INTERNATIONAL JOURNAL OF ENVIRONMENT

Volume-10, Issue-1, 2020/21 ISSN 2091-2854

Received: 15 May 2021 *Revised: 21 June 2021* *Accepted: 28 June 2021*

WATER QUALITY MODELLING OF RIVER MAHANADI USING PRINCIPAL COMPONENT ANALYSIS (PCA) AND MULTIPLE LINEAR REGRESSION (MLR)

Chandra Sekhar Matli[1*] and Nivedita[2]

[1,2]Department of Civil Engineering, National Institute of Technology, Warangal, India

*Corresponding author: mcsnitw@gmail.com

Abstract

Surface water quality is one of the critical environmental concerns of the globe and water quality management is top priority worldwide. In India, River water quality has considerably deteriorated over the years and there is an urgent need for improving the surface water quality. The present study aims at use of multivariate statistical approaches for interpretation of water quality data of Mahanadi River in India. Monthly water quality data pertaining to 16 parameters collected from 12 sampling locations on the river by Central Water Commission (CWC) and Central Pollution Control Board (CPCB) is used for the study. Cluster analysis (CA), is used to group the sampling locations on the river into homogeneous clusters with similar behaviour. Principal component analysis (PCA) is quite effective in identifying the critical parameters for describing the water quality of the river in dry and monsoon seasons. PCA and Factor Analysis (FA) was effective in explaining 69 and 66% of the total cumulative variance in the water quality if dry and wet seasons respectively. Industrial and domestic wastewaters, soil erosion and weathering, soil leaching organic pollution and natural pollution were identified as critical sources contribution to pollution of river water. However, the quantitative contributions were variable based on the season. Results of multiple linear regression (MLR) are effective in explaining the factor loadings and source contributions for most water quality parameters. The study results indicate suitability of multivariate statistical approaches to design and plan sampling and sampling programs for controlling water quality management programs in river basins.

Keywords: Water Pollution; Multivariate Analysis; Cluster Analysis; Factor Analysis; Receptor modelling

DOI: http://dx.doi.org/10.3126/ije.v10i1.38417

1. Introduction

Water is considered as one of the vital resources that supports life on the planet. Surface water sources, mostly rivers, lakes reservoirs are subjected to water pollution by various natural and anthropogenic sources and hence, are subjected to water quality degradation. Natural sources include erosion, weathering, dissolution of soil minerals, etc., while anthropogenic sources include wastewater from domestic and industrial activities, agricultural runoff, etc., (Singh et al., 2005). Surface water quality is of great importance as it supports life on the earth and public health in particular (Iscen et al., 2008). However, surface water quality is mostly influenced by hydrological and meteorological factors and human intervention in the hydrological cycle. In view of the above, surface water quality management is very challenging for the stakeholders (Bhaduri et al., 2001). In spite of advances in water quality monitoring in recent times, identification of representative and reliable samples is still perplexing (Salve et al., 2001). In this scenario, studies on spatial-temporal variations and source apportionment are worthwhile in management of surface waters (Shrestha et al., 2007).

Most often, water quality data is huge and can reveal lot of information individually and on inter relationships among variables. Hence, while analysing large data, the problem becomes complex to interpret and infer reliable conclusions findings (Kazi et al., 2009). In this context, for analysing complex data, research suggests use of statistical approaches such as Cluster Analysis (CA), Discriminant Analysis (DA), Principal Component analysis / Factor Analysis (PCA/FA) and Absolute Principal Component Score–Multiple Linear Regression (APCS-MLR) (Singh et al., 2005). Considering the above, the purpose of the study reported in this paper, aimed to adopt some of the above mathematical tools to understand the spatial variations and the major sources of water pollution in Mahanadi River basin. Large quantity of data collected during a 10 year (2001–2011) monitoring period at twelve different sites for sixteen water quality parameters, and for monsoon and dry seasons (about 22,000 observations) were subjected to CA, PCA/FA and APCS - MLR techniques to indicate variations in water quality at sampling sites and to identify natural and anthropogenic sources of pollution.

2. Materials and methods

2.1. Study area

Mahanadi River Basin majorly drains the states of Chhattisgarh and Odisha and smaller portions of Jharkhand, Maharashtra and Madhya Pradesh. It is one of important east flowing river in Peninsular India. River drains over an area of 141,600 km^2, which is about 4% of the total geographical area of the country.

The river basin lies between latitude 19º 21'N and 23º 35'N and longitudes 80º 30' E and 86º 50'E. The River Mahanadi originates at an elevation of about 442 m above MSL in Dhamtari district of Chhattisgarh and drains into Bay of Bengal. During its course of about 851 km, a number of tributaries (Seonath, Hasdeo, Mand, Ib, Bhadar, Jonk, Ong and Tel) join the River on both sides of the banks. Tropical monsoon climate with average annual temperature ranging between 15.8 ℃ to 28.7 ℃ prevail in the river basin. The average annual rainfall is about 1360 mm which mostly occurs during the months of June to September with occasional cyclonic storms with heavy rainfall (Dileep et al., 2013). The predominant soil types include red and yellow soils, mixed red and black soils (laterite soils). The major land uses in the basin are agriculture area, forest reserves, mining areas and urban centres. Figure 1 shows the location of the study area and the water quality monitoring locations on the River Mahanadi.

