

Finding the Number of Natural Clusters in Groundwater Data Sets Using the Concept of Equivalence Class

Fernando António Leal Pacheco^(a)

^(a)Secção de Geologia, Universidade de Trás-os-Montes e Alto Douro, 5000 Vila Real, Portugal
Fax: (059) 320480
E-mail: fpacheco@utad.pt

Paper Number 96-112

ABSTRACT

Cluster Analysis has numerous scientific and practical applications. This paper presents a computer program to find an adequate (natural) number of clusters and to isolate anomalous samples in a data set. The program stands on an algorithm that is based on the mathematical concept of equivalence class and uses the framework of the graph theory to identify equivalence classes in multivariate data bases. This type of clustering algorithm is particularly useful when one is dealing with groundwater data sets, because anomalies are frequent in these sets, and because the number of groups that is present is often impossible to estimate; it will depend on the combined effect of many factors, including geology, morphology, climate and pollution. As an example of the utility of this program, a set of groundwater samples is clustered, and the average chemistry of nine identified equivalence classes is related to weathering reactions of plagioclase in a Portuguese granitoid area.

Key words: Cluster Analysis, Groundwater Data Set, Equivalence Class, Graph Theory.

INTRODUCTION

Cluster Analysis is the art of finding groups in data. Some 30 years ago, biologists and social scientists began to look for systematic ways to find groups in their data sets, and because computers were becoming available the resulting algorithms could actually be implemented. Now, clustering methods are applied in many domains, including geosciences, artificial intelligence, pattern recognition, medical research, marketing, and many more.

There are two main types of clustering techniques, namely partitioning and hierarchical methods; in the classification literature the vast majority of algorithms is of either type (Hartigan, 1975; Everitt, 1977; Kaufman and Rousseeuw, 1990).

Conventional partitioning methods construct clusters from a data set. The number of clusters, k , is given by the user, and each object must belong to one group only. In order to obtain the k clusters, classical methods, like H-means (Forgy, 1965) or k-means (MacQueen, 1967), start with an arbitrary partition (samples are randomly distributed by the k groups) and proceed by exchanging samples between clusters until a predefined function is optimized. The results will depend on the initial partition and on the order the samples are exchanged. More recent algorithms, as for example k-medoid of Kaufman and Rousseeuw (1987), select k representative objects in the data set, and the corresponding clusters are then found by assigning each remaining object to the nearest representative object. Fuzzy methods (e.g., fuzzy k-means of Bezdek, 1974) also construct k clusters, but they avoid hard decisions by using the fuzziness principle: instead of deciding that an object belongs to cluster 1, fuzzy methods can, for example, decide that 70% of the object belongs to cluster 1, 20% to cluster 2 and 10% to cluster 3; this

means that the object should probably be assigned to cluster 1, but there is still doubt about whether it should be assigned to cluster 2 or 3.

Conventional and fuzzy methods need *a priori* good estimates of the number of groups present in the data set. This is often impossible when one is dealing with groundwater data sets because the number of groups that is present will depend on: (1) The number of rock types in the area; (2) The degree of chemical weathering of the various rock types; (3) Inputs from sources other than water-rock interactions. All these factors affect the water composition and in combination may generate a high number of groups.

Pacheco and Van der Weijden (1996) developed an algorithm, the Reflexive, Symmetric and Transitive (RST) algorithm, which tackles the problem of finding the number of natural clusters in groundwater data sets. The algorithm uses the definition of equivalence class to split a data base into sets of densely related water samples (the relation being determined by their chemistries), connected by reflexive, symmetric and transitive relations. The present paper, (1) Makes a comprehensive review of the concepts behind the RST algorithm, introducing the framework of the graph theory in the identification of equivalence classes; (2) Provides a full description of the algorithm, as in Pacheco and Van der Weijden (1996) the RST algorithm is only briefly described in one of the Appendices; (3) Discusses the nature of the classes found by the RST algorithm by comparing them with the results obtained with Principal Components Analysis; (4) Presents a computer program (EQCLASS) for finding equivalence classes using the RST algorithm; and (5) Shows an example of results and their application to a practical study of water-rock interaction in a portuguese granitoid area (Fundão, central Portugal).

THE RST ALGORITHM

Concepts Behind the Algorithm: the Graph Theory

As stated by the graph theory (e.g., Christofides, 1975), a graph G is a collection of points or vertices x_1, x_2, \dots, x_n (denoted by the set X) and a collection of lines a_1, a_2, \dots, a_m (denoted by the set A) joining all or some of these points. The graph G is then fully described by the doublet (X, A) . If the lines in A have a direction, which is usually shown by an arrow, they are called arcs, and the resulting graph is called a directed graph. If otherwise the lines have no orientation, they are called links, and the graph is nondirected or symmetric. A typical graph (Figure 1) will have both arcs and links and will be denoted as mixed. An alternative and often preferable way to describe a graph is by specifying a set X of vertices and a correspondence Γ which shows how the vertices are related to each other; for the graph shown in Figure 1, the number of vertices related with x_1 is $\Gamma(x_1) = \{x_2, x_3\}$, with x_2 $\Gamma(x_2) = \emptyset$, with x_3 $\Gamma(x_3) = \{x_2, x_4\}$, and so forth.

A subgraph G_s contained in G ($G_s \subset G$) is made of a subset of vertices $X_s \subset X$ and the subset of lines $A_s \subset A$ joining those vertices; in Figure 1, the set $\{x_1, x_2, x_3\}$ plus the set $\{a_1, a_2, a_3\}$ is a subgraph. A subgraph is said to be complete whenever exists an arc joining each pair of vertices; (e.g., the set $\{x_1, x_2, x_3\}$ and corresponding arcs). If a complete subgraph is symmetric, then it may be referred to as an equivalence class, as the properties of equivalence classes would apply to the vertices in the subgraph, namely symmetry and transitivity (Equations 1a,b):

$$\Gamma(x_i) \supset \{x_j\} \Rightarrow \Gamma(x_j) \supset \{x_i\}, \text{ for all } (x_i, x_j) \quad (1a)$$

$$\Gamma(x_i) \supset \{x_j\} \wedge \Gamma(x_j) \supset \{x_k\} \Rightarrow \Gamma(x_i) \supset \{x_k\}, \text{ for all } (x_i, x_j, x_k) \quad (1b)$$

where \supset means “contains”. The set $\{x_6, x_7, x_8, x_9\}$ and associated links is an equivalence class of the graph shown in Figure 1.

