GROUPING OF LAKES ON THE BASIS OF WATER QUALITY VARIABLES. AN APPLICATION TO NOVA SCOTIA AND NEWFOUNDLAND LAKES

S.R. Esterby[1], A.H. El-Shaarawi[1],
G.D. Howell[2] and T.A. Clair[2]

NWRI Contribution No. 89-24

[1] Lakes Research Branch
National Water Research Institute
Canada Centre for Inland Waters
Burlington, Ontario
Canada L7R 4A6

[2] Water Quality Branch
Atlantic Region
Moncton, New Brunswick

May 1989

ABSTRACT

Many lakes in eastern Canada are sensitive to long range transport of atmospheric pollutants because of their low buffering capacity. Thus, it is important to assess long term changes in water quality. Due to the large number of lakes, a method was needed for choosing a subset of lakes to monitor regularly. Preliminary surveys were conducted in Nova Scotia and Newfoundland to determine the water quality characteristics of lakes in the region. This paper describes the division of the set of lakes into groups, with lakes in a group having similar water quality parameter values, by means of cluster analysis and principal component ordination. The characteristics of the groups are shown by graphical procedures and summary statistics, and this characterization is used to both determine the number of groups and describe the final choice. The membership of the groups was subsequently assessed in terms of the influences of terrestrial weathering, marine aerosols, anthropogenically derived mineral acids and natural organic acids.

RÉSUMÉ

Beaucoup de lacs dans l'est du Canada sont sensibles au transport à longue distance de polluants atmosphériques, en raison de leur faible capacité tampon. Il est donc important d'évaluer les variations à long terme de la qualité de l'eau. Étant donné le grand nombre de lacs, il a fallu trouver une méthode pour choisir un sous-ensemble de lacs devant faire l'objet d'une surveillance régulière. Des études préalables ont été effectuées en Nouvelle-Écosse et à Terre-Neuve pour déterminer les caractéristiques qualitatives de l'eau des lacs dans la région. Le présent document décrit la division de l'ensemble des lacs en groupes grâce à l'analyse typologique et à l'ordination selon les composantes principales, les lacs d'un même groupe présentant des valeurs semblables pour les paramètres qualitatifs de l'eau. Les caractéristiques des groupes sont révélées à l'aide de méthodes graphiques et de statistiques sommaires, cette caractérisation étant utilisée à la fois pour déterminer le nombre de groupes et décrire le choix final. La composition des groupes a été ultérieurement évaluée en fonction des effets de l'érosion terrestre, des aérosols marins, des acides minéraux d'origine anthropogène et des acides organiques naturels.

MANAGEMENT PERSPECTIVE

In the monitoring of water quality to assess temporal responses to the long range transport of air pollutants, it is important to sample lakes of varying characteristics since different types of lakes would be expected to show different responses. To meet economical restraints, a limited number of lakes must be chosen for sampling over time. These requirements can be satisfied by dividing the population of lakes into subsets or strata, with the lakes in each subset similar to each other, and then choosing a small number of lakes from each subset. A method for determining the subsets is applied to two sets of water quality data from headwater lakes of Eastern Canada. The report illustrates the importance of examining the characteristics of the subsets determined by clustering and ordination methods in both choosing the number of subsets and understanding the characteristics of the final choice. The statistical methods used were highly successful in determining subsets that could be explained in terms of geological and hydrological characteristics.

PERSPECTIVES DE GESTION

Pour la surveillance de la qualité de l'eau visant à évaluer les réponses dans le temps au transport à longue distance de polluants atmosphériques, il est important d'échantillonner des lacs présentant des caractéristiques variables, car différents types de lacs donneront probablement des réponses différentes. Pour des raisons de coût, seul un nombre limité de lacs pourra être échantillonné dans le temps. Il faudra diviser la population de lacs en sous-ensembles ou strates, les lacs de chaque sous-ensemble étant semblables, puis choisir un petit nombre de lacs dans chaque sous-ensemble. Une méthode de caractérisation des sous-ensembles est appliquée à deux ensembles de données qualitatives de l'eau, provenant de lacs d'amont de l'est du Canada. Le rapport montre l'importance de l'examen des caractéristiques des sous-ensembles obtenus par analyse typologique et ordination à la fois pour le choix du nombre de sous-ensembles et pour la justification des caractéristiques du choix final. Les méthodes statistiques employées étaient très efficaces et ont permis de déterminer des sous-ensembles qui pouvaient se justifier en termes de caractéristiques géologiques et hydrologiques.

INTRODUCTION

The assessment of the potential long term changes in water quality due to the long range transport of air pollutants (LRTAP) has been of major interest in eastern Canada because of the high sensitivity of water bodies located on nonbuffering bedrocks with thin podzolic soils. In order to monitor temporal responses in surface water quality, it was necessary to design a monitoring network. Headwater lakes were considered to be the most suitable type of water body because of minimal human disturbance and because of quick reflection of changes in water quality resulting from the small basin sizes and the rapid flushing rates. The large number of these lakes, makes it economically impossible to monitor each lake on a regular basis, and hence, it was decided to choose a small number of representative lakes for monitoring. Preliminary surveys of the Nova Scotia and Newfoundland headwater lakes were conducted to determine the basic attributes of the target population, and to stratify the population into a set of homogeneous subpopulations. To reduce the effects of the within lake spatial variability, the measurements were taken at autumnal and vernal overturn, when the lakes are well mixed, and hence, the within lake variability is minimal. The purpose of this paper is to document the rationale and statistical techniques used as a basis for selecting the monitoring lakes. Multivariate statistical techniques (clustering and principal components) are used to describe the attributes of the target population in Nova Scotia and

Newfoundland, and to define clusters of lakes of similar attributes. From each cluster a predetermined number of lakes were chosen at random for future monitoring.

METHODS

Lake Surveys

The data were collected from southwestern Nova Scotia and insular Newfoundland during independent surveys of the population of lakes by Environment Canada and the Department of Fisheries and Oceans. The Newfoundland samples were collected from 152 lakes during the late summer and fall of 1980, while the 89 Nova Scotian lakes were sampled during the summer and fall of 1981. The water quality variables were measured using standard analytical methods (Environment Canada, 1979) and the variables used in the present analyses are shown in Tables 1 and 5 for Nova Scotia and Newfoundland respectively. Water colour was not measured in the samples from lakes 63 to 89 in Nova Scotia. For the 1980 Newfoundland lake survey, sulphate concentrations were determined using the methyl thymol blue method whereas the 1981 Nova Scotia lakes survey had sulphate concentrations determined by ion chromatography. The Nova Scotia sulphate and calcium concentrations and the Newfoundland sulphate concentrations are sea-salt corrected values.

The Nova Scotia lakes are concentrated in southwestern Nova Scotia, with a large number of sites situated in Kejimkujik National

Park and in the vicinity of Yarmouth (Figure 1). The Newfoundland lakes are widely dispersed, and, with the exception of the Avalon Peninsula, cover the entire island (Figure 2). The identification numbers used for the lakes in the tables have been plotted treating longitude and latitude as rectilinear coordinates, as have the cluster memberships later (Figures 4 and 12). To relate this to map locations, undistorted plots, without lake numbers, have been inset in Figures 1 and 2.

Statistical Methods

The observations can be represented in a matrix $X = (x_{ij})$ where $i=1,2,\ldots,n$, $j=1,2,\ldots m$ and x_{ij} is the value of the jth variable measured on the ith lake. Correlation coefficients were calculated for all pairs of variables using the n lakes. From this the correlation matrix was formed and it was used in the principal component analysis to reduce the number of variables. A principal component is given by the linear combination of the original variables and an eigenvector of the correlation matrix. The number of variables is reduced by using only those eigenvectors corresponding to large eigenvalues (Morrison, 1967). To obtain groupings of the lakes graphically, the component scores of the n lakes for the first three principal components were plotted and groups were formed by inspection.

