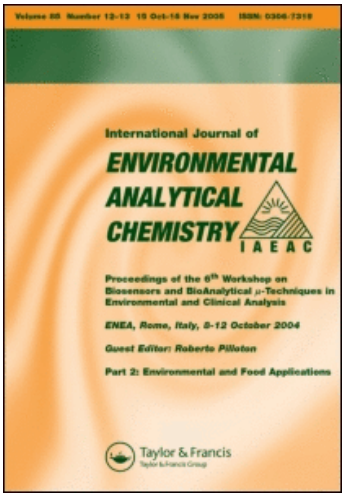

International Journal of Environmental Analytical Chemistry
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713640455

Cluster Analysis as a Tool in the Study of Groundwater Quality

G. Rauret[a]; R. Rubio[a]; F. X. Rius[b]; M. S. Larrechi[b]
[a] Department of Analytical Chemistry, Faculty of Chemistry, University of Barcelona, Barcelona, Spain
[b] Faculty of Chemistry of Tarragona, University of Barcelona, Tarragona, Spain

To cite this Article Rauret, G. , Rubio, R. , Rius, F. X. and Larrechi, M. S.(1988) 'Cluster Analysis as a Tool in the Study of Groundwater Quality', International Journal of Environmental Analytical Chemistry, 32: 3, 255 — 268
To link to this Article: DOI: 10.1080/03067318808079116
URL: http://dx.doi.org/10.1080/03067318808079116

Intern. J. Environ. Anal. Chem., Vol. 32, pp. 255–268
Reprints available directly from the publisher
Photocopying permitted by licence only

Printed in Great Britain

Cluster Analysis as a Tool in the Study of Groundwater Quality

G. RAURET and R. RUBIO

Department of Analytical Chemistry, Faculty of Chemistry, University of Barcelona, 08028-Barcelona, Spain

and

F. X. RIUS and M. S. LARRECHI

Faculty of Chemistry of Tarragona, University of Barcelona, 43005-Tarragona, Spain

(*Received 15 May 1987; in final form 22 August 1987*)

A study of the vulnerability of the river Tenes aquifer has been carried out by means of cluster analysis. On the basis of eleven features measured for forty-seven groundwater samples, pattern recognition techniques allow the visualization of several types of waters. Moreover, the natural differences occurring between manual and pump sampling procedures have arisen. Alkalinity and pH and to a lesser extent sulphates and conductivity have been identified as the features which contribute to the differentiation of the water samples according to the sampling procedure. The withdrawal from the chemometric analysis of these variables leads to the presence of a sole type of water quality in which those samples that show the influence of the surface pollution can be distinguished.

KEY WORDS: Water analysis, groundwater quality, water sampling, cluster analysis.

INTRODUCTION

In the context of a wide research program on the chemical characterization of surface and ground waters of the Besòs basin (Catalonia, Spain),[1,2] a study on the vulnerability of the aquifer of the

Tenes river, tributary of the Besòs river, has been undertaken. The main deterioration of the aquifer might be due to both the influence of the river water, highly polluted because of the industrial effluents discharged on it, and the presence of fertilizers used in the agricultural areas.

The first step in water analysis is always sampling. Very often the way of doing sampling of groundwaters is imposed by the characteristics of the wells. An important factor to take into account is that chemical properties of the waters sampled may depend not only on sampling methods[3,4] but also, to a great extent, on the chemical equilibrium established between the sample and environment. This equilibrium may be of different nature according to the way of doing sampling, either by pumping or manually. In the former it can be assumed that water samples are in equilibrium with the aquifer, while in the latter there exists greater possibilities of occurring interactions between the collected waters and the environment, so the samples may be of somewhat different composition from the aquifer water. Having in mind this fact, two strategies may be followed: The first one would consist of disregarding those wells that require manual sampling, thinking of the distortion of the information obtained from them. The second one would be to look for, and remove, the causes of this distortion. Pursuing the first strategy signifies that several sampling points have to be omitted and consequently some information will be lost. When choosing the second procedure, a multivariate data analysis has to be used.

Very recently, Bartels *et al.*[5] have used pattern recognition methods to classify several types of surface waters on the basis of physicochemical variables. In this paper these methods[6–9] have also been used in order to find changes in the groundwater quality of the area studied. Moreover, the physicochemical parameters which show greater variability according to the sampling procedure used (pumping or manual) have been located and those parameters which can be used in the study of groundwater quality independently of the way of doing sampling have been determined.

