

# Evaluation of graphical and multivariate statistical methods for classification of water chemistry data

Cüneyt Güler · Geoffrey D. Thyne · John E. McCray  
A. Keith Turner

**Abstract** A robust classification scheme for partitioning water chemistry samples into homogeneous groups is an important tool for the characterization of hydrologic systems. In this paper we test the performance of the many available graphical and statistical methodologies used to classify water samples including: Collins bar diagram, pie diagram, Stiff pattern diagram, Schoeller plot, Piper diagram, Q-mode hierarchical cluster analysis, K-means clustering, principal components analysis, and fuzzy k-means clustering. All the methods are discussed and compared as to their ability to cluster, ease of use, and ease of interpretation. In addition, several issues related to data preparation, database editing, data-gap filling, data screening, and data quality assurance are discussed and a database construction methodology is presented.

The use of graphical techniques proved to have limitations compared with the multivariate methods for large data sets. Principal components analysis is useful for data reduction and to assess the continuity/overlap of clusters or clustering/similarities in the data. The most efficient grouping was achieved by statistical clustering techniques. However, these techniques do not provide information on the chemistry of the statistical groups. The combination of graphical and statistical techniques provides a consistent and objective means to classify large numbers of samples while retaining the ease of classic graphical presentations.

**Résumé** Un système robuste de classification pour répartir des échantillons de chimie de l'eau en groupes homogènes est un outil important pour la caractérisation des hydrosystèmes. Dans ce papier nous testons les performances des nombreuses méthodes graphiques et statistiques disponibles utilisées pour réaliser une classifi-

cation des échantillons d'eau; ces méthodes sont les suivantes: les diagrammes en barres de Collins, en camembert, de Stiff, de Schoeller, de Piper, l'analyse hiérarchique en grappe en mode Q, le regroupement de moyennes K, l'analyse en composantes principales et le regroupement flou de moyennes K. Toutes ces méthodes sont discutées et comparées quant à leur aptitude à regrouper et leur facilité de mise en œuvre et d'interprétation. En outre, plusieurs points relatifs à la préparation des données, à l'édition des bases de données, à la reconstitution de données manquantes, à l'examen des données et au contrôle de validité des données sont discutés et une méthodologie d'élaboration d'une base de données est proposée.

L'utilisation de techniques graphiques a démontré qu'elle présente des limites par rapport aux méthodes multidimensionnelles, pour les jeux importants de données. L'analyse en composantes principales est utile pour réduire les données et pour évaluer la continuité/recouvrement des groupes ou le groupement/similitude dans les données. Le groupement le plus efficace est assuré par les techniques statistiques de regroupement en grappes. Cependant, ces techniques ne fournissent pas d'information sur le chimisme des groupes statistiques. La combinaison de techniques graphiques et statistiques donne les moyens solides et objectifs de faire une classification d'un grand nombre d'échantillons tout en conservant la facilité des représentations graphiques classiques.

**Resumen** Disponer de un esquema sólido de clasificación química de muestras de agua en grupos homogéneos es una herramienta importante para la caracterización de sistemas hidrológicos. En este artículo, contrastamos la utilidad de muchas metodologías gráficas y estadísticas disponibles para clasificar muestras de aguas; entre ellas, hay que citar el diagrama de barras de Collins, diagramas de sectores, diagrama de Stiff, gráfico de Schoeller, diagrama de Piper, análisis jerárquico de conglomerados en modo-Q, conglomerados de K-medias, análisis de componentes principales, y conglomerados difusos de k-medias. Se discute todos los métodos, comparándolos en función de su capacidad para establecer agrupaciones, de su facilidad de uso y de su facilidad de interpretación. Además, se discute varios aspectos relacionados con la entrada de datos, edición de bases de da-

Received: 25 July 2001 / Accepted: 23 February 2002  
Published online: 9 May 2002

© Springer-Verlag 2002

C. Güler (✉) · G.D. Thyne · J.E. McCray · A.K. Turner  
Colorado School of Mines,  
Department of Geology and Geological Engineering,  
1500 Illinois Street, Golden, CO 80401, USA  
e-mail: cguler@mines.edu  
Tel.: +1-303-2169770, Fax: +1-303-2733859

tos, extrapolación de datos en series incompletas, visualización de datos, y garantía de calidad de los datos, y se presenta una metodología para elaborar una base de datos.

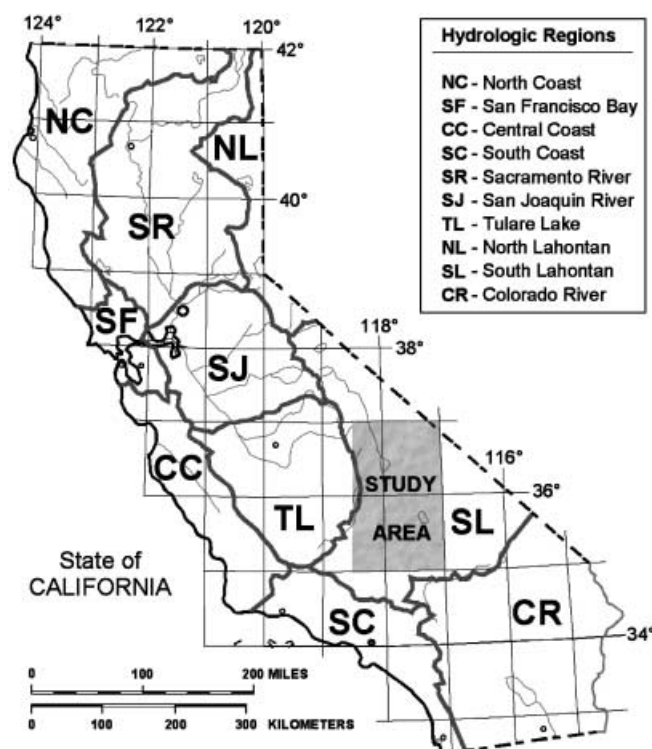
Se demuestra que el uso de técnicas gráficas padece limitaciones respecto a los métodos multivariados para conjuntos de datos numerosos. El análisis de componentes principales es útil para reducir el número de datos y establecer la continuidad/superposición de grupos o agrupaciones/similaridades en los datos. Los resultados más efectivos se logran mediante técnicas estadísticas de agrupamiento; sin embargo, éstas no proporcionan información sobre la química de los grupos estadísticos. La combinación de técnicas gráficas y estadísticas posibilita un enfoque coherente y objetivo para clasificar números elevados de muestras y, a la vez, mantener la facilidad de las presentaciones gráficas convencionales.

**Keywords** Classification techniques · Cluster analysis · Database construction · Fuzzy k-means clustering · Water chemistry

## Introduction

The chemical composition of surface and groundwater is controlled by many factors that include composition of precipitation, mineralogy of the watershed and aquifers, climate, and topography. These factors combine to create diverse water types that change spatially and temporally. In our study area, which lies within the south Lahontan hydrologic region of southeastern California (Fig. 1), there is a wide variety of climatic conditions (high alpine to desert), hydrologic regimes (alluvial basin-fill aquifers, fractured rock aquifers, and playas) and geologic environments (igneous rocks, volcanic rocks, metamorphic rocks, sedimentary deposits, evaporites, and mineralized zones). Thus, the samples from the area could potentially represent a variety of water types providing an opportunity to test the performance of many of the available graphical and statistical methodologies used to classify water samples.

The use of major ions as natural tracers (Back 1966) has become a very common method to delineate flow paths in aquifers. Generally, the approach is to divide the samples into hydrochemical facies (aka water types), that is groups of samples with similar chemical characteristics that can then be correlated with location. The spatial variability observed in the composition of these natural tracers can provide insight into aquifer heterogeneity and connectivity, as well as the physical and chemical processes controlling water chemistry. Thus, a robust classification scheme for partitioning water chemistry samples into homogeneous groups can be an important tool for the successful characterization of hydrogeologic systems. A variety of graphical and multivariate statistical techniques have been devised since the early 1920s in order to facilitate the classification of waters, with the ultimate goal of dividing a group of samples into similar



**Fig. 1** Location of the study area

homogeneous groups (each representing a hydrochemical facies). Several commonly used graphical methods and multivariate statistical techniques are available including: Collins bar diagram, pie diagram, Stiff pattern diagram, Schoeller semi-logarithmic diagram, Piper diagram, Q-mode hierarchical cluster analysis (HCA), K-means clustering (KMC), principal components analysis (PCA), and fuzzy k-means clustering (FKM). This paper utilizes a relatively large data set to review these techniques and compare their ease of use and ability to sort water chemistry samples into groups.

## Hydrogeologic Setting

The study area is part of the Basin and Range Province of the southwestern USA and extends from 35–37° of latitude north and from 117–118.5° of longitude west (Fig. 1). The area comprises a portion of the Sierra Nevada mountain range, which is the recharge area, and adjoining alluvial basins, which are arid. Because of the modern arid climate, surface water is scarce in the area and groundwater is the only source of drinking and household use water (Berenbrock and Schroeder 1994). Thus, effective management of the groundwater resources requires an accurate model for the aquifer characteristics, groundwater flow directions, recharge mechanisms, discharge mechanisms, and water chemistry processes.

In the basin and range groundwater system, water flows from recharge areas in the mountains to discharge areas in the adjacent valleys (Maxey 1968). This local flow system is often modified by local geologic, physio-

graphic, and climatic factors. During the Pleistocene and Holocene epochs, the valley floors in the study area were periodically occupied by a chain of lakes stretching from Mono Lake in the north to Lake Manley in Death Valley (Duffield and Smith 1978; Lipinski and Knochenmus 1981). Present-day valley floors are occupied by playas, known in different localities as “salt lakes,” “soda lakes,” “alkali marshes,” “dry lakes,” or “borax lakes”, where the majority of groundwater discharges by evapotranspiration (Lee 1912; Fenneman 1931; Dutcher and Moyle 1973). Minor discharge also occurs by other ways including discharge from springs, seeps, and pumping from wells. The groundwater in the area occurs in two porosity regimes: (1) intergranular porosity found mostly in alluvial basin-fill aquifers, and (2) fracture porosity found in the mountain watersheds. The alluvial basin-fill aquifers can be further divided into two components: a shallow saline aquifer (<150 m depth), and a deep (610 m), locally confined aquifer that extends throughout the area (Dutcher and Moyle 1973).

## Methods

The available major solute data (spring, surface, and well water) for the area was compiled for this study in order to create a comprehensive database, called SLH-DATA (south Lahontan hydrochemical database), for the

classification of waters into hydrochemical facies representing “water types”. The data were arranged in rows (for sampling locations) and in columns (for chemical parameters). The entire database consists of chemical analyses of 152 spring samples, 153 surface samples, and 1,063 well (groundwater) samples, including temporal samples (samples collected over a period of time at the same location). Sources of the data are presented in Table 1. In the case of multiple samples from the same location, the more recent and/or the more complete sample data were included in the statistical analysis unless evaluation of temporal effects was desired. Database construction procedures and comparison of the results from the various statistical and graphical techniques is discussed in detail in the following sections. Detailed analysis of the graphical and statistical water groups in terms of the physical and chemical factors that control water chemistry is not the focus of this paper. Instead, we are interested here in the ability of available techniques to classify a diverse set of samples into distinct groups.

Of the 39 hydrochemical variables (consisting of major ions, minor ions, trace elements, and isotope data) in the compiled database, 11 variables (specific conductance, pH, Ca, Mg, Na, K, Cl, SO<sub>4</sub>, HCO<sub>3</sub>, SiO<sub>2</sub>, and F) occur most often and, thus, were used in our evaluation. It is usually assumed that adequate quality assurance (QA) and quality control (QC) measures were performed

**Table 1** Data sources used to create the SLHDATA database. NWIS: US Geological Survey National Water Storage and Retrieval System

Code number	Data sources	Number of samples		
		Surface	Spring	Well
1	Barnes et al. (1981)	—	1	—
2	Berenbrock (1987)	—	—	194
3	Berenbrock and Schroeder (1994)	—	—	108
4	Buono and Packard (1982)	—	—	3
5	California State University, Bakersfield, unpublished data	51	38	74
6	Dockter (1980a)	—	—	1
7	Dockter (1980b)	—	—	2
8	Feth et al. (1964)	—	1	—
9	Font (1995)	—	3	18
10	Fournier and Thompson (1980)	1	8	3
11	Hollett et al. (1991)	—	—	5
12	Houghton (1994)	10	2	25
13	Hunt et al. (1966)	—	1	—
14	Johnson (1993)	—	—	13
15	Johnson et al. (1991)	—	—	6
16	Lamb et al. (1986)	—	—	42
17	Lopes (1987)	—	14	14
18	Maltby et al. (1985)	—	—	45
19	McHugh et al. (1981)	70	8	—
20	Melack et al. (1985)	10	—	—
21	Miller (1977)	2	7	2
22	Moyle (1963)	—	1	137
23	Moyle (1969)	—	24	96
24	Moyle (1971)	—	—	57
25	Ostdick (1997)	6	14	14
26	Robinson and Beetem (1975)	—	1	—
27	US Bureau of Reclamation (1993)	—	—	33
28	US Geological Survey NWIS QW data	—	28	154
29	Whelan et al. (1989)	—	1	9
30	No source information available	3	—	8

at the time of original data collection and analysis; however, we screened the data to verify that they were usable (see below for further discussion). The data collection methods, which are similar are described in detail for most of the data sources, or documented in US Geological Survey's "Techniques of water-resources investigations" manuals (e.g., Brown et al. 1970; Wood 1981).

