Environ Monit Assess (2016) 188:591
DOI 10.1007/s10661-016-5590-y

The contribution of cluster and discriminant analysis to the classification of complex aquifer systems

G. P. Panagopoulos · D. Angelopoulou · E. E. Tzirtzilakis · P. Giannoulopoulos

Received: 26 January 2016 / Accepted: 14 September 2016

Abstract This paper presents an innovated method for the discrimination of groundwater samples in common groups representing the hydrogeological units from where they have been pumped. This method proved very efficient even in areas with complex hydrogeological regimes. The proposed method requires chemical analyses of water samples only for major ions, meaning that it is applicable to most of cases worldwide. Another benefit of the method is that it gives a further insight of the aquifer hydrogeochemistry as it provides the ions that are responsible for the discrimination of the group. The procedure begins with cluster analysis of the dataset in order to classify the samples in the corresponding hydrogeological unit. The feasibility of the method is proven from the fact that the samples of volcanic origin were separated into two different clusters, namely the lava units and the pyroclastic–ignimbritic aquifer. The second step is the discriminant analysis of the data which provides the functions that distinguish the groups from each other and the most significant variables that define the hydrochemical composition of the aquifer. The whole procedure was highly successful as the 94.7 % of the samples were classified to the correct aquifer system. Finally, the resulted functions can be safely used to categorize samples of either unknown or doubtful origin improving thus the quality and the size of existing hydrochemical databases.

Keywords Multivariate statistical analysis · Groundwater · Hydrogeology · Hydrochemistry · Lesvos · Greece

Abbreviations

CA	Cluster analysis
DA	Discriminant analysis
PCA	Principal component analysis
FA	Factor analysis
CCDA	Combined cluster and discriminant analysis
CART	Classification and regression tree
BRT	Boosted regression tree
RF	Random forest classification

Introduction

During the last decades, many researchers discovered the usefulness of multivariate statistical analysis methods for the investigation and discrimination of sources of variation in water quality. This statistical technique which embodies principal component analysis (PCA), factor analysis (FA), cluster analysis (CA), and discriminant analysis (DA) has been widely used in hydrology and hydrogeology. Several

G. P. Panagopoulos (✉) · E. E. Tzirtzilakis
Department of Mechanical Engineering, Technological Educational Institute of Western Greece, M. Alexandrou 1, 26334 Patras, Greece
e-mail: gpanagopoulos@teimes.gr

D. Angelopoulou
Hellenic Open University, GR-26335 Patras, Greece

P. Giannoulopoulos
Division of Hydrogeology, Institute of Geology and Mineral Exploration, GR-13677 Athens, Greece

Published online: 27 September 2016

studies concern river waters (Zhang et al. 2011; Ajorlo et al. 2013; Li et al. 2015; Muangthong and Shrestha 2015; Phung et al. 2015; Tanos et al. 2015), lake waters (Moment and Zehr 1998; Papatheodorou et al. 2006; Yang et al. 2010; Kovács et al. 2014), or groundwater (Lambrakis et al. 2004; Papatheodorou et al. 2007; Panagopoulos and Panagiotaras 2011; Matiatos et al. 2014; Omonona et al. 2014; Sun 2014; Qian et al. 2016), while others focus in the salinization process (Petalas and Anagnostopoulos 2006; Mondal et al. 2010; Mondal et al. 2011; Arslan 2013). Filzmoser et al. (2012) and Asante and Kreamer (2015) pointed out the importance of data transformation prior to the application of standard DA methods. Rao and Srinivas (2006) tested different cluster algorithms to determine their effectiveness in regionalization and recommended the use of Ward's linkage method and K-means algorithms. Lin and Wang (2006) performed CA and DA of hydrological factors in one step, proposing a new method.

Usually, the hydrogeologists face problems in processing the hydrochemical data, because they do not know which aquifer unit a pumping well taps, since its lithological column is not always available. Such problems are more intense in cases of multilayer aquifer systems, where two or more aquifers are superimposed, rendering the hydrogeochemical interpretation of data as a complex and debatable issue. The design of a reliable water management plan requires the accurate identification of the origin of the groundwater samples in order to be deduced on proper hydrogeological and hydrochemical information.

Recently, some researchers used more sophisticated machine learning models for classifying water samples to groups. Classification and regression tree (CART) is a non-parametric regression technique that "grows" a decision tree based on a binary partitioning algorithm that recursively splits the data until groups either are homogeneous or contain not less observations than a user-defined threshold (Aertsen et al. 2010). Boosted regression tree (BRT) combines regression from CART to boosting method to create a combined modeling approach. Boosting is a forward, stepwise procedure, where tree models are fitted interactively to a subset of the training data (Aertsen et al. 2010). Random forest classification (RF) is a bagging based method, which generates a large number of trees using bootstrapping, and each tree is then handled using a randomized subset of the predictors (Breiman 2001). Naghibi et al. (2016) compared the results of CART, BRT, and RF in groundwater spring potential mapping and concluded that BRT gave the most accurate results (81.03 %), followed by CART and RF (78.70 and 71.19 %, respectively). On the other hand, Baudron et al. (2013) achieved very high accuracy (90.6 %) in the classification of a large number of water samples applying RF model compared with the results produced using CART (88 %) and linear DA (84.8 %).

This paper provides a simple multivariate statistical procedure for the identification of the hydrogeological unit in which a groundwater sample was taken based on commonly available major ions geochemistry. For this aim, a combination of unsupervised (CA) and supervised (linear DA [LDA]) statistical methods was used. Hierarchical CA was used to categorize the samples into preconceived groups and DA to construct equations for the discrimination between groups and the evaluation of the results. By applying the proposed technique, the classification of unknown origin's samples to the appropriate hydrogeological group with a very high degree of reliability is feasible.

Study area

Lesvos is a typical Mediterranean island, laying in the northeastern Aegean Sea (Fig. 1). It is the third largest island of Greece covering a total area of 1632 km^2 with a population of approximately 100,000 inhabitants. Lesvos is a concessive island with respect to the availability of water resources, contrary to the rest of the Aegean islands, because of the presence of permeable rocks hosting large aquifer bodies as well as the relatively high precipitation height. According to the data of the meteorological network of Lesvos, the mean annual rainfall height and temperature is 521 mm and 16.7 °C, respectively (hydrological period 2003–2013). The rainiest period is between December and February, whereas the period of lowest precipitation is between June and August.

