Chemometrics and Intelligent Laboratory Systems, 11 (1991) 229–249
Elsevier Science Publishers B.V., Amsterdam

Correspondence analysis applied to environmental data sets: a study of Chautauqua Lake sediments

Fernando Avila

Department of Mathematics, University of Arizona, Tucson, AZ 85721 (USA) and Departamento de Matematicas, Universidad de Sonora, Hermosillo (Mexico)

Donald E. Myers *

Department of Mathematics, University of Arizona, Tucson, AZ 85721 (USA)

(Received 26 December 1990; accepted 22 May 1991)

Abstract

Avila, F. and Myers, D.E., 1991. Correspondence analysis applied to environmental data sets: a study of Chautauqua Lake sediments. *Chemometrics and Intelligent Laboratory Systems*, 11: 229–249.

Correspondence analysis (CA) is a multivariate technique suitable for data matrices with nonnegative entries. We show its use on environmental data sets and compare our results with those found in previous studies of the same data sets. To perform CA we use an interactive PC program that uses several diagnostics to help find a solution.

INTRODUCTION

Correspondence analysis (CA) is a multivariate technique suitable for data matrices with nonnegative entries, where, traditionally, the rows of the data matrix represent samples/individuals and columns are counting variables. CA is similar to principal components analysis in that it is based on the eigenstructure of a certain matrix and in that it can be used as a dimension reduction technique. However, CA also emphasizes the graphical representation of the results and, because there is a correspondence between row analysis and column analysis, both samples and variables are routinely graphed on the same type of factorial planes. Another feature of CA is that it allows the use of supplementary variables (or samples) which can be projected onto the space generated by the variables (samples).

In this paper we show that CA can be developed in a form suitable for the analysis of data matrices where the variables are measured on a ratio scale. We give an example of its use on environmental data sets where the samples are distributed in space. In our approach, CA is presented as an exploratory technique developed from a purely algebraic point of view. Like other multivariate methods CA can be used as a dimension reduction tool or as a way of finding sample clusters, with the advantage that both tasks can be done simultaneously. It differs from similar techniques, such as biplots and principal coordinates,

0169-7439/91/$03.50

in the emphasis given to the use of graphs and diagnostics that help in the interpretation of the results.

Since we do not assume any underlying probability distribution for the data our interpretation of the results from CA cannot be validated by the classical inferential tests. However, if the results from CA are seen as an algebraic model of the data, then goodness-of-fit measures can be defined and used for the purpose of assessing the degree of approximation to the entries, rows or columns, of the original data matrix. In this paper we provide several such measures, some of which are analogous to the ones used in other methods and some of which are specific to CA. We also show the use of supplementary variables and samples as an aid to the interpretation of the results.

The algorithm for performing CA with goodness-of-fit measures has been implemented in an interactive program that runs on a PC. It differs from other CA programs in that it allows the use of intermediate results for decision purposes instead of asking the user for a predefined (number of factors) model.

PREVIOUS STUDIES

Chautauqua Lake is a narrow, 24-kilometer-long lake in northwestern New York State. It is a warm, shallow lake which is constricted near its midpoint.

The lake was intensively sprayed with sodium arsenite, used as an herbicide, from 1955 to 1963 and there have been several studies [1–4] describing the occurrence and investigating the sources of arsenic found in the lake sediments.

In the first study [1] the possible relationships between the arsenic concentrations and other measured parameters are presented. In this study it is suggested that arsenic has become associated with the lake sediments as a consequence of its use as an herbicide, and that it is being slowly released.

In the second study [2] 98 sediment grab samples were analyzed. The samples were collected during 1972. The sampling pattern consisted of transits made at half-mile intervals with three or four samples taken along each transit. Additional samples were taken in areas of special interest, such as a 23 meter hole, which is the deepest part of the lake.

The samples were analyzed by neutron activation analysis. The concentrations of europium, sodium, manganese, potassium, bromine, arsenic, gallium, lanthanum, hafnium, cesium, terbium, scandium, iron, tantalum and antimony were determined. Analyses were performed to establish particle size distributions and to obtain the percentage of sand, silt and clay present in each sample. In addition, the percentage of organic matter in the sediments was determined.

The highest linear correlation between arsenic and variables characterizing particle size occurs with percent clay ($r = 0.65$). A comparison with arsenic levels in bedrock samples from sites near and around the lake appears to support the assertion that the levels of arsenic found in the lake sediments are not the result of naturally high concentrations in the rock and soils of that area. The conclusion is that the greater concentration of arsenic with increasing amount of clay in the sediments reflects the importance of ion exchange potential of these minerals to attract and retain arsenic. It is also concluded that there has been a loss of the arsenic in the coarse grained sediments near shore where spraying actually occurred.

In the next study [3] the linear correlation coefficients were used to describe the geochemical profile of the sediments. It was also found that for other elements a relationship is indicated between their concentration and the clay fraction of the sediment. The greatest pairwise correlation coefficients ($r \geq 0.8$) were those of the pairs cesium–scandium, cesium–antimony, scandium–antimony and lanthanum–clay. In addition, cesium, scandium, antimony and tantalum had similar contour plots.

Sodium and hafnium had average concentrations which were significantly higher in the lake sediments than the source beds. They are the only two elements that show a positive correlation with sand and an intermediate negative correlation with silt.

The fourth paper [4] describes the results of two methods of multivariate analysis on the set of 79 samples with complete data for all the variables. The variables were the concentrations of the 15 elements analyzed in the previous studies, percent sand, percent silt, percent clay, percent organic matter, water depth above the sample, and several parameters describing the grain size distribution. Thirty-two variables were used in the analysis.

The results of this last paper were described [4] as follows. Factor analysis was performed on the variables. The procedure did not converge to a good fit [4], and the five-factor solution gave the best fit to the data. The common factors accounted for 66.7% of the total system variance, while the remaining variance was contained in the unique factors.

For some of the variables the unique factor had the highest loading. For example, manganese showed a communality of 0.19, implying that manganese concentrations are not linearly related to the common factors [4]. The nature of the unique factor was not explained.

The second multivariate procedure described [4] is hierarchical cluster analysis which was applied to the samples. The resulting dendogram classified samples as belonging to one of four clusters [4]. The average values of the variables and of the factors scores over each of the clusters were then calculated to help in the description of the nature of the sedimental sources and the processes acting on the sediments.

THE DATA

There are two basic data sets that will be used in the present paper. They were constructed using the values found in refs. 2 and 3.

The first data set, referred to henceforth as CHAU98, consists of the concentrations of 15 chemical elements measured in 98 grab samples identified by a code of the form G##. This set was published and analyzed by Hopke et al. [3].

The second data set, referred to henceforth as CHAU88, is a subset of the first data set consisting of the 88 samples for which there were recorded percentages of sand, silt and clay found in the sediment. These values were published by Ruppert et al. [2] and were appended to each of the 88 samples of CHAU98 having them.

The geographical coordinates of the samples are not available. There are also values for organic matter and water depth above the sample for most of the samples, but their inclusion in the data sets did not add to the interpretation of the results from CA, and they will not be used in this paper.

CORRESPONDENCE ANALYSIS

Correspondence analysis (CA) can be explained in a variety of ways [5,6]. Although all the approaches are mathematically equivalent, each one highlights a different aspect of CA.

For example, when data are frequencies and the input matrix is a two-way contingency table, CA can be explained as a generalization of the chi-square test of independence. This approach gives a probabilistic flavor to the technique.

When data are measured on a ratio scale, which is usually the case in the environmental sciences, CA is not adequately explained by a probabilistic interpretation. This is a valid criticism of some previous applications of CA in the earth sciences [7–10], and some attempts [11,12] have been made to present CA which have eschewed the traditional frequentist approach.

Our development tends back to the more traditional approaches, but we think it is more natural in terms of the data to which it will be applied and in terms of software development. It is also the simplest way of defining several 'goodness-of-fit' measures which will be used in what we call 'error analysis'. We present CA as an algebraic approximation method which provides a bilinear model for the data in terms of two sets of 'factors' which can be plotted and interpreted in a similar way to principal components. The similarity of CA to other methods, biplots [13] for example, is well known, but we feel there are some significant differences [6] that help it stand on its own and which can prove advantageous for some applications.

Briefly put, given an input matrix **F**, with nonnegative entries f_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, p$, summing to one (just for convenience), CA gives a representation of $\mathbf{F}$ in terms of two sets of vectors called factors. This representation can be used for algebraic (dimension reduction) or geometrical (interpretation) purposes.

Explicitly, CA of matrix $\mathbf{F}$ means finding a set of scalars $\lambda_1, \lambda_2, \ldots, \lambda_{p-1}$ (assume $p \leq n$), a set of $p-1$ vectors $(\phi_{1,1}, \phi_{2,1}, \ldots, \phi_{p,1}), \ldots, (\phi_{1,p-1}, \phi_{2,p-1}, \ldots, \phi_{p,p-1})$ in $\mathbf{R}^p$, and a set of $p-1$ vectors $(\psi_{1,1}, \psi_{2,1}, \ldots, \psi_{n,1}), \ldots, (\psi_{1,p-1}, \psi_{2,p-1}, \ldots, \psi_{n,p-1})$ in $\mathbf{R}^n$ such that each element of $\mathbf{F}$ can be represented by a bilinear form

$$f_{ij}=f_{i+}f_{+j}\left(1+\sum_{k=1}^{p-1}\sqrt{\lambda_k}\,\phi_{ik}\psi_{jk}\right) \tag{1}$$

where f_{i+} and f_{+j} are defined as $\Sigma_j f_{ij}$ and $\Sigma_i f_{ij}$, respectively, and where each of the truncated bilinear expansions

$$f_{ij}=f_{i+}f_{+j}\left(1+\sum_{k=1}^{K}\sqrt{\lambda_k}\,\phi_{ik}\psi_{jk}\right), \quad K \leq p-1 \tag{2}$$

gives the best approximation to f_{ij} in a well-defined, least square sense.

