Multivariate statistical techniques and water quality assessment: Discourse and review on some analytical models
Kumar Manoj, Pratap Kumar Padhy
Department of Environmental Studies, Institute of Science, Visva-Bharati University, Santiniketan, 731235, Birbhum, West Bengal, India
pkpadhy@visva-bharati.ac.in
doi: 10.6088/ijes.2014050100053

ABSTRACT

Regular monitoring and comprehensive assessment of water quality and its associated processes require sophisticated analytical models to reveal concealed instruments controlling their properties. This information is essential to design monitoring frameworks and sustainable management of the water resources. Intelligent data analysis techniques like multivariate statistical models can greatly assist in water quality management programs. This paper provides basic knowledge of the five multivariate data mining approaches, namely, cluster analysis, principal component analysis, factor analysis, multiple linear regression analysis and discriminant analysis, and highlights their applications in the characterization and classification of the surface water quality. The applicability of multivariate tools for the river basin management is the principal focus of this communication. Furthermore, this literature review also presents some of the basic concepts of the newly employed source apportionment receptor modeling technique involving multiple linear regression (MLR) and absolute principal component scores (APCS-MLR model) for extensive water quality assessment.

Keywords: Cluster analysis, Discriminant analysis, Environmetrics, Factor analysis, Multiple linear regression analysis, Principal component analysis.

1. Introduction

The multi-usage of water resources, such as rivers and lakes, for drinking, agricultural and industrial purposes, fisheries and energy productions depend considerably on their quality (Iscen et al., 2008). This quality, defined in terms of physical, chemical and biological compositions, is governed by both natural (precipitation, watershed geology, topography, climate) and anthropogenic (point and non-point sources like urban and industrial activities, other domestic activities, agricultural runoff) factors (Mustapha and Abdu, 2012). Polluted surface waters critically alter the balanced ecosystem which is essential for the beneficial interactions of the living things and the environment (Iscen et al., 2008). Pollution leads to disturbance of harmony in the nature. In addition to assessment of quality of the aquatic systems, identification of the factors controlling their behavioural properties is increasingly becoming inherent part of the water quality management programmes. However, due to spatial and temporal variations in hydrochemical and biological properties, continuous and regular monitoring programmes are required to have reliable information about the water quality (Singh et al., 2005). The database generated is large and complex and, therefore, its analysis requires some advanced and sophisticated analytical tools. The water quality data contains many concealed behavioural properties of the surface water body and, also the information about the region. Interpreting these hidden factors is essential for water quality
management. For example, to design a river basin monitoring network. Investigation of water quality is important to assess the health of the ecosystem in order to take necessary management steps required to control pollution of water resources and their sustainable utilization. Multivariate statistical techniques can reveal the concealed information present in large quality data matrices and avoid misinterpretation of the data obtained from the monitoring network. When multivariate statistical models are used in environmental science, the branch is broadly called environmetrics or chemometrics. In water quality studies, the applied part can be designated as hydrometrics. Although, a review of scientific literature shows several studies on environmetric techniques, a systematic communication depicting basic information about most commonly used multivariate data mining approaches especially with reference to water quality data is hardly available. Moreover, environmetric works on freshwater systems are needed to put together to arrive at their discrete applications in water management. To bridge this gap, in this communication, we present the application of five multivariate statistical techniques, namely, cluster analysis (CA), principal component analysis (PCA), factor analysis (FA), multiple linear regression analysis (MLR) and discriminant analysis (DA) in water quality data assessment. Furthermore, to make the review work more specific, only most utilizable surface water resources have been considered. The manuscript is divided into two parts. The first part deals with basic knowledge about five considered multivariate environmental data mining techniques, and the second part, which is a literature survey, discusses their application in surface water quality analysis. A short note on the newly applied absolute principal component score-multiple linear regression (APCS-MLR) receptor model on hydrological studies is also supplied in this communication.

2. Multivariate statistical techniques

Statistical analyses can be categorized into univariate, bivariate and multivariate techniques. A common example of univariate statistics is the computation of arithmetic mean. Correlation and simple regression analyses, on the other hand, are bivariate statistical tools. Relating to a multivariate analysis, it simply involves simultaneous examination of more than two variables (Hair et al., 2010). All statistical systems concerned with simultaneous analysis of multiple measurements on many different variables constitute the multivariate analysis (Johnson and Wichern, 2007; Hair et al., 2010). This system is also called variable-directed multivariate analysis. However, in environmental studies, identification of relationships between samples are equally as important as revealing their relationships between the variables. The former defines the sample-directed multivariate analysis (Mazlum et al., 1999). Environmetric techniques use multivariate statistical models for better interpretation of the data quality. Some commonly used multivariate statistical models for environmental data analysis are: CA, PCA, FA, MLR and DA. These intelligent data mining techniques are very helpful in pattern recognition and exploratory data analysis driving hidden information from the dataset (Kanade and Gaikwad, 2011). This section deals with data pre-treatment techniques and various multivariate approaches.