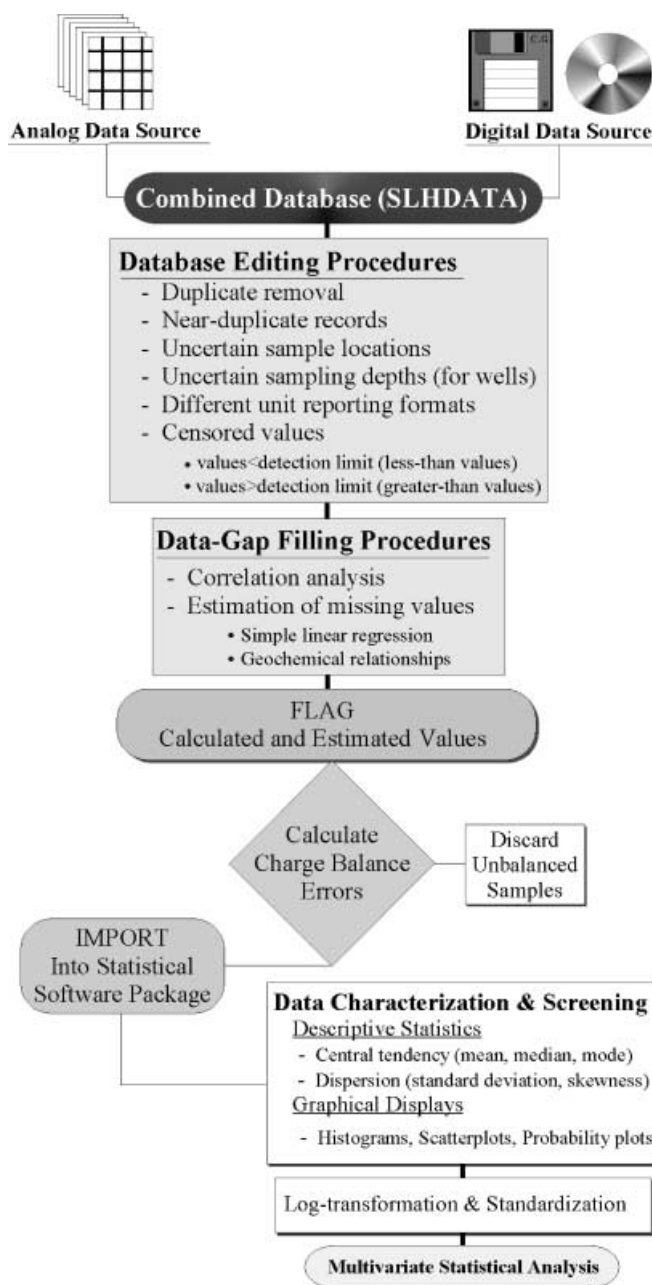
The Statistica Release 5.0 (StatSoft, Inc. 1995) commercial software package was utilized for the basic statistical analyses performed. Microsoft Excel 97 (Microsoft Corporation 1985) and RockWorks (RockWare, Inc. 1999) were used for the graphical analyses. Classification of the data was also performed using FuzME (Fuzzy k-Means with Extragrades; Minasny and McBratney 1999). The techniques used include cluster analysis (HCA and KMC), principle components analysis (PCA), fuzzy k-means clustering (FKM), and a variety of graphical methods. Detailed technical descriptions of HCA, KMC, and PCA techniques and a description of the FKM technique are provided in StatSoft, Inc. (1997) and Bezdek et al. (1984), respectively.

### Database Editing

Figure 2 is a flow chart that summarizes the methodology used for compiling the hydrochemical database. If reported, field measurements of alkalinity and pH were used for the construction of the database. Otherwise, laboratory measurements of these variables were used. Some of the individual data sets contained the same sample, or apparent near-duplicate analyses for minor elements. In general, there were more discrepancies for iron than for any other minor element. This was probably because of the different ways in which iron concentrations were expressed, or the convention being used was not clearly stated. We have chosen not to use any minor element data for our study because of these sorts of problems.

Samples with uncertain locations were located using reports and maps or eliminated from the database when such information was not available. Locations of sites that had only the name of the well or spring were determined as accurately as possible, usually to several hundred meters and always within 1 km. Well-water samples without sampling depths were retained in the database, but eliminated from the multivariate statistical analysis.

Units of measurement were sometimes inconsistent between different data sets. All values were converted to an internally consistent format (all units are in  $\text{mg L}^{-1}$  in the SLHDATA database). Common reporting units for data sets were weight-per-volume units [milligrams per liter ( $\text{mg L}^{-1}$ ) and micrograms per liter ( $\mu\text{g L}^{-1}$ )], equivalent-weight units [milliequivalents per liter ( $\text{meq L}^{-1}$ ) and microequivalents per liter ( $\mu\text{eq L}^{-1}$ )], or weight-per-weight units [parts per million (ppm), parts per billion (ppb), and parts per trillion (ppt)]. Conversion factors for calculation of a unit from the other units are given by Hem (1989).



**Fig. 2** Methodology used for compiling and editing the SLHDATA database

### Censored values

Water chemistry data are frequently censored, that is, concentrations of some elements are reported as non-detected, less-than or greater-than. These values are created by the lower or upper detection limit of the instrument or method used. Censored data are not appropriate for many multivariate statistical techniques. Therefore, the non-detected, less-than, and greater-than values must be replaced with unqualified values (Farnham et al. 2002). In our database there were no censored values for the 11 variables used in this study, however, because this is often not the case we briefly discuss methods to deal with this situation.



A number of techniques have been suggested for replacement of a censored value including replacement of the less-than values by 3/4 times the lower detection limit and the greater-than values by 4/3 times upper detection limit (VanTrump and Miesch 1977). An alternative is replacement of less-than values by 0.55 times the lower detection limit and the greater-than values by 1.7 times upper detection limit (Sanford et al. 1993). For data where the proportion of the censored values is >10%, another method that was devised by Sanford et al. (1993) can be used. This method estimates the mean of the normal distribution using a maximum likelihood estimation method. Then, this estimated mean is used to derive an estimated replacement value.

#### *Data-gap filling procedures – estimation of the missing values*

Usually the effective use of many of the methods requires complete water analyses (no missing data values). Missing data values may make the use of graphical water chemistry techniques impossible, or limit the quality of the statistical analysis. During the statistical analysis, most statistical software packages replace those missing values with means of the variables, or prompt the user for case-wise deletion of analytical data, both of which are not desirable. This can bias statistical analyses if these values represent a significant number of the data being analyzed.

There are statistical methods and chemical relationships that can be employed to estimate missing data values. For instance, missing conductance data can be calculated from total dissolved solids (TDS) data by using a simple linear regression method. In our database, a significantly ( $p < 0.001$ ) high correlation coefficient ( $r = 0.984$ ) was found to exist between these two variables. The  $p$ -value is the significance probability for testing the null hypothesis that true correlation in the population is zero. A small value of  $p$  (e.g.,  $p < 0.001$ ) indicates that there is a significant correlation. Thus, missing potassium (K) values were estimated by utilizing the linear relationship between potassium and sodium (Na), which had a significantly ( $p < 0.001$ ) high correlation coefficient of 0.904.

Missing bicarbonate ( $\text{HCO}_3$ ) data can be calculated from alkalinity values and pH. Inverting the problem, missing pH values can be calculated by using Eq. (1) if the  $\text{CO}_3$  and  $\text{HCO}_3$  values were reported:

$$K_2 = \frac{[\text{CO}_3^{2-}] \cdot 10^{-\text{pH}}}{[\text{HCO}_3^-]} \quad (1)$$

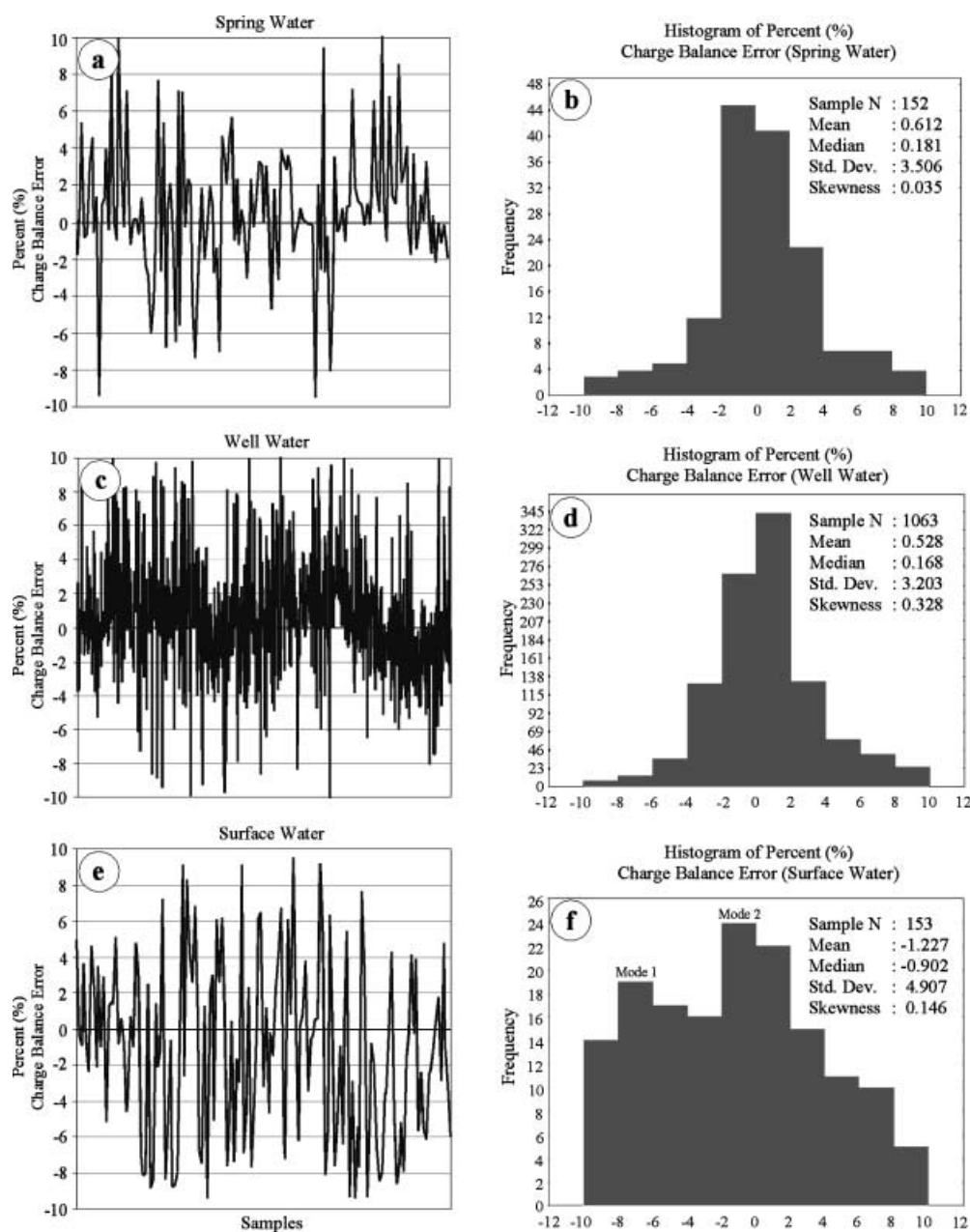
The same relationship can also be used to calculate the missing carbonate ( $\text{CO}_3$ ) values if pH was reported. Finally, if there were no means of establishing a value, a value of “–9,999” was entered for the missing value, indicating that no data were available for that entry. In our data set there were very few (3%) samples with censored or missing values.