Grouping of the lakes by a numerical procedure using the original variables was done by means of cluster analysis. When the units of the variables were of different magnitudes, the variables were first standardized by subtracting the mean for that variable and dividing by its standard deviation. A non-hierarchical nearest-centroid clustering procedure (Anderberg, 1973) was used. Initial seed-points were determined as points equally spaced within the range of values for each variable. For a specified number of clusters, on the first pass, the procedure assigns each lake to the cluster to which it is nearest in terms of the Euclidean distance between the variables values for that lake and the vector of cluster centroids. The cluster centroids are then recalculated as the means of each variable for all lakes in a cluster. The process of reassignment and updating of centroids is continued until there is no change in cluster-membership on two successive iterations. The clustering was performed for the number of clusters equal to 2 to 7 or 8. For each number of clusters, the proportion of the total variation for all variables accounted for by the clusters was calculated (Hartigan, 1975) as a ratio of sums of squares.

The geographic distribution of the clusters and the characteristics of the clusters in terms of the variables were examined. The number of the cluster to which a lake was assigned was plotted at the position of the lake on a plot with longitude and latitude treated as rectilinear coordinates. The mean for each variable in each cluster was plotted against cluster number for the varying numbers of clusters. Frequency distributions of the variables were also obtained and given as histograms for larger clusters and a simple plot showing the number of lakes in an interval of variable values for small clusters. These three forms of characterization, plus the percentage of explained variation were used to determine the nature of the groups obtained for different numbers of clusters.

RESULTS AND DISCUSSION

Nova Scotia

The correlation coefficients between the 5 variables which were measured in all 89 Nova Scotia lakes are very similar for the 89 lakes and for the subset of 62 lakes for which 7 variables were measured (Table 1). Alkalinity, calcium and chloride are highly correlated ($r \geq 0.89$) and, for the 62 lakes, conductivity is also highly correlated with these three. Colour shows negative or no correlation with all the other variables. This result is expected as the sources of the ions represented by the first 4 variables are terrestrial weathering or marine aerosols and thus their concentrations are largely a function of hydrologic regime. Water colour is a consequence of the processes controlling the production and delivery of organic matter from the bogs to the surface waters, which are different from chemical weathering and marine aerosol transport.

The results of the principal components analysis further reflect this situation. The first three principal components account for over ninety percent of the variance and the first principal component is dominated by the weathering and marine aerosol variables for both the

89 and the 62 lakes (Table 2). The most important variables for the second principal component are sulphate and pH for all 89 lakes and water colour, pH and sulphate for the subset of 62 lakes. The difference in signs observed for pH and sulphate is consistent with mineral acidification. Similarly, the magnitude of the coefficients for colour suggests that natural organic acids are also involved in the acidification of these lakes. Thus, component two could be thought of as the acidification component with both natural organic acids and anthropogenically derived mineral acids (LRTAP) contributing to the overall acidification. The third components for both the 89 lake and the 62 lake analyses are dominated by pH and sulphate. However, as the signs of the elements of the eigenvector are the same for both variables, these two components probably reflect non-acid sulphate loading, either through terrestrial weathering or atmospheric deposition, rather than a mineral acidification process.

The groups obtained by inspection of the plots of the scores for the lakes on the first and second and the first and the third principal components (Figure 3) are reasonably consistent. The most noticeable feature is the large group which corresponds to high scores on the first principal component and thus low concentrations of the highly correlated variables alkalinity, calcium, chloride, and for the subset of 62 lakes, conductivity. This grouping of lakes represents those sites which are furthest from the coast and thus are only marginally influenced by marine aerosols. Based on the 62 lakes this group is compact when the first and third principal components are plotted, but elongated for the first and second components (Figures 1c and d), showing that there is more variability associated with the mineral and organic acidification variables than the non-acid sulphate.

The cluster analysis was also performed on all 89 lakes and on the subset of the 62 lakes for which the 7 variables were measured (Table 3). The variables were standardized prior to clustering with the exception of the analysis using only alkalinity and sulphate. In all three cases, the first cluster was large and was not subdivided until 5 clusters were used. For 4 or fewer clusters, this first cluster corresponds to the group of inland sites identified from the principal component plots (Table 4 and Figures 4a and 4c). For small numbers of clusters, the variables which are highly positively correlated (Table 1) tend to increase in concentration with cluster number (Figure 5). As the number of clusters is increased, the other variables show larger differences in their cluster means. When seven clusters are formed, clusters 1 and 2 contain the lakes which comprised cluster 1 when only 4 clusters were formed (Table 4 and Figures 4b and 4d). As seen most readily in plots using standardized units (Figure 6), clusters 1 and 2 now have different water colour, pH and sulphate concentrations while the alkalinity, calcium, chloride and conductivity are quite similar. As these sites are all situated inland on granite or slate bedrocks, only minor variability in the marine aerosol and terrestrially weathered variables would be expected. However, as production of organic matter and the biological cycling of sulphate are basin specific processes, these parameters and their concomitant effect on pH are expected to show the most variability.

The frequency functions (Figures 7 and 8) provide more detailed information about the variable values in each cluster. The nearly equal mean pH, when the 89 lakes were separated into 4 clusters using the 5 variables (Figure 5), corresponds to the overlapping ranges (Figure 7). However, when the large first cluster is split, in the division into 7 clusters, the apparent small difference in mean pH in the resulting clusters 1 and 2 (Figure 5) results from groups with very little overlap in the range of pH values (Figure 8). Sulphate also shows an improved separation of frequency distributions by splitting cluster 1 but not as marked as for pH. As would be expected, alkalinity, calcium and chloride exhibited strong cluster overlap when cluster 1 was split. Thus it is clear that the division of cluster 1 accounts for differences in pH and sulphate concentrations. Similar results are obtained when the seven variable, 62 lake data set is split into seven clusters (Figure 9), with the exception that the importance of including water colour is emphasized. Figure 9 indicates that clusters 1 and 2 have little overlap in water colour, a shift in level for pH and sulphate, and very similar frequency functions for alkalinity, calcium and chloride. Thus the division of cluster 1 using the seven variables accounts for differences in pH, water colour and sulphate.

Clusters 3 to 7 generally have small memberships and account for differences in most of the parameters considered. The lakes in these clusters are situated very near to the coast in an area of diverse geology and soil overburden and thus reflect the chemical variability induced by different marine aerosol and terrestrial weathering regimes.

Newfoundland

The eight variables separate into two groups on the basis of their correlation coefficients (Table 5). Alkalinity, conductivity and calcium are highly positively correlated ($r \geq 0.83$) and to a lesser extent these variables are positively correlated with pH ($r \geq 0.56$) and conductivity with chloride ($r = 0.69$). This grouping is similar to that observed for Nova Scotia, with the exception that chloride is not highly correlated with the first three variables, and it suggests that the observed relationships are due predominately to terrestrial weathering processes rather than marine aerosol influence. Logarithmic decay of chloride (Ogden, 1982) and conductivity (Howell, 1986) with distance from the coast has been documented for Nova Scotian lakes. Howell (1986) did not observe this relationship for Newfoundland lakes and suggested that strong orographic effects remove the marine influence very near to the coast. This may explain why chloride is not highly correlated with the other major ions. Water colour is positively related to sulphate and total organic carbon (TOC) and negatively related to the other variables. This good relationship between colour and TOC is to be expected as water colour is a commonly used indicator of organic carbon concentration. The relationship between sulphate and TOC or colour is merely an artifact of the analytical technique used to measure sulphate. The methyl thymol blue sulphate method is subject to colour interference and has been shown by several authors (Kerekes et al., 1984; Howell and Pollock, 1986) to overestimate sulphate concentrations in organic waters.

