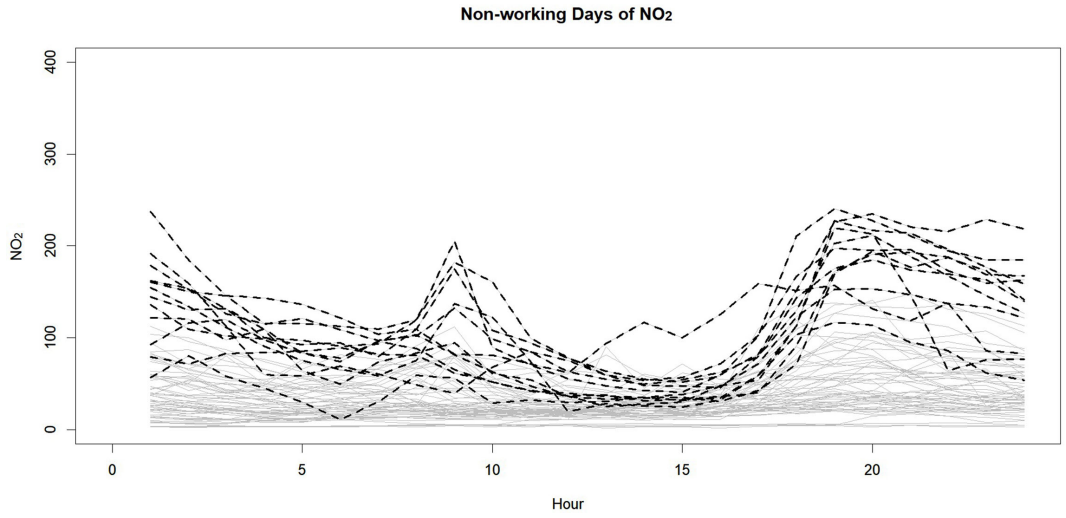


Figure 14. Functional outlier representation of NO_2 (Non-Working Days).

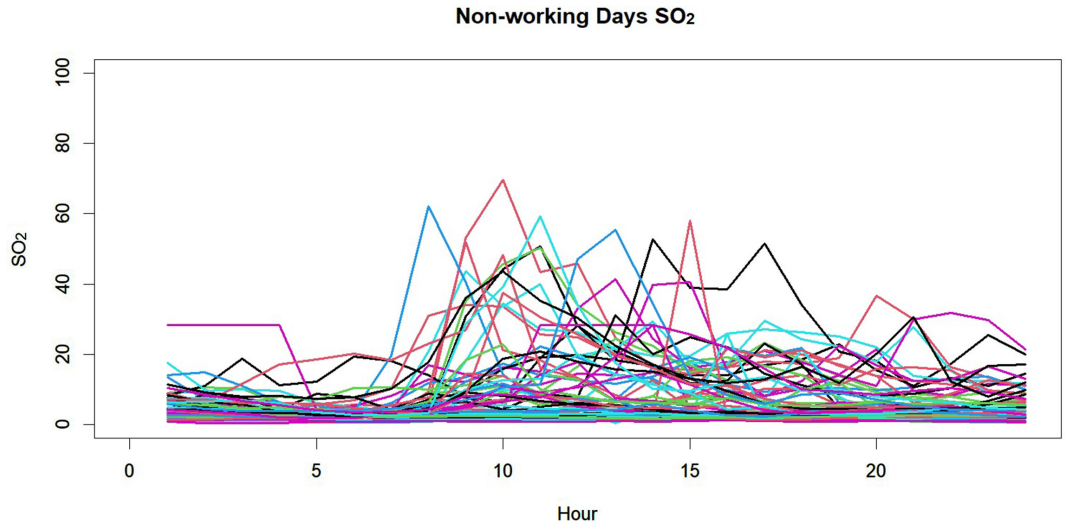


Figure 15. Functional representation of SO_2 concentration (Non-Working Days).