EXPERIMENTAL

Forty-seven groundwater samples (Table 1) were analyzed corresponding to fourteen different wells. Thirteen samples had been

Table 1 Groundwater samples, wells and sampling dates (see Figure 2)

Water sample	*Sampling*		*Water sample*	*Sampling*		*Water sample*	*Sampling*	
	Point	*Date*		*Point*	*Date*		*Point*	*Date*
1	AJ[a]	Oct. 83	17	CC[a]	Nov. 83	33	MB[b]	Jan. 84
2	RS[a]	Oct. 83	18	SP[a]	Nov. 83	34	MO[b]	Jan. 84
3	CV[a]	Oct. 83	19	PV[b]	Nov. 83	35	PV[b]	Jan. 84
4	MA[a]	Oct. 83	20	CO[a]	Nov. 83	36	DA[a]	Jan. 84
5	DA[a]	Oct. 83	21	AJ[a]	Dec. 83	37	CV[a]	Jan. 84
6	TI[a]	Oct. 83	22	RS[a]	Dec. 83	38	RS[a]	Jan. 84
7	CC[a]	Oct. 83	23	CV[a]	Dec. 83	39	ES[b]	Feb. 84
8	SP[a]	Oct. 83	24	MA[a]	Dec. 83	40	AL[b]	Feb. 84
9	PV[b]	Oct. 83	25	DA[a]	Dec. 83	41	MB[b]	Feb. 84
10	CO[a]	Oct. 83	26	TI[a]	Dec. 83	42	MO[b]	Feb. 84
11	AJ[a]	Nov. 83	27	CC[a]	Dec. 83	43	PV[b]	Feb. 84
12	RS[a]	Nov. 83	28	SP[a]	Dec. 83	44	TI[a]	Feb. 84
13	CV[a]	Nov. 83	29	PV[b]	Dec. 83	45	DA[a]	Feb. 84
14	MA[a]	Nov. 83	30	CO[a]	Dec. 83	46	CV[a]	Feb. 84
15	DA[a]	Nov. 83	31	ES[b]	Jan. 84	47	RS[a]	Feb. 84
16	TI[a]	Nov. 83	32	AL[b]	Jan. 84			

[a]Pump sampling.
[b]Manual sampling.

manually collected (manual sampling with a Niskin bottle) and thirty-four had been extracted by means of an automatic mechanical system (pump sampling). The sampling period spanned from October 1983 up till February 1984 with at least two sample extractions from each well. The area studied, (Figure 1), corresponding to the aquifer of the river Tenes is shown in Figure 2 where three distinct zones have to be distinguished: zone A is mainly industrial, the amount of water supplied by pumping the wells is considerable and within the studied area the river water suffers a serious progressive deterioration as it runs farther south because of the industrial effluents discharged without any treatment. Zone B is principally an urban area with a large percentage of manual extracted water which give rise to a liquid supply considerably lower than zone A. Zone C is a rural area in which pump sampling is predominant although the amount of pumped water is much smaller than that of zone A.