The first principal component is dominated by the three most highly correlated variables, alkalinity, conductivity and calcium, and to a lesser extent, chloride (Table 6). This is similar to the first component for the Nova Scotia lakes (Table 2), with the exception that it appears to emphasize terrestrial weathering processes over the marine aerosol involvement. However, the positive coefficient for pH and negative coefficients for colour, sulfate and TOC are similar to the second component for Nova Scotia, but the natural organic acidification seems more dominant here. The second component represents water colour, sulphate and total organic carbon, and, because the coefficients for pH or alkalinity are small, does not appear to represent the processes of natural and anthropogenic acidification. Although portions of Newfoundland are geologically sensitive to LRTAP, the geographic location of the island results in low atmospheric deposition of acidic compounds which may explain why the first two components do not clearly indicate the acidification process. The interference problem with the sulfate analysis is another possible explanation. Although the first two principal components account for 76.5 percent of the variance and represent all the variables, the third and fourth components account for enough variance to be retained in the analysis. These latter two components do not have any obvious interpretation.

The plots of the first three principal components (Figure 10) show that most of the lakes form one large group. The group extends from low values on the first and second principal components to high values on both. When represented on the first and third principal axes, the variation in this large group is mostly along the first principal component. Most of the lakes not in this group have higher scores on the first principal component, although some are also separate from the group due to their score on the second or third component. Unlike Nova Scotia, there is no obvious separation of the coastal and inland sites which further emphasizes that marine aerosols may not have a major influence on the chemistry of Newfoundland lakes.

Although the cluster analysis was performed for the sets of 2 and 7 variables used for the Nova Scotia lakes (Table 7), only the clusters based on all eight variables are reported here. The percentage of variation explained by the clusters, for the number of clusters between 2 and 8, is similar for the analysis using 7 or 8 variables and the use of only two variables, sulphate and alkalinity resulted in a small number of clusters. The variation explained by the clusters did not change appreciably by adding the eighth cluster.

Based on the eight variables, for five or fewer clusters, between 106 and 146 of the 152 lakes were located in the first cluster. This large cluster disappeared when 6 clusters were formed and considerable relocation of lakes among clusters occurred. The first occurrence of a second large cluster, when 4 clusters were formed, is due to the separation of lakes with high total organic carbon and colour (Figure 12b). This is cluster 4, when there are 4 clusters and cluster 2 for the 5 to 7 clusters. The small clusters which are obtained by forming a large number of clusters consist of lakes which have higher or lower values for one or more of the variables than do the lakes in the large clusters.

The plots of the location of the clusters for 5 and 7 clusters (Figure 11) indicate that the large clusters do not form separate geographic regions. For 7 clusters (Figure 11b), clusters 4 and 6 are restricted to portions of the Great Northern Peninsula which are underlain by the geological formations of limestone and dolomite. Cluster 2 does not occur in the north-western part of the province while none of the small clusters are observed in the southwest.

The cluster means for alkalinity, pH, conductivity and calcium vary together (Figure 12a). For 7 clusters, these variables have low cluster means for clusters 1, 2, 5, and 7. Mean conductivity and calcium are high in clusters 3 and 4, and, in cluster 3, mean alkalinity is also high. Clusters 2, 3 and 5 all have high mean water colours and total organic carbon concentrations and cluster 3 has high mean chloride concentration. Cluster 7 has higher mean alkalinity, pH, conductivity, chloride and calcium concentrations than cluster 1. Cluster 6 consists of a single lake which has very high conductivity, calcium and alkalinity and low water colour.

Clusters 1, 2 and 7 contain 133 of the 152 lakes with the major difference among these three being higher water colour readings in the lakes in cluster 2 than those in clusters 1 and 7, as shown by the small overlap of ranges (Figure 13). There are considerable differences in the frequency distributions of total organic carbon and sulphate for cluster 2 lakes. The frequency distributions also show that cluster 7 differs from clusters 1 and 2 with higher values and skewness to the right for alkalinity, conductivity and calcium, and with slightly higher pH.

Although the cluster 2 sites have higher water colours and therefore higher potential for natural organic acidification, the underlying geology is less sensitive than for the cluster 1 sites and this may explain why the pH values for the cluster 2 lakes tend to be slightly higher than those of the cluster 1 lakes. The separation of cluster 1 and 7 agrees well with the geological structure of insular Newfoundland. The cluster 7 lakes are situated in areas underlain by sandstone, siltstone, shale, limestones and dolomite, while the cluster 1 sites are situated in areas underlain by granite. These cluster 1 sites are concentrated along the south coast and on the western side of the Great Northern Peninsula and represent the most sensitive LRTAP receptor lakes on insular Newfoundland.

The groups formed by inspection of the principal component plots and by cluster analysis are compared in Table 8. There is good agreement by the two methods for the large groups, but the plots do not separate the small groups as well as the cluster analysis. Plotting of the cluster groups indicates that there is no distinct geographical separation of lakes types, but, when viewed on a geological context, the cluster groups exhibit the expected spatial distribution.

SUMMARY

The applicability of cluster analysis and principal components ordination to divide a set of lakes into well-characterized subsets has been demonstrated. In the choice of the groups, the cluster analysis was more useful since it removed lakes, with different characteristics in only one or a few parameters, from the large clusters. The principal components analysis helped in the understanding of the nature of the larger groups through the relationships between variables. The importance of characterizing the variable values within clusters, both during the choice of the number of clusters and for the final set of clusters, was illustrated.

The lakes from both provinces were divided into 2 (Nova Scotia) or 3 (Newfoundland) large clusters with important differences between the parameters of the groups. A number of small clusters accounted for less common lake types. For both provinces, water colour was an important variable in the determination of the membership of the large clusters. In Nova Scotia both large clusters had similar characteristics for alkalinity, calcium, chloride, and, to a lesser extent, conductivity, the parameters related to terrestrial weathering and marine aerosol. They, however, showed little overlap in the range of values for water colour, and also differed in pH and sulfate. For the Newfoundland lakes, two large clusters displayed the difference in acidity related parameters together with similarity in weathering and marine related parameters, as found for Nova Scotia, and a third large cluster, characteristic of less susceptible, lakes had higher pH values and low colour.

The rationale behind dividing the lakes into subsets with different characteristics and sampling over time from each subset, is to permit separation of effects due to natural processes from those due to atmospheric deposition. The identification of sets of lakes with similar sensitivity but different levels of acidification due to natural organic acids should permit the separation of changes due to atmospheric deposition from those due to natural acidification.

REFERENCES

Anderberg, M.R. 1973. Cluster Analysis for Applications. Academic Press, New York, 359 pp.

Environment Canada. 1979. Analytical Methods Manual. Inland Waters Directorate, Water Quality Branch, Ottawa, Ontario.

Hartigan, J.A. 1975. Clustering Algorithms. Wiley, New York, 351 pp.

Howell, G.D. 1986. Water Quality Characterization of 78 Acid Precipitation Monitoring Lakes in Atlantic Canada. Inland Waters Directorate. Water Quality Branch Report IWD-AR-WQB-86-109. July, 1986. 119 pp.

Howell, G.D. and T.L. Pollock. 1986. Sulphate, Water Colour and Dissolved Organic Carbon Relationships in Organic Waters of Atlantic Canada. In: A.H. El-Shaarawi and R.E. Kwiatkowski (eds.). Statistical Aspects of Water Quality Monitoring. Developments in Water Science. Vol. 27:53-63.

Kerekes, J., Howell, G. and Pollock, T. 1984. Problems Associated with Sulphate Determination in Coloured Humic Waters in Kejimkujik National Park, Nova Scotia, Canda. Verh. Internat. Verein. Limnol. 22:1811-1817.

Morrison, D.F. 1967. Multivariate Statistical Methods. McGraw-Hill, New York, 338 pp.

Ogden. 1982. Seasonal Mass Balance of Major Ions in Three Small Watersheds in a Maritime Environment. Water, Air and Soil Pollution, 17:119-130.

Table 1. Correlation coefficients for the Nova Scotia lakes for which five variables and seven variables were measured.