2.1 Normal distribution and homoscedasticity

Assumptions of normal distribution of variables and homoscedasticity or homogeneity of variance are two intrinsic parts of many statistical procedures such as multivariate approaches (Osborne, 2002). Although, data normality is a distinct requirement for parametric analyses to avoid Type I or II error, according to some authors, non-parametric analyses can also suffer when assumption of normality is violated (Zimmerman, 1995; Zimmerman, 1998; Osborne, 2002). In normal distribution, the distribution of individuals or measured values is
symmetrical about their mean value which leads to a bell shaped curve. Since, the distribution is symmetrical about the mean; it becomes equal to the median of the distribution (Gun et al., 2003). In other words, the mean and the median are same for the dataset. Many multivariate statistical techniques strictly work on the assumption that equal variance of the population error exists (Hair et al., 2010). Homoscedasticity assumption denotes constancy of the disturbance variance at each observation (Holgersson and Shukur, 2003). In homoscedastic data, a noticeable constant variance of the error terms over a range of calculated variables is obtained, while heteroscedastic data have non-constant (increasing or modulating) variance of the error terms (Hair et al., 2010). Several statistical methods are available to check the data for normality and homoscedasticity. The skewness and kurtosis are commonly used tools to examine normal distribution of the individuals or objects. Skewness of a frequency distribution is simply the departure from symmetry or an asymmetrical distribution denotes skewness. A positive skewness has less frequency towards higher values of the variable and consequently has the longer tail of the distribution curve whereas a negative skewness has more frequency towards higher values of the variable and consequently the tail is towards the lower values of the distribution curve. A positively skewed distribution shows mean > median > mode and a negatively skewed distribution display mean < median < mode. Kurtosis describes frequency distribution in terms of degree of “peakedness” or steepness. When the distribution is high, has a narrow peak and long tails, it is called leptokurtic (positive kurtosis), while distribution with a low, wide peak and short tails is called platykurtic (negative kurtosis). In case of normal distribution both skewness and kurtosis approach zero magnitude (Gun et al., 2003). Statistical plots like P-P plot, Q-Q plot and box-and-whisker plot are also utilized to evaluate normality assumption. The normal P-P plot (percentile-percentile plot or probability-probability plot) is a graphical representation where theoretical percentiles of a normal distribution on the Y-axis are plotted against the observed empirical sample percentiles on the X-axis (Totton and White, 2011).

The P-P plot demonstrates outcome of observed cumulative probability versus expected cumulative probability. A normally distributed data will have plotted points positioned along the diagonal straight line of the graph. The Q-Q plot is designed to determine marginal distribution of variable observations in a sample. Here, the sample quantile is plotted against the expected quantile when the observations would be normally distributed. The assumption of normal distribution is accepted if the plotted points, similar to the P-P plot, closely approach the straight line and are approximately linear. A plot showing deviation of points from the straight line points to departure from normality (Johnson and Wichern, 2007). Box-and-whisker plots are greatly applied to visually display distribution of variables in analyzed samples. The plot is a summary of five components viz. minimum, 25th percentile (1st or lower quartile), 50th percentile (portraying the central tendency median), 75th percentile (3rd or upper quartile) and maximum of the data spread (Park, 2008). The plot also represents outliers as closed circles flanking the whisker lines. Thus, the box-and-whisker plots effectively display skewed nature of data and the presence of potential outliers (Totton and White, 2011). For a normally distributed data, the plot will exhibit equal proportions around the median or, in other words, its 25th and 75th percentiles will be symmetrical (Park, 2008; Totton and White, 2011). Moreover, the plot’s mean and median will fall at the same line in the middle of the box (Park, 2008). In addition to the above mentioned methods, the Kolmogorov-Smirnov test, Shapiro-Wilk test and Anderson-Darling test are also frequently used to probe normality assumption (Thyne et al., 2004; Park, 2008).

2.2 Transformations and standardization

Kumar Manoj, Pratap Kumar Padhy
International Journal of Environmental Sciences, Volume 5 No.3, 2014
Data transformation is necessary to improve normality of analyzed variable and achieve homoscedasticity. Three frequently utilized transformation methods are square root transformation, inverse transformation and logarithmic transformations to convert original values to new magnitudes. As it is not possible to take the square root of a negative value, a constant should be added to the observed values of the variable to shift the minimum value of the data spread above zero, and preferably to one (Osborne, 2002). In inverse transformation, the reciprocal of the data value is computed. Although, a range of log transformations employing different bases are possible, transformations with base 10 is most regular, particularly, when the range is extreme. The natural logarithm, having constant e (= 2.7182818) as the base, can also be applied to transform data (Osborne, 2002; Osborne, 2010). Additionally, the Box-Cox transformation is also advocated (Osborne, 2010). Data standardization to standard scores, also called z-scores, is an essential procedure in multivariate modeling. The principal objective of the z-score standardization is to bring the variances of the dataset closer. The standardization procedure does not disturb the distribution of the data; rather it simply scales and shifts the data so that equal weights are assigned to them and variances between the parameters are reduced (Yidana et al., 2012). It scales the data to unit standard deviation centred about a zero magnitude mean (Hair et al., 2010). The z-score standardization computed is expressed as:

\[ z = \frac{(x - \bar{x})}{sd} \]  

(1)

Where \( x \), \( \bar{x} \) and \( sd \) are the observed value, mean value and standard deviation of each parameter of the dataset. Data standardization is essential to perform CA, PCA and FA. For DA transformed data, sometimes also called raw data, is used. MLR also functions on the principles of normality and homoscedasticity and, therefore, transformation is necessary for a non-normal data.

2.3 Cluster analysis

CA is a classification technique which can be used to determine existence of natural subgroups or classes of individuals in a given dataset. Here, individuals are grouped based on their similarities and dissimilarities. Each group represents a cluster of individuals or objects with homogeneous properties separated from a heterogeneous large population. The analysis contains two fundamental steps or choices: a proximity measure and a group-building algorithm. While the proximity measure (determined by a distance or similarity matrix and the resulting similarity coefficients) checks closeness or homogeneity of the objects, the group-building algorithm assigns groups to the objects based on assessments of the former so that objects in the same group are intimately homogeneous and significant differences exist between different groups (Härdle and Simar, 2003). Clustering can be hierarchical or non-hierarchical (Manoj et al., 2013a). The hierarchical method is a stepwise procedure where objects combine or split in a series of steps making a tree like structure (dendrogram). It is of two types, agglomerative and divisive or splitting. In the agglomerative clustering, each object begins as its own cluster which is successively combined to the next closest object forming a new cluster.

