A Comparative study of outlier detection of Yamuna River Delhi India by Classical Statistics and Statistical Quality Control

[1,*]Mohammad Ahmad, [2]Ahteshamul Haq, [3]Abdul Kalam and [4]Sayed Kifayat Shah

[1,3]Faculty of Science, Beijing University of Technology, China
[2]Department of Statistics & Operations Research Aligarh Muslim University, Aligarh-202002, India
[4]College of Economics and Management, Beijing University of Technology, Beijing China

[1,*]mahmador@gmail.com, [2]a.haq@myamu.ac.in, [3]faisal.stats@gmail.com, [4]skifss_20@qq.com
*Corresponding author

Abstract

Water quality control aids in preventing pollution, public health, and the preservation and improvement of the biological integrity of water bodies. Water quality involves many variables and observations, some of which are outside of the acceptable range. An observation that apart from the rest of the data or looks diverge from other observation of the sample in which it occurs. In this paper, we proposed two methodologies for detecting outliers for the Yamuna River water quality data with three variables Chemical Oxygen Demand (COD), Bio-chemical Demand Oxygen (BOD) and PH, at three different locations did comparison of these two methodologies. These two methodologies are based on Descriptive Statistics and Statistical Process Control (SPC). A few outliers are present in the data. The outcome shows how far the outlier detection method has progressed and better knowledge of the various outlier methodologies and provide a clear path for future outlier detection methods for researchers.

Keywords: Classical Statistical Analysis, Statistical Process Control, Outlier, Yamuna Water Quality

I. Introduction

The Great Ganga plain is home to around 0.5 Billion population due to the sufficient freshwater availability (Misra [1]). The River Yamuna, its largest tributary, has around 1370 kilometres and originates from the Yamunotri Glacier of Uttar Kashi in Uttar Pradesh (Agarwal et al. [2]). It has several tributaries (Tons, Giri river), which provides fresh water to the mountainous regions. Whereas the Yamuna River flows through the densely populated regions of the plain, including Delhi, Haryana and Uttar Pradesh. While traversing around the megacity of Delhi, which is one of the highly polluted cities (Anand et al. [3]), it covers around 22 km of stretch and receives large quantities (3000 MLD) of partially treated and untreated industrial and domestic waste through twenty-two major drains. The important pollution monitoring stations are located in Kudesi, Nizamuddin and ITO, where three water pollution parameters are constantly monitored: COD, BOD and Negative logarithm of Hydrogen ion concentration (PH). Various wastewater treatment plants are constructed using these parameters; despite this, a significant amount of the untreated water (~1341 MLD) is discharged into the Yamuna River (CPCB, 2004-05). These data mainly indicate the point source, i.e. industrial pollution, although the diffusive sources such as the domestic wastewater supply (washing, cattle wading, cooking, defecation etc.) contribute significant yet unaccounted pollutants to the Yamuna River.

Apart from the source dependency, the levels of the contaminants in the Yamuna River also rely

on the climatic/weather fluctuations such as monsoon rainfall and surface water temperatures. During the heavy rainfalls, the levels dilute, whereas hot summers restrict the vertical water mixing in the river and thus induce the contaminants' spike. The climatic conditions and human-induced pollution thus affect the overall water quality of the Yamuna River and adjoining water reservoirs which are the primary source of drinking water for a significant number of the population living under poverty in and near the Delhi region (Sharma et al. [4]; Bhargava [5]). There might also be information on irregular processes, such as emissions in observations that are not excessively high but deviate from surrounding values. Outliers may merely be noisy observations or, instead, they would suggest atypical activity in the system. These abnormal values are significant and can lead to helpful knowledge or important results and selecting the most effective mitigation techniques or steps. SPC is a process that must work around the goal or nominal dimensions of the quality features with little variability. SPC is a powerful set of problem-solving methods that are useful in achieving process reliability and improving capacity by reducing variability. It is essential to develop and maintain a normal variation pattern through continuous process monitoring. A disturbance has occurred if there is a divergence from the usual fluctuation, and the process must be adjusted. Statistical process control provides data collection, measurement, recording, analysis, and decision making methods. The process is statistically controlled when all disturbances or specific causes of variance are removed. The SPC concept determines the central line, upper control limit, and lower control limit. The process is out of control if the point is above the upper control limit (UCL) or below the lower control limit (LCL) (Torres et al. [6]; Kamalov and Leung [7]). They provide a unique outlier identification method based on principal component analysis and kernel density estimation. The suggested technique is designed to solve the problems associated with high-dimensional data by projecting the original data into a smaller area and calculating anomaly scores for each data point based on the data's intrinsic structure. (Muniz et al. [8]) The study uses oxygen and turbidity as indicator variables to develop a new method for spotting outliers in water quality monitoring metrics. Until now, techniques relied on treating the various parameters as a vector with concentration values as its components. Horn et al. [9] proposed a physician-determined healthy sample, improvement in reference interval estimation utilizing a new outlier identification technique is investigated. The impact of incorporating non-healthy individuals in the sample as determined by a physician is assessed.

