Evaluation of Water Quality by Factor Analysis

Evaluation of Water Quality by Factor Analysis

By

N. Okendro Singh

Guidance

Dr. P.C. Mahanta

Evaluation of Water Quality by Factor Analysis

Published by : Dr. P.C. Mahanta
Director, NRC CWF, Bhimtal, Uttarakhand (INDIA)

CONTENTS

डा. एस. अय्यप्पन
उप महानिदेशक (मत्स्य)

Dr. S. AYYAPPAN
Deputy Director General (Fisheries)

भारतीय कृषि अनुसंधान परिषद्
कृषि अनुसंधान भवन - II
पूसा, नई दिल्ली 110 012
INDIAN COUNCIL OF AGRICULTURAL RESEARCH
KRISHI ANUSANDHAN BHAVAN-II
PUSA, NEW DELHI 110 012

Dated: 8th January, 2008

Foreword

Water quality is a primary factor in any kind of fishery. In the recent past, water quality of different water bodies, especially at high altitude region, have been drastically changed in their physico-chemical characteristics due to human activities and natural causes. Therefore, water quality monitoring program becomes an inseparable part of fish and fishery activities. Many researchers have collected water quality data of various water bodies from time to time however, information-generating systems followed in the above cases are usually lack of using those efficient tools of processing, analysis, etc. As a result such programs often fails to win managerial and social support. Thus, sophisticated analytical tools, recently developed, could be valuable attributes to analyze existing water quality data. Moreover, monitoring professions should adopt new data analysis, interpretation and reporting procedures to their existing monitoring systems, where the historical focus has been simply collecting data.

There is no doubt about the usefulness of this document "Evaluation of Water Quality by Factor Analysis" which covers many unusual features such as conceptualization of terms and description of factor analysis method in a simplest way. In addition to above, the technique involved is illustrated by an example. I hope this document will be of great analytical significance for the water quality assessment of various water bodies of upland. It is a matter of appreciation that NRCCWF has taken

डा. पी. सी. महन्ता
निदेशक

Dr. P.C. Mahanta
Director

भाकृअनुप
ICAR

राष्ट्रीय शीतजल मात्स्यिकी अनुसंधान केन्द्र
National Research Centre on Coldwater Fisheries

Dated: 7th January, 2008

Preface

The water quality of various water bodies located at high altitude has been deteriorating day-by-day due to various reasons. Even it is threatening the fishery sector of this region otherwise; proper management practices of various water bodies are taken up timely. Therefore, devolvement of a sound and valid water quality-monitoring plan is quite concerned for those water bodies in the present scenario. Of late, multivariate statistical technique is emerged to develop effective water quality monitoring planning that involves minimum resources. But, it is hardly found using the above multivariate technique in the analysis of water quality data particularly, in the coldwater sector of the country.

The basic purpose of this document is to provide a technique on analysis of water quality data with a suitable statistical tool. This document will help to develop a well-designed water quality-monitoring plan for various water bodies located at high altitude. It is hoped that the publication will be immensely useful for researchers who are working in the monitoring of water quality. The effort of Shri N. Okendro Singh in bringing out this publication is praiseworthy.

ACKNOWLEDGEMENTS

It is a great privilege to express my gratitude to Dr. P.C. Mahanta, Director, NRC on Coldwater Fisheries, Bhimtal for his keen interest, guidance and encouragement during the preparation of this bulletin.

I am indebted to Dr. Prakash Nautiyal, Reader, Department of Zoology, H.N.B. Garhwal University, Srinagar, Garhwal, Uttarakhand for critically examining the results and discussion part of the example given in this bulletin and offering constructive comments and helpful suggestions to improve upon the findings.

Special thanks are due to my senior colleagues Dr. Shyam Sunder, Principal Scientist, NRC CWF, Bhimtal and Dr. Madan Mohan, Principal Scientist, NRC CWF, Bhimtal for their guidance, help rendered without which it would be incomplete to prepare this bulletin.

Dated: January 2008

(N. Okendro Singh)

Evaluation of Water Quality by Factor Analysis

1. INTRODUCTION

Anthropogenic activities and many natural phenomena lead to exert a strong pressure on freshwater resources (e.g. riverine systems). This has resulted to increase demands on policy and decision makers at various levels to develop well-fundamental strategies and solutions. In water management and policy, there is an increasing need for assessment methodologies for diagnosis and prognosis purposes in which an integrated water system approach is considered (Witmer, 1995). Such assessment methodologies should aggregate operational monitoring data to a comprehensive, strategic, preferably simple quantitative form, to support the policy and decision making process.

