International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

# Assessment and Studies on Physiochemical and Biological Parameter of Ganga River at local Barrage using Environmental and Multivariate Statistical Techniques

**Kshitij Upadhyay**

Department of Civil Engineering, Rajkiya Engineering College, Chandpur, Bijnor-246725, India

**Abstract:** ***Water quality assessments are an essential procedure in monitoring programs and are used to collect baseline environmental data. They are particularly important in developing regions where people often cannot access adequate supplies of water and effective water resource management is critical for future development. Here multivariate statistical techniques such as cluster analysis (CA), Principal component analysis(PCA), Discriminant analysis(DA) and Factor analysis(FA) were applied for the temporal and spatial variation and the interpretation of large complex water quality data set to predict variation in the various physiochemical properties and their remedies. Primary objective of cluster analysis (CA) is to assemble the object based on their characteristics they possess. Principal component analysis (PCA) was used to investigate the origin of each water quality parameter in the GANGA Basin and identified the major factor affecting the water quality. The major variations are related to the anthropogenic activities such as irrigation variation, construction activities, clearing of land and domestic waste disposal and natural processes such as erosion of river bank and runoff. Discriminant analysis (DA) was applied to the dataset to maximise the similarities between groups relative to within group variance of the parameter. DA provides better result with great discriminatory ability. Thus, this study illustrates the usefulness of multivariate statistical techniques for analysis and interpretation of complex data sets, and in water quality assessment, identification of pollution sources/factors and understanding temporal/spatial variations in water quality for effective river water quality management.***

**Keywords:** Ganga river, physiochemical parameter, cluster analysis, factor analysis, principal component analysis, discriminate analysis

## 1. Introduction

Ganga is the India's largest river basin; it covers 26 percent of landmass and support 43 percent of its population. There are 29 major cities, 70 towns and 1000's of villages located along the Ganges basin. There are 900 persons per km in Ganga's basin. This intense human population has a negative impact on the Ganga's basin. This is the major cause of Ganga's pollution.

Our major concern on the problem areas, that needs to be addressed in order to find a comprehensive solution to GANGA river pollution.

1) The inadequate flow of water in the river, needed to dilute the assimilate waste.
2) The growing quantum of untreated sewage discharged from cities along river.
3) The lack of enforcement against point source pollution from industries discharging waste into the river.

## 2. Study Area

The present study will be carried out in Bijnor district to evaluate water quality of river Ganga. There are two sites upstream and downstream. One location is near Railway bridge at BalawaliGhat, Bijnor located at $29^0$27'N $78^0$27'E / $29.45^0$N $78.45^0$E. It has an average elevation of 282 meters (925ft). Other location is near vill.-RasoolpurBhawar, Amroha located $28^0$54'15.95"N $78^0$28'3.10"E.


**Figure 1:** Bijnor City River Map

## 3. Methods

River water quality data sets were subjected to four multivariate techniques: cluster analysis (CA), principal component analysis (PCA), factor analysis (FA)and discriminant analysis (DA) (Wunderlin et al.,2001; Simeonov et al., 2003;Singh etal., 2004, 2005). DAwas applied to raw data, whereas PCA, FA and CAwere applied to experimental data, standardized through z-scale transformation to avoid misclassifications arising from the different orders of magnitude of both numerical values and variance of the parameters analyzed (Liu et al., 2003; Simeonovet al.,2003). All mathematical and statistical computations were madeusing Microsoft Office Excel 2010.

**Cluster Analysis**

Cluster analysis is a multivariate method which aims to classify a sample of data on the basis of the characteristics they possess. In cluster analysis the datasets are divided into no of groups in which the data are homogeneous within groups but heterogeneous to other groups.Cluster analysis includes a broad suite of techniques designed to find groups of similar itemswithin a data set. Hierarchical agglomerative clustering is the most common approach, which provides intuitive similarity relationships between any one sample and the entire data set, and is typically illustrated by a Dendrogram (tree diagram). The spatial variability of water quality in the whole river basin was determined from CA, using the linkage distance, reported as $D_{centriodal}$ which represents the centroidal distance of any two clusters.

**Important About Clustering**

1) There is no objective function here.
2) There is no dependent variable. Its just an implementation of input data for natural grouping.
3) Cluster analysis has no mechanism for differentiating between relevant and irrelevant variables.

**Principal Component Analysis**

Principal component analysis is a data reduction technique. It is introduced by Pearson (1901) and Hotelling H in 1933. It consist two basis linear dimension and orthogonality of the new dimension (PCS).In a typical PCA however, there are more than two variables i.e. more than two dimensions. If the second principal component will be both perpendiculars to the first, and along the line of second next greatest variation, the third principal component will be along the line of the following greatest variation and perpendicular to the first two principal components. The same applies to the N dimensions under analysis. With several variables the computation is more complicated but the basic principle to express two or more variables by a single factor remains the same. By multiplying the original data-set by the principal components, the data is rotated so that the components form the new perpendicular axes and the objects lying exactly on the axes have now only one coordinate, i.e. are captured by one variable only.

**Principal component analysis (PCA)**

1) Reduce the dimensionality of a data set by finding a new set of variables, smaller than the original set of variables
2) Retains most of the sample's information.
3) Useful for the compression and classification of data.