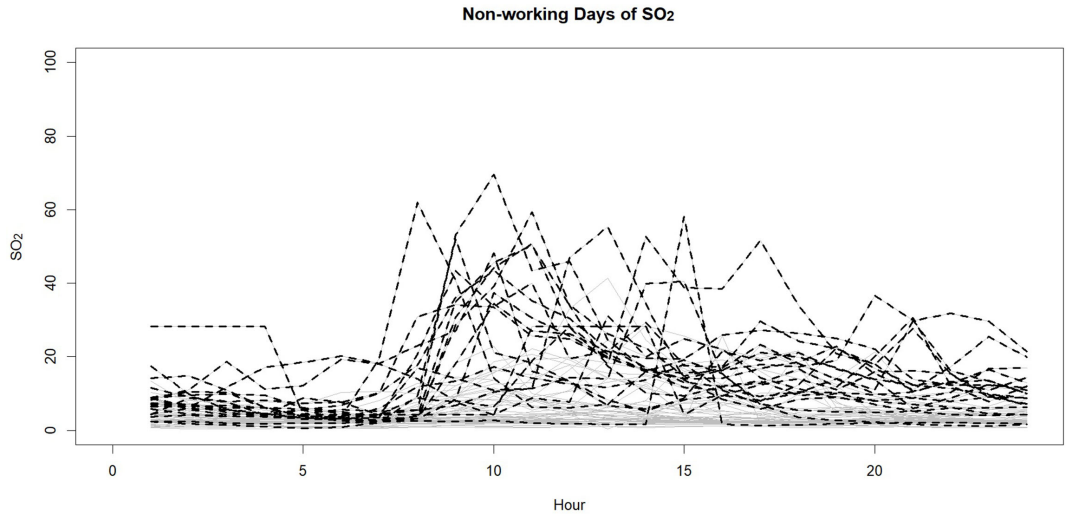


Figure 16. Functional outlier representation of SO_2 (Non-Working Days).

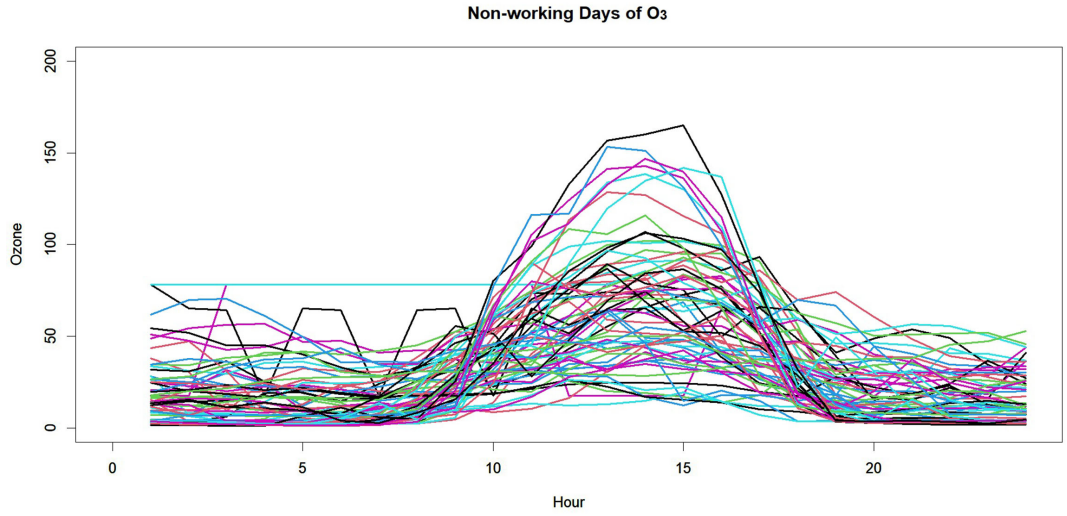


Figure 17. Functional representation of O_3 concentration (Non-Working Days).