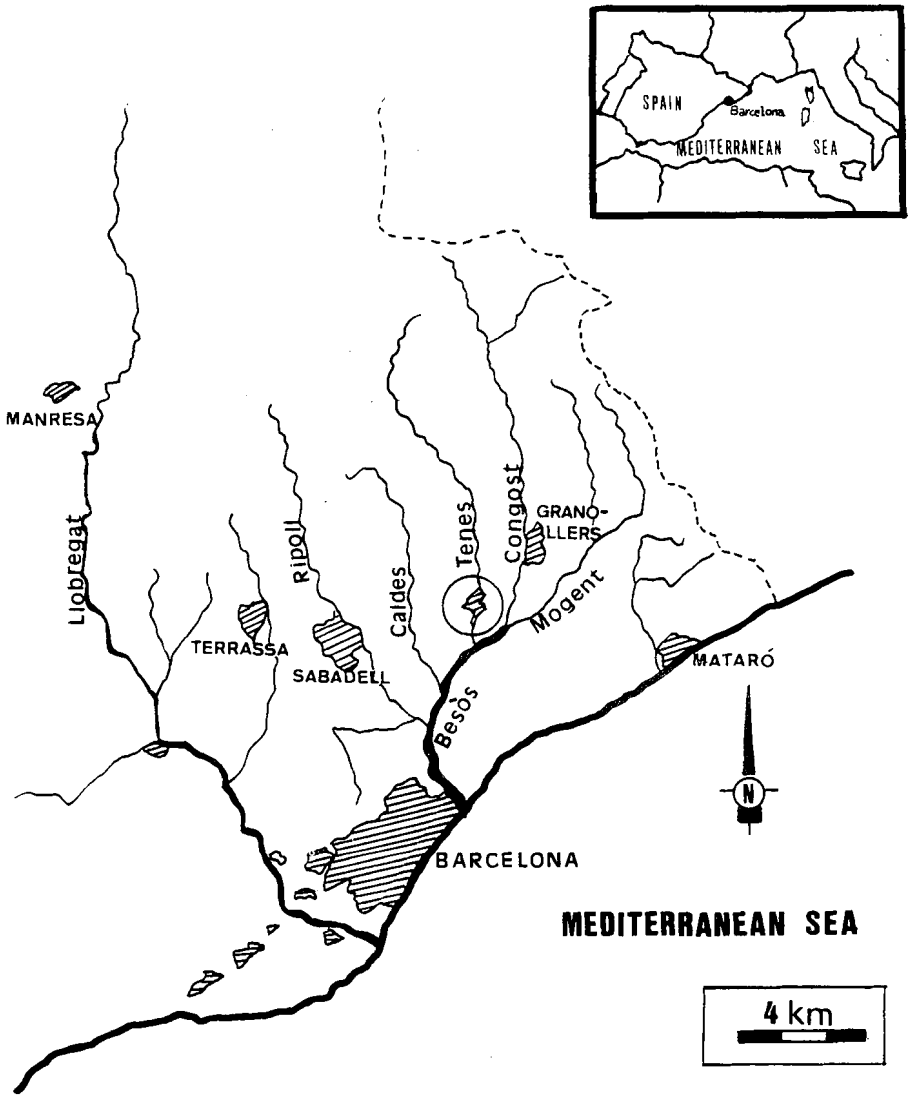

Figure 1 General area of the Besòs basin. The studied area has been encircled.

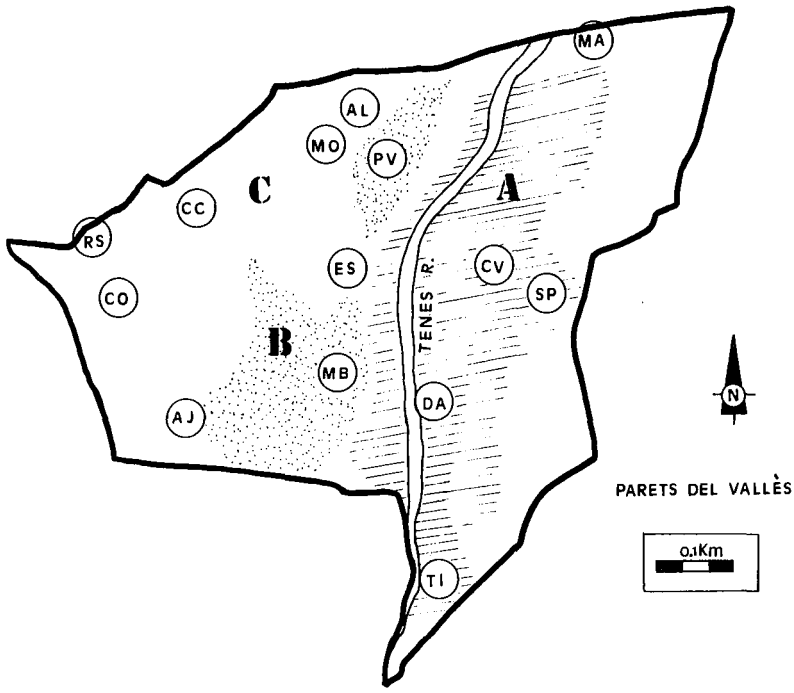

Figure 2 Studied area of the river Tenes (Parets del Vallès) with sampling points and different zones shown.

Eleven parameters corresponding to major, minor and trace components: NO_3^-, Cl^-, SO_4^{2-}, alkalinity, pH, conductivity, Ca^{2+}, Mg^{2+}, Na^+, K^+ and NO_2^- were determined according to standard analytical methods.[10] Other variables were measured: the contribution of temperature was not considered relevant for the chemometric analysis and heavy metals (Cu, Cr, Cd, Hg and Pb) had concentration levels very often below the detection limits of the analytical methods used (a.a.s and i.c.p-a.e.s) so they were not taken into account in the present work. Chemical results have been recently reported elsewhere.[11] Calculations were performed by using well-known pattern recognition packages: ARTHUR 81[12], SPSS-X[13] and CLUSTAN.[14] All programs were run on a IBM 3083 computer of the University of Barcelona.

PATTERN RECOGNITION TECHNIQUES

In order to know the vulnerability of the aquifer, the first objective has been the detection of different groups of water samples in a zone where only one groundwater quality would be expected. Before this unsupervised analysis of the water samples (objects), the usefulness of the variables has been tested. In the same way as the cluster analysis of the objects in the 11-dimensional space defined by the variables reflect similarities in the quality of waters, a cluster analysis of the variables in the 47-dimensional space of the objects exhibit similarities in the behavior of the measured parameters. Two or more parameters with identical location in the hyperspace would imply the same contribution of the variables to the natural disposition of the objects and therefore there would be an excess of information. The number of variables can be reduced with minimum loss of information by selecting one variable representative of each tight cluster. Therefore the initial data matrix has been transposed, the new objects have been autoscaled and the resulting features have been studied by non linear mapping and principal components analysis.

Figure 3 shows a representation (non-linear map) of the 11 parameters in the pseudothree-dimensional space in which the coordinates are non-linear combinations of the coordinates in the 47-dimensional space defined by the water samples. The three-dimensional principal components plot (not shown) gives very similar results. As can be seen in Figure 3, of all variables considered, pH, Ca^{2+}, NO_3^-, SO_4^{2-}, K^+ and NO_2^- display the most singular behaviour while all other features do not cluster in a dense grouping indicating the different contribution of these features to the natural distribution of the water samples. Consequently, no reduction of variables has been considered at this step and the initial data set is arranged in a 47×11 data matrix.