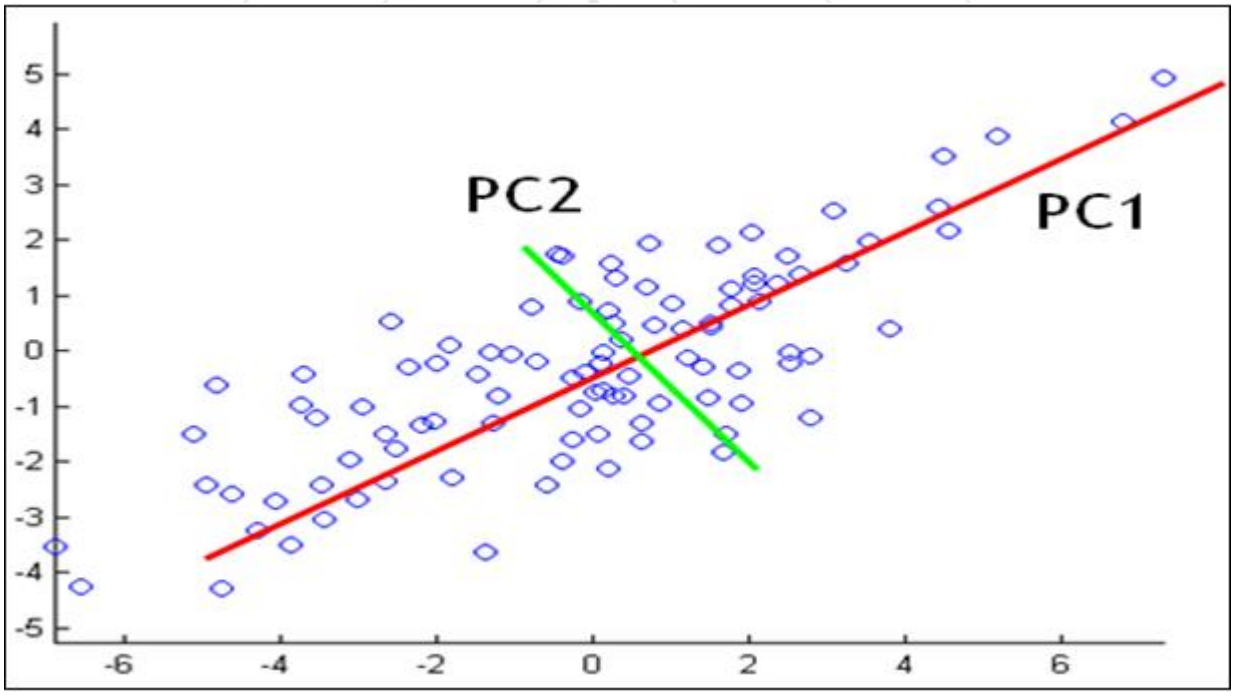


**Figure 2:** Discrimnent Analysis

**Discrimnent Analysis**

Discriminant analysis is a technique for analyzing data when the criterion or dependent variable is categorical and the predictor or independent variables are interval in nature.

The Discriminant Analysisprocedure is designed to help distinguish between two or more groups of data based on a set of *p* observed quantitative variables. It does so by constructing discriminant functionsthat are linear combinations of the variables. The objective of such an analysis is usually one or both of the following:-

1) To be able to describe observed cases mathematically in a manner that separates them into groups as well as possible.
2) To be able to classify new observations as belonging to one or another of the groups.
3) Development of discriminant functions, or linear combinations of the predictor or independent variables, which will best discriminate between the categories of the criterion or dependent variable (groups).
4) Examination of whether significant differences exist among the groups, in terms of the predictor variables.

5) Determination of which predictor variables contribute to most of the intergroup differences.
6) Classification of cases to one of the groups based on the values of the predictor variables

**Factor Analysis**

Factor analysis is a multivariate method used for data reduction and gives the correlation between factor and variable. The purpose of factor analysis is to describe, if possible the covariance in term of a under laying, but unobservable random qualities called factors. Factor analysis attempts to represent a set of observed variables $X_1$, $X_2$ …. $X_n$in terms of numbers of common factors plus a factor which is unique to each variable. The common factors are hypothetical variables which explain why a number of variables are correlated with each other,it is because they have one or more factors in common.

**Assumption**

Factor analysis is designed for interval data, although it can also be used for ordinal data. The variables used in factor analysis should be linearly related to each other. This can be checked by looking at scatter plots of pairs of variables. Obviously the variables must also be at least moderately correlated to each other, otherwise the number of factors will be almost the same as the number of original variables, which means that carrying out a factor analysis would be pointless.

**Factor analysis model**

If the observed variables are $X_1$, $X_2$…. $Xn$, the common factors are $F_1$, $F_2$… $F_m$and the unique factors are $U_1$, $U_2$ …$U_n$, the variables may be expressed as linear functions of the factors:

$X_1 = a_{11}F_1 + a_{12}F_2 + a_{13}F_3 + \ldots + a_{1m}F_m + a_1U_1$
$X_2 = a_{21}F_1 + a_{22}F_2 + a_{23}F_3 + \ldots + a_{2m}F_m + a_2U_2$
….
$X_n = a_{n1}F_1 + a_{n2}F_2 + a_{n3}F_3 + \ldots + a_{nm}F_m + a_nU_n.$

When the coefficients are correlations, i.e., when the factors are uncorrelated, the sum of the squares of the loadings for variable $X_1$, namely $a_{11}^2 + a_{12}^2 + \ldots + a_{13}^2$, shows the proportion of the variance of variable *X1* which is accounted for by the common factors. This is called the communality.