**Table 2** Charge balance (CB) statistics for the individual data sources

Code no. <sup>a</sup>	Years collected	–	+	CB error range		Median	Mean ( $\pm 1\sigma$ )
1	1981	1	0	–1.04	–	–1.04	–1.04 ( $\pm 0.00$ )
2	1977–1984	82	112	–10.31	7.75	0.35	0.16 ( $\pm 2.87$ )
3	1987–1989	54	54	–9.43	8.66	–0.11	0.24 ( $\pm 3.95$ )
4	1968–1980	2	1	–2.05	0.96	–1.31	–0.80 ( $\pm 1.57$ )
5	1994–1998	61	102	–9.44	10.04	1.33	1.88 ( $\pm 3.73$ )
6	1978	0	1	–	0.32	0.32	0.32 ( $\pm 0.00$ )
7	1978	1	1	–0.02	0.52	0.25	0.25 ( $\pm 0.38$ )
8	1959	1	0	–1.14	–	–1.14	–1.14 ( $\pm 0.00$ )
9	1989–1994	14	7	–3.70	3.03	–1.25	–0.83 ( $\pm 2.01$ )
10	1974–1979	7	5	–5.71	9.41	–0.30	0.82 ( $\pm 4.31$ )
11	1945–1978	1	4	–0.12	5.30	1.69	2.03 ( $\pm 2.15$ )
12	1993–1994	12	25	–8.39	10.40	0.87	0.90 ( $\pm 4.01$ )
13	Unknown	0	1	–	0.34	0.34	0.34 ( $\pm 0.00$ )
14	1993	9	4	–8.94	8.16	–1.43	–1.04 ( $\pm 4.34$ )
15	1990	4	2	–3.30	8.33	–0.28	0.56 ( $\pm 4.34$ )
16	1986–1991	34	8	–5.14	2.05	–2.20	–1.82 ( $\pm 1.80$ )
17	1986	10	18	–2.65	10.39	0.84	1.35 ( $\pm 2.92$ )
18	1984–1985	30	15	–6.02	8.51	–0.86	–0.92 ( $\pm 2.97$ )
19	1979	64	14	–9.45	9.10	–5.00	–4.02 ( $\pm 4.38$ )
20	1982	7	3	–6.17	4.75	–2.68	–2.01 ( $\pm 3.56$ )
21	1967–1972	4	7	–2.31	4.67	0.79	0.89 ( $\pm 2.02$ )
22	1917–1960	28	110	–2.34	10.13	0.13	1.13 ( $\pm 2.23$ )
23	1917–1967	17	103	–9.55	8.74	1.89	2.00 ( $\pm 2.28$ )
24	1916–1969	22	35	–3.81	8.84	0.20	0.74 ( $\pm 1.94$ )
25	1996	17	17	–8.70	10.32	0.03	1.14 ( $\pm 5.32$ )
26	1965	1	0	–0.09	–	–0.09	–0.09 ( $\pm 0.00$ )
27	1990–1992	8	25	–3.13	6.88	2.38	1.99 ( $\pm 2.55$ )
28	1945–1990	119	63	–8.10	9.92	–0.94	–0.75 ( $\pm 2.69$ )
29	1976–1987	1	9	–0.76	9.99	6.54	5.18 ( $\pm 3.52$ )
30	1989–1996	3	8	–10.04	9.58	2.59	2.35 ( $\pm 5.63$ )

<sup>a</sup> Data sources corresponding to code numbers are listed in Table 1. The + and – columns refer to the number of analyses from a data source having positive and negative CB errors

**Fig. 3** Distribution of percent charge balance error for **a, b** spring, **c, d** well, and **e, f** surface water samples



#### Charge balance error

The edited chemical analyses in the SLHDATA database were tested for charge balance (Freeze and Cherry 1979):

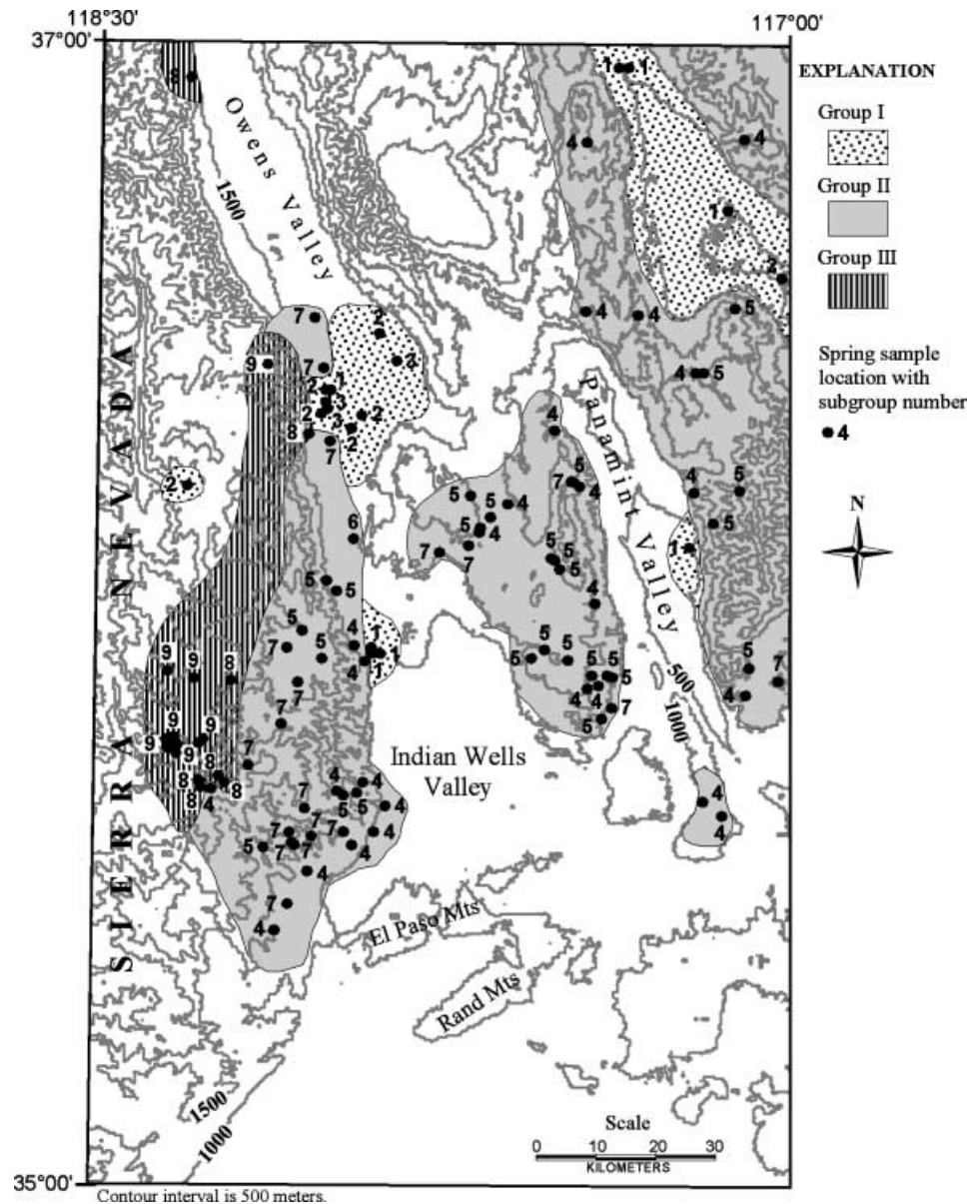
$$\% \text{Charge Balance (CB) Error} = \frac{\sum z \cdot m_c - \sum z \cdot m_a}{\sum z \cdot m_c + \sum z \cdot m_a} \cdot 100 \quad (2)$$

where  $z$  is the absolute value of the ionic valence,  $m_c$  the molality of cationic species and  $m_a$  the molality of the anionic species. Conventions and assumptions used in balancing the analyses included:

1. When bicarbonate and carbonate data were not given, alkalinity, if available, was used to estimate a bicarbonate concentration.
2. In the few cases where the calcium and magnesium data were missing, hardness was used to estimate the sum of calcium and magnesium concentrations.

Calculated charge balance errors are less than or equal to  $\pm 10.4\%$  for SLHDATA database, which is an acceptable error for the purpose of this study (Table 2). Samples with errors greater than  $\pm 10.4\%$  were not used. For the spring-water and well-water (groundwater) data, errors are evenly distributed between positive and negative values, and, thus, are not systematic (Fig. 3a–d). The charge balance errors of the surface water data showed a bimodal distribution and had a skewed distribution (Fig. 3e, f). Accordingly, the surface water samples were further studied. The samples were split into those from

**Fig. 4** Map view of HCA-derived subgroup and group values for the spring water samples



the Sierra Nevada mountain block and the Indian Wells-Owens Valley area (see Fig. 4 for locations). The Indian Wells-Owens Valley area samples have charge balance errors that approach a normal distribution and range from  $-6\%$  to  $+10\%$ , whereas the Sierra Nevada samples had a strongly skewed distribution. The Sierra Nevada data included 78 samples collected for the Domeland Wilderness study (McHugh et al. 1981), which were identified as the source of the skewed distribution and indicates a systematic error in that particular set of analyses. However, the error is not sufficient to remove the data set from the database.

#### Data screening

The purpose of data screening is to evaluate the distribution characteristics of each variable in the database.

We used univariate and bivariate statistical methods to assess each variable independently, and the relationship between variable pairs. The physical and chemical properties were evaluated using central tendency (mean, median, mode) and dispersion (standard deviation, skewness), and by graphical displays such as histograms, scatter plots, probability plots, and box plots. Based on these analyses, decisions were made concerning the need for, and selection of, appropriate transformations to achieve a better approximation of the normal distribution. This is important because most of the statistical analyses assume that data are normally distributed.

The data screening showed that the data used in this study were universally skewed positively; the data contained a small number of high values. Most naturally occurring element distributions follow this pattern (Miesch

1976). The data were log-transformed (except for pH) so that they more closely corresponded to normally distributed data. Then, all the 11 variables were standardized by calculating their standard scores ( $z$ -scores) as follows:

$$z_i = \frac{x_i - \bar{x}}{s} \quad (3)$$

where  $z_i$  = standard score of the sample  $i$ ;  $x_i$  = value of sample  $i$ ;  $\bar{x}$  = mean;  $s$  = standard deviation.

Standardization scales the log-transformed data to a range of approximately  $-3$  to  $+3$  standard deviations, centered about a mean of zero. In this way, each variable has equal weight in the statistical analyses. Otherwise, the Euclidean distances will be influenced most strongly by the variable that has the greatest magnitude (Judd 1980; Berry 1995). Besides normalizing and reducing outliers, these transformations also tend to homogenize the variance of the distribution (Rummel 1970). The raw data (with data-gaps filled) were used for the graphical analyses, whereas the transformed (log-transformed and standardized) data were used for the hierarchical cluster analysis (HCA), K-means cluster analysis (KMC), principal components analysis (PCA), and fuzzy k-means clustering (FKM).

## Results

The fundamental aim of the techniques compared here is to identify the chemical relationships between water samples. Samples with similar chemical characteristics often have similar hydrologic histories, similar recharge areas, infiltration pathways, and flow paths in terms of climate, mineralogy, and residence time. Table 3 shows

the various techniques and the required input data. For brevity, only the 152 spring water samples are discussed in the following text. The other subsets of the complete database produced similar results. A preliminary analysis of temporal effects, based on examination of individual analyses, suggested that relatively little change occurred in the water quality of samples with time. This indicates that the spatial variability is the most important source of variation in the data, rather than the temporal factor. This conclusion was later tested and confirmed as discussed in the statistical methods section. For that reason, we did not include samples from temporal series to statistical analysis. This reduced the total number of samples to 118. The fundamental aim of the techniques compared here is to identify the chemical relationships between water samples. Samples with similar chemical characteristics often have similar hydrologic histories, similar recharge areas, infiltration pathways, and flow paths in terms of climate, mineralogy, and residence time. Table 3 shows the various techniques and the required input data. For brevity, only the 152 spring water samples are discussed in the following text. The other subsets of the complete database produced similar results. A preliminary analysis of temporal effects, based on examination of individual analyses, suggested that relatively little change occurred in the water quality of samples with time. This indicates that the spatial variability is the most important source of variation in the data, rather than the temporal factor. This conclusion was later tested and confirmed as discussed in the statistical methods section. For that reason, we did not include samples from temporal series to statistical analysis. This reduced the total number of samples to 118.

