Application of chemometric techniques in the assessment of groundwater pollution in a suburban area of Chennai city, India

A. Ramesh Kumar[1] and P. Riyazuddin[2,*]

[1]Chemical Laboratory, Central Groundwater Board, South Eastern Coastal Region, E1, Rajaji Bhavan, Besant Nager, Chennai 600 090, India
[2]Department of Analytical Chemistry, University of Madras, Guindy Campus, Chennai 600 025, India

Chemometric techniques such as factor analysis (FA), cluster analysis (CA) and discriminant analysis (DA) were applied to the groundwater quality data in a tannery-polluted area of Chennai city, India. Groundwater samples were collected from 65 dug wells during pre- and post-monsoon seasons and analysed for 25 parameters. FA applied to the datasets pertaining to the pre- and post-monsoon seasons resulted in eight and nine factors explaining 78.7 and 79.7% of the total variance of the respective datasets. Though FA identified two polluting processes (major ion pollution and tannery pollution factors) explained by factors 1 and 2, it could not explain the remaining factors. Three major clusters, i.e. unpolluted, moderately polluted and severely polluted groups, were obtained through CA on the basis of similarities between them. But during the post-monsoon season, the clustering of unpolluted and moderately polluted groups was not clear. Spatial DA by standard mode classified the cases into three groups with 95.4 and 87.7% correct assignations for the two seasons. DA by stepwise mode suggested that electrical conductivity was the discriminating variable with 69.2% correct assignation of cases. FA identified temporal changes in water quality, due to groundwater recharge after monsoon. Changes in water quality were mainly attributed to reduction of pollution load of tannery pollution factor. However, little changes in the major ion pollution factor were observed. DA by stepwise mode predicted that Mg, SiO_2, pH and Cd were the most important parameters to discriminate the two seasons.

Keywords: Cluster analysis, discriminant analysis, factor analysis, groundwater pollution.

THE chemical composition of groundwater is controlled by many factors, including composition of precipitation, mineralogy of the aquifers, climate, topography and anthropogenic activities. These factors combine to create diverse water composition that varies temporally and spatially. Additionally, these factors may lead to contamination of groundwater with diverse constituents, resulting in severe environmental and socio-economic problems. Hydrochemical characterization of such systems requires many variables to be determined that would result in large datasets. The use of graphical methods such as Piper, Stiff, Schoeller and Collins diagrams to interpret such datasets becomes inadequate, because these methods consider only major ionic constituents. Constituents such as fluoride, silica and trace metals cannot be included; hence, these graphical methods are unsuitable for many environmental studies[1,2].

On the other hand, multivariate treatment of environmental data is widely used to characterize and evaluate water quality, and is useful for evidencing temporal and spatial variations caused by natural and anthropogenic factors[3–9]. Multivariate techniques such as factor analysis (FA), cluster analysis (CA) and discriminant analysis (DA) are powerful techniques to identify the underlying processes that control groundwater chemistry, grouping of samples of similar composition and origin, and to predict the variables that differentiate the sampling stations temporally and spatially. These techniques have widely been used as unbiased methods in the analysis of groundwater quality data to characterize groundwater composition influenced by natural and anthropogenic factors[2,10–18]. In the present article, the large dataset obtained from the hydrochemical analysis of groundwater samples collected from a sub-urban area of Chennai city, India has been subjected to FA, CA and DA, with the objective of assessment of: (a) groundwater pollution and its temporal and spatial variation; (b) hidden factors explaining the various processes/sources and (c) grouping of sampling stations according to the extent of pollution. Finally, the efficiency of these three chemometric techniques to afford data reduction, identification of sources/processes of groundwater pollution and classification of sampling stations on the basis of compositional differences and their origin has been evaluated.

Review of multivariate statistical techniques

A summary of the chemometric techniques used in this study is given here. More details on theoretical aspects[19–25] and

*For correspondence. (e-mail: riyazdr@yahoo.co.uk)

on environmental data analysis[26–28] are published elsewhere.

Factor analysis

The purpose of FA is to reduce the analytical data of each sample, which are intercorrelated to a smaller set of 'factors' that are then interpretable. The factors group correlated concentrations together and they can be associated directly or indirectly with some specific source or process. The method consists of three steps, namely data standardization, factor extraction and rotation of factor axes.

Prior to analysis, the initial data were standardized by *z*-scale transformation as

$$z = \frac{x_{ji} - \bar{x}_j}{s_j},$$

where x_{ji} indicates the original value of the measured parameter, $\bar{x}_j$ the average value of the parameter j and s_j the standard deviation of j. FA takes data contained in a correlation matrix and rearranges them in a manner that better explains the structure of the underlying system that produced the data. The starting point of FA is to generate a new group of variables from the initial dataset (the so-called factors) that are a linear combination of the original variables. The principal components (PC) extraction has been used in this procedure. The first factor obtained explains the biggest part of the variance. The following factors explain repeatedly smaller parts of the variance. Factor loadings show how the factors characterize the variables. High factor loadings (close to 1 or –1) indicate strong relationship (positive or negative) between the variable and the factor describing the variable. Then the factor loadings matrix is rotated to an orthogonal simple structure according to the varimax rotation technique. The result of this operation is high factor loadings (close to 1 or –1) obtained for the variables correlated in the factor and low factor loadings (close to 0) obtained for the remaining variables. In order to determine the number of factors to be retained, the Kaiser criterion is followed. The factors, which best describe the variance of the analysed data (eigen value >1) and can be reasonably interpreted, are accepted for further analysis. The measure of how well the variance of a particular variable is described by a particular set of factors is called 'communality'. Finally, factor scores are calculated for each sample and plotted as a scatter diagram. Extreme positive factor scores (>+1) reflect sampling stations most affected by the process and extreme negative (<–1) scores reflect those unaffected by the process explained by the factor. Near-zero scores reflect sampling stations affected to an average degree by the process.

Ideally, if a FA is successful, the number of factors will be small, communalities are high (close to 1) and the factors will be readily interpretable in terms of particular sources or process.

The disadvantage of FA is the difficulty of distinguishing the processes which cause similar differentiation of groundwater chemistry. Moreover, a priori knowledge of various processes affecting the sampling stations is required for successful application of FA.