Singh et al. [10] sought to bring together an organized and generic overview of several outlier detection strategies. Sim et al. [11] concentrate on spotting potential outliers using the commonly known boxplot software. Outliers are subsets of observations inconsistent with the rest of the observations in a data collection. They find outliers by building a box plot with a lower fence (LF) and an upper fence (UF) (UF). Chakraborti et al. [12] presented phase I parametric control charts for univariate variables. Akarupu et al. [13] conducted a study on five aspects of water quality utilizing Statistical Quality Control methodologies applied to real 2014 data gathered for a water treatment facility in the United States. Fu and Wang [14] introduced several statistical approaches for evaluating water quality data. Three common graphs, boxplots, Q-Q plots, and scatter plots, which provide relevant summary information about datasets, are employed to give insight into datasets. Grubbs [15] study was mainly written as an explanatory and instructive essay on the difficulty of finding outlier observations in an extensive experimental effort. In this work, they solely look at tests of significance. Wang et al. [16] gave a thorough and structured overview of the advancement of outlier identification algorithms from 2000 to 2019. Martinez and colleagues [17] offered the one-class peeling (OCP) approach, a customizable framework for detecting numerous outliers in multivariate data that integrates statistical and machine learning methodologies (Di et al. [18]).

This work presents a novel technique for detecting outliers in water quality monitoring parameters by employing turbidity, conductivity, and ammonium as indicator variables.

II. Methodology

A collection of systems, such as water quality measurements, are available to analyze environmental data. These systems may discover unusual data items using classical analysis, patterns, differences between neighbouring network stations, and predicted values concerning the sampling position. For classical analysis, the data is only statistically evaluated. Today, automated analysis techniques are needed for the amount of data that has been accumulated in environmental databases. The study technique presented here is focused on knowledge discovery of information in databases (KDD) (Chan et al. [19]), which provides a complete data extraction procedure as well as a transparent methodology for preparing data and evaluating the findings produced. The knowledge discovery database provides an iterative and collaborative way of looking for models, patterns, and parameters that are beneficial for outlier detection, categorization, and/or prediction to create information and aid decision-making. We apply statistical techniques to detect the outliers; classical statistical analysis and SPC.

I. Classical Statistical Analysis

Individual time series, descriptive statistics, box plots, and so forth, the classical statistical analysis tracks water quality, decides the importance if any of them falls beyond the limits: quartiles, interquartile range, and evaluate the trend. In general, traditional statistical techniques explain the measurable property distribution (descriptive statistics) and assess the reliability of the sample drawn from the starting population (inferential statistics). Thus, classical analysis is based on continually measuring the characteristics of an item and attempting to forecast the frequency with which the measurement process is repeated stochastically or randomly with a certain conclusion. Properties may be evaluated repeatedly for the same object or only once per object. The classical statistical analysis seeks to assess the empirical frequency distribution that yields the absolute or relative frequency of occurrence of each of the numerous potential outcomes of repeated measurement of an object's property (discrete case) or object class (Torres et al. [6]). If the distribution function is employed in the event of an indefinitely repeated and arbitrarily reliable computation and each outcome is different, then the relative frequency of a particular occurrence will not be very informative.