The application of two clustering methods: Ward's hierarchical agglomerative procedure and minimal spanning tree, and a display method: principal components analysis to the autoscaled variables is shown respectively in Figures 4–6. Several common trends can be observed for all assayed methods. First, the natural difference between those samples collected by manual sampling and those obtained by pump sampling. This results have been related more to the differences in water composition than to the sampling procedure.

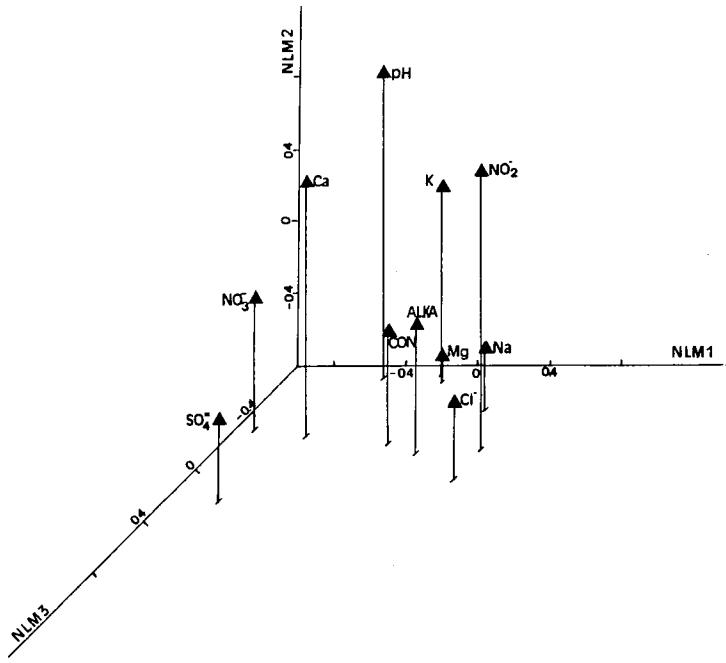

Figure 3 Three dimensional non-linear map of the eleven features in the 47-dimensional space of water samples.

They give evidence of the known fact that a water column in equilibrium with the atmosphere is less equilibrated with the aquifer, and therefore its composition is different, than the samples obtained by sampling from mechanically extracted water. Second, the presence of defined wells whose waters have a significantly different quality from all other groundwater samples. Several causes might be the origin of these results. Samples named no. 6, 16, 26 and 44 belong to only one well (TI) placed in the lower part of zone A where the aquifer is in contact with the most deteriorated surface waters. Samples 7, 17 and 27 have been collected from the CC well placed in the rural zone C. They have an abnormally high concentration of NO_3^- and a SO_4^{2-} level higher than the overall mean value. Samples 9, 19 and 29 from the PV well display a lower salinity than all other waters, their pH is higher as a consequence of a lower buffer

