© IWA Publishing 2020
973

Water Practice & Technology Vol 15 No 4
doi: 10.2166/wpt.2020.078

Application of Environmetrics tools for geochemistry, water quality assessment and apportionment of pollution sources in Deepor Beel, Assam, India

Siddhant Dash*, Smitom Swapna Borah and Ajay S. Kalamdhad

Department of Civil Engineering, Indian Institute of Technology Guwahati, Guwahati 781039, Assam, India

*Corresponding author. E-mail: dash.siddhant93@gmail.com

Abstract

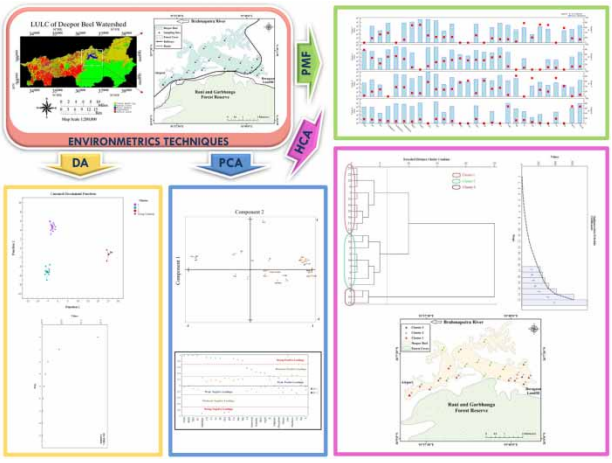
The present study uses four Environmetrics tools: hierarchical cluster analysis (HCA), discriminant analysis (DA), principal component analysis (PCA), and positive matrix factorization (PMF) for the assessment of water quality and geochemistry of Deepor Beel, Assam, India. The hierarchical clustering classified the 23 sampling locations into three clusters, classifying them as sites of high, low, and moderate contamination respectively. The DA of the water quality dataset resulted in 9 parameters (EC, TDS, TSS, PO_4^{3-}, Na^+, Mg, Cd, Pb and OrgN), primarily responsible for the discrimination of the clusters. PCA was then employed on the normalized dataset for the identification of potential pollution sources. PCA yielded two significant principal components, describing anthropogenic and natural factors defining the water contamination. Finally, PMF was employed on the dataset matrix, with four pre-defined factors. Leaching from Boragaon landfill site, surface water runoff, discharge of effluents from the industries in the wetland and discharge from Basistha River were found to be the major contributors. The results of this study provide a comprehensive correlation between water quality parameters and their sources, which would thereby assist in better planning and management of wetland restoration.

Key words: environmetrics, geochemistry, multivariate statistics, PMF receptor model, water quality assessment

Highlights

- First of its kind extensive monitoring and assessment of water quality for Deepor Beel, India.
- HCA, DA and PCA were employed for categorization of sampling sites, cluster validation and principal source identification.
- PMF provided quantitative contribution of each parameter for water pollution in Deepor Beel.
- PMF also identified 4 significant factors contributing towards wetland contamination.

Graphical Abstract



Water Practice & Technology Vol 15 No 4
doi: 10.2166/wpt.2020.078

NOMENCLATURE AND ABBREVIATIONS

BOD_5	5-day Biochemical Oxygen Demand
NH_3^-	Ammonia
Cd	Cadmium
Ca^{2+}	Calcium
COD	Chemical Oxygen Demand
Cl^-	Chloride
Cr	Chromium
Cu	Copper
DO	Dissolved oxygen
EC	Electrical conductivity
F^-	Fluoride
Fe	Iron
Pb	Lead
Mg	Magnesium
Mn	Manganese
NO_3^-	Nitrate
OrgN	Organic nitrogen
PO_4^{3-}	Phosphate
K^+	Potassium
Na^+	Sodium
SO_4^{2-}	Sulphate
TA	Total alkalinity
TDS	Total dissolved solids
TH	Total hardness
TKN	Total Kjeldahl nitrogen
TSS	Total suspended solids
WQ	Water quality

INTRODUCTION

Rapid and significant changes in the land-use and land-cover (LULC) patterns across the globe have been observed to have profound effects on the natural water bodies. Numerous studies citing the alteration of the natural state of our water bodies, both quantitative as well as qualitative, have raised serious concerns among researchers regarding global water sustainability (Shi *et al.* 2017; McCartney *et al.* 2018). The quantitative revival of the water bodies can be achieved through change in our land-use patterns and judicious utilization of our water resources as well. The qualitative restoration process, however, is a very tedious task and requires a more comprehensive and detailed study on the pollution sources and contaminating factors responsible for the continuous degradation of the water quality. This is owed to the dynamics of the natural systems. Thus, a continuous monitoring and assessment process, both spatial and temporal, for quality check purposes becomes inevitable. Continuous monitoring provides a reliable yet sophisticated water quality (WQ) dataset, which becomes quite ineffective when it comes to interpretation, owing to its complexity (Vega *et al.* 1998; Simeonov *et al.* 2003; Iscen *et al.* 2008; Chow *et al.* 2016; Hajigholizadeh & Melesse 2017; Singh *et al.* 2019). However, in recent years an Environmetrics approach (that is, application of various statistical techniques), has eased how the water quality datasets are understood, including the classification of spatially distributed monitoring sites and the primary factors or contaminants responsible for the deterioration of the water quality (Jha *et al.* 2014; Machiwal & Jha 2015; Bodrud-Doza *et al.* 2016; Chow *et al.* 2016; Kumar *et al.* 2017).

Water Practice & Technology Vol 15 No 4
doi: 10.2166/wpt.2020.078

Hierarchical clustering of sampling sites through cluster analysis (CA) and the identification of probable pollution sources through principal component analysis (PCA) have been widely used and accepted. The use of discriminant analysis (DA) as a supervised pattern recognition tool for the recognition of the most significant water quality variables accountable for spatial and temporal variability has also been used more recently compared to the other two methods (Hajigholizadeh & Melesse 2017). However, these statistical tools cannot solely quantify the contribution of potential pollution sources. For this purpose, various receptor models, such as the positive matrix factorization (PMF) model, can be made use of. The PMF models were initially applied to the dataset pertaining to the atmospheric pollution to determine how much is the contribution of various pollution sources. Only in recent times, have they been applied to the water quality (WQ) datasets along with PCA and CA for quantifying the contributions of pollution sources (Zhao *et al.* 2013; Mustaffa *et al.* 2014; Chen *et al.* 2015; Gholizadeh *et al.* 2016).

Deepor Beel is one of the largest freshwater wetlands of the Brahmaputra Valley, located in the heart of Guwahati, India. It can be considered as the lungs of the bustling city, assimilating a large amount of municipal and industrial wastewater along with agricultural runoff daily (MoEF 2008). The dilution and dispersion of pollutants in the wetland reduces the pollutants' concentration in the waters, whose ultimate fate lies in the Brahmaputra river. However, in recent times, Deepor Beel has been in a sorrowful state, primarily because of various anthropogenic reasons. Once a 40.14 km^2 spread wetland is now just a 4.1 km^2 plot of wetland with human dwellings and industries surrounding it from almost all directions (MoEF 2008). Thus, there is a need for an intensive scientific research with continuous monitoring of the wetland water quality to ensure that the conditions do not deteriorate any further and at the same time help the authorities in the decision-making process to implement better restoration programs.

Given the above considerations, the vast WQ dataset matrix generated as a result of continuous monitoring from 23 sampling locations for 28 parameters from October 2017 to February 2019 (about 24,000 observations) were subjected to (i) hierarchical clustering through CA for identification of the sampling locations having similarities and classifying them accordingly; (ii) confirmation of the grouping of sites from CA through discriminating functions obtained from DA; (iii) identification of presumable contributing pollution sources based on the results obtained through PCA and (iv) quantification of the contribution from various sources through source apportionment from a PMF model. The results obtained from this study are expected to be of substantial help for efficient water management in Deepor Beel.