Geological and hydrogeological setting

The geological bedrock of Lesvos is composed of an autochthonous metamorphic series, which is overthrusted by two allochthonous units representing the volcano-sedimentary and the ophiolitic nappe. These formations are mainly exposed in the

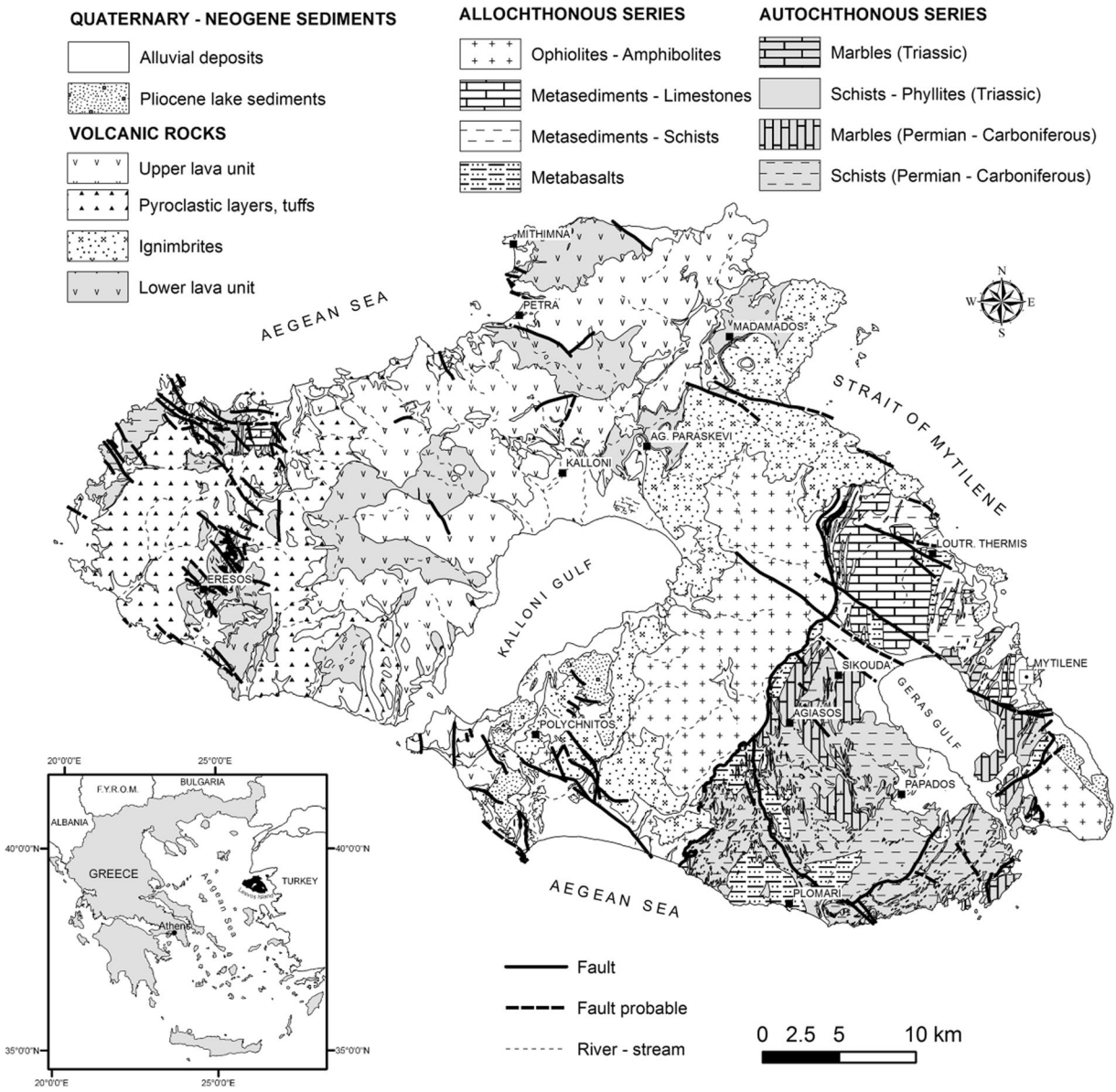

Fig. 1 Geological map of the study area (modified by Hecht 1972)

southeastern part of the island (Fig. 1). The autochthonous series consists of a Permo-Carboniferous unit with schists and marbles and a Triassic unit with marbles and phyllites (Hecht 1972). The volcano-sedimentary tectonic nappe is represented by a unit of green schists and meta-basaltic tuffs and a sedimentary unit with crystallic limestones and dolomites (Katsikatsos et al. 1982). The tectonic nappe of ophiolites has been further distinguished in two different units: the upper unit of ultramafic rocks, such as peridotites, pyroxene–peridotites, and olivinites, with variable degree of serpentinization and the lower unit with amphibolites and amphibolitic schists (Katsikatsos et al. 1982).

The largest part of the island is occupied by volcanic rocks of Late Miocene age, which have been assembled in four general units (Hecht 1972): (i) a lower unit with latitic and andesitic lavas, (ii) pyroclastic layers with lapilli tuffs and tuff breccias, (iii) ignimbrite sheets, and (iv) an upper unit of dacitic, latitic, and latitandesitic lavas.

The Pliocene deposits of Lesvos are composed of marly limestones, sandstones, and conglomerates.

Finally, the Quaternary alluvial deposits are composed of alluvial fans, debris cones, and loose sediments of silts, sands, and pebbles.

From a hydrogeological point of view, a very important aquifer was developed in the metamorphic basement of Lesvos (Fig. 2). In a regional scale, the marbles of the autochthonous series and the limestones and dolomites of the volcano-sedimentary nappe could be considered as a uniform karstic aquifer of high capacity. This carbonate aquifer is mainly confined due to the presence of impermeable schist layers, except of the sites structured only by limestones and dolomites where the karstic aquifer is unconfined.

Generally, the ophiolitic sequence of Lesvos (Fig. 2) constitutes a poor aquifer or a practically impermeable formation. However, the upper layers of peridotites have suffered from strong alteration (serpentiniosis) and weathering, forming unconfined aquifers of low capacity. On the other hand, high capacity ophiolitic aquifers occur in faulted and fractured zones due to the development of high secondary porosity.

The volcanic rocks of Lesvos (Fig. 2) constitute high-capacity aquifers which are confined by impermeable volcanic materials. These aquifers are characterized by double porosity. Especially, the lava flows have considerable pore space, and combined with columnar joints, as well as faults and fractures of tectonic origin, they form high productive aquifer bodies. The ignimbritic and pyroclastic layers could be classified as permeable geological units, especially when tectonic stress has formed fractures and faults acting as passages that allow water to move vertically and horizontally.

Finally, the alluvial deposits (Fig. 2) host extended unconfined aquifers which are recharged mainly from precipitation as well as from laterally inflows from the adjacent metamorphic and volcanic aquifers.

