Grenze International Journal of Engineering and Technology, Jan Issue

Determination of the Principal Factors of River Water Quality of Mahanadi Basin, Odisha using Multivariate Techniques

Abhijeet Das
Temporary Contractual Faculty, Department of Civil Engineering, Odisha University of Technology and Research (O.U.T.R), Bhubaneswar, Odisha-751003, India
Email: das.abhijeetlaltu1999@gmail.com

***Abstract*—In this work, cluster analysis (CA) and principal component analysis (PCA)/factor analysis (FA) were used to analyse the quality of water to detect and measure the causative factors affecting the surface quality of the Mahanadi River in Odisha. 20 water quality characteristics comprising 19 sampling stations were employed in a 1-year (2020–2021) dataset for this purpose. To reflect the various water quality criteria of the river system, hierarchical CA was used to arrange the 19 sampling locations into three comparable groups. Highly polluted Cluster III is claimed to include Cuttack D/s (R-8), Choudwar D/s (R-19), and Paradeep (R-9). The data matrix's total variance was explained by the PCA/FA method's five most significant variables, which accounted for 93.9 % of it. TSS, TKN, EC, TDS, B, TH, Cl^- and SO_4^{2-} have all been correlated with (loading > 0.7) in the 1st PC, which accounted for 43.133 % of the total variance. COD, NH_3-N, Free NH_3, SAR; pH, F were all linked to the 2nd PC, which accounted for 23.055 % of total variation, whereas the 3rd, 4th, and 5th PCs, while accounting for 12.866 %, 8.603 %, and 6.241 % of total variation, respectively. Five separate VFs with eigenvalue > 1 were justified by the principle of Maximum Variance Rotation of the PC i.e., VF-1 (40.191%), VF-2 (21.358%), VF-3 (14.593%), VF-4 (9.082%), and VF-5 (8.675 %) of total variance (93.899%). The factors responsible of river water contamination/deterioration were discovered to be point source contamination discharges from human impacts caused by the discharge of agricultural residues, home and industrial wastewater. Additionally, it comprised of nutrients (agricultural runoff), alkalinity, hardness, EC, solids as well as soil weathering, leaching from landfills and industrial waste disposal sites. Comparing the findings of the CA and PCA/FA approaches revealed some notable disparities in the estimated contributions for each probable source of pollution. The PCA/FA procedure was found to be more physically reasonable in the end for the current investigation. According to expectations, the findings could be highly helpful to the local government in managing and controlling pollution and better safeguarding vital riverine water quality.**

***Index Terms*— Cluster analysis (CA), Principal component analysis (PCA), Mahanadi River, Water quality, Eigenvalue, Human impacts.**

I. INTRODUCTION

Since it serves as the material foundation for the existence of all Earth species, water is the most important prerequisite for maintaining the planet's natural ecosystem over the long term (Nowak et al., 2018). The gap

Grenze ID: 01.GIJET.9.1.648

between availability and demand for water resources has widened as a result of increased water use, necessitating stricter regulations for the use of and preservation of waterbodies (Kazi et al., 2009). Society, irrigation, human consumption and industry, rivers are the most important source of water. Information on the quality and fluctuation of river water is critical for successful operation of these water resources. The majority of river contamination originates through commercial and residential wastewater along with drainage in farming (Carpenter et al., 1998). River pollution must be minimized and controlled, as well as reliable state of water quality, for effective management, given that rivers are the dominant contributor of inland water for domestic, industrial, and agricultural requirements. It is crucial in the assimilation and removal of urban and industrial wastewater along with runoff from agricultural land. Rainwater is a monsoonal activity influenced mostly by the catchment's weather. River flow is affected by seasonal variations in rainfall, stream flow, interflow, and groundwater flow, along with pumping into and out of, as a result, pollution accumulation (Gashi S et al., 2016). Seasonal changes in both human and natural processes that affect the river's hydrology that result in multiple functionalities between seasons which are influenced by fluctuations due to natural and human mechanisms (Vega et al., 1998). As a result, for integrated approach of these water supplies, continual monitoring and evaluation of river's water quality is essential (Singh et al., 2005). The quality of rainwater has always been driven by environmental processes in a given vicinity meteorological contributions, erosion, and degradation of volcanic activity elements, and even some manmade impacts such as increased water resource exploitation and city, commercial, and farming activity (Carpenter et al., 1998). While urban and industrial waste outflow is a persistent pollute sources, surface water runoff is a cyclical phenomenon that is highly altered by the climate of the watershed (Vega et al, 1998). Twine et al., 2005 investigated the environmental influences like temperature fluctuations, soil degradation and rainfall as well as human involvement influence the quality of a region's surface water such as over-exploitation of waterways from residential and industrial sources. Disposal of metropolitan sewage and industrial wastewater is a constant threat to the environment, proper sewage discharge control is essential for enhancing water quality (Mohammaad et al., 2020). Sampling networks appear to be a useful data source for determining the river's water quality on a local and temporal perspective, as well as monitoring its quality. These connections provide a snapshot of the ecosystem's temporal state, as well as its seasonal and spatial change (Simeonov et al., 2003). Statistical approaches have all been used in the academic papers for this sort of investigation because they can predict new level of pollutants by assessing spatiotemporal variability in river water quality (Khan et al., 2016). Surface and fresh water quality are widely depicted and evaluated using multivariate data processing, which is especially useful for displaying seasonal temporal and spatial cycles induced by natural and anthropogenic influences (Helena et al., 2000). Simple methods for assessing environmental sustainability include multivariate analysis, such as cluster analysis (CA) and principal component analysis (PCA). The quality of Indian rivers' water, especially the Mahanadi, has deteriorated over the last several decades as a result of the continual point and non-point outlets release partially/untreated sewage water, urban runoffs, and sewages (Mishra and Kumar 2020). Some of the river's tributaries and distributaries are likewise heavily polluted. These results are useful for categorizing samples, reducing WQ parameters, and discovering the correlations between them (Praus, 2007). CA is a way of categorizing things (cases) resemblances within a class and differences between classes (clusters). The outcomes help with data analysis and trend discovery (Vega et al., 1998). (Bhardwaj et al. 2010) adopted PCA to identify the components of water quality (WQ) and potentially even the origin of contaminants released by household and commercial activities. PCA demonstrated that all physicochemical factors in the river Mahanadi basin contributed equally and strongly to WQ changes, while various connections seen between stations were shown by CA, indicating WQ features (Yerel & Ankara, 2012). Thareja et al. 2011 also illustrated that the PCA approach proved to be an effective tool for identifying crucial river WQ monitoring and indicators. Sharma et al. (2015) conducted PCA and CA components, as well as correlation analysis, to assess seasonal patterns, possible pollution causes, and the clustering of Ganga and Yamuna River monitoring locations in Uttarakhand (India). PCA is a robust technique for limiting the components quantity of a data set with several relevant variables while keeping as much diversity as desirable. The amount of data collected is decreased by converting it into a fresh range of orthogonal (non-correlated) variables termed as principal components (PCs), which are then sorted in declining order of importance. By twisting the PCA axis, one can create new variables called varifactors (VFs). By restricting the number of large and moderate coefficients, Varimax rotation distributes the PC loadings so that their dispersion is maximized (Richman, 1986). In addition to significant data reduction without compromising too much substance, the entire data set's variability can be characterized by a small number of VFs. Additionally, the usage of VFs to categories the variables under study as per their distinct qualities encourages data interpretation (Simeonov et al., 2003). It was discovered that quantitative methods like PCA and CA were helpful in finding the underlying connections between data on water quality, pinpointing the sources of pollution, and constructing clusters of identical monitoring stations with comparable characteristics. As a result, reliable identification of available scenarios of surface water quality deterioration is crucial for the regulation of water quality. As a result, this method can be used to find sources and factors that potentially affect water systems and result in changes to the water quality (Chattopadhay T et al., 2012). In the present research, a study was conducted on the Mahanadi River, setting up 19 detection stations and evaluating the features of 20 physicochemical water samples. It was decided to use a 1-year detection period. The influence of water quality indicators on fluctuations in surface water quality over time was examined using multivariate approaches as Pearson correlation strategy, PCA, and CA. These techniques also helped identify the sources of contamination that are affecting water quality.