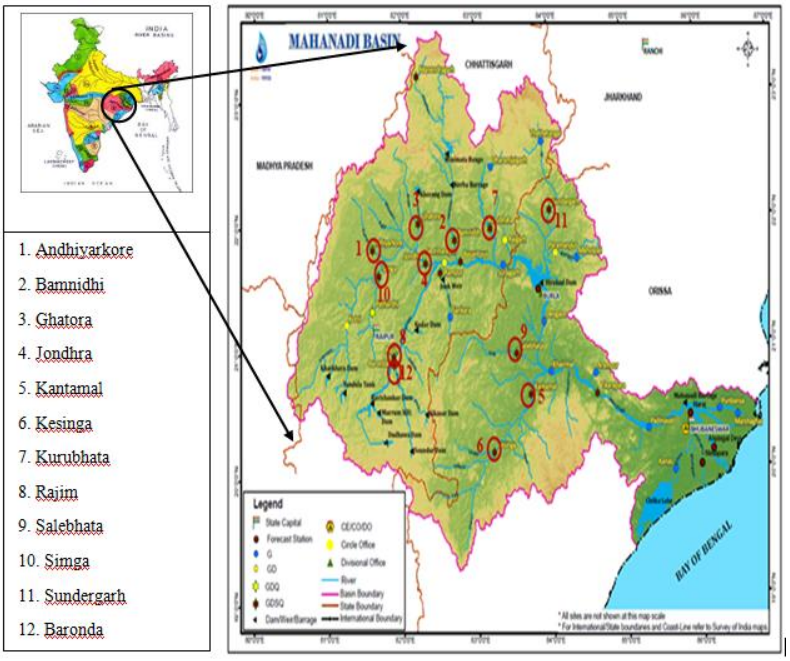

Figure 1. Map of study area and surface water quality monitoring stations in the Mahanadi river basin

2.2. Data Set Preparation

The Central Water Commission (CWC), Govt of India collect hydrological data and water quality data in river Mahanadi and its tributaries at 59 stations. The Mahanadi and Eastern Rivers Organisation under Central Water Commission (CWC), Bhubaneswar is engaged in collecting the discharge and water quality data of Mahanadi River. Standard Methods for examination of Water and Wastewater (APHA, 2017) were used for sampling and analysis of water quality at all locations on the river. The water quality data during the years from 2001 to 2011 from twelve monitoring stations provided by CWC was used for the present study. In order to overcome missing data problems, 16 water quality parameters were used though CWC collects data for 30 parameters. Considered parameters include pH, Electrical conductivity (EC), dissolved oxygen (DO),

Temperature, Biological Oxygen Demand (BOD), Chemical Oxygen Demand (COD), Nitrite + nitrate, Total Hardness, Fluoride (F), Boron (B), Calcium (Ca), Sodium (Na), Potassium (K), Chloride (Cl), Sulphate (SO_4), Bicarbonate (HCO_3). The complete data sets were divided into two sets to represent water quality of dry and monsoon seasons. Dry season includes the month of January, February, March, April, May, June and December. The Monsoon season includes the month of July, August, September, October and November.

2.3. Data Pre-Processing

The data were treated to replace few values which are either not detected or missing with half of its detection limit (Nasir et al., 2011) and the rest were filled with the geometric mean of the corresponding data set in order to facilitate statistical analysis. Normality test was performed using kurtosis and skewness tests since the multivariate statistical techniques requires normally distributed data (Chakrabarty and Sarma, 2011; Kumar et al., 2011; Zhou et al., 2007). Cantering, standardization and log-scaling methods are used for pre-treatment of data which was not normally distributed. Standardization options were adopted to increase the influence of variables with small variance and vice versa (Krishna et al., 2009; Kumar et al., 2014). Variables that were too low or high values were subjected to Log scaling (Felipe-Sotelo et al., 2007). This pre-treatment of the data sets was done using XLSTAT 2010. Data was standardised by z-scale transformation in order to overcome wrong classification due to different orders of magnitude of both numerical values and variance of the parameters considered for PCA ad CA (Simeonov et al., 2003). Microsoft Office Excel 2007 and SPSS 16 (Trial version) were used for the statistical analysis.

2.4. Cluster analysis

Cluster analysis is one of the classification techniques used for grouping or clustering data with similar nature and it is widely used multivariate statistical technique used to assess surface water quality. Analysis includes two steps: a proximity measure and a group-building algorithm. While the proximity measure (determined by a distance or similarity matrix and the resulting similarity coefficients) checks closeness or homogeneity of the objects, the group-building algorithm assigns groups to the objects based on assessments of the former so that objects in the same group are intimately homogeneous and significant differences exist between different groups (Kumar Manoj and Padhy, 2014). Dendrogram is demonstrates hierarchical agglomerative clustering (Simeonov et al., 2003). In this study, normalized data was used for Cluster Analysis. Ward's method uses variance approach to determine the distance between clusters in order to reduce the sum of squares (SS) of any two clusters that are formed in each step. Euclidean distance indicates the similarity between two samples and a distance denoted by variation of analytical values from the sample. Euclidean distance is expressed by Eq. (1).

$$d_{ij}^2 = \sum_{k=1}^{n} (z_{ik} - z_{jk})^2 \qquad \ldots \text{Eq. (1)}$$

where, d_{ij}^2 = square of Euclidean distance ; z_{ik} = k variable for the object i; z_{jk} = k variable for the object j; and n = number of variables.