This process is repeated continuously until there is only one cluster. The opposite procedure operates for the divisive clustering where the single cluster comprising all objects splits successively into finer groups until each cluster has a single member (Hair et al., 2010). The non-hierarchical clustering (also called partitioning clustering) does not engage any tree like building process; instead objects are assigned into the pre-specified number of clusters. Here,
exchange of objects between clusters takes place and consequently, groups assigned to objects may change until an optimum score is attained (Härdle and Simar, 2003; Hair et al., 2010). It involves initial partition of objects into pre-defined clusters; shifting of objects into other clusters showing closer mean vector than their present cluster; revision of relevant cluster mean vectors after each shift; and continuation of process until all objects belonging to a particular cluster display closeness to their own cluster mean vector as compared to other clusters (Landau and Everitt, 2004). Objects in the groups are distinguished based on some distance and similarity measures (Landau and Everitt, 2004). There are several distance measures available to perform cluster analysis like the Euclidean distance, squared Euclidean distance and the Mahalanobis distance. An inverse relationship converts distance into similarity measure (Hair et al., 2010). The method of computing the squared Euclidean distance can be expressed as:

\[ d_{ij}^2 = \sum_{k=1}^{n} (z_{ik} - z_{jk})^2 \]

Where, \( d_{ij}^2 \) is the squared Euclidean distance; \( z_{ik} \) is the value of \( k \) variable for the object \( i \); \( z_{jk} \) is the value of \( k \) variable for the object \( j \); and \( n \) is the number of variables (Ayeni, 2011).

### 2.4 Principal component analysis and factor analysis

When dealing with PCA and FA some careful approaches are needed as the terms are often confusing. As described in a recent work and references within two types of FA generally used are common factor analysis (CFA) and PCA (Kim, 2008). The procedure of performing these two analyses are almost identical in software systems (e.g., SPSS, XLSTAT). However, in most of the multivariate modeling environmental studies FA follows PCA (on the original component matrix). This concept of PCA and FA utilization is described in this article. Readers are advised to consult references given in the text for a comprehensive judgment as explaining all is beyond the scope of this communication. PCA is a powerful pattern recognition technique that aims to reduce multidimensionality of datasets. The newly obtained results in lower dimensions simplify data interpretation (Singh et al., 2005; West, 2009). Here, observed inter-correlated variables are transformed to a new set of uncorrelated variables organized into order of decreasing importance (Mazlum et al., 1999). These transformed uncorrelated or independent variables are called principal components (PCs) (Singh et al., 2005). The meaning of dimension reduction is to derive lesser components that account for the variability found in a relatively large number of measures. The PCs are defined in terms of linear combinations of the measurements and contain both common and unique variance (DeCoster, 1998). In other words, each orthogonal PC is a linear combination of the original variables and denotes a different variation source (Praus, 2007). These PCs disclose information about the most significant variables explaining the complete dataset by rendering exclusion of the variables considered less significant, while still retaining original information with minimum loss (Singh et al., 2005; Juahir et al., 2011). The technique uses correlation or covariance matrix to characterize dataset. As described by Mazlum and co-authors (including references within their work), when the variables have wide different units and are considered equally important, the use of correlation matrix is recommended. Moreover, to work with the correlation matrix variables should be scaled or
standardized to have unit variance (Mazlum et al., 1999). The PCs generated through PCA can be expressed as:

\[ y_{ij} = a_{i1}x_{1j} + a_{i2}x_{2j} + a_{i3}x_{3j} + \cdots + a_{im}x_{mj} \]  

(3)

Where \( y \) = component score, \( a \) = component loading, \( x \) = measured value of the variable, \( i \) = component number, \( j \) = sample number, and \( m \) = total number of variables. The component weights represent correlation between the PCs and the variables (Praus, 2007). In PCA, eigenvalues indicate significance of the components. The component which displays highest eigenvalue is taken as the most significant. The eigenvalues of magnitude one or more should be considered for multivariate analysis. These eigenvalues signify variations demonstrated by the observed variables (Praus, 2007). The first PC represents largest variability in the original dataset (Zare Gariji et al., 2011). Based on their absolute magnitude, component loading values greater than 0.75, between 0.75-0.5 and 0.5-0.3 are classified as strong, moderate and weak respectively (Nair et al., 2010; Manoj et al., 2013b). Kaiser-Meyer-Olkin (KMO) and Bartlett’s tests of sphericity are initially performed on the dataset to determine their appropriateness to undergo PCA. KMO test is a measure of sampling adequacy with values between 0 and 1. A value nearing 0 indicates unsuitability of the dataset for PCA whereas magnitude closer to 1 shows its suitability. Any value exceeding 0.60 is acceptable for PCA (Arumugam et al., 2010). However, some authors consider KMO values greater than 0.80 as desired to conduct PCA. On the other hand Bartlett’s test of sphericity examines whether the correlation matrix is an identity matrix (Eyduran et al., 2009). If the correlation matrix is an identity matrix then all variables become unrelated making PCA model inappropriate and unsuitable statistical tool for advanced data analysis. In other words, to work out PCA there must be some relationships between the variables. P <0.05 is considered significant for the Bartlett’s test (Nair et al., 2010). The PCs obtained are sometimes not readily interpreted and, therefore, are rotated to generate a new rotated component matrix from the original component matrix. This helps in easy interpretation of the data quality. VARIMAX approach is the most widely used rotation technique.