## 4. Result and Discussion

**Cluster Analysis**

It involves the plotting of data points on graph. In this plotting two physiochemical properties are located on two axes and data related to these properties are plotted. After the plotting of data for various physiochemical properties, formed clusters are as follows:

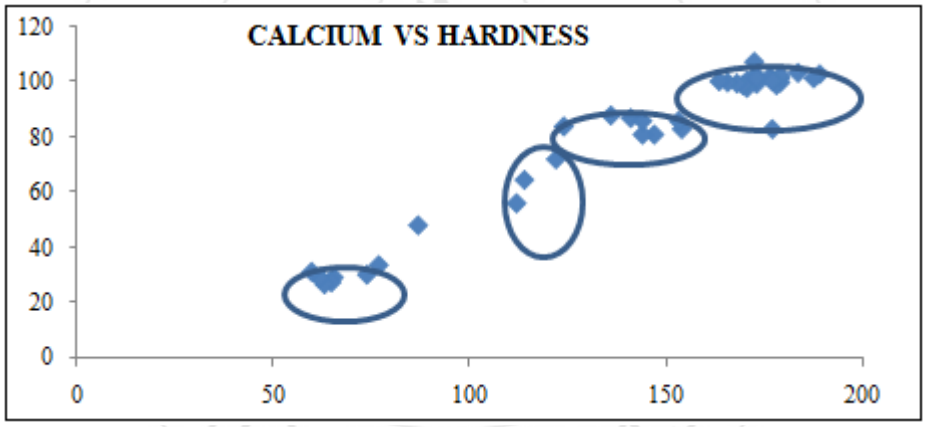


**Figure 3:** Cluster Representation

Using the centroidal distance measurement, the co-ordinate of the points representing the centroid of clusters is
$C_1$ (65.6, 29), $C_2$ (114, 64.5), $C_3$ (144, 86), $C_4$ (172.7, 102.7)

Now the distance matrix formed is

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | 0 | | | |
| 2 | 60.5 | 0 | | |
| 3 | 97.4 | 36.9 | 0 | |
| 4 | 130.5 | 69.45 | 33.2 | 0 |

Since the distance between the cluster 3 and 4 is minimum. Hence these two are combined to form a new cluster and again the distances are measured.

| | 1 | 2 | |
|---|---|---|---|
| (3, 4) | 0 | | |
| 1 | 97.4 | 0 | |
| 2 | 36.9 | 60 | 0 |

Again combining the (3, 4) and 2 and forming the distance matrix

| | 1 | |
|---|---|---|
| (2, 3, 4) | 0 | |
| 1 | 60 | 0 |

Using these distance matrixes the **DENDROGRAM TREE** representation of the chosen physiochemical properties is

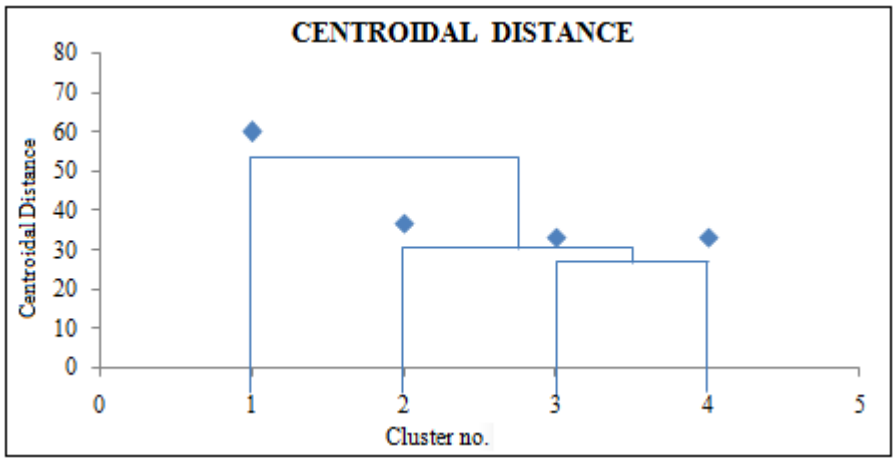


**Figure 4:** Dendrogram Tree Representation

**Factor Analysis**

Correlation between variable and factors are given by linear regression equation

$$X - \mu = \Lambda F + \delta$$

This equation is known as factor model.

So the variation between variable and factor are given by following example.

Here we have certain data of physiochemical property of water are given below.

| Sr .no. | $P^H$ | T.D.S | D.O | B.O.D | Hardness | calcium | Magnesium | Chloride |
|---|---|---|---|---|---|---|---|---|
| 1 | 7.5 | 410.0 | 6.8 | 1.0 | 189.0 | 103.0 | 86.0 | 12.0 |
| 2 | 7.5 | 420.0 | 6.6 | 1.0 | 179.0 | 100.0 | 79.0 | 11.0 |
| 3 | 7.6 | 450.0 | 7.2 | 1.33 | 172.7 | 102.7 | 70.0 | 14.7 |
| 4 | 7.6 | 437.5 | 7.3 | 1.25 | 178.0 | 99.0 | 79.0 | 13.0 |
| 5 | 7.6 | 434.0 | 7.4 | .96 | 176.4 | 102.0 | 76.0 | 12.4 |

These data can be normalising in to following manner. So for normalisation firstly we calculate mean of each parameter and then subtract mean from each parameter. Then adding square of this and calculate deviation.

So deviation $=\sqrt{(\Sigma x-\mu)^2}$

So normalisation = $( x-\mu ) / \sqrt{(\Sigma x-\mu)^2}$

If the variable are denoted by $X_1, X_2, X_3, X_4 \ldots\ldots X_N$.

So normalise data of variable are given below in to the table.

| | | | | | |
|---|---|---|---|---|---|
| $X_1$ | -0.0025 | -0.0056 | 0.0018 | 0.00518 | 0.0079 |
| $X_2$ | -0.843 | -0.974 | 0.882 | 0.932 | 0.737 |
| $X_3$ | -0.010 | -0.0435 | 0.0062 | 0.0310 | 0.067 |
| $X_4$ | -0.0045 | -0.0102 | 0.0099 | 0.0183 | -0.0294 |
| $X_5$ | 0.414 | -0.00189 | -0.283 | -0.132 | -0.522 |
| $X_6$ | 0.069 | -0.126 | 0.0609 | -0.303 | 0.131 |
| $X_7$ | 0.332 | 0.0946 | -0.358 | 0.129 | 0.398 |
| $X_8$ | -0.0257 | -0.152 | 0.0931 | 0.049 | -0.438 |

Now we calculate Eigen value or loading factor for different factor for made a scree plot between Eigen value and factor. Which give the information that how many variation occur in the physiochemical parameter of water from different factors.