**Table 3** Statistical and graphical techniques evaluated for the classification of water samples

Method	Cations used	Anions used	Other parameters	Input data and plotting units
Cluster analysis (HCA and KMC)	All major, minor and trace elements	All major, minor and trace elements	All applicable parameters Yes (1) or no (0) statements, discrete variables	Input: $z$ -scores of the log-transformed data Output: distance matrix (KMC) and dendrogram (HCA)
Principal components analysis (PCA)	All major, minor and trace elements	All major, minor and trace elements	All applicable parameters Yes (1) or no (0) statements, discrete variables	Input: $z$ -scores of the log-transformed data Output: PCA scores
Fuzzy k-means Clustering (FKM)	All major, minor and trace elements	All major, minor and trace elements	Same as above	Input: same as above matrix Output: membership
Piper diagram	Na + K, Ca, Mg	Cl, SO <sub>4</sub> , HCO <sub>3</sub> + CO <sub>3</sub>	n/a	Relative %meq L <sup>-1</sup>
Collins bar diagram	Na + K, Ca, Mg	Cl, SO <sub>4</sub> , HCO <sub>3</sub> (or HCO <sub>3</sub> + CO <sub>3</sub> )	n/a	Relative %meq L <sup>-1</sup> or meq L <sup>-1</sup>
Pie diagram	Na + K, Ca, Mg	Cl, SO <sub>4</sub> , HCO <sub>3</sub>	n/a	Relative %meq L <sup>-1</sup>
Stiff pattern diagram	Na (or Na + K), Ca, Mg Fe (optional)	Cl, SO <sub>4</sub> , HCO <sub>3</sub> CO <sub>3</sub> (optional)	n/a	meq L <sup>-1</sup>
Schoeller semi-logarithmic diagram	Na + K, Ca, Mg	Cl, SO <sub>4</sub> , HCO <sub>3</sub>	n/a	meq L <sup>-1</sup> in log-scale
Chernoff faces	Up to 20 parameters can be plotted			meq L <sup>-1</sup> or mg L <sup>-1</sup> Other parameters in their respective units



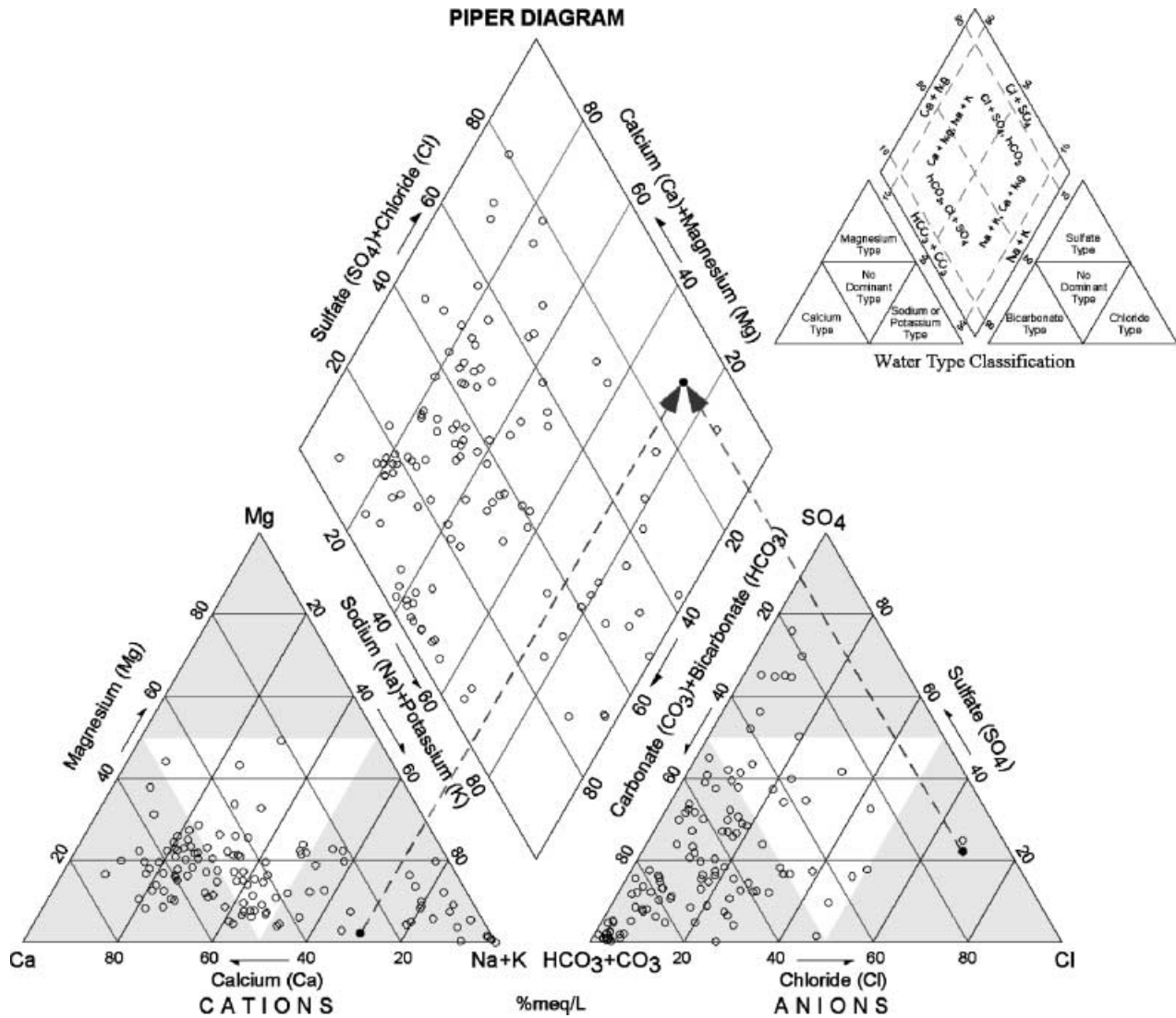


Fig. 5 Piper diagram of the 118 spring water samples

### Graphical Methods

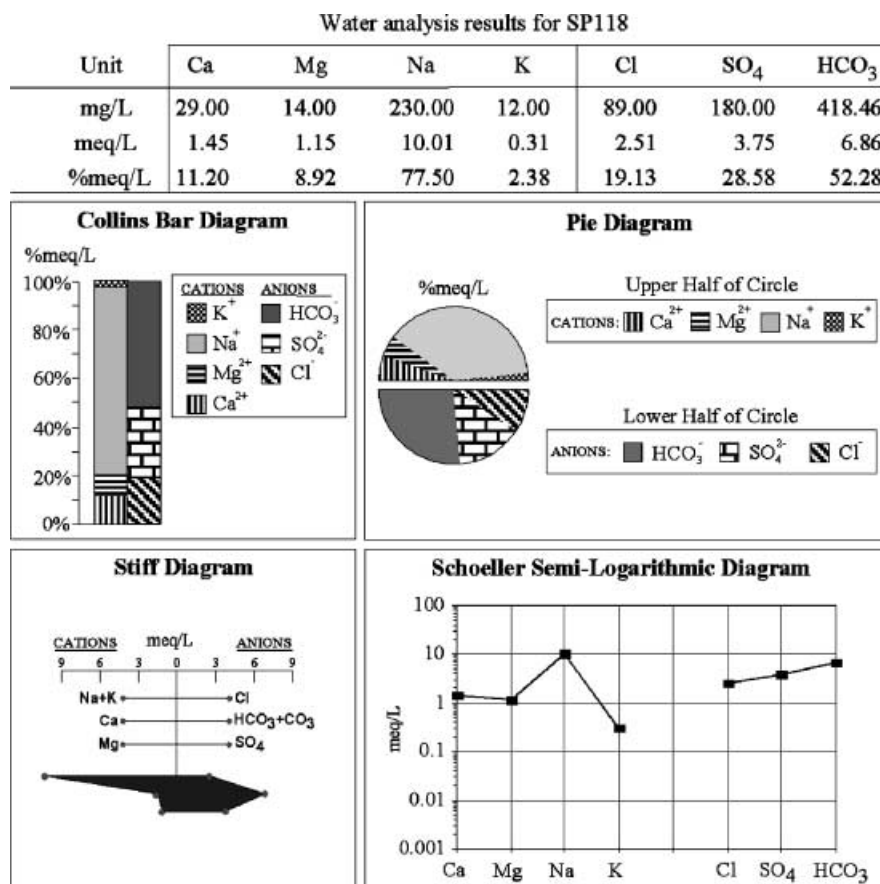
Most of the graphical methods are designed to simultaneously represent the total dissolved solid concentration and the relative proportions of certain major ionic species (Hem 1989). All the graphical methods use a limited number of parameters, usually a subset of the available data, unlike the statistical methods that can utilize all the available parameters. The Piper diagram (Piper 1944; Fig. 5) is the most widely used graphical form and it is quite similar to the diagram proposed by Hill (1940, 1942). The diagram displays the relative concentrations of the major cations and anions on two separate trilinear plots, together with a central diamond plot where the points from the two trilinear plots are projected. The central diamond-shaped field (quadrilateral field) is used to show overall chemical character of the water (Hill 1940; Piper 1944). Back (1961) and Back and Hanshaw (1965) defined subdivisions of the diamond field, which

represent water-type categories that form the basis for one common classification scheme for natural waters. The mixing of water from different sources or evolution pathways can also be illustrated by this diagram (Freeze and Cherry 1979). Symbol sizes can be scaled to TDS on the diamond-shaped field to add even more information (Domenico and Schwartz 1997).

Figure 5 shows the results of plotting the 118 spring samples on the Piper diagram. The data are broadly distributed rather than forming distinct clusters. Employing the water classification scheme of Back and Hanshaw (1965), the samples are classified into a variety of water types including Ca-HCO<sub>3</sub>, Ca-Mg-HCO<sub>3</sub>, Ca-Na-HCO<sub>3</sub>, Na-HCO<sub>3</sub>, Na-Ca-HCO<sub>3</sub>, Na-Cl and Ca-Mg-SO<sub>4</sub> types, with no dominant type. This diagram provides little information that allows us to discriminate between separate clusters (groups) of samples.

The Collins bar diagram (Collins 1923) and the pie diagram (Fig. 6) are easy to construct and present relative major ion composition in percent milliequivalents per liter (relative %meq L<sup>-1</sup>). The constituents can also

**Fig. 6** Plots for a single sample using several different graphical methods (Collins, pie, Schoeller and Stiff)



be plotted in meq L<sup>-1</sup> with an appropriate scaling. For the Collins bar diagram, major cations are plotted on the left and major anions are plotted on the right. For the pie diagram, the cations are plotted in the upper half and anions are plotted in the lower half of the circle. The pie diagram is usually drawn with a radius proportional to TDS.

The Stiff pattern (Fig. 6) is a polygon that is created from three (or four) parallel horizontal axes extending on either side of a vertical zero axis (Stiff 1951). In this diagram, cations are plotted on the left of the axes and anions are plotted on the right, in units of milliequivalents per liter (meq L<sup>-1</sup>). The Stiff diagram is usually plotted without the labeled axis and is useful making visual comparison of waters with different characteristics. The patterns tend to maintain its shape upon concentration or dilution, thus visually allowing us to trace the flow paths on maps (Stiff 1951).

The Schoeller semi-logarithmic diagram (Schoeller 1955, 1962; Fig. 6) allows the major ions of many samples to be represented on a single graph, in which samples with similar patterns can be easily discriminated. The Schoeller diagram shows the total concentration of major ions in log-scale.

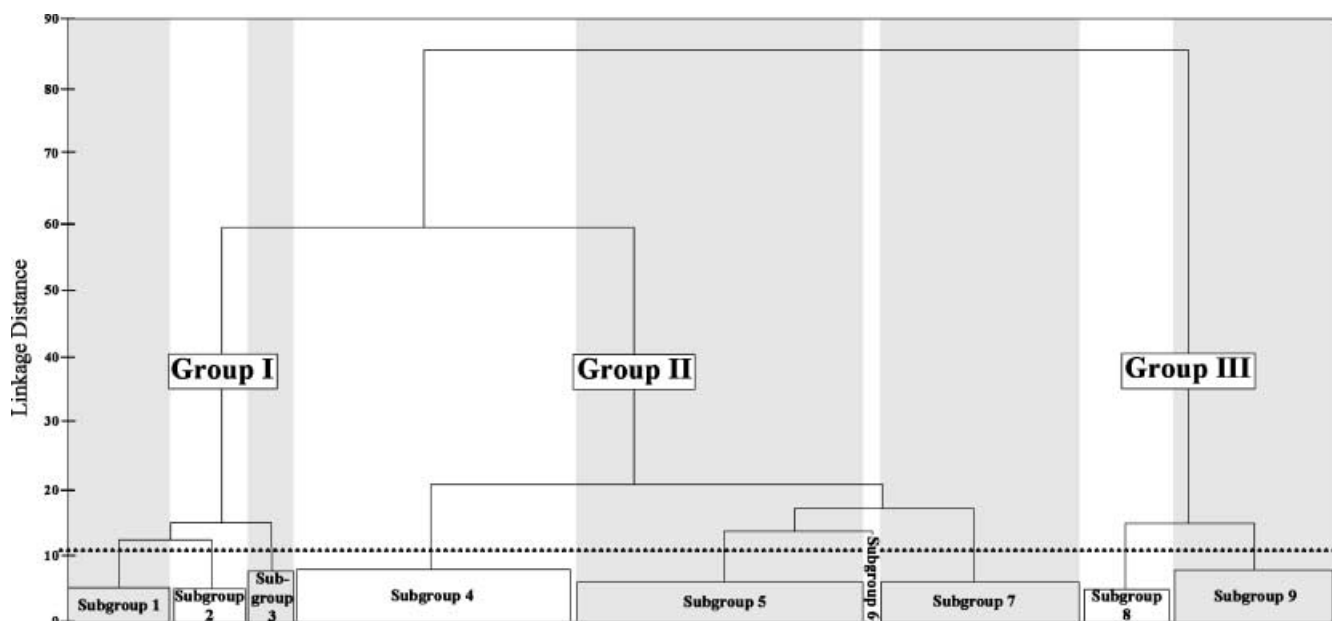
As we can see from Fig. 6, the Collins, pie, and Stiff methods produce a single diagram for each sample. Clearly, it is not practical to produce and manually sort 118 separate figures (e.g., Stiff diagrams), one for each

sample, in order to sort and classify large data sets. The choice of similarity would be based on the evaluation of the analyst, which is highly subjective. Therefore, we suggest that using purely graphical methods to group the samples is not efficient and can produce biased results. However, these methods are useful for presentation of maps showing hydrochemical facies, and software is available (e.g., RockWorks) to automatically and rapidly prepare such maps.