DATA ACQUISITION AND METHODOLOGIES ADOPTED

Study area and its WQ sampling and analysis

Deepor Beel (Figure 1) is situated in the heart of the Indian city of Guwahati, Assam. It holds an essential place in the international space owing to its inclusion as a Ramsar site under the Ramsar Convention of Wetlands (Ref. No. 1207) (Bhattacharyya & Kapil 2010). Two primary outlets act as sources of water for Deepor Beel, one being the Basistha River, joining the wetland from the south-eastern part, and the other being Khanajan, connecting Deepor Beel to the Brahmaputra River in the north. The entire watershed of the wetland has been delineated using ArcGIS (Version 10.2) and the land-use and land-cover (LULC) map prepared using ENVI (Version 4.7) (Figure 1(a)). It is clear from the LULC map that a significant portion of the watershed is either vegetation (Western part) or forest cover (the Rani and Garbhanga Forest Reserve in the South); however, the wetland is also surrounded by a dense urban built-up area, which has been a primary reason for its continuous deterioration, both in expanse as well as quality of water. The presence of a landfill site (Boragaon

Water Practice & Technology Vol 15 No 4
doi: 10.2166/wpt.2020.078

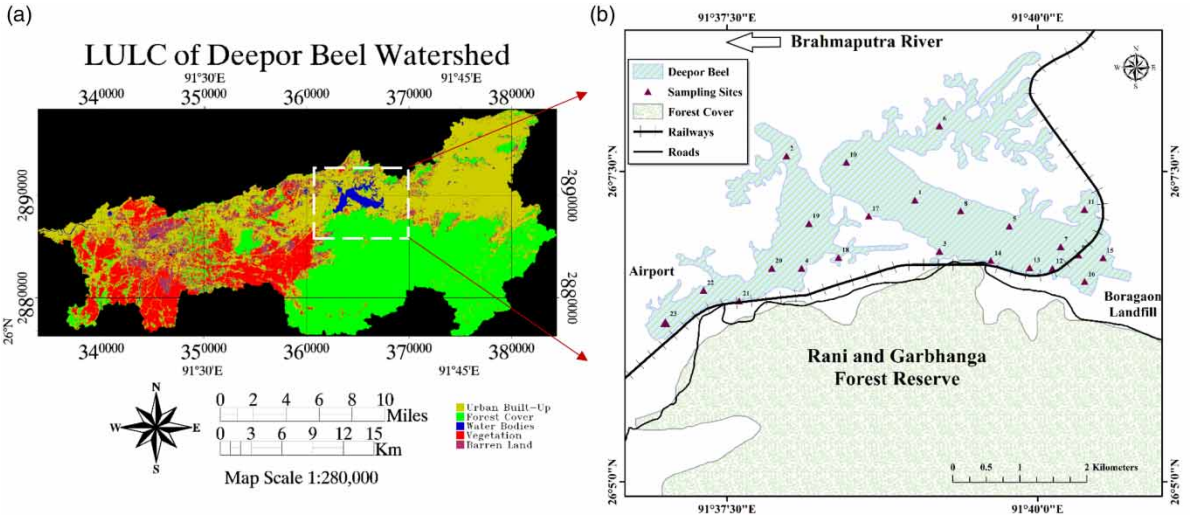


Figure 1 | Deepor Beel (study area) with (a) LULC classes and (b) sampling locations.

landfill in the East) and the growth of small and large scale industries (in the West) in recent times have further added to the degradation of water quality. Given these problems, there is a need for an immediate water quality monitoring program to identify the plausible pollution sources and determining which parameters are primarily responsible for the deterioration of the water quality and also to quantify their contribution.

Hence, based on the reconnaissance survey of the entire wetland, 23 sampling sites were chosen such that they are entirely accessible and are also proximate to various pollution sources (Figure 1(b)). Sampling was conducted continuously from October 2017 to February 2019, twice every month. The water samples were collected from a depth of 30 cm below the top water level in decontaminated and thoroughly scrubbed and washed (in laboratory-grade detergents) reagent bottles. Two bottles were utilized for carrying samples from each location with proper labelling; one for all the physiochemical analyses, and the other for metal analyses. The containers used for metal analyses of water samples were acidified with HNO_3 after the collection to avoid precipitation. To minimize the changes in the samples, chemically or biologically, they were preserved in an icebox during the transfer from the sampling sites to the laboratory and later in the refrigerator when analyses were not carried out. The entire processes of sample collection, preservation, transfer, and investigations complied with the standard protocol (APHA 2012). The WQ parameters investigated in this study, together with their units of measurements and the methodologies adopted for their determination, are listed in Table 1.

Hierarchical cluster analysis of WQ dataset

Hierarchical agglomerative clustering technique is a simple algorithm for partitioning the datasets into groups (Singh *et al.* 2004; Shrestha & Kazama 2007). Each dataset is first assumed to be a unique cluster. Each of these clusters is then joined step by step depending on their proximity to each other, till a stage is arrived at when all the datasets are placed under a single large cluster. The fundamental algorithm for hierarchical agglomerative clustering depends on large inter-cluster and low intra-cluster variance. The basic steps involved in the clustering process have been explained below:

Water Practice & Technology Vol 15 No 4
doi: 10.2166/wpt.2020.078

Table 1 | Statistical summary of physio-chemical water quality parameters of Deepor Beel, Assam

Parameters	Units of measurement	Analytical method	Max	Min	SD	Skewness	Kurtosis	IS:10500 (2012)*
DO	mg/L	Winkler's method	17.53	0.98	3.01	0.21	0.07	–
pH	pH units	Digital pH meter	8.51	5.31	0.52	0.64	0.76	6.5[a]–8.5[b]
EC	μS/cm	Digital conductivity meter	0.58	0.13	0.08	0.79	0.73	–
Turbidity	NTU	Nephelometric turbidimeter	114.50	0.30	20.26	1.65	2.56	1[a]–5[b]
TA	mg/L as $CaCO_3$	APHA titrimetric method	182.00	20.06	29.83	1.31	1.71	200[a]–600[b]
TH	mg/L as $CaCO_3$	APHA titrimetric method	150.00	30.00	21.18	1.76	3.40	200[a]–600[b]
BOD_5	mg/L	5-day BOD test	98.60	4.50	14.28	2.29	7.72	–
COD	mg/L	Closed reflux, titrimetric method	416.87	12.12	57.43	1.87	5.75	–
TDS	mg/L	Oven drying at 103–105°C	874.67	0.00	144.09	1.64	3.76	500[a]–2000[b]
TSS	mg/L		624.17	4.27	111.82	1.52	2.73	–
F^-	mg/L	Ion chromatograph (IC)	5.71	0.00	0.44	9.48	107.62	1.0[a]–1.5[b]
Cl^-	mg/L		123.37	3.77	12.76	4.15	24.50	250[a]–1000[b]
NO_3^-	mg/L		30.11	0.00	3.38	5.31	34.97	45
PO_4^{3-}	mg/L		4.99	0.00	1.13	1.61	1.57	–
SO_4^{2-}	mg/L		110.81	2.41	14.30	3.09	12.78	200[a]–400[b]
Na^+	mg/L	Flame photometer	48.20	0.67	9.92	1.37	1.11	–
K^+	mg/L		34.10	0.07	4.59	1.58	5.60	–
Ca^{2+}	mg/L		184.00	4.21	28.02	1.39	3.17	75[a]–200[b]
TKN	mg/L	APHA distillation method	42.30	2.38	6.87	1.10	1.46	–
Mg	mg/L	Atomic absorption spectroscopy (AAS)	9.26	0.01	2.73	0.29	1.23	30[a]–100[b]
Cr	mg/L		4.63	0.09	0.93	1.26	0.77	0.05
Cd	mg/L		0.11	0.00	0.03	− 0.07	1.43	0.003
Fe	mg/L		4.46	0.01	0.80	1.15	1.59	0.3
Mn	mg/L		0.76	0.01	0.16	0.48	0.36	0.1[a]–0.3[b]
Cu	mg/L		0.80	0.01	0.24	0.33	1.26	0.05[a]–1.5[b]
Pb	mg/L		0.79	0.01	0.12	1.96	5.56	0.01
NH_3^-	mg/L	Semi-automated colorimetry method	5.57	0.20	0.70	2.12	7.24	0.5
OrgN	mg/L	Kjeldahl method	39.83	2.18	6.35	1.12	1.59	–

It is also important to note that elements not having a permissible limit indicate **no relaxation** to their acceptable limit values.
[a]Requirement (*Acceptable limit*).
[b]*Permissible limit* in the absence of any alternate source.
*Standards prescribed by the Indian Standard Drinking Water – Specification {IS: 10500 (2012)} (*Second Revision*).

Step 1: computation of distances between individual data points

The matrix obtained is called the distance matrix, which provides the unit of distance between each dataset (Equation (1)). It is always a lower triangular matrix (m × m), with its diagonal elements being zero, which is quite apparent.

$$\begin{array}{c|ccccc} & P_1 & P_2 & P_3 & P_{m-1} & P_m \\ \hline P_1 & 0 & & & & \\ P_2 & d_{21} & 0 & & & \\ P_3 & d_{31} & d_{32} & 0 & & \\ P_{m-1} & d_{(m-1)1} & d_{(m-1)2} & \boxed{d_{(m-1)3}} & 0 & \\ P_m & d_{m1} & d_{m2} & d_{m3} & d_{m4} & 0 \end{array} \quad (1)$$

Water Practice & Technology Vol 15 No 4
doi: 10.2166/wpt.2020.078

Step 2: choosing the minimum value

From the distance matrix, the clusters closest (i.e. the minimum value in the matrix) to each other are categorized under a single cluster (in this case, let's suppose $d_{(m-1)3}$) (Equation (2)). A distance matrix is again created for $(m-1)$ observations $\{(m-1)\times(m-1)\}$.

	P_1	P_2	$[P_3, P_{m-1}]$	P_4
P_1	0			
P_2	d_{21}	0		
$[P_3, P_{m-1}]$	d_{a1}	d_{a2}	0	
P_4	d_{41}	d_{42}	d_{a3}	0

(2)

Step 3: repeat step 2

The clusters closest to each other are further categorized to a single group, and the process is repeated 'm' times till a single value is obtained; that is, the entire dataset is grouped into a single cluster.