2.5. Principal Component Analysis

PCA is a popular pattern recognition tool that focusses on reducing multidimensionality in data. PCA modifies the original variables to new, uncorrelated variables (axes), which are referred as Principal Components. However, the new variables are linear combinations of original variables. The new axes fall along the maximum variance direction. Thus, PCA results in indices that account for variance in the data significantly (Kumar Manoj and Padhy, 2014; Gajbhiye and Awasthi, 2015). Significant principal components are those with Eigen values more than 1. PCA describes the significant quality parameters due to spatial and seasonal variations (Singh et al., 2004). Principal Components indicate most critical parameters that describe data with not much loss of information. (Helena et al., 2000). Eq. (2) describes the expression for the principal component (PC).

$$y_{ij} = a_{i1}x_{1j} + a_{i2}x_{2j} + a_{i3}x_{3j} + \cdots \ldots \ldots \ldots \ldots + a_{im}x_{mj} \qquad \ldots\ldots \text{Eq. (2)}$$

where y and a are component score and loading respectively; x is variable measured; i is component number, j is sample number and m is number of variables.

The Kaiser-Meyer-Olkin (KMO) Test is generally used to find if data is suitable for Factor Analysis. The statistic determines the proportion of variance among variables that have equal variance. Data is better suited for Factor Analysis if the proportion is lower. KMO statistic values range between 0 and 1. As a rule of thumb, KMO values in the range of 0.8 - 1, indicate the sample is adequate; < 0.6 indicates inadequate sample (Hutcheson and Sofroniou, 1999). KMO values close to 0 indicate widespread correlations which is a problem for factor analysis.

Factor Analysis (FA) is a statistical technique used to minimize number of variables into fewer number of factors (Shrestha et al., 2008). The method uses maximum variance from all variables and groups them into a common score. The analysis indicates that there is true correlation between variables and factors. Principal Component Analysis is one of the popular methods used for factor analysis. New variables, referred as Vari-Factors (VF) are determined by rotating the axis defined by PCA. VFs are unobservable, hypothetical latent

variables while PC is a linear grouping of observable water quality parameters (Vega., 1998; Helena et al., 2000). Normalized water quality parameters are used for PCA to identify significant PCs which are subjected to varimax rotation to yield VFs (Singh et al., 2005). In the process, minimum factors that describe the same amount of information are obtained. Eq. (3) represents terms in FA.

$$Y_{ji} = a_{f1}f_{1i} + a_{f2}f_{2i} + a_{f3}f_{3i} + \ .....+ a_{fm}f_{mi} + e_{fi} \qquad \ldots \text{Eq. (3)}$$

where Y = measured water quality parameter; a = factor loading; f = factor score; e = residual term describing the errors or other source of variation, i = sample number and m = total number of factors.

2.6. Receptor Modelling (APCS-MLR)

Receptor model is a combination of Multiple Linear Regression model (MLR) and the Absolute Principal Component Scores (APCS) (Haji and Mellesse, 2016). The model assumes that the concentration of a contaminant under consideration is the sum of pollution components of various sources at the receptor location. Absolute scores (APCS) are used to determine pollutant source contributions. As z-transformed variables are used for PCA, normalized factor scores are obtained which are subsequently transformed to un-normalized APCS. Eq. (4) describes standardization of the concentrations of variables under consideration.

$$Z_{ij} = \frac{(x_{ij} - x_j)}{\sigma_j} \qquad \ldots \text{Eq. (4)}$$

where x_{ij} = concentration of water quality parameter (j) in sample i; x = mean concentration of variable j, and σ_j = standard deviation of variable j for all samples considered for the study.

Standardized variables are used for PCA which results in normalized factor scores (A_z) zero mean and unit standard deviation. To estimate absolute zero scores as given in Eq. (5), an artificial sample with zero concentration for all variables is introduced (Thurston and Spengler 1985).

$$(Z_0)_j = \frac{(0 - \overline{Xj})}{\sigma j} = -\frac{\overline{x}j}{\sigma j} \qquad \ldots \text{Eq. (5)}$$

Eq. (6) is used to compute the absolute zero factor scores (A_0) for each sample using the factor scores coefficients (S) obtained in PCA and Z_0.

$$(A_0)_f = \sum_{j=1}^{J} S_{fi}\,(Z_0)j \qquad \ldots \text{Eq. (6)}$$

Subtraction of absolute zero factor scores (A_0) of each sample from the appropriate normalized factor scores (A_z), the Absolute Principal Component Scores (APCS) are obtained (see Eq. (7)) (Thurston and Spengler, 1985).

where f =1, 2, ... , F.