	5 variables 89 lakes				7 variables 62 lakes					
	Alk.	SO_4	Ca	Cl	Alk.	Cond.	Colour	SO_4	Ca	Cl
pH	0.55	0.12	0.54	0.39	0.43	0.36	-0.58	0.17	0.49	0.37
Alkalinity		0.27	0.94	0.89		0.91	-0.36	0.29	0.95	0.92
Conductivity							-0.28	0.54	0.92	0.98
Colour								0.05	-0.33	-0.33
Sulphate			0.39	0.51					0.40	0.51
Calcium				0.93						0.94
Chloride										

Table 2. Principal components for the Nova Scotia lakes obtained from the correlation matrices.

	Standardized eigenvectors					
	5 variables 89 lakes			7 variables 62 lakes		
	1	2	3	1	2	3
pH	-0.64	0.68	1.00	-0.58	0.79	-0.56
Alkalinity	-0.97	0.22	-0.37	-0.97	-0.02	0.42
Conductivity				-0.99	-0.27	0.12
Colour				0.46	-1.00	0.12
Sulphate	-0.52	-1.00	0.65	-0.53	-0.65	-1.00
Calcium	-1.00	0.07	-0.27	-1.00	-0.07	0.23
Chloride	-0.97	-0.20	-0.36	-1.00	-0.22	0.15
Percent of total variance	67.4	18.8	11.7	63.3	18.7	10.5
Cumulative percent of total variance	67.4	86.2	97.9	63.3	82.0	92.5

Table 3. Variation explained by clusters for three combinations of variables and number of Nova Scotia lakes.

Number of Clusters	Percent of variation explained by clusters for (number of variables, number of lakes)		
	2,89	5,89 std	7,62 std
2	61	54	54
3	70	64	63
4	80	68	68
5	85	75	76
6	85	79	80
7	89	84	82
8	90	84	83

std indicates that the variables were standardized prior to cluster analysis

Table 4. Comparison of group membership obtained by clustering and principal component plots of the Nova Scotia Lakes.

	5 variables 89 lakes				7 variables 62 lakes			
	Cluster analysis		Principal components		Cluster analysis		Principal components	
Group	4 clusters	7 clusters	PC1,PC2	PC1, PC3	4 clusters	7 clusters	PC1,PC2	PC1,PC3
1	1-41,66-74 76-78,80-89	10,13,14,18,21 25,29,32,39 41,66,67,70-72 76-78,80-88	1-27,29-41 66-70,72-78 80-89	1-27,29-41 66-78,80-89	1-41,50,51	1,3-7,9,11-13 17,22-28,30,31 33,35-38	1-41	1-41,61
2	42,50,51,53 54,57,61 63-65,75,79	1-9,11,12 15-17,19,20 22-24,26-28 30,31,33-38,40 50,68,69,73,74,89	between groups 1&3 on PC1 axis 28,50,51,53 61,65,79	between groups 1&3 on PC1 axis 28,50,51,53 61,65,79	54-58,62	2,8,10,14-16 18-21,29,32 34,39-41,53 61	between groups 1&3 on PC1 axis 28,50,51,53 61	between groups 1&3 on PC1 axis 28,42,50,51 53,54,57,62
3	42,52,55,56 58,62	51,53,61,65 75,79	42-45,48,49 52,54-60,62-64	44,45,48,49 52,54,56,57 59,60,62,64	43,44,46-49 59-60	42,50,51,54 57,62	42-44,48,49 55,56,58,59 60	44-45,48,49 52,56,59,60
4	43,44,46-49 59,60	42,44,45,48 52,63,64	below group 3 on PC1 axis 46,47	below group 3 on PC1 axis 46,47	42,45,52,53 61	55,58	below group 3 on PC1 axis 46,47	below group 3 on PC1 axis 43,46,47
5		55,58		55,58		43,44,46-48		55,58
6		43,46,47,49 60		near the boundaries of group 3 42,43,63		42,52,56	near the boundaries of group 3 42,52,54,57 62	
7		54,56,57,59 62				49,59,60		

The numbers in the body of the table are lake numbers.

Table 5. Correlation coefficients between the eight variables measured for 152 Newfoundland lakes.

	Alk.	Cond.	Colour	SO_4	Ca	TOC	Cl
pH	0.68	0.56	-0.39	-0.34	0.59	-0.27	0.35
Alkalinity		0.83	-0.26	-0.11	0.94	-0.12	0.31
Conductivity			-0.19	0.12	0.88	-0.01	0.69
Colour				0.61	-0.20	0.77	-0.07
Sulphate					0.04	0.71	0.09
Calcium						-0.07	0.38
Total organic carbon							0.10
Chloride							

Table 6. Principal components for the Newfoundland lakes obtained from the correlation matrix.

Variable	Standardized eigenvector			
	1	2	3	4
pH	0.88	-0.14	-0.08	-1.00
Alkalinity	1.00	0.20	-0.45	0.02
Conductivity	0.98	0.46	0.15	0.34
Colour	-0.55	0.86	-0.19	-0.55
Sulphate	-0.29	0.98	-0.06	0.60
Calcium	0.98	0.34	-0.37	0.30
Total organic carbon	-0.37	1.00	-0.07	-0.47
Chloride	0.60	0.43	1.00	-0.11
Percent of total variance	46.7	29.8	10.0	6.4
Cumulative percent of total variance	46.7	76.5	86.5	92.9

Table 7. Variation explained by clusters of Newfoundland lakes for different numbers of variables.

Number of Clusters	Percent of variation explained by clusters for (number of variables, number of lakes)		
	pH & SO_4 2,152	7,152 std	8, 152 std
2	69	30	27
3	81	40	38
4	90	60	56
5	91	63	63
6	97	67	66
7	97	74	75
8	97	80	74

std indicates that the variables were standardized prior to cluster analysis.

Table 8. Comparison of group membership obtained by clustering and principal component plots for the 8 variables and 152 Newfoundland lakes.

	Cluster analysis		Principal Components	
Group	5 clusters	7 clusters	PC1, PC2	PC1,PC3
1	4,6-10,14,17-19,21 23,26-30,32-41,44-48 50,51,54-61,62,64 69-71,73-78,80,85-91 94-98,100,102-106 109-115,117,119-125 127-138,140,141 143-147,148,149 151,152	8-10,14,17-19,27 28,34,36-39,41 46,50,54,56,57 64,69,73,74,76 78,80,95,100,102 106,120,124,125 127-129,131,133-138 140,141,143,144 146-149,151,152	4,7-10,14,17-21 27-30,32-42 44-48,50-74 76-92,94-152	4,6-10,13,14,17-21 23-30,32-92 94-152
2	20,25,42,43,49 52,53,60,63,65-68 72,79,81-84,92,99 101,107,108,116,118 126,139,142,145,150	32,33,42,48,51 52,55,58,60,63 65,68,70-72,79 81-84,90,92,96 97,99,101,105,107 108,116,117,126,139 142,145,150		
3	12,13,15,22,93	12,15,22,93	higher PC2, high PC1 13,22,93	higher PC1, positive PC3 12,15,22,93
4	1,2,5,11,16,24,31	1,2,5,11,16	higher on PC1 axis than group 1 1,2,5,6,11,12,15,16 23,24,26,75	higher PC1, negative PC3 1-3,5,11,16,31
5	3	13,20,25,43,49 53,66,67,118	higher PC2, low PC1 25,43,59	
6		3		
7		4,6,7,21,23,24,26 29-31,35,40,44,45 47,59,61,62,75,77 85-89,91,94,98 103,104,109-115,119 121-123,130,132		

The numbers in the body of the plot are lake numbers

FIGURE CAPTIONS

Figure 1. Locations of lakes sampled in Nova Scotia.

Figure 2. Locations of lakes sampled in Newfoundland.

Figure 3. Scores of the Nova Scotia lakes on the first and second and first and third principal components for 89 lakes and 5 variables (a and b) and 62 lakes and 7 variables (c and d).

Figure 4. Cluster membership plotted at the lake location for the 4 and 7 clusters obtained when 89 lakes and 5 variables (a and b) and 62 lakes and 7 variables (c and d) were used.