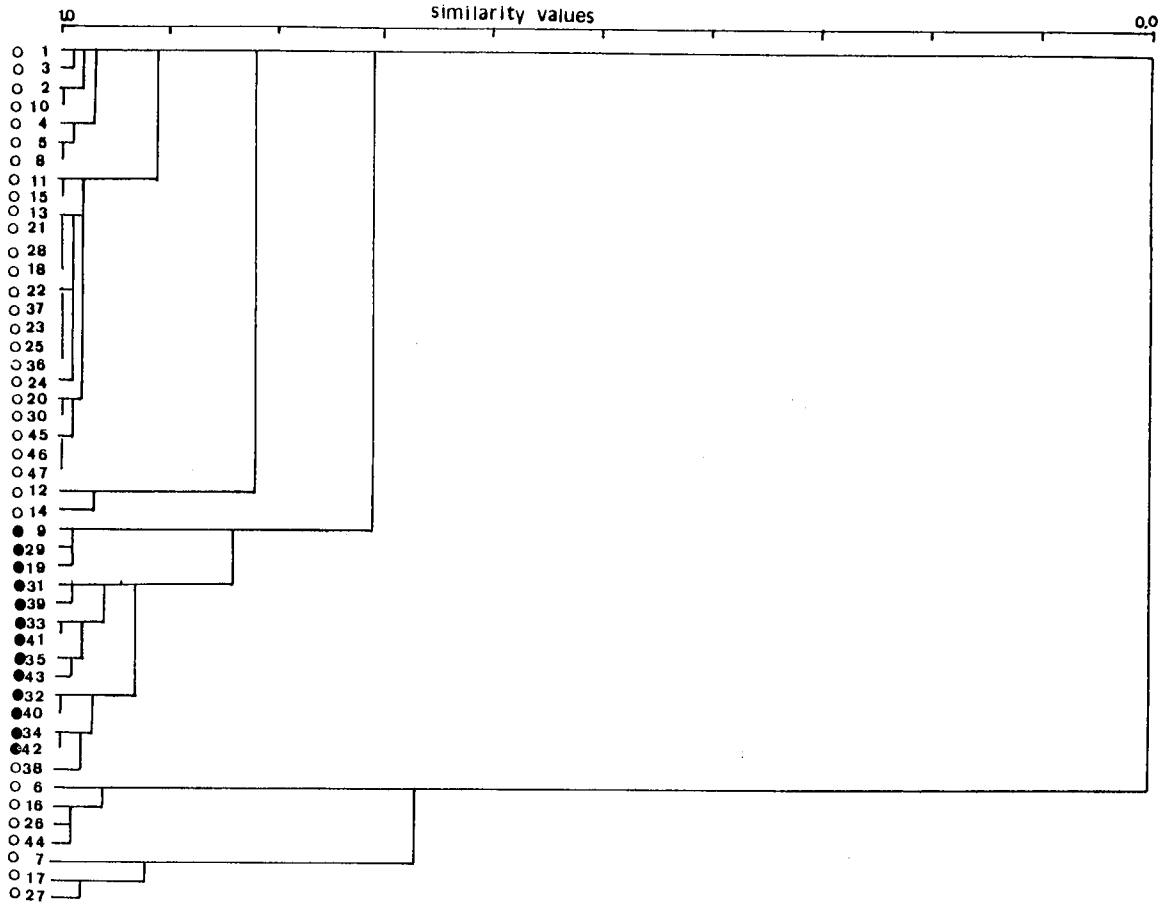

Figure 4 Dendogram of the 47 water samples obtained by using Ward's hierarchical agglomerative method. ○ pump sampling, ● manual sampling.

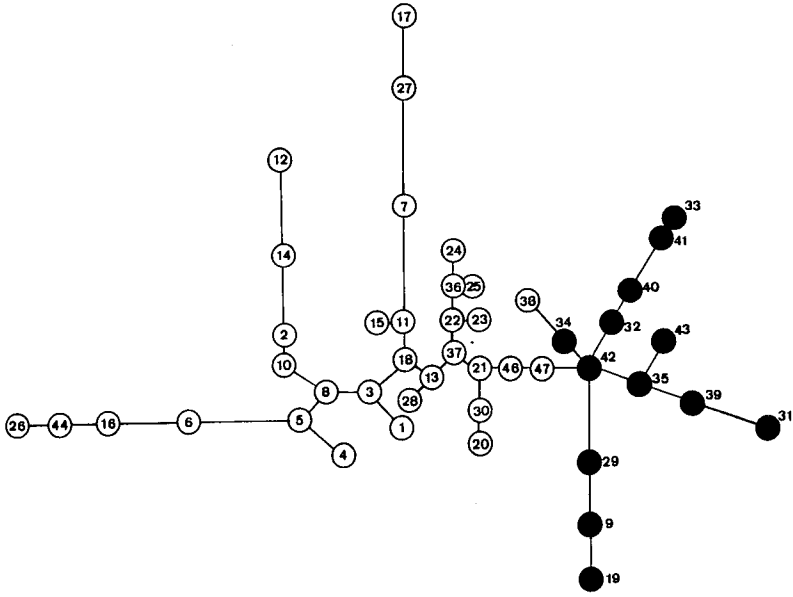

Figure 5 Minimal spanning tree of the water samples in the 11-dimensional space of the initial features. ○ pump sampling, ● manual sampling.

capacity. Samples 12 and 14 although both collected in November 1983, belong to different wells: RS and MA respectively, placed relatively far away from each other. Both samples show a serious punctual contamination of NO_2^-.

The existence of two different groups of waters corresponding to manual or pump sampling has been assessed by supervised analysis. Since the population of the manual sampling group is not large enough, no reasonable assumption about the probability densities of the group can be formulated and therefore, a non-parametric method of analysis has been used, the linear learning machine. The results obtained when all objects have been used to compute the decision surface, indicate the linear separability of the classes considered. In order to know the importance of the variables in the differentiation of the groups, two procedures of variable selection have been used. In Table 2 the individual importance of the features evaluated as the variance weights is listed. Clearly alkalinity and pH

