

CORRESPONDENCE ANALYSIS USED IN THE EVALUATION OF LAKEWATER CHEMISTRY IN THE ADIRONDACKS

HANNAN RASMUSSEN RHODES AND DONALD E. MYERS

Department of Applied Mathematics, University of Arizona, Tucson, AZ 85721, U.S.A.

SUMMARY

Correspondence analysis (CA) was applied to lakewater data in order to study the effects of acidic deposition on the geochemical composition of lakes in the Adirondacks. The lake chemistry data analyzed were taken from the Eastern Lake Survey – Phase I (ELS-I) conducted by the U.S. Environmental Protection Agency. CA was used to identify ‘outlying’ lake samples as well as ‘superfluous’ and ‘unresolved’ analytes. Correlational relationship among analytes were also examined.

KEY WORDS Correspondence analysis Eastern Lake Survey – Phase I data
Acidic deposition

INTRODUCTION

The impact of acidic deposition on the geochemical composition of lakes can be investigated statistically through regionally correlated relationships among chemical elements. However, a large number of analytes can be measured, making conventional correlational analysis as well as other data analyses difficult. Thus it is reasonable to ask if some analytes can be deleted from the list of analytes directly used in subsequent statistical analyses.

This paper describes the application of correspondence analysis (CA), which is an exploratory multivariate statistical technique, to lakewater data obtained from the Eastern Lake Survey – Phase I (ELS-I) conducted by the U.S. Environmental Protection Agency. The objectives of the CA were threefold. The primary objective was to obtain a reduction in the number of analytes in order to make subsequent data analyses and interpretation easier. This was accomplished by determining which analytes could be made supplementary without affecting the CA results. The second objective consisted of using CA as a tool for identifying outlying or atypical lake samples and analytes. The third objective was to explore relationships among the analytes by portraying the data geometrically and examining the plots for clusters of variables. This work is important both because of the results obtained and because there have been very few instances of the application of CA to environmental data.

THEORY OF CORRESPONDENCE ANALYSIS

Correspondence analysis is an exploratory multivariate statistical technique that exploits geometric properties of multidimensional data to reveal basic linear dependences. Graphical

0886–9383/91/030273–18\$09.00

© 1991 by John Wiley & Sons, Ltd.

Received 1 May 1990

Accepted (revised) 17 July 1990

displays are produced in which the rows and columns of a data matrix are depicted as row and column points projected simultaneously onto a two-dimensional factorial plane. The theory is well understood and discussed in many textbooks,¹⁻³ so only the main ideas are provided here.

Given a data matrix consisting of n rows of lake samples and p columns of analytes, two simultaneous analyses are performed: one in the p -dimensional space of analytes, \mathbb{R}^p , and one in the n -dimensional space of lake samples, \mathbb{R}^n . Only the analysis in \mathbb{R}^p is discussed here since the results in \mathbb{R}^n can be derived by using duality relationships between the two spaces.

Rather than working with the original data, the analyses are performed on row (and column) profiles. From an initial data matrix of non-negative entries, $\mathbf{X} = [x_{ij}]_{(n \times p)}$, a frequency matrix is computed by dividing each element of the data matrix by the sum of all the elements, i.e.

$$\mathbf{F} = [f_{ij}] = \left[\frac{x_{ij}}{x_{..}} \right]_{(n \times p)}$$

where

$$x_{..} = \sum_{i=1}^n \sum_{j=1}^p x_{ij}$$

Two normalizing matrices are constructed,

$$\mathbf{D}_n = \text{diag}[f_{i.}]_{(n \times n)}$$

where

$$f_{i.} = \sum_{j=1}^p f_{ij}$$

and

$$\mathbf{D}_p = \text{diag}[f_{.j}]_{(p \times p)}$$

where

$$f_{.j} = \sum_{i=1}^n f_{ij}$$

Each element of the frequency matrix is then weighted to obtain a matrix of row profiles,

$$\mathbf{D}_n^{-1} \mathbf{F} = \left[\frac{f_{ij}}{f_{i.}} \right]_{(n \times p)}$$

These row profiles define a cloud of n points in p -dimensional variable space. The center of gravity of the cloud is the point $(f_{.1}, \dots, f_{.p})$ and around this center are determined factorial axes. The axes are the principal axes of inertia within the cloud.

Since we are comparing variables which may have different orders of magnitude, we need a distance function in which the variables are weighted unequally. The usual Euclidean distance is unsuitable since it gives the same weight to each variable, so the chi-squared distance is used to eliminate the influence of variables with large absolute values compared to the rest. The chi-squared distance between variable j and variable k is given by

$$D^2(j, k) = \sum_{i=1}^n \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ki}}{f_{.k}} \right)^2$$

With distances in the cloud defined, we may set about finding the principal axes of inertia.

Let $\mathbf{u}_{(p \times 1)}$ be a unit vector for the chi-squared distance in \mathbb{R}^p , i.e. such that $\mathbf{u}^T \mathbf{D}_p^{-1} \mathbf{u} = 1$. The vector of the n projections of the row profiles on axis \mathbf{u} is $\mathbf{v}_{(n \times 1)} = \mathbf{D}_n^{-1} \mathbf{F} \mathbf{D}_p^{-1} \mathbf{u}$. To find the principal axis \mathbf{u} , maximize the weighted sum of squares of the projections $\mathbf{v}^T \mathbf{D}_n \mathbf{v} = \mathbf{u}^T \mathbf{D}_p^{-1} \mathbf{F}^T \mathbf{D}_n^{-1} \mathbf{F} \mathbf{D}_p^{-1} \mathbf{u}$ subject to the constraint $\mathbf{u}^T \mathbf{D}_p^{-1} \mathbf{u} = 1$. Applying standard Lagrangian multiplier techniques, we find that \mathbf{u} is an eigenvector of the $p \times p$ matrix $\mathbf{F}^T \mathbf{D}_n^{-1} \mathbf{F} \mathbf{D}_p^{-1}$, corresponding to the largest eigenvalue $\lambda = \mathbf{u}^T \mathbf{D}_p^{-1} \mathbf{F}^T \mathbf{D}_n^{-1} \mathbf{F} \mathbf{D}_p^{-1} \mathbf{u}$. The 'first factor' $\boldsymbol{\varphi}$ is obtained from the projection operator on the principal axis \mathbf{u} , $\boldsymbol{\varphi}_{(p \times 1)} = \mathbf{D}_p^{-1} \mathbf{u}$. Let \mathbf{u}_α denote the eigenvector corresponding to eigenvalue λ_α . \mathbf{u}_α is called the α th principal axis and $\boldsymbol{\varphi}_\alpha$ the α th factor. The vectors of the projections on principal axis α are then obtained from the scaled factor $\hat{\boldsymbol{\varphi}}_\alpha = \lambda_\alpha^{1/2} \boldsymbol{\varphi}_\alpha$. The data may be recreated using the reconstruction formula

$$f_{ij} = f_i \cdot f_j \left(1 + \sum_{\alpha} \lambda_{\alpha}^{1/2} \psi_{i\alpha} \varphi_{j\alpha} \right)$$

where $\psi_{\alpha} = \lambda_{\alpha}^{-1/2} \mathbf{D}_n^{-1} \mathbf{F} \boldsymbol{\varphi}_{\alpha}$ is the α th factor in the dual space \mathbb{R}^n .