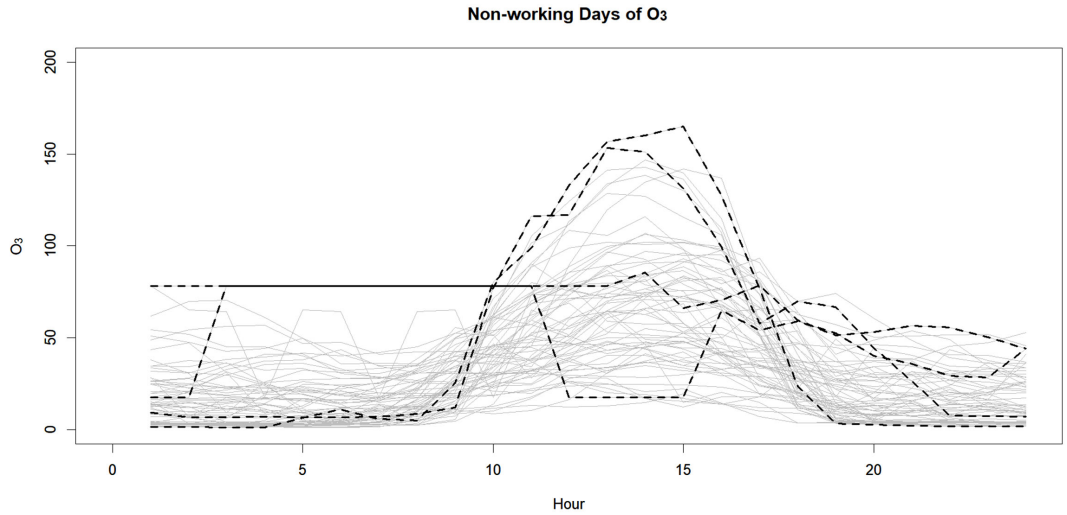


Figure 18. Functional outlier representation of O_3 (Non-Working Days).

The traditional method detects a large number of outliers, even identifying outliers in more than half of the data in certain circumstances. This problem happens regardless of the station or compound review. As seen in the previous figure of functional outlier detection, the number of outliers found by functional analysis is significantly smaller in both working and nonworking cases. It should also be highlighted that the outliers revealed by functional analysis are among those identified by the traditional method. The non-normality of the data is analyzed and may be the cause of this finding. When the data are characterized by their normality or when the sets under study have a known statistical distribution, the classical approaches are very robust. The effectiveness of this approach, however, declines when it is used to analyze sets without these characteristics. Since the traditional methods examine the data as punctual observations, the detection of outliers in this situation is complicated. When studying atmospheric pollution, the actual pollution caused by a particular substance is

dependent on a sequence of measurements that exceed the limits over a certain amount of time rather than a single measurement. The functional analysis, on the other hand, treats the data as a temporal series rather than as punctual observations, which is truly an appropriate solution to this case. Instead of an individual measurement, the airborne pollution caused by a certain compound is supplied by a group of observations that exceed the stated limitations over a specific time interval. This methodology explains why the classical analysis identifies a higher proportion of outliers than the functional analysis. Additionally, this methodology eliminates potential measurement errors, which are anomalous punctual measurements that are classified as outliers by conventional analysis. As a result, the results are more trustworthy when the instrumental error is removed in the first stage.

4. Conclusion

In this study, a number of mathematical techniques for identifying outliers in highly variable environmental data have been examined and contrasted. Real data from Kolkata, India, was used to validate these strategies. The database contains daily records of several air quality parameters (NO_2, SO_2, and O_3) from January 1 to December 31 (2019). But there are several missing values in the data sets. The FCM method was used for imputation to fill the missing values. The traditional statistical approach, implemented by SPC, nevertheless remains too simple when considering this scope. Although it offers intriguing statistical data, its discrete basis results in a number of flaws in the data set's time correlation structure. Additionally, it excludes all of the trends or outliers that exhibit behavior that deviates from the norm by having high or low values that are just below the limits. Although the idea of rational subgroups increases the loss of information, it does require the time series correlation of the data. False alarms are also caused by the data's non-normality. However, this approach can identify the most obvious outliers and offer an insightful graphical picture of the data's underlying trends. Based on the idea of functional depth, this study proposes an approach for the functional detection of outliers. As a result, it is possible to form decisions about the air quality in the research area by identifying outliers in a sample of impurities. This methodology is also contrasted with the traditional approaches to the research of outliers, with the result that it is more effective for the accurate identification of outliers because the probability of identifying an error in measurements as an outlier is reduced. In conclusion, future studies will concentrate on removing the requirement for percentiles to identify which functions are outliers. This will be attempted by putting a variety of classification techniques, like isolation forest or k-means, to the test.