The RST algorithm of Pacheco and Van der Weijden (1996) uses the concept of equivalence class as stated by Equations 1a,b to find the number of natural clusters in multivariate data sets, namely in groundwater data sets as in these cases the expected number of clusters is indeed hardly predictable. To accomplish that the algorithm operates in two main stages:

Stage 1 - *The algorithm builds a graph out of a data base*

As just said, the data base for the RST algorithm may be any multivariate data set representable by a matrix M of n rows denoting the objects and p columns denoting the variables used to describe those objects. Still, throughout this paper the objects will be referred to as groundwater samples and the variables as the concentrations of components dissolved in those samples (e.g., $[Na^+]$, $[K^+]$, etc). The graph in the present context is then a collection of water samples (which are the vertices) plus the similarities between them (the arcs and/or links) calculated on the basis of their chemistries (i.e., on the values of the above mentioned concentrations). Stage 1 proceeds in two main steps: In (Step 1) the similarities between the groundwater samples are calculated by a metric which sets them in the interval $]0,1]$. As seen from the lower limit of this interval, the similarity between water samples will never be zero, meaning that in theory there is always a link joining the samples. By accepting this, the resulting graph would always be an equivalence class, and the RST method would be useless (only one group would be identified). However, in practice, some similarities will be so low (when compared to others) that they can be easily assumed as they were zero. (Step 2) calculates the number of those samples that should be considered

as related with a particular sample given the high similarities between them (the so-called relevant relations); using terminology from the graph theory, Step 2 sets the $\Gamma(x_i)$ correspondence for every x_i in the data base. It ought to be mentioned that this second step imposes a structure to the data set, which may be artificial, but it will be as artificial as the structure imposed to data sets by all known clustering algorithms.

Stage 2 - *Equivalence classes are extracted from the graph built in Stage 1*

Now each water sample is connected to a set of other samples (i.e., a graph is defined). The main purpose of this second stage is to search for complete and nondirected subgraphs by testing them for the properties of symmetry and transitivity (Equations 1a,b). The stage starts by eliminating the non-symmetric relations (i.e., the arcs) and then proceeds by gathering samples for which the transitivity test is valid. Equivalence classes built this way must also assure that the samples belong to just one class.

The search for equivalence classes is dependent on the vertex that is chosen to start the searching as well as on the searching direction. Returning back to Figure 1, where several equivalence classes are represented (e.g., the sets $X_1=\{x_6,x_7,x_8,x_9\}$, $X_2=\{x_4,x_7\}$ and $X_3=\{x_6,x_8,x_9\}$, and corresponding links $A_1=\{a_7,a_8,a_9,a_{10},a_{11},a_{12}\}$, $A_2=\{a_6\}$ and $A_3=\{a_8,a_{10},a_{11}\}$), different results are obtained for different starting vertices. For example, if the searching starts on vertices x_6,x_8 or x_9 , one and the same equivalence class is identified, (X_1,A_1) . If on the other hand the starting point is x_4 , two equivalence classes are found, (X_2,A_2) and (X_3,A_3) . And finally, for a search starting at x_7 , the identified equivalence classes will depend on the direction of the searching: if it goes from x_7 to x_4 , then the results will be the same as those found for a starting point defined at x_4 ; otherwise, the results will be the same as those found for starting points defined at x_6,x_8 or x_9 . The remaining vertices

will split into groups of a single sample as they are not related by symmetric relations (all of them are joined by arcs or are isolated). In the case of a groundwater data set, depending on the values assumed by the original variables (the concentrations), the one-sample clusters may be gathered by the user into a group (or groups) of anomalous samples, if the values in one or more of those variables are abnormal, or into a group of scattered samples otherwise.

Stepwise Description

As mentioned above, the initial raw data for the RST algorithm consist of a matrix M_{np} , where n is the number of groundwater samples in the data base and p the number dissolved components describing the chemistry of those samples. The two consecutive stages and corresponding steps of the algorithm operate as follows:

Stage 1 - Building the graph

Step 1 - Setting up the similarities - The relation between two samples i and j is determined by a measure of similarity S_{ij} defined by:

$$S_{ij} = 1 / (1 + d_{ij}) \quad (2)$$

where,

d_{ij} = euclidian distance between two points

$$= \left[\sum_{k=1}^p w_k (M_{ik} - M_{jk})^2 \right]^{1/2}$$

M_{ik}, M_{jk} = values for the dissolved component k in samples i and j

p = number of components

w_k = weight given to component k

The transformation of the data by the use of the euclidian distance is scale variant, so different results may be obtained when one changes the scale in which the data are expressed.

This seems to reduce the applicability of the method, but the most common partitioning methods all use scale variant measures of similarities or distances to produce the clustering (Kaufman and Rousseeuw, 1990).

A water sample i is closely related to a water sample j if $S_{ij} \rightarrow 1$, and the two samples will probably end up in the same equivalence class. If, however, $S_{ij} \rightarrow 0$, the two samples are practically unrelated and they will end up in different equivalence classes. Among the $n-1$ S_{ij} 's of each sample, there is no general way to distinguish the values corresponding to $S_{ij} \rightarrow 1$ from those that have $S_{ij} \rightarrow 0$. For this reason, one has to decide on a criterion to mark the limits between the two sets of samples. To that end the next step is developed.

Step 2 - Setting up the relevant relations - In this step, the S_{ij} 's of each sample are separated into $S_{ij}=1$, for the related samples, and $S_{ij}=0$ for the unrelated samples. The following terminology was adopted:

raw signal - the $n-1$ S_{ij} 's of each sample sorted in ascending order;

noise - a function that describes the values of the S_{ij} 's for the unrelated samples;

true signal - the S_{ij} 's that will be set to $S_{ij}=1$ (the relevant relations);

filter - the method by which the true signals are separated from the noise.