| Parameter | Factor.1 | Factor.2 | Factor.3 | Factor.4 | Factor.5 | communalities |
|---|---|---|---|---|---|---|
| $X_1$ | -0.0038 | -0.0070 | 0.0046 | 0.0038 | 0.0065 | 0.000142 |
| $X_2$ | -0.99 | -1.121 | 0.735 | 0.785 | 0.59 | 3.741 |
| $X_3$ | -0.02 | -0.053 | -0.0037 | 0.021 | 0.057 | 0.0069 |
| $X_4$ | -0.0013 | -0.0070 | 0.013 | 0.0215 | -0.026 | 0.486 |
| $X_5$ | 0.519 | 0.1031 | -0.178 | -0.027 | -0.417 | 0.486 |
| $X_6$ | 0.102 | -0.093 | 0.094 | -0.27 | 0.164 | 0.1276 |
| $X_7$ | 0.212 | -0.025 | -0.478 | -0.009 | 0.278 | 0.351 |
| $X_8$ | 0.068 | -0.058 | 0.187 | 0.143 | -0.344 | 0.1817 |
| Eigen value | 1.312 | 1.28 | .84 | .704 | .748 | 4.896 |

Now the table of component and Eigen value are given below.

| Component | Eigen value | % variance | Cumulative % |
|---|---|---|---|
| 1 | 1.312 | .26 | .26 |
| 2 | 1.28 | .25 | .51 |
| 3 | 0.84 | .17 | .68 |
| 4 | 0.748 | .15 | .83 |
| 5 | 0.704 | .14 | .97 |

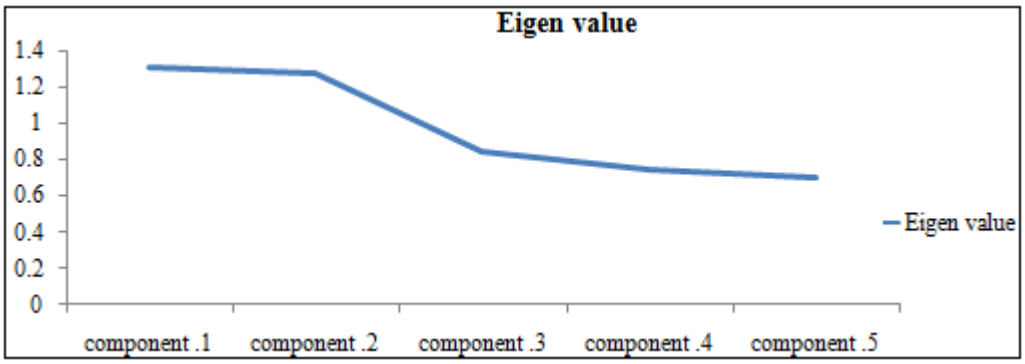


**Figure 5:** Scree Plot

This plot shows that how many variations occur in physiochemical parameter of water from different component.

Data values set of pH and DO.
(7.2, 7.6) , (7.1, 7.7) , (7.1, 8.1) , (7.2, 8.1) , (7.3, 8.4) , (7.2, 8.1) , (7.2, 8.0) , (7.2, 7.4)
Plotting the data values on the coordinate axis.

**Principal component analysis**
**1** Taking the two properties $P^H$ and DO for comparing there variation with time.

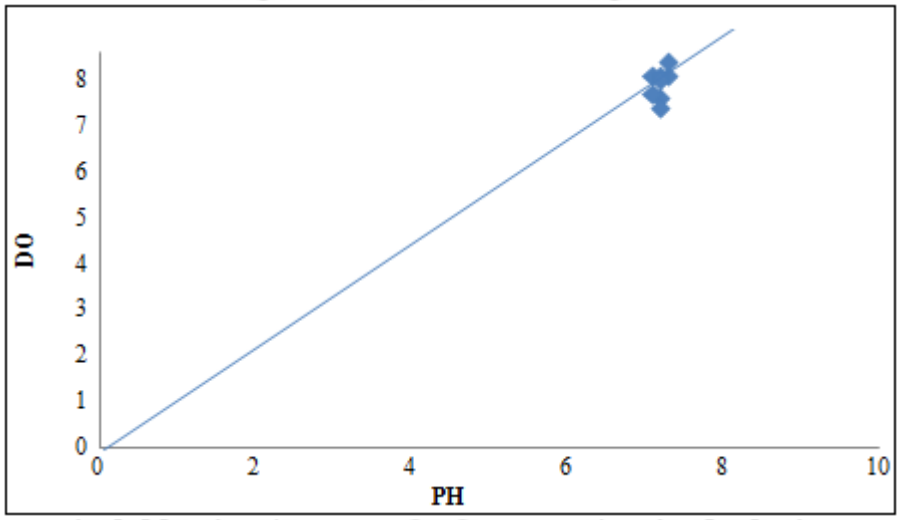


**Figure 6:** Scatter plot of PH and DO


**Figure 7**

First we need to calculate Covariance matrix
$Co.V = 1/n \sum(x-\mu)(x-\mu)^T$
Where x is a feature vector which is two dimensional, μ is the average (mean)

**Projection of data about principal component 1**
(7.2, 7.6) , (7.1, 7.7) , (7.1, 8.1) , (7.2, 8.1) , (7.3, 8.4) , (7.2, 8.1) , (7.2, 8.0) , (7.2, 7.4)

Mean

$$\begin{matrix} X_1 \\ X_2 \end{matrix} \begin{pmatrix} 7.2 \\ 7.6 \end{pmatrix} \begin{pmatrix} 7.1 \\ 7.7 \end{pmatrix} \begin{pmatrix} 7.1 \\ 8.1 \end{pmatrix} \begin{pmatrix} 7.2 \\ 8.1 \end{pmatrix} \begin{pmatrix} 7.3 \\ 8.4 \end{pmatrix} \begin{pmatrix} 7.2 \\ 8.1 \end{pmatrix} \begin{pmatrix} 7.2 \\ 8.0 \end{pmatrix} \begin{pmatrix} 7.2 \\ 7.4 \end{pmatrix} = \begin{pmatrix} 7.1875 \\ 7.925 \end{pmatrix}$$