The rotation modifies the correlation between the components and the original variables, so that in the new extracted components only the most important variables are included (Mazlum et al., 1999). These new groups of variables or components are called varimax factors or varifactors (VF’s). The new factor loadings (earlier component loadings) generated illustrates the correlation between the variables and the factors (Kothai et al., 2008). FA further reduces the role of less significant variables explaining dataset obtained from PCA (Singh et al., 2005). The objects displaying higher loading in each factor are interpreted as hallmarks of pollution source that it symbolizes (Kothai et al., 2008). Although, both PCA and FA are dimension reduction techniques, there are some fundamental differences between them. For example, while the PCs are simply based on the measured responses, FA assumes that measured responses are the outcomes of the underlying factors (DeCoster, 1998). FA is based on assumptions and has an error structure, while PCA is entirely a mathematical technique independent of assumptions (Mazlum et al., 1999). While in FA, the new groups of variables can include unobservable, hypothetical and latent variables, the PCs in PCA are simply the linear combinations of observable variables (Iscen et al., 2008). In former case, this happens because the new measured variable is expressed as a combination of factors and also involves a residual term (Singh et al., 2005). In other words, in FA, each new measured variable can be expressed as a linear combination of latent common factors and a single specific factor (Praus, 2007). In most of the environmental studies, PCA and FA are discussed as PCA/FA. Moreover, in environmental literature the term factor is commonly
used for both components and factors. The equation of FA is expressed as (Singh et al., 2005; Juahir et al., 2011):

$$ z_{jt} = \alpha_{j1}F_{1t} + \alpha_{j2}F_{2t} + \cdots + \alpha_{jm}F_{mt} + e_{jt} $$  \hspace{1cm} (4)

Where, \( z \) = determined value of a variable; \( a \) = factor loading; \( f \) = factor score; \( e \) = residual term accounting for errors or other sources of variation; \( i \) = sample number; \( j \) = variable number, \( m \) = total number of factors. PCA aims to explain total variation expressed in the correlation matrix, while FA defines correlation present in the common factor portion (Praus, 2007).

2.5 Multiple linear regression analysis

Regression analysis in simple terms, discloses average relationship between two variables, designated as dependent and independent variable, which makes estimation or prediction mathematically possible. The relationship may be linear or non-linear. A linear relationship is obtained when the dependent variable shows constant absolute change in response to change in the independent variable by one unit (Patel and Dhiman, 2011). Linear regression analysis is conducted to predict a dependent variable, also known as response or criterion variable, mathematically from one or more independent variables, also known as predictor or explanatory variables. In simple linear regression analysis, there is only one independent variable and is expressed as (Landau and Everitt, 2004):

$$ y_i = \beta_0 + \beta_1 x_i + e_i $$  \hspace{1cm} (5)

In MLR, which is basically an extension of the simple regression, there is more than one independent variable to predict the dependent variable, and both dependent and independent variables are metric ones. Mathematically MLR model is expressed as:

$$ y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_q x_{qi} + e_i $$  \hspace{1cm} (6)

Where, \( y \) = dependent variable, \( y_1, y_2, \ldots, y_n \) as the observed values, \( n \) = sample size; \( x_1, x_2, \ldots, x_q \) = explanatory or independent variables, \( x_{1i}, x_{2i}, \ldots, x_{qi} \) are the observed values, \( i = 1, \ldots, n \); \( e_i \) = residual or error for individual \( i \); \( \beta_0 = \text{constant; } \beta_{1,\ldots,q} = \text{multiple regression coefficients (Landau and Everitt, 2004). } \) MLR generates a set of best linear combination of independent variables called “regression variate or regression equation or regression model” that predict the dependent variable (Hair et al., 2010). Variations in the explanatory variables describe the variation in the dependent variable (Mazlum et al., 1999). The best equation selected is based on the highest multiple correlation coefficient (R), lowest standard deviation and magnitude of the F-ratio (Chenini and Khemiri, 2009). Thus, MLR can also reveal statistically significant variables of the system (Ott, 1998; Golfinopoulos and Arhonditis, 2002). To employ MLR for statistical modeling, two basic criteria required are: metric nature (appropriately transformed data) and a prior decision on selection of dependent and remaining independent variables (Hair et al., 2010). The predicted or estimated quantity of the dependent variable is a linear transformation of the independent variables such that the sum of squared deviations of the observed and predicted value of the dependent variable is a minimum. However, the computations of the predictive functions are highly complex as interrelationships among all the variables must be considered when assigning weights to the variables (Donia, 2011).
2.6 Discriminant analysis

MLR is used in situations where the dependent variable is a metric variable whereas in DA, the dependent variable is non-metric (nominal or categorical). It is also a classification and dimension reduction technique and is exploratory in nature. The basic objectives of DA are to differentiate between population groups and to assign new observations into the classified groups (Härdle and Simar, 2003). Thus, unlike cluster analysis, which creates groups or clusters from an unclassified data, DA operates when the dataset is already classified. A minimum of two groups are required for the DA to perform. When only two groups are involved, the technique is called “two-group DA” and when three or more groups are identified, the technique is called “multiple DA” (Hair et al., 2010). In data mining, DA is also employed to discover most important quantitative variables separating the groups and for testing the hypothesis on the differences between the groupings expected (Fernandez, 2002).

For effective data analysis, it is necessary to describe between-population as well as within-population variations. PCA describes overall variability occurring and may not be appropriate to obtain a clear understanding of between-group variations present. PCA tends to follow the direction of largest total variance, DA, on the other hand, capitalize on maximizing between-groups variation while minimizing within-group variation. Thus, DA becomes not only an effective technique to evaluate relationships between different clusters but also to perform the best discrimination of individuals into groups defined a priori (Jombart et al., 2010). The analysis derives a variate, also called discriminant function (DF), which is a linear combination of two or more independent or predictor variables with their calculated weightings. The DF defines the discriminant score computed for each object in the analysis (Hair et al., 2010; Ramayah et al., 2010). The maximum number of DFs generated is one less than the number of predictor variables or the number of groups, whichever is smaller (Ramayah et al., 2010). The variate of the DA is expressed as:

\[ z_{jk} = \alpha + w_1x_{1R} + w_2x_{2R} + \cdots + w_nx_{nR} \]  

(7)

Where, \( z_{jk} \) = discriminant z score of discriminant function j for object k; \( \alpha \) = intercept; \( w_i \) = discriminant weight or coefficient for the independent variable i; \( x_{IR} \) = independent variable i for object k.