The filtering method consists of substeps 2.1 to 2.3.

2.1) The $n-1$ relations are ranked in ascending order of their similarity to i and this row forms the raw signal of sample i . The sample j in position m on the raw signal is identified as samp_m ($j=\text{samp}_m$). This first substep is required prior to the application of the filter that will be defined in 2.3 (Equation 4). The samples j that are related to sample i are randomly distributed among the $n-1$ S_{ij} 's which makes it difficult to find these samples in the S_i array. By preceding ranking, the last elements of the array will be the ones to be joined with sample

i. The auxiliary array, *samp*, is used to save the original numbers of the samples (the *j*'s) before sorting the *S_{ij}*'s.

2.2) The first half of the population (lowest relations) is used to define a noise function:

$$\text{noise}_m = \begin{cases} \text{raw signal}_m & \text{if } m \leq (n-1)/2 \\ \text{raw signal}_{n-m} & \text{if } m > (n-1)/2 \end{cases} \quad (3)$$

$$m = 1, 2, 3, \dots, n-1$$

It is assumed that half of the lowest *S_{ij}*'s of each sample may not be transformed into relevant relations; by this method no cluster may have more than (n-1)/2 elements. However, when the number of equivalence classes is expected to be large (groundwater data sets), no sample is likely to have more than (n-1)/2 relevant relations, in which case no relevant relations are lost.

The noise function works as follows:

- The first half of the raw signal is considered to represent only noise (first equation of the noise function);
- The second half of the raw signal is considered to contain some noise; the higher the value of the *S_{ij}* the lower is its noise (second equation).

2.3) Now, a binary square matrix, the relevant matrix *R*, can be defined that represents the relevant relations of the *n* samples. The row *i* of matrix *R* (the true signal of sample *i*) is constructed by setting *R_{ij}*=1 for a relevant relation between *i* and *j* and *R_{ij}*=0 otherwise. For the calculation of the *R_{ij}*'s the following filter is defined:

$$R_{ij} = \text{nearest integer} \left\{ \frac{\text{raw signal}_m - \text{noise}_m}{\text{raw signal}_m} \right\}$$

$$m = 1, 2, 3, \dots, n-1$$

$$j = \text{samp}_m \quad (4)$$

Stage 2 - Identifying the equivalence classes

Step 3 - Setting up groups of water samples with symmetric and transitive relations - The non-zero R_{ij} 's set the correspondences Γ of the water samples; using the appropriate terminology, the samples j which are connected with sample i are given by $\Gamma(i)=\{j, \text{ for all } j\text{'s which have } R_{ij}=1\}$. From the previous steps it is not guaranteed that a relevant relation between sample i and sample j also exists between this sample j and sample i . This means that the symmetry of the relevant relations has to be tested (Equation 1a). In addition, the transitivity property has to be tested if more than two elements are to be joined in the same equivalence class. This is accomplished only if all the relations between the elements of that set of samples are relevant (Equation 1b). A computational implementation of Equations 1a,b is described in the consecutive substeps 3.1 to 3.9.

3.1) The symmetric relations are identified and saved in the elements above the main diagonal of

R :

$$\begin{aligned} R_{ij} &= R_{ij} * R_{ji} \\ i &= 1, 2, 3, \dots, n-1 \\ j &= i+1, \dots, n \end{aligned} \quad (5)$$

3.2) The transitive relations are identified. At the start of the transitivity test all samples have a status $R_{ii} = 1$ (ungrouped). This status changes to $R_{ii} = 0$ when sample i is included in one equivalence class. Only the first element of each class remains with its status unaltered.

3.3) To begin an equivalence class one looks for sample i with $R_{ii} = 1$.

3.4) For this sample i one considers the elements j ($j=i+1, \dots, n$) with $R_{ij} = 1$.

3.5) For this sample j the value of R_{ji} is tested to check whether j has already been included in another class. If $R_{ji}=0$, which means that sample j already belongs to another class, we assign $R_{ij}=0$ to guarantee that sample j will not be grouped with sample i ; otherwise sample j

is grouped with sample i (R_{ij} maintains the value of 1 and R_{ji} is set to 0). Testing the remaining samples k , one continues to preserve the transitivity between samples i, j, k by:

$$\begin{aligned} R_{ik} &= R_{ik} * R_{jk} \\ k &= j+1, \dots, n \end{aligned} \tag{6}$$

- 3.6) In case not all samples j with $R_{ij}=1$ are tested the procedure starts again at step 3.4.
- 3.7) The equivalence class initiated in 3.3 is complete. All samples j of row i with $R_{ij}=1$ belong to it and have $R_{jj}=0$, whereas $R_{ii}=1$.
- 3.8) This procedure must be completed for all samples i which kept $R_{ii}=1$. Subsequently another equivalence class is initiated, starting with step 3.3, until $i=n$.
- 3.9) The elements of each equivalence class are listed: the total number of rows with $R_{ii}=1$ defines the number of classes that have been identified; each class comprises samples j of those rows with $R_{ij}=1$

As already stated, the results described in steps 3.3 to 3.9 may in some cases depend on the starting sample and on the order in which the searching for equivalent relations is carried out; this is the case when the data set contains water samples that may belong to different but similar equivalence classes. In general, the equivalence classes will be defined according to the order in which the samples are inserted in the data base.

The Nature of the RST Classes

In general, the number of clusters present in groundwater data sets is expected to be high because the composition of groundwater is usually affected by many sources (e.g., atmospheric input, pollution) and processes (e.g., weathering, botanical uptake, ion-exchange) acting in combination. In addition, these sources and processes may be of different kinds which also

increases the number of groups (for example, pollution may be caused by agriculture, domestic effluents, etc, weathering may be related to the hydrolysis of silicates, dissolution of carbonates or evaporitic rocks, etc, botanical uptake is dependent on the tree species, etc, etc). But the nature of the clusters (i.e., their form and extent), and especially the separation between them on an euclidian space, is not predictable in advance. Once the chemistry of a cluster is dominated by a specific source or process not affecting the other clusters, then this cluster is supposed to form a disjoint set of samples. But when the contributions of weathering, pollution, etc to the water composition are similar, and therefore the differences between the chemistry of the clusters are narrow, then the distribution of the water samples on a p-dimensional space should reveal a picture of one big and probably elongated cloud, with our clusters forming a sequence of adjacent spots starting at one edge and ending at the other edge of that cloud, reflecting the transition of similar but still different water chemistries.