Cluster analysis

CA is an unsupervised pattern recognition technique that uncovers intrinsic structure or underlying behaviour of a dataset without making a priori assumption about the data, in order to classify the objects of the system into clusters based on their similarities. The main aim of CA is grouping objects (sampling stations) into classes (clusters), so that objects within a class are similar to each other but different from those in other classes. There are two major categories of CA: hierarchical and non-hierarchical. Hierarchical cluster analysis (HCA) is the most common approach in which clusters are formed sequentially, starting with the most similar pair of objects and forming higher clusters step by step. The process of forming and joining clusters is repeated until a single cluster containing all samples is obtained. The result can be displayed as a dendrogram and provides a visual summary of the clustering process, presenting a picture of the groups and its proximity with a dramatic reduction in dimensionality of the original data.

Here we perform CA using Ward's method on the normalized dataset. This method uses an analysis of variance approach to evaluate the distances between clusters, attempting to minimize the sum of squares of any two (hypothetical) clusters that can be formed at each step. The squared Euclidean distance usually gives the similarities between two samples and a distance can be represented by the 'difference' between analytical values from both the samples.

Though CA is simple and easy to perform, there are certain disadvantages. Similarity/dissimilarity measurements and linkage methods used for clustering greatly affect the outcome of the HCA results. Furthermore, interpretation of the dendrogram is a subjective evaluation and it does not give information about the distribution of the chemical constituents that form each group[1].

Discriminant analysis

DA is used to determine the variables that discriminate between two or more naturally occurring groups. DA builds up a discriminant function for each group as in equation (1):

$$f(G_i) = k_i + \sum_{j=1}^{n} w_{ij} p_{ij}, \quad (1)$$

where i is the number of groups (G), k_i the constant inherent to each group, n the number of parameters used to classify a set of data into a given group, and w_j the weight coefficient assigned by DA to a given selected parameter (p_j). The efficiency of these discriminant functions can then be checked with the same dataset using the cross validation method, which refers to the process of assessing the predictive accuracy of the model in a test sample relative to its predictive accuracy in the learning sample from which the model was developed. If the model performs as well in the test sample as in the learning sample, it is said to cross validate well, or simply to cross validate.

DA was applied to the raw dataset using the standard and stepwise modes to construct discriminant functions (DFs) to evaluate spatial and temporal variations in water quality. The site (spatial) and season (temporal) were the grouping (dependent) variables, while the measured parameters constituted the independent variables. In this study, two groups for temporal (pre- and post-monsoon seasons) and three groups for spatial (depending on the sampling stations) evaluation have been selected.

Materials and methods

Study area

The study area is located in Kancheepuram District, Tamil Nadu, and adjoins Chennai city (Figure 1). It has an aerial extent of 45 km^2, which includes several sub-urban places, among which Chrompet is known for its tanneries. A cluster of 152 tanneries is located at Chrompet. Though they were away from residential areas when the tanneries came up nearly a century ago, now they have become part of the city and are densely populated. The tanneries generate about 3000 m^3/day of effluents, which are treated at the common effluent treatment plant (CETP) and are finally disposed into the nearby Adyar river. Prior to the establishment of the CETP in 1995, the untreated effluents were disposed into the existing tanks and river through open sewerage system. In addition, the disposal of solid and liquid domestic waste into the tanks through open sewerage system causes severe groundwater pollution.

Geologically, the area is underlain by charnokite rocks of archean age. Groundwater occurs under un-confined to semi-confined conditions. The depth of sampled wells ranged from 3.32 to 14.15 m below ground level (bgl). The depth to water level ranged from 0.70 to 13.77 m bgl and 0.38 to 8.56 m bgl during pre- and post-monsoon seasons.

Sampling and analytical procedures

The sampling network and strategy were designed to cover a wide range of determinants at key sites, which reasonably represent the groundwater quality in the study area. In this study, the representative sampling sites were chosen in order to cover various anthropogenic activities, including waste disposal. The gathered background information provides sufficient details on these aspects.

Groundwater samples were collected from 65 dug wells in June 2004 and January 2005, representing pre- and post-monsoon seasons. Grab samples were collected at 30 cm

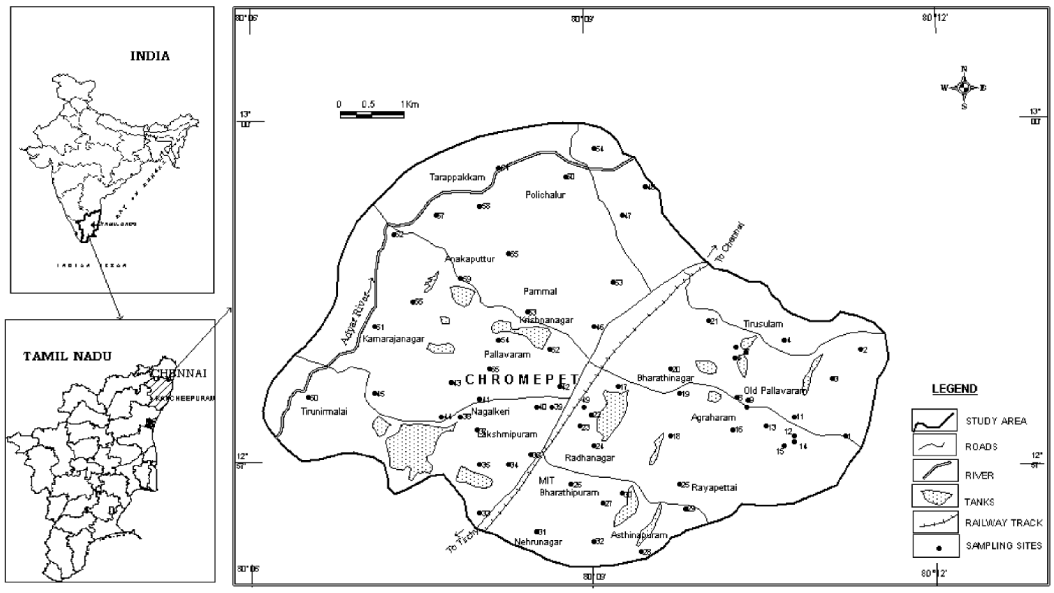

Figure 1. Location map of the study area showing sampling sites and other features.