3. Multivariate statistical techniques and water quality data analysis

In the preceding section, we described various multivariate data analysis techniques commonly used in environmental studies. In this section, the application of multivariate techniques in characterization and classification of surface water resources, with special reference to the river basin management is discussed. The discussion is in the form of synopsis of some important works done; as it is necessary to understand the applied parts of multivariate models. Strategic management of water resources is desperately needed as their pollution has become a major problem that can affect their availability for consumption. Many phenomena explain the need for advanced data mining tools in water quality management. For example, land use pattern transformation causes change in the types of pollutant loadings into the river system. Since, both point and non-point sources contribute to the pollution load, identification of origin of each pollutant becomes greatly difficult (Juahir et al., 2011). Source identification and the apportionment of contribution of different sources to the pollution load of each pollutant are some of the most basic requirements for the sustainable management of surface water resources.
Water quality dataset is multi-constituent and complex. Multivariate statistical techniques have greatly been applied in data mining to obtain all possible information about the water resources for their effective management. Both PCA/FA and MLR were applied to study water quality of the Jakara river (Nigeria) during investigation on the origin of pollutants impacting water quality and their contribution towards quality variation. The rotated component matrix, which consisted of five VFs, explained 83.1% of the total variance in the dataset and identified five different sources responsible for affecting water chemistry: interactions of ions and their compounds, erosion runoff, domestic wastes, dilution effect and agricultural runoff. Applied MLR further disclosed role and contribution of five most important studied parameters \( R^2 = 0.942 \) namely: dissolved oxygen, biochemical oxygen demand, suspended solids, total solids and chloride, responsible for quality variation in the river system. The multivariate technique clearly highlighted waste disposal into the river as chief cause of deteriorating water quality and, hence, a need to monitor human activities to minimize adverse effects on the river (Mustapha and Abdu, 2012). Assessment of surface water quality requires source identification of pollutants and the contribution of chemical constituents in explaining total variation.

Intelligent statistical tools are important medium to warn people and the environmental agencies in advance to take necessary measures to protect and sustain the freshwater resources from getting polluted. Klang river (Malaysia) water quality was modelled employing combined PCA/FA and MLR techniques. The two major objectives of the study were: to find out the pollution sources and their contribution in affecting the water quality and, to determine the most relevant pollution sources controlling water quality index (WQI) variability in the studied area. The VFs generated from PCA/FA were used to predict WQI employing MLR. WQI is commonly employed as a suitable ecological indicator of water pollution. Moreover, there are some WQIs which are specifically developed to determine drinking water suitability.

These are simple statistical tools to predict changes in water quality and trend analysis both spatially and temporally. In the initial step, PCA/FA identified nine possible pollution sources, which explained 72% of variability in the data matrix, responsible for influencing the river water chemistry, namely: soil erosion, organic pollution in the form of waste disposal activities (anthropogenic input), surface runoff, faecal waste, detergent, urban domestic pollution, industrial effluent, fertilizer waste and residential waste. To quantify contribution of each identified source towards WQI in order to discover main pollution sources receptor model APCS-MLR (absolute principal component scores-multiple linear regression) technique was applied. While PCA/FA showed soil erosion as the most important agent for variability in the Klang river, APCS-MLR revealed urban domestic pollution was the highest pollution contributor followed by anthropogenic input. The study showed that combination of PCA/FA and MLR can provide good performance in modeling techniques because of their efficacy in removing problems of collinearity and lessening the number of predictor or independent variables. According to the authors, since, the model is good for advanced forecasting of WQI, the environmental monitoring programme can be made more robust and efficient. The model is also helpful in reducing sampling campaigns and the reagents’ costs (Nasir et al., 2011).

A surface water body receives physicochemical and biological constituents from both point and non-point sources. A water body, for example a large canal, may receive water from many smaller streams. If the latter happens to be polluted, the former becomes polluted as
well. Multivariate techniques can be employed to investigate effects of water quality of the input sources to the water quality of the major receiving body. Drained agricultural water can be mixed with the fresh river water (keeping salinity below the recommended standard) to meet the supply of irrigation consumption. The El Salam canal project of Egypt is an excellent example in this respect. The agricultural drained water is reused by mixing with the fresh Nile water for a sustainable agricultural system. The canal receives Nile water (from the Damietta tributary branch) and waters of the El Serw drain and Bahr Hadous drain. MLR was performed between the El Salam canal water quality (before meeting the Suez canal) as the dependent variable and water quality of the Damietta branch, El Serw drain and Bahr Hadous drain as the independent variables to show the control of input concentrations on the endpoint concentration of the canal.

The most significant parameters identified responsible for affecting the water quality were: bacteriological (total coliform and faecal coliform), organic pollutants (Biochemical Oxygen Demand and Chemical Oxygen Demand), nutrients (total nitrogen and total phosphate), major ions (total dissolved solids) and heavy metals like copper and manganese. The analysis identified El Serw drain as mainly responsible for total coliform, faecal coliform, biochemical oxygen demand, total nitrogen and copper concentrations in the El Salam canal downstream, whereas Bahr hadous drain was essentially identified for total phosphate, total dissolved solids and manganese. The study showed that the quality of the agricultural drainage water was further lowered by mixing of domestic wastes, municipal sewage and industrial wastes and, hence, it was necessary to do treatment of drains carrying drained agricultural water before their discharge into the ultimate receiving body. The apportionment of pollution sources is essential for safe reuse of agricultural drained water in order to take some specific remedial measures and should be inherent part of strategic management of water resources (Donia, 2011).