II. Statistical Process Control

Outliers can be identified by using SPC to monitor the system. The analysis concentrates on significantly low and high readings even if the results do not meet the set limit. These techniques can examine individual or average maps to study individual observation. The dataset should be divided into reasonable subgroups (Shewart [20]). It is important to form rational subgroups because variation can be clustered, and variability can be easily detected in the presence of special causes. Unless it is impractical to utilize the rational subgroups, for instance, when a measurement repeatedly occurs in the same way, it differs only by laboratory or analytical error.

The method of gathering the data is the rational subgroup and generally be collected so that each one demonstrates the only intrinsic variety, which is the natural process of (common cause variation). It allows an additional source of variation (unique cause variation) to be established, which may affect the subgroups imperfectly where possible, to avoid unique cause variation. Moreover, if the mechanism is too violated, the limit of the control chart that defines the border is determined by the variability within each subgroup. For this reason, only subgroups that duplicate the common cause variation in the process should be gathered (Torres et al. [6]).

If the data is appropriately organized, a normality test must be performed. If the normality hypothesis is rejected, there are two possible ways to normalize the data. The first one is to use the modified techniques for non-normal distribution to convert in normality form or transform the data to normalize the data set (Chen [21]). The second technique is used for the transformation is Box-Cox transformation (Box and Cox [22]) is given as follows:

$$X_j^{\omega} = \begin{cases} X_j^{\omega}, & \omega \neq 0 \\ \log(X_j), & \omega = 0 \end{cases}$$

Where ω denotes the maximizes the profile likelihood function of the data X_j

It is possible to divide a classical process analysis into two stages, the first stage when a test has been conducted to remove normality and atypical measurement from the results, and the stage is the control stage, when the pattern is evaluated and when conditions outside of control are encountered. The first level specifies the central line (CL), UCL, and LCL. The control sample accurately defines the centre line, reflecting the objective value. Furthermore, the warning limit is placed at a distance of ±2σ from it and is the operation's standard deviation (Leavenworth and Grant [23]).

For SPC, the control chart of Shewart is most frequently used for its substantial success in detecting the significant changes in a process. It is more accurate to suggest that the control chart is a monitoring system for graphical statistical processes. In most cases, conventional control charts are designed to track process parameters when the underlying form of the distribution of processes is known. Despite these charts using the most recent samples, the minor or progressive improvement in the process is not established. There is a need for complementary rules; different rules have been established by different authors to identify particular deviations (Champ and Woodal [24]; Zhang et al. [25]) and to complement the initial rules. Using these supplementary rules (Western [26]) makes the control charts of Shewart more alert and contributes to a substantial capacity for a non-random sample to be detected.

III. Results and Discussion

In this document, the results of the two approaches are shown below. R and Minitab 16 were used to create all of the figures.

I. Classical Analysis

The traditional statistical method on water quality, time series, descriptive statistics, and box plot analysis was used to see if the value was outside the limit. The table displays the dataset's descriptive statistical parameters. Data is taken from ENVIS Centre on Hygiene, Sanitation, Sewage Treatment Systems and Technology sponsored by the Ministry of Environment, Forest and Climate Change Govt. of India [27].

Table 1: Summary of descriptive statistics of Yamuna river water quality analysis

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum	Range	IQR	Skewness	Kurtosis
KudesiCOD	12	0	72.50	6.20	21.48	28.0	65.00	78.00	86.0	106.00	78.00	21.00	-0.81	0.79
KudesipH	12	0	7.625	0.112	0.386	7.00	7.275	7.700	7.90	8.10	1.100	0.625	-0.57	-0.83
KudesiBOD	12	0	22.77	2.06	7.14	7.60	20.50	24.00	28.0	32.0	24.40	7.50	-1.03	0.82
NizamuddinpH	12	0	7.600	0.0728	0.2523	7.00	7.50	7.60	7.70	8.00	1.00	0.20	-0.90	2.40
NizamuddinCOD	12	0	66.33	2.85	9.87	48.0	58.00	68.00	75.00	80.0	32.00	17.00	-0.64	-0.37
NizamuddinBOD	12	0	21.17	1.54	5.34	14.0	16.75	21.00	23.75	32.0	18.00	7.00	0.53	0.24
ITOpH	12	0	7.733	0.1	0.347	7.00	7.525	7.800	8.00	8.10	1.100	0.475	-0.87	0.08
ITOCOD	12	0	71.33	4.97	17.21	40.0	57.00	72.00	85.0	96.0	56.00	28.00	-0.18	-0.45
ITOBOD	12	0	22.96	1.78	6.18	11.0	17.38	24.50	28.0	32.0	21.00	10.63	-0.55	-0.45