Monitoring is a sort of information system in which during a certain time on a systematic way data are being collected, handled, managed, analyzed and presented. The ultimate aim of monitoring is to provide information, not data. In the past, many monitoring programmes have been characterized by the "data rich, information poor syndrome" (DRIP-syndrome; Ward *et al.*, 1986). There should be more attention on the analysis and further use of collected data so that end product of monitoring is information. Water quality monitoring systems should be a balanced combination of data collection and information generation. This is illustrated in the monitoring cycle (Figure 1, Timmerman and Hendriksma, 1997), which illustrates that monitoring is a sequence of related activities that begins with the definition of information needs and ends (and starts again) with the use of information products. Too often water monitoring has been viewed as only the first three steps listed in Fig.1. In other words, once data are stored in a computer, the monitoring task is completed. Data are hereby viewed as the final product that is the general perception. However, at that point of obtaining water monitoring data, one is only half way towards the goal of having information about water systems.

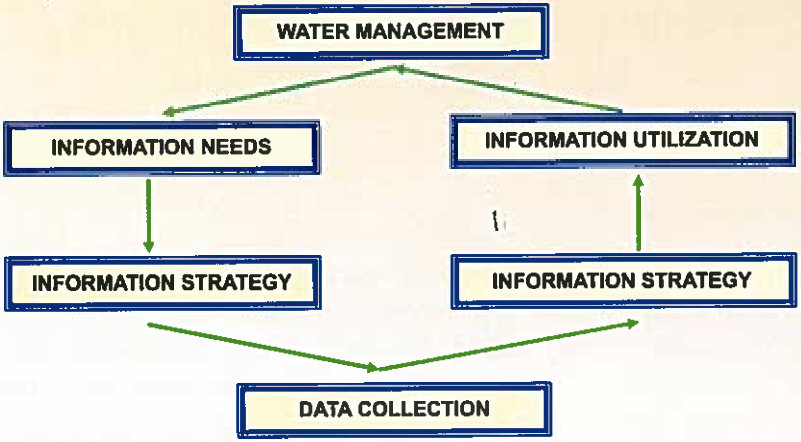

Fig 1. The information cycle (after Timmerman and Hendriksma, 1997)

This element of the monitoring process is regarded as a somewhat separate world of expertise. Recent developments in computing hardware and software made it possible for a broader public to use data more effectively and obtain almost instantaneously results of simple data analysis. These technological and scientific improvements in recent years have not been institutionalized in many monitoring programs, especially in the coldwater sector of this country.

2. MOTIVATION

Water quality management is prime concern for any kind of fishery and related activities since water quality determines to a great extent the success or failure of aquacultural activity. Consequently, water quality monitoring plan is required for proper management of water bodies. Water quality monitoring programs generally involve taking samples but sampling efforts viz. number of monitoring stations, sampling parameters, frequency requires are often restricted due to lack of resources. A well-designed water quality-monitoring plan is quite required to preserve scarce resources by minimizing the redundancy of nearby monitoring stations and the plethora of possible variables monitored, while at the same time maximizing the information content of the collected data. In the recent past, multivariate statistical techniques are emerged to polluting source and run-off from agricultural land, a seasonal phenomenon, largely affected by climate in the basin. The river discharge and subsequently the concentration of pollutants in river water are largely influenced by seasonal variations in hydrologic processes in the river basin (Vega *et al*., 1998). Moreover, the hydrochemistry of surface waters is largely influenced and determined by the natural processes and anthropogenic activities in the region (Carpenter *et al*., 1998). Since, rivers constitute the main inland water resources for fisheries, domestic, industrial and irrigation purposes, it is imperative to understand the hydro-chemical processes for prevention and control of the rivers pollution and to have reliable information on quality of water for effective management. In view of the spatial and temporal variations in hydrochemistry of rivers, regular monitoring programmes are required for reliable estimates of the water quality. This generally, results in a huge and complex data matrix comprised of a large number of physico-chemical parameters (Chapman, 1992), which are often difficult to interpret and drawing meaningful conclusions (Dixon and Chiswell, 1996). The multivariate statistical techniques can be appropriately used for meaningful data reduction and interpretation of multi-constituent chemical and physical measurements (Massart *et al*., 1988). Moreover, factor analysis is useful for identification of the factors that are influence on the water system and helps in detecting the possible sources of river pollution (Singh *et al*., 2005 and Singh *et al*., 2007).