Several methods exist for the determination of distances between clusters, as well as their linkages. Minimum, maximum, and mean distances between the clusters, together with single link, complete link, average link, centroids, and Ward's method (Wilks 2011) of clustering are some of the techniques mostly adopted. Each of the linkage methods has its significance. However, the use of Ward's method for agglomeration of clusters using the Squared Euclidean distance measure has been proved to be the most powerful, represented by Equations (3) and (4) respectively:

$$TD_{C_1\cup C_2} = \sum_{x\in C_1\cup C_2} D(x,\ \mu_{C_1\cup C_2})^2 \tag{3}$$

where TD denotes the total distance from the centroids; x denotes the distance between the clusters; and μ is the centre of the clusters C_1 and C_2

$$\text{Distance}\ \ (P_i, P_j) = \sum_{j=1}^{m} (Q_{1i} - Q_{2j})^2 \tag{4}$$

where P_i represents the i^{th} object and Q_{ij} the value of the j^{th} variable of the i^{th} object.

The results of hierarchical clustering is usually represented by a dendrogram which provides essential visual insight into the clustering process.

Discriminant analysis

Discriminant analysis is an appropriate technique for construction of categorically dependent values from statistically classified samples. It necessitates a prior knowledge of the statistical association of objects in any particular cluster or group for understanding which variable belongs to which group or cluster. For statistical analysis of a water quality dataset, the DA technique assists in determining the parameters that are primarily responsible for the discrimination of the clusters, by creating a discriminant score (Equation (5)) for each of the individuals of the raw WQ data.

$$\text{Discriminant Score,}\quad Z = l_i + \sum_{j=1}^{n} w_{ij}\, p_{ij} \tag{5}$$

Water Practice & Technology Vol 15 No 4
doi: 10.2166/wpt.2020.078

where, i and n signify the number of groups and parameters used for DA respectively; l denotes the constant value which is characteristic to each dataset; w_j represents the weight factor assigned by DA to each parameter p_j.

Another essential statistical tool used in DA is Wilk's Lambda (Equation (6)), for determining whether there is a clear distinction between the groups. Wilk's lambda (λ) represents the ratio of the variances measured within a group to the total variance. Hence, it can be established that lower values of λ are a clear indication that the groups are distinctive, and there is less overlapping between the clusters or groups and vice-versa.

$$\lambda = \frac{\text{Within group variance}}{\text{Total variance}} \tag{6}$$

DA was carried out on the raw dataset for both standard, as well as stepwise modes. The standard method involves a step-by-step build-up of the model of discrimination in which all the variables are assessed and evaluated to determine which one will contribute most to the discrimination between groups individually, at each step. That variable is then incorporated in the model, and the procedure starts again. On the other hand, in stepwise mode, all variables are included in the model initially, after which the variable contributing the minimum to the estimation of group membership is eliminated at each step. Thus, only the 'important' variables in the model which contribute the most to the discrimination between groups are kept as the result of a successful discriminant function analysis (Astel *et al.* 2009).

Principal component analysis

The principal component analysis (PCA) forms a powerful tool for minimizing the problem of data overfitting to a model, which has led to several confusions due to the inclusion of too many attributes and variables into it. The primary objective of the PCA is to convert a dataset having high dimensionality to a lower dimensionality dataset, without actually losing vital information, thus effectively reducing the problem of overfitting (Vega *et al.* 1998; Wunderlin *et al.* 2001; Chen *et al.* 2007). The reduced set of variables formed are termed as principal components (PCs), which depend on the number of attributes the dataset possesses, as well as its dimension. The number of PCs formed for a particular dataset is always less than or at a maximum scale, equal to the number of attributes given for model preparation. Also, the PCs formulated should be independent and orthogonal to each other. The necessary steps involved in the formulation of PCs is described below.

Step 1: compute the mean values of all the attributes

Suppose we have 'm' attributes (A_1, A_2, …, A_m), the values of those for 'n' sampling locations (variables) ($S_{11, \ldots,} S_{nm}$) can be represented as shown in Equation (7)., the first operation performed is the computation of the mean values for all the attributes ($\overline{A_1}$, $\overline{A_2}$, …, $\overline{A_m}$).

$$\begin{array}{ccccc} A_1 & A_2 & A_3 & \ldots & A_m \\ S_{11} & S_{12} & S_{13} & \ldots & S_{1m} \\ S_{21} & S_{22} & S_{23} & \ldots & S_{2m} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ S_{n1} & S_{n2} & S_{n3} & \ldots & S_{nm} \\ \hline \overline{A_1} & \overline{A_2} & \overline{A_3} & \ldots & \overline{A_m} \end{array} \tag{7}$$

Water Practice & Technology Vol 15 No 4
doi: 10.2166/wpt.2020.078

Step 2: preparation of the covariance matrix

The next step includes the preparation of the covariance matrix for all the attributes (Equation (8)) based on the formula given in Equation (9). The matrix formulated will be an (m × m) matrix (C).

$$C = \begin{vmatrix} \text{cov}(A_1, A_1) & \text{cov}(A_1, A_2) & \text{cov}(A_1, A_3) & \dots & \text{cov}(A_1, A_m) \\ \text{cov}(A_2, A_1) & \text{cov}(A_2, A_2) & \text{cov}(A_2, A_3) & \dots & \text{cov}(A_2, A_m) \\ \text{cov}(A_3, A_1) & \text{cov}(A_3, A_2) & \text{cov}(A_3, A_3) & \dots & \text{cov}(A_3, A_m) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \text{cov}(A_m, A_1) & \text{cov}(A_m, A_2) & \text{cov}(A_m, A_3) & \dots & \text{cov}(A_m, A_m) \end{vmatrix} \tag{8}$$

where the covariance is determined using Equation (9);

$$\text{cov}\ (A_x, A_y) = \sum_{i=1}^{m} \left\{ \frac{(x_i - \bar{x})(y_i - \bar{y})}{m - 1} \right\} \tag{9}$$

Step 3: determination of eigenvalues and eigenvectors

After the covariance matrix is generated, the eigenvalues and their corresponding eigenvectors are generated using Equations (10)–(13) respectively.

$$|C - \lambda I| = 0 \tag{10}$$

where I is the identity matrix, the size of which depends on the covariance matrix (in this case m × m).

$$I = \begin{vmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{vmatrix} \tag{11}$$

Solving Equation (10), we obtain the values $\lambda_1, \lambda_2, \lambda_3, \dots, \lambda_m$. Using these λ values, we compute the eigenvectors by substituting the λ's as shown;

$$\begin{aligned} [C - \lambda_1 I] \begin{Bmatrix} X_{11} \\ X_{12} \\ X_{13} \\ \vdots \\ X_{1n} \end{Bmatrix} &= 0 \\ [C - \lambda_2 I] \begin{Bmatrix} X_{21} \\ X_{22} \\ X_{23} \\ \vdots \\ X_{2n} \end{Bmatrix} &= 0 \end{aligned} \tag{12}$$

Water Practice & Technology Vol 15 No 4
doi: 10.2166/wpt.2020.078

Equation (12) will then provide the values for $\{X_{ij}\}_{(i,j)\ \varepsilon\ (1,m)}$. This will further lead to the formulation of the eigenvector matrix, represented by Equation (13).

$$\begin{pmatrix} X_{11} & X_{12} & X_{13} & \cdots & X_{1m} \\ X_{21} & X_{22} & X_{23} & \cdots & X_{2m} \\ X_{31} & X_{32} & X_{33} & \cdots & X_{3m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ X_{m1} & X_{m2} & X_{m3} & \cdots & X_{mm} \end{pmatrix} \tag{13}$$

Step 4: determination of PCs

PCs depend on the eigenvalues of the corresponding attributes. The eigenvalues are first considered in the descending order of their numerical values. The eigenvectors of the corresponding largest eigenvalue become PC1 and subsequently, other PCs are also determined (the corresponding eigenvectors) based on their descending order of eigenvalues.

Varimax rotation, Keiser-Meyer-Olkin (KMO) (Sahu *et al.* 2013) and Bartlett's sphericity (Dalal *et al.* 2010) were employed for the orthogonal rotation, evaluating the adequacy of sampling and estimating the applicability of PCA to the crude dataset, respectively. Eigenvalues greater than 1 were only considered for accountability of the factor loadings.

Positive matrix factorization analysis

The positive matrix factorization (PMF) is one of the methods for classification of objects using linear algebra (Paatero & Tapper 1994). The preliminary step involved in this method is the representation of document class by suitable vectors (Equation (14)).

$$d_j^T = [w_{1j},\ w_{2j},\ w_{3j},\ \ldots,\ w_{aj}] \tag{14}$$

where w_{ij} indicates the weighted frequency of the i^{th} item in the j^{th} class.

Considering a matrix $A_{a\times b}$; m and n denoting the dependent and independent variables respectively (Equation (15)),

$$A = \begin{array}{c|ccccc} & d_1 & d_2 & d_3 & \ldots & d_b \\ w_1 & & & & & \\ w_2 & & & & & \\ w_3 & & & & & \\ \vdots & & & & & \\ w_a & & & & & \end{array} \tag{15}$$

The following inferences can be established:

- $A_{a\times b}$ covers a multi-dimensional space ($A\ \varepsilon\ \Re^{a\times b}$).
- $A_{a\times b}$ may be sparse.
- This may lead to noisy features in additional data processing, ultimately providing poor classification.