The APCS are not concentrations of water quality parameters, however they can be converted to concentrations. The scores are proportional to source contributions. Mass concentrations are computed finding proportionality constants using MLR (see Eq. (8)) which takes factor scores as predictor variable. Contributions from each source for corresponding water quality parameter are now available for comparison with measured concentrations.

$$C_j = (r_0)_j + \sum_{k=1}^{F} rkj * \text{APCSj} \quad \ldots \text{Eq. (8)}$$

where, $(r_0)_j$ = constant term of multiple regression for parameter j; rkj = coefficient of multiple regression of the source k for parameter j; and $APCS_f$ is the absolute Principal Component Score; Combined term, r_{kj}*APCS = contribution of source k to C_j. Also, the average of the combined term indicates mean contribution of the sources (p).

3. Results and discussion

3.1. Spatial grouping

Cluster analysis detects similarity for grouping monitoring stations on the river network. Squared Euclidean distance obtained by using Ward's method on z-transformed data indicates the spatial similarity. Dendrogram given in Figure 2 indicates grouping of the twelve sampling sites on the river into three statistically significant clusters. Three groups were identified in cluster analysis that exhibited same characteristics and natural background source contributions. Cluster 1 (Andhiyarkore, Ghatora, Jondhra, Simga, Salebhata), corresponds to monitoring sites in the upper river basin, a relatively low pollution zone. Cluster 2 (Bamnidhi, Baronda) correspond to monitoring sites in the middle river basin. Cluster 3 (Kantamal, Kesinga, Kurubhata, Rajim)

corresponding to monitoring sites in middle and lower river basin and basically lie in an area comprising of red soils.

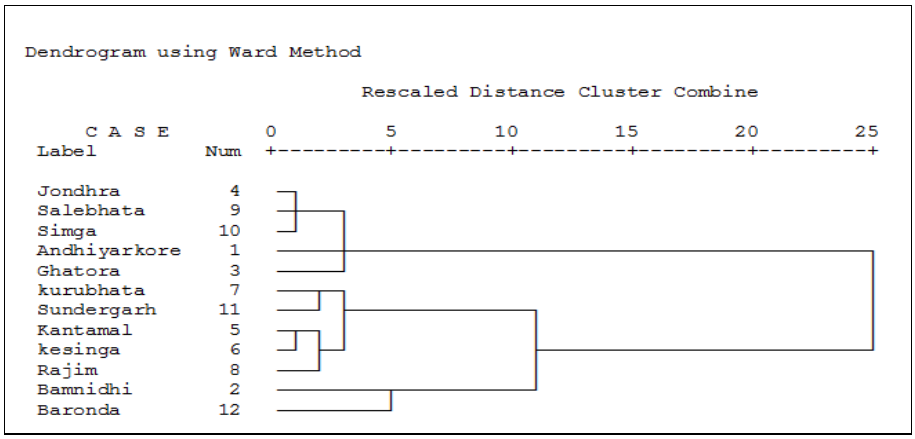

Figure 2. Dendrogram showing spatial clustering of sampling sites

Dendrogram is convenient for quick assessment of water quality as one sampling site in a cluster (not all monitoring sites) can serve as a good representative site for indicating water quality. Cluster analysis offers very useful in grouping sampling sites with similar behaviour and hence classification of surface waters of a river system. Also cluster analysis can be effectively used for optimizing future spatial sampling strategy. In the present study, data from three monitoring stations (one from each cluster) can serve the purpose of rapid water quality assessment and hence, reduced monitoring costs and risk of losing data for deriving significant outcome.

3.2. Principal Component Analysis (PCA) - Factor Analysis (FA)

Normalized water quality data (16 variables) of dry and monsoon seasons was used for Principal Component Analysis - Factor Analysis to find the factors influencing river water quality. Results of KMO and Bartllet's test of sphericity given in Table 1 were used to find if the data is suitable for PCA-FA. KMO values for dry and monsoon seasons were 0.896 and 0.831 respectively, while sphericity values from Bartlett's test were 8.67×10^3 and 4.25×10^3 ($p < 0.05$) respectively. Study results demonstrated significant relationships between water quality parameters and suitability of PCA analysis.

Table 1. KMO and Bartlett's test of MSA (Dry and Monsoon Seasons)

		Dry Season	Monsoon Season
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.896	0.831
Bartlett's Test of Sphericity	Approx. Chi-Square	8.67E+03	4.25E+03
	D_f	120	120
	Sig.	0	0

As per Kaiser Rule, principal components are identified (with Eigen value >1). For the dry season, four PCs (Principal Components), while for wet seasons five PCs were observed in the present study and are presented in Table 2 and 3 respectively. The PCs obtained sometimes could not be readily interpreted and, therefore, were rotated to generate a new rotated component matrix from the original component matrix. This helps in interpretation of the water quality data. VARIMAX approach is the most popular rotation technique. The rotation modifies the correlation between the components and the original variables, so that in the new extracted components, the important variables are included. These new groups of variables or components are termed varimax factors or vari-factors (VFs). The new factor loadings (earlier component loadings) generated illustrate the correlation between the variables and the factors. The objects displaying higher loading in each factor were interpreted as hallmarks of pollution source that it symbolizes. Factor loadings for 16 water quality parameters under study are given in Table 4. Factor loadings >0.75, [0.50–0.75] and [0.30–0.50] indicate strong, moderate and weak loadings respectively (Liu et al., 2003; Huang et al., 2010). Results indicate many of the factor loadings are above 0.75 indicating strong factor loadings for dry seasons, while many of them were under moderate factor loadings for monsoon season. This is perhaps due to the influence of dilution on water quality parameters during the monsoon season.