3. FACTOR ANALYSIS

Factor analysis is a multivariate statistical technique applied to a single set of variables when the researcher is interested in discovering which variables in the set form coherent subsets that are relatively independent of one another. Variables that are correlated with one another but largely independent of other subsets of variables are combined into factors. Factors are thought to reflect underlying processes that have created the correlations among variables. Thus, the basic idea of factor analysis is to combine several variables into a smaller set of independent variables without loosing the essential information from the original data set. Factor analysis is normally used to understand the correlation structure of collected data and identify the most important factors contributing to the data structure (Padro *et al*., 1993). There are two major types of factor analysis - exploratory and confirmatory. Exploratory factor analysis (EFA) may be described as orderly simplification of interrelated measures. EFA, factor structure of a set of observed variables. CFA allows the researcher to test the hypothesis that a relationship between observed variables and their underlying latent constructs exists. The researcher uses knowledge of the theory, empirical research, or both, postulates the relationship pattern a priori and then tests the hypothesis statistically. Here we restrict to 'exploratory factor analysis' or simply called 'factor analysis (FA)' only, if our purpose is seeking to describe and summarize a huge data of water quality.

Some authors refer to several different types of factor analysis, such as *R*-factor analysis, *Q*-factor analysis, etc. These simply refer to what is serving as the variables (the columns of the data set) and what is serving as the observations (the rows). According to Thompson (2000), different types of factor analysis are given below:

Type of Factor Analysis	Columns	Rows
	(What the factors explain)	(Measured by the columns)
R	Variables	Participants
Q	Participants	Variables
O	Occasions	Variables
P	Variables	Occasions
T	Occasions	Participants
S	Participants	Occasions

R-factor analysis is the most commonly used. In *R*-factor analysis, rows are cases or participants, columns are variables and cell entries are scores of the cases on the variables. Here the factors are clusters of variables on a set of entities, at a given point of time. *Q*-factor analysis also called 'inverse factor analysis' is factor analysis, which seeks to cluster the cases rather than the variables. That is, in *Q*-factor analysis the rows are variables and the columns are cases and the cell entries are scores of the cases on the variables. In this case the factors are clusters of entities for a set of variables. *Q*-factor analysis is used to establish the factional composition of a group on a set of issues at a given point of time. Other forms of factor analysis are seldom used although they have the same theoretical concept but the terminology and goals are different. Everything we have below refers to *R*-factor analysis.

few unobservable random variables $F_1, F_2, \ldots, F_m$ called common factors, and *p* additional sources of variation, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p$, called errors or, sometimes, specific factors. In particular, the orthogonal factor analysis model is

$$X_1 - \mu_1 = l_{11}F_1 + l_{12}F_2 + \ldots + l_{1m}F_m + \varepsilon_1$$
$$X_2 - \mu_2 = l_{21}F_1 + l_{22}F_2 + \ldots + l_{2m}F_m + \varepsilon_2$$

.

.

.

$$X_p - \mu_p = l_{p1}F_1 + l_{p2}F_2 + \ldots + l_{pm}F_m + \varepsilon_p$$

or, in matrix notation,

$$X - \mu = LF + \varepsilon$$

where *X* is a matrix of order *p*x1, *L* is of order *p*x*m*, *F* is of order *m*x1 and ε is of order *p*x1. The coefficient l_{ij} is called the loading of the *i* th variable on the *j* th factor, so the matrix *L* is the matrix of factor loadings.

The unobservable random vectors *F* and ε satisfy that i) *F* and ε are independent, ii) *E* (*F*)=0, Cov (*F*)=*I* and iii) *E* (ε)=0, Cov (ε)=ψ, where ψ is a diagonal matrix.

Further it can be seen for the orthogonal factor model that

$$\mathrm{Cov}(X) = LL' + \psi$$

or,

$$\mathrm{Var}(X_i) = l_{i1}^2 + l_{i2}^2 + \ldots + l_{im}^2 + \psi_i$$

and

$$\mathrm{Cov}(X_i, X_k) = l_{i1}l_{k1} + l_{i2}l_{k2} + \ldots + l_{im}l_{km}\,.$$

That proportion of the variance of the *i* th variable contributed by the *m* common factors is called the *i* th communality while the proportion of Var(X_i) due to the

5. ADEQUACY OF DATA

The adequacy of the data is evaluated on the basis of the results of a Kaiser-Meyer-Olkin (KMO) sampling adequacy test and Bartlett's test of Sphericity (homogeneity of variance). KMO is a ratio of the observed correlation coefficients to the sum of the observed correlation coefficients and the partial correlation coefficients. It is generally used to evaluate whether the relationship between variables is truly reflective of an underlying process. The KMO measure of sampling adequacy provides an index (between 0 and 1) of the proportion of variance among the variables that might be common variance. Kaiser (1974) suggested that the value of KMO sampling adequacy test less than 0.5 is probably not amenable to useful factor analysis.