Table 1. Details of analytical methodology and basic statistics of groundwater samples collected from the study area

Variable	Symbol	Method	Units	Detection limit	Pre-monsoon				Post-monsoon			
					Minimum	Maximum	Mean	SD	Minimum	Maximum	Mean	SD
pH	pH	Potentiometry	pH	–	6.96	8.18	7.59	0.26	6.5	8.20	7.38	0.32
Electrical conductivity	EC	Electrolytic	μmhos/cm	–	605	11660	3171	1851	600	10260	2677	1585
Total hardness (as $CaCO_3$)	TH	EDTA titrimetric	mg/l	5	135	3580	964	583	176	2490	800	446
Calcium	Ca	EDTA titrimetric	mg/l	2	18	776	193	130	16	563	177	105
Magnesium	Mg	by difference	mg/l	1	22	399	117	74	13	263	86	53
Sodium	Na	Flame photometry	mg/l	1	41	2600	343	334	31	2050	274	265
Potassium	K	Flame photometry	mg/l	1	BDL	170	12	25	BDL	150	10	20
Silica	SiO_2	Spectrophotometry	mg/l	5	13	92	53	14	21	118	67	21
Acidity as $CaCO_3$	Acidity	Titrimetry	mg/l	5	5	60	31	13	5	285	43	36
Bicarbonate	HCO_3	Titrimetry	mg/l	6	73	650	376	126	27	902	397	164
Chloride	Cl	Argentometry	mg/l	4	71	3205	693	579	57	2921	551	449
Sulphate	SO_4	Nephlo–turbidimetry	mg/l	5	46	1606	306	294	28	1090	201	196
Nitrite nitrogen	NO_2	Spectrophotometry	mg/l	0.05	BDL	11.75	0.39	1.98	BDL	8.52	0.27	1.39
Nitrate	NO_3	Spectrophotometry	mg/l	0.05	2	452	133	116	0.26	381	105	92
Fluoride	F	SPADNS	mg/l	0.1	0.23	1.48	0.67	0.23	0.13	1.25	0.60	0.23
Phosphate phosphorus	PO_4–P	Spectrophotometry	mg/l	0.05	BDL	0.30	0.07	0.05	BDL	0.41	0.06	0.09
Total dissolved solids	TDS	Gravimetry	mg/l	10	400	8170	2164	1263	380	7190	1832	1115
Copper	Cu	ET–AAS	mg/l	0.005	BDL	0.29	0.02	0.05	BDL	0.02	0.00	0.00
Cadmium	Cd	ET–AAS	mg/l	0.005	BDL	0.04	0.00	0.00	BDL	0.01	0.00	0.00
Iron	Fe	ET–AAS	mg/l	0.010	BDL	7.88	0.39	1.03	BDL	1.92	0.12	0.33
Lead	Pb	ET–AAS	mg/l	0.020	BDL	0.22	0.02	0.04	BDL	BDL	0.07	0.02
Chromium	Cr	ET–AAS	mg/l	0.005	BDL	0.29	0.02	0.04	BDL	0.25	0.02	0.04
Zinc	Zn	ET–AAS	mg/l	0.005	BDL	7.99	0.19	0.99	BDL	2.50	0.06	0.30
Manganese	Mn	ET–AAS	mg/l	0.010	BDL	2.05	0.22	0.40	BDL	0.85	0.09	0.17
Nickel	Ni	ET–AAS	mg/l	0.010	BDL	0.36	0.01	0.04	BDL	0.15	0.01	0.02

SD, Standard deviation; BDL, Below detection limit.

below the water level using a water sampler. Samples for major ions and other inorganics were collected in 1 l precleaned polypropylene bottles. Samples for trace metal analysis were collected separately, filtered at site using 0.45 μm membrane filter and acidified to pH < 2 using ultrapure grade HNO_3. The samples were immediately transported to the laboratory under low-temperature conditions in ice-boxes and stored in the laboratory at 4°C until analysis. All the samples were analysed for 25 parameters according to the standard methods of APHA–AWWA–WEF[29] and were completed within a month. Details of analytical methodology followed are given in Table 1.

Hydrochemistry of major ions

The ionic composition of the groundwater is dominated by major cations (Ca^{2+}, Mg^{2+}, Na^+ and K^+) and anions (HCO_3^-, Cl^-, SO_4^{2-} and NO_3^-). Plotting of major ionic constituents in the Piper diagram (not shown) indicated that 61% of samples is $CaCl_2$-type, 31% of NaCl-type and about 8% of mixed-type during pre-monsoon season. During post-monsoon season 72% is of $CaCl_2$-type, 23% of NaCl-type and about 5% of $CaHCO_3$-type. These show that the recharge of groundwater after monsoon decreased the concentration of major ions of Na^+, Cl^- and SO_4^{2-}. On the basis of compositional differences and field observations, the samples have been divided into three groups, i.e. unpolluted (group 1), moderately polluted (group 2) and severely polluted (group 3).

Data treatment and chemometric analysis

Chemometric analysis of the data was performed using FA, CA and DA techniques. FA and CA were performed on standardized (*z*-scale transformation) experimental datasets in order to avoid misclassification due to wide differences in data dimensionality. The *z*-scale transformation renders the data normalized with mean and variance of zero and one respectively. Standardization tends to increase the influence of variables whose variance is small and reduce the influence of variables whose variance is large. Furthermore, standardization procedure eliminates the influence of different units of measurement and renders the data dimensionless. All the statistical computations were made using SPSS 10.1 software. Basic statistics of the hydrochemical variables of the groundwater samples are shown in Table 1.

Results and discussion

Factor analysis

The Bartlett's sphericity test carried out on the correlation matrix shows a calculated $\chi^2 = 2495.9$ and 2356.4 for the pre- and post-monsoon seasons respectively, which is greater than the critical value $\chi^2 = 387.3$ ($P = 0.0005$ and 300 degrees of freedom), thus proving that the PC extraction can achieve a significant reduction of the dimensionality of the original dataset.

FA was applied separately to the hydrochemical dataset pertaining to pre- and post-monsoon seasons. Tables 2 and 3 summarize the sorted FA results, including the variable loadings, eigen values and variance explained by each factor for the two seasons. The factor loadings were sorted according to the criteria of Liu *et al.*[16], i.e. strong, moderate and weak, corresponding to absolute loading values of >0.75, 0.75–0.50 and 0.50–0.30 respectively. Loading values <0.30 are insignificant and not shown here. The communalities of the variables for the two seasons are given in Table 4.