Table 2 .Total Variance Explained (Dry Period)

Component	Initial Eigen values			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	7.027	43.921	43.921	6.509	40.678	40.678
2	1.569	9.808	53.73	1.692	10.576	51.254
3	1.365	8.532	62.261	1.669	10.428	61.683
4	1.136	7.101	69.362	1.229	7.679	69.362
5	0.968	6.047	75.41			
6	0.783	4.892	80.301			
7	0.563	3.521	83.822			
8	0.535	3.346	87.168			
9	0.445	2.781	89.949			
10	0.43	2.69	92.638			
11	0.365	2.283	94.922			
12	0.282	1.764	96.686			
13	0.224	1.402	98.088			
14	0.161	1.003	99.091			
15	0.12	0.748	99.839			
16	0.026	0.161	100			

Table 3. Total Variance Explained (Monsoon Period)

Component	Initial Eigen values			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5.196	32.477	32.477	3.668	22.926	22.926
2	1.571	9.817	42.294	2.787	17.422	40.348
3	1.329	8.305	50.6	1.47	9.19	49.537
4	1.325	8.28	58.88	1.46	9.123	58.66
5	1.139	7.119	65.998	1.174	7.338	65.998
6	0.891	5.569	71.567			
7	0.82	5.125	76.692			
8	0.68	4.247	80.939			
9	0.593	3.707	84.646			
10	0.518	3.236	87.882			
11	0.485	3.03	90.913			
12	0.4	2.502	93.415			
13	0.358	2.237	95.652			
14	0.343	2.145	97.797			
15	0.294	1.835	99.632			
16	0.059	0.368	100			

3.3. Identification of Potential Pollution Sources

For the dry period, PCA is successful in explaining 69% of the total cumulative variance in the water quality data used for the study. Similar findings were reported by Mustapha et al., 2013 in a study on Jakarta River basin. The factor loadings and predicted sources of pollution are presented in Table 4 and 5. PC 1 exhibits 40.6 % of the total variance, indicating significant positive loadings on EC, COD, Total Hardness, Ca, Na, K, Cl, SO_4, HCO_3, weak loadings on BOD and Fluoride and negative loading on DO. COD is an indicator of organic pollution from sources such as partially / untreated domestic and industrial wastewater from urban areas. (Singh et al., 2005; Chen et al., 2015). EC indicates the presence of dissolved solids. Hardness is contributed by multivalent cations which result from dissolution of sedimentary rocks, seepage and run off from soils. Ca is constituent of limestone and chalk while SO_4 is found in soil and rock minerals. PC1 also indicates decomposition process and hence, negative DO loading and positive BOD. Thus, PC1is to be interpreted as one type of combined namely, soil leaching and industrial pollution (Fahmi et al., 2011).

Table 4. Varimax Rotated Factor Loadings

Parameters	Component (Dry season)				Component (Monsoon season)				
	PC1	PC2	PC3	PC4	PC1	PC2	PC3	PC4	PC5
pH	0.123	-0.067	0.071	**0.671**	**0.441**	-0.064	0.167	-0.269	0.083
EC	**0.896**	0.094	0.238	0.03	**0.777**	0.308	-0.096	0.097	0.04
DO	-0.17	-0.787	0.101	0.017	0.047	-0.293	-0.066	**-0.554**	-0.466
Temp	-0.016	0.04	0.007	**0.824**	0.094	-0.082	-0.034	-0.082	**0.882**
NO_2+NO_3	-0.021	**0.788**	0.073	0.027	0.353	-0.211	-0.099	**0.628**	-0.295
BOD	0.35	**0.535**	0.442	-0.188	0.019	0.09	0.04	**0.699**	0.001
COD	**0.81**	-0.019	0.251	-0.052	**0.575**	0.377	0.197	0.237	-0.131
TH	**0.907**	0.189	0.156	0.011	**0.842**	0.342	-0.038	0.068	0.012
F	0.374	0.188	**0.576**	0.02	0.486	-0.111	**0.586**	0.17	0.136
B	0.019	-0.094	**0.862**	0.122	-0.079	0.129	**0.88**	-0.051	-0.064
Ca	**0.877**	0.17	0.264	0.01	**0.868**	0.264	0.086	0.097	0.021
Na	**0.863**	0.097	-0.023	0.103	0.306	**0.783**	-0.019	-0.03	-0.062
K	**0.684**	-0.02	0.093	0.165	0.198	**0.633**	0.18	0.267	-0.137
Cl	**0.802**	0.108	0.005	0	0.245	**0.758**	0.134	-0.026	0.13
SO_4	**0.746**	0.037	-0.306	0.065	0.214	0.66	-0.413	0.054	0.055
HCO_3	**0.859**	0.137	0.247	-0.03	**0.679**	0.403	-0.174	0.063	0.023