Two coefficients are computed for every sample i and variable j , showing how well a particular sample or variable is reconstructed as well as how much the samples and variables contribute to particular factors. The absolute contribution of variable j to principal axis α is defined by

$$C_{\alpha}^A(j) = f_j \boldsymbol{\varphi}_{\alpha j}^2$$

and satisfies $\sum_{j=1}^p C_{\alpha}^A(j) = 1$. The relative contribution of principal axis α to variable j is defined by

$$C_{\alpha}^R(j) = \frac{\hat{\boldsymbol{\varphi}}_{\alpha j}^2}{\sum_{\alpha} \hat{\boldsymbol{\varphi}}_{\alpha j}^2}$$

and satisfies $\sum_{\alpha} C_{\alpha}^R(j) = 1$. The absolute contributions indicate the proportion of inertia of a principal axis explained by each variable; the relative contributions indicate the part of the inertia of a variable explained by a principal axis.

Variables not used in the formation of the principal axes can be made supplementary and projected onto the space generated by the retained factors by creating a new point j with profile

$$\frac{x_{ij}}{x_{.j}}$$

where

$$x_{.j} = \sum_{i=1}^n x_{ij}$$

Point j is then projected onto principal axis α , i.e. $\hat{\boldsymbol{\varphi}}_{\alpha j} = \lambda_{\alpha}^{-1/2} \mathbf{D}_p^{-1} \mathbf{F}^T \psi_{\alpha}$.

DATA ANALYSIS

Eastern Lake Survey – Phase I (ELS-I) data set

The lake chemistry data analyzed in this study were taken from the Eastern Lake Survey – Phase I (ELS-I) which is part of the National Surface Water Survey (NSWS) initiated in 1983 by the United States Environment Protection Agency (EPA) as part of the National Acid

Precipitation Assessment Program (NAPAP). ELS-I was designed to statistically describe the chemical status of lakes in areas of the eastern United States containing the majority of low-alkalinity lakes (acid-neutralizing capacity (ANC) $\leq 400 \mu\text{eq l}^{-1}$). Complete details on statistical design, lake selection, analytical methodologies, field methods and quality assurance protocols are given elsewhere,⁴⁻⁸ so only the key design features are presented here. Reference 7 also provides some results and discussions of associations among primary chemical analytes relevant to lake acidification processes.

A stratified design was used wherein sample lakes were allocated equally among strata. The first level of stratification was three regions of the eastern U.S. (the Northeast, the Upper Midwest and the Southeast) expected to be the most susceptible to change as a result of acidic deposition. The second stratification level was obtained by dividing each region into subregions exhibiting geographical homogeneity with respect to water quality, physiography, vegetation, climate and soils. Finally, each subregion was further subdivided into three alkalinity map classes (ANC $< 100 \mu\text{eq l}^{-1}$, $100-200 \mu\text{eq l}^{-1}$, ANC $> 200 \mu\text{eq l}^{-1}$) yielding the third stratification level.

Each lake with surface area 4 ha or more appearing on 1:250 000-scale U.S. Geological Survey (USGS) topographic maps was assigned a unique number and entered into a computer file in numeric order. The lakes to be sampled were then selected statistically from each stratum by systematic sampling of the ordered list following a random start.

ELS-I was conducted in the fall of 1984 from 7 October to 14 December. The fall season was chosen since the spatial variation within a lake is reduced at this time. Water samples were collected from a depth of 1.5 m and the sampling was done at the apparently deepest part of the lake to provide a sample from the dominant water mass. A single index sample representing the essential characteristics of the lake was chosen in order to maximize both lake number and spatial coverage on a large geographic scale. A total of 26 analytes thought to influence or be influenced by surface water acidification and several physical attributes were measured for each lake.

Data analysis overview

The analysis presented here is restricted to region 1, the Northeast, which is divided into five subregions: the Adirondacks (1A), the Poconos/Catskills (1B), Central New England (1C), Southern New England (1D) and Maine (1E). Nineteen of the 26 concentrations from the PC version of the ELS-I data set were selected for the CA: equilibrated pH, acid neutralizing capacity (ANC), sulfate (SO_4^{2-}), calcium (Ca^{2+}), extractable aluminum (Al^{3+}), dissolved organic carbon (DOC), magnesium (Mg^{2+}), sodium (Na^+), potassium (K^+), ammonium (NH_4^+), nitrate (NO_3^-), chloride (Cl^-), fluoride (F^-), iron (Fe^{3+}), manganese (Mn^{2+}), silica (SiO_2), total phosphorus (P), equilibrated dissolved inorganic carbon (DIC) and bicarbonate (HCO_3^-). Bicarbonate was not measured directly but was estimated from the pH and DIC values. The seven measurements not included in the analysis are: turbidity, colour, measured conductivity, calculated conductivity, closed pH, closed DIC and total Al. Although CA does not require that the variables be measured in the same units, it should be noted that the reported units are not commensurate: mg l^{-1} was used for DIC, DOC and SiO_2 ; Fe, Al, Mn and P concentrations were given in $\mu\text{g l}^{-1}$; while the remaining analytes were reported in $\mu\text{eq l}^{-1}$. Also, since CA requires non-negative data, a constant of 50 was added to ANC in order to offset the negative alkalinity values.

The CAs were carried out using the CORRES program developed by Fernando Avila Murillo at the University of Arizona. Output from the program includes:

- (1) a list of eigenvalues along with the percentage of variation explained by each associated factor
- (2) a table with the weight of each variable and its absolute and relative contributions to the set of retained factors
- (3) an analysis of the error by the set of retained factors, giving the sum of relative contributions for the different variables as well as a chi-squared norm estimate of the error of reconstruction
- (4) graphical representations of the data obtained by plotting sample co-ordinates and variable factors in two-dimensional factorial planes
- (5) projections of supplementary variables onto a factorial axis.

Note that, as a result of the normalization performed in CA, a trivial eigenvalue with value 1.0 is obtained; this first eigenvalue is not listed in the output. Because of the stratified design of the ELS-I survey, separate CAs were performed on each of the three alkalinity map classes within the five subregions. The various CA runs were compared and contrasted by visually examining the results for changes in the absolute and relative contributions, fluctuations in the sum of relative contributions and movements in the graphical displays. For consistency, the first seven factors were always retained. The CAs for the various subregions show that the seven retained factors explain at least 96% of the total variation in the data. The geographical co-ordinates of the lakes were used as supplementary variables throughout the analyses.