FA rendered eight significant factors (eigen value >1) explaining 78.7% of the total variance of the pre-monsoon dataset. Factor 1 explains 19.9% of the variance and is characterized by strong positive loadings (>0.90) of TH and Ca, and strong loadings by Mg, EC, TDS and Cl. Na and SO_4 show weak loadings. Factor 2 explains 17.8% of the variance and has strong loadings of Na and moderate loadings of EC, TDS, SO_4, Cu, Cr, Cd and Cl, and weak loadings of HCO_3, Mg and acidity. These two factors explaining almost equal variance individually, account for 37.7% of the total variance. Considerable overlapping of variables (i.e. EC, TDS, Cl, SO_4 and Na) is observed. Hence, the underlying processes explaining these two factors are mixed. Further, major ionic constituents that are highly correlated to EC and TDS mainly contribute to factor 1. Hence, factor 1 may be termed as the 'major ion pollution factor'. Factor 2 is contributed by Na, Cl, SO_4 and trace metals such as Cd, Cr and Cu. The sources of these trace metals, especially chromium, are the tanneries located in the area; hence factor 2 could be termed as the 'tannery pollution factor'. The sources of major ionic constituents are the poor domestic sewerage system and the tanneries; hence, factors 1 and 2 could be collectively called as pollution factors.

The factors 3–8 account for 41% of the variance of the dataset; however, the variable loadings of factors 3–8 are not clear (Table 2). Hence the possible sources associated with these factors could not be explained.

FA of the post-monsoon data rendered nine significant (eigen value >1) factors explaining 79.7% of the total variance (Table 3). Factor 1 (major ion pollution factor) explains 21% of the variance and has strong loadings of TH, Ca, TDS, Mg, EC and Cl. Similar to the pre-monsoon season, Na and SO_4 showed weak loadings. Factor 2 (tannery pollution factor) explains 12.7% of variance and has strong loadings of Na and Cr, moderate loadings of SO_4, TDS and EC, and weak loadings of Cl and Cu. The variance explained by the two factors accounts for 33.9% of the total variance. Similar to pre-monsoon season, overlapping of variables is observed. Also, the variable

Table 2. *R*-mode varimax rotated factor loadings of water quality parameters during pre-monsoon season

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8
TH	**0.952**							
Ca	**0.922**							
Mg	**0.827**	0.356						
Cl	**0.814**	0.510						
EC	**0.759**	0.612						
TDS	0.749	0.620						
SO_4	0.474	0.700						
Na	0.358	**0.861**						
Cu		0.744		0.354				
Cr		0.735						
Cd		0.661		0.488				
HCO_3		0.437				0.744		
Acidity		0.334				0.623	0.304	
Fe			**0.964**					
Zn			**0.961**					
Pb				0.722				
NO_2–N				0.721				0.311
F					**0.866**			
Mn					**0.858**			
PO_4–P						0.723		
pH							–0.807	
Silica							**0.776**	
NO_3								**0.765**
K								0.553
Ni								0.684
Eigen values	4.983	4.456	2.066	1.696	1.655	1.651	1.590	1.587
% Variance	19.9	17.8	8.3	6.8	6.6	6.6	6.4	6.3
Cumulative (%)	19.9	37.7	46.0	52.8	59.4	66.0	72.4	78.7

Table 3. *R*-mode varimax rotated factor loadings of water quality parameters during post-monsoon season

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8	Factor 9
TH	**0.963**								
Ca	**0.918**								
Mg	**0.846**								
Cl	**0.836**	0.492							
TDS	**0.807**	0.546							
EC	**0.806**	0.546							
SO_4	0.450	0.679				0.398			
Na	0.453	**0.821**							
Cr	**0.805**								
Cu	0.377		0.610						
Fe			**0.898**						
Zn			**0.874**						
Acidity				0.736					
HCO_3				0.719					
Silica				0.433			0.666		
F				0.393					
Pb				–0.378				0.390	–0.394
pH					0.643				
NO_3					0.628			–0.483	
F					–0.616	0.310			
NO_2–N					0.600				
Ni						0.732			
Cd						0.728	–0.305	–0.330	
K							**–0.783**		
Mn								**0.820**	
PO_4–P									**0.906**
Eigen values	5.310	3.173	2.074	1.973	1.797	1.611	1.436	1.384	1.173
%Variance	21.2	12.7	8.3	7.9	7.2	6.4	5.7	5.5	4.7
Cumulative (%)	21.2	33.9	42.2	50.1	57.3	63.7	69.5	75.0	79.7

loadings of factors 3–9 are not clear, though they account for 45.6% of the total variance.

Comparison of FA for the two seasons shows the effect of groundwater recharge caused by monsoon on the two processes associated with the two factors. It appears that the major ion pollution factor shows little change during the post-monsoon season, but there is considerable reduction in pollution load caused by the tannery pollution factor. This is evidenced by the substantial reduction in the concentration of trace metals notably, Cd, Cr and Cu, and major ionic constituents such as Na, Cl and SO_4 during the post-monsoon season. The loading pattern of factors 3–8 during pre-monsoon season and factors 3–9 during post-monsoon season is not clear and indicates the absence of correlation with other variables. Variables which are known to be geogenic, i.e. F, silica, K and HCO_3 are grouped together with that of anthropogenic variables, i.e. Pb, Zn, Cu, NO_3 and NO_2. Hence the processes or sources associated with these factors are highly localized and contributed by geogenic and anthropogenic sources.

The factor score plots of the first two factors for the pre- and post-monsoon seasons are shown in Figure 2 *a* and *b*, respectively. Comparison of the factor score plots for the two seasons shows the effect of dilution caused by recharge on the hydrochemical variables. The score plots for the two seasons show almost the same grouping of samples. The samples affected by the two factors (factor score >1) are well identified for the two seasons. During the pre-monsoon season, most of the samples are clustered around the origin, indicating contamination by the two processes to an average extent. Only a few samples were not affected by the two processes and have high negative scores (>–1). The clustering of samples around the origin is less pronounced during the post-monsoon season, indicating the effect of dilution caused by rainfall.