The statistical parameter in table 1 shows the limits and not more than the decided limit. The following step in classical data analysis is to present a time series of monthly data of water quality from 2019 to 2020 (one year) (Fig. 1) ranging of Kudesi COD (28.00,106.00), Kudesi PH(7.00,8.100),

Kudesi BOD(7.60,32.00), Nizamuddin COD(48.00,80.00), Nizamuddin PH(7.00,8.00), Nizamuddin BOD(14.00,32.00), ITO COD(40.00,96.00), ITO PH(7.00,8.10), ITO BOD(11.00,32.00).

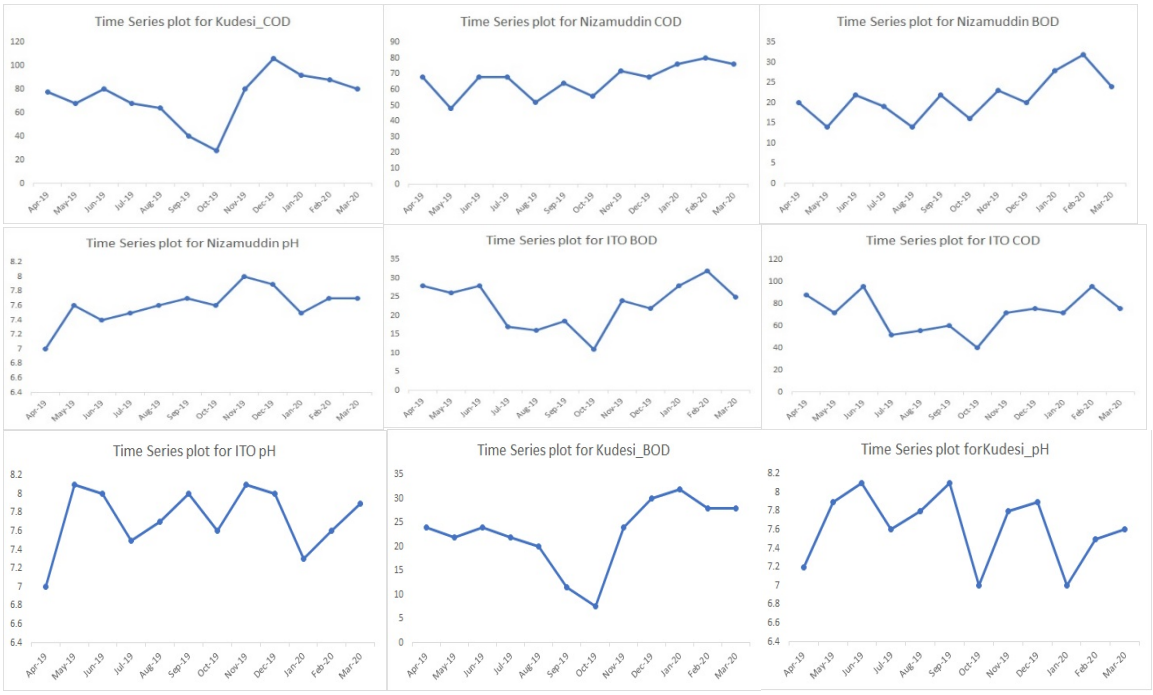

Figure1: Time series plot for Kudesi, Nizamuddin, ITO (COD, BOD and PH)

Figure 2 is a boxplot that graphically depicts COD, BOD, and PH data at various locations concentrated by quartiles. In the below fig., there are no outliers detected in the data except Kudesi COD, BOD and Nizamuddin PH.