If $\mathbf{F}$ is a two-way contingency table, then this representation of the entries of $\mathbf{F}$ is a generalization of the statistic used in the chi-square test of independence, comparing f_{ij} to $f_{i+}f_{+j}$. However, even if the data matrix is not a contingency table, formula (1) still makes sense as a matrix factorization.

The following theorem is well known [14] and its proof follows directly from the theorem on the singular value decomposition (also known as the Eckart–Young decomposition) of a matrix. It provides the existence of the factors and defines the goodness-of-fit criterion to be used. It also provides a duality principle unique to CA.

THEOREM. *let* $\mathbf{F}$ *be an* $n \times p$ (*assume* $n \geq p$) *matrix with non-negative entries* f_{ij} *such that* $\Sigma_{ij} f_{ij} = 1$.

Let $\mathbf{D}_p$ *and* $\mathbf{D}_n$ *be diagonal matrices with diagonal entries* $f_{+j} = \Sigma_i f_{ij}$ *and* $f_{i+} = \Sigma_j f_{ij}$, *respectively. These sums are called the weights of the variables and of the samples.*

Let $\mathbf{1}_n$ *be a vector in* $\mathbf{R}^n$ *with all the coordinates equal to one.*

Then there exist $(p-1)$ *triplets* $(\lambda_1, \psi_1, \phi_1), \ldots, (\lambda_{p-1}, \psi_{p-1}, \phi_{p-1})$, *where* $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{p-1} \geq 0$, $\psi_1, \ldots, \psi_{p-1}$ *are vectors in* $\mathbf{R}^n$ *and* $\phi_1, \ldots, \phi_{p-1}$ *are vectors in* $\mathbf{R}^p$, *such that:*

(i) *For every* k, $l = 1, \ldots, p-1$

$$\psi_k^t \mathbf{D}_n \psi_k = \delta_{kl} \tag{3}$$

and

$$\phi_k^t \mathbf{D}_p \phi_k = \delta_{kl} \tag{4}$$

(ii) *For every* $k = 1, \ldots, p-1$

$$\mathbf{F}\mathbf{D}_p^{-1}\mathbf{F}^t\psi_k = \lambda_k \mathbf{D}_n \psi_k \tag{5}$$

and

$$\mathbf{F}'\mathbf{D}_n^{-1}\mathbf{F}\phi_k = \lambda_k \mathbf{D}_p \phi_k \tag{6}$$

(iii) *For every* $k = 1, \ldots, p-1$

$$\psi_k = \frac{\mathbf{D}_n^{-1}\mathbf{F}\phi_k}{\sqrt{\lambda_k}} \tag{7}$$

and

$$\phi_k = \frac{\mathbf{D}_p^{-1}\mathbf{F}^t\psi_k}{\sqrt{\lambda_k}} \tag{8}$$

(iv) *If for every* $K = 1, \ldots, p-1$,

$$\mathbf{F}_K = \mathbf{D}_n\left(\mathbf{1}_n\mathbf{1}_p^t + \sum_{k=1}^{K}\sqrt{\lambda_k}\,\psi_k\phi_k^t\right)\mathbf{D}_p \tag{9}$$

then

$$\mathrm{Tr}\left[(\mathbf{F}-\mathbf{F}_K)\mathbf{D}_p^{-1}(\mathbf{F}-\mathbf{F}_K)'\mathbf{D}_n^{-1}\right] = \|\mathbf{F}-\mathbf{F}_K\|^2 = \sum_{k=K+1}^{p-1}\lambda_k \tag{10}$$

The ψs and the ϕs are called *factors* and sometimes *standardized scores*; if we multiply the factors by the square roots of the λs we get the *coordinates* which can be plotted in the usual Cartesian system.

The factors are unit vectors (in the norms induced by the matrices $\mathbf{D}_n$ and $\mathbf{D}_p$) and can be obtained from an eigenvalue–eigenvector problem, namely eqs. (5) and (6). It can be seen that only one set of factors needs to be obtained, since the other set can be computed from the *transition formulas* (7) and (8), which express a *duality principle* that distinguishes CA from other methods, and which allows the simultaneous analysis of row and column structures.

In CA there are only $p-1$ nontrivial factors. There is one factor, corresponding to an eigenvalue equal to zero or one (depending on the matrix used to extract them), which is discarded since it represents the induced 'correlation' due to the closure of the data. The optimality of the nontrivial factors is expressed in (iv) of the theorem. The case $K=p-1$ gives the *reconstruction formula* (1). When $K<p-1$ factors are kept, we can estimate the error of the approximation when the model $\mathbf{F}_K$ is used, by looking at the matrix norm (called a Frobenius norm) of the difference $\mathbf{F}-\mathbf{F}_K$.

There are several quantities that help in the interpretation of the output:

(i) The *cumulative percentage of variation*

$$\sum_{k=1}^{K} \lambda_k \Big/ \sum_{k=1}^{p-1} \lambda_k \tag{11}$$

which is a global measure of fit when K factors are retained; each λ giving the contribution of a particular factor. Note that this is related to the Frobenius norm of $(\mathbf{F}-\mathbf{F}_K)$ and its name comes from the terminology used in PCA, although here the term variation does not refer to 'variance' in the usual statistical sense.

(ii) For every $k=1,\ldots,p-1$

$$\mathbf{RC}^k(j)=\frac{\lambda_k\phi_{jk}^2}{\sum_{l=1}^{p-1}\lambda_l\phi_{jl}^2}, \quad j=1,\ldots,p \tag{12}$$

and

$$\mathbf{RC}^k(i)=\frac{\lambda_k\psi_{ik}^2}{\sum_{l=1}^{p-1}\lambda_l\psi_{il}^2}, \quad i=1,\ldots,n \tag{13}$$

These are called the *relative contributions*, or squared correlations, of factor k with column j or row i. They provide a measure of the row or column variation explained by a particular factor. We normally quote the relative contributions as percentages and refer to them as correlations by an abuse of the language because they are not correlations in the strict statistical sense. However, they measure the size of the projection onto a particular factor which is directly related to the angle between the factor and the given row or column. Note that the sum of the relative contributions for a particular factor is equal to one.

(iii) For every $k=1,\ldots,p-1$

$$\mathbf{AC}^k(j)=f_{+j}\phi_{jk}^2, \quad j=1,\ldots,p \tag{14}$$

and

$$\mathbf{AC}^k(i)=f_{i+}\psi_{ik}^2, \quad i=1,\ldots,n \tag{15}$$

These are called the *absolute contributions* of column j or row i to factor k. They help in understanding the composition of a particular factor, and are quoted as percentages. By an abuse of the language we say that a particular factor 'is made of' certain variables/samples if they have a high absolute contribution to that factor. Note that the sum of the absolute contributions for a particular row or column is equal to one.

(iv) For every $j=1,\ldots,p$

$$\mathbf{EP}(j)=\sum_{i=1}^{n} f_{i+}\left(\sum_{k=K+1}^{p-1}\sqrt{\lambda_k}\,\psi_{ik}\phi_{jk}\right)^2 \tag{16}$$

and for every $i=1,\ldots,n$

$$\mathbf{EP}(i)=\sum_{j=1}^{p} f_{+j}\left(\sum_{k=K+1}^{p-1}\sqrt{\lambda_k}\,\psi_{ik}\phi_{jk}\right)^2 \tag{17}$$

These are called the *error profiles* for column j or row i when K factors are kept. They are a measure of the error when the matrix $\mathbf{F}$ is 'reconstructed' by the matrix $\mathbf{F}_K$. Note the identity

$$\mathrm{Tr}\left[(\mathbf{F}-\mathbf{F}_K)\mathbf{D}_p(\mathbf{F}-\mathbf{F}_K)'\mathbf{D}_n\right]$$
$$=\sum_{i=1}^{n}\sum_{j=1}^{p} f_{i+}f_{+j}\left(\sum_{k=K+1}^{p-1}\sqrt{\lambda_k}\,\psi_{ik}\phi_{jk}\right)^2 \tag{18}$$

(v) *Supplementary elements*, which can be either rows or columns. A given supplementary row

$(f_{s1}, f_{s2}, \ldots, f_{sp})$ can be projected onto the kth principal axis, with its projection (coordinate) being equal to

$$\hat{\psi}_{sk} = \sum_{j=1}^{p} \frac{f_{sj}}{f_{s+}} \phi_{kj} \tag{19}$$

Analogously, for a supplementary column $(f_{1s}, f_{2s}, \ldots, f_{ns})'$ its projection onto the kth principal axis is

$$\hat{\phi}_{sk} = \sum_{i=1}^{n} \frac{f_{is}}{f_{+s}} \psi_{ik} \tag{20}$$

RESULTS FROM CORRESPONDENCE ANALYSIS

Whenever an exploratory method, such as CA, is used, there is the possibility that the results are an artifact of the method. To determine the stability of the results it is necessary to apply the method on several data sets or variations of the same data set. On the subject of stability, Critchley [15] suggests two possible courses of action: either 'robustify' the method in some way, or perform standard analyses in parallel with appropriate diagnostic statistics and graphical displays. We follow the second course of action suggested, using the criteria developed in the preview section as guiding instruments for successive analyses. Our main objective is to show that CA produces the same 'good' results as the combined application of factor analysis and cluster analysis, and that it helps explain the 'bad' results of these other methods. We will be using CA both as a dimension reduction method and as a clustering technique. We think that a very nice feature of CA is that all the results can be put in graphical form, which makes interpretation easier than by just looking at a set of numerical tables. We should, however, be aware of possible pitfalls when interpreting the graphs.