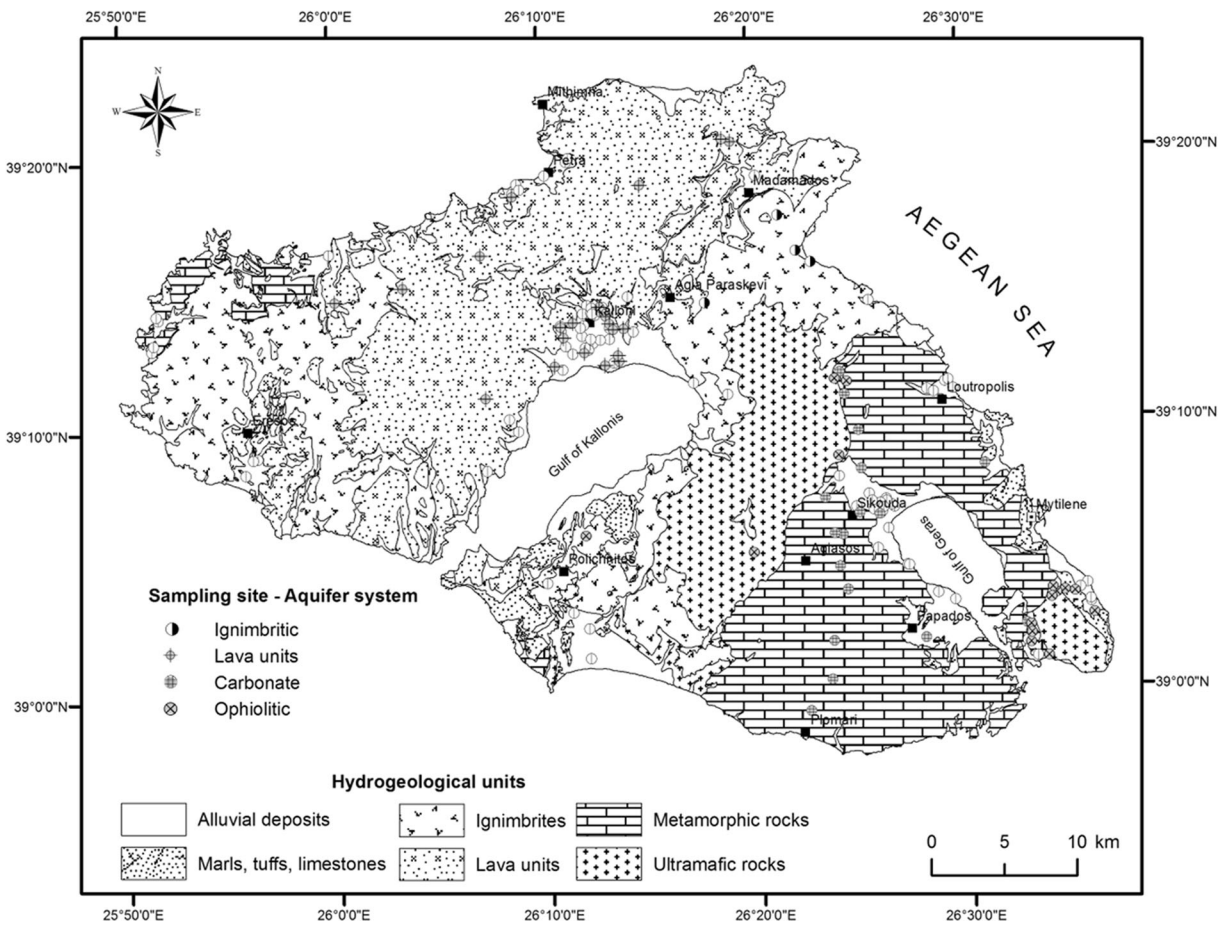

Fig. 2 Hydrogeological map of Lesvos Island showing the sampling locations

Materials and methods

Sampling

In this study, the hydrochemical data published in various reports of the Institute of Geological and Mineral Exploration of Greece (I.G.M.E.) (Giannoulopoulos and Lappas 2010) were used. The hydrochemical network is composed by 57 sampling locations (Fig. 2) consisting of wells and boreholes, with variable depths ranging from 30 to 200 m below ground level. The samples are split in four distinct groups, representing different hydrogeological units according to all the available hydrogeological and hydrochemical information obtained by the database of I.G.M.E. (Geological maps, cross sections, borehole data, etc) as well as in situ observations. The first group comprises samples derived from boreholes screened in the carbonate aquifers, while the second group contains samples belonging to the ophiolitic aquifer of Lesvos. The third group of cases represents the volcanic aquifer of the island and specifically the lava units. The fourth group has volcanic origin which contains samples from the pyroclastic and ignimbritic aquifers. Consequently, each sample classification was known in advance.

The sampling campaign took place during May 2006, and the chemical analyses of major ions were performed at the laboratory of I.G.M.E. according to APHA (1998) methods. The samples were collected at the well-pump outflow in two polyethylene bottles, after at least 1 h pumping. All samples were filtered on-site through 0.45-μm pore size filters. The first sample of 0.5 L volume was acidified with 2 mL of 65 % HNO_3 for cation analysis. The second non-acidified aliquot (1 L) was used to determine bicarbonates (HCO_3^-) and chloride (Cl^-) by titrimetry. Sulfates (SO_4^{2-}) and nitrates (NO_3^-) were determined using a spectrophotometer. Ca^{2+} and Mg^{2+} were measured using the atomic absorption method in a sprectrophotometer, while Na^+ and K^+ were measured in a flame photometer. After the removal of the outliers, the total number of data examined (observations × variables) was 456. The identification of outliers was made by using the Grubbs test according to which a case is labeled as outlier if its value deviates from the mean value for more than 3.*s* times, where *s* is the standard deviation. The deleted outliers were four. Table 1 presents a statistical report of the major ion concentrations of each aquifer.

Chemometric methods

Hierarchical cluster analysis

CA consists of different techniques that classify observations in groups (clusters) in such a way that the resulting groups are distinct from each other and group members are very similar to each other (Davis 1986; Afifi and Clark 1996; Brown 1998). This procedure that starts with *n* groups (each containing one case) and results in one group containing all (*n*) cases is called *hierarchical cluster*. Series of data describing the concentration of water sample in chemical elements are the input variables of the algorithm. The procedure transforms the data in order to standardize deviation to unity before computing proximities.

The next step is to cluster the data depending on their values. The basic criterion for any clustering is distance. There is a variety of technics for measuring and determining similarity. For the type of data used in this analysis, the Euclidean distance is considered the best choice to measure similarity. Objects that are near each other should belong to the same cluster, and objects that are far from each other should belong to different clusters since they are considered to be different.

The decision of which clusters will be merged is based on the minimization of the loss of information from joining two groups. This is the basic idea in Ward's method that is used for continuous valued data and provides more equal in number groups. It is a very efficient method and it uses the Euclidean distance to calculate an analysis of variance approach to evaluate the distances between clusters. The basic output of this procedure is a dendrogram presenting the path of merging clusters that guides the researcher to select the number of clusters. The data are best described with a number of clusters that have large distances between them.