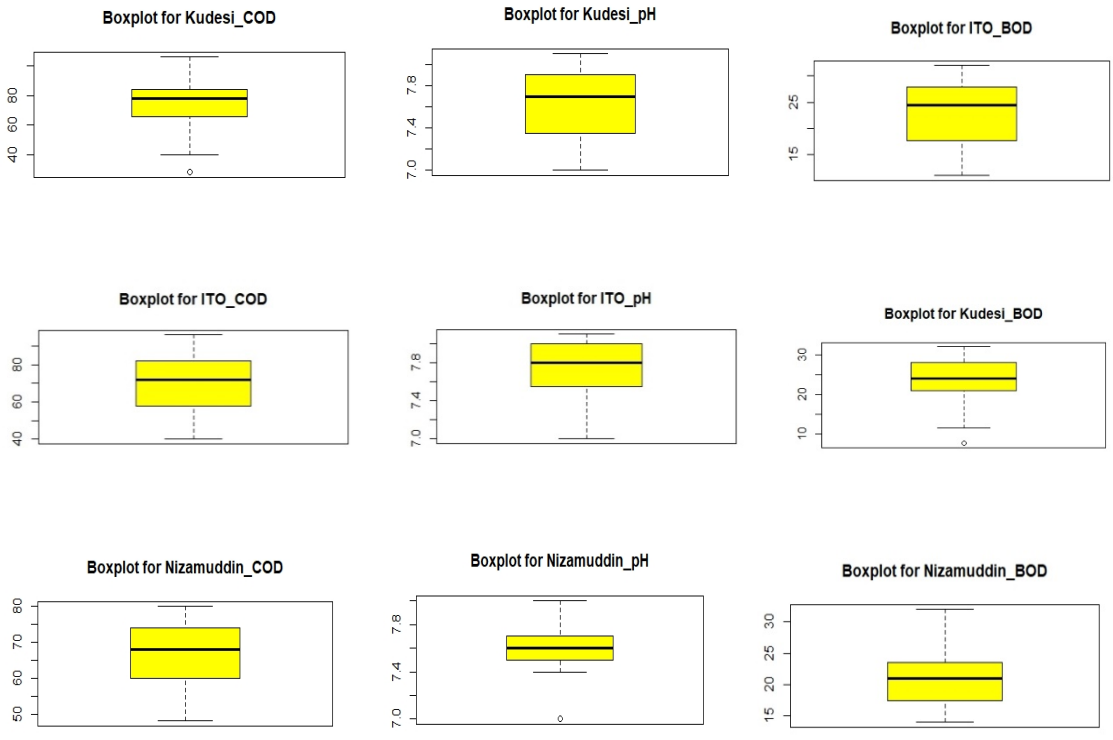

Figure 2: Box plot for Kudesi, Nizamuddin and ITO (COD, BOD, PH)

II. SQC Analysis

Individuals' IMR charts show each observation or measurement as a distinct data point that stands independently (subgroup size = 1). The analysis of the findings given in Fig. 3 reveals that few have a false alarm, i.e. outlier.