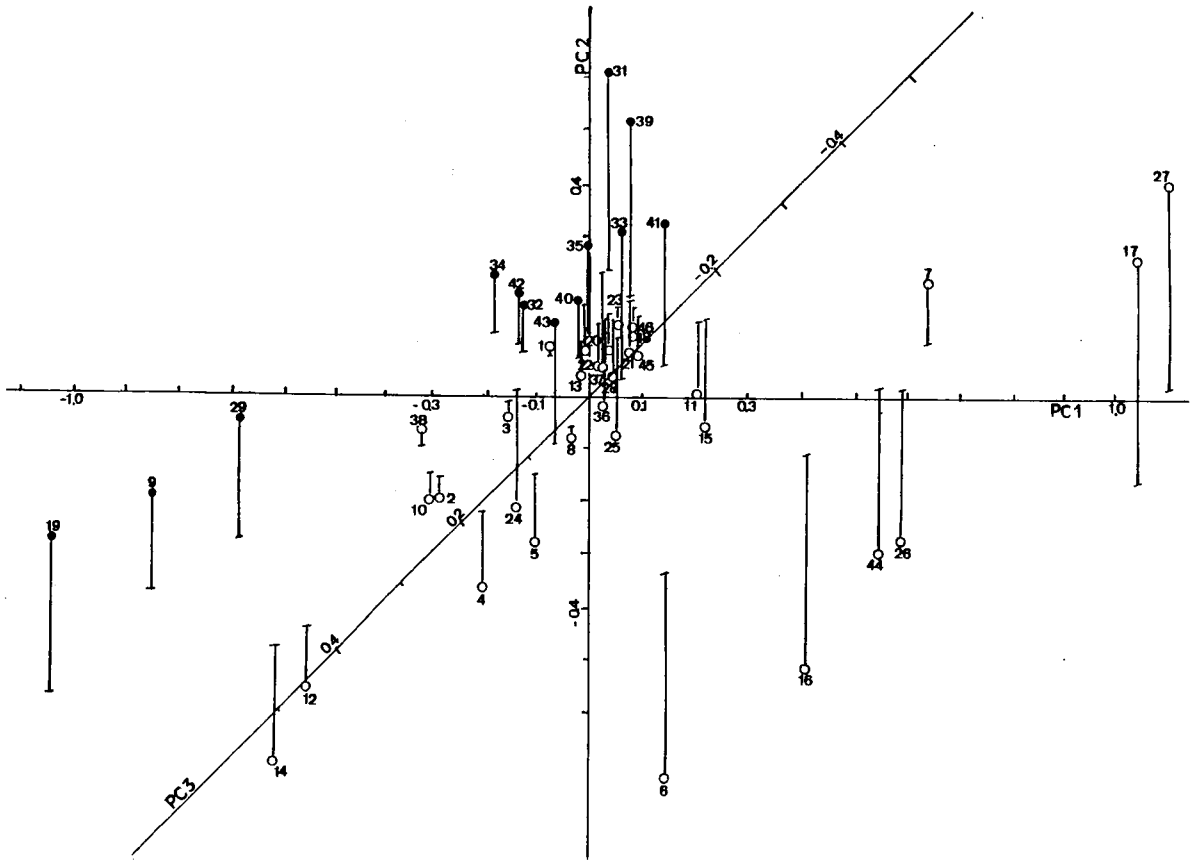

Figure 6 Principal components projection of the water samples distributed in the 11-dimensional space of the initial features. 79.5% of total variance preserved. ○ pump sampling, ● manual sampling.

Table 2 List of features and importance in the differentiation of sampling procedures

Feature	*Symbol*	*Variance weight*	*Feature*	*Symbol*	*Variance weight*
Conductivity[a]	CON	1.58	Sodium[c]	Na^+	1.24
pH	pH	2.34	Potassium[c]	K^+	1.00
Calcium[b]	Ca^{2+}	1.03	Nitrite[c]	NO_2^-	1.03
Magnesium[b]	Mg^{2+}	1.97	Sulphate[c]	SO_4^{2-}	1.25
Chloride[c]	Cl^-	1.57	Alkalinity[b]	AIKA	3.64
Nitrate[c]	NO_3^-	1.23			

[a]In $\mu S\,cm^{-1}$.
[b]In $mg\,l^{-1}$ of $CaCO_3$.
[c]In $mg\,l^{-1}$.

are the most relevant parameters although the absolute weight values indicate that their importance is not considerable. Subroutine SELECT of ARTHUR81 which chooses variance weighted features which have been decorrelated from those previously chosen, has given rise to the variable selection listed in Table 3. These results basically coincide with those reported by the stepwise selection procedure of subprogram DISCRIMINANT of SPSS-X listed also in Table 3. A 100% classification score is obtained when four basic features are selected: alkalinity, sulfates, pH and conductivity. In the context of the global program, the reasons for the pH increment in the waters from the manual sampling wells are being studied at present. This pH increases can give rise to a larger precipitation of $CaCO_3$ which in turn would signify a decrease of both alkalinity and conductivity as encountered in the present study.

When these features are withdrawn from the initial data set, the linear separability of the classes according to the linear learning machine method is no longer accomplished. In order to visualize the distribution of the water samples in the new 7-dimensional space a

Table 3 Feature selection

Procedure	*Selected features*
SELECT[a]	ALKA, SO_4^{2-}, CON, pH, Mg^{2+}, Na^+
DISCRIMINANT[b]	ALKA, SO_4^{2-}, pH, CON

[a]Variance weight criterion, Tolerance level = 1.001.
[b]WILKS, MAXMINF and RAO criterions, FIN = FOUT = 3.0.