Figure 5. The Nova Scotia cluster means, in the original units, for the number of clusters, k, equal to 2 to 7, and for 89 lakes and 5 variables (a) and 62 lakes and 7 variables (b). The cluster number, ℓ, number of lakes in cluster ℓ, n_ℓ, and the percentage of variation explained by k clusters, R_k, are also shown.

Figure 6. The Nova Scotia cluster means in standardized units, for 7 clusters and both combinations of numbers of lakes and variables. The cluster number ℓ, number of lakes in cluster ℓ, n_ℓ, and the percentage of variation explained by 7 clusters, R_7, are also shown.

Figure 7. Frequency functions of the water quality parameters shown separately for each of the clusters when 4 clusters were determined from the 89 Nova Scotia lakes using the 5 variables.

Figure 8. Frequency functions of the water quality parameters shown separately for each of the clusters when 7 clusters were determined from the 89 Nova Scotia lakes using the 5 variables.

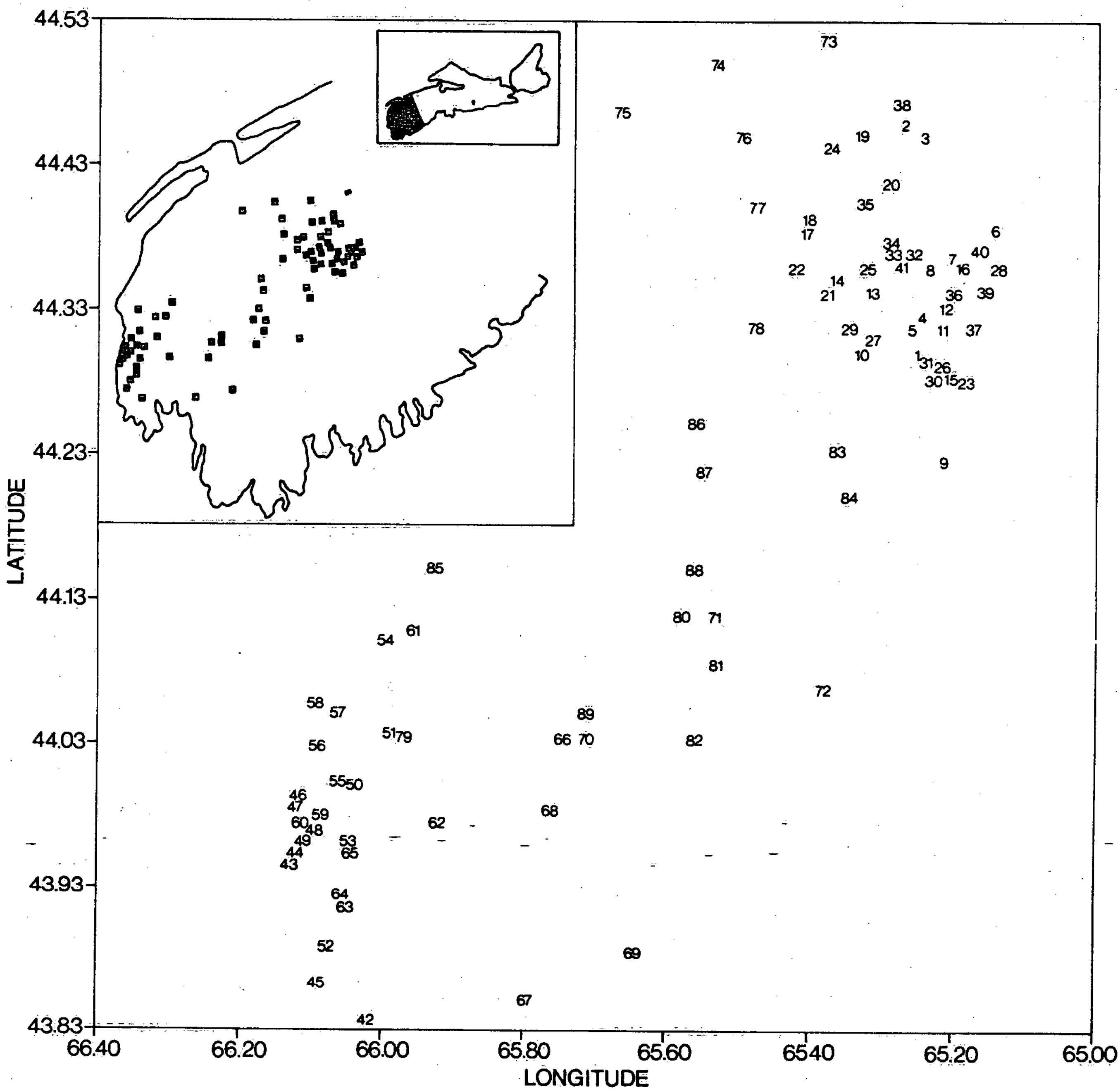
Figure 9. Frequency functions of the water quality parameters shown separately for each of the clusters when 7 clusters were determined from the 62 Nova Scotia lakes using the 7 variables.

Figure 10. Scores of the Newfoundland lakes on the first and second and first and third principal components for the 152 lakes and 8 variables.

Figure 11. Cluster membership plotted at lake location for the 5 and 7 clusters obtained from the Newfoundland lakes.

Figure 12. The Newfoundland cluster means, in the original units, for the number of clusters, k, equal to 2 to 7. The cluster number, ℓ, the number of lakes in cluster ℓ, n_ℓ, and the percentage of variation explained by k clusters, R_k, are also shown.

Figure 13. Frequency functions of the water quality parameters shown separately for each of the clusters when 7 clusters were determined from the Newfoundland lakes.