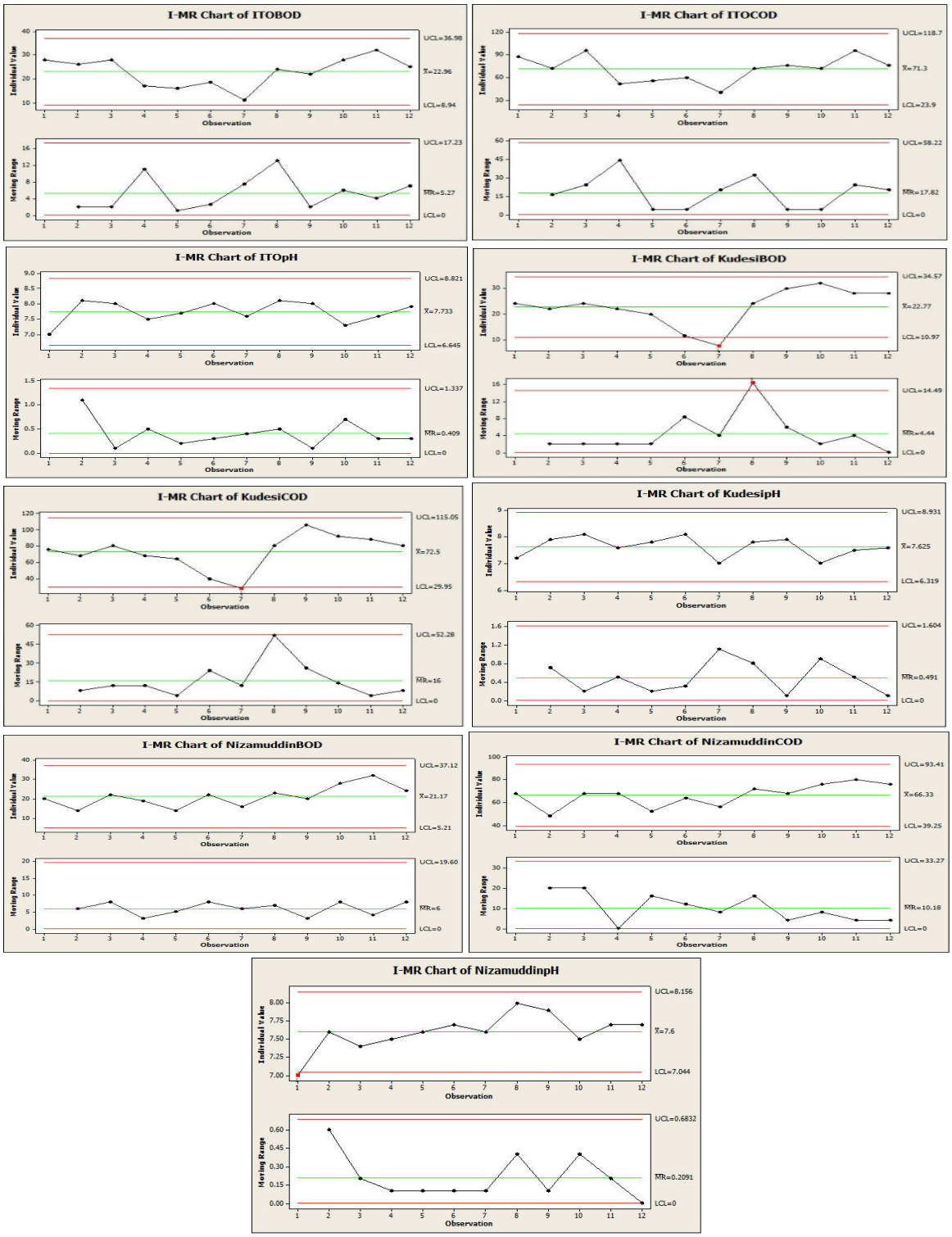

Figure 3: Individual moving range chart for Kudesi, Nizamuddin and ITO(COD, BOD, PH).

The Xbar chart of each observation or measurement of each data point is displayed in fig. 4. In this examination, we observed an outlier detected in the Kudesi COD at sample 7, Kudesi BOD and Nizamuddin PH.