TABLE 1

Eigenvalues and variation explained by factors

CHAU98		CHAU93	
Eigenvalues	% Variation	Eigenvalues	% Variation
0.14415	63.4	0.07370	76.4
0.04883	21.5	0.01516	15.7
0.03398	14.9	0.00715	7.4
0.00017	0.1	0.00015	0.2
0.00015	0.1	0.00013	0.1
0.00010	0.0	0.00009	0.1

TABLE 2

Absolute and relative contributions (AC and RC, respectively) of the variables for the first three factors

Weight	Factor 1 (λ = 0.14)		Factor 2 (λ = 0.05)		Factor 3 (λ = 0.03)	
	AC(1)	RC(1)	AC(2)	RC(2)	AC(3)	RC(3)
Eu 0.00001	0.0	6.7	0.0	2.3	0.0	16.5
Na 0.07200	78.2	95.2	5.0	2.0	9.6	2.8
Mn 0.01869	1.4	4.4	79.1	82.9	17.5	12.8
K 0.13130	7.3	26.3	15.6	19.2	63.6	54.4
Br 0.00011	0.0	0.0	0.0	5.0	0.0	12.6
As 0.00024	0.0	2.7	0.0	7.9	0.0	7.7
Ga 0.00018	0.0	3.6	0.0	0.9	0.1	12.2
La 0.00026	0.0	8.8	0.0	4.5	0.1	41.5
Hf 0.00016	0.0	29.6	0.0	0.9	0.0	0.1
Cs 0.00006	0.0	2.4	0.0	15.4	0.0	8.3
Tb 0.00001	0.0	3.7	0.0	5.7	0.0	5.3
Sc 0.00039	0.0	0.0	0.0	24.7	0.0	2.8
Fe 0.77654	13.1	85.7	0.2	0.4	9.0	13.9
Ta 0.00002	0.0	8.5	0.0	4.0	0.0	6.1
Sb 0.00003	0.0	2.9	0.0	9.2	0.0	0.0

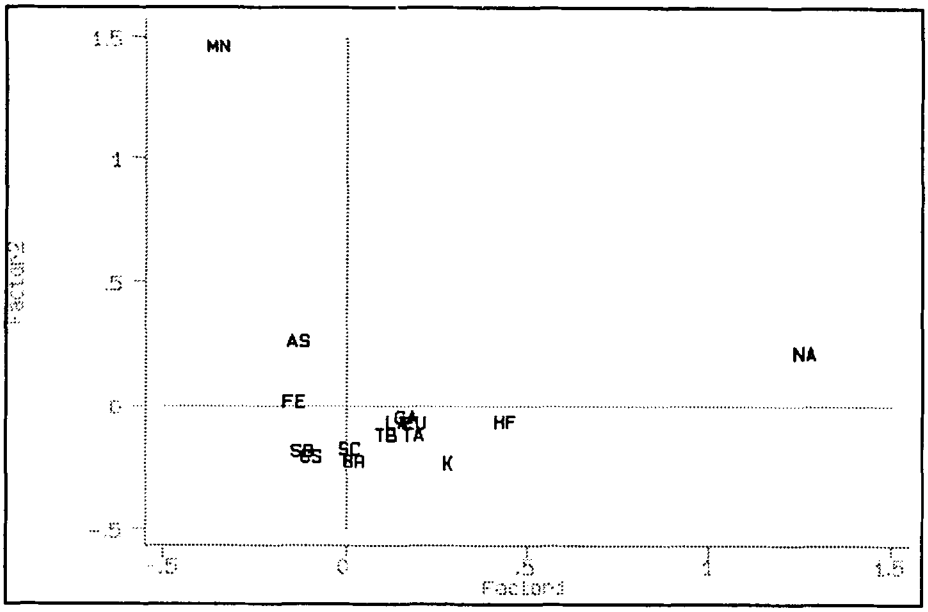

Fig. 1. Factor 2 vs. Factor 1 with 98 samples. The variables.

CA for CHAU98

CA was performed on the set CHAU98 using the 15 chemical variables. The values that were reported as below a certain threshold were given values equal to half the threshold value, which is a common practice. CA was also performed after setting those values equal to zero and latter after

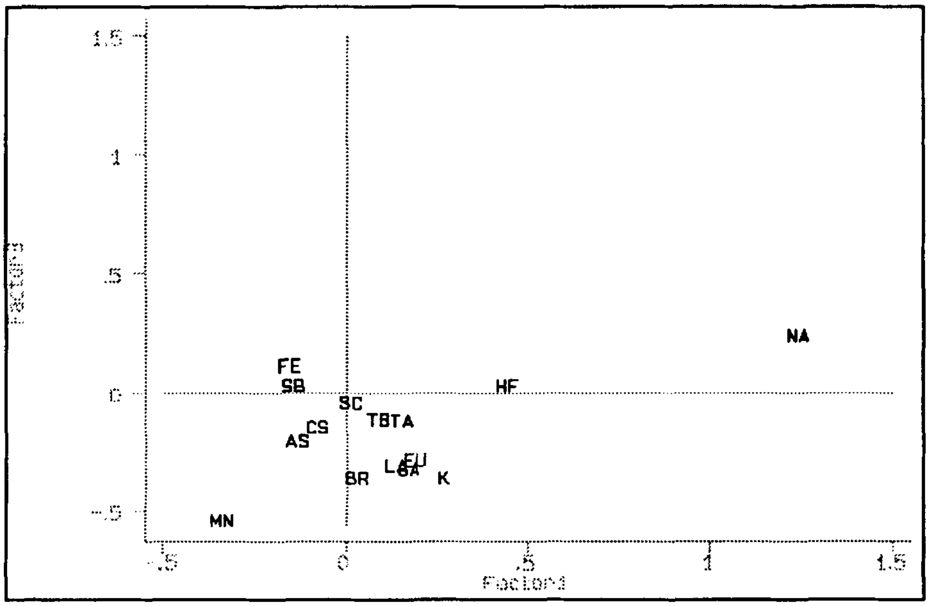

Fig. 2. Factor 3 vs. Factor 1 with 98 samples. The variables.

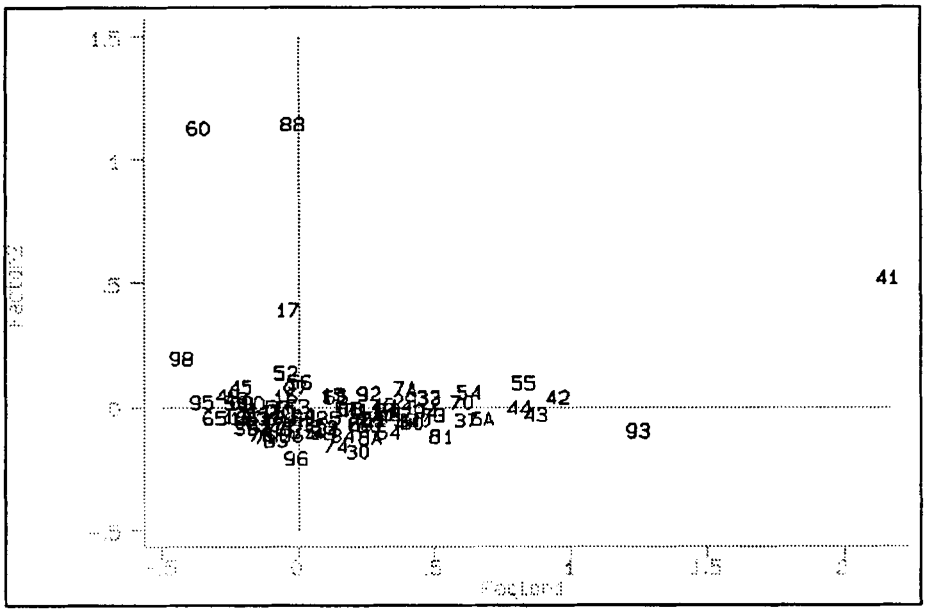

Fig. 3. Factor 2 vs. Factor 1 with 98 samples. The samples.

setting them equal to the threshold values. Essentially the same results were obtained in the three cases. This was due to the fact that the variables having large values, several orders of magnitude larger than the others, had the most influence on the results.

Three factors accounted for 99.75% of the total variation (Table 1), with the first one accounting for 63.35%. We analyze first the results for the variables starting with the composition of the factors (Table 2).

The first factor is composed of sodium (78%), iron (13%) and potassium (7%). Sodium and iron are very well correlated with this factor.

The second factor, accounting for 21.5% of the total variation, is composed mainly of manganese (83%) and this element is the only one that correlates highly with this factor. Potassium also contributes (16%) to the formation of this factor.

The third factor, which accounts for 14.9% of the total variation, is composed of potassium (64%), manganese (18%), sodium (9%) and iron (9%). These are the only elements that contribute to this factor, but only potassium and lanthanum have relative contributions above 40%.

We identify the elements that contribute most to the formation of the factors. For this data set these have the highest weights. We then analyze the graphical displays generated by CA (Table 3, Figs. 1 and 2).