44.53
44.43
44.33
44.23
44.13
44.03
43.93
43.83
LATITUDE
66.40
66.20
66.00
65.80
65.60
65.40
65.20
65.00
LONGITUDE

FIGURE 1

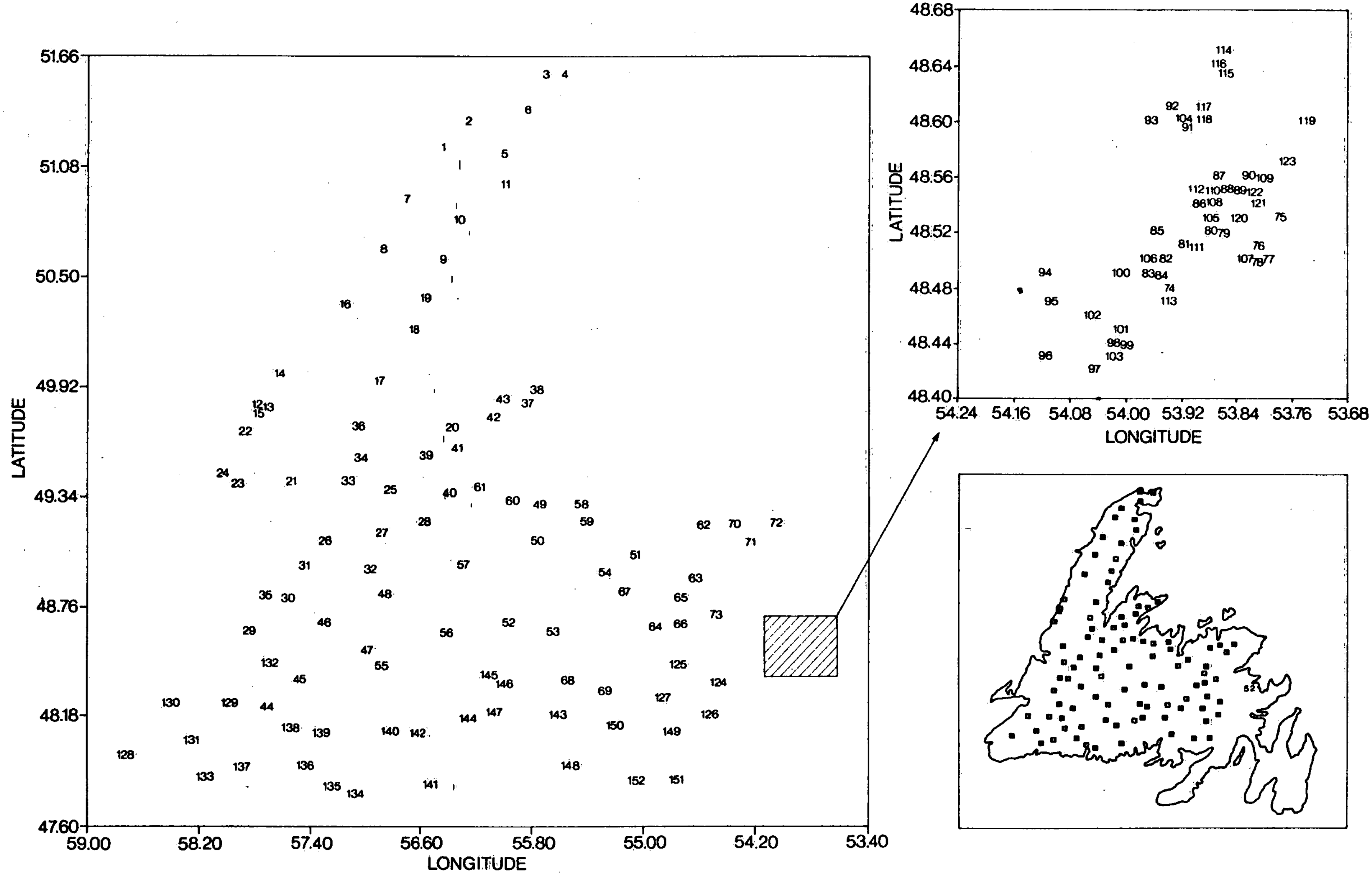

LATITUDE
51.66
51.08
50.50
49.92
49.34
48.76
48.18
47.60
59.00
58.20
57.40
56.60
55.80
55.00
54.20
53.40
LONGITUDE
LATITUDE
48.68
48.64
48.60
48.56
48.52
48.48
48.44
48.40
54.24
54.16
54.08
54.00
53.92
53.84
53.76
53.68
LONGITUDE

FIGURE 2

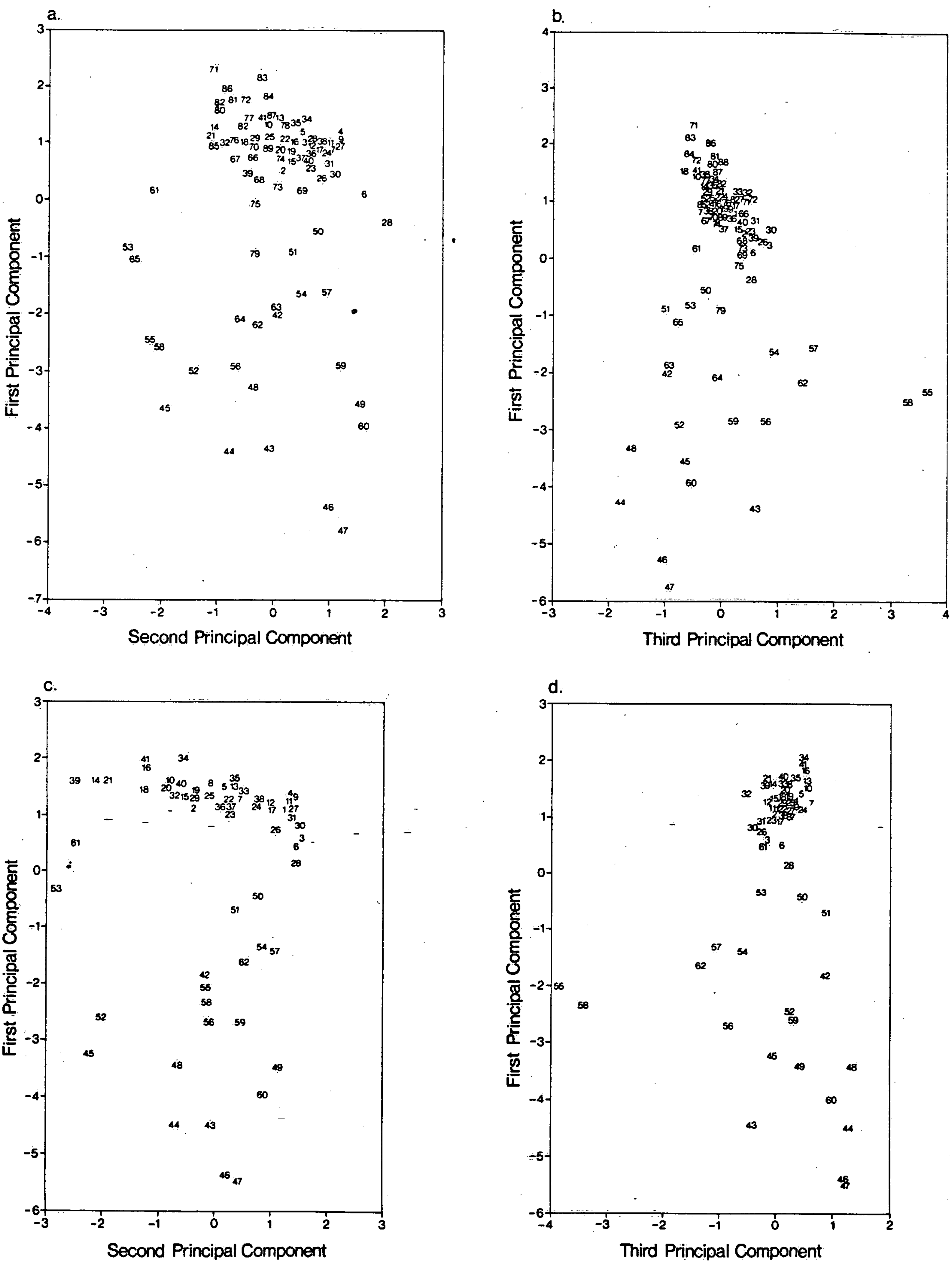

a.
First Principal Component
Second Principal Component
b.
First Principal Component
Third Principal Component
c.
First Principal Component
Second Principal Component
d.
First Principal Component
Third Principal Component