Table 5. Predicted Sources of Pollution

DRY SEASON		
Principal Component	Typical Loadings	Predicted Sources
PC 1	EC, COD, Total Hardness, Ca, Na, K, Cl, SO4, HCO3	Industrial and domestic wastewaters, Soil leaching
PC 2	BOD, NO2+NO3, DO	Organic pollution
PC 3	F-, B	Soil leaching + weathering
PC 4	pH, Temp	Natural pollution
MONSOON SEASON		
PC 1	EC, Total Hardness, Ca, HCO3, COD, pH	Industrial and domestic wastewaters, Soil leaching
PC 2	Na,K,Cl	Soil erosion, Agricultural runoff and weathering
PC 3	F-,B	Soil leaching
PC 4	BOD, NO2+NO3	Organic pollution
PC 5	Temp	Natural pollution

PC 2 accounts for about 10.5% of the total variance, with strong positive loadings for BOD and NO_2+NO_3 and also strong negative loading on DO. This is obvious because organic matter contains nitrogenous matter and organic matter depletes DO in waters, hence negative loading. PC 2 represents organic pollution which can be attributed to wastewater inflows into the river. PC 3 explains about 10% of total variance with strong positive loadings for Boron and moderate loadings on F-. Boron probably represents weathering of rocks, possible marine deposits, sea water intrusion etc. (Simenov et al., 2003). As in the present study area there is no influence on marine systems, it is mostly due to dissolution of rock minerals. Fluoride is usually contributed

by cement plants, chemical and metallurgical industries (Huang et al., 2010). However, in the study region, higher F- levels are reported in ground waters though no significant concentrations are found in river water. Low Fluoride concentrations are contributed by local soils through the run-off (Huang et al., 2010). So, this can be interpreted as soil leaching. Low concentrations of Fluoride in river water indicate not much contributions from cement industries which are around. PC 4 explains for 7.6% of the total variance, with significant loadings for pH and Temperature. As the average pH is around 7.6 and average temperature during the dry period is around 24ºC, this PC can be attributed to natural pollution. However, PC 4 is not very significant in the present study.

For the monsoon period, PCA is effective in explaining 66% of total cumulative variance in the water quality parameters for the study. Study results indicate that PC 1 accounts for 23% of total variance with significant positive loadings for EC, Total Hardness, Ca, HCO3 and Temperature and modest loadings on COD, weak loadings on pH and NO2+NO3. This factor can perhaps indicate contributions from dissolution of soil minerals in the surface runoff and diluted wastewater inflows from domestic and industries. PC 2 explains about 17% of total variance with strong positive loadings for Na, K and Cl. These loadings are contributed by minerals granite and other rocks, while potassium is perhaps contributed from agricultural runoff. However, the concentration of K at all monitoring sites is less indicating dominant contributions from rock minerals by soil weathering and erosion. PC 3 describes 9% of total variance with strong positive loadings for Fluoride and Boron indicating contributions from soil leaching. PC 4 accounts for 9% of the total variance with strong positive loadings for BOD and NO2+NO3, negative loadings on DO indicating organic pollution. PC 5 describes 7% of total variance with significant positive loadings for temperature indicating thermal pollution due to tropical climate.

3.4. Source Apportionment using APCS-MLR Modelling

Absolute Principal Component Scores - Multiple Liner Regression (APCS-MLR) modelling is found to be effective for identifying the pollutant inputs from each source to water quality parameters and subsequently used for source apportionment (Chen et al., 2013). Although factor loadings and scores are useful indicators for relative comparison, these cannot be applied to quantitative estimations of contributions. In MLR modelling, it is assumed that concentrations of each water quality parameter is equal to the sum of contributions by several sources at the receptor location. Results of any study are reliable if $n \geq m + 50$ (where n is the number of samples and m is the number of pollutants analysed) (Thurston and Spengler (1985). In the present study, data considered satisfies the above condition and hence results are reliable. The principal components with Eigen value >1 are considered (Kaiser's criteria) to indicate water quality parameters with significant impact. Results of APCS-MLR Modelling for dry and monsoon seasons are presented in Table 6.

Table 6. Source Apportionment Results of APCS-MLR Modelling

Parameters	Dry Season						Monsoon Season						
	Un explained	S1	S2	S3	S4	R^2	Un explained	S1	S2	S3	S4	S5	R^2
pH	1.00	1.07	1.32	1.35	1.07	0.47	2.16	2.99	0.15	0.69	1.53	0.19	0.37
EC	51.50	240.21	6.67	33.69	1.67	0.86	11.62	132.71	48.60	8.45	7.40	2.54	0.72
DO	1.81	0.44	4.62	0.36	0.04	0.65	0.17	0.08	0.75	0.19	3.25	2.19	0.62
Temp	6.02	0.12	0.47	0.02	17.06	0.68	3.29	0.89	0.76	0.22	0.78	21.05	0.80
NO2+NO3	0.11	0.00	0.34	0.02	0.01	0.63	-	0.11	0.04	0.01	0.20	0.06	0.60
BOD	0.03	0.48	1.28	0.94	0.11	0.64	0.46	0.01	0.04	0.01	0.73	-	0.49
COD	2.62	26.27	0.23	4.97	1.42	0.72	2.73	10.55	4.83	0.29	3.14	0.90	0.58
TH	2.88	96.45	10.58	7.06	0.66	0.88	3.79	57.16	16.75	0.64	0.97	0.09	0.83
F	0.01	0.08	0.03	0.13	0.00	0.51	0.01	0.12	0.01	0.16	0.03	0.01	0.60
B	0.00	0.00	0.00	0.03	0.00	0.76	0.00	0.00	0.00	0.01	0.00	0.00	0.80
Ca	1.63	20.47	0.93	3.40	1.10	0.86	0.32	13.94	2.81	0.55	0.33	0.41	0.80
Na	1.71	15.36	1.21	0.14	1.62	0.76	1.42	2.16	7.44	0.03	0.06	0.01	0.70
K	0.84	2.22	0.02	0.21	0.42	0.50	0.38	0.26	1.49	0.21	0.44	0.15	0.56
Cl	3.34	18.82	2.25	0.30	-	0.65	1.78	2.10	9.45	0.69	0.08	0.90	0.66
SO4	1.30	12.11	0.21	3.73	0.41	0.65	0.73	1.45	6.87	3.94	0.11	0.11	0.65
HCO3	2.15	102.40	15.02	22.31	1.14	0.80	8.96	54.51	26.96	5.84	0.88	0.19	0.65