One possibility that can be used to illustrate the ideas expressed in the previous paragraph is to calculate the first and second principal components of the data set in question (the details of Principal Components Analysis are beyond the scope of this paper and can be found elsewhere, for example in the book of Jackson, 1991), make a cross-plot, and show the distribution of the clusters on this plot. That was done for a data set pertaining to the chemical composition of shallow groundwaters from a zoned granitoid plutonite in central Portugal (the Fundão plutonite) published by Van der Weijden and others (1983). This data set was extensively studied by Pacheco and Van der Weijden (1996) who have defined ten clusters using the RST algorithm (plus one cluster of polluted samples, another of samples with abnormally high concentrations in bicarbonate, and a third of scattered samples, all gathered “by hand” from samples left isolated by the algorithm) and interpreted nine of them using a geochemical mass

balance model and the available geological information. In the present paper, Principal Components Analysis was applied to samples of the Fundão data set, those belonging to the nine interpreted clusters plus the cluster of polluted samples. The variables used in the analysis were the same used by Pacheco and Van der Weijden (1996) to produce the clustering: Na^+ , K^+ , Mg^{2+} , Ca^{2+} , HCO_3^- , Cl^- , SO_4^{2-} and NO_3^- . The samples scores on the first and second principal components (pc_1 and pc_2 , $71.4\% + 10.0\% = 81.4\%$ of the data variation) are shown in Figure 2. Different symbols were used to represent the clusters that Pacheco and Van der Weijden (1996) have associated with (1) A granitic satellite of the plutonite (group 2, filled triangles); (2) The granodiorites forming the body of the plutonite (groups 1,3 and 4, open circles representing the most alkaline facies, and groups 5, 8 and 9, filled squares representing the chalk-alkaline facies); (3) The dike swarm of basic rocks cutting the plutonite (groups 6 and 10, open triangles); and (4) Pollution (bullets). Despite the scatter, its clear from Figure 2 that each spot have a definite position within the factor space. However, no specific sites could be found neither for clusters 1,3 and 4 within the open circles area, nor for clusters 5, 6, 8, 9 and 10 within the filled squares and open triangles areas. The reason for this may have opposite interpretations: (1) The differences between the chemistries of those clusters are artificial; (2) Those differences are real but too narrow to be detected by eigenvector techniques such as Principal Components Analysis. I believe the second interpretation is the right one.

The order in which the spots appear in Figure 2 is essentially conditioned by the first principal component. Using four different groundwater data sets from crystalline rocks (granites and schists), including the Fundão data set, Pacheco (in press) noticed that pc_1 is usually related with the samples electrical conductivities (Ec); in the present case, a Pearson correlation coefficient of 0.99 was found between pc_1 and Ec for 99.95% probability. Looking at the

sequence of spots from the left- to the right-hand side of the diagram shown in Figure 2, its apparent that the samples scores are higher when the associated rocks are more weatherable (amphibolites are surely less resistant to the alteration than granodiorites which in turn are less resistant to weathering than granites). In other words, according to Figure 2, waters become more concentrated (with higher Ec's) when their parent rocks dissolve more quickly. This is additional validation of the RST results.

PROGRAM DESCRIPTION

The program EQCLASS (see Appendix) performs an RST analysis on a multivariate data set, i.e., it finds the number of natural clusters present on that set. The code is written in FORTRAN (MicroSoft™ F32) and includes a central program with 4 subroutines (SIMILARITY, RANK, FILTER and RST) which do:

SIMILARITY - Sets up the similarities between the samples of the data set (Stage 1, Step 1 of the RST algorithm);

RANK (adapted from the INDEXX routine of Press and others, 1989) - Ranks the similarities of each sample in ascending order, a substep (2.1) required prior to the calculation of the relevant relations;

FILTER - Calculates the relevant relations (substeps 2.2 and 2.3);

RST - Identifies the equivalence classes (Stage 2, Step 3).

The program starts asking for the input and output filenames. The input file is an ASCII file which must have the following structure:

- 1) First line - the number of samples (n) and the number of variables (p) separated by space(s);
- 2) Next p lines - the name of each variable (8 characters);
- 3) Next n lines - the number and the variable scores of each sample separated by space(s).

After performing the RST analysis, the identified equivalence classes are written in the output file and the program ends.

APPLICATIONS

The RST algorithm, like any other clustering algorithm, was designed to find groups in data. This particular method is of interest to all those, in any field of research, who cannot by any means estimate *a priori* the number of clusters present in their data sets. As mentioned before, Pacheco and Van der Weijden (1996) applied the RST algorithm to a set of spring and well samples collected in Fundão, a granitoid and agricultural area at central Portugal. The algorithm was used in combination with a novel weathering algorithm (the Silica-Bicarbonate - SiB - algorithm also described in that study) to assess the contributions made by chemical weathering and anthropogenic inputs to the composition of nine groups of shallow groundwaters in that area. After that, Pacheco and others (submitted) applied both the RST and the SiB algorithms to a groundwater data set from a granitoid and forested area in northern Portugal (the Chaves-Vila Pouca de Aguiar region), and using these algorithms they could relate the chemistry of nine equivalence classes with climatic variations within the area, differences in the mineral chemistry

of two granitic facies, large-scale faulting, changes in the forest biomass, ion-exchange reactions and agricultural pollution.

Example

As an example of how the results obtained with the RST algorithm can be used in the field of hydrogeochemistry, the nine equivalence classes that could be interpreted by Pacheco and Van der Weijden (1996) using the SiB algorithm (average chemical compositions depicted in Table 1) will now be re-interpreted by a classical graphical method (Garrels, 1967).

The Fundão plutonite was studied by Portugal Ferreira (1982) and Portugal Ferreira and others (1985). The main lithological types are granodiorites although some granites appear in places. These units are cut by a dike swarm of amphibolites and metadiabases. Plagioclase in the granites is albite, in the granodiorites varies in composition between oligoclase and andesine due to crystal zonation, and in the dikes is andesine. The soils in the Fundão area have a homogeneous mineralogical composition defined by the association quartz + feldspar + biotite + halloysite (Costa and others, 1971).