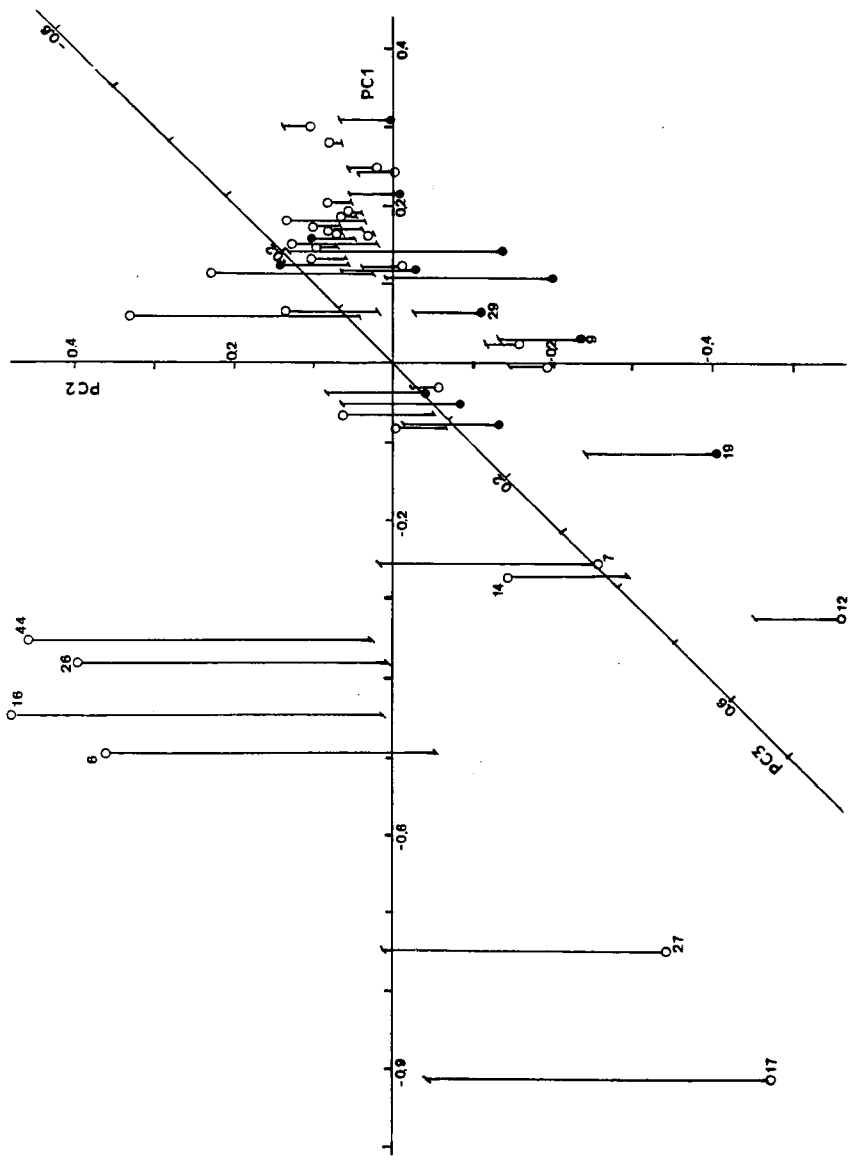

Figure 7 Principal components plot of the water samples in the 7-dimensional space of reduced features. 86.1% of total variance preserved. ○ pump sampling, ● manual sampling.