Table 4. Communalities of variables of the factor model

Variable	Pre-monsoon	Post-monsoon
EC	0.979	0.990
TDS	0.977	0.990
TH	0.968	0.974
Fe	0.943	0.821
Zn	0.933	0.812
Cl	0.932	0.951
Na	0.917	0.915
Ca	0.906	0.886
Mg	0.869	0.853
Mn	0.862	0.737
HCO_3	0.823	0.728
pH	0.792	0.611
F	0.788	0.772
Acidity	0.757	0.686
SO_4	0.755	0.847
Cd	0.737	0.752
NO_2	0.725	0.538
Cu	0.724	0.717
NO_3	0.711	0.753
PO_4–P	0.703	0.859
Silica	0.656	0.801
Pb	0.601	0.738
Cr	0.618	0.715
Ni	0.591	0.754
K	0.414	0.732

It is interesting to note the unrotated principal component analysis (PCA) loadings, in that the pollution factors explained by factors 1 and 2 are merged to a single PC. During pre-monsoon season, PC 1 is contributed by strong loadings of EC, TDS, Cl, TH, Mg, Na and SO_4, and moderate loadings of Ca, acidity, Cu, Cd and Cr, and explains 31% of variance. On the other hand, moderate loadings of Ca and moderate negative loading of HCO_3 characterize PC 2, and other variables show weak loadings and explain only 9% of variance. During post-monsoon season, PC 1 accounted for 29.8% of total variance with strong loadings of EC, Cl, TDS, Na, SO_4, TH, Ca and Mg, moderate loadings of HCO_3 and Cr, and weak loadings of Ni, acidity and Cu. PC 2 was characterized by moderate positive loading of NO_3, weak loadings of Fe, Pb and Zn, and explained 9.5% of the total variance. Thus varimax rotation refined the PC model by separating the two pollution processes explained by factors 1 and 2. A similar result of FA has been reported by Helena *et al.*[13].

Cluster analysis

CA was performed on the standardized (*z*-scale) dataset for the two seasons separately by Ward's method using squared Euclidean distance as similarity measure. The resulting dendrogram is shown in Figure 3 *a* and *b*. CA was also performed on variables for the two seasons separately to find out the grouping of variables. The dendrogram for pre-monsoon season is shown in Figure 4. Grouping of variables was similar for the two seasons. However, CA performed on the combined dataset for both seasons, with the objective of classifying the sampling stations based on temporal variation was unsuccessful.

The dendrograms for the two seasons (Figure 3 *a* and *b*) consist of several groups, and each group consists of several sub-groups and singletons. However, for the sake of interpretation, it could be classified into three major clusters, as shown in Table 5. During the pre-monsoon season, the three clusters consisted of 23, 29 and 13 members. About 56% of the cluster I members belonged to the unpolluted group, cluster II was highly mixed comprising 41 and 59% of moderately and severely polluted samples and cluster III comprised 100% of severely polluted samples. During the post-monsoon season clusters I and II were mixed; however, 95% of the cluster III members were from severely polluted samples. It appears that the grouping of cases in CA is mainly based on the major ionic characteristics.

CA performed on variables indicates the groups of variables that behave similarly and/or have similar origin

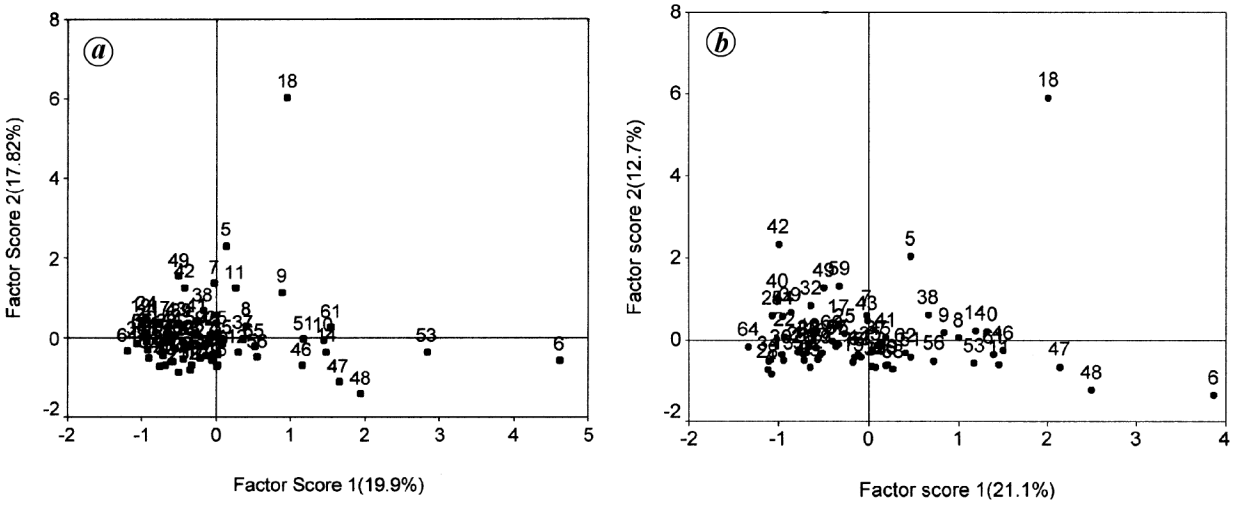

Figure 2. Factor score plot of factors 1 and 2 for (***a***) pre-monsoon and (***b***) post-monsoon seasons.

Table 5. Clustering of cases during pre- and post-monsoon seasons

Cluster	Case	Total no.	Group membership[a]		
			1	2	3
Pre-monsoon					
I	13–2, 22–16, 33–62, 52–28	23	13	7	3
II	41–49, 38–42, 12–46, 36–17, 23–24, 25–65, 26–35, 48,3, 32, 39, 60	29	–	12	17
III	5–18, 51–14, 53, 10–47, 9–7, 6, 8	13	–	–	13
Post-monsoon					
I	45–19, 3–35, 33–23, 15–16, 26–50, 17	21	4	10	7
II	21–55, 31, 54–20, 59–40, 4–64, 28, 65	23	9	8	6
III	5–32, 11, 7–61, 14, 46–56, 9–48, 47, 6, 42–49, 51–53, 39, 8–18	21	–	1	20

[a]Number of samples belonging to the original classification based on sampling location.