Data Availability Statement

The data has been made publicly available by the Central Pollution Control Board: https://cpcb.nic.in/ which is the official portal of Government of India. They also have a real-time monitoring app: https://app.cpcbccr.com/AQI_India/.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Bhunia, G.S., Ghosh, A. and Shit, P.K. (2022) Comprehensive Spatio-Temporal Analysis of Ambient Air Quality of Kolkata Municipal Corporation, Kolkata (West Bengal, India) during 2017-2020. *Arabian Journal of Geosciences*, **15**, Article No. 1782. https://doi.org/10.1007/s12517-022-11081-7

[2] Kumar, K. and Pande, B.P. (2022) Air Pollution Prediction with Machine Learning: A Case Study of Indian Cities. *International Journal of Environmental Science and Technology*, **20**, 5333-5348.

[3] Bera, B., Bhattacharjee, S., Sengupta, N. and Saha, S. (2021) $PM_{2.5}$ Concentration Prediction during COVID-19 Lockdown over Kolkata Metropolitan City, India Using MLR and ANN Models. *Environmental Challenges*, **4**, Article ID: 100155. https://doi.org/10.1016/j.envc.2021.100155

[4] Rady, A., Beheary, M., El-Metwally, M. and Zahran, A. (2023) Comparative Analysis of PM10 Emission Rates from Controlled and Uncontrolled Cement Silos in Concrete Batching Facilities. *Open Journal of Air Pollution*, **12**, 67-77. https://doi.org/10.4236/ojap.2023.122004

[5] Anyikwa, S., Ndukwe, O., Umeojiakor, T., Nnaji, P. and Amadi, N. (2022) Monitoring and Evaluation of Air Pollution at Ohaji/Egbema Flow Station and Its Environs via GPS in Ohaji Egbema Lga, Imo State Nigeria. *Detection*, **9**, 37-49. https://doi.org/10.4236/detection.2022.94004

[6] Sosa Donoso, J.R., Flores, M., Naya, S. and Tarrío-Saavedra, J. (2023) Local Correlation Integral Approach for Anomaly Detection Using Functional Data. *Mathematics*, **11**, Article 815. https://doi.org/10.3390/math11040815

[7] Sancho, J., Iglesias, C., Piñeiro, J., Martínez, J., Pastor, J.J., Araújo, M. and Taboada, J. (2016) Study of Water Quality in a Spanish River Based on Statistical Process Control and Functional Data Analysis. *Mathematical Geosciences*, **48**, 163-186. https://doi.org/10.1007/s11004-015-9605-y

[8] Dombeck, D.A., Graziano, M.S. and Tank, D.W. (2009) Functional Clustering of Neurons in Motor Cortex Determined by Cellular Resolution Imaging in Awake Behaving Mice. *Journal of Neuroscience*, **29**, 13751-13760. https://doi.org/10.1523/JNEUROSCI.2985-09.2009

[9] Sancho, J., Martínez, J., Pastor, J.J., Taboada, J., Piñeiro, J.I. and García-Nieto, P.J. (2014) New Methodology to Determine Air Quality in Urban Areas Based on Runs Rules for Functional Data. *Atmospheric Environment*, **83**, 185-192. https://doi.org/10.1016/j.atmosenv.2013.11.010

[10] Sancho, J., Pastor, J.J., Martínez, J. and García, M.A. (2013) Evaluation of Harmonic Variability in Electrical Power Systems through Statistical Control of Quality and Functional Data Analysis. *Procedia Engineering*, **63**, 295-302. https://doi.org/10.1016/j.proeng.2013.08.224

[11] Martínez, J., Saavedra, Á., García-Nieto, P.J., Piñeiro, J.I., Iglesias, C., Taboada, J. and Pastor, J. (2014) Air Quality Parameters Outliers Detection Using Functional Data Analysis in the Langreo Urban Area (Northern Spain). *Applied Mathematics and Computation*, **241**, 1-10. https://doi.org/10.1016/j.amc.2014.05.004

[12] Majumdar, D., Purohit, P., Bhanarkar, A.D., Rao, P.S., Rafaj, P., Amann, M. and

Srivastava, A. (2020) Managing Future Air Quality in Megacities: Emission invEntory and Scenario Analysis for the Delhi Metropolitan City, India. *Atmospheric Environment*, **222**, Article ID: 117135. https://doi.org/10.1016/j.atmosenv.2019.117135

[13] Ahmad, M., Haq, A., Kalam, A. and Shah, S.K. (2022) A Comparative Study of Outlier Detection of Yamuna River Delhi India by Classical Statistics and Statistical Quality Control. *Reliability*: *Theory & Applications*, **17**, 430-438.