To design monitoring network for the river systems, simply physicochemical and biological analyses of the water quality are insufficient. Assessments involving reasons for the spatial and temporal variations, identification of pollution sources (natural, anthropogenic or mixed ones) and their relative contributions towards monitored parameters, most significant parameters responsible for variations in space and time, relationships between quality parameters and the sampling sites are essential. In this regard, two very useful works – on the Gomti river (India) and the Langat river basin (Malaysia) – provide excellent case studies to employ environmetric techniques in water management. Four multivariate statistical tools, namely, CA, PCA, FA and DA along with the APCS-MLR model were applied to the Gomti river water quality dataset involving thirty one parameters. CA classified sampling locations (based on similarities between them) into three distinct regions – upper catchments (UC), middle catchments (MC) and lower catchments (LC) corresponding to low river pollution, high river pollution and moderate river pollution respectively, which clearly marked the role of the Lucknow city (the capital of the Indian state of Uttar Pradesh) and adjoining areas on the river water quality. DA applied to evaluate spatial and temporal variations in water quality, rendered a considerable dimension reduction of the data. DA was performed using standard, forward stepwise and backward stepwise modes. Forward stepwise mode begins with the most significant environmental variable where they are included step-by-step until no significant changes are obtained. On the other hand, in the backward stepwise mode, monitored variables are removed step-by-step beginning with the less significant one until no significant changes occur (Singh et al., 2005). The temporal DA of the river water quality revealed that only five parameters (most significant), namely, temperature, alkalinity, chloride, sodium and potassium, accounted for most of the variations and, therefore, were
Multivariate statistical techniques and water quality assessment: discourse and review on some analytical models

Kumar Manoj, Pratap Kumar Padhy

International Journal of Environmental Sciences, Volume 5 No.3, 2014

responsible for discriminating three monitored seasons – winter, summer and monsoon. Thus, the anthropogenic pollution, chiefly due to wastewater discharge into the river, was not found to discriminate between the seasons, which showed that it was a regular source throughout the year. Furthermore, the spatial DA showed existence of significant differences between three catchment regions, and identified only ten discriminating parameters, namely, river discharge, pH, biochemical oxygen demand, chloride, fluoride, phosphate, ammonical nitrogen, nitrate, total kjeldahl nitrogen and zinc, responsible for variations in water quality. PCA/FA was applied individually on the each catchment regions recognized. FA performed on the PCs (obtained through PCA) generated seven VFs each for the UC and MC regions and six VFs for the LC region, explaining 74.4%, 73.4% and 81% of the total variance respectively.

A comparative study of the compositional patterns between water samples analyzed is essential for the river basin management as it leads to disclosing factors that influence each one. Application of PCA/FA suggested possible pollution sources in each of the three recognized catchment regions. Overall four major groups of parameters were identified; namely, trace metals group; organic pollution group; nutrient parameters group and; alkalinity, hardness, conductivity and solids group. Presence of metals were attributed to leaching from soil and industrial waste disposal sites; organic pollutants had their origin from point sources like industrial effluents and municipal wastes; nutrients in river water represented non-point sources such as atmospheric deposition and agricultural run-off; and finally, the fourth group corresponded to soil erosion/leaching and the subsequent runoff process. Similarly, for the temporal analysis, PCA/FA identified four groups of parameters which explained maximum seasonal variability, namely – soluble salts, toxic metals, nutrients and organic pollutants. These observations show that PCA/FA, in addition to identifying possible pollution sources/origin of parameters, also identifies parameters having greatest contribution to spatial and temporal variations in water quality. Furthermore, APCS-MLR receptor modeling technique applied to investigate contributions of the identified possible sources to the studied water quality parameters found leaching at the industrial waste disposal sites, disposal of domestic and industrial wastes, agricultural run-off, natural weathering and surface run-off from catchment areas as predominant sources. Additionally, miscellaneous unidentified sources were also attributed by the receptor model to the river water pollution in all the three catchment regions (Singh et al., 2005).

In a similar study conducted on the Langat river basin (Malaysia) to identify the roles played by land use activities on the spatial water quality variations, involving historical data of twenty three water quality parameters, CA classified the river basin into low pollution source (LPS), moderate pollution source (MPS) and high pollution source (HPS) regions on the basis of similar characteristics and natural backgrounds. DA applied to evaluate spatial variations among the three regions rendered significant data reduction. The forward stepwise mode identified six parameters, namely, dissolved oxygen, biochemical oxygen demand, pH, ammonical nitrogen, chloride, and *E. coli* as the most significant parameters discriminating the sampling stations. The backward stepwise mode, on the other hand, identified seven discriminating parameters (having coliform as the seventh ingredient). The most significant parameters responsible for discriminating the spatial or temporal groups also indicate high variation with respect to their spatial or temporal distribution. These parameters then become principal objects of further investigations from a large original dataset. PCA/FA applied individually on the three identified regions resulted in seven VFs for the HPS region and six VFs each for the MPS and LPS regions, explaining 80.8%, 70.9% and 78.7% of the total variance in the data matrix respectively. For the HPS region the major sources of pollution

Kumar Manoj, Pratap Kumar Padhy

International Journal of Environmental Sciences, Volume 5 No.3, 2014

617
recognized by PCA/FA were: discharges of wastewater treatment plants (municipal sewage and sewage treatment plants), domestic wastewater, industrial effluents, runoffs from fields having high load of soil and waste disposal activities, metallic roofs of houses and buildings, and agricultural runoffs like livestock waste and chemical fertilizers. For the MPS region, pollution sources varied from industrial, domestic, municipal, commercial to agricultural runoff. Conversion of forest or agriculture land to urban areas and the associated surface runoff also emerged as recent pollution sources. Lastly, for the LPS region, major identified sources contributing to the river basin pollution were: domestic wastes, agricultural runoff, and runoff from the forest areas, construction activities and village houses with zinc roofs (Juahir et al., 2011). The study showed FA as a powerful statistical tool to discover similarities among variables or samples.