Variance of corresponding data (x-μ)

$$\begin{pmatrix} .0125 \\ -0.325 \end{pmatrix} \begin{pmatrix} -0.0875 \\ -0.225 \end{pmatrix} \begin{pmatrix} -0.0875 \\ -0.175 \end{pmatrix} \begin{pmatrix} 0.0125 \\ 0.175 \end{pmatrix} \begin{pmatrix} 0.1125 \\ 0.475 \end{pmatrix} \begin{pmatrix} 0.0125 \\ 0.175 \end{pmatrix} \begin{pmatrix} 0.0125 \\ 0.075 \end{pmatrix} \begin{pmatrix} 0.0125 \\ -0.525 \end{pmatrix}$$

Finding $(x-\mu)(x-\mu)^T$

$$\begin{pmatrix}0.0125\\-0.325\end{pmatrix}\begin{pmatrix}0.0125,-0.325\end{pmatrix}=\begin{pmatrix}1.5625\times10^{-4} & -4.0625\times10^{-3}\\-4.0625\times10^{-3} & 0.105625\end{pmatrix}$$

$$\begin{pmatrix}-0.0875\\-0.225\end{pmatrix}\begin{pmatrix}-0.0875,-0225\end{pmatrix}=\begin{pmatrix}7.65625\times10^{-3} & 0.0196875\\0.0196875 & 0.050625\end{pmatrix}$$

$$\begin{pmatrix}-0.0875\\0.175\end{pmatrix}\begin{pmatrix}-0.0875,0.175\end{pmatrix}=\begin{pmatrix}7.65625\times10^{-3} & -0.0153125\\-0.0153125 & 0.030625\end{pmatrix}$$

$$\begin{pmatrix}0.0125\\0.175\end{pmatrix}\begin{pmatrix}0.0125,0.175\end{pmatrix}=\begin{pmatrix}1.5625\times10^{-4} & 2.1875\times10^{-3}\\2.1875\times10^{-3} & 0.030625\end{pmatrix}$$

$$\begin{pmatrix}0.1125\\0.475\end{pmatrix}\begin{pmatrix}0.1125,0.475\end{pmatrix}=\begin{pmatrix}0.012656 & 0.0534375\\0.0534375 & 0.225625\end{pmatrix}$$

$$\begin{pmatrix}0.0125\\0.175\end{pmatrix}\begin{pmatrix}0.0125,0.175\end{pmatrix}=\begin{pmatrix}1.5625\times10^{-4} & 2.1875\times10^{-3}\\2.1875\times10^{-3} & 0.030625\end{pmatrix}$$

$$\begin{pmatrix}0.0125\\0.175\end{pmatrix}\begin{pmatrix}0.0125,0.175\end{pmatrix}=\begin{pmatrix}1.5625\times10^{-4} & 2.1875\times10^{-3}\\2.1875\times10^{-3} & 0.030625\end{pmatrix}$$

$$\begin{pmatrix}0.0125\\-0.525\end{pmatrix}\begin{pmatrix}0.0125,-0.525\end{pmatrix}=\begin{pmatrix}1.5625\times10^{-4} & -6.5625\times10^{-3}\\-6.5625\times10^{-4} & 0.275625\end{pmatrix}$$

$$\text{CoV.X}=\begin{pmatrix}3.59371875\times10^{-3} & 0.390625\times10^{-4}\\0.390625\times10^{-4} & 0.0975\end{pmatrix}$$

Eigens values corresponding to this covariance matrix.
$|A-\lambda I|=0$

$(3.59371875\times10^{-3}-\lambda)(0.0975-\lambda)=0.15258789\times10^{-4}$
$\lambda_1=.09766$
$\lambda_2=0.0034315$

Eigen vectors corresponding to these Eigen values

$AX=\lambda IX$

Vectors are $\begin{pmatrix}1\\234.81\end{pmatrix}$ and $\begin{pmatrix}1\\4.1528\end{pmatrix}$

Finding,
Trace(S) of these eigen values.
TRACE (S) = $\sum\lambda = 0.1010915$

So, Total variance accounted by first principal component = $\lambda_1$/ trace(s)
= 0.966
Total variance accounted by second principal component = $\lambda_2$/trace(s)
= .034

So **96.6%** variance accounted only by first principal component and other **3.4%** accounted by second principal component.

It is clear that all the variance can be explained only by first principal component and second may neglect.

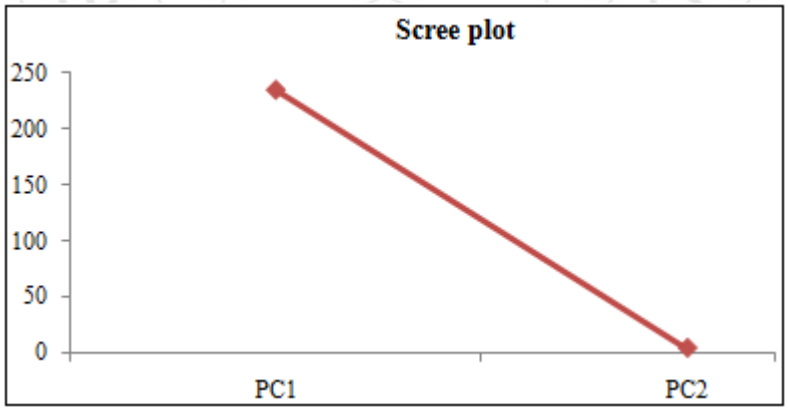


**Figure 8:** Scree plot of principal components

**Discriminant Analysis**

Projection direction that maintain seperaility between two classes:
n -no of d-dimensional feature vectors
$x_1, x_2, x_3 - - - - - - - - x_n$
and these feature vector partition into two different sets.