Preliminary CAs were applied to each of the original data sets. If a data set contained lake samples exhibiting unusually high (or low) concentrations of an analyte, the samples were deleted and CA was performed again. The results were compared and if significant differences were observed the lakes were classified as 'sample outliers' and excluded in subsequent CA runs.

Analytes, which were not primarily responsible for the composition of the first seven factors but which were well represented by these factors, were made supplementary one at a time. Additional CAs were performed with each variable projected onto the new space generated by the retained factors. If the results did not change, then the information lost by deleting the variables was not crucial and the analytes were classified as 'superfluous' variables. Some analytes were poorly represented even by the seven retained factors, possibly indicating behavior different from the other analytes. Analytes with a sum of relative contributions from the first seven factors less than 50 were identified as 'unresolved' variables.

The sample and variable points were projected onto the factorial plane composed of the two primary factorial axes where at least 65% of the information was reconstructed in all cases. A visual inspection of the plots provided the following information: proximity between lake samples is due to similar behavior with respect to analytes; proximity between variables is due to correlation between analytes; lake samples rich in one analyte will be displayed close to the point representing that variable. The graphical displays obtained for the various subregions were compared and examined for clusters of variables. Other two-dimensional factorial planes depicting variables which were poorly represented in the plane of the first two factors were also produced. If the cumulative relative contributions from the two displayed factors were less than 30% the variables were excluded from conclusions.

RESULTS

CAs were performed on all five subregions of the Northeast, but only the results obtained for subregion 1A, the Adirondacks, are presented in this paper. The alkalinity map classes for

subregion 1A are depicted in Figure 1, while Figure 2 shows the location of the sampled sites. A total of 155 lakes were sampled in subregion 1A: 57 in alkalinity map class 1A1, 51 in 1A2 and 47 in 1A3. Table 1 lists the variables identified as unresolved and superfluous in each of the three alkalinity map classes. The CA results are presented in Tables 2–10 and the main results for each alkalinity map class are summarized below.

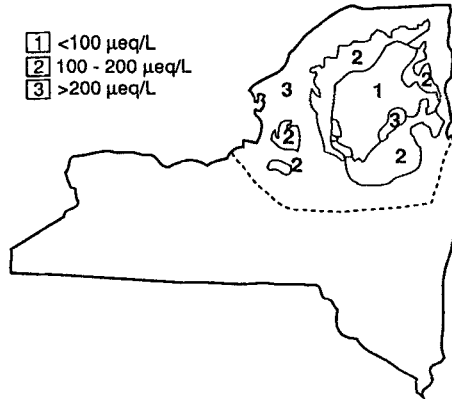


Figure 1. Alkalinity map classes in subregion 1A; ----- subregion boundaries

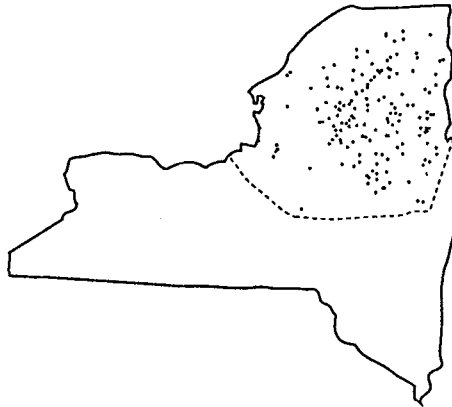


Figure 2. Lake sample sites in subregion 1A; ----- subregion boundaries

Table 1. 'Superfluous' and 'Unresolved' variables

Alkalinity map class	Superfluous	Unresolved
1A1	pH, DIC, DOC, Mg, NO ₃ , F	K, NH ₄ , P, SiO ₂
1A2	pH, DIC, DOC, P	K, NH ₄ , NO ₃ , F, SiO ₂
1A3	pH, DIC, DOC, Na, K, NO ₃ , F, SiO ₂	NH ₄ , P

Table 2. Decomposition of variation in terms of factors (1A1)

Eigenvalue	Per cent variation	Cumulative variation
0.3247073	58.77	58.77
0.0951364	17.22	75.99
0.0619817	11.22	87.21
0.0299951	5.43	92.64
0.0156359	2.83	95.47
0.0093812	1.70	97.17
0.0051361	0.93	98.09
0.0030727	0.56	98.65
0.0025960	0.47	99.12
0.0017104	0.31	99.43
0.0011057	0.20	99.63
0.0007215	0.13	99.76
0.0004944	0.09	99.85
0.0003131	0.06	99.91
0.0002834	0.05	99.96
0.0001592	0.03	99.99
0.0000400	0.01	99.99
0.0000289	0.01	100.00

Table 3. Absolute and relative contributions with respect to the set of retained factors (1A1)

Analyte	Weight	C ₁ ^A	C ₁ ^R	C ₂ ^A	C ₂ ^R	C ₃ ^A	C ₃ ^R	C ₄ ^A	C ₄ ^R	C ₅ ^A	C ₅ ^R	C ₆ ^A	C ₆ ^R	C ₇ ^A	C ₇ ^R
Fe	0.08470	3.9	13.4	84.3	86.1	0.3	0.2	0.6	0.2	0.0	0.0	0.3	0.0	0.8	0.0
pH	0.00700	0.0	2.3	0.0	2.5	0.1	13.9	1.0	48.0	0.2	4.8	0.9	12.9	0.0	0.0
ANC	0.15709	9.2	88.0	0.2	0.6	4.4	8.0	1.8	1.6	2.6	1.2	0.4	0.1	0.0	0.0
DIC	0.00157	0.1	62.5	0.0	8.6	0.0	4.9	0.1	7.2	0.0	1.8	0.0	1.1	0.0	0.1
DOC	0.00461	0.0	0.3	0.3	34.7	0.0	0.8	0.1	3.8	0.0	0.0	1.5	15.6	0.2	1.3
Al	0.17020	56.3	92.7	10.1	4.9	0.3	0.1	15.0	2.3	0.1	0.0	0.0	0.0	0.1	0.0
Ca	0.15637	5.3	73.5	0.6	2.3	2.8	7.4	2.2	2.8	12.4	8.3	1.2	0.5	4.9	1.1
Mg	0.05546	2.1	64.1	0.0	0.0	1.3	7.5	0.0	0.0	5.8	8.4	0.7	0.6	5.1	2.4
Na	0.05675	2.9	25.6	0.5	1.3	30.8	52.5	0.2	0.2	45.1	19.4	2.5	0.6	0.7	0.1
K	0.00910	0.0	2.4	0.1	3.8	0.0	0.1	2.2	37.1	0.1	0.6	0.1	0.4	0.0	0.0
NH ₄	0.00228	0.1	24.7	0.0	0.7	0.0	0.7	0.6	12.4	0.1	1.0	0.0	0.0	0.0	0.0
SO ₄	0.13654	0.5	11.1	2.2	13.7	2.2	8.6	27.4	53.1	0.0	0.0	17.4	10.6	0.5	0.2
HCO ₃	0.09170	13.1	80.4	0.3	0.6	2.9	3.4	23.7	13.4	1.4	0.4	7.5	1.3	0.3	0.0
Cl	0.02956	2.6	17.9	0.3	0.7	54.5	71.5	0.1	0.1	27.7	9.2	2.5	0.5	0.7	0.1
NO ₃	0.00316	0.5	22.0	0.7	8.7	0.0	0.2	2.2	8.8	0.1	0.2	7.5	9.5	61.7	42.6
F	0.00337	0.0	2.6	0.0	2.6	0.0	0.4	1.6	46.3	0.0	0.0	0.1	1.1	0.0	0.0
P	0.00425	0.1	5.8	0.2	4.8	0.0	0.5	0.1	0.8	1.1	4.8	1.6	4.4	13.8	20.4
SiO ₂	0.00332	0.0	1.6	0.0	1.4	0.1	6.7	0.0	0.1	0.0	0.3	0.0	0.3	3.3	12.6
Mn	0.02295	3.4	46.2	0.1	0.3	0.3	0.7	21.1	26.5	3.3	2.2	55.8	21.9	7.8	1.7