### Multivariate Statistical Techniques

Another approach to understanding the chemistry of water samples is to investigate statistical relationships among their dissolved constituents and environmental parameters, such as lithology, using multivariate statistics (Drever 1997). Statistical associations do not necessarily establish cause-and-effect relationships, but do present the information in a compact format as the first step in the complete analysis of the data and can assist in generating hypothesis for the interpretation of hydrochemical processes.

Statistical techniques, such as cluster analysis, can provide a powerful tool for analyzing water-chemistry data. These methods can be used to test water quality data and determine if samples can be grouped into distinct populations (hydrochemical groups) that may be significant in the geologic context, as well as from a statistical



**Fig. 7** Dendrogram from the HCA for the 118 spring water samples. *Line of asterisks* defines “phenon line”, which is chosen by analyst to select number of groups or subgroups

point of view. Cluster analysis was successfully used, for instance, to classify lake samples into geochemical facies (Jaquet et al. 1975). Alther (1979), Williams (1982), and Farnham et al. (2000) also applied cluster analysis to classify water-chemistry data.

The assumptions of cluster analysis techniques include homoscedasticity (equal variance) and normal distribution of the variables (Alther 1979). Equal weighing of all variables requires the log-transformation and standardization ( $z$ -scores) of the data, as discussed above. Comparisons based on multiple parameters from different samples are made and the samples grouped according to their “similarity” to each other. The classification of samples according to their parameters is termed Q-mode classification. This approach is commonly applied to water-chemistry investigations in order to define groups of samples that have similar chemical and physical characteristics because rarely is a single parameter sufficient to distinguish between different water types.

Both the hierarchical cluster analysis (HCA) and K-means clustering (KMC) were used to classify the samples into distinct hydrochemical groups based on their similarity. In order to determine the relation between groups, the  $r \times c$  data matrix ( $r$  samples with  $c$  variables) is imported into a statistics package. The Statistica (StatSoft, Inc. 1995) has seven similarity/dissimilarity measurements and seven linkage methods and supports up to 300 cases for the amalgamation process in the cluster analysis. Individual samples are compared with the specified similarity/dissimilarity and linkage methods and then grouped into clusters.

The linkage rule used here is Ward’s method (Ward 1963). Linkage rules iteratively link nearby points (sam-

ples) by using the similarity matrix. The initial cluster is formed by linkage of the two samples with the greatest similarity. Ward’s method is distinct from all other methods because it uses an analysis of variance (ANOVA) approach to evaluate the distances between clusters. Ward’s method calculates the error sum of squares, which is the sum of the distances from each individual to the center of its parent group (Judd 1980) and forms smaller distinct clusters than those formed by other methods (StatSoft, Inc. 1995).

Similarity/dissimilarity measurements and linkage methods used for clustering greatly affects the outcome of the HCA results. After careful examinations of available combinations of similarity/dissimilarity measurements, it was found that using Euclidean distance (straight line distance between two points in  $c$ -dimensional space defined by  $c$  variables) as similarity measurement, together with Ward’s method for linkage, produced the most distinctive groups where each member within the group is more similar to its fellow members than to any member from outside the group. The HCA technique does not provide a statistical test of group differences; however, there are tests that can be applied externally for this purpose (e.g., Student’s  $t$ -test). It is also possible in HCA results that one single sample that does not belong to any of the groups is placed in a group by itself. This unusual sample is considered as residue.

HCA classifies the data in a relatively simple and direct manner, with the results being presented as a dendrogram, an easily understood and familiar diagram (Davis 1986). In the present case, we selected the number of groups based on visual examination of the dendrogram (Fig. 7). The resulting dendrogram was interpreted to have classified the 118 spring water samples into three major groups (I–III) and nine subgroups (1–9) using 11 variables; this, however, is a subjective evaluation. Greater or fewer groups could be defined by moving the

**Table 4** Mean water chemistry of the spring water subgroups determined from HCA. pH (standard units); specific conductance ( $\mu\text{Siemens cm}^{-1}$ ), mean concentrations ( $\text{mg L}^{-1}$ )

Group	Sub-group	<i>n</i> <sup>a</sup>	pH	S. cond.	Ca	Mg	Na	K	Cl	SO <sub>4</sub>	HCO <sub>3</sub>	SiO <sub>2</sub>	F	TDS
I	1	10	7.92	1,657.00	53.99	32.01	261.23	22.01	224.16	175.51	453.39	46.07	2.20	1,063.45
	2	7	7.04	6,264.17	70.69	75.10	1,287.00	77.64	1,165.71	433.71	1,541.14	101.83	1.70	4,347.86
	3	4	9.12	4,160.00	4.40	3.44	967.75	77.25	464.50	336.25	1,122.25	50.50	—	3,206.34
II	4	27	7.70	855.10	95.91	29.09	70.76	3.92	46.34	211.92	274.82	36.44	1.24	646.22
	5	28	7.92	550.19	62.96	15.20	32.87	3.20	25.64	68.69	220.58	25.10	0.26	344.17
	6	1	8.08	400.00	25.43	5.76	44.57	5.83	40.90	16.49	155.00	0.61	0.00	200.00
	7	18	7.09	397.79	42.26	8.57	29.28	2.80	15.20	37.96	177.50	37.34	0.59	308.11
III	8	8	8.03	272.57	22.38	1.81	30.99	1.03	5.61	19.70	124.37	32.29	2.62	205.18
	9	15	7.24	92.50	11.59	1.21	12.10	0.92	1.23	0.82	70.56	22.23	0.75	70.73

<sup>a</sup> Number of samples within subgroups

dashed horizontal line (phenon line) up or down. In addition, the dendrogram does not give information about the distribution of the chemical constituents that form each group: a distinct limitation when compared with the graphical techniques. The differences among subgroups defined by the HCA (Fig. 7) were determined to be statistically significant ( $p < 0.001$ ), except the subgroups 2 and 3 of group I, which were significant only at  $p < 0.05$ .

Table 4 shows the means for each of the parameters produced by the HCA analysis. These values reveal some trends between the major groups. Group I samples all have significantly higher TDS than group II or III samples. Subgroup 6 has only one member (Table 4, Fig. 7), a sample that is distinguished by an abnormally low SiO<sub>2</sub> value. This value is probably an analytical or typographic error, and was removed from the database. Groups II and III also appear to be separated based on TDS. The basis for the division into subgroups is not so apparent. For instance, subgroups 1 and 2 appear to be distinguished from subgroup 3 by the lower pH values and higher Ca and Mg values. However, the differences between subgroups 5 and 7 are subtle.

At this point, it is fair to ask if these clusters of samples have any physical significance/meaning, or are just a statistical result. The relationship of the statistically defined clusters of samples to geographic location was tested by plotting the subgroup value for each sample on a site map (Fig. 4). The figure shows that there is a good correspondence between spatial locations and the statistical groups as determined by the HCA. For instance, the spring samples composing group I are usually found close to playa or nearby discharge areas on the basin floors and have the highest TDS concentration in the area (Table 4). Group II samples are mostly located below the 2,000-m contour line in the Sierra Nevada and also found at the ranges surrounding the valleys (Fig. 4). Group III samples plot above the 2,000-m contour line in the high Sierra Nevada and are characterized by low TDS concentrations (Table 4). The majority of recharge to the basin-fill aquifers occurs from areas where group II and III samples are located. It appears that the technique can provide valuable information to help define the hydrologic system. For instance, the high degree of

spatial and statistical coherence in this data set could be used to support a model of hydrochemical evolution where the changes in water chemistry are a result of increasing rock–water interactions along hydrological flow paths.

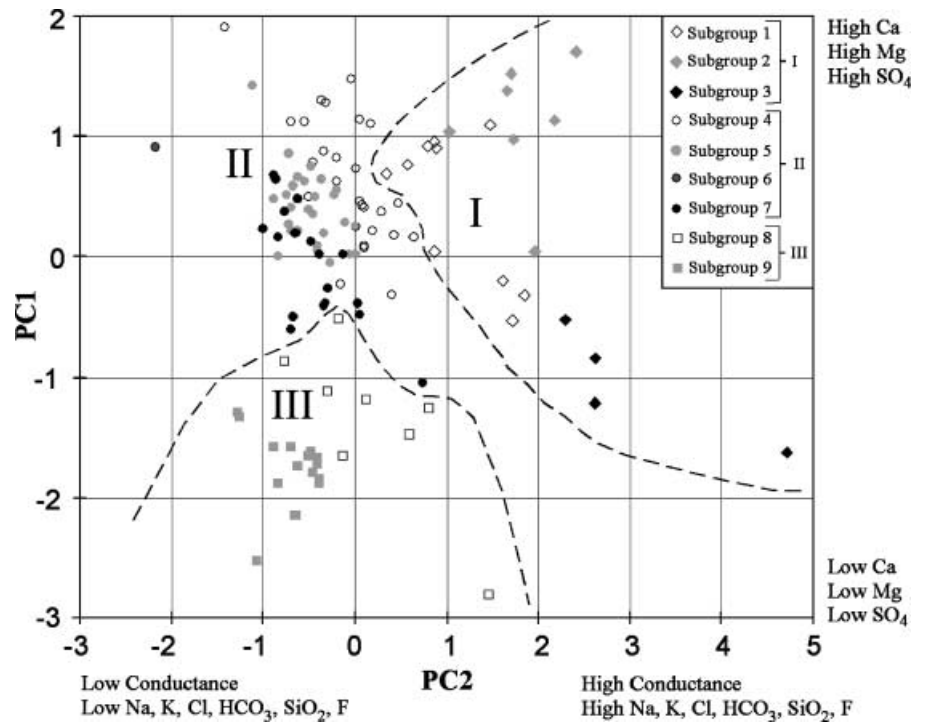
K-means clustering (KMC) has also been used to classify water samples into distinct hydrochemical groups (Johnson and Wichern 1992). This method of clustering is different from the HCA because the number of clusters is pre-selected at the start of the analysis, producing a subjective bias. The KMC method will produce exactly K different clusters with the greatest possible distinction. Computationally, this method can be thought of as an analysis of variance (ANOVA) in reverse. The clustering starts with K random clusters, and then moves objects between those clusters with the goal to (1) minimize variability within clusters, and (2) maximize variability between clusters (StatSoft, Inc. 1995). Unlike HCA, the results from KMC cannot be presented in a dendrogram for a quick visual assessment of the results. Instead, the results are presented in a large table that shows members of clusters and their distances from respective cluster centers.

As discussed previously, we did not include samples from temporal series because the preliminary analysis showed that there was little temporal variation. To verify that analysis, we included the entire 152 spring samples in an hierarchical cluster analysis (HCA) and examined the resulting dendrogram. The dendrogram had the temporal series samples placed together, suggesting that little change occurs in the water quality with time period of sampling. This agrees with the preliminary analysis that spatial variability is the most important source of variation in the data.

Another type of data analysis sometimes used is principal components analysis (PCA). This technique reduces the number of dimensions present in data (reducing 11 variables to 2 variables in our study). The PCA-defined new variables can then be displayed in a scatter diagram, presenting the individual water samples as points in a lower-dimensional (generally 2-D) space. This technique, strictly speaking, is not a multivariate statistical technique, but a mathematical manipulation that may



**Fig. 8** Plot of the principal components analysis showing the distribution of HCA-derived classification of samples for the spring water



provide a certain amount of insight into the structure of the data matrix (Davis 1986) by reducing the dimensions of the data matrix. Figure 8 shows the results of the principle components analysis of the 118 samples. The first principal component (PC1) contains 54.5% of the total variance and the second component (PC2) represents 14.5% of the total variance. Although there appears to be reasonable statistical discrimination between the three major groups as defined by HCA, there is no objective means to clearly distinguish boundaries between the groups or subgroups, nor does this type of analysis provide any information about chemical composition. This method was used to investigate the degree of continuity or clustering of the samples and to determine if overlapping water types exist within the data. The scatter of points in Fig. 8 suggest that there is continuous variation of the chemical and physical properties of the samples.

The HCA clustering scheme was also repeated using just the two principal components scores (reduced two-dimensional data). The resulting classification differs very little from the first HCA classification, suggesting that employing PCA has not improved the clustering results here. However, other data sets may benefit because using lower dimensional data (defined by PCA) may improve the clustering results by reducing the redundancy in the data. The use of variables that have specific relationships can cause undesirable redundancies in cluster analysis. For instance, TDS (related to total ions present and also specific conductance), alkalinity (related to bicarbonate), and hardness (related to calcium and magnesium) were not used in our cluster analyses (HCA and KMC) because they are directly related.

### Fuzzy k-Means Clustering

Geological and hydrochemical systems are sometimes too complex to analyze easily using conventional graphical or statistical methods. Often the chemical and physical properties of the natural system vary continuously, rather than abruptly. In other words, these underlying physical and chemical processes do not always produce discrete outcomes. Because of this continuity, statistical clusters may not be well separated and instead may form a sequence of overlapping clusters. Therefore, methods related to "fuzzy logic" may be useful for modeling and classification purposes.