The use of an appropriate dimension-reduction tool, like the Singular Value Decomposition (SVD) proves to be the most effective solution. This form the chief elementary principle of the PMF

Water Practice & Technology Vol 15 No 4
doi: 10.2166/wpt.2020.078

technique, that accounts for only non-negative items in the matrix ($A_{a\times b}$) database. On the other hand, any other item having a negative value is rendered to be zero (Al-Dabbous & Kumar 2015; Li *et al.* 2015; Liu *et al.* 2015). The PMF technique works on the principle that a non-negative matrix ($A_{a\times b}$) can be factorized into two non-negative matrices W and H (Equation (16)) such that

$$A_{a\times b} = W_{a\times k}H_{k\times b} \ \{k \ll \min(a, b)\} \tag{16}$$

where W and H represent the basis matrix and weight matrix respectively.

The following minimization problem is used in the PMF model for categorization of classes (Equation (17)):

$$\text{Minimize } f(W, H) = \frac{1}{2}\|A - WH\|^2 \tag{17}$$

A multiplicative update algorithm is usually employed for solving this minimization problem, which is carried out using the following steps:

Step 1: initialise W & H

W and H are randomly initialised with non-negative values as represented in Equation (18).

$$\begin{aligned} W &= \text{rand}(a, k) \\ H &= \text{rand}(k, b) \end{aligned} \tag{18}$$

Step 2: update

W and H are updated as in Equation (19).

$$\begin{aligned} H_{ij} &\leftarrow H_{ij}\frac{(W^TA)_{ij}}{(W^TWH)_{ij} + \varepsilon} \\ W_{ij} &\leftarrow W_{ij}\frac{(AH^T)_{ij}}{(WHH^T)_{ij} + \varepsilon} \end{aligned} \tag{19}$$

where error (ε) is initialised with an insignificant value to avoid division by zero.

Step 3: iterate

Iterations are carried out using step 2 till W and H converge with each other and become stable.

All the statistical techniques were carried out using various software; Hierarchical clustering, DA (both standard and stepwise) and PCA using IBM – SPSS Statistics (Version 25) and the PMF model was formulated using EPA – PMF (Version 5.0).

RESULTS AND DISCUSSION

Descriptive statistics of WQ dataset

A summary of the overall descriptive statistics of various WQ parameters is shown in Table 1. It was observed that most of the elements exceeded the acceptable limit as set by IS:10500 (2012), which is

Water Practice & Technology Vol 15 No 4
doi: 10.2166/wpt.2020.078

an indicator for water contamination of the wetland. Only Cd exhibited negative skewness, while all other parameters exhibited positive skewness, with a majority lying in the highly positively skewed (skewness > 1) category. This indicates that they have a long right-hand tail curve, when normally distributed. As far as the kurtosis of the observed dataset is concerned, about 57.2% of the parameters were observed to be of the Plytokurtic category (Kurtosis < 3), thus indicating a flatter, normally distributed curve. On the other hand, the rest were found to be Leptokurtic (kurtosis > 3), indicating a higher peak representation of the normally distributed data.

Hierarchical clustering of sampling locations and its validation

The hierarchical clustering of the 23 sampling locations of Deepor Beel resulted in classifying them into three major groups (clusters 1, 2, and 3), represented through a dendrogram (Figure 2(a)). The clustering was then validated using the agglomeration schedule coefficients, which generated a plot like a scree plot, only backward (Figure 2(b)). The number of agglomerative schedule coefficients represents the amount of heterogeneity in a cluster solution (known as the Stopping Rule in hierarchical clustering). Hence, our main objective was to maximize the heterogeneity (distinction) for obtaining better clustering results. A single factor solution (single cluster) is the most distinct, with the next best being the split-up of the cluster into two parts; that is, the formation of two clusters. However, often, two numbers of clusters are not considered to be ideal. Hence, the check is carried out for three clusters, then four, and so on until no significant jump (flatter slope) arises, wherein the distinction is considered unsuitable. In the current study, it was observed that a steep slope was produced until the third cluster process, after which the slope of the curve became relatively flatter, thus indicating a lack of difference in the homogeneity among the clusters. This can also be visualized in Figure 2(a), where distinct clusters were observed until the third clustering, after which the distance between the clustering schedule decreased significantly with the formulation of new clusters. This provided a relatively unclear picture about the further clustering process. Thus, three clusters (Figure 2(a)) were considered for the analysis. The sites categorized into three clusters were then plotted on a GIS platform for better visualization of the sampling locations (Figure 2(c)). It was observed that the three locations belonging to cluster 3 were in the closest proximity to Boragaon landfill site, thus providing evidence for maximum possible contamination among all the locations. Furthermore, cluster 2 categorizes all the sites located in the middle portion of the wetland, which has remained devoid of any significant anthropogenic contamination, thus indicating the sites having the least pollution. Lastly, the locations categorized under cluster 1 were observed to be primarily belonging to the areas proximate to the industrial complex (Western side of the wetland), or between the landfill site and the middle area. This is evidence of a moderate level of pollution, the sections between the landfill and the central portion of the wetland acting as a transition zone between high and medium contamination, respectively.

Discriminant analysis

The discriminant analysis (DA) technique was employed on the raw WQ dataset to determine the spatio-temporal variation of the WQ parameters. Standard and stepwise modes were used for constructing the classification functions (CFs) (Table 2). It was observed that only 9 (EC, TDS, TSS, PO_4^{3-}, Na^+, Mg, Cd, Pb and OrgN) out of 28 parameters (based on Fisher's Linear Discriminating Functions) were responsible for the discrimination, as well as the spatial variability among the three clusters. The generation of the CFs was validated using the classification matrix (Table 3), which provided correct classifications among the discriminating functions using the stepwise mode of DA for 73.9% of the cross-validated grouped cases, thereby providing an excellent result for the spatial variations. The

Water Practice & Technology Vol 15 No 4
doi: 10.2166/wpt.2020.078

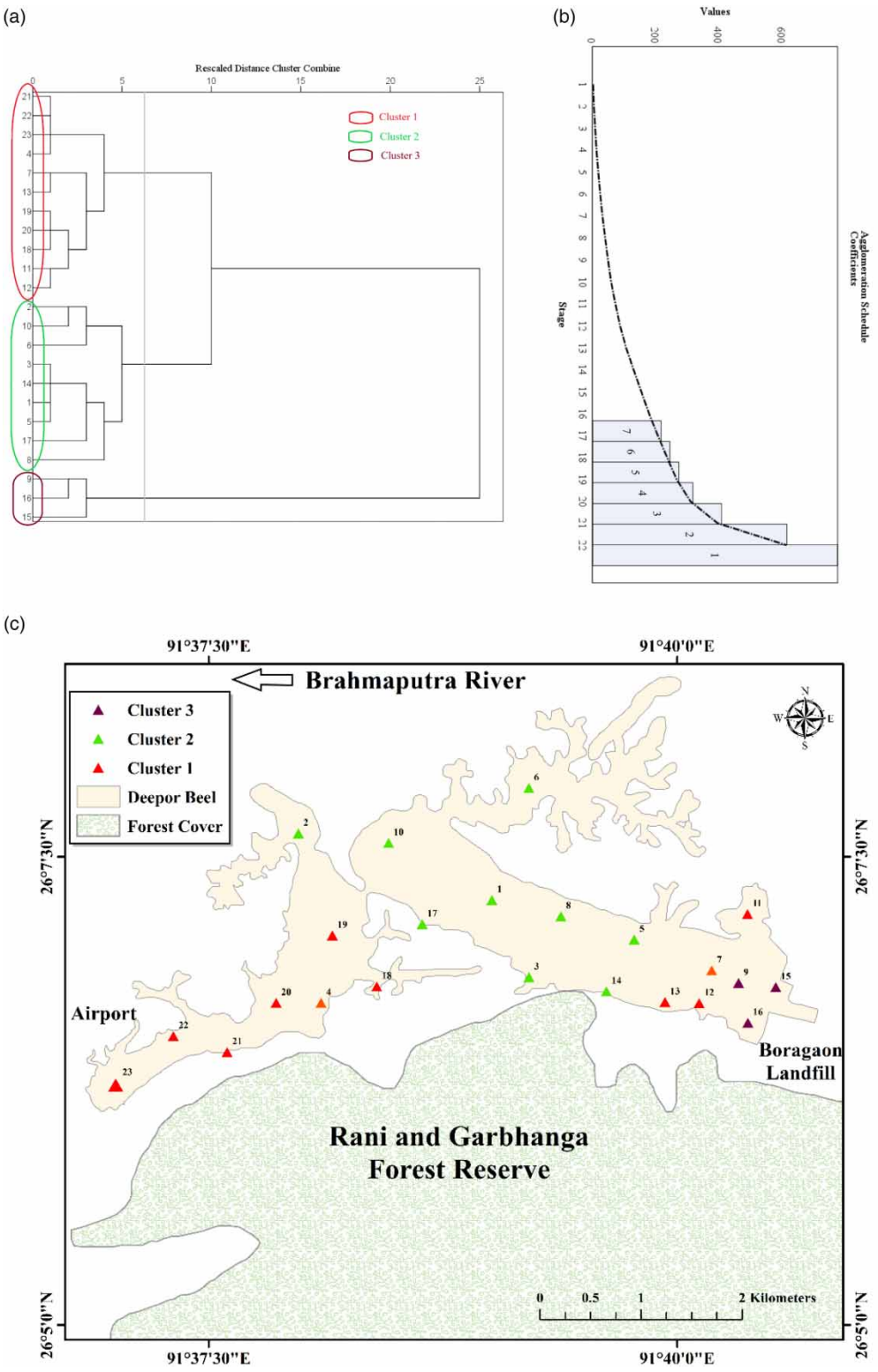


Figure 2 | Hierarchical clustering: (a) dendrogram representation, (b) validation of clustering, and (c) representations of sampling locations through GIS plotting.

scores of the two functions were plotted (Figure 3(a)) along with the Wilk's lambda (λ) values of the discriminating functions (Figure 3(b)). The values of λ varied from 0 to 0.14, which is numerically insignificant, thus indicating that the classes or groups formed are distinctive and there is minimum overlapping between them.