Plagioclase is the most important weathering reactant among the primary minerals present in the various rocks of the Fundão plutonite. During weathering, plagioclase have altered to halloysite, and the waters collected in the different lithotypes are expected to have chemical compositions related to the weathering of their plagioclase types. Garrels (1967) showed that the bicarbonate to silica mole ratio is a good diagnostic parameter for particular water-mineral interactions. Still, one should bear in mind that this ratio may be strongly upset by sizable input of limestone dust, application of calcium carbonate on agricultural land (not done in the area),

precipitation of silica and/or bicarbonate, and selective uptake of nitrate in exchange with bicarbonate.

Figure 1 is a plot of the mole ratio bicarbonate/silica ($\text{HCO}_3^-/\text{H}_4\text{SiO}_4$) vs. the HCO_3^- in mg/l. It shows that waters collected in the granites (bullets) are indeed related to the weathering of albite, and are clearly separated from waters collected in the granodiorites (open squares) or in the dikes (filled squares). Plagioclase in the granodiorites may vary in composition from oligoclase to andesine and these changes are reflected in the chemistries of clusters 3, 4, (1, 8, 9), 5. Clusters 1, 8, 9 have similar $\text{HCO}_3^-/\text{H}_4\text{SiO}_4$ ratios, but cluster-8 waters are obviously more polluted than cluster-5 waters and much more than cluster-1 waters (cf. Table 1, last row). Cluster-6 and cluster-10 waters fall down between the lines that represent the alteration of andesine into halloysite and Ca-montmorillonite. These waters apparently represent alteration to a product averaging in composition somewhere between those end members, which presumably would represent a mixture of the two phases. Alteration to intermediate smectite type clays (Ca-montmorillonite) may result from stagnant conditions of flow, characteristic for the circulation in faults or dikes, as early pointed out by Tardy and others (1971).

CONCLUSIONS

EQCLASS is a computer program based on the RST algorithm which can be used for finding the number of natural clusters present in a multivariate data set. It is very effective in the clustering of groundwater data sets as in this case the number of groups is large and hardly predictable. An example of the utilization of this program shows that a large number of spring a well samples collected in a granitoid area can be represented by a limited number of clusters, and that the

average chemistry of those clusters can be related to different water-mineral interactions involving the weathering of plagioclase.

ACKNOWLEDGMENTS

This research was partly performed during my stay in Utrecht (Holland). This has occurred in the period October 1990-June 1991 and was supported by an ERASMUS scholarship. I wish to thank C. H. Van der Weijden of the Utrecht University for his supervision during that period and for his advice and constructive comments on the RST algorithm. I am also grateful to the reviewers and to the editor for their constructive remarks and suggestions on an earlier version of this paper.

REFERENCES

- Bezdek, J.C., (1974). Cluster validity with fuzzy sets. *Journal of Cybernetics*, **v. 3**, **p. 58-72**.
- Christofides, N., (1975). *Graph theory, an algorithmic approach*. Academic Press, London, 400p.
- Costa, C.V., Pereira, L.G., Portugal Ferreira, M.R., and Santos Oliveira, J.M., (1971). Distribuição de oligoelementos nas rochas e solos da região do Fundão. *Memórias e Notícias* (Publicações do Museu e Laboratório Mineralógico e Geológico da Universidade de Coimbra), **v. 71**, **p. 1-37**.
- Deer, W.A., Howie, R.A., and Zussman, J., (1962). *Rock-forming minerals*. Longmans, **v. I-IV**.
- Everitt, B., (1977). *Cluster Analysis*. Heinemann Educational Books, London, 122p.
- Forgy, E.W., (1965). Cluster Analysis of multivariate data, efficiency vs. interpretability of classifications (abstract). *Biometrics*, **v. 21**, **p. 768-769**.
- Garrels, R.M., (1967). Genesis of some ground waters from igneous rocks. In: Abelson, P.H. (ed), *Researches in geochemistry*, **v. 2**, **p. 405-420**, Wiley, New York.
- Hartigan, J., (1975). *Clustering Algorithms*. Wiley Interscience, New York, 351p.

Jackson, J. E., (1991). *A user's guide to principal components*. John Wiley & Sons, New York, 569p.

Kaufman, L., and Rousseeuw, P.J., (1987). Clustering by means of medoids. In: Dodge, Y. (ed), *Statistical data analysis based on the L_1 norm*, p.405-416, Elsevier, Amsterdam.

Kaufman, L., and Rousseeuw, P.J., (1990). *Finding groups in data*. John Wiley & Sons, New York, 342p.

MacQueen, J., (1967). Some methods for classification and analysis of multivariate observations. In: Le Cam, L. and Neyman, J. (eds), 5th Berkeley Simp. Math. Statist. Prob., p. 281-297.

Morel, F.M.M., and Hering, J.G., (1993). *Principles and applications of aquatic chemistry*. John Wiley & Sons, New York, 588p.

Pacheco, F.A.L., (in press). Application of Correspondence Analysis in the assessment of groundwater chemistry. *Mathematical Geology*, paper number 96-100.

Pacheco, F.A.L., and Van der Weijden, C. H., (1996). Contributions of water-rock interactions to the composition of groundwater in areas with sizable anthropogenic input. A case study of the waters of the Fundão area, central Portugal. *Water Resources Research*, v.32, no.12, p.3553-3570.

Pacheco, F.A.L., Sousa Oliveira, A., Van der Weijden, A.J., and Van der Weijden, C.H., (submitted). Weathering, biomass production and groundwater chemistry in an area of dominant anthropogenic influence, the Chaves-Vila Pouca de Aguiar region, North of Portugal. *Water, Air and Soil Pollution*.

Portugal Ferreira, M.R., (1982). A magmatic arc in the Iberian Segment of the Hercynian Chain, I- the northwest-southeast lineament between Oporto (Portugal) and Zarza la Major (Spain). *Memórias e Notícias* (Publicações do Museu e Laboratório Mineralógico e Geológico da Universidade de Coimbra), v. 94, p. 31-50.