**We take two parameters i.e.$P^H$ and DO.**

Sample for class $\omega_1$ $\begin{pmatrix}7.2 & 7.3 & 7.2 & 7.2 & 7.2 & 7.2\\8.1 & 8.4 & 8.1 & 8.0 & 7.4 & 7.4\end{pmatrix}$

Sample for class $\omega_2$ $\begin{pmatrix}7.2 & 7.4 & 7.4 & 7.1 & 7.1 & 7.2\\7.9 & 6.7 & 7.2 & 6.9 & 7.4 & 8.4\end{pmatrix}$

Firstly we started with the projection direction for best representation, and then we have to find out eigen vectors of the covariance matrix of data elements.\

**Calculation of covariance matrix**

$$\begin{matrix} & & & & & & & & & & & & \text{Mean}(\mu) \\ \begin{pmatrix} 7.2 & 7.3 & 7.2 & 7.2 & 7.2 & 7.2 & 7.2 & 7.4 & 7.4 & 7.1 & 7.1 & 7.2 \\ 8.1 & 8.4 & 8.1 & 8.0 & 7.4 & 7.4 & 7.9 & 6.7 & 7.2 & 6.9 & 7.4 & 8.4 \end{pmatrix} & & & & & & & & & & & & \begin{pmatrix} 7.225 \\ 7.658 \end{pmatrix} \end{matrix}$$

$$(x-\mu) = \begin{pmatrix} -0.025 & 0.075 & -0.025 & -0.025 & -0.025 & -0.025 & -0.025 & 0.175 & 0.175 & -0.125 & -0.125 & -0.025 \\ 0.442 & 0.742 & 0.442 & 0.342 & -0.0258 & -0.0258 & 0.242 & -0.958 & -0.458 & -0.758 & -0.258 & 0.742 \end{pmatrix}$$

$$M_1 = \begin{pmatrix} -0.025 \\ 0.442 \end{pmatrix} \begin{pmatrix} -0.025 & 0.442 \end{pmatrix}$$

$$M_2 = \begin{pmatrix} 0.075 \\ 0.742 \end{pmatrix} \begin{pmatrix} 0.075 & 0.742 \end{pmatrix}$$

$$M_3 = \begin{pmatrix} -0.025 \\ 0.442 \end{pmatrix} \begin{pmatrix} -0.025 & 0.442 \end{pmatrix}$$

$$M_4 = \begin{pmatrix} -0.025 \\ 0.342 \end{pmatrix} \begin{pmatrix} -0.025 & 0.342 \end{pmatrix}$$

$$M_5 = \begin{pmatrix} -0.025 \\ -0.258 \end{pmatrix} \begin{pmatrix} -0.025 & -.0.258 \end{pmatrix}$$

$$M_6 = \begin{pmatrix} -0.025 \\ -0.025 \end{pmatrix} \begin{pmatrix} -0.025 & -0.0258 \end{pmatrix}$$

$$M_7 = \begin{pmatrix} -0.025 \\ 0.242 \end{pmatrix} \begin{pmatrix} -0.025 & 0.242 \end{pmatrix}$$

$$M_8 = \begin{pmatrix} 0.175 \\ -0.958 \end{pmatrix} \begin{pmatrix} 0.175 & -0.958 \end{pmatrix}$$

$$M_9 = \begin{pmatrix} 0.175 \\ -0.458 \end{pmatrix} \begin{pmatrix} 0.175 & -0.458 \end{pmatrix}$$

$$M_{10} = \begin{pmatrix} -0.125 \\ -0.758 \end{pmatrix} \begin{pmatrix} -0.125 & -0.758 \end{pmatrix}$$

$$M_{11} = \begin{pmatrix} -0.125 \\ -0.258 \end{pmatrix} \begin{pmatrix} -0.125 & -0.258 \end{pmatrix}$$

$$M_{12} = \begin{pmatrix} 0.175 \\ -0.458 \end{pmatrix} \begin{pmatrix} 0.175 & -0.458 \end{pmatrix}$$

$$\text{Covariance matrix } (M_i) = \begin{pmatrix} 0.00723 & -0.00472 \\ -0.00472 & 0.31144 \end{pmatrix}$$

**Calculation of Eigen value ($\lambda$):**

$$\begin{pmatrix} 0.00723 - \lambda & -0.00472 \\ -0.00472 & 0.31144 - \lambda \end{pmatrix}$$

$$2.2516 - 0.31867\lambda + \lambda^2 = 0$$

$$\lambda_1 = 0.1593$$

$$\lambda_2 = 0.1593$$

**Calculation of Eigen vector:**

$AX = \lambda X$

Where $\lambda$ is Eigen value and X is Eigen vector.

$$\begin{pmatrix} -0.152105 & -0.00472 \\ -0.00472 & 0.01521 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = 0$$

After the calculation, the value of $X_1$ and $X_2$ is $\begin{pmatrix} 32 \\ 1 \end{pmatrix}$