Table 4. Analysis of error by the set of retained factors (1A1)

Analyte	Norm of the error of reconstruction	Sum of relative contributions
Fe	0·0045	100·0
pH	0·0099	84·5
ANC	0·0139	99·4
DIC	0·0070	86·2
DOC	0·0196	56·5
Al	0·0032	100·0
Ca	0·0309	95·9
Mg	0·0426	83·1
Na	0·0101	99·7
K	0·0317	44·5
NH ₄	0·0286	39·6
SO ₄	0·0206	97·3
HCO ₃	0·0163	99·5
Cl	0·0812	99·9
NO ₃	0·0247	91·8
F	0·0223	53·1
P	0·0450	41·5
SiO ₂	0·0322	22·9
Mn	0·0122	99·4

Table 5. Decomposition of variation in terms of factors (1A2)

Eigenvalue	Per cent variation	Cumulative variation
0·3071259	51·04	51·04
0·152402	25·33	76·37
0·0756435	12·57	88·95
0·0250608	4·17	93·11
0·0100305	1·67	94·78
0·0097901	1·63	96·41
0·0058289	0·97	97·37
0·0044515	0·74	98·11
0·0037761	0·63	98·78
0·0029875	0·50	99·24
0·0011749	0·20	99·43
0·0011379	0·19	99·62
0·00009674	0·16	99·78
0·00006322	0·11	99·89
0·00004482	0·07	99·96
0·00001566	0·03	99·99
0·00000444	0·01	100·00
0·00000230	0·00	100·00

Table 6. Absolute and relative contributions with respect to the set of retained factors (1A2)

Analyte	Weight	C_1^A	C_1^R	C_2^A	C_2^R	C_3^A	C_3^R	C_4^A	C_4^R	C_5^A	C_5^R	C_6^A	C_6^R	C_7^A	C_7^R
Fe	0.08980	13.7	42.9	0.0	0.0	73.7	56.8	0.0	0.0	0.3	0.0	1.5	0.1	0.2	0.0
pH	0.00618	0.0	14.0	0.1	11.6	0.1	8.7	1.4	38.8	1.2	12.5	0.0	0.2	0.0	0.0
ANC	0.17000	3.9	42.4	9.0	49.3	0.1	0.4	3.2	2.9	7.2	2.6	2.6	0.9	0.0	0.0
DIC	0.00188	0.0	21.3	0.1	48.3	0.0	2.4	0.1	12.0	0.0	0.3	0.1	4.1	0.0	0.3
DOC	0.00466	0.2	33.9	0.0	4.3	0.1	6.1	0.1	1.4	0.1	0.8	0.1	0.4	0.0	0.0
Al	0.11799	43.0	81.6	8.1	7.6	16.7	7.8	19.1	2.9	0.3	0.0	0.2	0.0	0.0	0.0
Ca	0.16937	3.3	45.3	4.5	30.3	0.7	2.4	1.2	1.3	24.8	11.0	3.7	1.6	7.8	2.0
Mg	0.06100	0.7	24.9	0.8	14.3	0.0	0.0	1.1	3.2	20.5	23.2	0.1	0.1	30.4	20.0
Na	0.06553	6.5	36.5	20.5	57.1	0.1	0.1	2.0	0.9	12.6	2.3	0.2	0.0	8.8	0.9
K	0.00843	0.0	2.8	0.0	3.9	0.0	1.0	1.4	19.3	1.9	10.5	0.1	0.5	1.1	3.4
NH ₄	0.00208	0.1	18.5	0.0	1.6	0.0	0.0	0.0	0.5	0.1	0.6	1.1	6.0	0.9	2.9
SO ₄	0.11243	1.6	21.6	0.6	3.7	7.5	25.1	34.2	37.8	9.5	4.2	8.1	3.5	6.5	1.7
HCO ₃	0.10776	8.1	55.7	9.4	32.0	0.0	0.0	18.9	10.5	1.9	0.4	1.2	0.3	0.7	0.1
Cl	0.04901	14.6	38.3	46.1	59.9	0.8	0.5	1.0	0.2	2.6	0.2	0.2	0.0	5.0	0.2
NO ₃	0.00178	0.2	15.7	0.0	0.4	0.0	0.3	0.5	2.7	1.0	2.3	2.4	5.5	3.9	5.3
F	0.00317	0.0	10.4	0.0	3.3	0.0	0.0	0.9	15.7	0.9	6.2	0.6	3.9	0.6	2.3
P	0.00840	0.8	22.4	0.0	0.4	0.0	0.2	2.1	5.0	2.4	2.4	57.9	54.8	9.4	5.3
SiO ₂	0.00256	0.1	15.8	0.0	2.8	0.0	1.0	0.0	0.3	0.1	0.8	0.0	0.2	1.2	5.7
Mn	0.01796	3.1	49.6	0.5	4.2	0.0	0.1	12.8	17.0	12.6	6.7	20.0	10.4	23.5	7.2

Table 7. Analysis of error by the set of retained factors (1A2)

Analyte	Norm of the error of reconstruction	Sum of relative contributions
Fe	0.0057	100.0
pH	0.0114	86.0
ANC	0.0212	98.4
DIC	0.0059	88.8
DOC	0.0273	47.0
Al	0.0054	100.0
Ca	0.0369	94.0
Mg	0.0353	85.9
Na	0.0342	97.9
K	0.0330	41.3
NH ₄	0.0356	30.0
SO ₄	0.0231	97.7
HCO ₃	0.0215	99.0
Cl	0.0266	99.4
NO ₃	0.0540	32.2
F	0.0285	41.9
P	0.0315	90.4
SiO ₂	0.0301	26.7
Mn	0.3004	95.1

Table 8. Decomposition of variation in terms of factors (1A3)