TABLE 3

Coordinates of the variables on the first three factors

	Factor 1	Factor 2	Factor 3
Eu	0.20246	−0.11850	−0.31686
Na	1.25117	0.18340	0.21302
Mn	−0.32954	1.43785	−0.56412
K	0.28229	−0.24114	−0.40576
Br	0.01882	−0.24219	−0.38582
As	−0.13760	0.23743	−0.23425
Ga	0.17865	−0.08911	−0.32756
La	0.15670	−0.11175	−0.34103
Hf	0.44977	−0.07917	−0.01951
Cs	−0.08458	−0.21426	−0.15728
Tb	0.11022	−0.13698	−0.13182
Sc	0.00625	−0.20658	−0.06955
Fe	−0.15596	−0.01067	0.06281
Ta	0.17677	−0.12102	−0.14996
Sb	−0.12084	−0.21629	0.01176

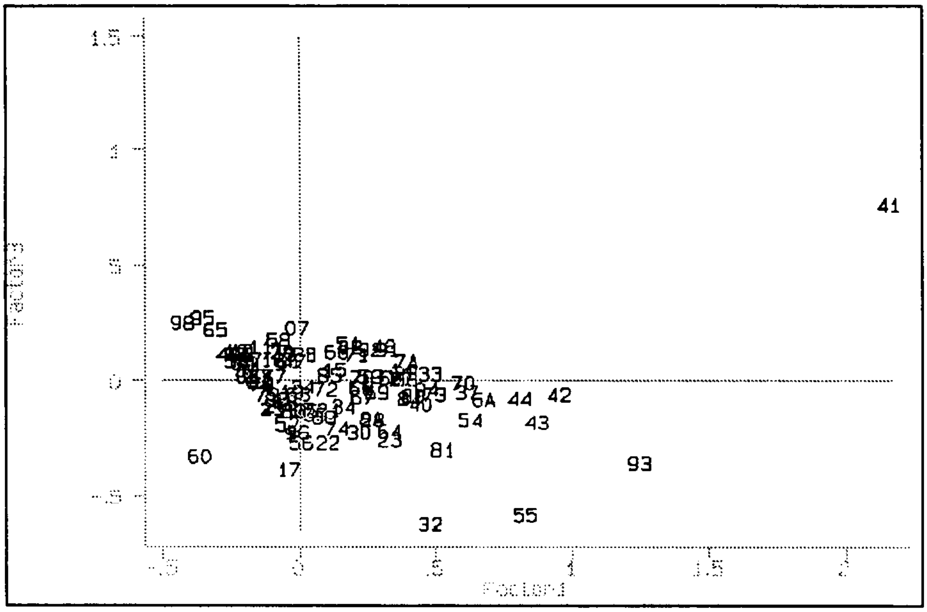

Fig. 4. Factor 3 vs. Factor 1 with 98 samples. The samples.

In the first factor we note the opposition of sodium (positive) and iron (negative). In the second factor we see that only manganese, sodium and arsenic have positive coordinates. In the third factor only sodium has a significant positive coordinate.

We now analyze the results for the samples (Figs. 3 and 4). Sample G41 is clearly anomalous; it has the highest value of sodium but typical values for the rest of the variables. Sample G93 has the lowest value of iron, and low values for several other variables. Sample G60 has the highest value of manganese followed by sample G98, which also has the highest value of iron. Sample G88 has a high value of manganese and the lowest value of potassium.

It seems clear that the composition of the first three factors is the result of the influence of a small group of samples. This influence is related to the sample location. For example, samples G60 and G98 were taken from the deepest part of the lake, where there are iron–manganese nodules [3], whereas samples G41 and G93 were collected in the sandy northern part of the lake.

CA for CHAU93

CA was performed on CHAU98 with samples G41, G93, G60, G98 and G88 deleted from the analysis and made supplementary. We call this data set CHAU93.

The first three factors account for 99.5% of the total variation, but there is a change in the distribution of the variation among them (Table 1). The first factor accounts for 76.4% and the second accounts for 15.7%. This reduces the variation accounted for by the third factor to half of what it was in the previous analysis. The results for CHAU93 are displayed in Figs. 5–8.

The first factor is still composed of sodium (55%) and iron (19%), but now has a contribution by potassium (26%), and no other variable contributes to its formation. Those variables are the only ones that correlate well to this factor, although lanthanum (37%) and hafnium (29%) have some correlation to it.

The second factor is now composed principally of sodium (32%), manganese (23%) and potassium (42%). No element is well correlated to this factor,

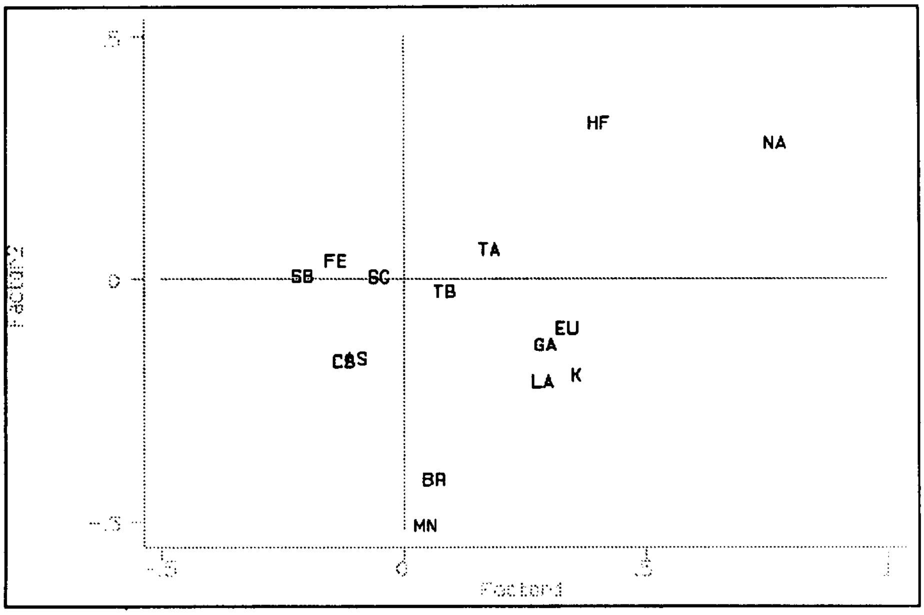

Fig. 5. Factor 2 vs. Factor 1 with 5 supplementary samples. The variables.

but some have correlations between 10 and 40%. Only sodium and hafnium have large positive coordinates on this factor.

The third factor is composed of manganese (75.5%) and potassium (17%), but only manganese is well correlated (61%) to it. It is clear that much

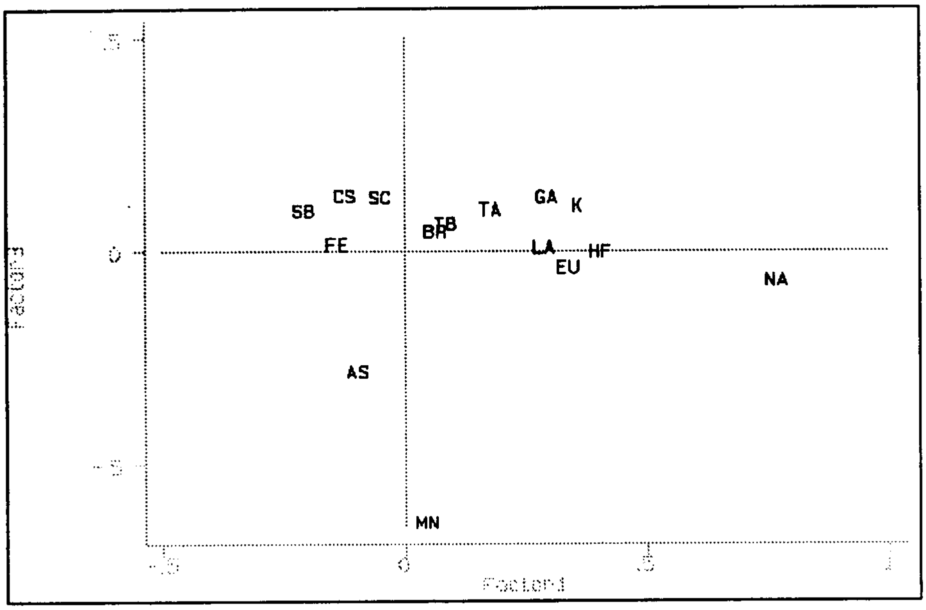

Fig. 6. Factor 3 vs. Factor 1 with 5 supplementary samples. The variables.

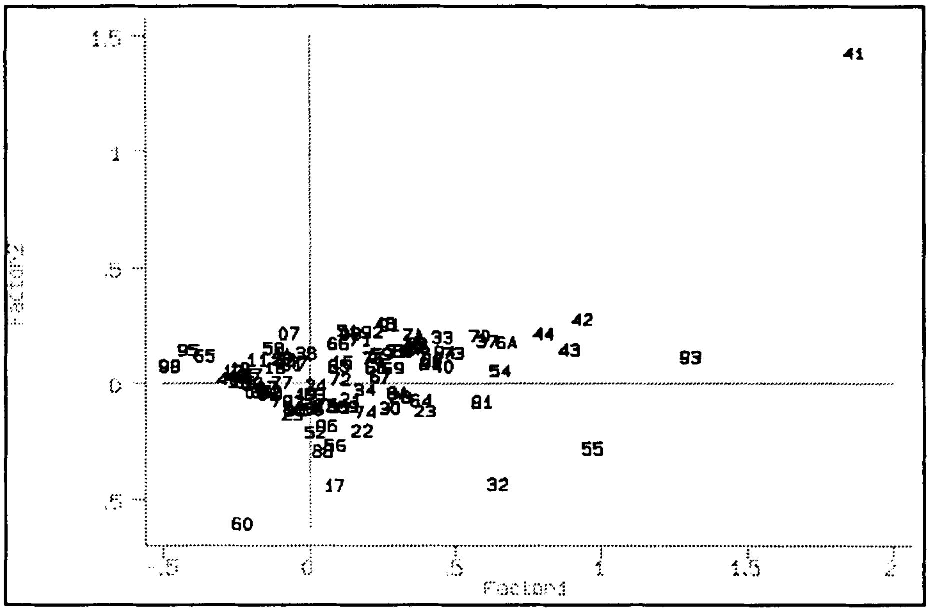

Fig. 7. Factor 2 vs. Factor 1 with 5 supplementary samples. The samples.