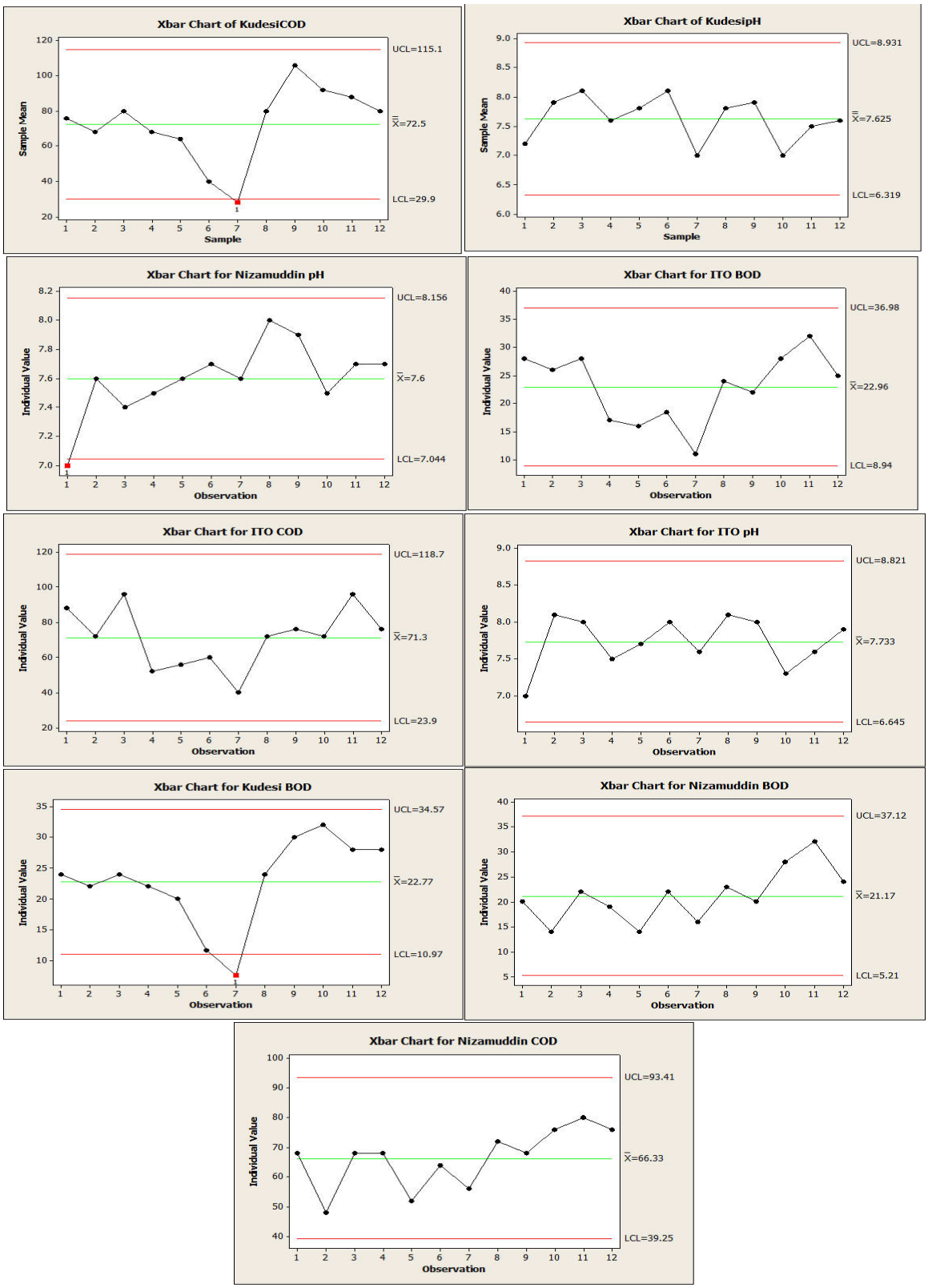

Figure 4: Xbar chart for Kudesi, Nizamuddin, ITO (COD, BOD, PH).

IV. Conclusion

We employed two approaches to analyze water pollution and outliers' data from the urban river Yamuna at three different locations, Kudesi, Nizamuddin, and ITO, using COD, BOD, and PH in Delhi, India. The data were collected monthly from April 2019 to March 2020 with monthly measurements. Firstly, we applied a classical approach by analyzing the data with descriptive statistics such as mean, range, Q1, Q2, Q3 and IQR, time series, and box plot. Secondly, adopted an SPC to learn the approximately normal data, gathered by month with different control charts such as IMR and Xbar chart. A novel method and set of instruments to efficiently access resident water pollution are necessary to effectively enable water pollution abatement and give genuine water quality circumstances. The classical approach is oversimplified despite giving helpful decision-making information. It has many flaws in the data's time correlation structure, including failing to find true outliers months with behaviour deviating from the norm simply because the points do not exceed the bound values. As a result, more complex and modern methodologies can better understand water pollution incidents. SPC is an advanced methodology for identifying outliers in pollution episodes. We create a model and graph it, and this method marks them as outliers. It only works with discrete explanations and cannot extract data in a continuous format. This document outlines a simpler method for environmentalists to discover outliers. We can include it in this functional outlier identification technique for future use. In general, as compared to traditional statistical analysis, SPC is the most effective method for detecting outliers.

References

[1] Misra, A. K. (2010). A river about to die: Yamuna. *Journal of water resource and protection*, *2*(5), 489.

[2] Agarwal, T., Khillare, P. S., & Shridhar, V. (2006). PAHs contamination in bank sediment of the Yamuna River, Delhi, India. *Environmental monitoring and assessment*, *123*(1), 151-166.

[3] Anand, C., Akolkar, P., & Chakrabarti, R. (2006). Bacteriological water quality status of river Yamuna in Delhi. *Journal of environmental biology*, *27*(1), 97-101.