Eigenvalue	Per cent variation	Cumulative variation
0.1669996	49.77	49.77
0.0673210	20.06	69.83
0.0403326	12.02	81.85
0.0208755	6.22	88.07
0.0140377	4.18	92.26
0.0107794	3.21	95.47
0.0049781	1.48	96.95
0.0040477	1.21	98.16
0.0020672	0.62	98.77
0.0010112	0.30	99.07
0.0009415	0.28	99.36
0.0007394	0.22	99.58
0.0005143	0.15	99.73
0.0003980	0.12	99.85
0.0002297	0.07	99.92
0.0002014	0.06	99.98
0.0000575	0.02	99.99
0.0000222	0.01	100.00

Table 9. Absolute and relative contributions with respect to the set of retained factors (1A3)

Analyte	Weight	C_1^A	C_1^R	C_2^A	C_2^R	C_3^A	C_3^R	C_4^A	C_4^R	C_5^A	C_5^R	C_6^A	C_6^R	C_7^A	C_7^R
Fe	0.02345	13.9	48.9	1.5	2.2	45.9	39.0	3.0	1.3	24.1	7.1	6.4	1.4	0.1	0.0
pH	0.00307	0.7	56.5	0.0	0.0	0.6	11.7	2.0	20.5	0.3	2.2	0.0	0.0	0.0	0.1
ANC	0.25844	4.1	67.1	4.3	28.6	0.2	0.7	0.0	0.1	0.0	0.0	0.2	0.2	0.1	0.1
DIC	0.00290	0.0	38.8	0.1	31.4	0.0	0.7	0.0	5.4	0.0	3.1	0.0	0.5	0.0	1.1
DOC	0.00198	0.5	53.7	0.0	0.0	0.6	14.5	0.4	5.1	0.0	0.3	0.2	1.1	1.4	4.1
Al	0.03295	46.8	81.9	0.4	0.3	40.0	16.9	0.2	0.0	3.2	0.5	1.9	0.2	3.5	0.2
Ca	0.22199	1.7	42.4	1.7	16.7	0.1	0.6	2.4	7.4	5.3	11.1	9.5	15.4	0.3	0.2
Mg	0.06978	0.1	1.5	0.1	0.4	0.5	1.8	2.1	4.3	28.3	38.8	49.6	52.3	0.4	0.2
Na	0.04648	0.1	0.4	31.4	94.2	0.1	0.1	0.1	0.1	0.4	0.2	0.0	0.0	0.0	0.0
K	0.00531	0.3	30.7	0.0	0.6	0.2	5.0	0.8	8.7	0.2	1.7	1.0	5.6	1.2	3.1
NH ₄	0.00104	0.2	19.3	0.0	0.2	0.1	3.2	0.0	0.2	0.0	0.1	0.3	2.1	0.3	1.1
SO ₄	0.06176	11.6	58.0	1.4	2.7	4.7	5.7	42.0	26.1	14.4	6.0	3.0	1.0	1.0	0.1
HCO ₃	0.21331	8.4	75.8	4.8	17.6	1.2	2.7	1.7	1.9	0.5	0.4	0.1	0.1	0.0	0.0
Cl	0.04170	0.7	2.8	53.3	92.6	0.3	0.3	3.6	1.9	2.4	0.9	0.0	0.0	0.2	0.0
NO ₃	0.00120	1.3	26.8	0.5	3.9	1.5	7.5	3.7	9.3	0.4	0.7	0.1	0.1	79.3	47.8
F	0.00133	0.7	58.2	0.0	0.1	0.1	1.7	1.3	14.5	0.0	0.1	0.0	0.0	1.9	4.9
P	0.00321	0.1	3.3	0.2	2.6	0.8	7.3	0.1	0.6	0.9	2.8	1.1	2.8	9.1	10.3
SiO ₂	0.00122	0.6	49.4	0.1	1.9	0.0	0.5	0.0	0.2	0.7	4.7	0.0	0.1	0.2	0.5
Mn	0.00886	8.2	47.9	0.4	0.9	3.1	4.3	36.6	26.8	18.9	9.3	26.5	10.0	0.9	0.2

Table 10. Analysis of error by the set of retained factors (1A3)

Analyte	Norm of the error of reconstruction	Sum of relative contributions
Fe	0.0059	99.9
pH	0.0135	90.9
ANC	0.0180	96.8
DIC	0.0059	80.8
DOC	0.0188	78.9
Al	0.0029	100.0
Ca	0.0204	93.7
Mg	0.0920	99.2
Na	0.0328	95.2
K	0.0290	55.4
NH ₄	0.0316	26.1
SO ₄	0.0111	99.6
HCO ₃	0.0168	98.5
Cl	0.0238	98.5
NO ₃	0.0180	96.1
F	0.0198	79.5
P	0.0556	29.7
SiO ₂	0.0305	57.2
Mn	0.0127	99.4

Alkalinity map class 1A1

Two lakes were identified as sample outliers: sample 1A1-019, Quiver Lake in NY, which exhibited very high NH₄ and Fe concentrations, and sample 1A1-064, Mt. Arab Lake in NY, which had a high concentration of P. When these samples were deleted from the data set, the coefficient of variation was reduced from 1.54 to 1.14 for Fe, from 1.50 to 0.86 for P and from 1.08 to 0.66 for NH₄. The composition of factors 5, 6 and 7 changed.

Factor 1, which accounts for more than half (59%) of the variation in the data, is composed mainly of Al (56%) with smaller contributions from HCO₃ (13%) and ANC (9%). The second and third factors explain 17% and 11% of the variability respectively. Fe (84%) is almost entirely responsible for factor 2 with a small contribution from Al (10%); Cl (55%) and Na (31%) make up factor 3. The fourth factor is responsible for only 5% of the variation and separates the samples rich in HCO₃ and Al from those rich in SO₄ and Mn. The absolute contributions for these variables are 24%, 16%, 27% and 21% respectively.

The two-dimensional factorial planes providing the most information about the relationships among the analytes are shown in Figures 3–5. Surprisingly, Na and Cl were poorly reconstructed by the first two factors but were well represented in the factor 1 versus factor 3 plane (Figure 3) and clearly form a cluster. A cluster consisting of ANC, DIC, Ca, Mg and HCO₃ is also evident in this plane. pH and SO₄ are grouped together in the factor 1 versus factor 4 plane (Figure 4), which also shows that Al and Mn do not form a cluster as suggested in Figure 3. Fe and DOC, which were poorly represented in the aforementioned planes, shows up in the plane composed of factors 2 and 4 (Figure 5) but are not part of any clusters. F is also represented well but apparently not grouped with other analytes.