(Figure 4). The dendrogram shows two major clusters, one including major ionic constituents such as Cl, Na, SO_4, TH, Ca, Mg plus EC and TDS, and the other including trace metals and constituents such as acidity, HCO_3, NO_2, NO_3, pH, silica and F. The grouping of variables was similar for the two seasons; however, the similarity levels of the variables were changed in the post-monsoon season. The grouping pattern predicted by CA is consistent with the factor loading pattern of the two seasons predicted by FA.

Discriminant analysis

DA was performed on the raw dataset comprised of 25 parameters for the two seasons separately. Temporal DA was performed on the combined dataset after grouping the cases season-wise (pre-monsoon – group 1 and post-monsoon – group 2). The standard mode for building discriminant function coefficients based on entering all the variables was used. In the stepwise mode variables were included step-by-step beginning with the more significant, until no significant changes were obtained.

The standard mode DA formed two DFs using all variables except Mg and Mn, and Mg for the pre- and post-monsoon seasons respectively, as they failed the tolerance test. The DFs were tested using Wilks' lambda and chi-squared tests. The classification result showed that 95.4 and 87.7% of the original grouped cases correctly classified during the pre- and post-monsoon seasons. The stepwise mode DA showed that 69.2% of the original

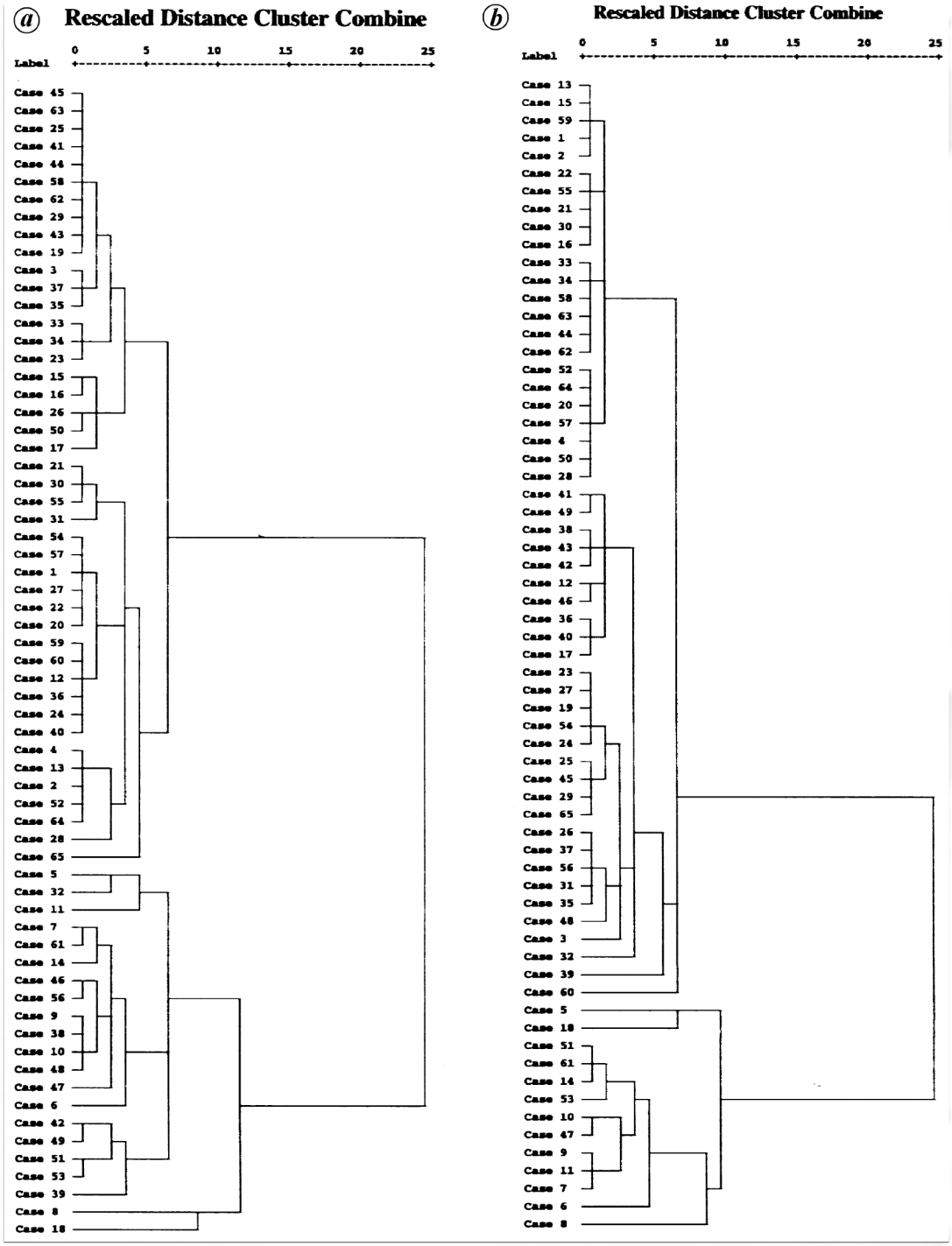

Figure 3. Dendrogram showing grouping of sampling stations during (***a***) pre-monsoon and (***b***) post-monsoon seasons.

grouped cases correctly classified for both seasons utilizing the variable EC only. Thus DA suggests that EC is the most important parameter to discriminate between the three groups of sampling stations. The scatter plot of DFs 1 and 2 for the two seasons is shown in Figures 5 *a* and *b*. Among the three groups, group 3 is clearly separated from group 1, whereas little overlap is seen between groups 1 and 2, and 2 and 3.

DA performed on the combined dataset (temporal DA) using the standard mode utilized 25 variables and classi-

fied the cases with 79.2% right assignations. The stepwise mode DA utilizing variables Mg, pH, SiO_2 and Cd, gave 72.3% right assignations, suggesting that these variables are important to discriminate between pre- and post-monsoon seasons.