of the variation of manganese was due to a small set of samples, and with their deletion the coefficient of variation of this element was reduced by more than half its previous value, from 246% in CHAU98 to 96% in the present case. The deletion of these samples also had an effect on the varia-

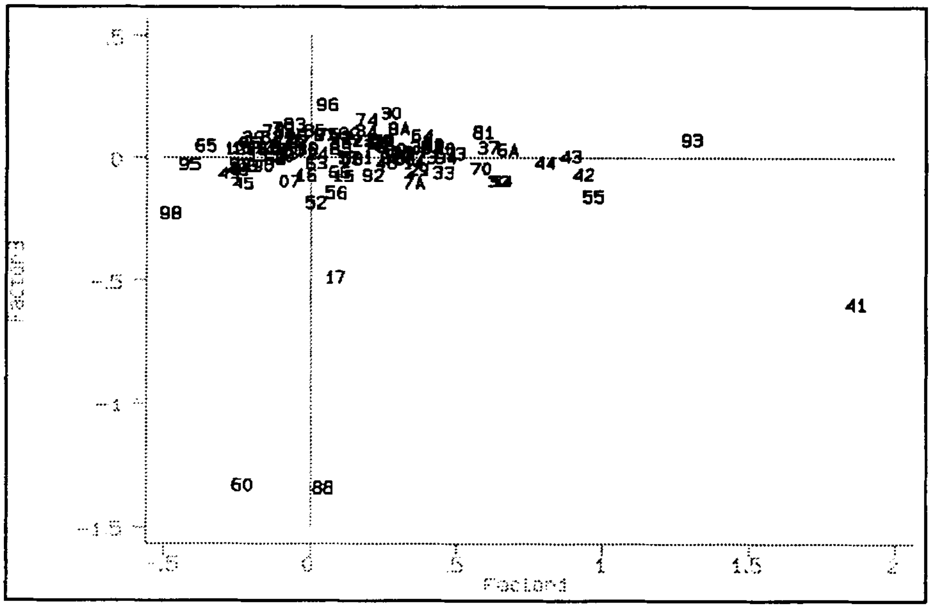

Fig. 8. Factor 3 vs. Factor 1 with 5 supplementary samples. The samples.

tion of sodium and produced a switch in factors two and three from the previous analysis which essentially become factors three and two in this analysis (compare Figs. 1 and 6, and 2 and 5).

The outstanding samples in this analysis are G17 which has the highest value of manganese; G95 with the highest values of arsenic and iron; G65 which has the highest values of cesium, scandium, tantalum and antimony, a large value of iron and a low value of sodium; and G55 having the lowest values of iron, tantalum and scandium. Sample G95 is in the deepest part of the lake, sample G65 is in the center of the lake and sample G55 is on the shore.

The supplementary samples G41, G60, G88, G93 and G98 are projected onto the factorial planes and when graphed with the rest of the samples their position on the graphs is essentially the same as when they contributed to the formation of the factors. It may be concluded that the main effect that these samples produce is a twofold increase of the variation of manganese, and when they are deleted from the analysis, the decrease in the 'noise' produces a switch in the ordering of two factors: the second factor becomes the third factor and vice versa. The fact that manganese is almost uniquely responsible for the original second factor implies that there is no significant linear relationship between this element and the others. One reason for this lack of 'correlation' is the spatial distribution of manganese in the lake.

CA for a modified CHAU98

CA was performed on CHAU98 with sodium, manganese, potassium and iron made supplementary. In this analysis, these variables do not contribute to the formation of the factors, but are projected onto the factorial planes determined by the other elements.

The results (Table 4) show a first factor with a predominance of arsenic, a second factor with hafnium and bromide, a third factor composed of gallium and scandium, bromide contributing to the fourth and fifth factor, and lanthanum making more than 50% of the fifth factor. Five factors are needed to account for 97% of the variation which is now more 'distributed' among factors and among elements. We plot the first three factors

TABLE 4

Absolute and relative contributions of variables when Na, Mn, K and Fe are supplementary

Weights		Factor 1		Factor 2		Factor 3		Factor 4		Factor 5	
		(λ = 0.16)		(λ = 0.07)		(λ = 0.07)		(λ = 0.04)		(λ = 0.02)	
		AC(1)	RC(1)	AC(2)	RC(2)	AC(3)	RC(3)	AC(4)	RC(4)	AC(5)	RC(5)
Eu	0.00503	0.0	0.0	0.0	1.8	0.1	2.2	0.8	16.5	0.7	8.9
Br	0.07367	0.1	0.4	26.4	46.8	5.9	9.6	32.9	28.6	26.7	14.5
As	0.16076	74.2	95.1	7.6	4.4	0.2	0.1	1.0	0.3	0.9	0.2
Ga	0.12238	1.2	3.7	9.4	13.4	52.6	68.2	15.4	10.8	8.9	3.9
La	0.17817	0.0	0.0	3.9	12.0	5.3	14.6	13.8	20.7	55.8	52.4
Hf	0.11115	12.9	31.8	50.7	56.0	3.1	3.1	13.2	7.1	5.3	1.8
Cs	0.04430	0.6	10.3	1.4	11.5	5.5	40.0	1.5	5.9	1.4	3.4
Tb	0.00568	0.1	18.7	0.1	7.8	0.1	3.5	0.1	3.4	0.0	0.0
Sc	0.26495	9.6	41.6	0.0	0.1	22.1	39.1	16.6	15.9	0.0	0.0
Ta	0.01131	0.5	34.4	0.1	5.0	0.0	0.0	0.0	0.4	0.0	0.2
Sb	0.02261	0.8	12.0	0.2	1.3	5.3	33.6	4.7	15.9	0.2	0.5
Na	0.00000	0.0	18.6	0.0	46.9	0.0	11.1	0.0	17.7	0.0	0.7
Mn	0.00000	0.0	64.7	0.0	2.4	0.0	11.0	0.0	0.3	0.0	16.2
K	0.00000	0.0	29.0	0.0	0.2	0.0	6.2	0.0	13.2	0.0	16.6
Fe	0.00000	0.0	8.4	0.0	0.2	0.0	0.1	0.0	6.3	0.0	56.9

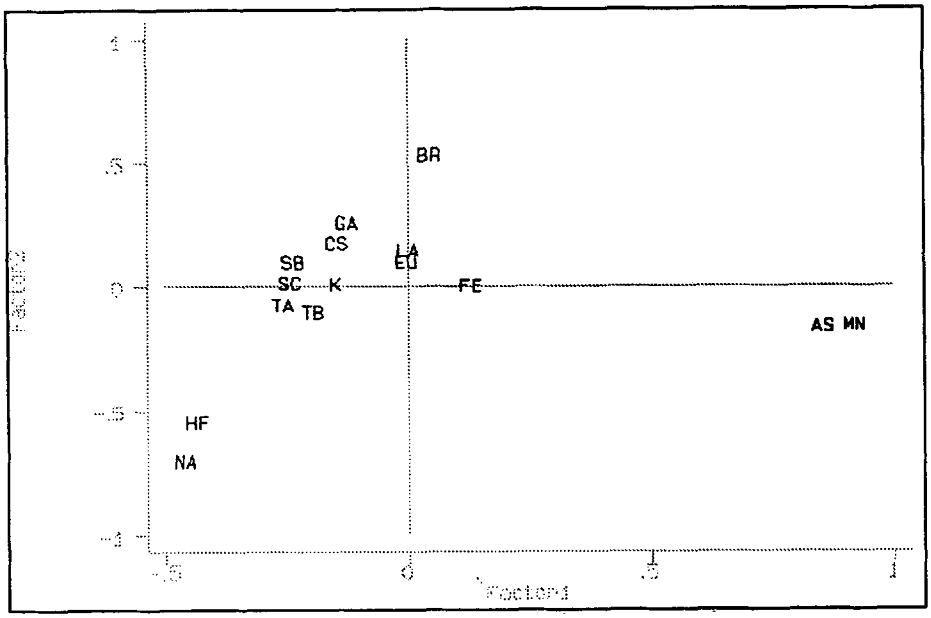

Fig. 9. Factor 2 vs. Factor 1 with Na, Mn, K and Fe supplementary. The variables.

looking simultaneously at the resulting graphs (Figs. 9–11 and 12–14). The plots suggest some interesting conclusions.

Arsenic is the nearest to manganese (related to depth) and far from sodium (related to shore locations); this was seen before (Fig. 1), but it is

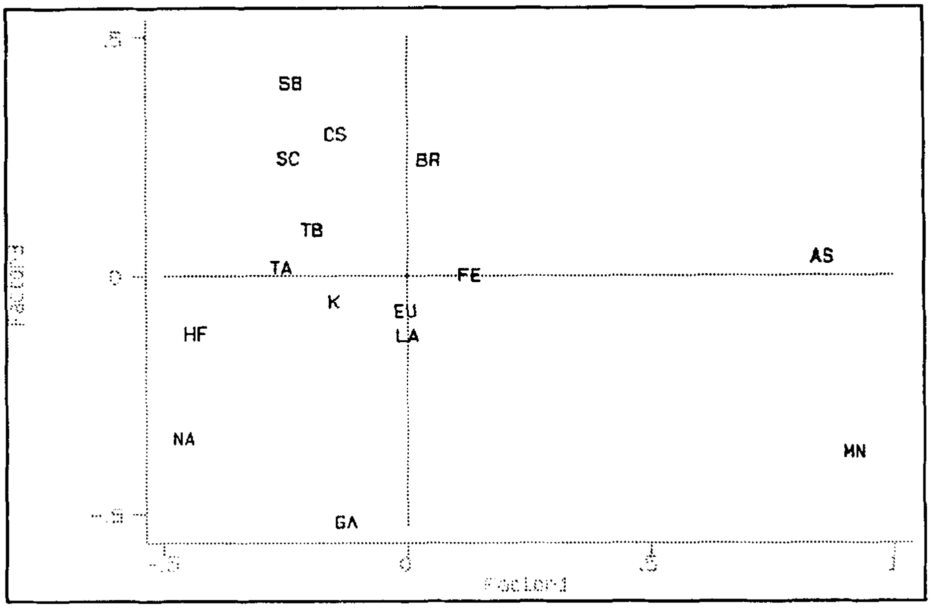

Fig. 10. Factor 3 vs. Factor 1 with Na, Mn, K and Fe supplementary. The variables.