[14] Haque, M. and Singh, R.B. (2017) Air Pollution and Human Health in Delhi, India: A Case Study. *Climate*, **5**, Article 77. https://doi.org/10.3390/cli5040077

[15] Torres, J.M., Nieto, P.G., Alejano, L. and Reyes, A.N. (2011) Detection of Outliers in Gas Emissions from Urban Areas Using Functional Data Analysis. *Journal of Hazardous Materials*, **186**, 144-149. https://doi.org/10.1016/j.jhazmat.2010.10.091

[16] Di Blasi, J.P., Torres, J.M., Nieto, P.G., Fernández, J.A., Muñiz, C.D. and Taboada, J. (2013) Analysis and Detection of Outliers in Water Quality Parameters from Different Automated Monitoring Stations in the Miño River Basin (NW Spain). *Ecological Engineering*, **60**, 60-66. https://doi.org/10.1016/j.ecoleng.2013.07.054

[17] Beevers, S.D., Carslaw, D.C., Dajnak, D., Stewart, G.B., Williams, M.L., Fussell, J.C. and Kelly, F.J. (2016) Traffic Management Strategies for Emissions Reduction: Recent Experience in London. *Energy and Emission Control Technologies*, **4**, 27-39. https://doi.org/10.2147/EECT.S69858

[18] Matías, J.M., Ordóñez, C., Taboada, J. and Rivas, T. (2009) Functional Support Vector Machines and Generalized Linear Models for Glacier Geomorphology Analysis. *International Journal of Computer Mathematics*, **86**, 275-285. https://doi.org/10.1080/00207160801965305

[19] Ben-Gal, I. (2005) Outlier Detection. In: Maimon, O. and Rokach, L., Eds., *Data Mining and Knowledge Discovery Handbook*, Springer, Boston, 131-146. http://mse415.free.fr/outliermethods.pdf https://doi.org/10.1007/0-387-25465-X_7

[20] Martinez, W.G., Weese, M.L. and Jones-Farmer, L.A. (2020) A One Class Peeling Method for Multivariate Outlier Detection with Applications in Phase I SPC. *Quality and Reliability Engineering International*, **36**, 1272-1295. https://doi.org/10.1002/qre.2629

[21] García-Nieto, P.J. (2001) Parametric Study of Selective Removal of Atmospheric Aerosol by Coagulation, Condensation and Gravitational Settling. *International Journal of Environmental Health Research*, **11**, 149-160. https://doi.org/10.1080/09603120020047528

[22] Karaca, F., Alagha, O. and Ertürk, F. (2005) Statistical Characterisation of Atmospheric PM10 and PM2. 5 Concentrations at a Non-Impacted Suburban Site of Istanbul, Turkey. *Chemosphere*, **59**, 1183-1190. https://doi.org/10.1016/j.chemosphere.2004.11.062

[23] Torres, J.M., Pastor Perez, J., Sancho Val, J., McNabola, A., Martinez Comesana, M. and Gallagher, J. (2020) A Functional Data Analysis Approach for the Detection of Air Pollution Episodes and Outliers: A Case Study in Dublin, Ireland. *Mathematics*, **8**, Article 225. https://doi.org/10.3390/math8020225

[24] Rigueira, X., Araújo, M., Martínez, J., García-Nieto, P.J. and Ocarranza, I. (2022) Functional Data Analysis for the Detection of Outliers and Study of the Effects of the COVID-19 Pandemic on Air Quality: A Case Study in Gijón, Spain. *Mathematics*, **10**, Article 2374. https://doi.org/10.3390/math10142374

[25] Amiri, M. and Jensen, R. (2016) Missing Data Imputation Using Fuzzy-Rough Methods. *Neurocomputing*, **205**, 152-164.