Discriminant analysis

DA builds one or more functions that predict a group membership (Davis 1986; Afifi and Clark 1996; Brown 1998). Series of water sample hydrochemical data and their grouping are the input variables of the algorithm. DA combines the information from hydrochemical data into functions that are used in order to decide in which group a new sample belongs. DA is meaningful only if the original classification of cases in groups is preconceived. This is the reason why CA is used first in order to ensure in which group each point sample belongs.

Table 1 Average, maximum, and minimum values with standard deviations of groundwater quality parameters at different aquifer systems in Lesvos

Parameter	Aquifer system				
		Carbonate $n = 17$	Ophiolitic $n = 15$	Lava units $n = 21$	Ignimbritic $n = 4$
Ca^{2+}	Average	71.3	34.5	33.6	17.7
	Max	98.0	62.5	67.2	33.1
	Min	48.9	6.4	10.4	6.4
	SD	13.3	19.2	14.4	13.1
	Skewness	0.4	0.0	0.3	1.0
Mg^{2+}	Average	22.7	85.8	14.1	8.8
	Max	49.9	157.0	19.9	17.9
	Min	5.5	53.0	2.9	1.0
	SD	14.5	30.7	4.5	7.0
	Skewness	0.3	1.3	−1.2	1.3
Na^{+}	Average	19.4	23.2	36.8	142.1
	Max	35.3	35.6	83.0	152.0
	Min	6.1	7.6	17.0	134.0
	SD	8.2	9.3	17.7	8.3
	Skewness	0.1	−0.1	2.8	−1.7
K^{+}	Average	0.9	1.3	6.3	4.5
	Max	2.6	4.7	12.6	7.8
	Min	0.4	0.4	2.2	2.2
	SD	0.6	1.2	2.7	2.5
	Skewness	1.3	2.2	0.5	0.5
HCO_3^{-}	Average	267.9	480.0	140.8	264.5
	Max	421.0	824.0	190.0	303.0
	Min	154.0	333.0	68.3	222.1
	SD	78.8	121.8	37.4	35.4
	Skewness	0.2	1.7	0.9	0.5
Cl^{-}	Average	38.4	45.8	62.0	102.7
	Max	69.1	81.5	149.0	138.0
	Min	14.2	15.6	24.8	74.5
	SD	17.3	17.1	35.0	29.3
	Skewness	0.3	0.2	1.1	0.3
SO_4^{2-}	Average	32.1	17.7	22.2	32.7
	Max	53.0	43.2	52.0	58.6
	Min	11.4	0.5	5.4	3.1
	SD	13.0	13.0	14.7	23.3
	Skewness	0.1	0.7	0.7	0.3
NO_3^{-}	Average	9.8	17.2	7.6	7.0
	Max	37.2	83.7	34.8	24.8
	Min	0.1	0.0	0.1	0.1
	SD	11.7	24.0	9.4	11.9
	Skewness	1.0	1.9	2.0	1.7

Concentrations are expressed in mg/L

A number of linear coefficients (eigenvectors), as many as different variables, are extracted from a sample of cases (for which group membership is known) as linear combinations of predictor variables. Predictor variables are selected in such a way so as to provide the best discrimination between groups, and the procedure takes care of the minimization of misclassification by maximizing the variance between groups to the variance within-group. During the procedure, several tests have been made to test the significance of each variable included in the equations and the validation of the procedure.

Of great importance is the structure of each group, where the correlation of the predictor variables and the discriminant scores (loadings produced by canonical functions) indicate the most significant variable for each function and therefore for each group. The contribution of each variable to the corresponding discriminant function is presented in a structure matrix.

The prior assignment of a sample case to a group is made considering equal probabilities, and it is the starting point of the analysis. When the analysis is completed, the final step is to form the discriminant functions with the calculated coefficients. A case is categorized in the group which the corresponding discriminant function has the greatest value. These equations can be used for the classification of any other new sample case in one of the groups.

Cross validation is achieved through the *leave-one-out classification*. This technique, called also the Jackknife technique, calculates the discrimination equations from all but one case and uses them to predict the membership of that case. Since the procedure is repeated for every case, this technique produces more unbiased estimations and ensures the minimization of incorrect classification. The overall percentage of succeeded classifications indicates the validity of the model.

Conclusively, the data analysis is performed using the algorithm depicted in the flowchart of Fig. 3.

Results and discussion

Combined cluster and discriminant analysis (CCDA) has been used from Lambrakis et al. (2004) in order to find the significant parameters that influence the quality of groundwater in an aquifer system. CA classified the samples into two groups, and DA certified that this classification was 100 % correct. Furthermore, DA indicated that SO_4^{2-} and NO_3^- are responsible for the discrimination among groups, so the contamination of groundwater from fertilizers defines the hydrochemical character of the aquifer. Pati et al. (2014) applied CCDA in time series of hydrochemical data of two stations intended to develop a water quality index. CA classified the data into three major groups and, DA generated the corresponding discriminant functions. The hit ratio of DA was about 90 %. Kovács et al. (2014) and Tanos et al. (2015) used CCDA in order to refine monitoring networks by aggregating similar sampling locations. Thus, they achieved a cost cut without significant loss of information.

The novelty of our study is the use of CCDA in complicated hydrogeological regimes where reliable information about the hydrochemical origin of water samples is missing. In such cases, CA can be used in order to categorize the samples into groups which have similar hydrochemical properties corresponding to specific hydrogeological units. LDA can be used for the validation of the results of CA, according to the hit ratio. LDA provides the procedure to categorize samples of unknown or doubtful origin into the correct aquifer system. Furthermore, LDA evaluates the results, providing the most significant parameters that explain the hydrogeochemical properties of the aquifers. Finally, another important benefit of CCDA is the use of methods that are incorporated in well known and wide used software like SPSS, Statgraphics, Minitab, etc. The proposed method was executed using the Statistical Package for Social Science (SPSS v.20).

The hierarchical CA was applied by using Ward's linkage method for sample classification and the square Euclidean distance as the measure of similarity. The variables entered in CA were carefully selected to be related but not to be correlated in order to avoid a misleading solution. In this context, the variables Ca^{2+}, Mg^{2+}, Na^+, K^+, HCO_3^-, Cl^-, SO_4^{2-}, and NO_3^- from each sampling point were included in the analysis. Moreover, in order to equalize the effect of variables measured on different scales, the data values were standardized prior to the clustering process.