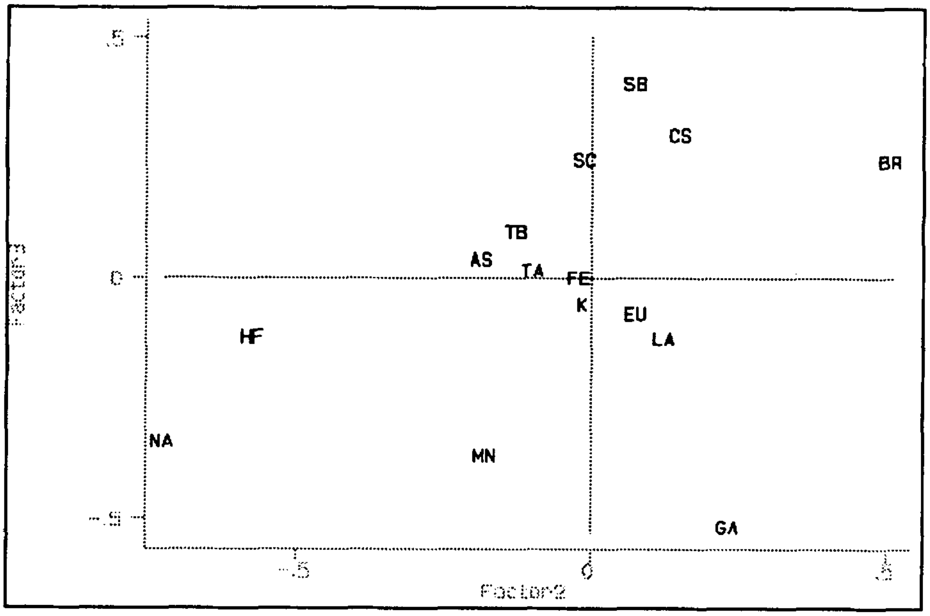

Fig. 11. Factor 3 vs. Factor 2 with Na, Mn, K and Fe supplementary. The variables.

evident in this analysis. It is also clear that arsenic is not well correlated to any of the other elements.

Hafnium, on the other hand, is the only element found near sodium. It is also removed from the other elements.

Scandium, cesium and antimony form a recog-

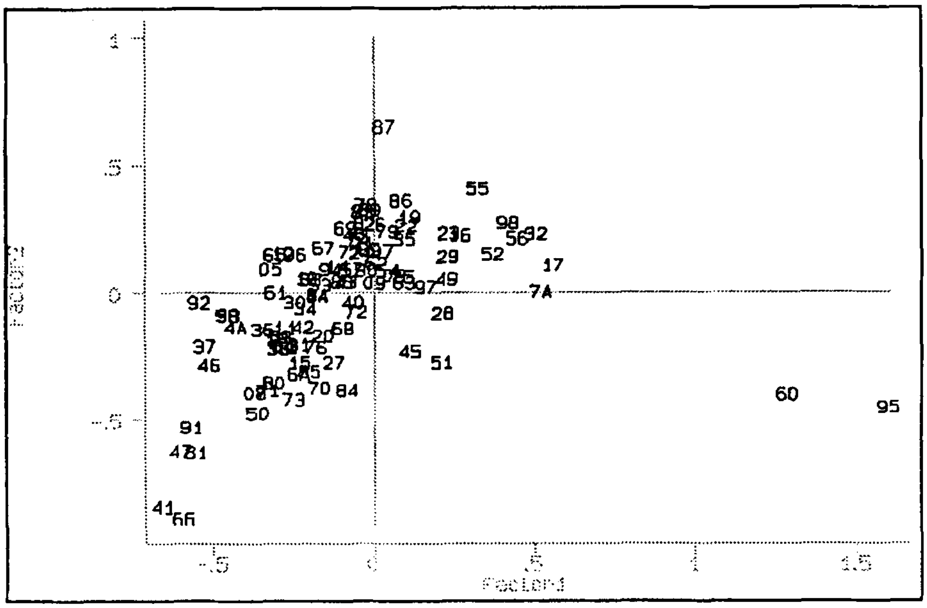

Fig. 12. Factor 2 vs. Factor 1 with Na, Mn, K and Fe supplementary. The samples.

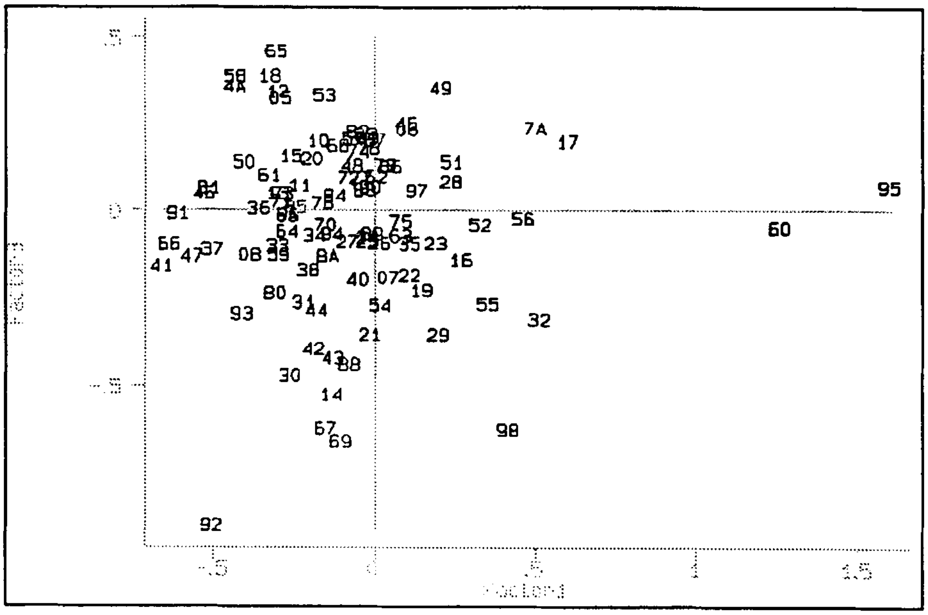

Fig. 13. Factor 3 vs. Factor 1 with Na, Mn, K and Fe supplementary. The samples.

nizable cluster, which apparently could also include tantalum and terbium, although we will see in the discussion of error profiles that this would be wrong.

The plots of the samples show some anomalous points: G95 (highest value of arsenic), G60 (high value of arsenic), G41 (high value of hafnium, low values of others), G66 (high value of hafnium, low

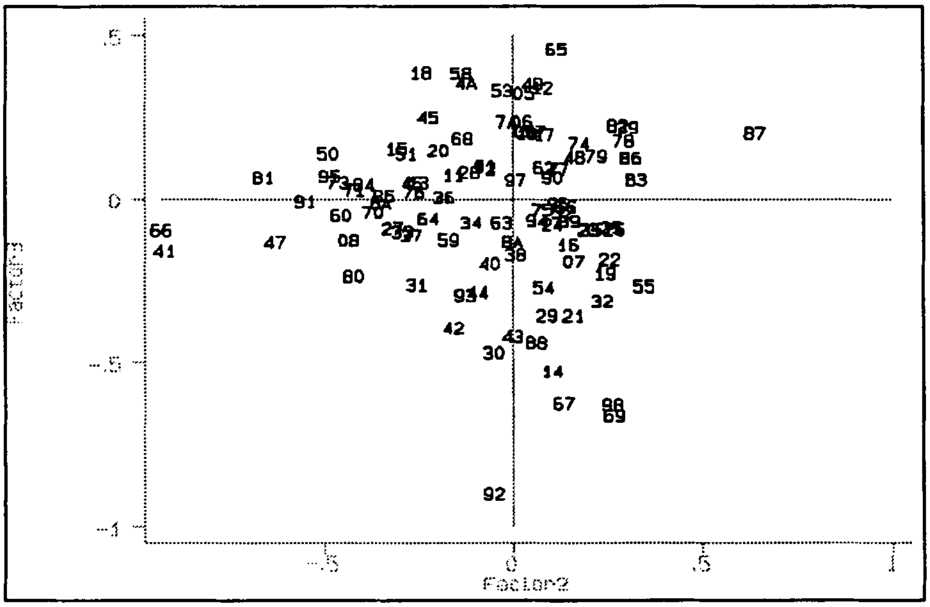

Fig. 14. Factor 3 vs. Factor 2 with Na, Mn, K and Fe supplementary. The samples.

TABLE 5

Error profiles (EP) and cumulative relative contributions (CRC) for the variables when three factors are retained

	EP	CRC
Eu	0.452E+00	25.6
Na	0.419E−06	100.0
Mn	0.689E−05	100.0
K	0.350E−05	100.0
Br	0.975E+00	17.6
As	0.584E+00	18.2
Ga	0.731E+00	16.7
La	0.127E+00	54.7
Hf	0.474E+00	30.6
Cs	0.221E+00	26.0
Tb	0.282E+00	14.6
Sc	0.125E+00	27.6
Fe	0.514E−07	100.0
Ta	0.301E+00	18.5
Sb	0.446E+00	12.1

values of others), G87 (highest value of bromide), and G92 (highest value of gallium). We note the effects of G41 and G60 on the previous analyses.