FIGURE 3

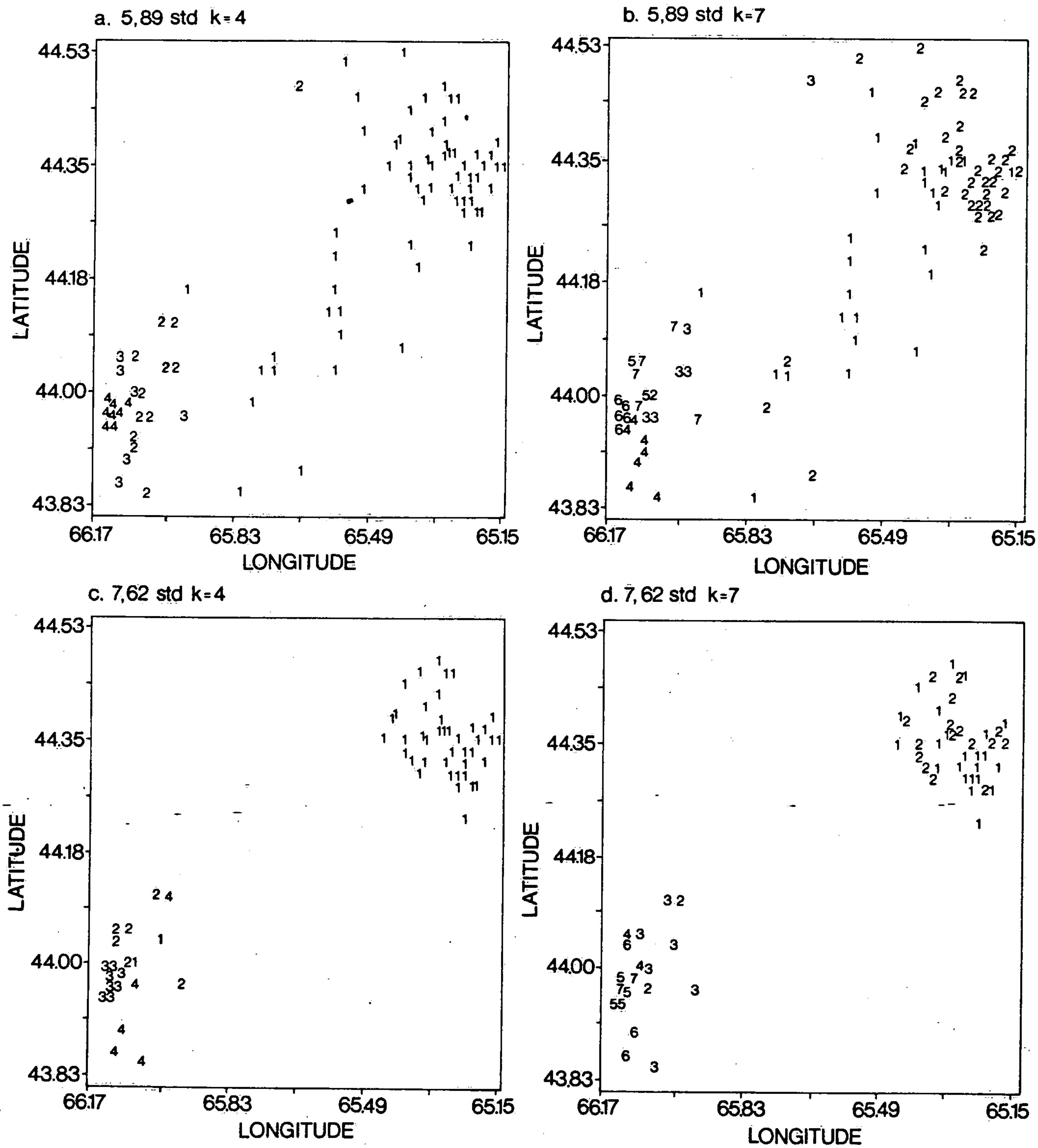

a. 5,89 std k=4
b. 5,89 std k=7
c. 7,62 std k=4
d. 7,62 std k=7
LATITUDE
LONGITUDE
44.53
44.35
44.18
44.00
43.83
66.17
65.83
65.49
65.15

FIGURE 4

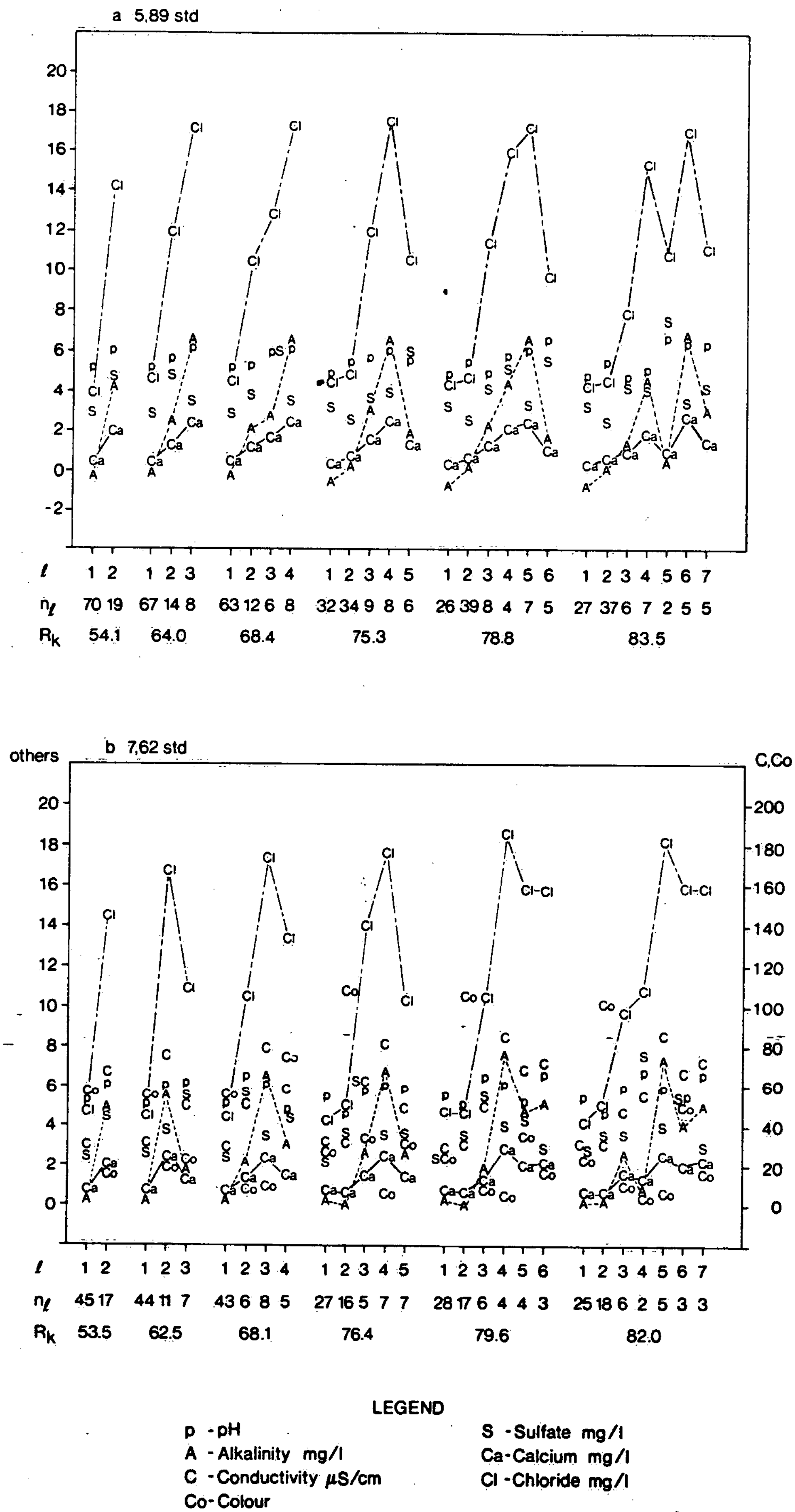
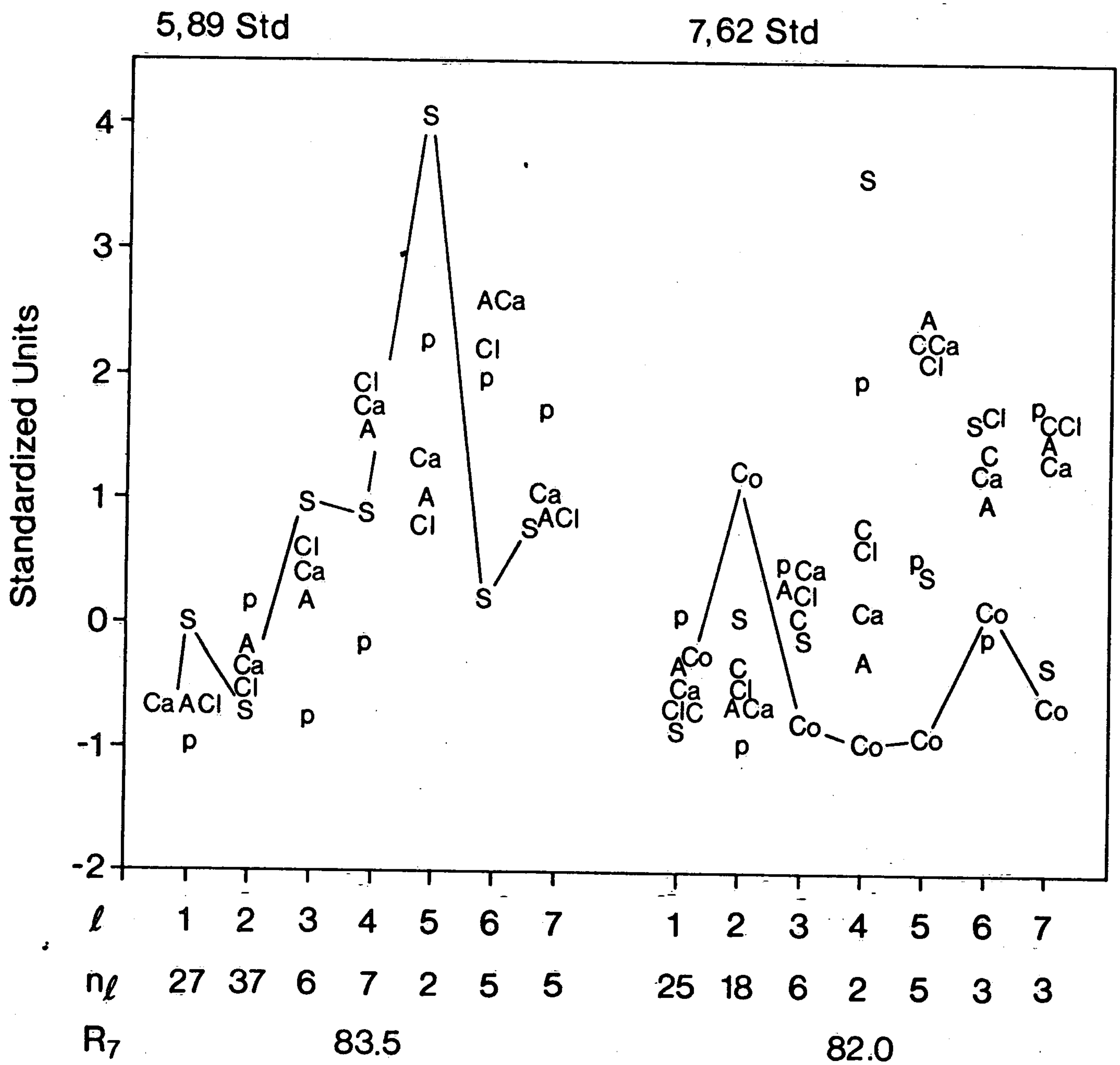