Bartlett's test of Sphericity is used to test that variables in the matrix are uncorrelated, an undesirable result. If there are three or more groups, a test for equality of variances, that is the Bartlett's test statistic:

$$B = Q/h$$

Under

$$H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$$

With

$$Q = (n-k)\log S_p^2 - \sum_{i=1}^{k}(n_i - 1)\log S_i^2 \ ; \ i = 1,2,\ldots,k\,,$$

and

$$h = 1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\frac{1}{n_i - 1} - \frac{1}{n-k}\right),$$

where n_i is the size of the sample from the *i* th population, $n = \sum_{i=1}^{k} n_i$ is the total sample size, $S_1^2, S_2^2, \ldots, S_k^2$ are the unbiased estimators of the variances for each of the *k*

6. METHODS OF ESTIMATION

Given observations $x_1, x_2, ..., x_n$ on p generally correlated variables, factor analysis seeks to determine a few important common factors. Suppose the sample covariance matrix S is an estimator of the unknown population covariance matrix Σ. If the off diagonal elements of S are small or those of the sample correlation matrix R essentially zero, the variables are not related and a factor analysis will not prove useful. If Σ appears to deviate significantly from a diagonal matrix then a factor model can be entertained and the initial problem is one of estimating the factor loadings l_{ij} and specific variances ψ_i. There are different methods in literature for the estimation of factor loadings and specific variances. Some of the techniques are discussed below:

Principal Component Analysis (PCA)

The goal of PCA is to extract maximum variance from the data set with each component. The principal components are ordered, with the first component extracting the most variance and the last component the least variance. Since the components are orthogonal, their use in other analyses may greatly facilitate interpretation of results. PCA is the solution of choice for the researcher who is primarily interested in reducing a large number of variables down to a smaller number of components.

Principal Axis Factoring (PAF)

PAF method differs from PCA in that estimates of communality, instead of ones, are in the positive diagonal of the observed correlation matrix. These estimates are derived through an iterative procedure, with squared multiple correlations of each variable with all other variables or, the absolute value of the maximum correlation of that variable with any of the others or, the corresponding diagonal element from the inverse of the correlation matrix used as the starting values in the iteration. PAF is generally used when the research purpose is to identify latent variables, which contribute to the common variance of the set of measured variables, excluding variable-specific variance.

Maximum Likelihood Factoring (MLF)

of a variable minus its communality). MLF generates a chi-square goodness-of-fit test. The researcher can increase the number of factors one at a time until a satisfactory goodness of fit is obtained. However, for large samples, even very small improvements in explaining variance can be significant by the goodness-of-fit test and thus lead the researcher to select too many factors.

Image Factoring

This is based on the correlation matrix of predicted variables rather than actual variables, where each variable is predicted from the others using multiple regression. Interpretation of results of image factoring is not so easy because, loadings represent covariances between variables and factors rather than correlations.

Alpha Factoring

Alpha factoring is based on maximizing the reliability of factors, assuming variables are randomly sampled from a universe of variables. All other methods assume cases to be sampled and variables fixed. Probably the greatest advantage to the procedure is that it focuses the researcher's attention squarely on the problem of sampling variables from the domain of variables of interest. Disadvantages stem from the relative unfamiliarity of most researchers with the procedure and the reason for it.

Unweighted Least Squares Factoring (ULSF)

This method minimizes the sum of squared differences between observed and estimated correlation matrices, not counting the diagonal. Communalities are derived from the solution rather than estimated as part of the solution. Thus, ULSF may be considered as a special case of PAF in which communalities are estimated after the solution.

Generalized (Weighted) Least Squares Factoring

This method also seeks to minimize (off-diagonal) squared differences between observed and reproduced correlation matrices but in this case weights are applied to the variables. Variables that are not as strongly related to other variables in the set are not as important to the solution.

7. NUMBER OF FACTORS

There are several rules for determining how many factors are appropriate for a particular dataset. Some of them are discussed below:

Kaiser Criterion

According to Kaiser criterion, take as many factors as there are eigenvalues >1 for the correlation matrix. Hair, *et al*. (1998) reports that this rule is good if there are 20 to 50 variables, but it tends to take too few if there are <20 variables, and too many if there are >50. Stevens (2002) reports that it tends to take too many if there are >40 variables and their communalities are around 0.4. It tends to be accurate with 10-30 variables and their communalities are around 0.7.