Application of fuzzy logic in Earth sciences is still in its early stages. On this topic, there are only a small number of papers published in the areas of geophysics, geology, petroleum, and geotechnical engineering. For example, McBratney and Moore (1985) applied fuzzy sets to climatic classification and, later, McBratney and deGruijter (1992) and Odeh et al. (1992) used the Fuzzy k-means approach for classification of soils. Nordlund (1996) applied a rule-based Fuzzy logic to model deposition and erosion processes.

Traditional Aristotelian logic (binary logic) imposes sharp boundaries (Sibigtroth 1998); however, fuzzy logic has no sharp boundaries (Fang and Chen 1990). Fuzzy logic is basically a multi-valued logic that allows intermediate values to be defined between conventional evaluations like yes/no, 0/1, true/false, black/white, and so on (Zadeh 1965; McNeill and Freiburger 1993). Fuzzy logic also allows for formalization of qualitative statements, which are widely used in Earth sciences. Both fuzziness and probability describe uncertainty numerically; however, probability treats yes/no occurrences and

**Table 5** First 20 rows of the FKM membership matrix for the spring water data. Class memberships are equivalent to HCA subgroups

Sample no.	Class <sup>a</sup>	Membership								
		1	2	3	4	5	6	7	8	9
SP33	9	0.000	0.002	0.001	0.003	0.033	0.001	0.000	0.004	<b>0.955</b>
SP35	9	0.000	0.002	0.001	0.003	0.043	0.001	0.000	0.005	<b>0.944</b>
SP36	9	0.000	0.003	0.001	0.003	0.110	0.001	0.000	0.006	<b>0.875</b>
SP34	9	0.000	0.001	0.000	0.001	0.020	0.000	0.000	0.002	<b>0.975</b>
SP52	5	0.001	0.008	0.004	0.015	<b>0.736</b>	0.002	0.001	0.017	0.216
SP32	5	0.000	0.002	0.001	0.003	<b>0.687</b>	0.001	0.000	0.005	0.300
SP29	9	0.000	0.002	0.001	0.004	0.203	0.001	0.000	0.005	<b>0.783</b>
SP31	5	0.001	0.003	0.002	0.004	<b>0.952</b>	0.001	0.000	0.006	0.031
SP27	9	0.000	0.001	0.000	0.001	0.009	0.000	0.000	0.002	<b>0.987</b>
SP28	5	0.000	0.002	0.001	0.002	<b>0.948</b>	0.000	0.000	0.003	0.043
SP37	5	0.000	0.001	0.000	0.001	<b>0.988</b>	0.000	0.000	0.001	0.008
SP116	9	0.002	0.038	0.016	0.074	0.068	0.005	0.001	0.077	<b>0.718</b>
SP86	7	0.019	0.021	0.026	0.014	0.004	0.081	<b>0.813</b>	0.018	0.005
SP47	5	0.000	0.002	0.001	0.003	<b>0.935</b>	0.001	0.000	0.006	0.050
SP30	5	0.001	0.005	0.002	0.008	<b>0.939</b>	0.001	0.000	0.009	0.034
SP24	9	0.047	0.074	0.064	0.076	0.113	0.037	0.011	0.110	<b>0.468</b>
SP38	5	0.003	0.019	0.011	0.021	<b>0.619</b>	0.006	0.002	0.057	0.262
SP21	9	0.022	0.142	0.151	0.101	0.033	0.050	0.008	0.217	<b>0.274</b>
SP20	8	0.004	0.202	0.183	0.060	0.005	0.051	0.004	<b>0.469</b>	0.021
SP25	9	0.001	0.008	0.004	0.010	0.033	0.002	0.000	0.015	<b>0.928</b>

<sup>a</sup> Class memberships on the basis of which the rows were selected are in boldface

is inherently a statistical method. Fuzziness deals with degrees and is a non-statistical method (Zadeh 1965).

One approach to fuzzy classification, and probably the best and most commonly used, is fuzzy c-means (Bezdek 1981), later renamed to fuzzy k-means (FKM) by deGrujter and McBratney (1988). This method minimizes the within-class sum of square errors. In this technique, samples may not be a 100% member of a group, instead the membership of samples are graded (partitioned) between groups. For example, a water sample may be mostly a member of a certain group, but it may be also a partial member of other groups. The analysis produces membership grades for each sample between 0 and 1. The higher the membership value for a group, the more closely the sample resembles the other members of this group. The FKM method does not impose any limitations on the number of samples or objects that can be clustered in one batch. Some clustering programs limit the amount of samples that can be clustered in one batch (e.g., MVSP: Kovach 1990, 100 samples; and Statistica: StatSoft, Inc. 1995, 300 samples). Others use a two-step approach (pre-clustering and clustering) to cluster samples (SAS Institute Inc. 1988). In this respect, FKM may provide a better tool for clustering a larger data set (e.g., combination of spring, surface, and well-water data) with overlapping or continuous clusters.

We employed the program FuzME (Minasny and McBratney 1999), which uses Brent's algorithm (Press et al. 1992), when searching for an optimal value (deGrujter and McBratney 1988). In this method, a parameter called "fuzziness exponent" ( $f$ ) is selected before application of the method. It determines the degree of fuzziness of the final solution, which is the degree of overlap between groups. With the minimum meaningful value of  $f=1$ , the solution is a hard partition, that is, the result obtained is not fuzzy at all. As  $f$  approaches infinity ( $\infty$ ) the solution ap-

proaches its highest degree of fuzziness (Bezdek 1981). For most data,  $1.5 \leq f \leq 3.0$  produces satisfactory results (Bezdek et al. 1984). The fuzzy k-means algorithm is applied as follows (Minasny and McBratney 1999):

1. Choose the number of classes  $K$  (which is equivalent to HCA subgroups), with  $1 < K < n$ .
2. Choose a value for the fuzziness exponent  $f$ , with  $f > 1$ .
3. Choose a definition of distance in the variable-space (Euclidean, diagonal, or Mahalanobis distance).
4. Choose a value for the stopping criterion  $e$  (e.g.,  $e=0.001$  gives reasonable convergence).
5. Initialize with random memberships or with memberships from a hard K-means partition (e.g., HCA or KMC).

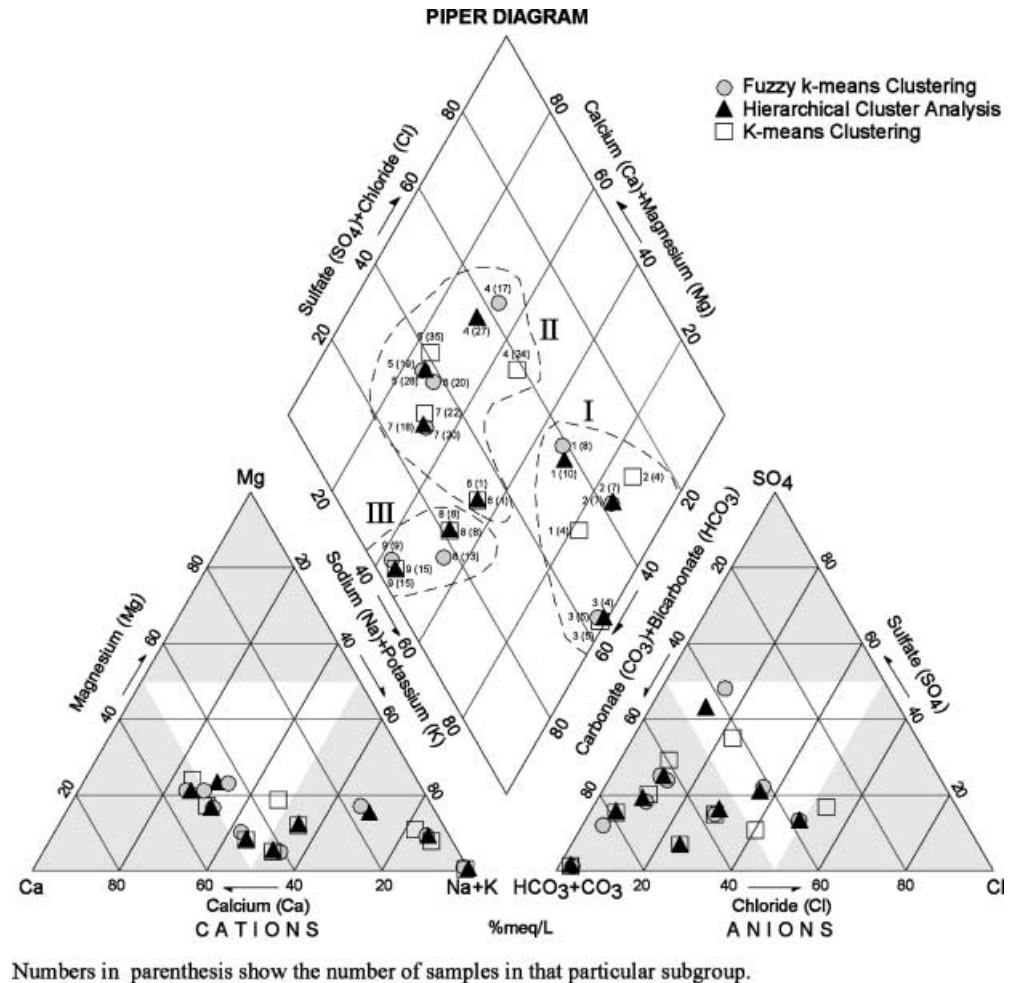
Odeh et al. (1992) suggested methods for choosing fuzziness exponent and number of classes. For our study, a value of 1.5 was used for the fuzziness exponent ( $f$ ) and Euclidean distance was chosen as the distance measure. Like the KMC, the selection of the optimal number of groups was based on the results of the HCA technique (nine subgroups).

In the FKM method, the results are strongly influenced by those variables that have large variances. Therefore, log-transformed and standardized data matrix were used as input data for the FKM analysis. The FKM analysis reduced the original  $118 \times 11$  data matrix to a  $9 \times 11$  matrix of class centers. Table 5 shows the membership matrix for the first 20 samples (the complete table is too large to present).

## Discussion

A direct comparison of the results of the three types of cluster analysis (HCA, KMC, and FKM) is difficult be-

**Fig. 9** Piper diagram of the nine subgroup means for the clusters defined by the three different methods



cause only the HCA technique produces a graphical output. Therefore, we plotted the HCA-, KMC- and FKM-defined means for each subgroup on a Piper diagram. Figure 9 shows that the HCA-, KMC- and FKM-defined means for each group overlap for most of the subgroups, showing the similar results obtained for all three methods. For instance, the FKM analysis placed 97% of the samples within the same three major groups defined by HCA method, whereas 79% of the samples are placed exactly into same subgroups. However, in both the KMC and FKM analysis, we had pre-selected the number of groups (in our case that number was based on the nine defined by the HCA results). The similarity of the results for all three techniques suggests that the pre-selection of the number of groups strongly influences the outcome. This is a serious limitation that means the investigator is required to have performed some type of preliminary analysis when employing the KMC and FKM techniques, which could then bias the results of the statistical analysis.