Water Practice & Technology Vol 15 No 4
doi: 10.2166/wpt.2020.078

Table 2 | Classification functions for discriminant analysis of the spatial variations along Deepor Beel

Classification function coefficients

Standard mode				**Stepwise mode**		
	Cluster			**Cluster**		
	1.00	**2.00**	**3.00**	**1.00**	**2.00**	**3.00**
DO	− 279.321	− 217.418	− 517.047			
pH	− 244.949	621.552	− 1,946.884			
EC	44,886.626	56,994.354	17,176.709	15,857.332	13,221.401	13,926.393
Turbidity	11.420	− 16.576	58.262			
Alkalinity	278.840	227.627	389.087			
Hardness	40.375	− 16.160	136.897			
BOD_5	− 516.219	− 422.640	− 648.016			
COD	69.615	31.753	120.657			
TDS	54.132	44.286	79.294	− 5.362	− 4.031	− 2.896
TSS	− 1.473	.097	− 1.858	− .418	− .177	.180
F^-	1,972.947	1,646.654	2,419.821			
Cl^-	− 81.130	− 56.241	− 109.744			
NO_3^-	− 66.961	− 47.201	− 57.353			
PO_4^{3-}	517.219	489.412	1,596.487	524.481	569.092	943.303
SO_4^{2-}	3.645	13.609	− 20.632			
Na^+	301.637	326.161	218.933	92.884	86.114	118.290
K^+	187.663	85.017	440.358			
Ca^{2+}	− 69.354	− 49.078	− 110.620			
TKN	98.193	20.189	203.213			
Mg	524.014	499.660	488.266	553.758	466.641	490.048
Cr	726.350	713.277	582.055			
Cd	20,755.572	20,470.825	28,886.247	64,908.078	63,699.964	95,741.628
Fe	− 3,866.103	− 3,853.944	− 3,543.238			
Mn	4,096.467	4,151.461	3,587.756			
Cu	18,679.943	18,892.675	17,487.073			
Pb	4,560.400	4,596.263	3,600.740	− 13,594.566	− 12,600.372	− 16,764.891
NH_3^-	− 422.741	− 441.312	− 448.577			
OrgN	72.609	73.789	75.372	60.170	55.527	72.409
(Constant)	− 25,656.941	− 24,058.597	− 29,344.87	− 4,213.607	− 3,517.647	− 6,156.426

Fisher's linear discriminant functions.

Identification of latent pollution sources

PCA employed on the normalized WQ dataset resulted in the formation of two primary principal components (PC 1 and PC 2), with eigenvalues >1. KMO and Bartlett's sphericity test results (Table 4) revealed the measure of sampling adequacy to be 0.742, with 253 degrees of freedom and a chi-square value of approximately 3,743.39 and a value of p close to zero ($p < 0.05$). This indicates the validity of the PCA due to their statistical significance between the variables (Kaiser 1974). Figure 4 represents the rotation factor matrix (Varimax rotation) of the two PCs. Values of PCs less than ± 0.3 were considered to have weak loadings, while those higher than ± 0.7 were regarded as PCs having a significant contribution with strong loading values. The PC values lying between $\pm$ (0.3–0.7) were considered to have a moderate impact. It was clearly observed that COD, BOD_5, TDS, EC, Cd, Cr, Fe, Pb, Cu, Mn, and Mg exhibited strong positive loadings for PC 1, indicating substantial

Water Practice & Technology Vol 15 No 4
doi: 10.2166/wpt.2020.078

Table 3 | Classification matrix for discriminant analysis of the spatial variation along Deepor Beel

Classification results[a,c]

			Predicted group membership			
		Cluster	1.00	2.00	3.00	Total
Original	Count	1.00	11	0	0	11
		2.00	0	9	0	9
		3.00	0	0	3	3
	% Correct	1.00	100.0	0	0	100.0
		2.00	0	100.0	0	100.0
		3.00	0	0	100.0	100.0
Cross-validated[b]	Count	1.00	8	1	2	11
		2.00	0	7	2	9
		3.00	1	0	2	3
	% Correct	1.00	72.7	9.1	18.2	100.0
		2.00	0	77.8	22.2	100.0
		3.00	33.3	0	66.7	100.0

[a]100.0% of original grouped cases correctly classified.
[b]Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.
[c]73.9% of cross-validated grouped cases correctly classified.

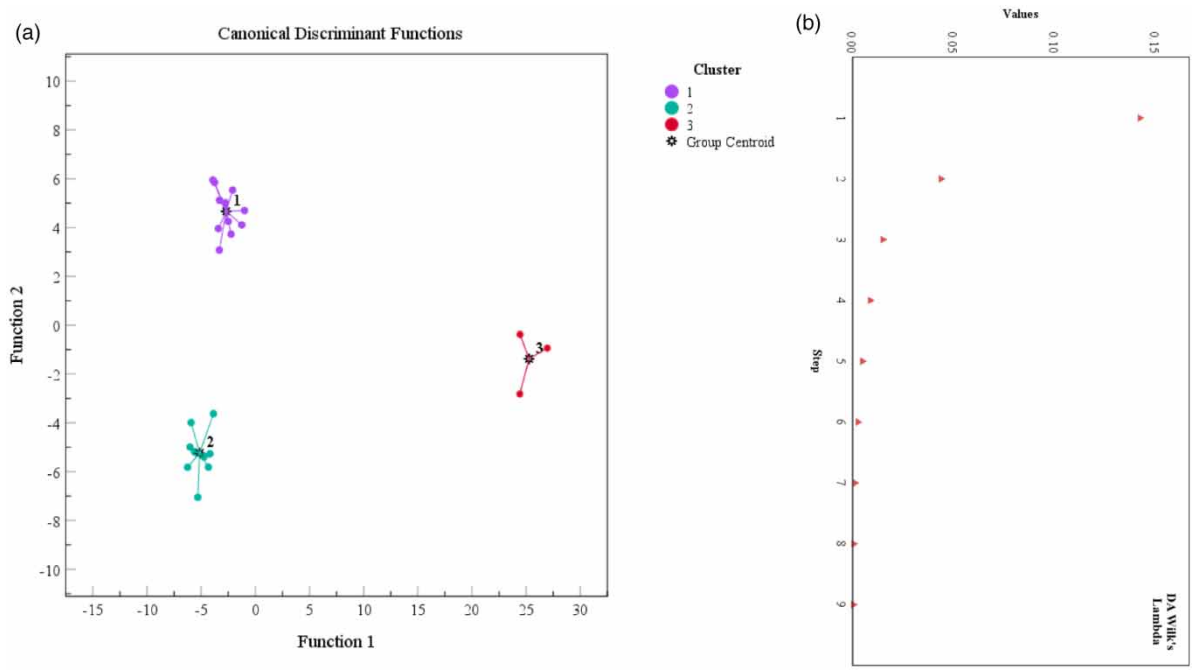


Figure 3 | (a) Scatter plot for the spatial discrimination analysis of water quality variations across three clusters (DA stepwise mode), (b) Wilk's Lambda values for the 9 discriminating parameters.

Table 4 | The results of KMO and Bartlett's sphericity test (obtained through PCA)

KMO and Bartlett's Test[a]

Kaiser-Meyer-Olkin Measure of Sampling Adequacy		0.742
Bartlett's Test of Sphericity	Approximate Chi-Square	3,743.387
	Degrees of Freedom	253
	Significance level	0.000

[a]Based on correlations.