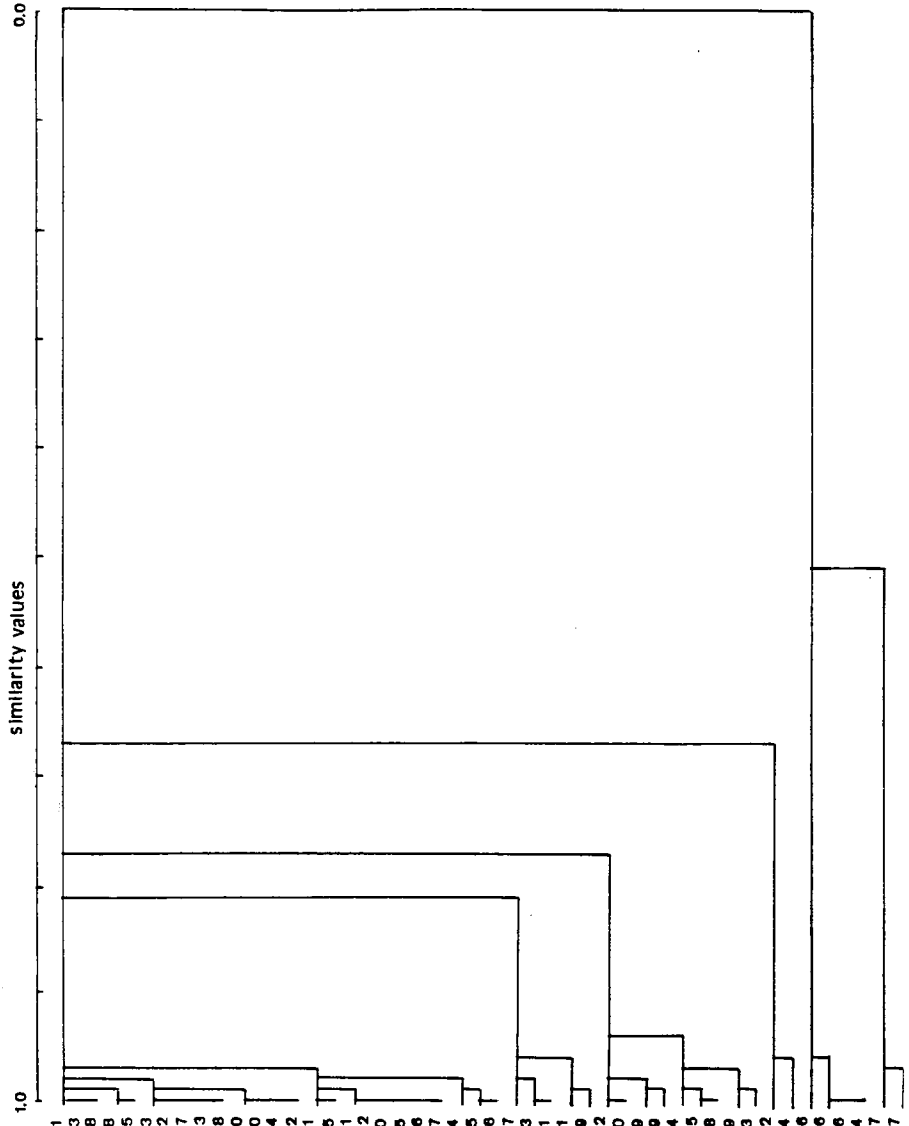

Figure 8 Ward's hierarchical clustering of sample waters in the 7-dimensional space. ○ pump sampling, ● manual sampling.

principal components analysis has been carried out and plotted in Figure 7. Samples are randomly distributed independently of the sampling procedure although a clear distinction is still observed for those samples which previously were characterized as of different quality (samples no. 6, 16, 26, 44, 17, 27, 12, 14). In a similar way, the dendogram obtained when applying Ward's hierarchical agglomerative method (Figure 8) shows a recombination of the initially well defined groups, but clearly, samples no. 17 and 27 are of different quality in a similar way as samples no. 6, 16, 26 and 44. Therefore, the effects of the surface pollution on the quality of the aquifer waters do remain.

The well TI placed farthest south around the river bed in the studied area A displays the most serious deterioration of the groundwater quality with respect to all other wells located in the same area. At present several sampling points farther south in the aquifer are being examined in order to verify the existence in the same aquifer of a water type completely different, and of inferior quality, from that studied in the present work.

References

1. R. Rubio, J. Huguet and G. Rauret, *Water Res.* **18,** 423 (1984).
2. G. Rauret and R. Rubio, *Butll. Soc. Cat. Cièn.* **VI,** 239 (1985).
3. G. F. Lee and R. A. Jones, *Journal WPCF* **55,** 92 (1983).
4. J. Josephson, *Environ. Sci. Technol.* **15,** 993 (1981).
5. J. H. M. Bartels, T. A. H. M. Janse and F. W. Pijpers, *Anal. Chim. Acta* **177,** 35 (1985).
6. K. Varmuza, Pattern Recognition in Chemistry (Springer-Verlag, Berlin, 1980).
7. D. L. Massart, A. Dijkstra and L. Kaufman, *Evaluation and Optimization of Laboratory Methods and Analytical Procedures* (Elsevier, Amsterdam, 1978).
8. D. L. Massart and L. Kaufman, *The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis* (John Wiley, New York 1983).
9. R. C. Jurs and T. R. Isenhour, *Chemical Applications of Pattern Recognition* (Wiley, New York, 1975).
10. APHA-AWWA-WPCF, *Standard Methods for the Examination of Water and Wastewater* (American Public Health Association, New York 1980) 15th ed.
11. G. Rauret, R. Rubio and L. Matia, *Butll. Soc. Cat. Cièn.* **VII,** no. 2, 251 (1986).
12. A. M. Harper, D. L. Duewer, B. R. Kowalski and J. L. Flashing. In: *Chemometrics: Theory and Applications*, (B. R. Kowalski, ed.) (American Chemical Society, Washington D.C., 1977), pp. 14–52.
13. N. H. Nie, C. H. Hull, J. G. Jenkins, K. Steinbrenner and D. H. Bendt, *Statistical Package for the Social Sciences* (McGraw-Hill, New York, 1975) 2nd ed.
14. D. Wishart, *CLUSTAN User Manual* (Edinburgh University, 1978). 3rd ed.