II. Study Area

In Chhattisgarh's Raipur district in the far south-east, the Mahanadi River appears to be a little pool about 6 kilometres from Pharsiya Township in the Amarkantak Mountains of the Bastar plains. It runs for 494 kilometres in Odisha, out of a total length of 851 kilometers. In Odisha, the main tributaries are Ib, Ong, Tel, Hariharjore, and Jeera, while the primary distributaries are Kathajodi, Kuakhai, Devi, and Birupa. Nearly 400 kilometres from the mouth of the Mahanadi, the multipurpose Hirakud Dam in Sambalpur is situated directly in the middle of the main stream. Near Bagra, the river Ib joins the Mahanadi and flows to the left of the Hirakud reservoir. The river travels south from Sambalpur until it meets Ong and Tel. The river travels eastward again at Sonepur, the Mahanadi's greatest tributary, Tel, finally joining the Bay of Bengal. The river runs through the Eastern Himalayas before reaching the coastal region and forming the delta, cutting through a 60-kilometer-long "Satkosia Gorge" bordered by high hills and tropical vegetation. The river finally exits the Eastern Ghats in the vicinity of Naraj, some 10 km away west of Cuttack. The Kathjodi and Mahanadi, two sizable distributaries, split the riverbed around Naraj, beginning the deltaic flow. It has created a massive delta by passing through the towns of Cuttack and Puri and a variety of distributaries from westward. The Mahanadi River receives minimal external rainfall replenishment and gets its water primarily from sewer systems in cities and wastewater treatment plants. It is the primary inland water supply used for home, industrial, and irrigation uses, as well as incorporating or eliminating urban and industrial pollutants and farming runoff (Yidana et al., 2011). As a result, river pollution must be avoided and controlled, and accurate water quality data must be available for successful management. Graphical representation of River Mahanadi along with sampling sites is being represented using Arc GIS 10.3 software as shown in (Figure 1).

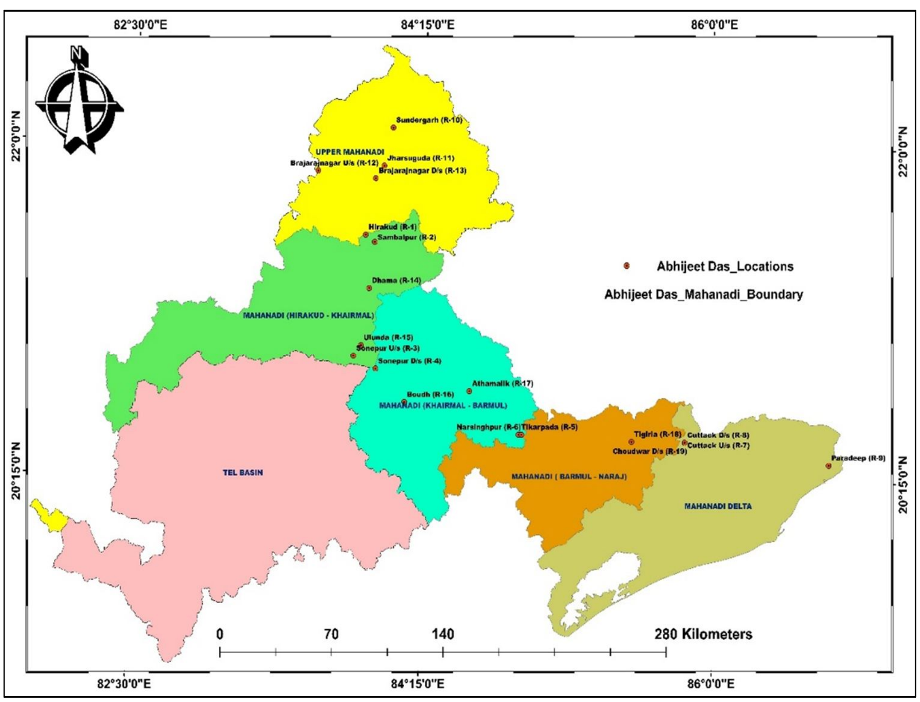

Figure 1. Map showing water quality monitoring sites of Mahanadi River

III. Sampling Collection and Pre-Treatment

Water samples were collected, preserved, and transferred to the laboratory according to standard methods for the Mahanadi River's water quality monitoring programme (APHA, 1992). A mercury thermometer was used to test the temperature of the water on the spot. All of the other parameters were evaluated in the lab using standard techniques (APHA, 1992). The data collection included 20 water quality metrics that were observed over the course of 1 year i.e. 2020-2021 by 19 sampling sites (Figure 1) namely Hirakud (R-1), Sambalpur (R-2), Sonepur U/s (R-3), Sonepur D/s (R-4), Tikarpada (R-5), Narsinghpur (R-6), Cuttack U/s (R-7), Cuttack D/s (R-8), Paradeep (R-9), Sundergarh (R-10), Jharsuguda (R-11), Brajarajnagar U/s (R-12), Brajarajnagar D/s (R-13), Dhama (R-14), Ulunda (R-15), Boudh (R-16), Athmalik (R-17), Tigiria (R-18) and Choudwar D/s (R-19) . This investigation looked at a number of elements like pH, Dissolved oxygen (DO), Biochemical oxygen demand (BOD), Chemical oxygen demand (COD), Total coliform (TC), Total suspended solids (TSS), Total alkalinity (TA), ammonia-nitrogen (NH_3-N), Free ammonia (Free NH_3), Total Kejdahl Nitrogen (TKN), Electrical Conductivity (EC), Total Dissolved Solids (TDS), Boron (B), Sodium Absorption Ratio (SAR), Total Hardness (TH), Chloride (Cl-), Sulphate (SO_4^{2-}), Fluoride (F-), Nitrate (NO_3^-) and Iron (Fe). Careful standardization, systematic blank observations, spiked and duplicate samples were used to assure the analytical data quality. Water samples were obtained using a standard open cartridge sampler in order to analyze the water's quality (1.5 litre capacity). To ensure that the data is representative, this sampler may collect water samples at varied water depths. Before taking a sample of the water, the 2-litre polythene bottle was cleaned with a metal-free solution, rinsed many times with distilled water, steeped in 10% nitric acid for 24 hours, and then rinsed once more with ultra-pure water. The water samples were kept in an insulated chiller and kept at 4°C on the day they were collected before being transferred to the lab for analysis. All reagents used for ease of understanding were analytical grade for greater efficiency and correctness (Merck, India). Some samples were sent to water quality laboratory of Central Water Commission (C.W.C), Bhubaneswar for analysis of physicochemical indices for examination and proper treatment. All these verification purposes are being carried out with extra care and precautions so that accuracy should be maintained.