Scree Plot

The Cattell scree test plots the factors as the X-axis and the corresponding eigenvalues as the Y-axis. It takes the number of factors corresponding to the last eigenvalue before they start to level off. Hair, *et al*. (1998) reports that it tends to keep one or more factors more than Kaiser's criterion. Stevens (2002) reports that both Kaiser and Scree are accurate if n>250 and communalities 0.6.

Fixed % of Variance Explained

It keeps as many factors as are required to explain 60%, 70%, 80-85%, or 95%. There is no general consensus and one should check what is common in our field. It seems reasonable that any decent model should have at least 50% of the variance in the variables explained by the common factors.

A Priori

If we have a hypothesis about the number of factors that should underlie the data, then that is probably a good (at least minimum) number to use.

Parallel Analysis

It is also known as Humphrey-Ilgen parallel analysis is often recommended to the X-axis and cumulative eigenvalues on the Y-axis is plotted. Where the two lines intersect determines the number of factors to be extracted.

In practice, there is no single best rule to use and a combination of them is often used successfully, so when we have no a priori hypothesis, check different methods and use the closest thing to a majority decision.

8. ROTATION METHODS

Rotation serves to make the output more understandable and is usually necessary to facilitate the interpretation of factors. The sum of eigenvalues is not affected by rotation, but rotation will alter the eigenvalues of particular factors and will change the factor loadings. A decision is required between orthogonal and oblique rotation. In orthogonal rotation, the factors are uncorrelated. Orthogonal solutions offer ease of interpreting, describing, and reporting results; yet the strain 'reality' unless the researcher is convinced that underlying processes are almost independent. The researcher who believes that underlying processes are correlated uses an oblique rotation. In oblique rotation the factors may be correlated, with conceptual advantages but practical disadvantages in interpreting, describing, and reporting results. Some of the important rotation methods are discussed below:

Varimax Rotation

This is an orthogonal rotation of the factor axes to maximize the variance of the squared loadings of a factor (column) on all the variables (rows) in a factor matrix, which has the effect of differentiating the original variables by extracted factor. Each factor will tend to have either large or small loadings of any particular variable. A varimax solution yields results, which make it as easy as possible to identify each variable with a single factor.

Quartimax Rotation

This method is an orthogonal alternative, which minimizes the number of factors needed to explain each variable. This type of rotation often generates a general factor on which most variables are loaded to a high or medium degree. Such a factor structure

Direct Oblimin Rotation

It is the standard method when one wishes a non-orthogonal (oblique) solution that is, one in which the factors are correlated. This will result in higher eigenvalues but diminished interpretability of the factors.

Promax Rotation

This method is an alternative non-orthogonal (oblique) rotation method, which is computationally faster than the direct oblimin method and therefore is sometimes used for very large datasets.

Among the above rotation methods, varimax rotation is the most common option for many researchers.

9. INTERPRETATION OF FACTORS

To interpret a factor, we need to understand the underlying dimension that unifies the group of variables loading on it. In both orthogonal and oblique rotations, loadings are obtained from the factor-loading matrix, but the meaning of the loadings is different for the two rotations. After orthogonal rotation, the values in the loading matrix are correlations between variables and factors. The researcher decides on a criterion for meaningful correlation collects together the variables with loadings in excess of the criterion, and searches for a concept that unifies them. But, after oblique rotation, the process is the same, however, the interpretation of the values in the above matrix called as pattern matrix in this case, is no longer straightforward. The loading is not a correlation but is a measure of the unique relationship between the factor and the variable. A variable may correlate with one factor through its correlation with another factor rather than directly.

The choice of the cutoff for size of loading to be interpreted is a matter of researcher preference (Tabachnick and Fidell, 2007). The greater the loading, the more the variable is a pure measure of the factor.

10. ILLUSTRATIONS

The catchment area of the river (29⁰17′36″ -29⁰27′48″ N; 29⁰49′ -79⁰26′ E) ranges in elevation from 500-2610 m asl. The upper course of the river flows along a steep gradient (14 - 23 m/ km) cutting a gorge across the Shiwalik hills. Sub-tropical pines and deciduous forests are found in the upper and middle catchment. In the lower course, the hills are mostly with cultivation on the terraces while Sal and riverine forests occur on the upper reaches. Water quality data observed for physico-chemical examination from three different sampling sites at monthly interval from the Gaula River has been used in this study (see Sunder, *et al*., 1991 for further details of data collection). The three sampling sites selected were first at the proposed Jamrani dam site, second at two km upstream of the HMT factory and third at the three km downstream of the HMT factory. The data set has been analyzed by SPSS 12.0 version available at NRC on Coldwater Fisheries, Bhimtal.