Analysis of the error profiles

The error profiles can be seen as a partition of the global variation explained by the discarded factors. Although this variation may be small, a particular entry (row, column or element) of the input matrix could be poorly reconstructed when using only some of the factors.

For CHAU98 the first three factors 'explained' almost all of the variation which is almost due to the overwhelming presence of a small number of samples and variables. However, as can be seen in Table 5, the other variables would be poorly reconstructed with only three factors, and since most of them are not well 'correlated' to the factorial space defined by sodium, manganese, potassium and iron, i.e. the first three factors, their variation is not explained by this solution.

When sodium, manganese, potassium and iron are made supplementary, it is seen that the variation of the remaining variables can be factored and this factoring provides another way to look at possible relationships between elements. We see in Table 6 that five factors will not reconstruct europium, terbium, tantalum and antimony, although antimony is correlated to the factorial space. We conclude that terbium and tantalum should not be clustered with scandium and cesium.

We can use the error profiles to find the number of factors needed to reconstruct a particular element. For example, it is seen that four factors are needed to reconstruct arsenic, including the first three from the previous analysis.

For the samples, a box plot of the error profiles (Figs. 15 and 16) will show those samples that are poorly reconstructed using the retained factors. These samples would not be outliers in the traditional sense, and probably would not show well on the plots of the projections onto the factorial planes, but they would be related to the variables that dit not contribute to the factors. For example, samples G87 and G92 which were discussed before and G73A which has high values of bromide and arsenic show as poorly reconstructed samples when using the three factors dominated by sodium, manganese, potassium and iron. These same three samples are not poorly reconstructed when sodium, manganese, potassium and iron are supplementary because the composition of the factors change; instead, sample G27, which has the highest value of europium, appears as the poorest reconstructed sample.

TABLE 6

Error profiles and cumulative relative contributions for the variables when Na, Mn, K and Fe are supplementary and five factors are retained

	EP	CRC
Eu	0.238E+00	29.5
Br	0.158E−03	100.0
As	0.112E−05	100.0
Ga	0.911E−04	100.0
La	0.348E−03	99.7
Hf	0.911E−03	99.8
Cs	0.581E−01	71.1
Tb	0.132E+00	33.5
Sc	0.470E−02	96.6
Ta	0.114E+00	39.9
Sb	0.168E+00	63.1
Na	*********	95.0
Mn	*********	94.6
K	*********	65.2
Fe	*********	71.9

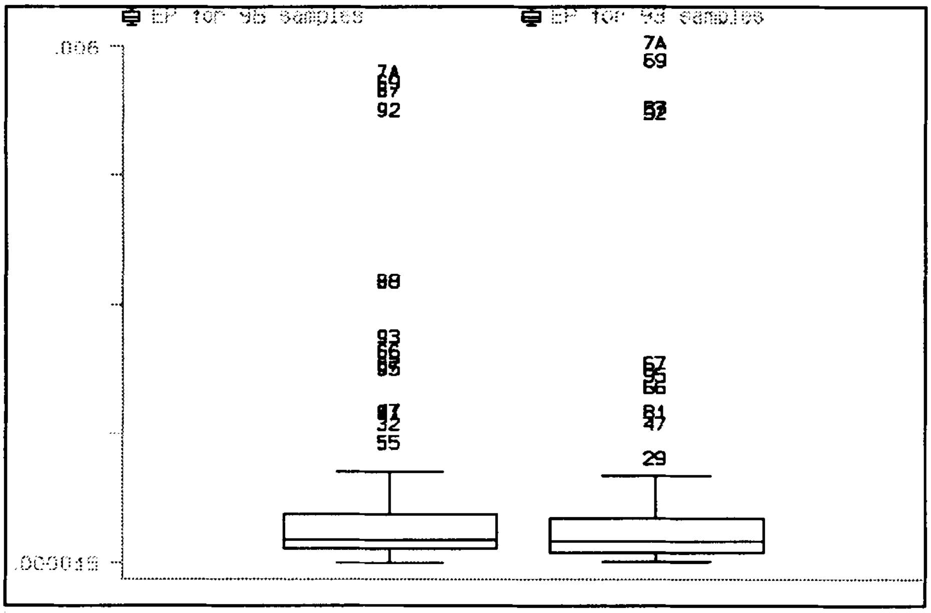

Fig. 15. Error profiles for CHAU98 and CHAU93 when three factors are retained. The samples.

CA for CHAU88

CA was performed on CHAU88, using sand, silt and clay content as main variables. Sodium, manganese, potassium, gallium, lanthanum, scandium, iron, hafnium, bromide and arsenic were made supplementary and projected onto the factorial plane determined by the main variables (Figs. 17 and 18).

The values of the soil variables add up to one

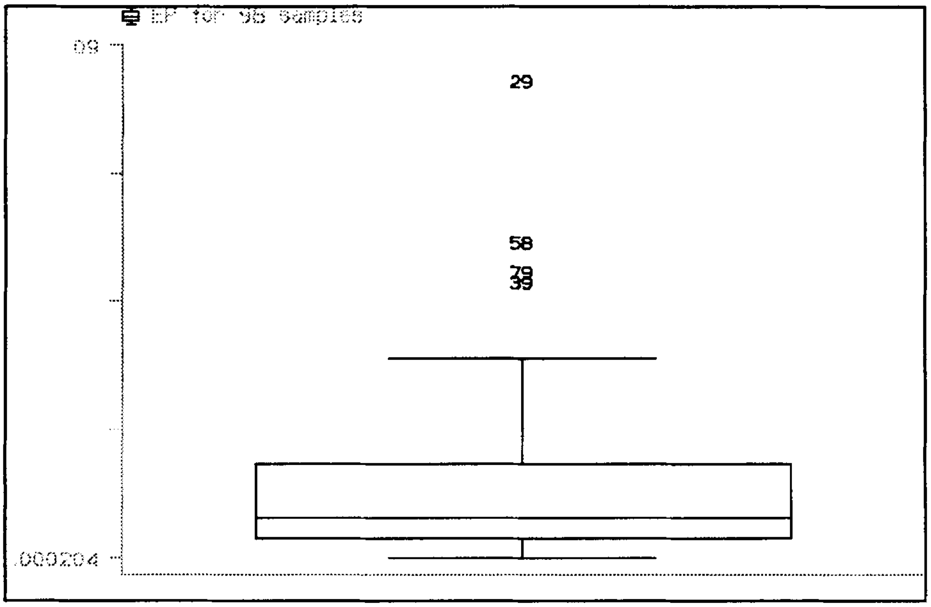

Fig. 16. Error profiles for CHAU98 when Na, Mn, K and Fe are supplementary and five factors are retained. The samples.

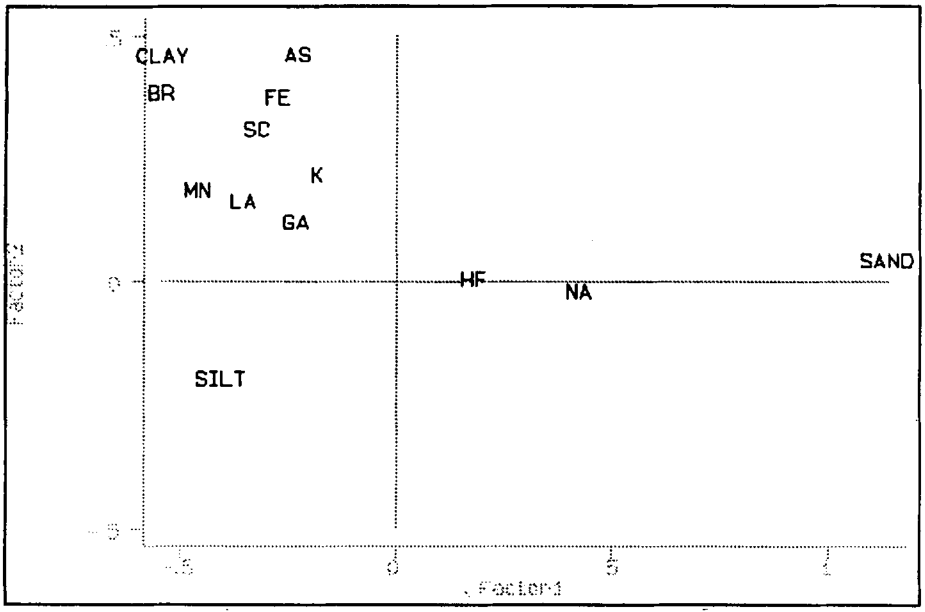

Fig. 17. Variables projected onto space of soil variables.

hundred percent on each sample, and thus the dimensionality of the variable space is two. Since CA starts by making that type of normalization on the input matrix, the linear dependency of the variables will not produce any spurious correlations; in fact, the dimension of the factor's space

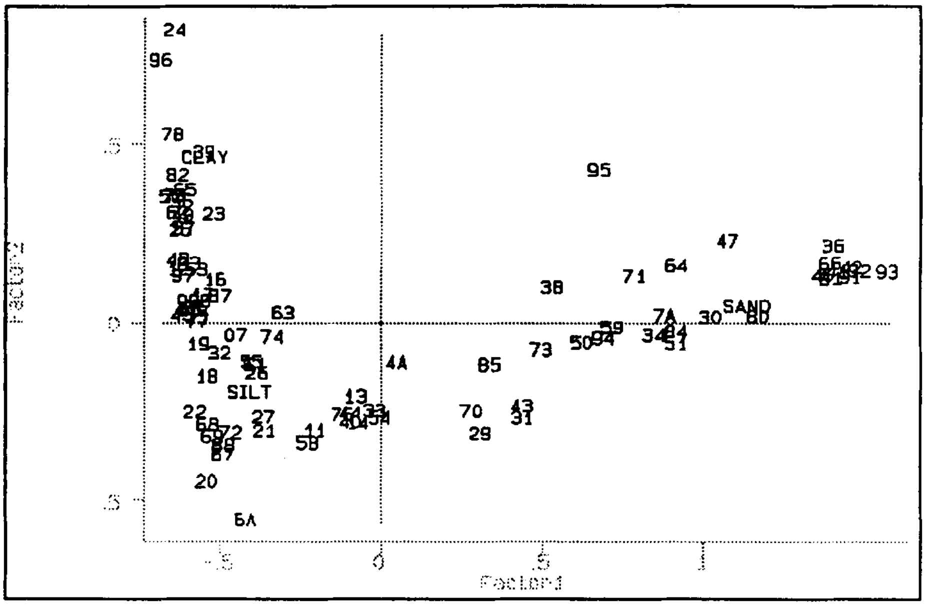

Fig. 18. 88 samples projected onto space of soil variables.

will be precisely two and the plot of the projection onto the factorial plane will reproduce 100% of the variation. The error profiles will therefore be equal to zero.