IV. Data Treatment and Multivariate Statistical Methods

Although water sampling was accomplished at all locations every year, due to adverse weather, certain locations were unable to be surveyed, and the missing data was filled by the average value. The fundamental statistics of the one-year water quality data frame were conducted and executed well in (Table 1**)**. Prior to conducting the multi-variable data study, each factor was examined for compliance to the normal distribution by investigating skewness and kurtosis characteristics. The results typically follow a normal distribution. According to the findings, all constituents fall into a range that is close to or equal to the normal curve. Skewness and kurtosis were calculated in the ranges of 1.87 to 4.36 and -0.84 to 19, respectively. To accommodate for the varied orders of magnitude and intensity values of numerous water quality indicators, all feature selection were z-scaled standardized with average 1 and variability 0 for CA and PCA. All datasets were evaluated and the results to investigate the factors that affected sources of contamination at various spatial scales. An extensive volume of water quality of data needs to be discussed for efficient pollution control and water sustainability. Experimental data were subjected to multidimensional research of river water quality datasets by PC, CA and FA/PCA and were standardized by z-scale conversion to prevent misinterpretation owing to significant discrepancies in attribute values. For leadership to be successful, river pollution must be reduced and trustworthy water quality data must be mastered. All mathematical and statistical calculations were performed using Excel 2013 and MATLAB 2016.

A. Pearson's Correlation (Pc)

In reality, the coefficient of correlation is used to quantify the interdependencies and level of correlations between variables. The number +1 denotes a perfect association between the two at a significant threshold of $p < 0.05$, whereas 1 denotes a perfect link between the variables despite their inverted fluctuations (Mudgal et al. 2009), and a value of zero implies that the variables have no link (Mudgal et al. 2009). In general, coefficient values of $r > 0.7$ are seen to indicate high correlation, while r values between 0.5 and 0.7 are thought to indicate moderate correlation and r values below 0.5 are thought to indicate weak correlation.

B. Cluster Analysis (Ca)

CA is an unsupervised classifier that unveils the structural information or fundamental behavioral patterns of a data frame without attempting to make any approximations in order to identify or conglomerate the system's

TABLE I. DESCRIPTIVE STATISTICS DATA OF SURFACE WATER PARAMETERS OF MAHANADI BASIN

Sl. No	Variables	Min	Max	Avg	Standard Deviation	Skewness	Kurtosis
1	**PH**	7.74	7.92	7.82	0.05	0.02	-0.84
2	**DO**	7.26	7.83	7.68	0.14	-1.87	3.66
3	**BOD**	1.05	2.40	1.36	0.34	1.84	4.19
4	**TC**	1212.40	42529.20	5151.25	9193.97	4.15	17.64
5	**TSS**	28.63	74.90	39.33	11.56	1.89	4.06
6	**TA**	70.40	100.90	85.71	8.24	0.07	-0.61
7	**COD**	6.76	21.88	11.25	3.96	1.63	2.52
8	**NH_3-N**	0.51	1.93	0.66	0.31	4.03	16.97
9	**Free NH_3**	0.02	0.06	0.03	0.01	2.04	5.71
10	**TKN**	3.28	11.80	5.73	2.07	1.49	2.94
11	**EC**	138.10	7779.35	580.86	1743.33	4.36	18.99
12	**TDS**	82.30	13230.60	812.80	3007.19	4.36	19.00
13	**B**	0.03	0.55	0.08	0.12	4.02	16.85
14	**SAR**	0.41	16.59	1.34	3.69	4.36	18.99
15	**TH**	51.20	2195.20	186.45	486.60	4.35	18.97
16	**CL□**	9.65	4904.91	269.23	1122.58	4.36	19.00
17	**S0□$^{2-}$**	4.97	376.07	26.43	84.68	4.36	18.99
18	**F**	0.26	1.00	0.37	0.17	3.26	11.13
19	**NO□$^{-}$**	1.29	2.70	2.00	0.41	-0.22	-0.66
20	**Fe^{++}**	0.60	2.61	1.31	0.46	1.04	2.41

objects based on their relative position or similarity without making any predictions (Vega et al., 1998). The hierarchical CA strategy employed in this research is the most often utilized clustering method. Through sequential agglomeration, this strategy categorizes the data that is the closest or most identical, and then organizes these clusters into coherent groups. The Euclidean distance is used to determine whether two samples are comparable, and the "distance" is alluded to as "difference" between the two samples results (Zhang et al., 2011). The Ward technique was employed, and the squared Euclidean distance was adopted as a measure of similarity in a hierarchical aggregate CA on a scaled data structure. Using analysis of variance, the sum of squares of the two main groups produced in each stage was adjusted to account for the size of the clusters. Using homology, CA groups water quality records, producing a spatiotemporal structure of samples. It creates a dynamic depiction, displaying a snapshot of each unit and those nearby, despite the fact that the original data's dimensions have been drastically reduced.

C. Principal Component Analysis (PCA)/ Factor Analysis (Fa)

PC provides details on the most crucial aspects of the data gathering, enabling for reduction in data with minimal original data loss (Helena et al., 2000). It's a robust image recognition strategy that seek to explain the variation of a wide number of interrelated parameters by reducing them to a smaller number of unrelated variables (principal components). This technique is used to extract the eigenvalues and eigenvectors of original variables from the covariance matrix. When interrelated variables are multiplied by the eigenvector, uncorrelated (orthogonal) variables, or PCs, are created (a set of coefficients or loadings or weights). Its key aim is to drastically restrict the number of variables needed to properly account for the variation in the initial data while also changing a substantial percentage of highly associated elements into independent or unconnected variables (LiH et al. 2017). The basic principle behind principal component analysis is to create the "optimal" fitting line for n points, with the minimum squared of the distance away between these n points and the line, and this line is known as the first principal component (Gonzalez MJG et al., 2018). FA is a multivariate modelling approach that illustrates a few possible indices to demonstrate the covariance connection between multiple aspects (factors) (Ebrahimiyan M et al. 2013). PCA's contribution to the involvement with less key factors is further decreased by twisting the axis designated by PCA, and a new set of variables known as VFs is recovered. To find significant PCs from scaled variables, PCA was utilized (water-quality data frame) and to limit the contribution of low-value variables even more; these PCs were then rotated (raw) to generate VFs. The factor in this study is a good approximation of the initial variables, which were formed by adopting the maximum variance criterion (Ding S et al. 2011). In general, only factors with eigenvalues greater than one are considered in the analysis.