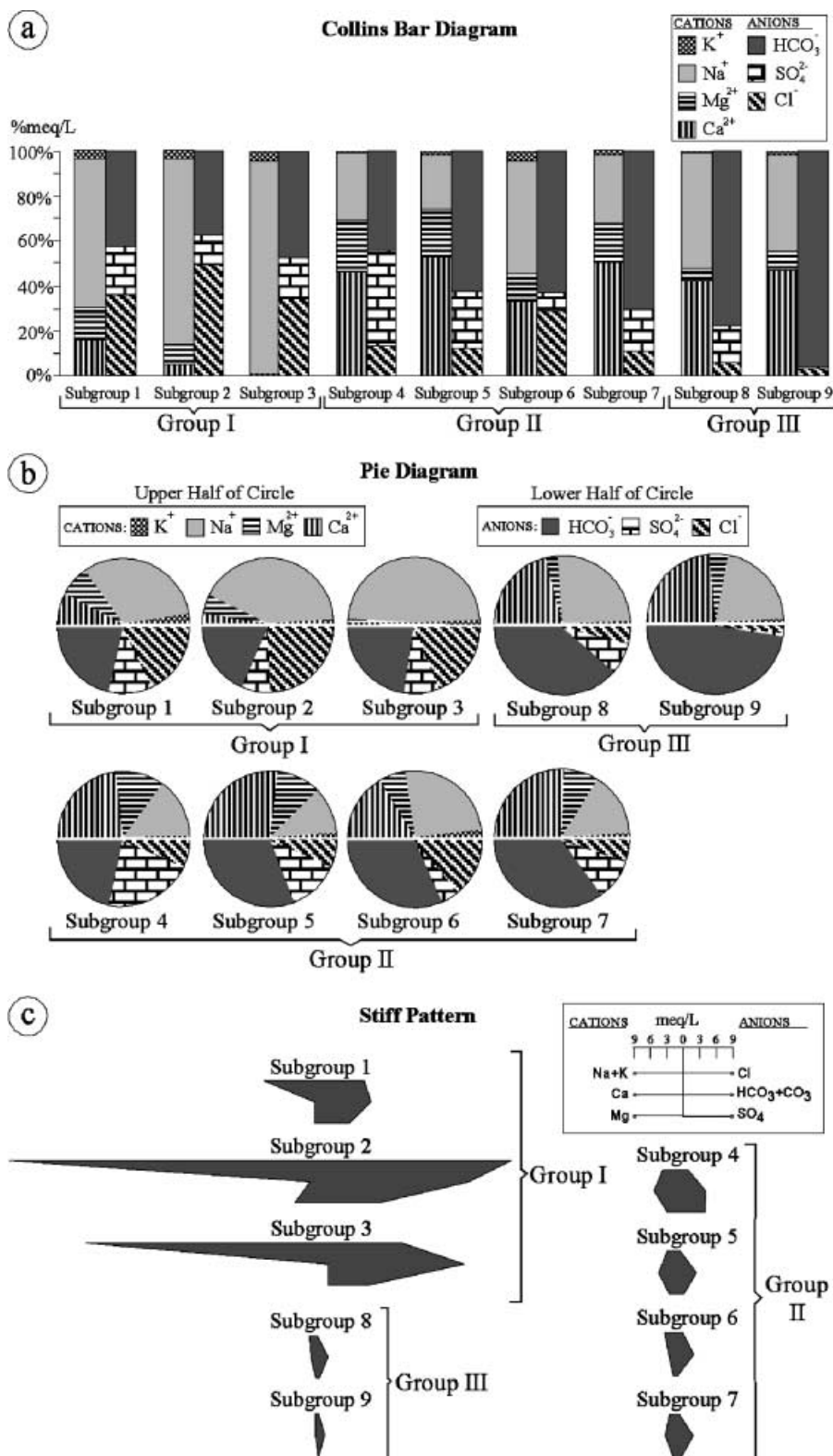
The efficiency and semi-objective nature of the statistical techniques makes these techniques superior to the graphical methods in order to group samples based on water chemistry data. However, the graphical methods

provide valuable information about the chemical nature of the groups. By combining the two techniques we can gain additional information that neither technique by itself can offer.

Figure 10a–c shows the mean values for each of the nine subgroups (defined by HCA) on Collins bar, pie, and Stiff diagrams, respectively. Each graphical technique shows distinct visual differences between the subgroups, while providing information about the chemical composition of each group. In Fig. 11, all the samples are plotted on Schoeller semi-logarithmic diagrams for each subgroup. This plot illustrates the difficulty in using purely graphical means to cluster samples. The patterns of subgroups 1 and 2 are distinctive, but subgroup 3 does not appear related. However, although subgroups 4–7 show a distinct pattern that differs from the other subgroups, it would be difficult to discriminate between samples belonging to subgroups 4–7.

Although previously not utilized in the classification of water samples, we have included an example of icon plots that can be used to represent and visually discern similarities between water samples. The basic idea of icon plots is to represent individual water samples as graphical objects where values of variables are assigned

**Fig. 10** Plots of **a** Collins bar diagram, **b** pie diagram, and **c** Stiff pattern using subgroup means defined by HCA



to specific features or dimensions of the objects. The assignment is such that the overall appearance of the objects changes as a function of the configuration of values. Thus, the objects are given visual “identities” that are unique for configurations of values. One of the most

elaborate type of icon plot is Chernoff faces (Chernoff 1973), which can be used to plot up to 20 parameters for one water sample. Chernoff faces were plotted for the subgroup means from the cluster analysis (Fig. 12). This technique also provides an effective visualization of a



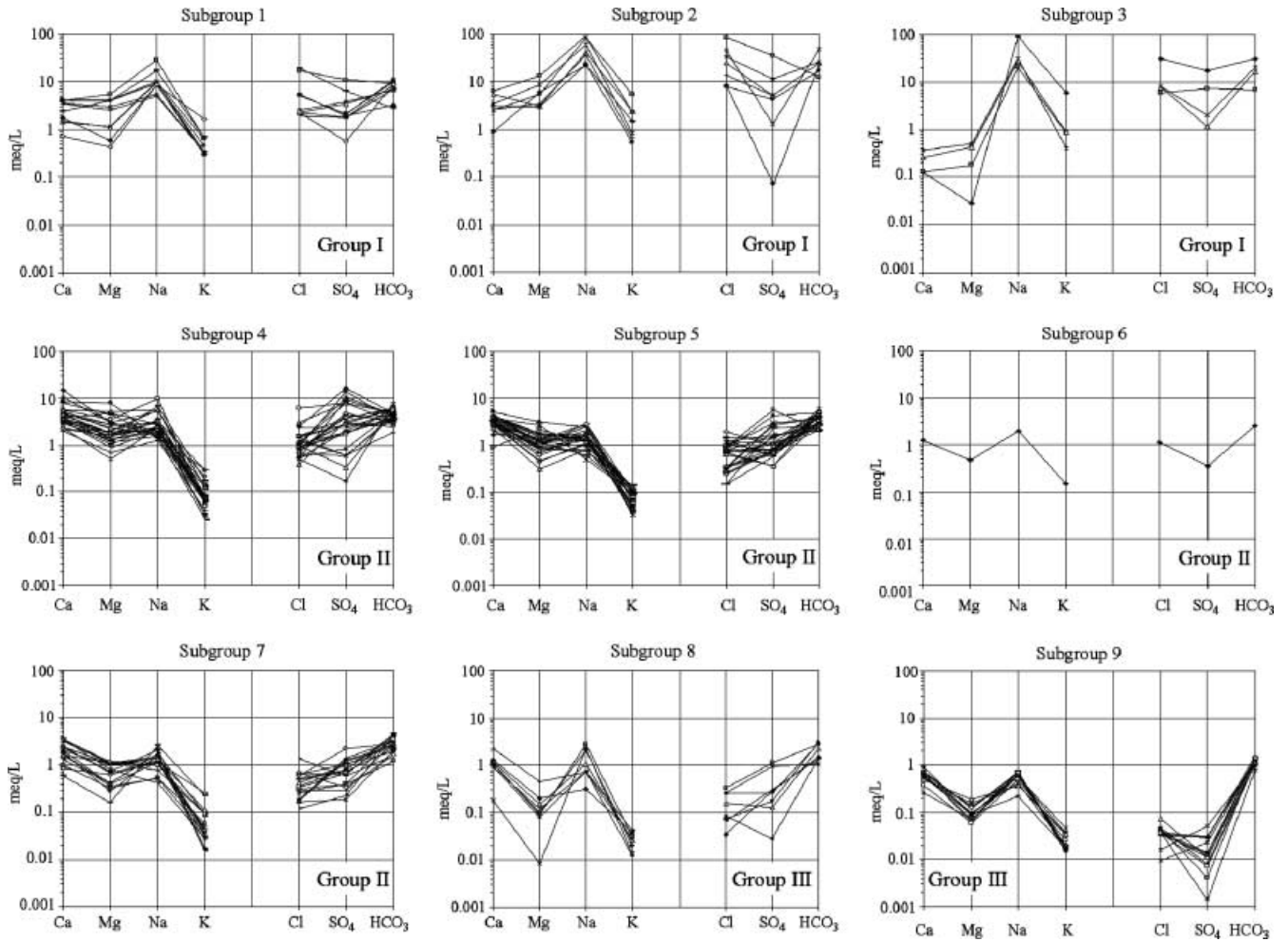
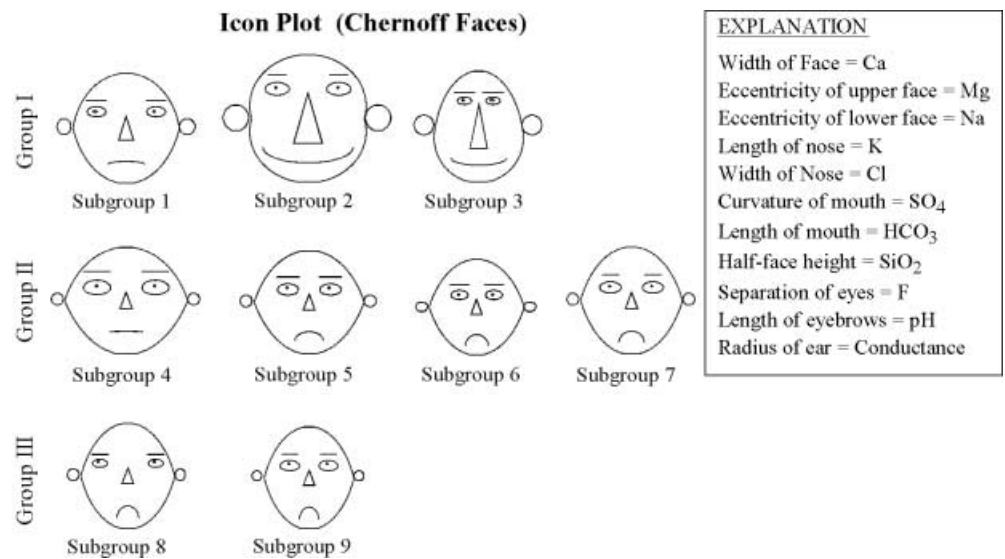


Fig. 11 Plots of all spring water samples by using Schoeller diagram (subgroups and groups defined by HCA)

Fig. 12 Chernoff faces (subgroups and groups defined by HCA)



small number of water samples having different characteristics. The different physical and chemical characteristics of the samples are shown by the changes in facial features. Parameter values are represented in schematic humanlike faces such that the values for each variable are represented by the variations of specific facial features (StatSoft, Inc. 1995). Examining such plots may help to discover specific clusters of both simple relations and interactions between variables.

## Summary and Conclusions

Each technique that has been discussed in this paper has advantages and disadvantages in clustering and displaying water samples using typical chemical and physical parameters. The graphical techniques can provide valuable and rapidly accessible information about the chemical composition of water samples such as the relative proportion of the major ions; however, these techniques have some serious limitations when used alone. All the graphical techniques use only a portion of the available data. Minor constituents ( $0.01\text{--}10\text{ mg L}^{-1}$ ; e.g., boron, fluoride, iron, nitrate, strontium) and trace constituents ( $<0.1\text{ mg L}^{-1}$ ; e.g., aluminum, arsenic, barium, bromide, chromium, lead, lithium) are not used. From a water-quality standpoint, the presence of one of these minor or trace elements may be important because small amounts can pose threats to human health. Some of these minor and trace constituents behave more conservatively in the groundwater, thus, they can be used more efficiently to classify waters (Farnham et al. 2000). Some graphical techniques can display only one sample (or a mean) per diagram (e.g., Collins bar, pie, Stiff), whereas others can display multiple samples (e.g., Piper, Schoeller). Neither type is particularly useful to produce distinct grouping of samples because there is no objective means to discriminate the groups or to test the degree of similarity between samples in a group. Collins bar, pie, and Stiff diagrams are probably the best to help distinguish between small numbers of samples that have distinct chemical differences. For a large number of samples these diagrams are unwieldy. In this study, neither the Piper nor Schoeller diagrams could group all the similar water samples (based on statistical measures) from our data set.

Unlike the graphical classification techniques, multivariate statistical techniques can use any combination of chemical and physical parameters (e.g., temperature) to classify water samples. The HCA technique was judged more efficient than the KMC and FKM techniques because it offers a semi-objective graphical clustering procedure (dendrogram), which does not require pre-determining the final number of clusters. However, none of the statistical techniques offered easily accessible information about the chemical composition of the samples in the clusters (groups). That is, these methods are very efficient at grouping water sample by physical and chemical similarities, but the results are not immediately useful for identifying trends or processes relevant to hydrogeochemical problems.

Combining the two approaches appears to offer a methodology that retains the advantages while minimizing the limitations of either approach. Using the HCA analysis to initially cluster the samples (into groups and subgroups) provides an efficient means to recognize groups of samples that have similar chemical and physical characteristics. The technique also allows discrimination of samples that have extreme values for closer evaluation. These statistical groups have distinct spatial patterns in the study area, providing the spatial discrimination desired when determining hydrochemical facies. The mean for each of the required chemical parameters are then plotted on diagrams, e.g., a Piper diagram, offering easily accessible information on the chemical differences between the groups and potential information about the physical and chemical processes in the watershed. The use of the hierarchical cluster analysis (HCA) in conjunction with a multi-sample graphical technique such as the Piper plot offers a robust methodology with consistent and objective criteria to efficiently classify large numbers of water samples based on common chemical and physical parameters.

**Acknowledgements** The authors would like to thank reviewers, Dr. Patrice deCaritat and Dr. K.H. Johannesson, for their thorough critique. This study was fully sponsored by the Ministry of National Education of Turkey.