Figure 4 pictures the dendrogram produced using Ward's method. It is obtained that the samples of the study area can be classified in four distinct groups. Each sample is correctly classified into one of the

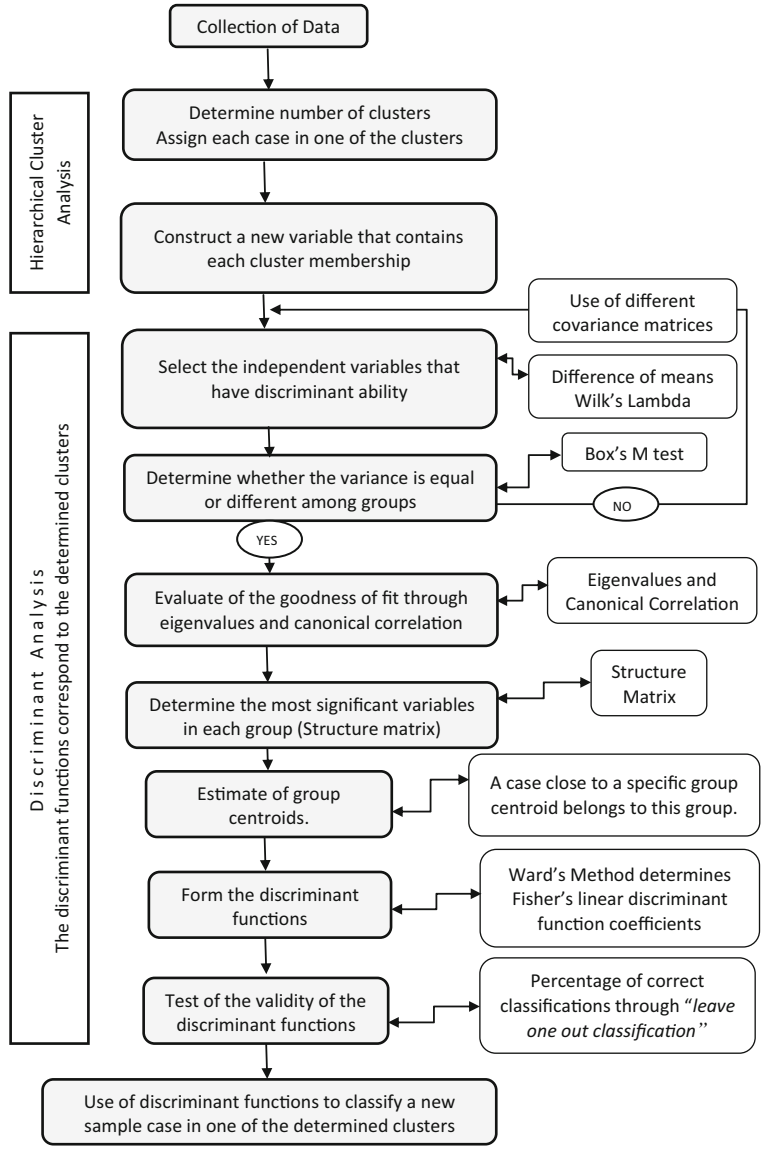

Fig. 3 Flowchart showing the steps of CCDA

four groups defined by I.G.M.E (100 % correct classification). The formation of this group indicates the discriminating power of CA and its usefulness in the hydrogeochemical research. Cluster membership of each case was saved as a new variable in order to be used in DA.

DA uses as independent variables the same variables applied in CA, which are data values of Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}, HCO_3^{-}, Cl^{-}, SO_4^{2-}, and NO_3^{-} from each sampling point. The choice of variables included in discriminant equations is tested according to its contribution to the prediction. The first test is whether a variable discriminates groups or, in other words, if there are differences among group means. Since all variables have different group means, as it is shown in the last column of Table 2, they are included in the model. The discriminant ability of each predicted variable is tested in Table 2, where the variables that have smaller values on Wilk's lambda are those who contribute more to the discriminant functions. The last column presents the significance of the null hypothesis of equality of means in each group. It is apparent that only NO_3^{-} has equal means among groups (sig = 0.293 > 0.05

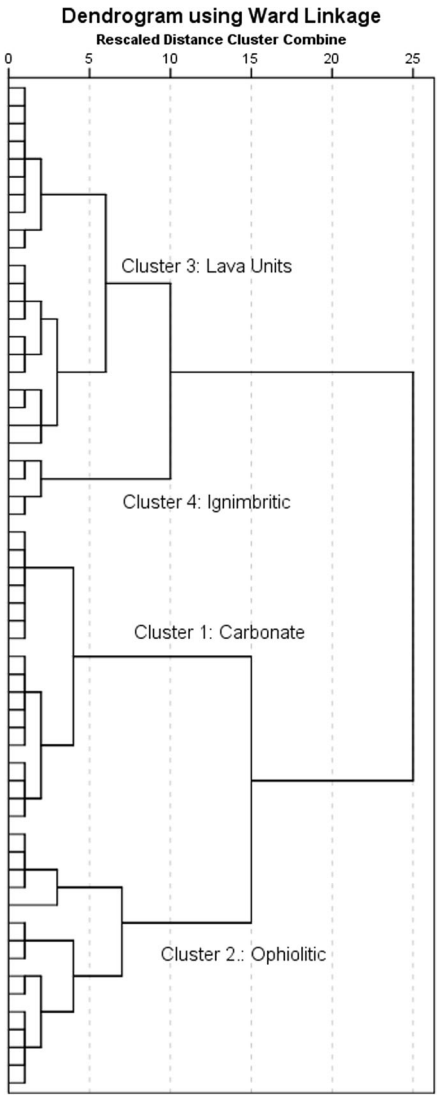

Fig. 4 Dendrogram presenting the merging path of clusters using Ward's method and Euclidean Distances

Table 2 Tests of equality of group means

	Wilks' Lambda	*F*	df1	df2	Sig.
Ca^{2+}	0.399	26.639	3	53	0.000
Mg^{2+}	0.243	55.086	3	53	0.000
Na^{+}	0.143	105.744	3	53	0.000
K^{+}	0.346	33.325	3	53	0.000
HCO_3^{-}	0.251	52.647	3	53	0.000
Cl^{-}	0.692	7.851	3	53	0.000
SO_4^{2-}	0.842	3.321	3	53	0.027
NO_3^{-}	0.933	1.273	3	53	0.293

in Table 2), and therefore, it is removed from the analysis. This is an expected result since nitrate contamination of groundwater has anthropogenic and non-geogenic origin. Consequently, nitrate concentration in each sampling point depends on human activities and not on the geochemical and mineralogical compositions of the aquifer.

The Box's *M* test showed that variances among classes were not equal, but further test using different separate matrices did not alter the results. Corresponding to eigenvalues, tests also showed that the calculated discriminant functions have considerable discriminated power.

Canonical discriminant functions are produced from the discriminant functions and actually test the discriminant ability of the discriminant functions. The number of canonical functions is always one less than the number of discriminant functions. The percentages of the variance described by each discriminant function are presented in Table 3. The first canonical discriminant function describes 63 % of total variance, the second 21 %, and the last one 16 %, respectively. In the same table, it is verified that the independent variables used in discriminant functions are highly correlated to the corresponding predicted variable.