V. Results and Discussion

Over the course of 1 year (2020-2021), water quality monitoring of the Mahanadi River was examined out at nineteen different locations. The linkage between the variables is found using the Correlation analysis matrix, which is also used to calculate the goodness of fit. These correlations show that waste water from both home and industrial sources, as well as its organic load, is discharged into the river. The correlation empirical findings are taken into account in the ensuing interpretation. Correlation matrix analysis (Table 2) demonstrates the high positive association exist between (TH, Cl^-, SO_4^{2-})-B, , EC-TSS, TDS-TSS, SO_4^{2-}- Cl^-, B-TSS, TH-TSS, Cl^--TSS, SO_4^{2-}-TSS, Fe^{++}-TSS, (NH_3-N)-COD, Free NH_3-COD, F^--COD, TC-BOD, TKN-TSS ,Free NH_3 – NH_3-N, F^-- NH_3-N, (TDS, B, TH, Cl^-, SO_4^{2-})-EC, (TH, Cl^-, SO_4^{2-}, B)-TDS, F^--SAR, and (Cl^-, SO_4^{2-})-TH. This denotes that the parameters are changing in a straight proportional manner. Some parameters were discovered to have a strong negative correlation, indicating that they vary in inverse proportionality. These connected characteristics might be viewed as the primary cause of water quality fluctuations throughout time. The non-significant association suggests that anthropogenic sources play a role in the catchments. The data on water quality were then applied to a variety of multivariate statistical methods in order to investigate trends in time and space. Cluster analysis generated a dendrogram (Figure 2), dividing the river's 19 sampling sites into three statistically significant groupings. Because the sites in these groups have comparable characteristics and natural background source types, the clustering technique yielded three distinct groupings of sites, each of which is highly persuasive. Cluster I (R-1, R-3, R-5, R-7, R-10, R-11, R-12, R-15, R-16 and R-18), Cluster II (R-2, R-4, R-6, R-13, R-14 and R-17) and Cluster III (R-8, R-19 and R-9) corresponds to a relatively less polluted zones, moderately polluted zones and highly polluted zones respectively. The low-pollution sector, which was in the lower region, was unpolluted from outside. The waterway has surpassed the cleansing of manmade wetlands and even its self-purification potential, indicating that it is in the low-pollution zone. The industrial effluent released into the river primarily polluted the medium pollution zones. In case of highly-pollution zones, it was the direct dumping of neighboring rural residential sewage into rivers and streams. As urban living standards have improved, household water demand has outpaced sewage treatment plant capacity, leads to poor water quality. In case of cluster I, at these sites, they transport less pollution to the reservoir. Rajnandagaon, Bhillai, Durg, Shimoga, Raipur, Bilaspur, and Korba are just a few of the major sites and sectors along the Mahanadi River's banks, and these alone contribute a large pollution load to the reservoir. The river in Odisha discharges its impurities into the reservoir. The river runs through a region near Sonepur (U/s) that is devoid of large urban communities or waste water outfalls. At this juncture, the Mahanadi and its two significant right bank branches, Ong and Tel, unite. As a consequence, Sonepur U/s, which is positioned directly below the Ong confluence, has remarkable water quality. Despite the fact that Sonepur is the district headquarters and hosts all of the district's operations, the water quality has not deteriorated. There is no industry or urban habitation on the river's 102-kilometer run from Sonepur D/s to Tikarpada (two minor towns in sub-divisions - Boudh and Athamalik), and there seem to be no large wastewater outfalls in this area. In the late 1990s, a paper mill operated in Sundergarh, Jharsuguda, and Brajarajnagar U/s, but it has been shut down since December 1998, and because none of the three municipalities is a big metropolitan center and there is little organized residential wastewater disposal to the river. Jharsuguda's role as the state's industrial hub has lately increased. Industrial activity, on the other hand, have a minor impact. The remaining sites are pollution-free. TC is the metric that causes the water quality to deteriorate in all of these sample sites. In cluster II, with a population of over 1.5 million people, Sambalpur is the state's largest city. The district, division, and lakhs headquarters are all located directly downstream of the Hirakud reservoir (about 5 km). The Mahanadi in Sambalpur is employed for bathing and wastewater (untreated) discharge in addition to providing drinking water, which is prompting the water supplies in Sambalpur D/s to deteriorate. Other locations, such as Dhama and Athamalik, encounter or tolerate moderate pollution loads, while Brajarajnagar D/s contributes less to water quality due to the shutdown of the paper mill. Sonepur D/s is connected to the Tel basin, which has a large yearly total flow and relatively low load. Furthermore, although being the administrative center, it is still a tiny town with minimal evidence of urban expansion (population around 19000). The remaining locations, Narshingpur, Dhama, and Athamalik, have a moderate level of pollution. Except for Sambalpur D/s, where BOD and TC threaten water quality, TC is the dominant water quality degrading metric at all sites. In cluster III, it is classified as extremely contaminated. The river enters its deltaic zone between Narsinghpur and Cuttack (approximately 56 km), which is marked by high population density and intensive agricultural activity. As a result, the river water ingress into Cuttack D/s has deteriorated, in terms of TC, FC, TKN, and EC. The river receives a large amount of dirty water from the city (population: 5.35 lakhs), causing the water condition at Cuttack D/s to degrade even more. When it comes to Choudwar D/s, it used to be the home of important industrial businesses like a textile mill, a large pulp paper industry, and a charge chrome factory with a thermal power plant. At the moment, only the charge chrome enterprise is in existence, and water contamination is just a small concern. Many industrial clusters in Paradeep, such as oil refineries, iron and steel, thermal power, fertilizer, and breweries, have influenced and thus exerted impacts on water and riverine discharges, which have contributed pollutants carried by agricultural runoff, factory flush outs, mine discharges, and state and city discharges from its catchment to the estuarine sea. The seashore is constantly under strain from both natural and anthropogenic causes due to the significant pollution impact. It could be caused by overfishing, sporadic