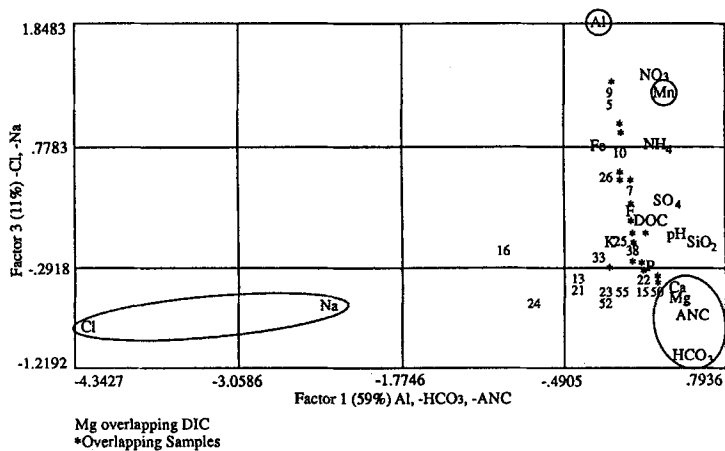


Figure 3. Sample co-ordinates and variable factors in the factor 1–factor 3 plane (1A1)

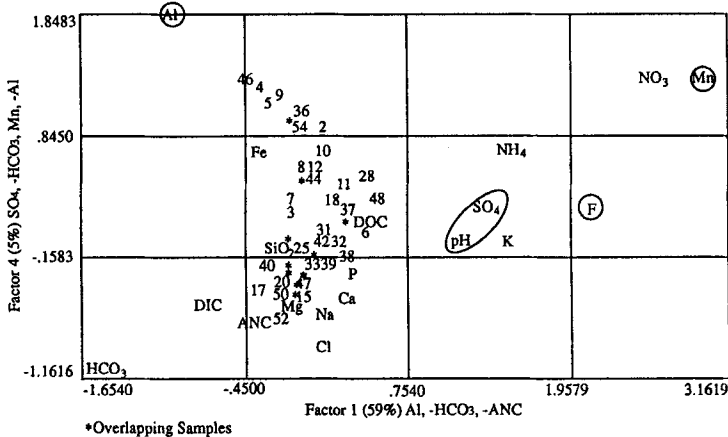


Figure 4. Sample co-ordinates and variable factors in the factor 1–factor 4 plane (1A1)

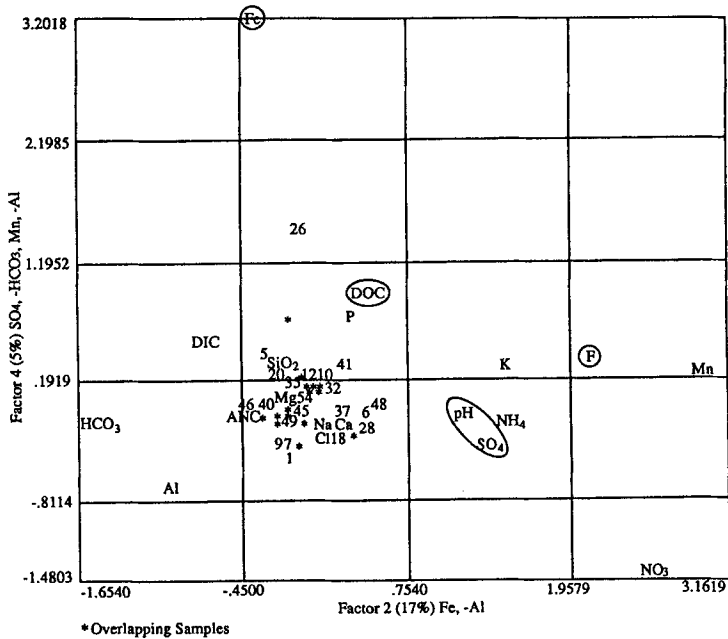


Figure 5. Sample co-ordinates and variable factors in the factor 1–factor 4 plane (1A1)

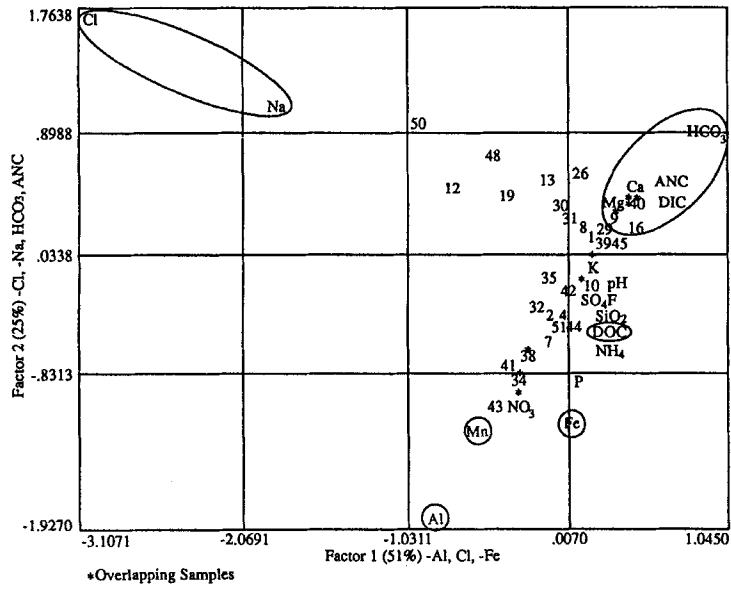


Figure 6. Sample co-ordinates and variable factors in the factor 1-factor 2 plane (1A2)

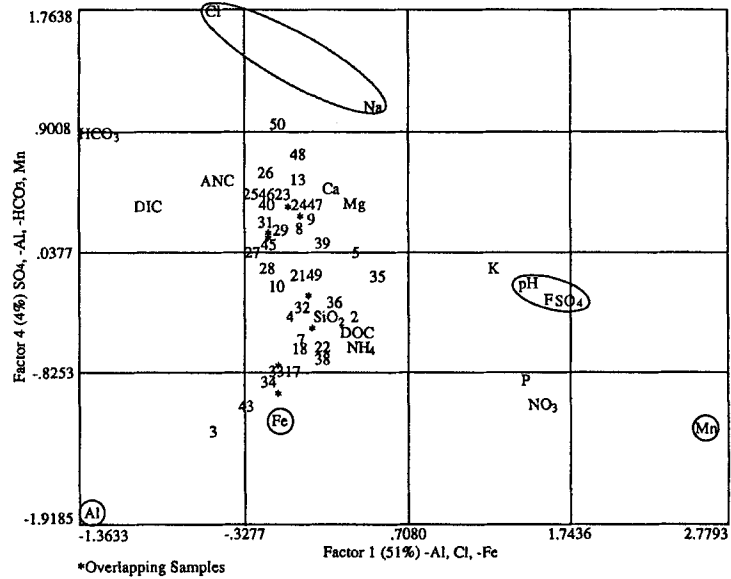


Figure 7. Sample co-ordinates and variable factors in the factor 1-factor 4 plane (1A2)