From above, we can conclude that the management of water resources requires designing strategic sampling locations. CA technique performs reliable classification of surface waters considered under the monitoring network. All sampling stations grouped together represent some similar features and backgrounds and need not be sampled each time, only one representative site in each cluster may provide enough information about the water quality of all sites of the monitoring network. It reduces number of sampling sites and, therefore, cost of environmental programmes. CA also provides a visual summary of intra-relationship among the water quality parameters and assists in better understanding of controlling factors (Pejman et al., 2009). Application of CA to classify water quality parameters is available in literatures (Bhattacharyya et al., 2013; Manoj et al., 2013b). Assembling of objects into groups based on their similarities and dissimilarities makes interpretation easier. All objects in a group share similar features which make them different from objects of other groups. Thus, the underlying information about water quality behaviour and pattern is disclosed and recognized. DA identifies most important indicator parameters responsible for variation in water quality of the pre-defined spatial or seasonal groups. These identified parameters account for most of the spatial or temporal variations in the surface water quality. APCS-MLR when applied on these indicator parameters can identify the relative contributions of identified possible sources, worked out by PCA/FA, towards quantification of water quality parameters/pollutants. PCA/FA expediently describes reasons for spatial and temporal variations. Thus, application of multivariate tools in surface water quality management is of immense importance.

The role of PCA/FA is not limited to identify all possible sources of pollution or origin of water quality parameters. These statistical techniques can also recognize essential parameters which affect the chemistry of each spatial or seasonal group of surface water. A parameter contributing significantly to water quality variation in one season, may be less effective or not significant at all for the other one (Pejman et al., 2009; Zare Gariji et al., 2011). The same applies to the water quality data in space. While DA identifies most important parameters contributing to spatial or seasonal variations of water quality, PCA/FA recognizes factors or origins responsible for the variations (Zare Gariji et al., 2011). To recognize major parameters controlling seasonal variations in river water quality is crucial for the river basin management. PCA/FA was utilized to identify and evaluate important seasonal (spring, summer, autumn and winter) water quality parameters in Haraz river basin (Iran). Each parameter with a strong correlation coefficient value (>75%; factor loading >0.75) was defined as a significant contributor to the seasonal variations. The results identified temperature, nitrate, total solids and discharge as the most significant parameters which influenced water quality variations for all four seasons. While nitrate significantly contributed to water quality variations for all four seasons, total phosphate contributed to three seasons.
This showed considerable influx of inorganic nutrients into the river system from the orchard and agricultural areas and sewage almost throughout the year. Biochemical oxygen demand and faecal coliform were the most important parameters causing variations for three seasons (out of four) indicating entry of domestic wastewater into the Haraz river for a considerable time of the year. Presence of total solids with strong factor loadings in all four seasonal groups demonstrated continuous erosion effects resulting from soil cultivation, other high agricultural activities and more decisively river sand mining activities in the river basin (Pejman et al., 2009). The study showed that seasonal variations of water quality parameters must be considered when designing and implementing environmental monitoring plans for the river basin management which is made easier by the multivariate techniques. In a related study, multivariate tools were used to provide functional information for management of water resources at the Chehelchay river watershed scale.

The evaluation of temporal variations of the river pollution showed significant seasonal (spring, summer, fall and winter) changes in water quality. Application of standard DA determined electrical conductivity, chloride and sodium as the most significant quality parameters discriminating four seasons and accounting for most of the seasonal variations. The backward DA mode, on the other hand, suggested five discriminating parameters, namely, conductivity, chloride, bicarbonate, sulphate and total hardness. Thus, out of twelve parameters, DA produced significant data reduction. PCA/FA applied separately for four different seasons revealed that, in addition to the geochemical aspects, seasonal regime of the Chehelchay water quality was controlled by a variety of factors – dilution effect, evapo-transpiration and the base-flow effect, non-point pollution sources such as agricultural activities and floods operated flushing of soil salts. The multivariate investigation led the authors to propose regulation of water extraction from the river for irrigation of rice paddy fields of the floodplains so that concentration of chemicals in the river water during the summer season could be reduced. Furthermore, the authors also recommended establishment of riparian vegetation in the watershed as flushing of soil salts during the fall and winter led to increased river water pollution (Zare Gariji et al., 2011).

Multivariate statistical approaches have also been applied on the wetland systems to obtain better information concerning their quality. The Wetlands are of immense importance having significant ecological values and their study holds vast significance as they are vital factors in the management of the river basins worldwide. CA, PCA and FA were employed to evaluate surface water quality variations of the Uluabat Lake (Turkey), a Ramsar site. Application of CA revealed two different groups of similar water quality characteristics which showed that the aquatic system was affected by different levels of pollution. PCA and FA when performed determined three latent factors controlling 77.4% of the total variance in the data matrix. The first factor, namely, the microbiological factor, revealed intensive sewage discharge into the lake. The second factor was organic-nutrient factor and it showed pollution from domestic waste and nutrients input. The third factor, called physicochemical factor, represented complex physical and chemical reactions between dissolved oxygen, organic matter and nitrogenous products. Pollution load of the Mustafa Kemal Pasa river feeding the Uluabat Lake was identified as the central reason for deteriorating water quality of the latter (Iscen et al., 2008).