Data Analysis

Factor analysis using a principal axis factoring of extraction method and varimax rotation of physico-chemical parameters of the Gaula River has been conducted. Also, correlation matrix is chosen because the covariance method has problems when the variables are measured on widely different scales. When the above procedure of factor analysis attempted to extract 5 factors, the communality of a variable exceeded 1.0 and then the extraction has been terminated. The communality of a variable is the proportion of the variance that is explained by the common factors. Thus, the communality of a variable cannot exceed 1.0. Then, factor analysis is again carried out using principal component analysis of extraction method while the others remain unchanged. Although a number of factor extraction methods are available in literature, principal axis factoring (PAF) and principal component analysis (PCA) are the most commonly used extraction methods for factor analysis. In PCA, the total variance in the data is considered. The diagonal of the correlation matrix consists of unities, and full variance is brought into the factor matrix. Here the factors are called components. In PAF, the factors are estimated based only on the common variance. Communalities are inserted in the diagonal of the correlation matrix. Thus, factors or components are the dimensions (or latent variables) identified with clusters of variables, as computed using factor analysis through the extraction method of PAF or PCA respectively. Tabachnick and Fidell (2007) used the term 'factor' to refer to both the criteria for conducting factor analysis. However, Bartlett's test of shpericity is highly significant (p < 0.001), indicating sufficient correlation between the variables to proceed with the analysis (Table 1). All the extracted communalities are reasonably high (say, > 0.5) and acceptable (Table 2).

Table 1. KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.507
Bartlett's Test of Sphericity	Approx. Chi-Square	312.515
	df	78
	Sig.	.000

Table 2. Communalities

	Initial	Extraction
Air Temperature	1.000	.934
Water Temperature	1.000	.878
Depth	1.000	.783
Flow of Water	1.000	.860
pH	1.000	.806
Dissolved Oxygen	1.000	.679
Free Carbon Dioxide	1.000	.802
Total Alkalinity	1.000	.807
Chlorides	1.000	.723
Calcium	1.000	.591
Magnesium	1.000	.751
Total Dissolved Solids	1.000	.918
Specific Conductance	1.000	.958

Extraction Method: Principal Component Analysis.

be used because subsequent eigenvalues are all < 1. Scree plot shown in Fig 2 is also a useful tool to decide number of components.

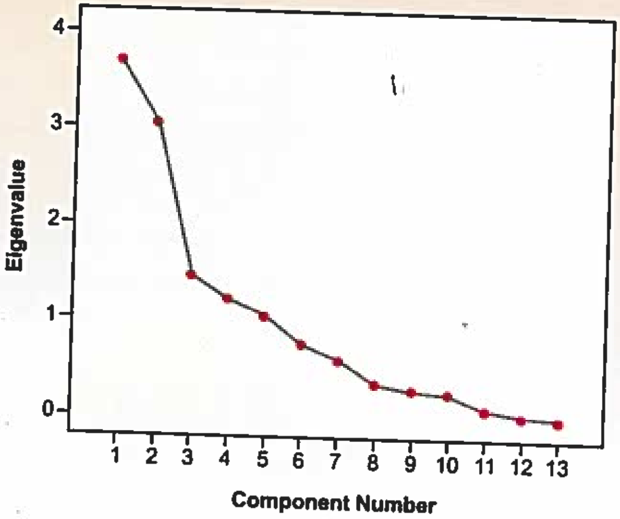

Fig 2. Scree Plot of the eigenvalues for the factor analysis

Component loadings are used to measure correlation between variables and the components. A loading close to ± 1 indicates a strong correlation between a variable and the component, while a loading close to zero indicates weak correlation. Evans *et al*. (1996) considered that those variables exhibited a rotated absolute loading value greater than 0.75 are strongly loaded on a component. Unrotated solutions of component loadings are not suitable for interpretation purposes since the variables generally tend to load on multiple components. The components are rotated with the used of varimax rotation, which is a standard rotation method (Kaiser, 1958). In the present case, only those absolute factor loadings > 0.6 are considered for interpretation purposes.