https://doi.org/10.1016/j.neucom.2016.04.015

[26] García-Laencina, P.J., Sancho-Gómez, J.L. and Figueiras-Vidal, A.R. (2010) Pattern Classification with Missing Data: A Review. *Neural Computing and Applications*, **19**, 263-282. https://doi.org/10.1007/s00521-009-0295-6

[27] Tang, J., Zhang, G., Wang, Y., Wang, H. and Liu, F. (2015) A Hybrid Approach to Integrate Fuzzy C-Means Based Imputation Method with Genetic Algorithm for Missing Traffic Volume Data Estimation. *Transportation Research Part C*: *Emerging Technologies*, **51**, 29-40. https://doi.org/10.1016/j.trc.2014.11.003

[28] Tang, J., Yu, S., Liu, F., Chen, X. and Huang, H. (2019) A Hierarchical Prediction Model for Lane-Changes Based on Combination of Fuzzy C-Means and Adaptive Neural Network. *Expert Systems with Applications*, **130**, 265-275. https://doi.org/10.1016/j.eswa.2019.04.032

[29] Hasnain, A., Sheng, Y., Hashmi, M.Z., Bhatti, U.A., Ahmed, Z., and Zha, Y. (2023) Assessing the Ambient Air Quality Patterns Associated to the COVID-19 Outbreak in the Yangtze River Delta: A Random Forest Approach. *Chemosphere*, **314**, Article 137638. https://doi.org/10.1016/j.chemosphere.2022.137638

[30] Chao, Q.I.A.N., Jian-Xun, C.H.E.N., Yan-Bin, L.U.O. and Liang, D.A.I. (2016) Random Forest Based Operational Missing Data Imputation for Highway Tunnel. *Journal of Transportation Systems Engineering and Information Technology*, **16**, 81-87.

[31] Ghose, M.K., Paul, R. and Banerjee, S.K. (2005) Assessment of the Impact on Human Health of Exposure to Urban Air Pollutants: An Indian Case Study. *International Journal of Environmental Studies*, **62**, 201-214. https://doi.org/10.1080/0020723042000275123

[32] Anon (2002) Report of the Committee Constituted by the Order of the Honourable High Court, Calcutta for Recommending Measures to Check the Pollution in the City of Calcutta. Health Effects of Air Pollution: A Study of Kolkata, Department of Environment, Government of West Bengal and West Bengal Pollution Control Board, Kolkata.

[33] Kumar, B. and Singh, R.B. (2003) Urban Development and Anthropogenic Climate Change: Experience in Indian Metropolitan Cities. Manak Publication Pvt. Ltd., New Delhi.

[34] Singh, R.B. and Haque, S. (2016) Urban Ambient Air Quality and Respiratory Health in Kolkata: A Dispensary Level Analysis. *European Urban and Regional Studies*, **2**, 7-21.

[35] Bhaumik, S. (2007) Air Pollution Suffocates Calcutta. BBC News. http://news.bbc.co.uk/2/hi/south_asia/6614561.stm

[36] Mondal, R., Sen, G.K., Chatterjee, M., Sen, B.K. and Sen, S. (2000) Ground-Level Concentration of Nitrogen Oxides (NO_x) at Some Traffic Intersection Points in Calcutta. *Atmospheric Environment*, **34**, 629-633. https://doi.org/10.1016/S1352-2310(99)00216-2

[37] Ghose, M.K., Paul, R. and Banerjee, S.K. (2004) Assessment of the Impacts of Vehicular Emissions on Urban Air Quality and Its Management in Indian Context: The Case of Kolkata (Calcutta). *Environmental Science & Policy*, **7**, 345-351. https://doi.org/10.1016/j.envsci.2004.05.004

[38] Mukherjee, A., Mukherjee, G., Banerji, U. and Mukhopadhyay, S.P. (1998) Occupational Exposure of the Traffic Personnel of Calcuta to Lead and Carbonmonoxide. *Pollution Research*, **17**, 359-362.