Conclusion

Groundwater quality dataset of a tannery-polluted suburban area of Chennai city, India was analysed using chemometric techniques (FA, HCA and DA) for variations in compositional differences, temporal and spatial variations caused by anthropogenic factors. FA identified two polluting processes, namely major ion pollution factor and tannery pollution factor, responsible for groundwater pollution in the area. The polluting processes associated with factors 3–8 and factors 3–9 during pre- and post-monsoon seasons could not be identified because variable loadings of these factors are not clear. FA predicted that temporal changes in water quality are due to the reduction of pollution load caused by the tannery pollution factor. However, FA did not give considerable data reduction, because it requires 21 and 16 variables (factor loadings >0.7), which correspond to 16 and 36% data reduction to explain 78.7 and 79.7% variances for pre- and post-monsoon seasons respectively. Further, it could not differentiate between the unpolluted and moderately polluted stations clearly. CA resulted in three major clusters, which correspond to the three groups of sampling stations. Though it formed well-defined clusters for unpolluted and severely polluted groups, grouping of cases into the moderately polluted group is not clear, particularly for the post-monsoon season. Similarly, temporal CA performed on the combined dataset was unsuccessful. CA performed on variables resulted in two major clusters, one comprising major ionic constituents and the other comprising minor constituents and trace metals. DA rendered best results for both temporal and spatial analysis.

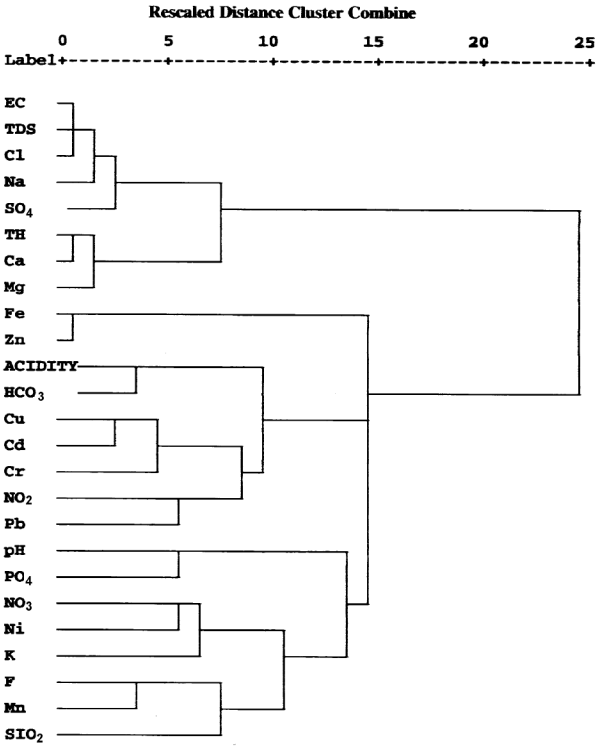

Figure 4. Dendrogram showing grouping of variables during pre-monsoon season.

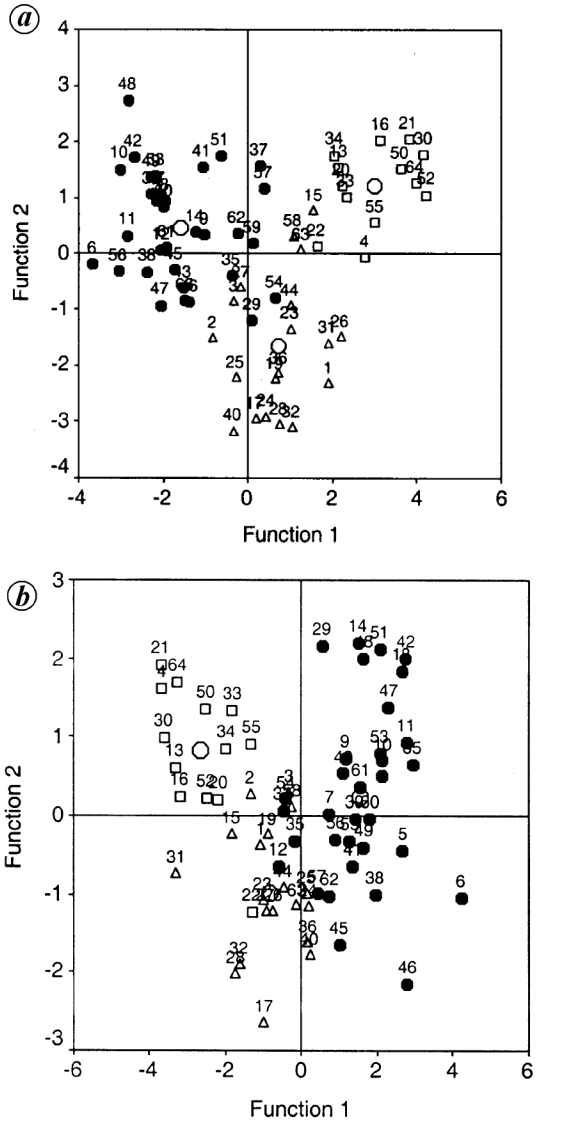

Figure 5. Bivariate plot of discriminant functions 1 and 2 during (***a***) pre-monsoon and (***b***) post-monsoon seasons; □, Group 1 unpolluted; △, Group 2 (moderately polluted); ●, Group 3 (severely polluted); ○, Group centroids.

Spatial DA by standard mode gave 95.4 and 87.7% correct assignations of the classified groups using two DFs. Stepwise-mode DA assigned 69.2% cases correctly using the variable EC only. Temporal DA performed on the combined dataset indicates that Mg, pH, SiO_2 and Cd are important discriminating parameters and assigned 72.3% cases correctly. This study reveals that FA is more effective in identifying the compositional differences of water-quality data and, DA is more effective in grouping the sampling stations based on the extent of pollution and its spatial and temporal variations.

The findings of the study indicate the need for proper industrial planning and the safe disposal of industrial and urban waste, which would otherwise lead to severe environmental degradation. The persistence of heavy metals in the groundwater indicates that contaminated aquifers require several years to flush out the contaminants naturally. Though several 'pump and treat' techniques are available to make the water fit for its intended use, aquifer remediation techniques[30,31] are suitable for this type of small area.