Results and Discussion

water temperature and also, strong negative loading of pH. Thus, it basically represents the physical parameters group. Temperature affects the physical, chemical and biological processes in water bodies, and therefore, has an important role in determining the concentration of various water quality variables. Thus, water temperature may be considered as an indicator variable of this component. The second component that explains about 19% of the total variance (Table 3) has strong positive loading on chloride and moderate loading on magnesium and calcium. According to Thresh *et al.* (1994), high chloride content of water is an index of pollution from animal origin. Magnesium and calcium are generally found high when chloride concentrations are high. Chloride may be considered as an indicator variable of this component.

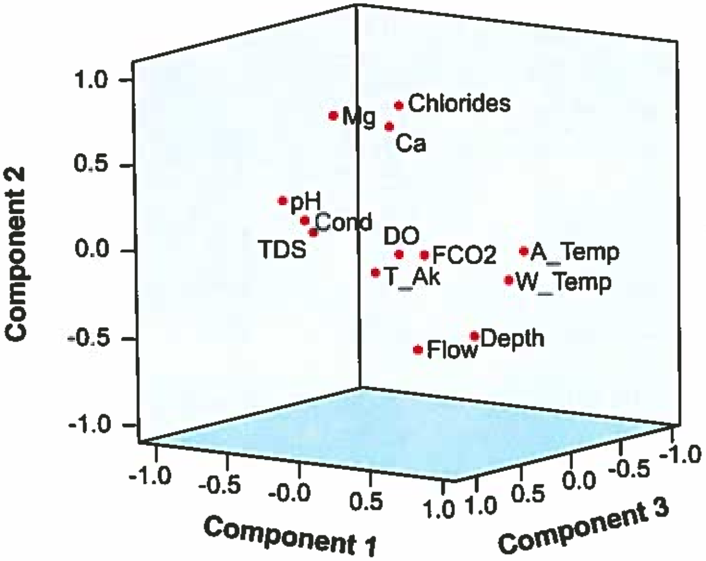

Fig 3. Factor plot in rotated factor space

The third component incorporates those water quality variables that are characteristics of wastewater discharges into the river since it explains about 19% of

Table 3. Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.691	28.388	28.388	3.691	28.388	28.388	2.608	20.061	20.061
2	3.048	23.443	51.832	3.048	23.443	51.832	2.445	18.809	38.870
3	1.466	11.277	63.109	1.466	11.277	63.109	2.437	18.748	57.618
4	1.230	9.463	72.572	1.230	9.463	72.572	1.730	13.304	70.922
5	1.056	8.124	80.696	1.056	8.124	80.696	1.271	9.775	80.696
6	.775	5.960	86.656						
7	.607	4.667	91.323						
8	.379	2.912	94.235						
9	.307	2.365	96.600						
10	.267	2.056	98.656						
11	.110	.845	99.501						
12	.048	.367	99.868						
13	.017	.132	100.000						

Extraction Method: Principal Component Analysis.

explains about 13% of the total variance. In general, negative loading on flow of water similar to dissolved oxygen is expected because, high speed of flow of water, in usual case, increases the level of dissolved oxygen. However, oxygen depletion often results during high speed of flow of water exceptionally in this river. The water flow of this river is quite high especially during the rainy season wherein organic matters from various sources are added to the river water, resulting in depletion of dissolved oxygen level. After the monsoon months, flow of water is gradually reduced from month to month and dissolved oxygen content approaches to a stable level based on the volume of water. Dissolved oxygen content, which plays a vital role in supporting aquatic life in running waters may be considered as the key parameter of this component. The fifth component has strong loading on free carbon dioxide alone that

Table 4. Rotated Component Matrix (a)

	Component				
	1	2	3	4	5
Air Temperature	.920	.059	.171	.107	.211
Water Temperature	.784	-.128	.132	.398	.268
pH	-.750	.194	.186	-.054	.410
Chlorides	.049	.820	.160	-.147	.021
Magnesium	-.473	.699	.071	.137	.122
Calcium	-.004	.696	.191	.085	-.251
Depth	.438	-.515	.003	.489	-.294
Specific Conductance	-.091	.243	.910	.248	-.016
Total Dissolved Solids	-.039	.191	.895	.281	.008
Total Alkalinity	.315	-.035	.791	-.282	.040
Dissolved Oxygen	-.193	-.116	-.170	-.763	-.129
Flow of Water	.106	-.602	.067	.682	-.130
Free Carbon Dioxide	.095	-.065	-.004	.046	.887

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization.

a Rotation converged in 6 iterations.