fish culture, pollution from industries, port activities such as discharge and dredging, marine transport, associated barge discharges and accidents, and sediment drifting from riverine systems. In case of Paradeep, multiple parameters like TC, EC, SAR, and Cl^- threaten the water system. To examine the constituent trends among the studied water samples and find the variables that affect each one, FA/PCA was applied to the entire dataset. To test whether the data were accurate and valid, the Kaiser-Meyer-Olkin (KMO) and Bartlett's sphericity examination were used to study the relationship and partial link between elements. The KMO statistic (LoF, Hong et al. 2012) appears to have a value between 0 and 1. The factor analysis data is judged to have a superior effect when the KMO value is greater than 0.7. The KMO outcome is 0.716, and Bartlett's sphericity results as 1321.21 ($p < 0.05$), demonstrating that PCA may potentially reduce complexity. In order to identify the critical variables influencing each water sample, FA/PCA report examines and compares the composition patterns between water samples using reliable and valid data (Koutsias et al. 2009). 5 principal components (PCs) were retrieved from the PCA of all sets (Table 3), accounting for 93.9 % variance and possessing eigenvalues higher than 1. The 1st PC (43.133% variability) was linked with loading >0.7 with TSS, TKN, EC, TDS, B, TH, Cl^-, SO_4^{2-}. However, the 2nd, 3rd, 4th, 5th PCs although compensating for the variance explained of 23.055%, 12.866%, 8.603% and 6.241% respectively and loadings of COD, NH_3-N, Free NH_3, SAR; pH, F; BOD and TC; TA and NO_3^-; and Fe, DO and pH respectively. It quantifies emissions linked to enterprises like the development of handmade paper, chemical raw materials, and chemical products, along with the primary industrial industries present at the sampling sites, by combining the localized industrial structure and distributions. A scree plot (Figure 3) is employed to compute the proportion of PCs to be maintained (Jackson 1991; Vega et al., 1998). After the eighth Eigen value, the plot indicated a considerable shift in slope. Andrew's plot (Figure 4) illustrates how higher-intensity loadings are more susceptible to pollution and can impede human and animal growth and lifestyle. The higher loading in brown color indicates that the primary downgrading parameters responsible for station Paradeep are TC, EC, SAR, Cl, while the key downgrading parameters responsible for Cuttack (D/s) are FC, TKN, and TC, which clearly demonstrates slight to moderate pollution. This map explains the loading of each parameter and the amount to which it can influence the position of all sampling sites in this way. To put it another way, this graphic encourages the use of the varimax rotation technique, which produces new indices that are primarily made up of a subset of the variables used initially. The original variables will be split up into numerous distinct groups by the PCA rotation axis, which will generate a new sequence of factors with each having a unique influence, primarily consisting of similarity measure variables (Dominick D et al., 2010). Therefore, the current Mahanadi River database's quantitative study i.e., FA minimizes the influence of negligible factors that PCA contributes. Five distinct varifactors (VFs) with Eigen values > 1 were described by the maximum variance rotation of the PC (Original), which accounted for 93.899 % variance. A scree plot (Figure 5) is also employed to compute the proportion of VFs to be maintained (Jackson 1991; Vega et al., 1998). After implementing the maximum variance approach to rotate the data, the value of PC was exposed even more, and the original variable's involvement in VF became evident (Table 4). Lieu et al. (2003) categorized the factor loadings classified as "strong," "moderate," and "weak," with comparative loading values of > 0.75, 0.75-0.5, and 0.50-0.30, respectively. VF-1 (40.191 percent of the entire variability) had positive components that are quite robust on Cl^-, TDS, SO_4^{2-}, EC, TH, B, TSS, Fe, and TKN, indicating metal content and solid waste decomposition. This is because the primary source of river drainage is residential sewage facilities, the primary contaminant is domestic pollution sources, and the major source of heavy metal contamination is metal factories upstream of the tributary. The source was industrial wastewater contamination, since VF-2 (21.358 percent of total variation) exhibited substantial positive loadings on NH_3-N, SAR, F, Free NH_3, and COD. This is related to the main paper business in Brajarajnagar D/s, and while the mill is closed, some paper-making activities continue on a lesser scale, degrading the quality and posing a health risk. Some metal goods and steel casting manufacturing industries near the river also contribute to water quality degradation, which can harm or endanger human health. High loadings (significant) on TC and BOD and strong loadings (negative) on DO were seen in VF-3 (14.593 percent of total variance), which were primarily manifested as pollution from heavy industries, fertilizers, thermal power plants, agricultural runoff, industrial flush outs, mine discharges, municipal discharges, and other sources that contain organic substances and encourage the growth of decomposers that consume large amounts of them. This clustering suggests that the contamination came from waste water discharged by the chemical sector, which is linked to a number of chemical and metal-producing businesses across the river. VF-4 showed a high nitrate loading and a moderate TA loading, indicating that the contamination emanated from municipal, industrial, and sewage discharges, accounting for 9.082 percent of the entire variance. High nitrate levels promote eutrophication because nitrate can enter water bodies. The intestinal mucosa produces less nitrite, which subsequently interacts with the circulation to produce methemoglobin, hindering oxygen delivery, high nitrate intake or consumption offers a significant risk. VF-5 (8.675% of entire variation) has a high positive factor on PH, making it mildly alkaline in nature. It's because home sewage contains detergents and personal hygiene products. The findings revealed that most alterations were constituted of dissolved salts (natural) and organic pollutants (artificial). The principal contaminants are Cl^-, TDS, SO_4^{2-}, EC, TH, B, TSS, Fe, and TKN. Hazardous contaminants arise mainly from the discharge of municipal wastewater, paper manufacturing wastewater, fabrication wastewater, and other commercial effluents. The FA can pinpoint the components that most cause fluctuations in quality of water.