Portugal Ferreira, M.R., Ivo Alves, E., and Regêncio Macedo, C.A., (1985). A zonalidade interna de um plutonito, estruturas condicionantes e idades de evolução (plutonito do Fundão, Portugal central) *Memórias e Notícias* (Publicações do Museu e Laboratório Mineralógico e Geológico da Universidade de Coimbra), v. 99, p. 167-186.

Press, W.H., Flannery, B.P., Teukolsky, S.A., and Vetterling, W.T., (1989). *Numerical recipes in pascal*. Cambridge University Press, Cambridge, 759p.

Tardy, Y., (1971). Characterization of the principal weathering types by the geochemistry of waters from some European and African crystalline massifs. *Chemical Geology*, v. 7, p. 253-271.

Van der Weijden, C.H., Oosterom, M.G., Bril, J., Walen, C.G., Vriend, S.P., and Zuurdeeg, B.W., (1983). Geochemical controls of transport and deposition of uranium from solution. Case study: Fundão, Portugal. Technical Report, Utrecht University, Institute of Earth Sciences, Department of Geochemistry, EC contract 007.79.3 EXU NL.

APPENDIX

EQCLASS Source Code Listing

```
C      ***** EQCLASS PROGRAM *****
C
C      This is a FORTRAN program for finding a natural partition of a
C      data set. It uses the concept of equivalence class for performing
C      the clustering as presented in the RST algorithm (text).
C
C      The program was compiled by the Microsoft FORTRAN Visual Workbench
C
C      Author: Fernando António Leal Pacheco
C      Address: Secção de Geologia
C              Universidade de Trás-os-Montes e Alto Douro
C              5000 Vila Real
C              Portugal
C              Fax: (059) 320480; E-mail: fpacheco@utad.pt
C
C      ***** Main Program *****
C
C      Declaration/description of variables
C
C      PARAMETER (NINP=10,NOUT=20)          ! Input/Ouput ports
C      CHARACTER*20 INPUT,OUTPUT            ! Input/Output file names
C
C      INTEGER*2    NUMBER    [ALLOCATABLE](:) ! Sample numbers
C      CHARACTER*8  VARNAME   [ALLOCATABLE](:) ! Variable names
C      REAL*4       M         [ALLOCATABLE](:,:) ! Data matrix
C
C      REAL*4       S         [ALLOCATABLE](:) ! Similarities
C      INTEGER*2    SAMP      [ALLOCATABLE](:) ! Auxiliary array
C      INTEGER*1    R         [ALLOCATABLE](:,:) ! Relevant matrix
C
C      INTEGER P
C      CHARACTER*8 VN
C
C      Open the input and output files
C
C      WRITE (*,'(//////////,T5,A,/,T5,A,//////////)')
C      1'Welcome to EQCLASS program','Please follow instruction'
C      WRITE (*,'(T5,A,$)') 'Input Filename? '
C      READ (*,'(A)') INPUT
C      OPEN (NINP,FILE=INPUT)
C      WRITE (*,'(T5,A,$)') 'Output Filename? '
C      READ (*,'(A)') OUTPUT
C      OPEN (NOUT,FILE=OUTPUT)
C
C      Read from the input file the number of samples (N), the number of
C      variables (P), the variable names (VARNAME), the sample numbers
C      (NUMBER) and the variable values in each sample (M). First,
C      allocate memory to variables.
C
C      WRITE(*,'(T5,A)') 'Reading data from the input file...'
C      REWIND NINP
C      READ(NINP,*) N,P
C      ALLOCATE (NUMBER(N),VARNAME(P),M(N,P),S(N),SAMP(N),R(N,N))
```



```

DO J=1,P
  VN=' '
  READ (NINP,'(A)') VN
  ILEN=LEN_TRIM(VN)
  IN=8-ILEN+1
  VARNAME(J)=' '
  VARNAME(J)(IN:8)=VN(1:ILEN)
END DO
DO I=1,N
  READ (NINP,*) NUMBER(I),(M(I,J),J=1,P)
END DO
C
C Steps 1 and 2 of the RST algorithm: calculate the relevant matrix R.
WRITE(*,'(T5,A)') 'Calculating the relevant relations...'
DO I=1,N
  CALL SIMILARITY (I,M,N,P,S)
  CALL RANK (S,N,SAMP)
  CALL FILTER (I,S,SAMP,N,R)
END DO
C
C Step 3 of the RST algorithm: search for equivalence classes.
WRITE(*,'(T5,A)') 'Searching for the equivalence classes...'
CALL RST (R,N)
C
C Write the results on the output file.
C
WRITE(*,'(T5,A)') 'Writing the results...'
WRITE (NOUT,'(T5,A,2X,A)')
1'Equivalence classes identified on the input file ',INPUT
WRITE (NOUT,'(T5,A,T30,I4)') 'Total number of samples: ',N
WRITE (NOUT,'(T5,A,T30,I4,/ )') 'Total number of variables: ',P
NGROUPS=0
DO I=1,N
  IF (R(I,I).EQ.1) NGROUPS=NGROUPS+1
END DO
WRITE (NOUT,'(T5,A,T30,I4,/ )') 'Total number of groups: ',NGROUPS
IGROUP=0
DO I=1,N
  IF (R(I,I).EQ.1) THEN
    IGROUP=IGROUP+1
    NSAMPLES=0
    DO J=I,N
      IF (R(I,J).EQ.1) NSAMPLES=NSAMPLES+1
    END DO
    WRITE (NOUT,'(T5,A,2(I5,A))')
1 'Group ',IGROUP,':',NSAMPLES,' samples'
    WRITE (NOUT,'(T5,A)') 'Sample numbers and group information:'
    WRITE (NOUT,'(T5,16(A8,1X))') 'Number',(VARNAME(J),J=1,P)
    DO J=I,N
      IF (R(I,J).EQ.1) THEN
1 WRITE (NOUT,'(T5,I8,1X,20(F8.2,1X))')
        NUMBER(J),(M(J,K),K=1,P)
      END IF
    END DO
    WRITE (NOUT,'(/)')
  END IF
END DO
DEALLOCATE (NUMBER,VARNAME,M,S,SAMP,R)
STOP ' Normal end of program EQCLASS'
END