4. APCS-MLR receptor modeling technique

The receptor modeling approach for source apportionment of water quality contaminants employs a combination of statistical techniques of MLR and APCS. It can also be called
PCA-MLR (Kothai et al., 2008). The model is developed on the assumption that all possible pollution source components contribute linearly to the final concentration of the contaminant of interest at the receptor site. Here, APCS are used to estimate the contribution of sources to each pollutant (Singh et al., 2008). Since, the PCA is performed on z-transformed standardized variables; the normalized factor scores obtained cannot be implemented directly for quantitative source contributions. In simple term, the PCA in APCS-MLR model can be presented as:

\[
C_{ij} = \sum_{j=1}^{N} L_{ij} \times S_{ij}
\]

(8)

Where, \(C_{ij}\) = standardized concentration of pollutant; \(N\) = total number of pollution sources; \(L_{ij}\) = factor loadings; \(S_{ij}\) = factor scores. These normalized factor scores are rescaled and converted to un-normalized APCS. Details about this technique can be found in some earlier published works (Thurston and Spengler, 1985; Guo et al., 2004; Singh et al., 2005; Singh et al., 2008; Kothai et al., 2008; Pandit et al., 2011; Nasir et al., 2011). The contents in this communication have mostly been derived from some recent works, and references within, of Pandit et al. (2011) and Singh et al. (2008). In the initial step concentration of all water quality variables are standardized as \(Z_{ij}\):

\[
Z_{ij} = \frac{X_{ij} - \bar{X}_j}{\sigma_j}
\]

(9)

Where, \(X_{ij}\) = measured concentration of variable \(j\) in sample \(i\); \(\bar{X}_j\) = arithmetic mean concentration of variable \(j\); \(\sigma_j\) = standard deviation of variable \(j\) for all samples included in analysis. Now, PCA is conducted utilizing these standardized variables which ultimately give rise to normalized factor scores \(A_0\) with magnitudes of zero mean and unit standard deviation. Subsequently, an artificial sample having zero concentration for all the variables is added to the procedure to calculate absolute zero scores for each factor. The basic concept of the former is:

\[
(Z_0)_{ij} = \frac{0 - \bar{X}_j}{\sigma_j} = \frac{-\bar{X}_j}{\sigma_j}
\]

(10)

To obtain absolute zero factor scores \(A_0\) for each sample, the values of corresponding factor scores coefficients \(S\) determined from application of PCA on standardized variables and the values of \(Z_0\) calculated above are combined as:

\[
(A_0)_k = \sum_{j=1}^{J} S_{kj} \times (Z_0)_{ij}
\]

(11)

Where, \(J\) = total number of variables. The APCS for each sample in each component is then estimated by subtracting the values of absolute zero factor scores \(A_0\) of each sample from the corresponding PCA obtained normalized factor score values \(A_0\) of the standardized variables.
Multivariate statistical techniques and water quality assessment: discourse and review on some analytical models

\[ APCS_K = (A_d)_K - (A_o)_K \]  \hspace{1cm} (12)

Where, \( k = 1, 2, \ldots, K \). In the end, employing measured concentration of water quality parameters as dependent variables and mass concentrations of different sources as independent variables a multiple linear regression analysis is performed. The source contribution to \( C_j \) is computed as:

\[ C_j = (r_0)_j + \sum_{k=1}^{F} r_{kj} \times APCS_k \]  \hspace{1cm} (13)

Where, \( C_j \) = pollutant’s concentration, \( (r_0)_j \) = constant term of multiple regression for pollutant \( j \) (average contribution of the \( j^{th} \) pollutant from sources not determined by PCA/FA); \( r_{kj} \) = coefficient of multiple regression of the source \( k \) for pollutant \( j \); \( APCS_k \) = scaled value of the rotated factor \( k \) for the considered sample. The combined term \( r_{kj} \times APCS_k \) represents the contribution of source \( k \) to \( C_j \). Moreover, the mean of the product \( r_{kj} \times APCS_k \) on all samples represents the average contribution of the sources (N). Quantitative contributions from each identified source for individual water quality contaminant are compared with their original measured values (Singh et al., 2008; Pandit et al., 2011). The values for \( C_j, (r_0)_j, \) and \( r_{kj} \) have the dimensions of the original concentration measurements (Singh et al., 2005).

The applicability of APCS-MLR model in water quality characterization and classification is a relatively new concept as compared to its application in air quality assessment and monitoring studies. A few researches that utilized its potential in hydrological examinations systematically showed its remarkable role in water quality evaluation. Appraisal of contributions of all possible pollution sources to the quantity of aquatic pollutants is indeed an essential requirement for strategic design of management of scarce and precious freshwater resources. Dissemination of role of APCS-MLR receptor model in water management is recommendable. Therefore, we strongly advocate its employment in high quality water studies to the water researchers.

5. Conclusions

Physicochemical and biological components continuously interact with the environment. Their complex relationship cannot be deduced simply on the basis of concentration analysis and simple statistical tools. Multivariate statistical techniques are important and exploratory data analysis methods for enhanced interpretation of information within large and complex data matrix. They have a vast application in monitoring, assessment and consequently sustainable management of the surface water resources, especially the river basins. They suggest quick possible solutions to pollution hazards and present means to select the most appropriate pollution control schemes and policies. CA provides information about optimal sampling strategy and river monitoring network design. PCA/FA discloses most important parameters responsible for variation in the dataset and also identifies origin of pollutants and all possible water pollution sources. DA identifies specific parameters causing differences between groups in space and time. MLR evaluates relationships between water quality parameters to expose hidden meaningful information. APCS-MLR model does pollution apportionment and reveals contribution of all pollution sources in the quantification of each pollutant. Thus, a systematic approach employing a combination of multivariate statistical

Kumar Manoj, Pratap Kumar Padhy
International Journal of Environmental Sciences, Volume 5 No.3, 2014

621
Multivariate statistical techniques and water quality assessment: discourse and review on some analytical models

tools is essential for developing and designing suitable water management strategies. Multivariate techniques are equally effective in evaluating reasons for variations in both spatial and temporal studies.

Acknowledgement

The first author is thankful to the Council of Scientific and Industrial Research (CSIR), India, for providing Junior and Senior Research Fellowships.

6. References