Water Practice & Technology Vol 15 No 4
doi: 10.2166/wpt.2020.078

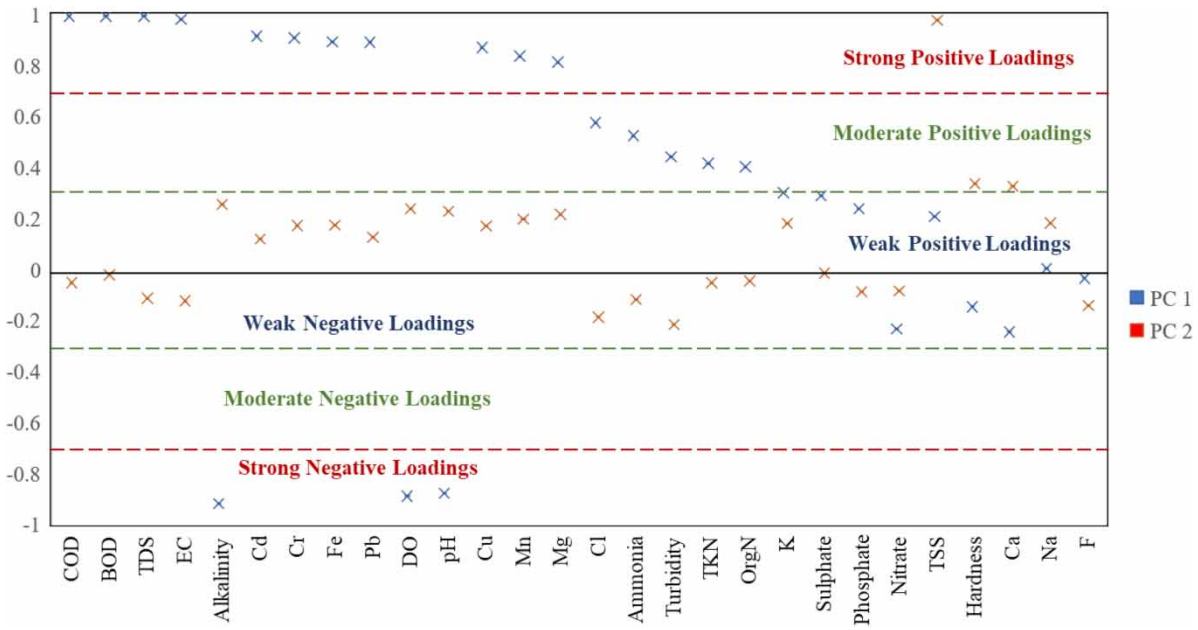


Figure 4 | Rotation factor matrix (Varimax with Kaiser Normalization).

anthropogenic interference such as the Boragaon landfill and the industries in the wetland. Strong negative loading of DO displayed good results in PC 1 as the values of DO vary inversely with the BOD/COD values. Furthermore, moderate loading values for Cl, NH_3^-, turbidity, TKN, and OrgN indicate a secondary source of organic contamination to the wetland (possibly from the Basistha River, which drains its water, primarily composing municipal wastewater, into Deepor Beel). PC 2, however, showed a strong positive loading for TSS while having moderate loadings for TH and Ca, indicative of a natural surface water runoff being the major probable contributor.

Figure 5 presents an overall summary of the PCA carried out for the WQ dataset. It provides the parameters (variables) that are associated more with a particular component along with the correlations existing between them. For example, it can be observed that BOD_5, EC, TDS, COD, and all the metals were observed to have a higher affinity towards PC 1. Similar is the case for TA, DO and pH. However, they lie on the opposite quadrant, thus displaying a highly negative correlation. Similarly, TSS, TH, Na^+, Ca^{2+}, NO_3^-, and F^- are affiliated more towards PC 2, with F^- and NO_3^- being negatively correlated to other variables. Figure 6 is a spatial representation of the PC loading values (both PC 1 and PC 2). Clearly, PC 1 had higher loading values near the landfill site and the industrial zone, representing sites having higher anthropogenic contamination factor (Figure 6(a)). Similarly, PC 2 showed higher values towards the central region of the wetland (Figure 6(b)), thus indicating natural factors of contamination of Deepor Beel.

Source apportionment using PMF model

The source apportionment and various factor contributions were estimated using the EPA – PMF software, which takes into account the concentration and uncertainty factors as input to the model. The resulting input parameters then generated a signal/noise ratio for each of the WQ parameters. Based on the S/N ratio, the parameters were classified as Strong ($S/N > 2$), Weak ($0.2 < S/N < 2$) and Bad ($S/N < 0.2$) respectively. Four factors; 1, 2, 3, and 4 were considered for the PMF analysis to assess each of their contributions to the water pollution of Deepor Beel, based on the iterations for each factor obtaining the minimum Q(Robust)/Qexp value (Table 5). The model was then simulated, considering all the input parameters and the four factors, providing

Water Practice & Technology Vol 15 No 4
doi: 10.2166/wpt.2020.078

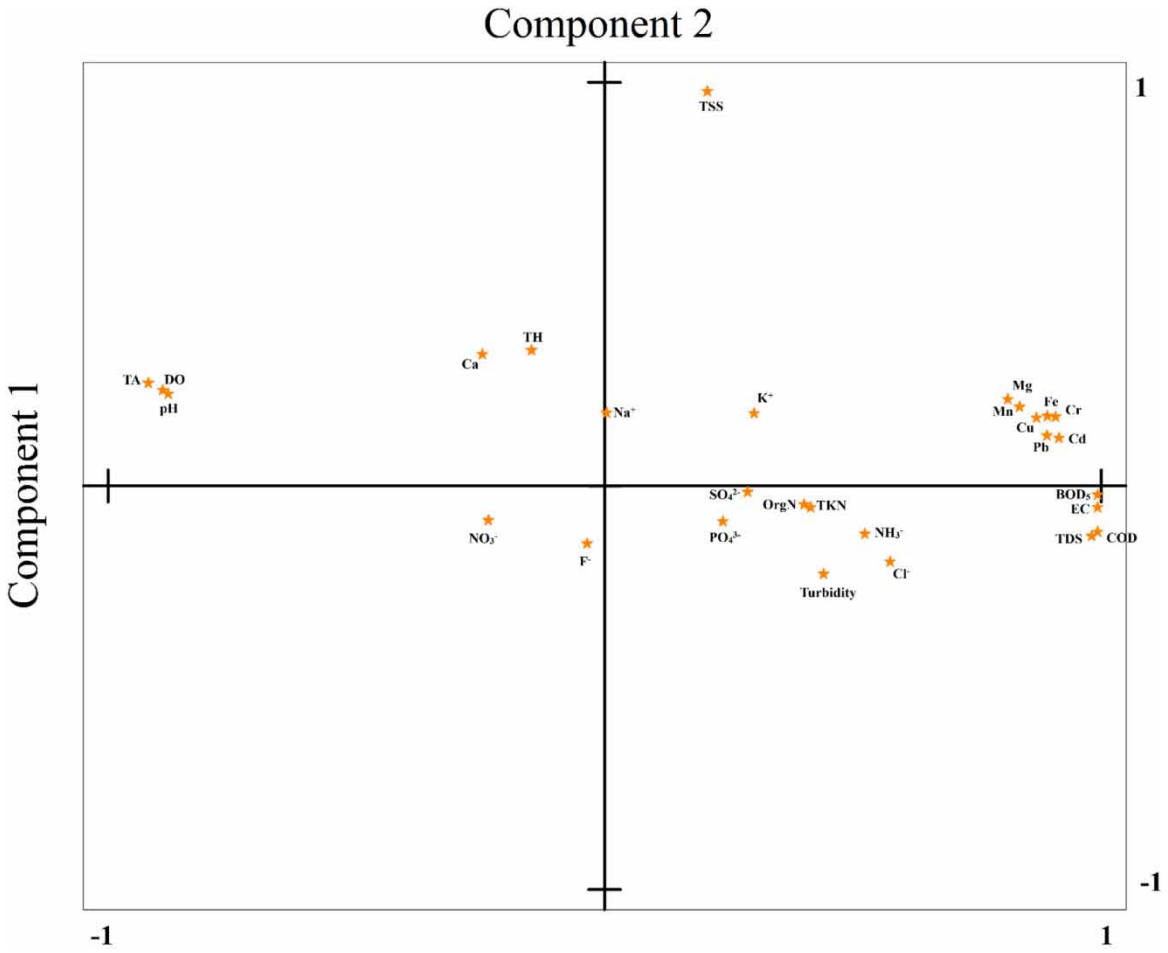


Figure 5 | PCA analysis results showing plot between the two PCs in a rotated space.

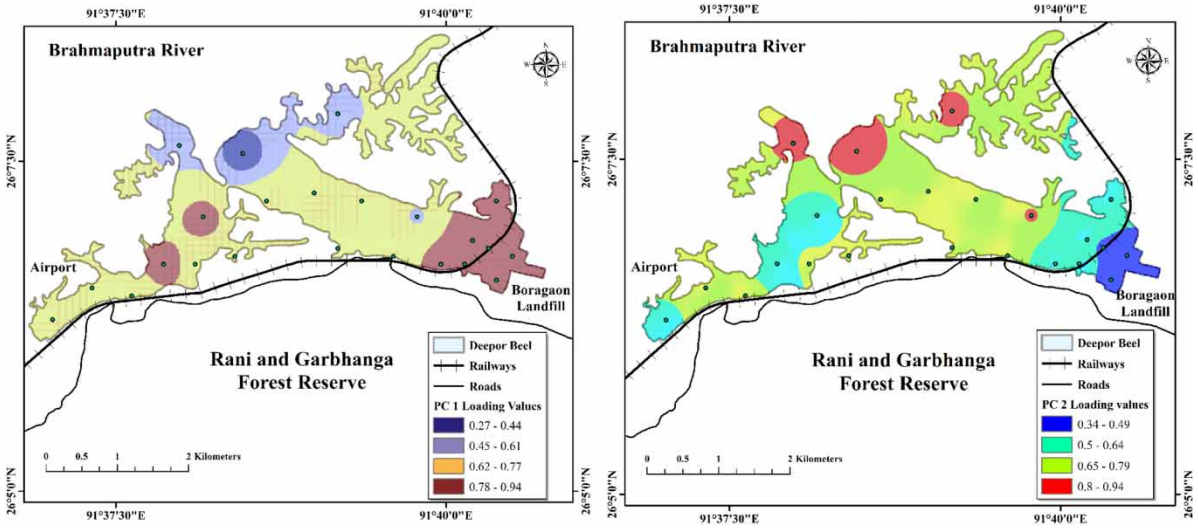


Figure 6 | GIS mapping showing the distribution of principal factors (a) PC 1 and (b) PC2 for each of the sampling locations.