```

```

C      ***** SUBROUTINES *****
C
C      SUBROUTINE SIMILARITY (I,M,N,IP,S)
C
C      Calculates the N-1 similarities of sample I. An arbitrary value of
C      2.00 identifies sample I in the S array.
C
C      REAL*4 M(N,IP),S(N)
C
C      DO J=1,N
C        IF (J.EQ.I) THEN
C          S(J)=2.00
C          CYCLE
C        END IF
C        DIJ=0.00
C        DO K=1,IP
C          DIJ=DIJ+(M(I,K)-M(J,K))**2
C        END DO
C        DIJ=SQRT(DIJ)                ! euclidian distance
C        S(J)=1.0/(1.0+DIJ)          ! similarity.
C      END DO
C      RETURN
C      END
C
C      -----
C      SUBROUTINE RANK(S,N,SAMP)
C
C      Indexes the array S of length N, i.e. outputs the array SAMP
C      such that S(SAMP(J)) is in ascending order for J=1,2,...,N.
C
C      REAL*4 S(N)
C      INTEGER*2 SAMP(N)
C
C      DO J=1,N
C        SAMP(J)=J
C      END DO
C      L=N/2+1
C      IR=N
C
C      CONTINUE
C
C      IF (L.GT.1) THEN
C        L=L-1
C        SAMPT=SAMP(L)
C        Q=S(SAMPT)
C      ELSE
C        SAMPT=SAMP(IR)
C        Q=S(SAMPT)
C        SAMP(IR)=SAMP(1)
C        IR=IR-1
C        IF (IR.EQ.1) THEN
C          SAMP(1)=SAMPT
C          RETURN
C        ENDIF
C      ENDIF
C      I=L
C      J=L+L
C
C      IF (J.LE.IR) THEN
C        IF (J.LT.IR) THEN
C          IF (S(SAMP(J)).LT.S(SAMP(J+1))) J=J+1
C        ENDIF

```

```

        IF (Q.LT.S(SAMP(J))) THEN
            SAMP(I)=SAMP(J)
            I=J
            J=J+J
        ELSE
            J=IR+1
        ENDIF
        GO TO 20
    ENDIF
    SAMP(I)=SAMPT
    GO TO 10
END
C -----
SUBROUTINE FILTER (I,S,SAMP,N,R)
C
C   Calculates the line I of the relevant matrix R.
C
    REAL*4 S(N)
    INTEGER*2 SAMP(N)
    INTEGER*1 R(N,N)
C
    REAL*4 RAWS [ALLOCATABLE] (:)
    REAL*4 NOISE [ALLOCATABLE] (:)
C
    ALLOCATE (RAWS(N),NOISE(N))
    DO m=1,N-1
        RAWS(m)=S(SAMP(m))
    END DO
    DO m=1,(N-1)/2
        NOISE(m)=RAWS(m)
    END DO
    DO m=(N-1)/2+1,N-1
        NOISE(m)=RAWS(N-m)
    END DO
    DO m=1,N
        R(m,m)=1
    END DO
    DO m=1,N-1
        J=SAMP(m)
        R(I,J)=NINT((RAWS(m)-NOISE(m))/RAWS(m))
    END DO
    DEALLOCATE (RAWS,NOISE)
    RETURN
END
C -----
SUBROUTINE RST (R,N)
C
C   Identifies sets of samples with symmetric and transitive relations
C   (classes of equivalence).
C
    INTEGER*1 R(N,N)

    DO I=1,N-1
        DO J=I+1,N
            R(I,J)=R(I,J)*R(J,I)
        END DO
    END DO
    DO I=1,N
        IF (R(I,I).EQ.0) CYCLE
        DO J=I+1,N

```

```

      IF (R(I,J).EQ.0) CYCLE
      IF (R(J,J).EQ.0) THEN
        R(I,J)=0
        CYCLE
      ENDIF
      DO K=J+1,N
        R(I,K)=R(I,K)*R(J,K)
      END DO
      R(J,J)=0
    END DO
  END DO
RETURN
END

```

TABLE LEGEND

Table 1: Average chemical composition of the nine equivalence classes that could be interpreted by Pacheco and Van der Weijden (1996) using the SiB algorithm. Original data set of Van der Weijden and others (1983). Square brackets denote concentrations, in $\mu\text{eq/l}$ for the ions and in $\mu\text{mol/l}$ for dissolved silica. Pollution = $[\text{Cl}^-] + [\text{SO}_4^{2-}] + [\text{NO}_3^-]$.

FIGURE CAPTIONS

Figure 1: Representation of a typical graph.

Figure 2: Cross-plot representing the first two principal components and the RST classes of the Fundão data set. The distribution of the water samples by the RST groups and the data for the Principal Components Analysis were compiled from Pacheco and Van der Weijden (1996). Only the data pertaining to the RST classes 1-6, 8-10 and to the anomalies in pollutants were considered. Different symbols were used to represent the RST classes. The pc_1 axis was split into two branches because the pc_1 scores of the pollution group are in general one order of magnitude higher than the scores of the other groups.

Figure 3: Plot of the $[HCO_3^-]/[H_4SiO_4]$ mole ratio *versus* the $[HCO_3^-]$ in mg/l for waters collected in the different lithological types of the Fundão plutonite. The numbers above the symbols represent the RST classes as presented in Table 1. The horizontal lines show the $[HCO_3^-]/[H_4SiO_4]$ ratios expected if the water chemistries were the result of various reactions for the alteration of plagioclase. These expected ratios were calculated taking into account the chemical compositions of the primary and secondary minerals as given in Deer and others (1962) and Morel and Hering (1993).