The structure matrix (Table 4) presents the hierarchical sorting of predicted variables for each function. The most significant variable for each function and therefore for each group is marked in the matrix with an asterisk. It is noted that each chemical parameter is characterized as most significant for a specific canonical function when it has the greatest discriminant ability, i.e., it is the greatest number in each row.

The carbonate sequence constitutes the first discriminated group and the most important variables for the first function that discriminates it from other groups are Na^{+} and Cl^{-}. This is expected as the groundwater of the aquifer is fresher and not contaminated by seawater.

Table 3 Eigenvalues and the goodness of fit of the model

Canonical function	Eigenvalue	% of Variance	Cumulative %	Canonical correlation
1	9.678	62.9	62.9	0.952
2	3.245	21.1	84.0	0.874
3	2.453	16.0	100.0	0.843

This is confirmed from Table 1, where the sodium and chloride contents are less in the carbonate aquifer compared to the others.

The second group of samples represents the ophiolitic aquifer of Lesvos, and the most important variables for the discriminant function are Mg^{2+} and HCO_3^-. This is in agreement with the hydrochemical data of Table 1, as these ions have increased concentrations in the ophiolitic aquifer compared to the other aquifers. The contribution of these ions to the discriminating power of the second function is related with the geochemical and mineralogical properties of the ophiolites. The basic mineralogical composition of these rocks is olivine [$(Mg,Fe^{2+})_2SiO_4$] + orthopyroxene ($MgSiO_3$) + serpentine [$(Mg,Fe)_3Si_2O_5(OH)_4$], indicating that the increased Mg^{2+} concentrations of the groundwater originate from the dissolution and disassociation of these minerals. Moreover, the increased content of Mg^{2+} and HCO_3^- ions in the groundwater samples of this group is attributed to the presence of magnesite ($MgHCO_3$) dykes. According to Chatzidimitriadis and Allagianis (1972), these dykes have been formed in faulted and fractured zones of the ophiolites during the serpentiniosis process.

The third discrimination function separates the volcanic sequences of Lesvos Island based on Ca^{2+}, K^+, and SO_4^{2-} concentrations of groundwater (Table 4). The lava units are high-K shoshonitic rocks associated with Ca-alkaline volcanism, while the general mineral assemblage contains Ca-plagioclase ($CaAlSi_3O_8$), K-feldspar ($KAlSi_3O_8$), and clinopyroxene ($CaMgSi_2O_6$). On the other hand, the pyroclastic layers and the ignimbritic sheets are acid volcanic materials containing mainly Na-plagioclase ($NaAlSi_3O_8$) and K-feldspars (Pe-Piper and Piper 1992; Kelepertsis 1993). Consequently, the aquifers of lava units are enriched in Ca^{2+} compared to the pyroclastic and ignimbritic aquifers which are enriched in Na^+ (Table 1).

There are several ways to predict a case membership in a group. One way is the use of information for each group centroid as presented in Table 5. Cases with function scores close to a group centroid are predicted as belonging to the same group. Thus, if a case has all function scores positive, it should belong to group 4, whereas, if a case has positive score only for the third function, it should belong to group 1.

The most significant way to predict a group membership is through the discriminant functions, one for each group (Table 6). Each case is classified in the group that the corresponding function has the greater value.

Table 6 presents the coefficients of the discriminant functions that in the end take the following form:

Table 4 Structure matrix

	Canonical function		
	1	2	3
Na^+	0.722*	0.436	0.360
Cl^-	0.211*	0.055	−0.044
HCO_3^-	−0.285	0.822*	0.021
Mg^{2+}	−0.316	0.753*	−0.355
Ca^{2+}	−0.245	−0.276	0.526*
K^+	0.310	−0.353	−0.475*
SO_4^{2-}	0.021	−0.079	0.258*

*Largest absolute correlation between each variable and any discriminant function

Table 5 Unstandardized canonical discriminant functions evaluated at group means

Cluster	Canonical function		
	1	2	3
1	−2.120	−0.917	1.895
2	−2.245	2.341	−0.983
3	1.662	−1.458	−1.266
4	8.703	2.777	2.279

Table 6 Fisher's linear discriminant functions coefficients for four groups using Ward's method

Parameter	Ward method			
	1	2	3	4
Ca^{2+}	0.385	0.197	0.082	−0.033
Mg^{2+}	0.023	0.097	−0.047	−0.179
Na^+	0.101	0.057	0.184	0.918
K^+	0.050	0.442	1.762	1.393
HCO_3^-	0.032	0.059	0.028	0.046
Cl^-	−0.082	−0.011	−0.016	−0.070
SO_4^{2-}	−0.084	−0.163	−0.064	−0.110
(Constant)	−17.776	−22.418	−12.159	−69.366

Table 7 The hit ratio of the discriminant model

	Total observations (*n*)	Group	Predicted group membership			
			Carbonate	Ophiolitic	Lava units	Pyroclastic–ignimbritic
Original group	17	Carbonate	17 (100 %)	0 (0 %)	0 (0 %)	0 (0 %)
	15	Ophiolitic	0 (0 %)	15 (100 %)	0 (0 %)	0 (0 %)
	21	Lava units	0 (0 %)	0 (0 %)	21 (100 %)	0 (0 %)
	4	Pyroclastic–ignimbritic	0 (0 %)	0 (0 %)	0 (0 %)	4 (100 %)
Cross validated[a]	17	Carbonate	17 (100 %)	0 (0 %)	0 (0 %)	0 (0 %)
	15	Ophiolitic	1 (7 %)	14 (93 %)	0 (0 %)	0 (0 %)
	21	Lava units	1 (5 %)	0 (0 %)	19 (90 %)	1 (5 %)
	4	Pyroclastic–ignimbritic	0 (0 %)	0 (0 %)	0 (0 %)	4 (100 %)

[a] Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case

$$D1 = -17.776 + 0.385\,Ca^{2+} + 0.023\,Mg^{2+} + 0.101\,Na^{+} + 0.050\,K^{+} + 0.032\,HCO_3^{-} - 0.082\,Cl^{-} - 0.084\,SO_4^{2}$$
$$D2 = -22.418 + 0.197\,Ca^{2+} + 0.097\,Mg^{2+} + 0.057\,Na^{+} + 0.442\,K^{+} + 0.059\,HCO_3^{-} - 0.011\,Cl^{-} - 0.163\,SO_4^{2-}$$
$$D3 = -12.159 + 0.082\,Ca^{2+} - 0.047\,Mg^{2+} + 0.184\,Na^{+} + 1.762\,K^{+} + 0.028\,HCO_3^{-} - 0.016\,Cl^{-} - 0.064\,SO_4^{2-}$$
$$D4 = -69.366 - 0.033\,Ca^{2+} - 0.179\,Mg^{2+} + 0.918\,Na^{+} + 1.393\,K^{+} + 0.046\,HCO_3^{-} - 0.070\,Cl^{-} - 0.110\,SO_4^{2-}$$

To assign a sample case in a group, the values D1–D4 from these equations are calculated and called discriminant scores. The sample case is assigned to the group with the higher discriminant score. This is performed for each case of the data, so all cases are classified in one of the four groups.