[4] Sharma, M. P., Singal, S. K., & Patra, S. (2008). Water quality profile of Yamuna river, India. *Hydro Nepal: Journal of Water, Energy and Environment*, *3*, 19-24.

[5] Bhargava, D. S. (1985). Water quality variations and control technology of Yamuna river. *Environmental Pollution Series A, Ecological and Biological*, *37*(4), 355-376.

[6] Martínez Torres, J., Pastor Pérez, J., Sancho Val, J., McNabola, A., Martinez Comesana, M., & 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*(2), 225.

[7] Kamalov, F., & Leung, H. H. (2020). Outlier detection in high dimensional data. *Journal of Information & Knowledge Management*, *19*(01), 2040013.

[8] Muñiz, C. D., Nieto, P. G., Fernández, J. A., Torres, J. M., & Taboada, J. (2012). Detection of outliers in water quality monitoring samples using functional data analysis in San Esteban estuary (Northern Spain). *Science of the Total Environment*, *439*, 54-61.

[9] Horn, P. S., Feng, L., Li, Y., & Pesce, A. J. (2001). Effect of outliers and non-healthy individuals on reference interval estimation. *Clinical chemistry*, *47*(12), 2137-2145.

[10] Singh, K., & Upadhyaya, S. (2012). Outlier detection: applications and techniques. *International Journal of Computer Science Issues (IJCSI)*, *9*(1), 307.

[11] Sim, C. H., Gan, F. F., & Chang, T. C. (2005). Outlier labeling with boxplot procedures. *Journal of the American Statistical Association*, *100*(470), 642-652.

[12] Chakraborti, S., Human, S. W., & Graham, M. A. (2008). Phase I statistical process control charts: an overview and some results. *Quality Engineering*, *21*(1), 52-62.

[13] AKARUPU, V., GUNKALA, S., PATTIGADAPA, S., PATTIGADAPPA, B. K., PONNAPALLI, S., & SEGALL, R. S. (2016). Statistical Quality Control and Improvement of Waste Water Treatment Plant. In *Proceedings of The 20th World Multi-Conference on Systemics, Cybernet and*

Informatics.

[14] Fu, L., & Wang, Y. G. (2012). Statistical tools for analyzing water quality data. *Water quality monitoring and assessment*, *1*, 143-68.

[15] Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, *11*(1), 1-21.

[16] Wang, H., Bah, M. J., & Hammad, M. (2019). Progress in outlier detection techniques: A survey. *Ieee Access*, *7*, 107964-108000.

[17] Martinez, W. G., Weese, M. L., & 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*(4), 1272-1295.

[18] Di Blasi, J. P., Torres, J. M., Nieto, P. G., Fernández, J. A., Muñiz, C. D., & 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.

[19] Chan, K. C. C., Wong, A. K. C., Piatesky–Shapiro, G., & Frawley, W. J. (1991). Knowledge Discovery in Databases.

[20] Shewhart, W. A. (1931). "*Economic control of quality of manufactured product"*. Macmillan And Co Ltd, London.

[21] 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*(7), 602-610.

[22] Box, G. E., & Cox, D. R. (1964). "An analysis of transformations". *Journal of the Royal Statistical Society: Series B (Methodological)*, *26*(2), 211-243.

[23] Leavenworth, R. S., & Grant, E. L. (2000). "*Statistical quality control"*. Tata McGraw-Hill Education.

[24] Champ, C. W., & Woodall, W. H. (1987). "Exact results for Shewhart control charts with supplementary runs rules". *Technometrics*, *29*(4), 393-399.

[25] Zhang, M. H., Lin, W. Y., Klein, S. A., Bacmeister, J. T., Bony, S., Cederwall, R. T., & Zhang, J. H. (2005). "Comparing clouds and their seasonal variations in 10 atmospheric general circulation models with satellite measurements". *Journal of Geophysical Research: Atmospheres*, *110*(D15).

[26] Western Electric, C. (1956). "Statistical quality control handbook".

[27] Water quality Status of River Yamuna Water quality available online: Water Quality Status of River Yamuna (sulabhenvis.nic.in) (accessed on March 2021).