Table 1

lithotype	granite	granodiorite						amphibolite, metadiabase	
equivalence class	2	4	3	1	8	9	5	10	6
number of samples	20	14	14	22	7	7	10	5	8
[Na ⁺]	301	419	381	422	565	476	656	748	600
[K ⁺]	17	27	24	32	25	29	30	36	41
[Mg ²⁺]	90	150	138	187	379	302	347	504	424
[Ca ²⁺]	80	156	192	292	628	433	616	916	682
[HCO ₃ ⁻]	252	382	433	579	497	606	720	844	1135
[Cl ⁻]	131	192	173	178	304	299	326	553	266
[SO ₄ ²⁻]	56	136	71	79	326	140	367	422	136
[NO ₃ ⁻]	51	140	44	53	472	153	150	385	78
[H ₄ SiO ₄]	472	591	534	610	506	615	676	498	629
[HCO ₃ ⁻]/[H ₄ SiO ₄]	0.53	0.65	0.81	0.95	0.98	0.99	1.07	1.69	1.80
Pollution	238	468	288	310	1102	592	843	1360	480

Figure 1

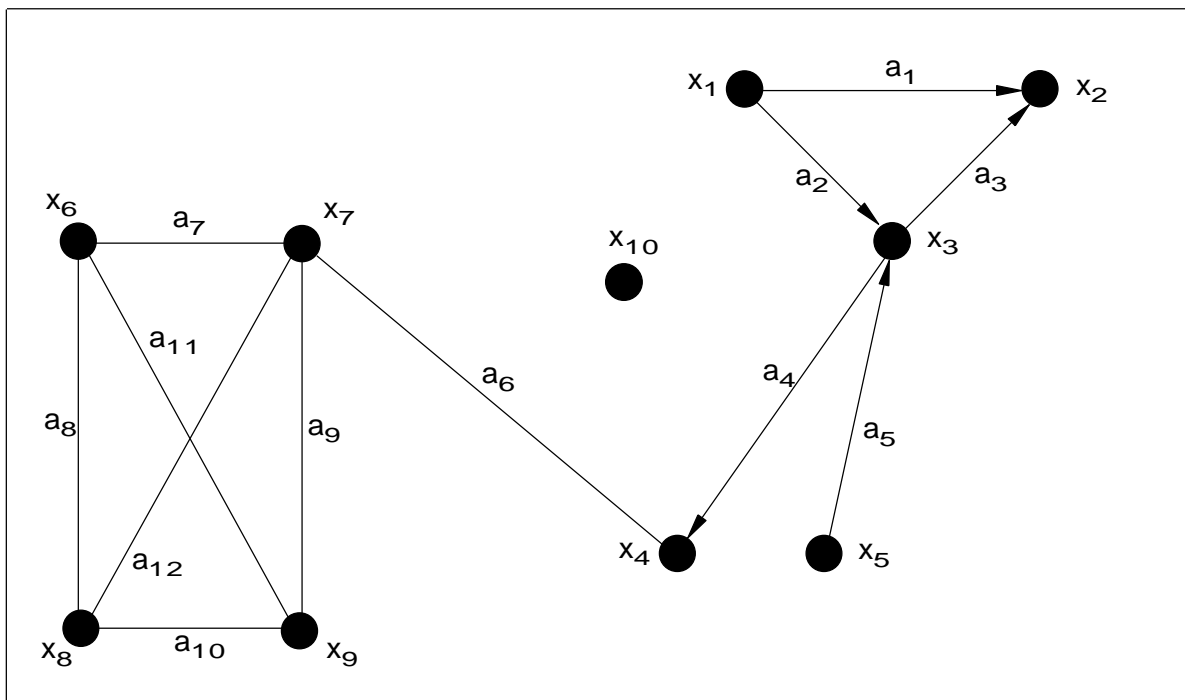


Figure 2

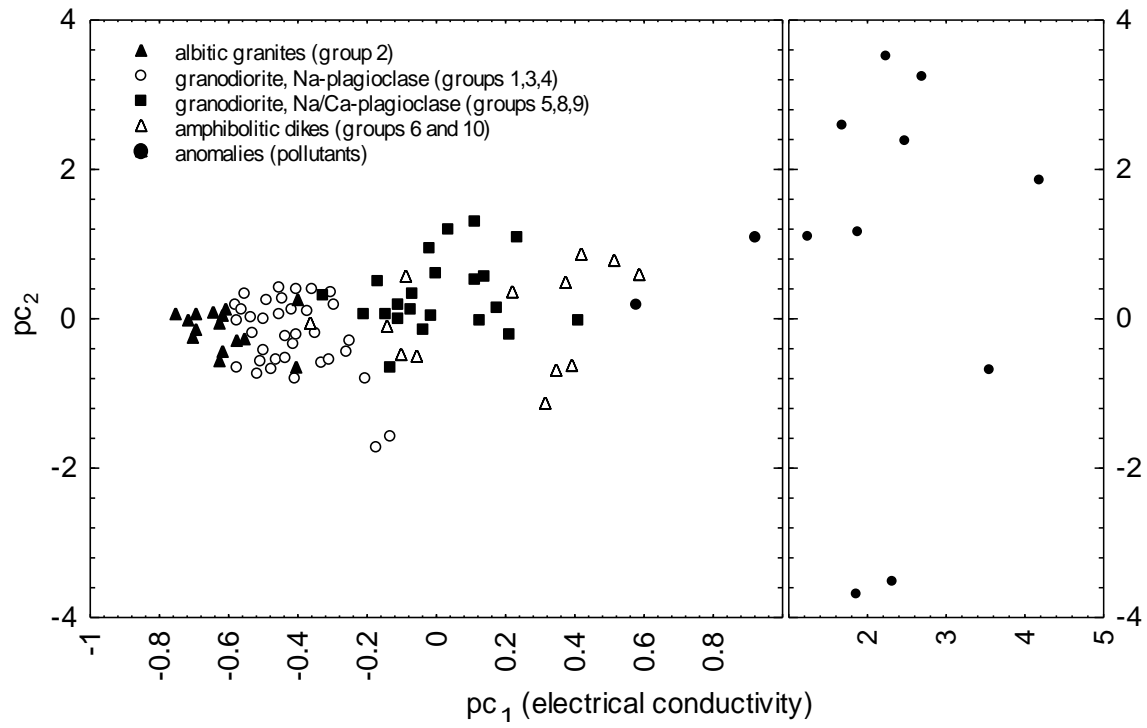


Figure 3