It is summarized that factor analysis is able to identify significant sources of water quality inputs to Gaula River. The first five components account for (81%) almost the total water quality variation. The largest source of variation (20%) appeared to be from water quality parameters associated with physical parameters. Additional inputs from the second component accounting for about 19% of animal waste discharges into the river. The third source of variation (19%) appeared to be associated with wastewater discharges; and the fourth component, accounting for 13%. A monitoring program could use a smaller set of variables to identify times for intensive sampling; water

REFERENCES

Carpenter, S.R., N.F. Caraco, D.L. Correll, R.W. Howarth, A.N. Sharpley and V.H. Smith (1998). Nonpoint pollution of surface waters with phosphorous and nitrogen. *Ecol. Appl.* **8**: 559-568.

Chapman, D. (1992) *Water quality assessment*. In: Chapman D. (Ed.), on behalf of UNESCO, WHO and UNEP, Chapman & Hall, London, pp. 585.

Child, D. (1990). *The essentials of factor analysis (2nd Edition)*. London: Cassel Laticaudus Muller and Henle.

Dixon, W. and B. Chiswell (1996). Review of aquatic monitoring program design. *Water Research*, **30** : 1935-1948.

Evans, C.D., T.D., Davies, Jr. P.J.W., M. Tranter and W.A. Kretser (1996). Use of factor analysis to investigate processes controlling the chemical composition of four streams in the Adirondack Mountains, New York. *Journal of Hydrology*, **185**: 297-316.

Hair, J.F. Jr., R.E. Anderson, R.L. Tatham and W.C. Black (1998). *Multivariate Data Analysis, (5th Edition)*. Upper Saddle River, NJ: Prentice Hall.

Johnson, D.E. (1998). *Applied multivariate methods for data analysis*. Pacific Grove, CA: Brooks/Cole Publishing.

Kaiser, H.F. (1958). The varimax criterion for analytic rotation in factor analysis. *Pyrometrical*, **23**: 187-200.

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

Lance, C.E., M. B. Marcus and C.M. Lawrence (2006). The sources of four commonly reported cutoff criteria: what did they really say? *Organizational Research Methods*, **9**(2): 202-220.

Massart, D.L., B.G.M. Vandeginste, S.N. Deming, Y. Michotte and L. Kaufman (1988). *Chemometrices: A Textbook*. Elsevier, Amsterdam.

Padro, R., E. Barrado, Y. Castrillejo, M.A. Valasco and M. Vaga, (1993). Study of the contents and speciation of heavy metals in river sediments by factor analysis. *Anal. Lett*. **26** : 1719-1739.

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

Singh, N.O., S. Kumar and P.C. Mahanta (2007). Multivariate statistical anlysis of water quality
in Narmada

Sunder, S., H.S. Raina, M. Mohan and B. Singh (1991). Studies on bio-ecology of a mountain river (the Gaula) in Kumaon. Unpublished Project Report, National Research Centre on Coldwater Fisheries, ICAR, Bhimtal

Tabachnick, B.G. and L.S. Fidell (2007). *Using multivariate statistics (5th Edition).* Pearson Education, Inc. USA.

Thompson, B. (2000). Q-technique factor analysis: one variation on the two-mode factor analysis of variables. In: L.G., Grimm and P. Yarnold (Eds.), *Reading and understanding more multivariate statistics*. Washington, D.C.: American Psychological Association.

Thresh, J.C., E.V. Suekling and J.F. Beale (1994). Chemical and zooplankton studies of lentic habitats in Northeastern new South Wales. *Australian Journal of Freshwater Research*, **21**: 11-33.

Timmerman, J.G. and J. Hendrikhsma (1997). Informatie op maat: een raamwerk voor waterbeheer. H_2O (30), **17** : 528-530. (In Dutch, English abstract).

Vega, M., R. Padro, E. Barrado and L. Deban (1998). Assessment of seasonal and polluting effects on the quality of river water by exploratory data analysis. *Water Research*, **32**: 3581-3592.

Ward, R.C., J.C. Loftis and G.B. McBride (1986). The "data rich but information-poor" syndrome in water quality monitoring. *Environmental Management*, **10**(3): 291-297.

Wilkinson, L., G. Blank and C. Gruber (1996). *Desktop data analysis with SYSTAT*. Upper Saddle River, NJ: Prentice-Hall.

Witmer, M.C.H. (1995). Information needs for policy evaluation. In: M.J. Adriaanse, J. van der Kraats, P.G. Stoks and R.C. Ward (Eds.), *Proceedings of the International Workshop Monitoring Tailor-Made I*, 1994, Beekbergen, the Netherlands, pp. 55-61.

[illegible] press
[illegible]
[illegible]