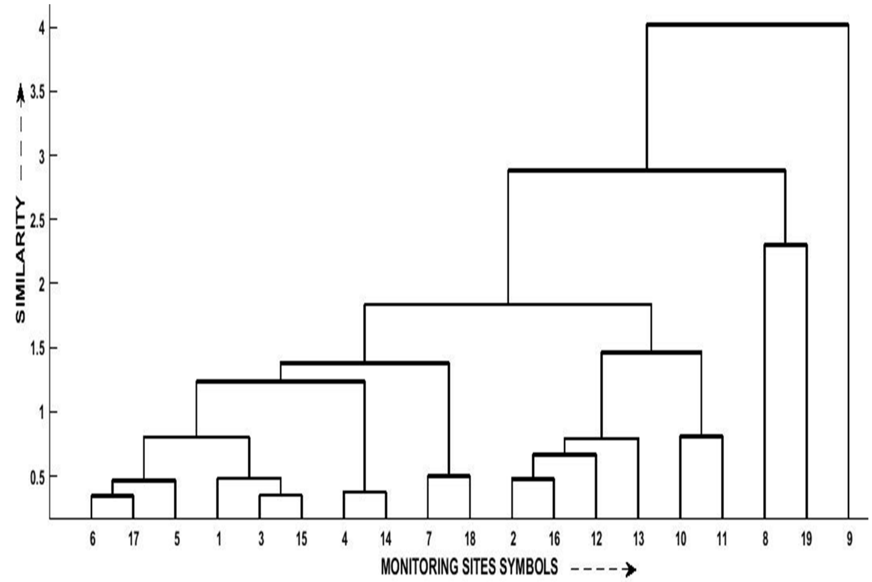

Figure 2. Dendrogram of CA of sampling stations

TABLE II. PEARSON CORRELATION MATRIX OF WATER QUALITY PARAMETERS

PARAMETERS	PH	DO	BOD	TC	TSS	TOTAL ALKALINITY	COD	NH3-N	FREE AMMONIA	TKN	EC	TDS	B	SAR	TH	CL_2	$S0_4^{2-}$	F	NO_3	FE
PH	1.00																			
DO	0.05	1.00																		
BOD	0.49	-0.59	1.00																	
TC	0.36	-0.77	0.77	1.00																
TSS	-0.23	-0.54	0.04	0.25	1.00															
TOTAL ALKALINITY	0.45	-0.15	0.34	0.14	-0.04	1.00														
COD	0.57	-0.39	0.66	0.62	0.17	0.50	1.00													
NH_3-N	0.29	0.13	0.15	0.12	0.07	0.06	0.72	1.00												

FREE NH_3	0.53	0.08	0.39	0.31	0.04	0.02	0.75	0.88	1.00											
TKN	-0.18	-0.54	0.13	0.15	0.73	0.23	0.16	-0.14	-0.10	1.00										
EC	0.05	-0.42	0.03	0.08	0.74	0.46	0.21	-0.06	-0.15	0.71	1.00									
TDS	0.05	-0.42	0.02	0.08	0.74	0.45	0.20	-0.07	-0.16	0.71	1.00	1.00								
B	0.10	-0.41	0.07	0.11	0.71	0.48	0.33	0.10	-0.01	0.67	0.98	0.97	1.00							
SAR	0.16	0.14	0.02	0.03	0.14	0.12	0.67	0.97	0.78	-0.08	0.04	0.03	0.20	1.00						
TH	0.06	-0.42	0.03	0.08	0.74	0.46	0.22	-0.05	-0.14	0.70	1.00	1.00	0.98	0.05	1.00					
CL^-	0.04	-0.42	0.02	0.08	0.75	0.45	0.20	-0.06	-0.16	0.71	1.00	1.00	0.98	0.04	1.00	1.00				
SO_4^{2-}	0.06	-0.42	0.03	0.08	0.74	0.45	0.21	-0.05	-0.15	0.71	1.00	1.00	0.98	0.04	1.00	1.00	1.00			
F	0.28	0.01	0.12	0.06	0.30	0.36	0.74	0.88	0.70	0.14	0.37	0.36	0.51	0.93	0.38	0.37	0.38	1.00		
NO_3^-	-0.10	-0.26	0.15	-0.08	0.08	0.52	0.25	0.09	-0.05	0.47	0.42	0.41	0.47	0.18	0.42	0.41	0.41	0.33	1.00	
FE	0.19	-0.20	-0.05	0.14	0.75	0.02	0.17	0.16	0.24	0.54	0.68	0.68	0.69	0.17	0.68	0.68	0.68	0.36	0.01	1.00

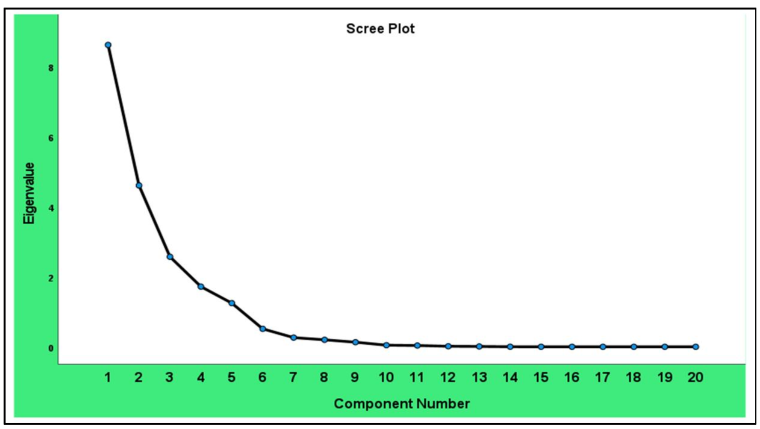

Figure 3. Scree plot of eigenvalue of all PCs

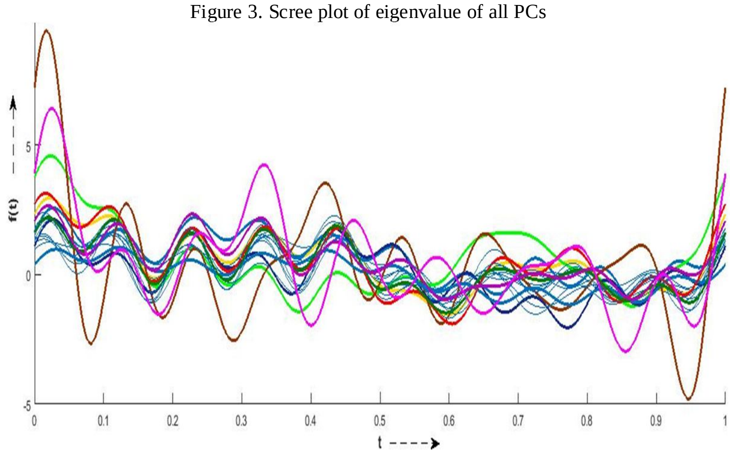

Figure 4. Andrew plot for all water quality parameters

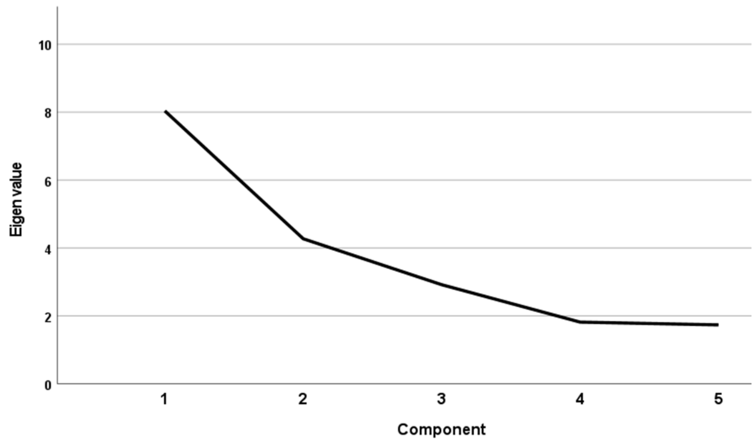

Figure 5. Scree plot of eigenvalue of all VFs

TABLE III. LOADINGS OF VARIABLES (20) ON SIGNIFICANT PRINCIPAL COMPONENTS (PCS)