## References

- Alther GA (1979) A simplified statistical sequence applied to routine water quality analysis: a case history. *Ground Water* 17:556–561
- Back W (1961) Techniques for mapping of hydrochemical facies. *US Geol Surv Prof Pap* 424-D, pp 380–382
- Back W (1966) Hydrochemical facies and groundwater flow patterns in northern part of Atlantic Coastal Plain. *US Geol Surv Prof Pap* 498-A, Washington, DC
- Back W, Hanshaw BB (1965) Chemical geohydrology. *Adv Hydrosci* 2:49–109
- Barnes I, Kistler RW, Mariner RH, Presser TS (1981) Geochemical evidence on the nature of the basement rocks of the Sierra Nevada, California. *US Geol Surv Water-Supply Pap* 2181, 13 pp
- Berenbrock C (1987) Ground-water data for Indian Wells Valley, Kern, Inyo, and San Bernardino Counties, California, 1977–84. *US Geol Surv Open-File Rep* 86-315, 56 pp
- Berenbrock C, Schroeder RA (1994) Ground-water flow and quality, and geochemical processes, in Indian Wells Valley, Kern, Inyo and San Bernardino Counties, California, 1987–88. *US Geol Surv Water-Resour Invest Rep* 93-4003, 59 pp
- Berry JK (1995) Spatial reasoning for effective GIS. *GIS World Books*, Fort Collins, Colorado
- Bezdek JC (1981) Pattern recognition with fuzzy objective function algorithms. Plenum Press, New York
- Bezdek JC, Ehrlich R, Full W (1984) FCM: the fuzzy c-means clustering algorithm. *Comput Geosci* 10:191–203
- Brown E, Skougstad MW, Fishman MJ (1970) Methods for collection and analysis of water samples for dissolved minerals and gases. *US Geol Surv Tech Water-Resour Invest*, TWI 5-A1, 160 pp
- Buono A, Packard EM (1982) Evaluation of increases in dissolved solids in ground water, Stovepipe Wells Hotel, Death Valley National Monument, California. *US Geol Surv Open-File Rep* 82-513, 23 pp

- Chernoff H (1973) The use of faces to represent points in k-dimensional space graphically. *J Am Stat Assoc* 68:361–368
- Collins WD (1923) Graphic representation of water analyses. *Ind Eng Chem* 15:394
- Davis JC (1986) *Statistics and data analysis in geology*, 2nd edn. Wiley, New York
- deGruijter JJ, McBratney AB (1988) A modified Fuzzy k-means for predictive classification. In: Bock HH (ed) *Classification and related methods of data analysis*. Elsevier Science, Amsterdam
- Dockter RD (1980a) Geophysical, lithologic and water-quality data from test well CU-2, Cuddeback dry lake, San Bernardino County, California. US Geol Surv Open-File Rep 80-1033, 1 sheet
- Dockter RD (1980b) Geophysical, lithologic, and water-quality data from test well CU-1, Cuddeback dry lake, San Bernardino County, California. US Geol Surv Open-File Rep 80-1034, 1 sheet
- Domenico PA, Schwartz FW (1997) *Physical and chemical hydrogeology*, 2nd edn. Wiley, New York
- Drever JI (1997) *The geochemistry of natural waters*, 3rd edn. Prentice-Hall, Upper Saddle River, NJ
- Duffield WA, Smith GI (1978) Pleistocene history of volcanism and the Owens River near Little Lake, California. *J Res US Geol Surv* 6:395–408
- Dutcher LC, Moyle WR Jr (1973) Geologic and hydrologic features of Indian Wells Valley, California. US Geol Surv Water-Supply Pap 2007, 30 pp
- Fang JH, Chen HC (1990) Uncertainties are better handled by Fuzzy arithmetic. *Am Assoc Petrol Geol Bull* 74:1128–1233
- Farnham IM, Stetzenbach KJ, Singh AK, Johannesson KH (2000) Deciphering groundwater flow systems in Oasis Valley, Nevada, using trace element geochemistry, multivariate statistics, and geographical information system. *Math Geol* 32:943–968
- Farnham IM, Stetzenbach KJ, Singh AS, Johannesson KH (2002) Treatment of nondetects in multivariate analysis of groundwater geochemistry data. *Chemometrics Intelligent Lab Sys* 60:265–281
- Fenneman NM (1931) *Physiography of western United States*. McGraw-Hill Book Company, New York
- Feth JH, Roberson CE, Polzer WL (1964) Sources of mineral constituents in water from granitic rocks, Sierra Nevada, California and Nevada. US Geol Surv Water-Supply Pap 1535-I, p 11–170
- Font KR (1995) Geochemical and isotopic evidence for hydrologic processes at Owens Lake, California. MSc Thesis University of Nevada, Reno, 219 pp
- Fournier RO, Thompson JM (1980) The recharge area for the Coso, California, geothermal system deduced from  $\delta D$  and  $\delta^{18}O$  in thermal and non-thermal waters in the region. US Geol Surv Open-File Rep 80-454, 24 pp
- Freeze RA, Cherry JA (1979) *Groundwater*. Prentice Hall, Englewood Cliffs, NJ
- Hem JD (1989) Study and interpretation of the chemical characteristics of natural water, 3rd edn. US Geol Surv Water-Supply Pap 2254, 263 pp
- Hill RA (1940) Geochemical patterns in Coachella Valley. *Trans Am Geophys Union* 21:46–49
- Hill RA (1942) Salts in irrigation waters. *Trans Am Soc Civil Eng* 107:1478–1493
- Hollett KJ, Danskin WR, McCaffrey WF, Walti CL (1991) Geology and water resources of Owens Valley, California. US Geol Surv Water-Supply Pap 2370-B, 77 pp
- Houghton BD (1994) Ground water geochemistry of the Indian Wells Valley. MSc Thesis, California State University, Bakersfield, 85 pp
- Hunt CB, Robinson TW, Bowles WA, Washburn AL (1966) Hydrologic basin, Death Valley, California. US Geol Surv Prof Pap 494-B, 138 pp
- Jaquet JM, Froidevoux R, Verned JP (1975) Comparison of automatic classification methods applied to lake geochemical samples. *Math Geol* 7:237–265
- Johnson JA (1993) Water resources data California water year 1993, vol 5, ground-water data. US Geol Surv Water-Data Rep CA-93-5, pp 97–270
- Johnson JA, Fong-Frydendal LJ, Baker JB (1991) Water resources data California water year 1991, vol 5, ground-water data. US Geol Surv Water-Data Rep CA-91-5, pp 61–62
- Johnson RA, Wichern DW (1992) *Applied multivariate statistical analysis*. Prentice Hall, Englewood Cliffs, NJ
- Judd AG (1980) The use of cluster analysis in the derivation of geotechnical classifications. *Bull Assoc Eng Geol* 17:193–211
- Kovach WL (1990) A multivariate statistical package. Institute of Earth Studies, University College of Wales, Aberystwyth, Wales SY23 3DB, UK
- Lamb CE, Keeter GL, Grillo DA (1986) Water resources data California water year 1986, vol 5, ground-water data for California. US Geol Surv Water-Data Rep CA-86-5, pp 34–185
- Lee CH (1912) Ground water resources of Indian Wells Valley, California. California State Conservation Commission Report, pp 403–429
- Lipinski P, Knochenmus DD (1981) A 10-Year plan to study the aquifer system of Indian Wells Valley, California. US Geol Surv Open-File Rep 81-404, 18 pp
- Lopes TJ (1987) Hydrology and water budget of Owens Lake, California. MSc Thesis, University of Nevada, Reno, 127 pp
- Maltby DE, Downing KT, Keeter GL, Lamb CE (1985) Water resources data California water year 1985, vol 5, ground-water data for California. US Geol Surv Water-Data Rep CA-85-5, pp 42–207
- Maxey GB (1968) Hydrogeology of desert basins. *Ground Water* 6:10–22
- McBratney AB, deGruijter JJ (1992) A continuum approach to soil classification by modified Fuzzy k-means with extra-grades. *J Soil Sci* 43:159–175
- McBratney AB, Moore AW (1985) Application of fuzzy sets to climatic classification. *Agric For Meteorol* 35:165–185
- McHugh JB, Ficklin WH, Miller WR (1981) Analytical results of 78 water samples from Domeland Wilderness and adjacent further planning areas (RARE II), California. US Geol Surv Open-File Rep 81-730, 17 pp
- McNeill D, Freiburger P (1993) Fuzzy logic – the revolutionary computer technology that is changing our world. Simon and Schuster, New York
- Melack JM, Stoddard JL, Ochs CA (1985) Major ion chemistry and sensitivity to acid precipitation of Sierra Nevada lakes. *Water Resour Res* 21:27–32
- Microsoft Corporation (1985) Microsoft Excel 97 spreadsheet software
- Miesch AT (1976) Geochemical survey of Missouri – methods of sampling, laboratory analysis and statistical reduction of data. US Geol Surv Prof Pap 954-A, 39 pp
- Miller GA (1977) Appraisal of the water resources of Death Valley, California–Nevada. US Geol Surv Open-File Rep 77-728, 68 pp
- Minasny B, McBratney AB (1999) FuzME (Fuzzy k-Means with Extragrades) version 1.0. Australian Centre for Precision Agriculture, McMillan Building A05, University of Sydney. <http://www.usyd.edu.au/su/agric/acpa>
- Moyle WR Jr (1963) Data on water wells in Indian Wells Valley area, Inyo, Kern, and San Bernardino Counties, California. California Dept Water Resour Bull 91-9, 243 pp
- Moyle WR Jr (1969) Water wells and springs in Panamint, Searles, and Knob Valleys, San Bernardino and Inyo Counties, California. California Dept Water Resour Bull 91-17, 110 p
- Moyle WR Jr (1971) Water wells in the Harper, Superior, and Cuddeback Valley areas, San Bernardino County, California. California Dept of Water Resour Bull 91-19, 99 pp
- Nordlund U (1996) Formalizing geological knowledge – with an example of modeling stratigraphy using Fuzzy logic. *J Sediment Res* 66:689–698

- Odeh IOA, McBratney AB, Chittleborough DJ (1992) Soil pattern recognition with Fuzzy c-means: application to classification and soil-landform interrelationships. *Soil Sci Soc Am J* 56:505–516
- Ostdick JR (1997) The hydrogeology of southwest Indian Wells Valley, Kern County, California: evidence for extrabasinal, fracture-directed groundwater recharge from the adjacent Sierra Nevada Mountains. MSc Thesis, California State University, Bakersfield, 128 pp
- Piper AM (1944) A graphic procedure in the geochemical interpretation of water-analyses. *Trans Am Geophys Union* 25: 914–923
- Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) Numerical recipes: the art of scientific computing. Cambridge University Press, Cambridge. <http://www.nr.com>
- Robinson BP, Beetem WA (1975) Quality of water in aquifers of the Amargosa Desert and vicinity, Nevada. *US Geol Surv Rep* 474-215, 64 pp
- RockWare, Inc. (1999) RockWorks plotting software. Golden, Colorado
- Rummel RJ (1970) Applied factor analysis. Northwestern University Press, Evanston
- Sanford RF, Pierson CT, Crovelli RA (1993) An objective replacement method for censored geochemical data. *Math Geol* 25:59–80
- SAS Institute, Inc. (1988) SAS/STAT user's guide, release 6.03 edn. SAS Institute Inc. Cary, NC, 1028 pp
- Schoeller H (1955) *Géochimie des eaux souterraines*. *Rev Inst Fr Pétrol* 10:230–244
- Schoeller H (1962) *Les eaux souterraines*. Massio et Cie, Paris, France
- Sibigtroth J (1998) A graphical introduction to fuzzy logic. <http://www.mcu.motsp.com/lit/tutor/fuzzy/fuzzy.html>
- StatSoft, Inc. (1995) STATISTICA for Windows (computer program manual), vol 3. StatSoft, Inc. Tulsa, OK
- StatSoft, Inc. (1997) Electronic statistics textbook. StatSoft, Inc. Tulsa, OK. <http://www.statsoft.com/textbook/stathome.html>
- Stiff HA Jr (1951) The interpretation of chemical water analysis by means of patterns. *J Petrol Tech* 3:15–17
- US Bureau of Reclamation (1993) Indian Wells Valley groundwater project. US Bureau of Reclamation Technical Report, vols I–II
- VanTrump G Jr, Miesch AT (1977) The US Geological Survey RASS-STATPAC system for management and statistical reduction of geochemical data. *Comput Geosci* 3:475–488
- Ward JH (1963) Hierarchical grouping to optimize an objective function. *J Am Stat Assoc* 69:236–244
- Whelan JA, Baskin R, Katzenstein AM (1989) A water geochemistry study of Indian Wells Valley, Inyo and Kern Counties, California, vol 1, geochemistry study and appendix A. Naval Weapons Center Technical Publication 7019, 54 pp
- Williams RE (1982) Statistical identification of hydraulic connections between the surface of a mountain and internal mineralized zones. *Ground Water* 20:466–478
- Wood WW (1981) Guidelines for collection and field analysis of ground-water samples for selected unstable constituents. *Tech Water-Resour Invest US Geol Surv*, 24 pp
- Zadeh LA (1965) Fuzzy sets. *Inform Control* 8:338–353