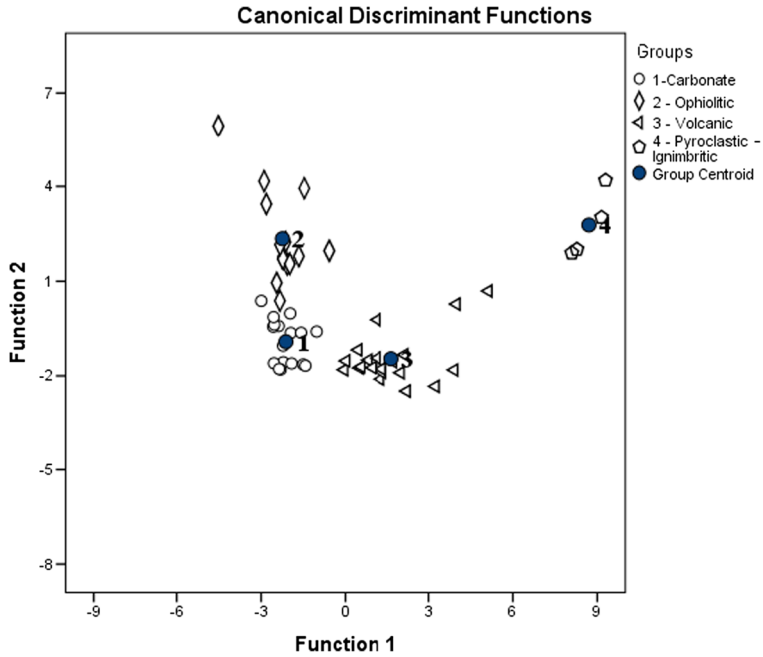

Fig. 5 The dispersion of four groups according to discriminant analysis as a projection in function-1 vs function-2 plane

The hit ratio of the discriminant model is computed by comparing the predefined grouping of all cases with the one predicted from the model. In our case, this scales to 100 % of successful classification of original grouped cases or 94.7 % of cross-validated grouped cases (Table 7).

This accuracy of the method is better than that obtained by Pati et al. (2014) who used CCDA on similar hydrochemical data with a hit ratio of approximately 90 %. This hit ratio is large enough, so we do not think that it is necessary, in this case, to use more complex machine learning methods despite the more accurate results when analyzing other datasets (Naghibi et al. 2016; Baudron et al. 2013).

The visualization of the discrimination is often very enlightening. In such graphs, the existence of outliers and wrong classifications can be observed, as well as the discriminating power of the method, as it is shown in Fig. 5. It is noted that the outliers were correctly classified using the four discriminant functions. The discrimination of groups is clear, and the distances between group centroids are distinct. This is happening because each aquifer unit has different geological origin and consequently different geochemical composition. As a result, the range values of the ion concentrations of the four aquifers do not overlap facilitating in that way the classification of the dataset. Nonetheless, the researcher should test the proposed CCDA method to such an easy dataset prior to try it with more complex cases.

The final outcome of the analysis is substantially the construction of a set of functions (D1–D4) that can be used for future sample cases in order to categorize them to a specific hydrogeological unit.

Conclusions

The CCDA method consists of CA and DA and was applied to hydrochemical data of major ions from 57 samples taken from wells and boreholes of Lesvos Island in Greece. CA classified the samples into four distinct groups with clear hydrochemical interpretation corresponding perfectly to the four main aquifer systems of Lesvos. It is noteworthy that CA manages to categorize the groundwater samples of volcanic origin into two subgroups which are composed of different volcanic aquifer systems, namely the lava units and the pyroclastic–ignimbritic aquifer.

DA proved the feasibility of CA, since 94.7 % of cross-validated grouped cases were correctly classified. The obtained discriminant functions provided the most significant ions which characterize the hydrogeochemical properties of each aquifer.

Finally, the proposed procedure could be very useful for the categorization of unknown or doubtful origin samples, especially in areas with complicated hydrogeological regimes. This way, CCDA could be very helpful for improving the quality and size of existing hydrochemical databases which constitute a common handicap for integrated water management plans.

References

Aertsen, W., Kint, V., van Orshoven, J., Özkan, K., & Muys, B. (2010). Comparison and ranking of different modelling techniques for prediction of site index in Mediterranean mountain forests. *Ecological Modelling, 221*, 1119–1130.

Afifi, A. A., & Clark, V. (1996). *Computer-aided multivariate analysis* (3rd ed.). London: Chapman & Hall.

Ajorlo, M., Abdullah, R. B., Yusoff, M. K., Halim, R. A., Hanif, A. H. M., Willms, W. D., & Ebrahimian, M. (2013). Multivariate statistical techniques for the assessment of seasonal variations in surface water quality of pasture ecosystems. *Environmental Monitoring and Assessment, 185*, 8649–8658.

American Public Health Association (APHA) (1998). Standard methods for examination of water and wastewater. American Public Health Association Inc, 20th Edition, Washington DC.

Arslan, H. (2013). Application of multivariate statistical techniques in the assessment of groundwater quality in seawater intrusion area in Bafra Plain, Turkey. *Environmental Monitoring and Assessment, 185*, 2439–2452.

Asante, J., & Kreamer, D. (2015). A new approach to identify recharge areas in the Lower Virgin River Basin and surrounding basins by multivariate statistics. *Mathematical Geoscience, 47*, 819–842.

Baudron, P., Alonso-Sarría, F., García-Aróstegui, J. L., CánovasGarcía, F., Martínez-Vicente, D., & Moreno-Brotóns, J. (2013). Identifying the origin of groundwater samples in a multi-layer aquifer system with random forest classification. *Journal of Hydrology, 499*, 303–315.

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32.

Brown, C. A. (1998). *Applied multivariate statistics in geohydrology and related sciences*. New York: Springer.

Chatzidimitriadis, E., & Allagianis, I. (1972). Final conclusions of magnesite study in “Aforades” area, Agiasos, Lesvos. Unpublished report, I.G.M.E., Athens (in Greek).

Davis, J. C. (1986). *Statistics and data analysis in geology*. New York: Wiley.

Filzmoser, P., Hron, K., & Templ, M. (2012). Discriminant analysis for compositional data and robust parameter estimation. *Comp Stat, 27*(4), 585–604.

Giannoulopoulos, P., & Lappas, I. (2010). Evaluation of water resources of Aegean, Quality measurements and development measures. Unpublished report, I.G.M.E., Athens (in Greek).