Parameters	**PCs values**				
	PC-1	**PC-2**	**PC-3**	**PC-4**	**PC-5**
PH	0.13	0.57	-0.27	0.28	0.68
DO	-0.52	0.03	0.71	0.25	0.33
BOD	0.19	0.47	-0.79	0.06	-0.02
TC	0.24	0.38	-0.8	-0.3	0
TSS	0.78	-0.15	0.04	-0.55	-0.1
Total Alkalinity	0.49	0.2	-0.18	0.75	0.09
COD	0.43	0.85	-0.26	0.07	-0.07
NH_3-N	0.13	0.89	0.37	-0.13	-0.14
Free NH_3	0.07	0.93	0.11	-0.22	0.08
TKN	0.75	-0.26	-0.12	-0.14	-0.27
EC	0.96	-0.22	0.07	0.06	0.09
TDS	0.96	-0.23	0.07	0.06	0.09
B	0.97	-0.07	0.12	0.07	0.05
SAR	0.21	0.81	0.48	-0.08	-0.23
TH	0.96	-0.21	0.08	0.07	0.09
CL□	0.96	-0.23	0.08	0.05	0.09
S0□$^{2-}$	0.96	-0.22	0.08	0.06	0.1
F	0.52	0.72	0.42	0.05	-0.11
NO□	0.47	0.02	0.03	0.56	-0.56
FE	0.71	0	0.2	-0.43	0.36
Eigen value	8.63	4.61	2.57	1.72	1.25
% Variance	43.13	23.06	12.87	8.6	6.24
Cumulative %	43.13	66.19	79.05	87.66	93.9

VI. CONCLUSION

A multitude of multivariate statistical tools were used in this work to examine the geographical and temporal fluctuations in the Mahanadi River's surface water quality. These methods were found to be useful in explaining synoptic fluctuations in water quality metrics that contributed appreciably to alterations in stream water chemistry and quality. The study discovered distinct geographical and time variability in water quality measures, in addition to the fact that surface water pollution is spatially heterogeneous due to anthropogenic influences. According to the similarities of river basin features and pollutants, cluster analysis (CA) categorized the 19 data points into three categories. Cluster III is said to be extremely contaminated which includes Cuttack D/s (R-8),

TABLE IV. LOADINGS OF VARIABLES (20) ON SIGNIFICANT VFS (WITH VARIMAX ROTATION)

PARAMETERS	Rotated Component Matrix				
	VF-1	VF-2	VF-3	VF-4	VF-5
PH	0.012	0.258	0.241	-0.086	0.900
DO	-0.383	0.127	-0.843	-0.113	0.242
BOD	-0.053	0.122	0.861	0.157	0.309
TC	0.054	0.098	0.936	-0.139	0.128
TSS	0.817	0.132	0.237	-0.263	-0.371
Total Alkalinity	0.323	0.071	0.123	0.673	0.558
COD	0.136	0.713	0.560	0.213	0.303
NH_3-N	-0.062	0.992	0.018	-0.007	0.036
Free NH_3	-0.128	0.879	0.231	-0.199	0.224
TKN	0.735	-0.065	0.267	0.174	-0.311
EC	0.979	-0.005	0.038	0.167	0.061
TDS	0.980	-0.016	0.037	0.161	0.057
B	0.951	0.154	0.052	0.202	0.072
SAR	0.032	0.981	-0.084	0.087	-0.052
TH	0.977	0.006	0.039	0.173	0.068
CL_2	0.980	-0.006	0.036	0.161	0.052
SO_4^{2-}	0.980	0.000	0.039	0.160	0.063
F	0.340	0.905	-0.045	0.189	0.102
NO_3	0.307	0.129	0.050	0.842	-0.153
FE	0.790	0.220	0.014	-0.431	0.068
Eigen value	8.038	4.272	2.919	1.816	1.735
% Variance	40.191	21.358	14.593	9.082	8.675
Cumulative %	40.191	61.548	76.142	85.223	93.899

Choudwar D/s (R-19) and Paradeep (R-9). Only a small sample of data is reduced, ignoring the fact that FA/PCA highlighted five parameters as being needed to explain 93.9 % of the data variability. The main pollutants identified are Cl^-, TDS, SO_4^{2-}, EC, TH, B, TSS, Fe, and TKN. However, the five VFs procured from the PC, which explained VF-1 (40.191 percent of total variance), VF-2 (21.358 percent of total variance), VF-3 (14.593 percent of total variance), VF-4 (9.082 percent of total variance), and VF-5 (8.675 percent of total variance), indicated that river water quality parameters were primarily divided into natural (dissolved salts) and manmade (organic substances) attribution. One of the possible reasons may be the effluent discharges of industrial and domestic wastewater, generally kept at a relatively steady level throughout the entire year. It may be stated that the main sources of river water contamination were point source contamination flows from human impacts caused by the discharge of agricultural byproducts and household and industrial wastewater. This data also revealed advancements in farming methods. The findings also demonstrated the need of performing multivariate statistical analyses on huge datasets in order to obtain more information about river water. It is desirable that both state and local agencies pay more attention and consideration in order to improve and protect the vulnerable river quality. Additional studies will be required to assess precisely the unidentified sources of pollution and variation of further water quality parameters that were not analysed in this study.

AUTHOR CONTRIBUTIONS

Abhijeet Das conceptualized the idea and led the writing of this paper. The first draft was written by me and then sent to the reviewers for their comments and suggestions. Abhijeet Das compiled all the edits and produced the final paper.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGMENT

The research was funded by the Odisha University of Technology and Research (O.U.T.R), Bhubaneswar, Odisha, India. Special thanks to Dr. Deba Prakash Satapathy, Professor, Department of Civil Engineering, O.U.T.R, Bhubaneswar for providing the necessary guidance and instrument facilities required for this work. The observational data were obtained from India Meteorological Department (IMD) and Central Water

Commission (CWC), Bhubaneswar, Odisha. I also thank the reviewers and the journal officials for providing insightful comments.

REFERENCES

[1] APHA, 1992. Standard Methods for the Examination of Water and Wastewater, eighteenth ed. American Public Health Association, Washington, DC.

[2] Bhardwaj, V., Singh, D. S., & Singh, A. K. (2010). Water quality of the Chhoti Gandak River using principal component analysis, Ganga plain, India. Journal of Earth System Science, 119(1), 117–127. Doi: 10.1007/s12040-010-0007-8.

[3] Carpenter, S.R., Caraco, N.F., Correll, D.L., Howarth, R.W., Sharpley, A.N., Smith, V.H., 1998. Nonpoint pollution of surface waters with phosphorus and nitrogen. Ecol. Appl. 8, 559–568.

[4] Chattopadhyay T, Sharina M, Davoust E, De T, Chattopadhyay AK. Uncovering the formation of ultra-compact dwarf galaxies by multivariate statistical analysis. The Astro- physical Journal. 2012 2012-05-10; 750(2):91.