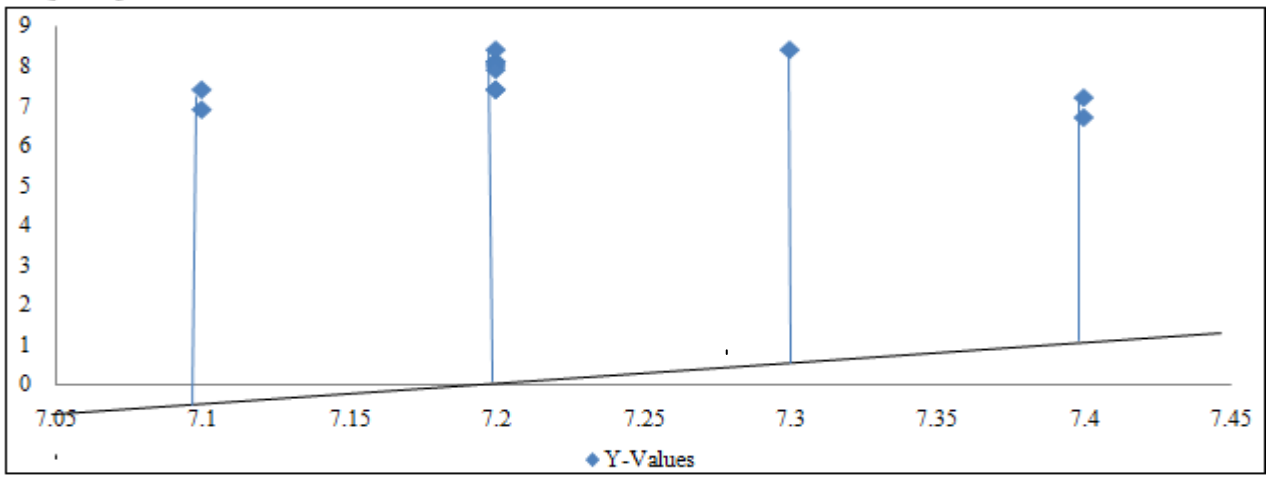


**Figure 9**

**Scatter plot of PH and DO**

**Scatter of projected data:**

$S_1$ = It is the within class scatter of the data points class $\omega_1$

$S_2$ = It is the within class scatter of the data points class $\omega_2$

$$S_w = S_1 + S_2$$

Where $S_w$ is the total within class scatter

$$W = S_W^{-1}(\mu_1 - \mu_2)$$

Where W is the direction of the projection or projection vector which maintains separability.

**The classes mean are:**

$\mu_1$

$$\begin{pmatrix} 7.2 & 7.3 & 7.2 & 7.2 & 7.2 & 7.2 \\ 8.1 & 8.4 & 8.1 & 8.0 & 7.4 & 7.4 \end{pmatrix} \quad \begin{pmatrix} 7.216 \\ 7.9 \end{pmatrix}$$

$\mu_2$

$$\begin{pmatrix} 7.2 & 7.4 & 7.4 & 7.1 & 7.1 & 7.2 \\ 7.9 & 6.7 & 7.2 & 6.9 & 7.4 & 8.4 \end{pmatrix} \quad \begin{pmatrix} 7.23 \\ 7.416 \end{pmatrix}$$

**Covariance matrix of the first class:**

$$(x - \mu_1) = \begin{pmatrix} -0.016 & 0.084 & -0.016 & -0.016 & -0.016 & -0.016 \\ 0.2 & 0.5 & 0.2 & 0.1 & -0.5 & -0.5 \end{pmatrix}$$

$$S_1 = \sum (x - \mu_1)(x - \mu_1)^t = \begin{pmatrix} 0.008335 & 0.0516 \\ 0.0516 & 0.84 \end{pmatrix}$$

**Covariance matrix of the second class:**

$$(x - \mu_2) = \begin{pmatrix} -0.03 & 0.17 & 0.17 & -0.13 & -0.13 & -0.13 \\ 0.484 & -0.716 & -0.216 & -0.516 & -0.016 & 0.984 \end{pmatrix}$$

$$S_2 = \sum (x - \mu_2)(x - \mu_2)^t = \begin{pmatrix} 0.1094 & -0.2318 \\ -0.2318 & 2.028 \end{pmatrix}$$

**Within class scatter matrix:**

$$S_w = S_1 + S_2$$

$$S_W = \begin{pmatrix} 0.1177 & -0.1802 \\ -0.1802 & -1.188 \end{pmatrix}$$

$$W = S_W^{-1}(\mu_1 - \mu_2)$$

$$S_W^{-1} = \begin{pmatrix} 11.071 & -1.6794 \\ -1.6794 & -1.0969 \end{pmatrix}$$

$$(\mu_1 - \mu_2) = \begin{pmatrix} -0.014 \\ 0.484 \end{pmatrix}$$

Projection direction is simply given by:

$$e = S_W^{-1}(\mu_1 - \mu_2) = \begin{pmatrix} -0.9676 \\ -0.50 \end{pmatrix}$$

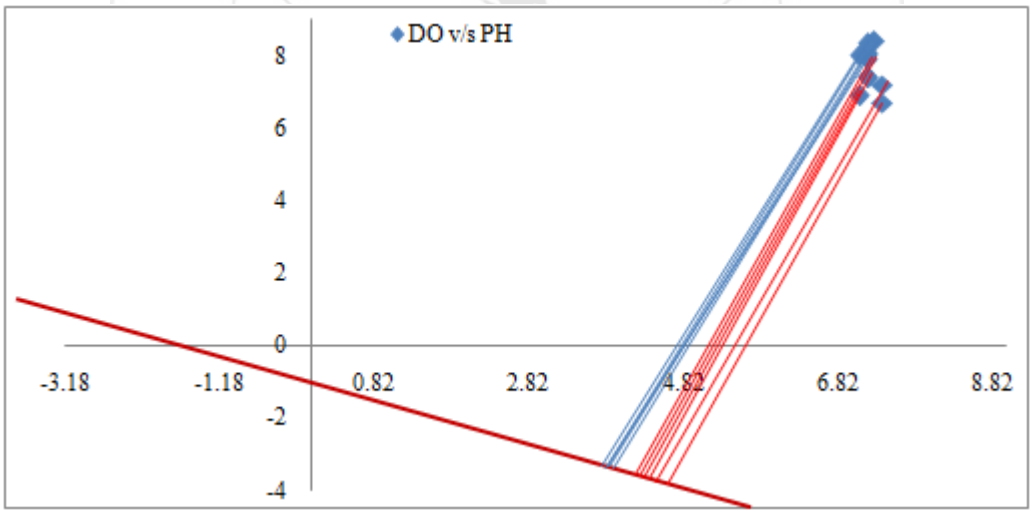


**Figure 10:** Projection about vector having high Eigen values

## 5. Conclusion and Result Discussion

**Cluster analysis**

- In cluster analysis the plotted **DENDROGRAM TREE** represents the conclusion drawn from the analysis.
- **A DENDROGRAM** is a branching diagram that represents the relationship of similarity among a group of data.
- The height of the branch points indicates how similar or different they are from each other: the greater the height, the greater the difference.
- In the **figure no. 4DENDROGRAM TREE** is plotted between the clusters formed by calcium and hardness, and the no of clusters.
- Is shows that the dissimilarity between the cluster 3 and 4 is less than the combined cluster (3, 4)and 2 and similarly the difference between cluster 1 and (2, 3, 4) is maximum.
- This dissimilarity shows that the monthly variation of calcium and hardness data in cluster 3 and 4 is less as compared to variation in (3, 4) and 2 and so on.
- In **DENDROGRAM TREE** representation only vertical height has significance. Horizontal distance does not represent anything, but only the cluster number.

**Principal Component**

The first principal component is strongly correlated with five of the original variables. The first principal component increases with increasing parameters. This component viewed as measure of the quality of parameters.The second principal component increases with increasing parameter variability. This suggests that places with one parameter high also tend to have better recreation of other.

It tends to conclude following from output:

1) The proportion of variance indicates how much of total variance is there in variance of a particular principal component. Hence PC1 variability explkain more than 75% of total variance of the data.
2) Considering rotations of PC, one can conclude parameters are directly related.
3) To show all rotation in one graph, one can show their relative contribution to total variation by multiplying each rotation by proportion of variance of that principal component.

**Discriminant analysis**

The discriminant analysis provides discriminant result based on the selected discriminant method. Discriminant analysis report contains the following sections:

1) When one select the wide linear discriminant method, a principal component report appeas.
2) The canonical plot shows the points and multivariate means that best Separate the two groups.

The biplot axes are the first two canonical variables. These define the two dimensions that provide maximum separation among groups.

The observations and the multivariate means of each group are represented as points on the biplot. They are expressed in terms of the first two canonical variables. The set of rays that appears in the plot represents the covariates. The rays show how each covariate loads onto the first two standardized canonical variables. The direction of a ray indicates the degree of association of that covariate with the first two groups.

**Factor analysis**

A Factor analysis is conducted on different physiochemical characteristics of water of Ganga River and variationin these characteristics are shown by scree plot. A scree plot displays the Eigen values associated with a component or factor in descending order versus the number of the component or factor. This scree plot show that how many variation in the physiochemical property of Ganga river water occur by each factor. So from fig.no.(5) show that monthly wise variation in the physiochemical property of water.

## References

[1] APHA (2005) Standard methods for the examination of water and wastewater. American Water Works Association, Environment Federation, Washington, DC
[2] Assessment of surface water quality using multivariate statistical techniques: A case study of the Fuji river basin, Japan S. Shrestha and F. Kazama (22 March 2006)
[3] Studies on distribution and fractionation of heavy metals in Gomti river sediments—a tributary of the Ganges, India Kunwar P. Singh, Dinesh Mohan, Vinod K. Singh and Amrita Malik (31 January 2005)
[4] Uttar Pradesh pollution control board manual and data, regional office, Bijnor
[5] River water quality assessment using environmentric techniques: case study of Jakara River Basin, Adamu Mustapha & Ahmad ZaharinAris&HafizanJuahir& Mohammad FiruzRamli&Nura Umar Kura (3 February 2013)
[6] Assessment of the surface water quality in Northern Greece, V. Simeonov[a], J.A. Stratis[b], C. Samara[c], G. Zachariadis[b], D. Voutsa[c], A. Anthemidis, M. Sofoniou[b] and Th. Kouimtzis[c]
[7] Statistical and Structural Approaches to Texture by ROBERT M. HARALICK.
[8] Helena, B., Pardo, R., Vega, M., Barrado, E., Ferna´ndez, J.M., Ferna´ndez, L., 2000. Temporal evolution of groundwater composition in an alluvial aquifer (Pisuerga River, Spain) by principal component analysis. Water Research 34, 807e816
[9] Abdul-Wahab, S.A., Bakheit, C.S., Al-Alawi, S.M., 2005. Principal component and multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations. Environmental Modelling & Software 20 (10), 1263e1271.
[10] BIS, 1991. Drinking water specification IS: 10500:1991. Bureau of Indian Standard, New Delhi, India.
[11] Jha, P.K., Suramanian, V., Stasawad, R., Griekan, V.R., 1990. Heavy metals in sediments of the Yamuna River (A tributary of the Ganges) India. Sci. Total Environ. 95, 7–27.
[12] Khwaja, A.R., Singh, R., Tandon, S.N., 2000. Monitoring of Ganga water and sediments vis-a`-vis tannery pollution at Kanpur (India): a case study. Environ. Monit. Assess. 68, 19–35
[13] Saika, D.K., 1987. Studies on the sorption properties of bed sediments of river Ganges, transport of some heavy metal ions. PhD Thesis, University of Roorkee, Roorkee, India.
[14] Srivastava, S.K., Gupta, V.K., Anupam, Mohan, D., 1994. Status of some toxic heavy metal ions in the upper reaches of river Gnages, Indian. J. Chem. Soc. 71, 29–34.
[15] Subramanian, V., Van Grieken, R., van Dack, L., 1987. Heavy metal distribution in the sediments of Ganges and Brahamputra rivers. Environ. Geol. Water Sci. 9, 93–108.