Higher R^2 values (most of them greater than 0.5) in most cases, indicate good consistency in explaining source contributions and hence source apportionment (Simeonov et al., 2003). Eq. (8) is used to compute contributions from unidentified sources. Predominant sources influencing river water quality in the study area are identified as Industrial and domestic wastewaters (S_1); Organic pollution (S_2); Soil leaching and weathering (S_3) and natural pollution (S_4).during dry season while in wet seasons the sources are Industrial and domestic wastewaters (S_1); soil erosion and weathering (S_2), soil leaching (S_3), organic pollution (S_4) and natural pollution (S_5).

4. Conclusions

Multivariate statistical techniques like CA, PCA/FA and APCS-MLR are used to identify the potential polluting sources and their influence on water quality parameters of Mahanadi River Basin. Results of CA demonstrated that the monitoring sites can be divided into 3 clusters with similar trends in water quality and background contributions. Principal components and factor loadings are successfully used for describing the water quality of the river. The study suggests that PCA can be effectively used for identification of critical water quality parameters while monitoring river water. Potential polluting sources in the river system are identified as industrial and domestic wastewaters, soil erosion and weathering, soil leaching, organic pollution and natural pollution with seasonal variations in quantitative contributions. Results demonstrate the applicability of the PCA, FA and APCS-MLR modelling for water quality studies in river basins. The correlation coefficient higher than 0.65 for most cases indicates the performance of the methods used for describing the water quality variations. Results can be useful in river water quality management and planning. Structural Equation Modelling is not attempted in this study due to want of time and data requirements.

Sustainable management of rivers requires immediate attention in many of the developing countries in order to prevent water quality degradation and to protect the aesthetic value of rivers for future generations.

Conflicts of interest

The authors declare no conflicts of interest.

Author contribution statements

Conceptualization; Formal analysis; Investigation; Resources: M Chandra Sekhar and Nivedita
Methodology; Software; Validation; Data Curation; Visualization ; Writing - Original Draft: Nivedita
Writing - Review & Editing: M Chandra Sekhar

References

APHA, 2017. Standard methods for the examination of water and wastewater, 23rd Edition, American Public Health Association, American Water Works Association, Water Environment Federation, ISBN: 9780875532875

Bhaduri, B., Minner, M., Tatalovich, S., Harbor, J, 2001. Long-term hydrologic impact of urbanization: A table of two models. J. Water Res. Plan. Man., 127, 13–19.

Chakrabarty, S. and Sarma, H.P., 2011. A statistical approach to multivariate analysis of drinking water quality in Kamrup district, Assam, India. Archives of Applied Science Research. 3 (5), 258-264.

Chen, H., Teng, Y., Yue, W., Song, L., 2013. Characterization and source apportionment of water pollution in Jinjiang River, China. Environ. Monit. Assess. 185, 9639–9650.

Chen, P., Li, L., Zhang, H., 2015. Spatio-temporal variations and source apportionment of water pollution in Danjiangkou Reservoir Basin, Central China. Water (Switzerland) 7, 2591–2611

Dileep K. Panda, A. Kumar, S. Ghosh, R.K. Mohanty, 2013. Streamflow trends in the Mahanadi River basin (India): Linkages to tropical climate variability, Journal of Hydrology, Volume 495, Pages 135-149, ISSN 0022-1694, https://doi.org/10.1016/j.jhydrol.2013.04.054.

Fahmi, Mohd & Mohd Nasir, Mohd Fahmi & Samsudin, Mohd & Mohamad, Isahak & Roshide, Mohammad & Awaluddin, Amir & Mansor, Muhd & Juahir, Hafizan & Ramli, Norlafifah, 2011. River Water Quality Modeling Using Combined Principle Component Analysis (PCA) and Multiple Linear Regressions (MLR): A Case Study at Klang River, Malaysia. World Applied Sciences Journal. 14: 73-82, ISSN 1818-4952; © IDOSI Publications, 2011

Felipe-Sotelo, J.M.A., Carlosena, A, Tauler, R., (2007). Temporal characterisation of river waters in urban and semi-urban areas using physico-chemical parameters and chemometric methods. Analytica Chemica Acta 2007, 583:128–137.