Hecht, J. (1972). *Geological map of Greece, scale 1:50,000, Mytilene-Plomari, Agia Paraskevi, Polychnitos and Eresos sheets*. Athens: I.G.M.E.

Katsikatsos, G., Mataragas, D., Migiros, G., & Triantafylli, E. (1982). Geological study of Lesvos Island, Unpublished report, I.G.M.E., Athens (in Greek).

Kelepertsis, A. E. (1993). Hydrothermal alteration of basic island-arc volcanic rocks north and south of Mytilini Town, Lesvos Island, Greece. *Terra Nova, 5*(1), 52–60.

Kovács, J., Kovács, S., Magyar, N., Tanos, P., Hatvani, I. G., & Anda, A. (2014). Classification into homogeneous groups using combined cluster and discriminant analysis. *Environ Model & Soft, 57*, 52–59.

Lambrakis, N., Antonakos, A., & Panagopoulos, G. (2004). The use of multicomponent statistical analysis in hydrogeological environmental research. *Water Research, 38*, 1862–1872.

Li, D., Huang, D., Guo, C., & Guo, X. (2015). Multivariate statistical analysis of temporal–spatial variations in water quality of a constructed wetland purification system in a typical park in Beijing, China. *Environmental Monitoring and Assessment, 187*, 4219. doi:10.1007/s10661-014-4219-2.

Lin, G.-F., & Wang, C.-M. (2006). Performing cluster analysis and discrimination analysis of hydrological factors in one step. *Advances in Water Resources, 29*, 1573–1585.

Matiatos, I., Alexopoulos, A., & Godelitsas, A. (2014). Multivariate statistical analysis of the hydrogeochemical and isotopic composition of the groundwater resources in northeastern Peloponnesus (Greece). *Sci Tot Environ, 476-477*, 577–590.

Moment, B., & Zehr, J. (1998). Watershed classification by discriminant analyses of lakewater-chemistry and terrestrial characteristics. *Ecological Applications, 8*(2), 497–507.

Mondal, N. C., Singh, V. P., Singh, V. S., & Saxena, V. K. (2010). Determining the interaction between groundwater and saline water through groundwater major ions chemistry. *Journal of Hydrology, 388*, 100–111.

Mondal, N. C., Singh, V. S., Saxena, V. K., & Singh, V. P. (2011). Assessment of seawater impact using major hydrochemical ions: a case study from Sadras, Tamilnadu, India. *Environmental Monitoring and Assessment, 177*, 315–335.

Muangthong, S., & Shrestha, S. (2015). Assessment of surface water quality using multivariate statistical techniques: case study of the Nampong River and Songkhram River, Thailand. *Environmental Monitoring and Assessment, 187*, 548. doi:10.1007/s10661-015-4774-1.

Naghibi, S. A., Pourghasemi, H. R., & Dixon, B. (2016). GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran. *Environmental Monitoring and Assessment, 188*, 44. doi:10.1007/s10661-015-5049-6.

Omonona, O. V., Onwuka, O. S., & Okogbue, C. O. (2014). Characterization of groundwater quality in three settlement areas of Enugu metropolis, southeastern Nigeria, using multivariate analysis. *Environmental Monitoring and Assessment, 186*, 651–664. doi:10.1007/s10661-013-3405-y.

Panagopoulos, G., & Panagiotaras, D. (2011). Understanding the extent of geochemical and hydrochemical processes in coastal karst aquifers through ion chemistry and multivariate statistical analysis. *Fres Env Bull, 20*(12A), 3270–3285.

Papatheodorou, G., Demopoulou, G., & Lambrakis, N. (2006). A long-term study of temporal hydrochemical data in a shallow lake using multivariate statistical techniques. *Ecological Modelling, 193*, 759–776.

Papatheodorou, G., Lambrakis, N., & Panagopoulos, G. (2007). Application of multivariate statistical procedures to the hydrochemical study of coastal aquifer: an example from Crete, Greece. *Hydrological Processes, 21*(11), 1482–1495.

Pati, S., Dash, M. K., Mukherjee, C. K., Dash, B., & Pokhrel, S. (2014). Assessment of water quality using multivariate statistical techniques in the coastal region of Visakhapatnam, India. *Environmental Monitoring and Assessment, 186*, 6385–6402.

Pe-Piper, G., & Piper, D. J. W. (1992). Geochemical variation with time in the Cenozoic high-K volcanic rocks of the island of Lesbos Greece: significance for shoshonite petrogenesis. *J Volcanol Geothermal Res, 53*, 371–387.

Petalas, C., & Anagnostopoulos, K. (2006). Application of stepwise discriminant analysis for the identification of salinity sources of groundwater. *Water Resources Management, 20*, 681–700.

Phung, D., Huang, C., Rutherford, S., Dwirahmadi, F., Chu, C., Wang, X., Nguyen, M., Nguyen, N. H., Do, C. M., Nguyen, T. H., & Dinh, T. A. D. (2015). Temporal and spatial assessment of river surface water quality using multivariate statistical techniques: a study in Can Tho City, a Mekong Delta area, Vietnam. *Environmental Monitoring and Assessment, 187*, 229. doi:10.1007/s10661-015-4474-x.

Qian, J., Wang, L., Ma, L., Lu, Y., Zhao, W., & Zhang, Y. (2016). Multivariate statistical analysis of water chemistry in evaluating groundwater geochemical evolution and aquifer connectivity near a large coal mine, Anhui, China. *Environmental Earth Sciences, 75*, 747. doi:10.1007/s12665-016-5541-5.

Rao, A. R., & Srinivas, V. V. (2006). Regionalization of watersheds by hybrid-cluster analysis. *Journal of Hydrology, 318*, 37–56.

Sun, L. H. (2014). Statistical analysis of hydrochemistry of groundwater and its implications for water source identification: a case study. *Arabian Journal of Geosciences, 7*, 3417–3425.

Tanos, P., Kovács, J., Kovács, S., Anda, A., & Hatvani, I. G. (2015). Optimization of the monitoring network on the River Tisza (Central Europe, Hungary) using combined cluster and discriminant analysis, taking seasonality into account. *Environmental Monitoring and Assessment, 187*, 575. doi:10.1007/s10661-015-4777-y.

Yang, Y.-H., Zhou, F., Guo, H.-C., Sheng, H., Liu, H., Dao, X., & He, C.-J. (2010). Analysis of spatial and temporal water pollution patterns in Lake Dianchi using multivariate statistical methods. *Environmental Monitoring and Assessment, 170*, 407–416. doi:10.1007/s10661-009-1242-9.

Zhang, X., Wang, Q., Liu, Y., Wu, J., & Yu, M. (2011). Application of multivariate statistical techniques in the assessment of water quality in the Southwest New Territories and Kowloon, Hong Kong. *Environmental Monitoring and Assessment, 173*, 17–27. doi:10.1007/s10661-010-1366-y.