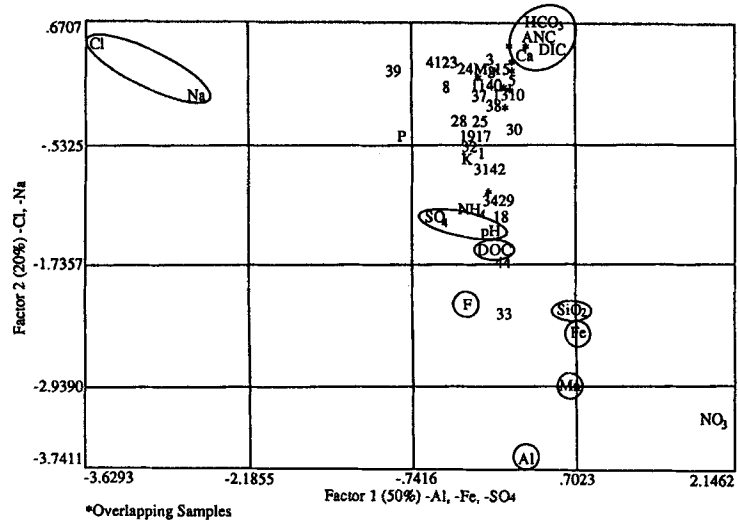


Figure 8. Sample co-ordinates and variable factors in the factor 1-factor 2 plane (1A3)

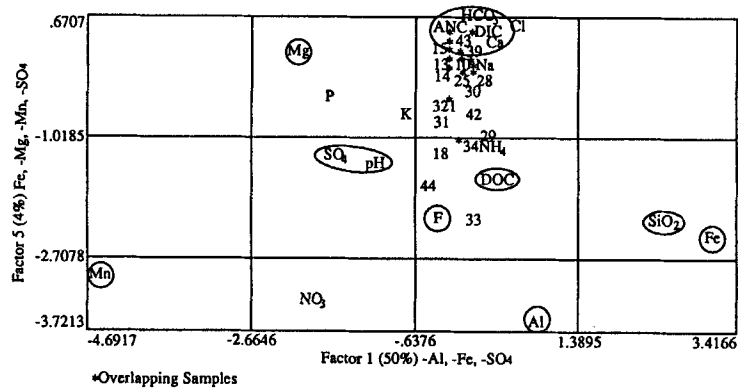


Figure 9. Sample co-ordinates and variable factors in the factor 1-factor 5 plane (1A3)

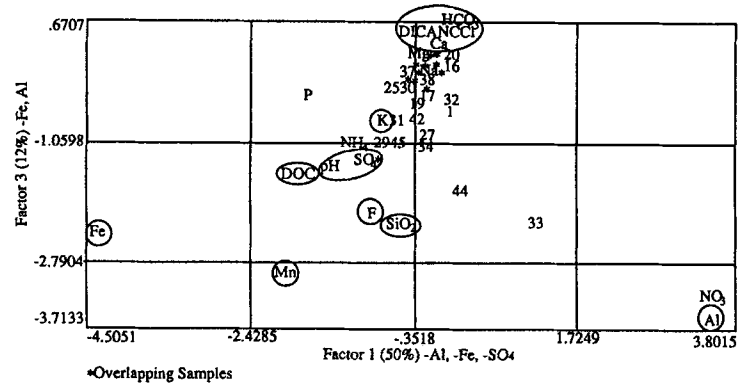


Figure 10. Sample co-ordinates and variable factors in the factor 1-factor 3 plane (1A3)

Alkalinity map class 1A2

No sample outliers were identified in this alkalinity map class. The first two factors account for 76% of the overall variation in the data. Al (43%) is the major contributor to factor 1, but Cl (15%) and Fe (14%) also contribute some. Factor 2 is composed of Cl (46%) and Na (21%) with smaller contributions from HCO₃ (9%) and ANC (9%). Fe, with 74% absolute contribution, is almost solely responsible for the third factor which explains 13% of the variability. Factor 4 accounts for only 4% variation and separates samples rich in SO₄ and Mn from those rich in Al and HCO₃ with absolute contributions of 34%, 13%, 19% and 19% respectively.

It is interesting that P, which was identified as unresolved in the analysis of alkalinity map classes 1A1 and 1A3, was very well represented in this analysis. NO₃ and F, on the other hand, were poorly reconstructed and were classified as unresolved variables.

Figure 6 shows the plot of the first two factorial axes where two main clusters are seen: (Na and Cl) and (ANC, DIC, Ca, Mg and HCO₃). As in alkalinity map class 1A1, Fe and DOC are well represented but are not clustered with other analytes. The factor 1 versus factor 4 plane displayed in Figure 7 shows pH and SO₄ grouped together. Note that Al and Mn are depicted in opposite corners of the plane and hence do not form a cluster.

Alkalinity map class 1A3

Sample 1A3-039, Woodman Pond in NY, which had a very high concentration of Mg, was classified as a sample outlier. With the sample deleted the coefficient of variation for Mg decreased from 1.25 to 0.95 and the compositions of factors 5 and 6 were changed.

Factor 1, which is responsible for 50% variation, is composed of Al (47%), Fe (14%) and SO₄ (12%). Factors 2 and 3 have essentially the same compositions as those obtained for factors 2 and 3 in the analysis of alkalinity map class 1A2 and explain 20% and 12% of the variability respectively. Factor 4 accounts for 6% variation and is made up of SO₄ (42%) and Mn (37%). The analytes responsible for factor 5 are Fe (24%) opposing Mg (28%), Mn (19%) and SO₄ (14%). Only 4% of the variability is explained by the fifth factor.

It is noteworthy that both K and SiO₂, which were classified as unresolved in the analysis of alkalinity map classes 1A1 and 1A2, had good quality of representation and were identified as superfluous.

Three main clusters are found in the plane of the first two factorial axes (Figure 8). However, Mg was very poorly represented by the two primary factors and could therefore not be grouped together with ANC, DIC, Ca and HCO₃ as it was in the graphical displays for alkalinity map classes 1A1 and 1A2. Mg is well reconstructed in the factor 1 versus factor 5 plane shown in Figure 9 but is clearly not part of the (ANC, DIC, Ca and HCO₃) cluster. This plane also shows that DOC is probably not part of the cluster including pH and SO₄ and that Al and Mn do not form a group. Fe and SiO₂ are displayed quite far apart in the plane composed of factors 1 and 3 (Figure 10), indicating that these two analytes most likely do not constitute a cluster as suggested in the previous two factor planes. As in the analysis of alkalinity map class 1A1, F was well represented but does not appear to be grouped with other analytes.

DISCUSSION

The fact that the reported measurement units are not commensurate was of some concern. Therefore additional CAs were run for each of the three alkalinity map classes after converting

all the concentrations to $\mu\text{g l}^{-1}$. Changes were seen in the composition of the seven retained factors as well as in the sum of the relative contributions for all the analytes. Different graphical displays were also obtained but the same main clusters of variables were evident. Chemists and hydrologists consulted agreed that the differences in the CA results were not critical as long as the main correlational relationships among the analytes are preserved. It was therefore decided to carry out the CA with the data in the reported units.