[39] Kazimuddin, A. and Banerjee, L. (2000) Fighting for Air.

http://www.downtoearth.org.in/coverage/fighting-for-air-18428

[40] Ghose, M.K. (2009) Air Pollution in the City of Kolkata: Health Effects Due to Chronic Exposure. *Environmental Quality Management*, **19**, 53-70. https://doi.org/10.1002/tqem.20245

[41] West Bengal Pollution Control Board (2005) Air Quality Management: Final Report. WBPCB in Collaboration with Asian Development Bank, Intercontinental Consultant and Technocrats Pvt. Ltd., New Delhi.

[42] Carslaw, D.C. and Ropkins, K. (2012) Openair—An R Package for Air Quality Data Analysis. *Environmental Modelling & Software*, **27**, 52-61. https://doi.org/10.1016/j.envsoft.2011.09.008

[43] Chan, K.C.C., Wong, A.K.C., Piatesky-Shapiro, G. and Frawley, W.J. (1991) Knowledge Discovery in Databases. MIT Press, Cambridge.

[44] Huang, J., Mao, B., Bai, Y., Zhang, T. and Miao, C. (2020) An Integrated Fuzzy C-Means Method for Missing Data Imputation Using Taxi GPS Data. *Sensors*, **20**, Article 1992. https://doi.org/10.3390/s20071992

[45] Montgomery, D.C. (2017) Design and Analysis of Experiments. John Wiley & Sons, Hoboken.

[46] Shewhart, W.A. (1931) Economic Control of Quality of Manufactured Product. Macmillan, London.

[47] Chen, Y.K. (2003) An Evolutionary Economic-Statistical Design for VSI X Control Charts under Non-Normality. *The International Journal of Advanced Manufacturing Technology*, **22**, 602-610. https://doi.org/10.1007/s00170-003-1612-3

[48] Grant, E. and Leavenworth, R. (1998) Statistical Quality Control. McGraw-Hill, New York.

[49] Champ, C. and Woodall, W. (1987) Exact Results for Shewhart Control Charts with Supplementary Runs Rules. *Technometrics*, **29**, 393-399. https://doi.org/10.1080/00401706.1987.10488266

[50] Zhang, M.H., Lin, W.Y., Klein, S.A., Bacmeister, J.T., Bony, S., Cederwall, R.T., Zhang, J.H., *et al.* (2005) Comparing Clouds and Their Seasonal Variations in 10 Atmospheric General Circulation Models with Satellite Measurements. *Journal of Geophysical Research*: *Atmospheres*, **110**, D15S02. https://doi.org/10.1029/2004JD005021

[51] Box, G. and Cox, D. (1964) An Analysis of Transformations. *Journal of the Royal Statistical Society*: *Series B* (*Methodological*), **26**, 211-243. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x

[52] Electric, W. (1956) Statistical Quality Control Handbook. AT&T Technologics, Indianapolis.

[53] Ramsay, J. and Silverman, B. (2005) Functional Data Analysis. Springer, New York. https://doi.org/10.1007/b98888

[54] Cuevas, A. and Fraiman, R. (1997) A Plug-in Approach to Support Estimation. *The Annals of Statistics*, **25**, 2300-2312. https://doi.org/10.1214/aos/1030741073

[55] Febrero, M., Galeano, P. and González-Manteiga, W. (2008) Outlier Detection in Functional Data by Depth Measures, with Application to Identify Abnormal NO_x Levels. *Environmetrics*: *The Official Journal of the International Environmetrics Society*, **19**, 331-345. https://doi.org/10.1002/env.878

[56] Cuevas, A., Febrero, M. and Fraiman, R. (2006) On the Use of the Bootstrap for Estimating Functions with Functional Data. *Computational Statistics & Data Analysis*, **51**, 1063-1074. https://doi.org/10.1016/j.csda.2005.10.012

[57] Febrero, M., Galeano, P. and González-Manteiga, W. (2007) A Functional Analysis of NO_x Levels: Location and Scale Estimation and Outlier Detection. *Computational Statistics*, **22**, 411-427. https://doi.org/10.1007/s00180-007-0048-x

[58] Peng, L. and Qi, Y. (2008) Bootstrap Approximation of Tail Dependence Function. *Journal of Multivariate Analysis*, **99**, 1807-1824. https://doi.org/10.1016/j.jmva.2008.01.018