[5] Ding S, Jia W, Su C, Zhang L, Liu L. Research of neural network algorithm based on factor analysis and cluster analysis. Neural Computing and Applications. 2011; 20(2):297–302.

[6] Dominick D, Juahir H, Latif MT, Zain SM, Aris AZ. Spatial assessment of air quality patterns in Malaysia using multivariate analysis. ATMOS ENVIRON. 2012; 60:172–81.

[7] Ebrahimiyan M, Majidi MM, Mirlohi A. Genotypic variation and selection of traits related to forage yield in tall fescue under irrigated and drought stress environments. GRASSFORAGE SCI. 2013; 68(1):59– 71.

[8] Gashi S, Bajmaku Y, Drini P. Comparsion of some Chemical Parameters of Urban and Industrial Water Discharges onto the River Lumbardh of Prizren. IFAC-PapersOnLine. 2016; 49(29):129–32.

[9] Gonza´lez MJG, Vallejo-Pascual M. The Application of Principal Component Analysis (PCA) for the Study of the Spanish Tourist Demand. Quaestiones geographicae. 2018 2018-01-01; 37(4):43–52.

[10] Helena, B., Pardo, R., Vega, M., Barrado, E., Fernandez, J.M., Fernandez, L., 2000. Temporal evo- lution of ground water composition in an alluvial aquifer (Pisuerga River, Spain) by principal component analysis. Water Res. 34, 807–816. http://dx.doi.org/10.1016/S0043- 1354(99)00225-0.

[11] Jackson, J.E., 1991. A User's Guide to Principal Components. Wiley, New York

[12] Kazi TG, Arain MB, Jamali MK, Jalbani N, Afridi HI, Sarfraz RA, et al. Assessment of water quality of polluted lake using multivariate statistical techniques: A case study. ECOTOX ENVIRON SAFE. 2009; 72(2):301–9. https://doi.org/10.1016/j.ecoenv.2008.02.024 PMID: 18423587

[13] Khan, M. Y. A., Gani, K. M., & Chakrapani, G. J. (2016). Assessment of surface water quality and its spatial variation. A case study of Ramganga River, Ganga Basin, India. Arabian Journal of Geosciences, 9(1), 1–9.

[14] Koutsias N, Mallinis G, Karteris M. A forward/backward principal component analysis of Landsat-7 ETM + data to enhance the spectral signal of burnt surfaces. ISPRS J PHOTOGRAMM. 2009; 64(1):37–46.

[15] Li H, Liu J, Liu R, Xiong N, Wu K, Kim T. A Dimensionality Reduction-Based Multi-Step Clustering Method for Robust Vessel Trajectory Analysis. SENSORS-BASEL. 2017 2017-08-04; 17(8):1792. https://doi.org/10.3390/s17081792 PMID: 28777353.

[16] Liu C, Lin K, Kuo Y. Application of factor analysis in the assessment of groundwater quality in a black- foot disease area in Taiwan. SCI TOTAL ENVIRON. 2003; 313(1–3):77–89. https://doi.org/10.1016/ S0048-9697(02)00683-6 PMID: 12922062.

[17] Lo F, Hong J, Lin M, Hsu C. Extending the Technology Acceptance Model to Investigate Impact of Embodied Games on Learning of Xiao-zhuan(). Procedia—Social and Behavioral Sciences. 2012; 64:545–54.

[18] Mishra, S., & Kumar, A. (2020). Estimation of physicochemical characteristics and associated metal contamination risk in the Narmada River. Environmental Engineering Research, 26(1), 1–11.

[19] Mohamad Asri MN, Mat Desa WNS, Ismail D. Combined Principal Component Analysis (PCA) and Hierarchical Cluster Analysis (HCA): an efficient chemo metric approach in aged gel inks discrimination. AUST J FORENSIC SCI. 2020 2020-01-02; 52(1):38–59.

[20] Mudgal KD, Kumari M, Sharma DK (2009) Hydro chemical analysis of drinking water quality of Alwar District, Rajasthan. Nat Sci 7(2):30–39

[21] Nowak B, Nadolna A, Stanek P. Evaluation of the potential for the use of lakes in restoring water resources and flood protection, with the example of the Noteć Zachodnia River catchment (Gniezno Lakeland, Poland). Meteorology Hydrology and Water Management. 2018 ,2018-05-14.

[22] Praus, P. (2007). Urban water quality evaluation using mul- tivariate analysis. Acta Montanistica Slovaca, 12(2), 150–158.

[23] Richman, M.B., 1986. Rotation of principal components. J. Climatol. 6, 293–335.

[24] Sharma, M., Kansal, A., Jain, S., & Sharma, P. (2015). Application of multivariate statistical techniques in determining the spatial temporal water quality variation of Ganga and Yamuna Rivers present in Uttarakhand State, India. Water Quality, Exposure and Health, 7(4), 567–581.

[25] Simeonov, V., Stratis, J.A., Samara, C., Zachariadis, G., Voutsa, D., Anthemidis, A., Sofoniou, M., Kouimtzis, Th., 2003. Assessment of the surface water quality in Northern Greece. Water Res. 37, 4119–4124.

[26] Singh KP, Malik A, Sinha S. Water quality assessment and apportionment of pollution sources of Gomti river (India) using multivariate statistical techniques—a case study. ANAL CHIM ACTA. 2005; 538(1– 2):355–74.
[27] Thareja, S., Choudhury, S., & Trivedi, P. (2011). Assessment of water quality Ganga River in Kanpur by using principal component analysis. Advances in Applied Science Research, 2(5), 84–91.
[28] Twine TE, Kucharik CJ, Foley JA. Effects of El Niño–Southern Oscillation on the Climate, Water Bal- ance, and Streamflow of the Mississippi River Basin. J CLIMATE. 2005 2005-11-15; 18(22):4840–6
[29] Vega, M., Pardo, R., Barrado, E., & Deban, L. (1998). Assessment of seasonal and polluting effects on the quality of river water by exploratory data analysis. Water Research, 32, 3581–3592.
[30] Yerel, S., & Ankara, H. (2012). Application of multivariate statistical techniques in the assessment of water quality in Sakarya River, Turkey. Journal of the Geological Society of India, 79(1), 89–93. doi:10.1007/s12594-012-0019-x
[31] Yidana SM, Sakyi PA, Stamp G. Analysis of the Suitability of Surface Water for Irrigation Purposes: The Southwestern and Coastal River Systems in Ghana. Journal of Water Resource and Protection. 2011; 03(10):695–710.
[32] Zhang X, Wang Q, Liu Y, Wu J, Yu M. Application of multivariate statistical techniques in the assess- ment of water quality in the Southwest New Territories and Kowloon, Hong Kong. ENVIRON MONIT ASSESS. 2011; 173(1–4):17–27. https://doi.org/10.1007/s10661-010-1366-y PMID: 20191317.