Gajbhiye Sharma S.K. and Awasthi M.K., 2015. Application of Principal Components Analysis for Interpretation and Grouping of Water Quality Parameters", International Journal of Hybrid Information Technology Vol.8, No.4 , pp.89-96.

Haji Gholizadeh, M., Melesse,M.A., 2016. Water quality assessment and apportionment of pollution sources using APCS-MLR and PMF receptor modeling techniques in three major rivers of south Florida, Sci Total Environ (2016)

Helena, B., Pardo, R., Vega, M., Barrado, E., Fernandez, J.M., Fernandez, L., 2000. Temporal evolution of groundwater composition in an alluvial aquifer (Pisuerga River, Spain) by principle component analysis. Water Res. 34, 807–816.

Huang, F., Wang, X., Lou, L., Zhou, Z., Wu, J., 2010. Spatial variation and source apportionment of water pollution in Qiantang River (China) using statistical techniques. Water Res. 44,1562–1572.

Hutcheson, G., Sofroniou, N., 1999. The Multivariate Social Scientist, Introductory Statistics Using Generalized Linear Models.

Iscen, C.F., Emiroglu, Ö., Ilhan, S., Arslan, N., Yilmaz, V., 2008. Ahiska, S. Application of multivariate statistical techniques in the assessment of surface water quality in Uluabat Lake, Turkey. Environ. Monit. Assess, 144, 269–276.

Kazi, T.G., Arain, M.B., Jamali, M.K., Jalbani, N., Afridi, H.I., Sarfraz, R.A., Baig, J.A., Shah, A.Q. 2009. Assessment of water quality of polluted lake using multivariate statistical techniques: A case study. Ecotox. Environ. Safety, 72, 301–309.

Krishna, A.K., Satyanarayanan, M., Govil, P. K., 2009. Assessment of heavy metal pollution in water using multivariate statistical techniques in an industrial area: a case study from Patancheru, Medak District. Andhra Pradesh, India. J Hazard Mater, 167:366–373

Kumar Manoj and Pratap Kumar Padhy, 2014. Multivariate statistical techniques and water quality assessment: Discourse and review on some analytical models. International Journal of Environmental Sciences Volume 5, No 3.

Kumar, A.S., Reddy, A.M., Srinivas, L., Reddy, P.M., 2014. Assessment of surface water quality in Hyderabad Lakes by using multivariate statistical techniques, Hyderabad-India. Environ.Pollut. 4, 14.

Kumar, P., Saxena, K.K., Singh, N.O., Nayak, A.K., Tyagi, B.C., Ali, S., Panney, N.N. and Mahanta, P. C., 2011. Application of multivariate statistical techniques for water quality characterisation of Sarada Sagar Reservoir, Ind. J. of Fisheries, 58 (4), 21-26.

Liu, C.W., Lin, K.H., Kuo, Y.M., 2003. Application of factor analysis in the assessment of groundwater quality in a blackfoot disease area in Taiwan. Sci. Total Environ., 313, 77–89.

Mustapha, A., Aris, A.Z., Juahir, H., Ramli, M.F., Kura, N.U., 2003. River water quality assessment using environmentric techniques: Case study of Jakara River Basin. Environ. Sci. Pollut. Res. 2013, 20, 5630–5644

Nasir, M.F. M., Samsudin, M.S., Mohamad, I., Awaluddin, M.R.A., Mansor, M.A., Juahir, H., Ramli, N., 2011. River water quality modeling using combined principle component analysis (PCA) and multiple linear regressions (MLR): a case study at Klang River, Malaysia. World Appl. Sci. J. 14, 73–82.

Salve, P.R., Gobre, T., Lohkare, H., Krupadam, R.J., Bansiwal, A., Ramteke, D.S., Wate, S.R., 2011. Source identification and variation in the chemical composition of rainwater at coastal and industrial areas of India. J. Atmos. Chem. 2011, 68, 183–198Sarita

Shrestha, S., Kazama, F. and Nakamura, T., 2008. Use of Principle component analysis, factor analysis and discriminant analysis to evaluate spatial and temporal variations in water quality of the Mekong River. Journal of Hydro informatics. 10 (1), 43-56.

Shrestha, S., Kazama, F., 2007. Assessment of surface water quality using multivariate statistical techniques: a case study of the Fuji river basin, Japan. Environ. Model. Softw.22, 464–475.

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

Singh, K.P., Malik, A., and Sinha, S., 2005. Water quality assessment and apportionment of pollution sources of Gomti river (India) using multivariate statistical techniques-A case study, Analytica Chimica Acta, 538(1-2), pp. 355-374.

Singh, K.P., Malik, A., Mohan, D., Sinha, S., 2004. Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of Gomti River (India) -A case study. Water Res. 38, 3980–3992.

Thurston, G.D., Spengler, J.D., 1985. A quantitative assessment of source contributions to inhalable particulate matter pollution in metropolitan Boston. Atmos. Environ. 19, 9–25.

Vega, M., 1998. Assessment of seasonal and polluting effects on the quality of river water by exploratory data analysis. Water Res. 32, 3581–3592.

Zhou, F., Liu, Y., Guo, H., 2007. Application of multivariate statistical methods to water quality assessment of the water courses in north western new territories. Hong Kong. Environ Monit Assess. 2007 Sept., 132 (1-3):1-13.