results, as shown in Figures 7 and 8. It can be seen that Factor 1 is primarily dominated by BOD_5 (69.3%), COD (71%), Cr (78.5%), Fe (71.9%), Mn (70.6%) and Pb (58.6%). This indicates a source rich in organic matter, as well as heavy metals. Thus, leaching from the Boragaon landfill site can

Water Practice & Technology Vol 15 No 4
doi: 10.2166/wpt.2020.078

Table 5 | Summary of PMF and EE diagnostics by run for water quality data of Deepor Beel

Diagnostic	2 factors	3 factors	4 factors	5 factors	6 factors	7 factors
Qexp	247.9	358.1	1,358.1	2,043.1	830.4	375.8
Q(True)	9,731.1	7,069.0	1,566.9	3,288.5	3,290.6	4,023.8
Q(Robust)	9,649.1	7,017.3	1,566.8	3,245.3	3,300.0	4,001.9
Q(Robust)/Qexp	38.92	19.59	1.15	1.59	3.97	10.64

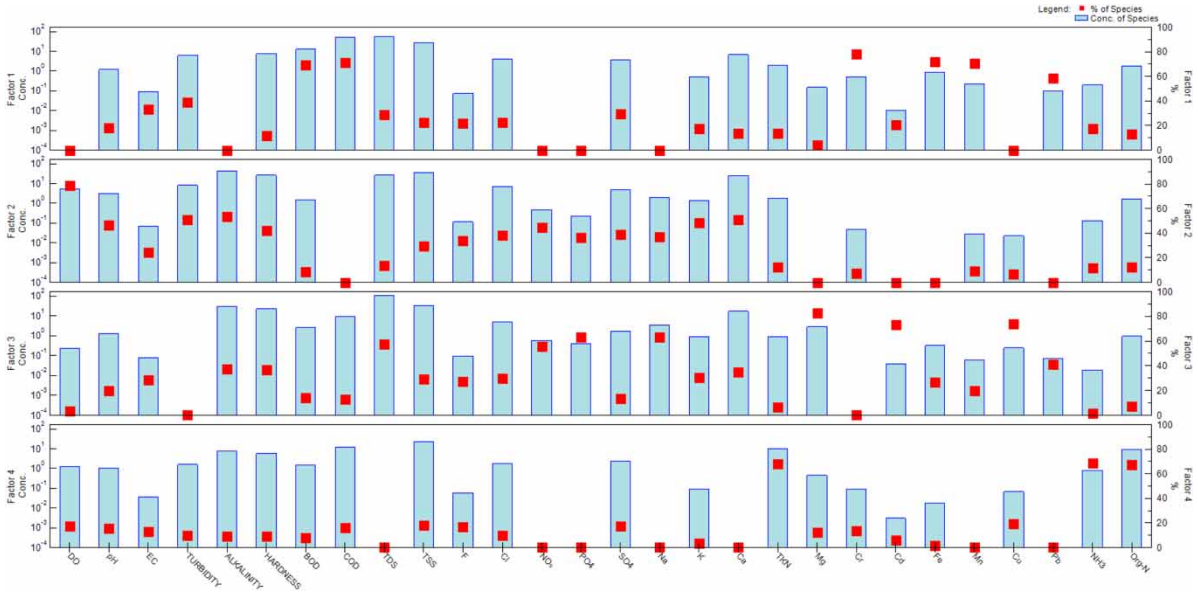


Figure 7 | PMF source profiling of 28 WQ parameters for Deepor Beel.

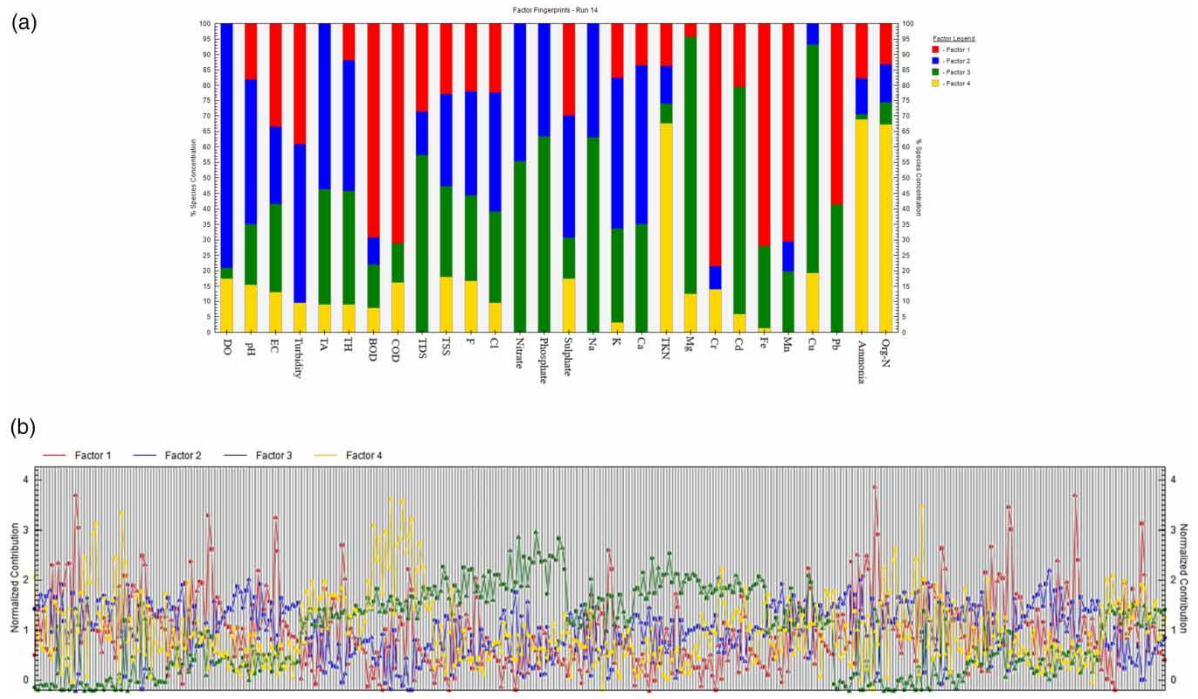


Figure 8 | (a) Factor Figureprints and (b) Normalized contribution profile of all the 28 parameters resulted from EPA – PMF model.

Water Practice & Technology Vol 15 No 4
doi: 10.2166/wpt.2020.078

be considered as the major source of contamination for Factor 1, since the municipal landfill site leachates are highly organic in nature owing to the various organic wastes (e.g. food wastes) being dumped into the landfill. Also, the inorganic wastes rich in heavy metals such as batteries, scrap and unused metals contribute to the leaching of heavy metals (Chu *et al.* 1994). Furthermore, DO (79.1%), pH (46.9%), TA (53.6%), TH (42.3%), Turbidity (51.4%), K (48.7%) and Ca (51.2%) were observed to have significant contribution to water contamination of Deepor Beel, pertaining to Factor 2. This is an indicator for a natural cause of pollution such as the surface water runoff. Similarly, for Factor 3, TDS (57.5%), NO_3^- (55.4%), PO_4^{3-} (63.5%), Na^+ (63%), Mg (83.2%), Cd (73.7%), and Cu (74%) were found to be the predominating parameters. Industrial wastewater effluents enriched with contagious trace metal elements along with cations and anions might be the primary factor responsible for this. Finally, TKN (67.6%), NH_3^- (68.8%) and OrgN (67.2%) were found to be the parameters affecting Factor 4 the most. This indicates contamination from a source primarily rich in nutrients. Thus, discharge from Basistha River (carrying large concentrations of domestic sewage) might be the primary factor responsible. Figure 8(a) and 8(b) represent the Factor Figureprints and the normalized contributions for all the 28 WQ parameters, providing a summary of the PMF results.