The effects of adding a constant to ANC in order to offset the negative alkalinity values were also examined by performing additional CAs after increasing all the ANC values by 100 and 200. One of the statistics computed for each variable is the coefficient of variation, which is defined as the ratio of the standard deviation and the mean. Adding a constant increases the mean by the constant but leaves the standard deviation unchanged, thus resulting in a decrease in the coefficient of variation. Table 11 lists the changes in the coefficient of variation for ANC in the three alkalinity map classes. Despite these decreases, the overall composition of the first seven factors remained unchanged since ANC is not one of the major contributors to any of the seven retained factors. However, small changes did occur in the relative contributions from the first two factors (Table 12) as well as in the sum of relative contributions for a few of the variables (Table 13). Since the CA results appear to be relatively insensitive to increasing ANC by different constants, adding a constant to make all the alkalinity values positive seems to be a reasonable approach.

Between alkalinity map classes the composition of the seven retained factors varied quite a bit. However, four factors were common to all three classes: one factor with aluminium as the

Table 11. Changes in the coefficient of variation for ANC

Alkalinity map class	ANC + 50	ANC + 100	ANC + 200
1A1	0.64	0.47	0.31
1A2	0.73	0.57	0.39
1A3	1.08	1.00	0.88

Table 12. Changes in the relative contributions from the first two factors

Alkalinity map class	Analyte	ANC + 50	ANC + 100	ANC + 200
1A1	ANC	89.3	84.5	67.8
	DIC	65.0	63.9	62.6
	HCO ₃	81.4	79.3	76.7
	Ca	77.8	79.2	80.3
1A2	ANC	91.7	87.5	73.1
	DIC	69.6	66.8	62.4
	HCO ₃	87.7	85.4	82.2
1A3	ANC	95.8	91.2	63.2
	HCO ₃	93.3	91.2	88.1
	Ca	46.7	52.0	59.0

Table 13. Changes in the sum of relative contributions

Alkalinity map class	Analyte	ANC + 50	ANC + 100	ANC + 200
1A1	ANC	99.4	98.5	96.7
	pH	86.0	87.4	90.5
	K	43.3	41.8	39.0
	F	52.6	50.7	46.6
1A2	ANC	98.4	97.2	96.1
	pH	86.0	88.3	92.1
	K	41.3	39.1	35.2
	P	90.4	88.0	83.8
1A3	ANC	96.7	92.1	80.9
	DIC	81.2	82.7	85.0
	K	58.9	57.2	53.3
	P	29.6	26.9	22.9

main contributor, a second factor composed almost entirely of iron, a third factor made up of sodium and chloride and a fourth factor showing sulfate and manganese.

The graphical displays of sample co-ordinates and variable factors projected in the plane of the first two factorial axes were also quite different. However, when other two-dimensional factor planes were examined, three main clusters were identified: one including ANC, DIC, Ca, Mg and HCO_3 , a second containing Na and Cl and a third composed of pH and SO_4 . These three clusters show analytes related to alkalinity, salinity and acidity respectively. The graphical displays also show that Fe, DOC, Al and Mn were well represented but appear to be uncorrelated with other analytes.

It is interesting that two of the six primary analytes (pH, ANC, sulfate, calcium, extractable aluminum and DOC) often thought the most relevant to lake acidification processes, namely pH and DOC, were among the variables identified as superfluous. This was not only observed in the three alkalinity map classes of subregion 1A but in all five subregions of the Northeast. However, it should be noted that pH and DOC are necessary to evaluate the potential toxicity of aluminium as well as to evaluate the relative contributions of minerals versus organic acidity in lakewater chemistry. Thus the interpretation of these findings should be confined to the ELS-I data set since other data sets may not provide the same outcome if subjected to a CA.

Although calcium, magnesium and bicarbonate were part of the alkalinity cluster, these analytes could not be made supplementary without changing the outcome of the CA. Similarly, with the exception of alkalinity map class 1A3, sodium, which was grouped with chloride, could not be eliminated without affecting the CA results.

It is noteworthy that two of the analytes which were not well represented by the seven retained factors, namely ammonium and phosphorous, are related nutrients. The different behavior of these as well as the other variables classified as unresolved should be further investigated. Greater insight might be provided by retaining more factors. With more factors, more analytes such as potassium, phosphorus, fluoride and silica would probably become superfluous, thus further reducing the number of analytes needed in subsequent data analyses.

NOTICE

Although the research described in this article has been funded wholly or in part by the U.S.

Environmental Protection Agency, it has not been subjected to Agency review and therefore does not reflect the view of the Agency and no official endorsement should be inferred.

REFERENCES

1. M. J. Greenacre, *Theory and Application of Correspondence Analysis*, Academic Press, London (1984).
2. L. Lebart, A. Morineau and K. Warwick, *Multivariate Descriptive Statistical Analysis, Correspondence Analysis and Related Techniques for Large Matrices*, Wiley, New York (1984).
3. M. David, M. Dagbert and Y. Beauchemin, 'Statistical analysis in geology: correspondence analysis method', *Q. Colorado School Mines*, **72**, 1 (1977).
4. P. Kanciruk, R. A. McCord, L. A. Hook and M. J. Gentry, *National Surface Water Survey, Eastern Lake Survey – Phase I, Data Base Report, EPA600/4-88/032*, U.S. Environmental Protection Agency, Las Vegas, NV (1988).
5. W. S. Overton, *National Surface Water Survey, Eastern Lake Survey – Phase I, Data Analysis Plan*, U.S. Environmental Protection Agency, Corvallis, OR (1986).
6. M. D. Best, L. M. Creelman, S. K. Drouse and D. J. Chaloud, *National Surface Water Survey, Eastern Lake Survey – Phase I, Quality Assurance Report, EPA600/4-86/011*, U.S. Environmental Protection Agency, Las Vegas, NV (1986).
7. R. A. Linthurst, D. H. Landers, J. M. Eilers, D. F. Brakke, W. S. Overton, E. P. Meier and R. E. Crowe, *Eastern Lake Survey – Phase I. Characteristics of Lakes in the Eastern United States. Volume I: Population Descriptions and Physico-Chemical Relationships EPA600/4-86/007a*, U.S. Environmental Protection Agency, Las Vegas, NV (1986).
8. W. S. Overton, P. Kanciruk, L. A. Hook, J. M. Eilers, D. H. Landers, D. J. Blick Jr., D. F. Brakke, R. A. Linthurst and M. D. DeHaan, *Eastern Lake Survey – Phase I. Characteristics of Lakes in the Eastern United States. Volume II: Lakes Sampled and Descriptive Statistics for Physical and Chemical Variables, EPA600/4-86/007b*, U.S. Environmental Protection Agency, Washington, DC (1986).