The two nontrivial eigenvalues account for 86% and 14% of the variation, respectively. The first factor separates sand from silt and clay, and the second factor separates silt from clay. Thus the soil variables plot (Fig. 17) as vertices of a triangle, with the supplementary variables falling 'inside' the triangle. From the graph it is easy to see that only sodium and hafnium are correlated to sand, with the rest of the variables lying near the vertex determined by clay. Note that this is particularly true of arsenic, suggesting that it is to be found in clay-rich sediments [2].

The projection of the samples (Fig. 18) onto the factorial plane shows several clusters. Various methods to separate clusters could be given based on a distance measure, but, since sand, silt and clay are plotted on different quadrants, we will use this simple criterion: samples belong to different clusters according to the signs of their coordinates, (+,+), (−,+), (−,−) and (+,−). This leads to the following clusters:

- A cluster of samples belonging to the northern portion of the lake, like samples G41, G42, G46, G47, C91, G92, and G93, or to shore locations, like samples G36, G66, G80 and G81. This cluster represents the group of samples from the sandy portions of the lake.
- A cluster of samples from locations in the deepest portions of the lake, for example samples G39, G52, G65, G78, G82. The plot shows that depth is well correlated to clay and the cluster also includes samples from clay-rich sediments, for example sample G24.
- A cluster of samples collected on the south end of the lake, for example, samples G67, G68, G69, G69A, G72, and samples collected in the narrow center portion of the lake, like samples G18, G19, G20, G21, G22, and G32. This cluster is correlated with silt.
- A cluster of samples from near-shore locations where there is a balance between sand and silt. Samples G29, G31, G43 and G70 belong to this group.

COMPARISON WITH PREVIOUS RESULTS AND METHODS

We emphasize here that the conclusions from the use of CA were obtained after a series of analysis and they are not to be taken as inferences in the traditional statistical sense. CA is an exploratory technique whose main objective is to describe patterns found in the data, without assuming a statistical model for the population from where the data was obtained. However, CA does assume an algebraic model which may not have a simple interpretation. A summary of our results and a comparison with those found in previous studies can be given as follows:

- CA discriminates sodium from the rest of the elements, relating this to the sand content of the sample. Only hafnium shows some similarity to sodium.
- CA projects most of the elements near silt or clay. Arsenic and bromide, in particular, are strongly dependent on the clay content of the sample.
- CA shows that scandium and cesium have similar profiles, and antimony is correlated to them.
- CA highlights the uniqueness of manganese due to the localized nature of its occurrence.

These results from CA are in agreement with the results that were found using factor analysis [4]. However no a priori number of factors have to be postulated in CA. Performing CA on different sets of variables, complemented with various supplementary variables, we found 'solutions' (i.e. models) with three, five and two factors, each solution highlighting specific aspects of the description of the data.

The correspondence between analyses for rows and columns allows us to also describe the interrelationships among samples. This we have done relating the behavior of certain variables to the effect of specific samples, and, correspondingly, using variables to explain sample clusters. The cluster found by CA are essentially the same ones proposed by Hopke [4], using a dissimilarity measure.

Through the use of diagnostics, CA was useful in identifying outliers and influential samples, and in isolating samples with distinctive features. This is not easy to do with other techniques unless an underlying model, such as multivariate normality for example, is assumed from the onset.

Finally, the emphasis on the graphical description of the output from CA made it easier to analyze the results. We did not plot samples and variables on the same diagram, although this is a common practice among early practitioners [7,10] of CA, since this has been the subject of major criticisms [16,17].

THE PROGRAM

The CA program that was used for these analyses was written in FORTRAN77 and compiled with a Microsoft$^{\mathrm{tm}}$ compiler. The program is based on the code given in the paper by David et al. [10], with many changes and additions. The main changes are that it runs on a microcomputer and that it is interactive; the main additions are the inclusion of more diagnostics and input and output options.

The program shows several screens with intermediate results from CA, allowing the user to change previous options. Only when the user is satisfied with the fit, or when certain limits are reached, are the results written to output files. The output includes:

- Basic descriptive statistics of the variables.
- List of eigenvalues and percentage of variation explained by each.
- Variable and/or sample coordinates and/or factors.
- Variable and/or sample weights, and absolute and relative contributions for each of the factors retained.
- Error profiles and sums of relative contributions for each variable and/or sample.
- A reconstructed matrix from the set of retained factors.
- Coordinates and relative contributions of supplementary variables on each of the retained factors.
- Files for plotting or graphs of factorial planes with the user specifying the factors to be plotted and what is to be plotted: variables and/or samples factors and/or coordinates.

The program computes the basic statistics for the variables and performs some checks on the input values to prevent overflow errors. In its present version it can handle matrices with up to ten thousand entries.

CONCLUSIONS

We have shown that correspondence analysis can be applied to environmental data sets where the variables are measured on a ratio scale and where the samples are collected in space.

Because of the dual scaling of rows and columns of the data matrix and the transition formulas, CA performs R-mode and Q-mode analysis simultaneously with results comparable to the joint use of factor analysis and cluster analysis. The solution found by CA, however, is easier to obtain and provides more information about the underlying structure of the data.

In the approach given in the present paper an emphasis is given to the graphical display of the results, and to the use of several diagnostic measures that give a goodness-of-fit evaluation and that are helpful in detecting influential samples and outliers, assessing the stability of the solution and the robustness of the method.

NOTICE

Although the research described in this article has been funded wholly or in part by the US Environmental Protection Agency through a Cooperative Research Agreement with the University of Arizona, it has not been subjected to Agency review and therefore does not reflect the views of the Agency and no official endorsement should be inferred.

REFERENCES

1 S.A. Lis and Ph.K. Hopke, Anomalous arsenic concentrations in Chautauqua lake, *Environmental Letters*, 5 (1973) 45–51.

2 D.F. Ruppert, Ph.K. Hopke, P. Clute, W. Metzger and D. Crowley, Arsenic concentrations and distribution in Chautauqua lake sediments, *Journal of Radioanalytical Chemistry*, 23 (1974) 159–169.

3 Ph.K. Hopke, D.F. Ruppert, P.R. Clute, W.J. Metzger and D.J. Crowley, Geochemical profile of Chautauqua lake sediments, *Journal of Radioanalytical Chemistry*, 29 (1976) 39–59.

4 Ph.K. Hopke, The application of multivariate analysis for interpretation of the chemical and physical analysis of lake sediments, *Journal of Environmental Science and Health*, All (1976) 367–383.

5 L. Lebart, A. Morineau and K.M. Warwick, *Multivariate Descriptive Statistical Analysis*, Wiley, New York, 1984.

6 M. Greenacre, *Theory and Applications of Correspondence Analysis*, Academic Press, London, 1984.

7 M. David, C. Campiglio and R. Darling, Progress in *R*- and *Q*-mode analysis: Correspondence analysis and its application to the study of geological processes, *Canadian Journal of Earth Sciences*, 11 (1974) 131–146, 603 and 1497–1499.

8 H. Teil, Correspondence factor analysis: an outline of its method, *Mathematical Geology*, 7 (1975) 3–12.

9 H. Teil and J.L. Cheminee, Application of correspondence factor analysis to the study of major and trace elements in the Erta Ale Chain (Afar, Ethiopia), *Mathematical Geology*, 7 (1975) 13–30.

10 M. David, M. Dagbert and Y. Beauchemin, Statistical analysis in geology: correspondence analysis method, *Quarterly of the Colorado School of Mines*, 72 (1977) 1–57.

11 D. Zhou, T. Chang and J.C. Davis, Dual extraction of *R*-mode and *Q*-mode factor solutions, *Mathematical Geology*, 15 (1983) 581–605.

12 J.C. Davis, *Statistics and Data Analysis in Geology*, Wiley, New York, 2nd ed., 1986.

13 K.R. Gabriel, The biplot graphic display of matrices with application to principal components analysis, *Biometrika*, 58 (1971) 453–467.

14 H.O. Lancaster, The structure of bivariate distributions, *Annals of Mathematics and Statistics*, 29 (1958) 719–736.

15 F. Critchley, Influence in principal components analysis, *Biometrika*, 72 (1985) 627–636.

16 M.J. Greenacre, The Carroll–Green–Schaffer scaling in correspondence analysis: a theoretical and empirical appraisal, *Journal of Marketing Research*, 26 (1989) 358–365.

17 L.A. Goodman, Some useful extensions of the usual correspondence analysis approach and the usual log–linear model approach in the analysis of contingency tables, *International Statistical Review*, 54 (1986) 243–309.