1. Guler, C., Thyne, G. D., McCray, J. E. and Turner, A. K., Evaluation of graphical and multivariate statistical methods for classification of water chemistry data. *Hydrogeol. J.*, 2002, **10**, 455–474.
2. Lambrakis, N., Antonakos, A. and Panagopoulos, G., The use of multicomponent statistical analysis in hydrogeological environmental research. *Water Res.*, 2004, **38**, 1862–1872.
3. Simeonov, V., Stefanov, S. and Tsakovski, S., Environmental treatment of water quality survey data from Yantra river, Bulgaria. *Microchim. Acta*, 2000, **134**, 15–21.
4. Ouyang, Y., Evaluation of river water quality monitoring stations by principal component analysis. *Water Res.*, 2005, **39**, 2621–2635.
5. Singh, K. P., Malik, A. and Sinha, S., Water quality assessment and apportionment of pollution sources of Gomati river (India) using multivariate statistical techniques – A case study. *Anal. Chim. Acta*, 2005, **538**, 355–374.
6. Kowalkowski, T., Szpejna, Z. and Buszewski, B., Application of chemometrics in river water classification. *Water Res.*, 2006, **40**, 744–752.
7. Ouyang, Y., Nkedi-Kizza, P., Wu, Q. T., Shinde, D. and Huang, C. H., Assessment of seasonal variations in surface water quality. *Water Res.*, 2006, **40**, 3800–3810.
8. Panda, U. C., Sundaray, S. K., Rath, P., Nayak, B. and Bhatta, D., Application of factor and cluster analysis for characterization of river and estuarine water systems – A case study: Mahanadi river (India). *J. Hydrol.*, 2006, **331**, 434–445.
9. Kennel, P. R., Seockheon, L., Kanel, S. R. and Khan, S. P., Chemometric application in classification and assessment of monitoring locations of an urban river system. *Anal. Chim. Acta*, 2007, **582**, 390–399.
10. Ruiz, F., Gomis, V. and Blasco, P., Application of factor analysis to the hydrogeochemical study of a coastal aquifer. *J. Hydrol.*, 1990, **119**, 169–177.
11. Melloul, A. and Collin, M., The principal components statistical method as a complementary approach to geochemical methods in water quality factor identification; application to the coastal plain aquifer of Israel. *J. Hydrol.*, 1992, **140**, 49–73.
12. Helena, B., Vega, M., Barrado, E., Pardo, R. and Fernandez, L., A case of hydrochemical characterization of an alluvial aquifer influenced by human activities. *Water Air Soil Pollut.*, 1999, **112**, 365–387.
13. Helena, B., Pardo, R., Vega, M., Barrado, E., Fernandez, J. M. and Fernandez, L., Temporal evolution of groundwater composition in an alluvial aquifer (Pisuerga river, Spain) by principal component analysis. *Water Res.*, 2000, **34**, 807–816.
14. Reghunath, R., Sreedhara Murthy, T. R. and Raghavan, B. R., The utility of multivariate statistical techniques in hydrogeochemical studies: An example from Karnataka, India. *Water Res.*, 2002, **36**, 2437–2442.
15. Farnham, I. M., Johanneson, K. H., Singh, A. K., Hodge, V. F. and Stetzenbach, K. J., Factor analytical approaches for evaluating groundwater trace element chemistry data. *Anal. Chim. Acta*, 2003, **490**, 123–138.
16. Liu, C.-W., Lin, K.-H. and Kuo, Y.-M., Application of factor analysis in the assessment of groundwater quality in a blackfoot disease area in Taiwan. *Sci. Total Environ.*, 2003, **313**, 77–89.
17. Singh, K. P., Malik, A., Singh, V. K., Mohan, D. and Sinha, S., Chemometric analysis of groundwater quality data of alluvial aquifer of Gangetic plain, North India. *Anal. Chim. Acta*, 2005, **550**, 82–91.
18. Dragon, K., Application of factor analysis to study contamination of a semi-confined aquifer (Wielkopolska buried valley aquifer, Poland). *J. Hydrol.*, 2006, **331**, 272–279.
19. Davis, J. C., *Statistical Data Analysis in Geology*, John Wiley, New York, 1986, 2nd edn.
20. Sharaf, M. A., Illman, D. L. and Kowalski, B. R., *Chemometrics*, Wiley, New York, 1986.
21. Mellinger, M., Multivariate data analysis: Its methods. *Chemometr. Intell. Lab. Syst.*, 1987, **2**, 29–36.
22. Wold, S., Esbensen, K. and Geladi, P., Principal component analysis. *Chemometr. Intell. Lab. Syst.*, 1987, **2**, 37–52.
23. Bratchell, N., Cluster analysis. *Chemometr. Intell. Lab. Syst.*, 1989, **6**, 105–125.
24. Schneeweiss, H. and Mathes, H., Factor analysis and principal components. *J. Multivar. Anal.*, 1995, **55**, 105–124.
25. Wackernagel, H., *Multivariate Geostatistics. An Introduction with Applications*, Springer, New York, 1995.
26. Birks, H. J. B., Multivariate analysis in geology and geochemistry: An introduction. *Chemometr. Intell. Lab. Syst.*, 1987, **2**, 15–28.
27. Wenning, R. J. and Erickson, G. A., Interpretation and analysis of complex environmental data using chemometric methods. *Trends Anal. Chem.*, 1994, **13**, 446–457.
28. Brown, S. D., Sum, S. T. and Despagne, F., Chemometrics. *Anal. Chem.*, 1996, **68**, 21R–61R.
29. APHA–AWWA–WEF, Standard methods for the examination of water and wastewater, Washington DC, 1998.
30. Seaman, J. C., Bertsh, P. M. and Schwallie, L., *In situ* Cr(VI) reduction with coarse-textured oxide-coated soil and aquifer systems using Fe(II) solutions. *Environ. Sci. Technol.*, 1999, **33**, 938–944.
31. Kostarelos, K., Reate, D., Dermatas, D., Rao, E. and Moon, D. H., Optimum dose of lime and flyash for treatment of hexavalent chromium-contaminated soil. *Water Air Soil Pollut., Focus*, 2006, **6**, 171–189.

ACKNOWLEDGEMENTS. A.R.K. thanks B. M. Jha, Chairman, CGWB, Faridabad and N. Varadaraj, Regional Director, CGWB, SECR, Chennai for permission to publish this work. The constructive comments of the reviewers are acknowledged.

Received 30 May 2007; revised accepted 5 March 2008