CONCLUSION

This study presented a detailed insight into the geochemistry and assessment of water quality as well as apportioning of various pollution sources contributing to the water contamination of Deepor Beel using different Environmetrics tools. Critical concluding remarks of the study are as follows:

- Hierarchical clustering categorized the 23 sampling locations of the study area into three statistically distinct clusters. The clustering process was further validated using the 'Stopping Rule'. It was observed that the sites closest to the Boragaon landfill, locations present in the central portion and the industrial zone of the wetland formed separate and distinct clusters based on the site similarities. They were considered as sites of high, low, and moderate contaminations, respectively.
- The discriminant analysis provided the significance of clustering by discriminating the parameters responsible. Nine (EC, TDS, TSS, PO_4^{3-}, Na^+, Mg, Cd, Pb and OrgN) out of 28 parameters were found to be responsible for the discrimination, as well as the spatial variability among the three clusters. Low Wilk's λ values also validated that the groups formed were distinctive with minimum overlapping between them.
- The principal component analysis applied to the normalized dataset yielded two principal components, which aided in the identification of potential pollution sources. PC 1 indicated anthropogenic contamination as the primary source, while PC 2 suggested natural phenomena such as surface water runoff to be the primary source. Spatial representation of the sites also revealed similar results.
- The WQ dataset matrix was finally subjected to factorization, generating a model for source apportionment. Leaching from Boragaon landfill site, surface water runoff, discharge of effluents from the industries in the wetland and discharge from Basistha River were found to be the significant factors for the pollution of Deepor Beel.
- The results obtained from this study would be of immense help for restoration and revival of Deepor Beel.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

Water Practice & Technology Vol 15 No 4
doi: 10.2166/wpt.2020.078

ACKNOWLEDGEMENTS

The authors extend their sincere thanks to the anonymous reviewers for their valuable suggestions, which helped improve the quality of the manuscript substantially.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

REFERENCES

Al-Dabbous, A. N. & Kumar, P. 2015 Source apportionment of airborne nanoparticles in a Middle Eastern city using positive matrix factorization. *Environmental Science: Processes & Impacts* **17**(4), 802–812.

APHA 2012 *Standard Methods for the Examination of Water and Wastewater*. American Public Health Association, American Water Works Association, Water Environment Federation, Washington, DC.

Astel, A., Małek, S. & Krakowian, K. 2009 Multivariate exploration and classification applied to the chemical composition of spring waters in sanctuary forest areas. *International Journal of Environmental and Analytical Chemistry* **89**(8–12), 597–620.

Bhattacharyya, K. G. & Kapil, N. 2010 Impact of urbanization on the quality of water in a natural reservoir: a case study with the Deepor Beel in Guwahati city, India. *Water and Environment Journal* **24**(2), 83–96.

Bodrud-Doza, M., Islam, A. T., Ahmed, F., Das, S., Saha, N. & Rahman, M. S. 2016 Characterization of groundwater quality using water evaluation indices, multivariate statistics and geostatistics in central Bangladesh. *Water Science* **30**(1), 19–40.

Chen, K., Jiao, J. J., Huang, J. & Huang, R. 2007 Multivariate statistical evaluation of trace elements in groundwater in a coastal area in Shenzhen, China. *Environmental Pollution* **147**(3), 771–780.

Chen, P., Li, L. & Zhang, H. 2015 Spatio-temporal variations and source apportionment of water pollution in Danjiangkou Reservoir Basin, Central China. *Water* **7**(6), 2591–2611.

Chow, M., Shiah, F., Lai, C., Kuo, H., Wang, K., Lin, C., Chen, T., Kobayashi, Y. & Ko, C. 2016 Evaluation of surface water quality using multivariate statistical techniques: a case study of Fei-Tsui Reservoir basin, Taiwan. *Environmental Earth Sciences* **75**(1), 6.

Chu, L., Cheung, K. & Wong, M. H. 1994 Variations in the chemical properties of landfill leachate. *Environmental Management* **18**(1), 105–117.

Dalal, S., Shirodkar, P., Jagtap, T., Naik, B. & Rao, G. 2010 Evaluation of significant sources influencing the variation of water quality of Kandla creek, Gulf of Katchchh, using PCA. *Environmental Monitoring and Assessment* **163**(1–4), 49–56.

Gholizadeh, M. H., Melesse, A. M. & Reddi, L. 2016 Water quality assessment and apportionment of pollution sources using APCS-MLR and PMF receptor modeling techniques in three major rivers of South Florida. *Science of the Total Environment* **566**, 1552–1567.

Hajigholizadeh, M. & Melesse, A. M. 2017 Assortment and spatiotemporal analysis of surface water quality using cluster and discriminant analyses. *Catena* **151**, 247–258.

IS:10500 2012 *Indian Standard Drinking Water–Specification (Second Revision)*. Bureau of Indian Standards (BIS), New Delhi.

Iscen, C. F., Emiroglu, Ö., Ilhan, S., Arslan, N., Yilmaz, V. & Ahiska, S. 2008 Application of multivariate statistical techniques in the assessment of surface water quality in Uluabat Lake, Turkey. *Environmental Monitoring and Assessment* **144**(1–3), 269–276.

Jha, D. K., Vinithkumar, N., Sahu, B. K., Das, A. K., Dheenan, P., Venkateshwaran, P., Begum, M., Ganesh, T., Devi, M. P. & Kirubagaran, R. 2014 Multivariate statistical approach to identify significant sources influencing the physico-chemical variables in Aerial Bay, North Andaman, India. *Marine Pollution Bulletin* **85**(1), 261–267.

Kaiser, H. F. 1974 An index of factorial simplicity. *Psychometrika* **39**(1), 31–36.

Kumar, M., Ramanathan, A., Tripathi, R., Farswan, S., Kumar, D. & Bhattacharya, P. 2017 A study of trace element contamination using multivariate statistical techniques and health risk assessment in groundwater of Chhaprola Industrial Area, Gautam Buddha Nagar, Uttar Pradesh, India. *Chemosphere* **166**, 135–145.

Li, H., Hopke, P. K., Liu, X., Du, X. & Li, F. 2015 Application of positive matrix factorization to source apportionment of surface water quality of the Daliao River basin, northeast China. *Environmental Monitoring and Assessment* **187**(3), 80.

Liu, Y., Wang, S., Lohmann, R., Yu, N., Zhang, C., Gao, Y., Zhao, J. & Ma, L. 2015 Source apportionment of gaseous and particulate PAHs from traffic emission using tunnel measurements in Shanghai, China. *Atmospheric Environment* **107**, 129–136.

Machiwal, D. & Jha, M. K. 2015 Identifying sources of groundwater contamination in a hard-rock aquifer system using multivariate statistical analyses and GIS-based geostatistical modeling techniques. *Journal of Hydrology: Regional Studies* **4**, 80–110.

Water Practice & Technology Vol 15 No 4
doi: 10.2166/wpt.2020.078

McCartney, M., Drechsel, P., Karg, H., Karg, H., Drechsel, P., Drechsel, P., Rao, K. C. & Gebrezgabher, S. 2018 Water quantity and hydrology. In: *Freshwater Ecology and Conservation: Approaches and Techniques* (Hughes, J. ed)) p. 67. Oxford University Press, Oxford, UK.

MoEF 2008 *Report on Visit to Deepor Beel in Assam: A Wetland Included Under National Wetland Conservation Management Programme of the Ministry of Environment and Forests*. Govt. of India.

Mustaffa, N. I. H., Latif, M. T., Ali, M. M. & Khan, M. F. 2014 Source apportionment of surfactants in marine aerosols at different locations along the Malacca Straits. *Environmental Science and Pollution Research* **21**(10), 6590–6602.

Paatero, P. & Tapper, U. 1994 Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics* **5**(2), 111–126.

Sahu, B. K., Begum, M., Khadanga, M., Jha, D. K., Vinithkumar, N. & Kirubagaran, R. 2013 Evaluation of significant sources influencing the variation of physico-chemical parameters in Port Blair Bay, South Andaman, India by using multivariate statistics. *Marine Pollution Bulletin* **66**(1–2), 246–251.

Shi, P., Zhang, Y., Li, Z., Li, P. & Xu, G. 2017 Influence of land use and land cover patterns on seasonal water quality at multi-spatial scales. *Catena* **151**, 182–190.

Shrestha, S. & Kazama, F. 2007 Assessment of surface water quality using multivariate statistical techniques: a case study of the Fuji river basin, Japan. *Environmental Modelling & Software* **22**(4), 464–475.

Simeonov, V., Stratis, J., Samara, C., Zachariadis, G., Voutsa, D., Anthemidis, A., Sofoniou, M. & Kouimtzis, T. 2003 Assessment of the surface water quality in Northern Greece. *Water Research* **37**(17), 4119–4124.

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 Research* **38**(18), 3980–3992.

Singh, K. R., Goswami, A. P., Kalamdhad, A. S. & Kumar, B. 2019 Assessment of surface water quality of Pagladia, Beki and Kolong river (Assam, India) using multivariate statistical techniques. *International Journal of River Basin Management*. (just-accepted), 1–43. https://doi.org/10.1080/15715124.2019.1566236

Vega, M., Pardo, R., Barrado, E. & Debán, L. 1998 Assessment of seasonal and polluting effects on the quality of river water by exploratory data analysis. *Water Research* **32**(12), 3581–3592.

Wilks, D. S. 2011 *Statistical Methods in the Atmospheric Sciences*. Academic Press, San Diego, CA, USA.

Wunderlin, D. A., del Pilar, D. M., Amé, M. V., Pesce, S. F., Hued, A. C. & de los Ángeles, B. M. 2001 Pattern recognition techniques for the evaluation of spatial and temporal variations in water quality. A case study: Suquía River Basin (Córdoba–Argentina). *Water Research* **35**(12), 2881–2894.

Zhao, J., Xu, Z., Liu, X. & Niu, C. 2013 Source apportionment in the Liao River basin. *China Environmental Science* **33**(5), 838–842.