a 5,89 std
ℓ 1 2 1 2 3 1 2 3 4 1 2 3 4 5 1 2 3 4 5 6 1 2 3 4 5 6 7
$n_ℓ$ 70 19 67 14 8 63 12 6 8 32 34 9 8 6 26 39 8 4 7 5 27 37 6 7 2 5 5
R_k 54.1 64.0 68.4 75.3 78.8 83.5
others
b 7,62 std
C,Co
ℓ 1 2 1 2 3 1 2 3 4 1 2 3 4 5 1 2 3 4 5 6 1 2 3 4 5 6 7
$n_ℓ$ 45 17 44 11 7 43 6 8 5 27 16 5 7 7 28 17 6 4 4 3 25 18 6 2 5 3 3
R_k 53.5 62.5 68.1 76.4 79.6 82.0
LEGEND
p - pH
A - Alkalinity mg/l
C - Conductivity μS/cm
Co - Colour
S - Sulfate mg/l
Ca - Calcium mg/l
Cl - Chloride mg/l

FIGURE 5



LEGEND

p - pH

A - Alkalinity mg/l

C - Conductivity μS/cm

Co - Colour

S - Sulfate mg/l

Ca - Calcium mg/l

Cl - Chloride mg/l

FIGURE 6

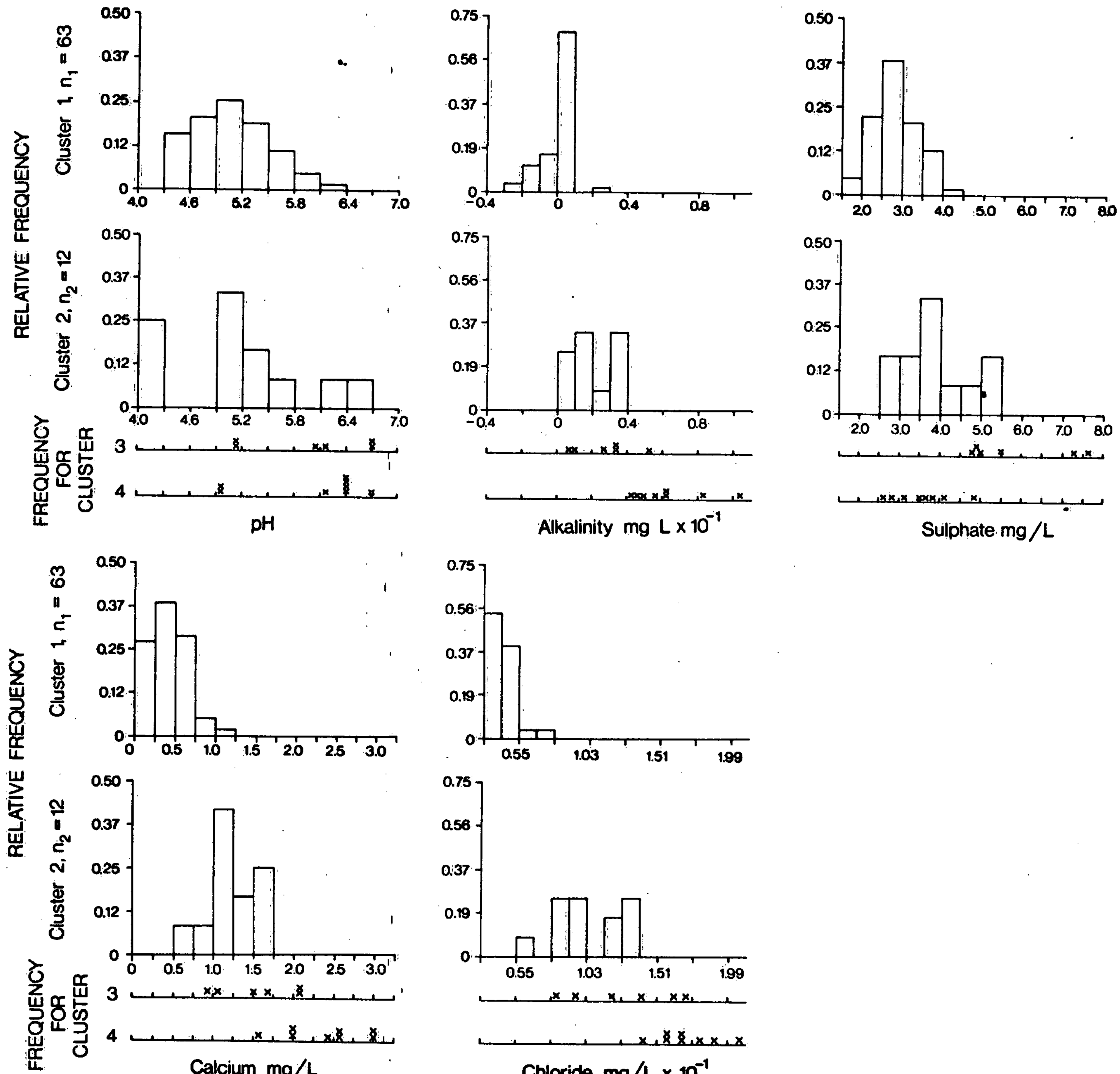

RELATIVE FREQUENCY
Cluster 1, n_1 = 63
Cluster 2, n_2 = 12
FREQUENCY FOR CLUSTER
3
4
pH
Alkalinity mg L x 10^{-1}
Sulphate mg/L
Calcium mg/L
Chloride mg/L x 10^{-1}

FIGURE 7

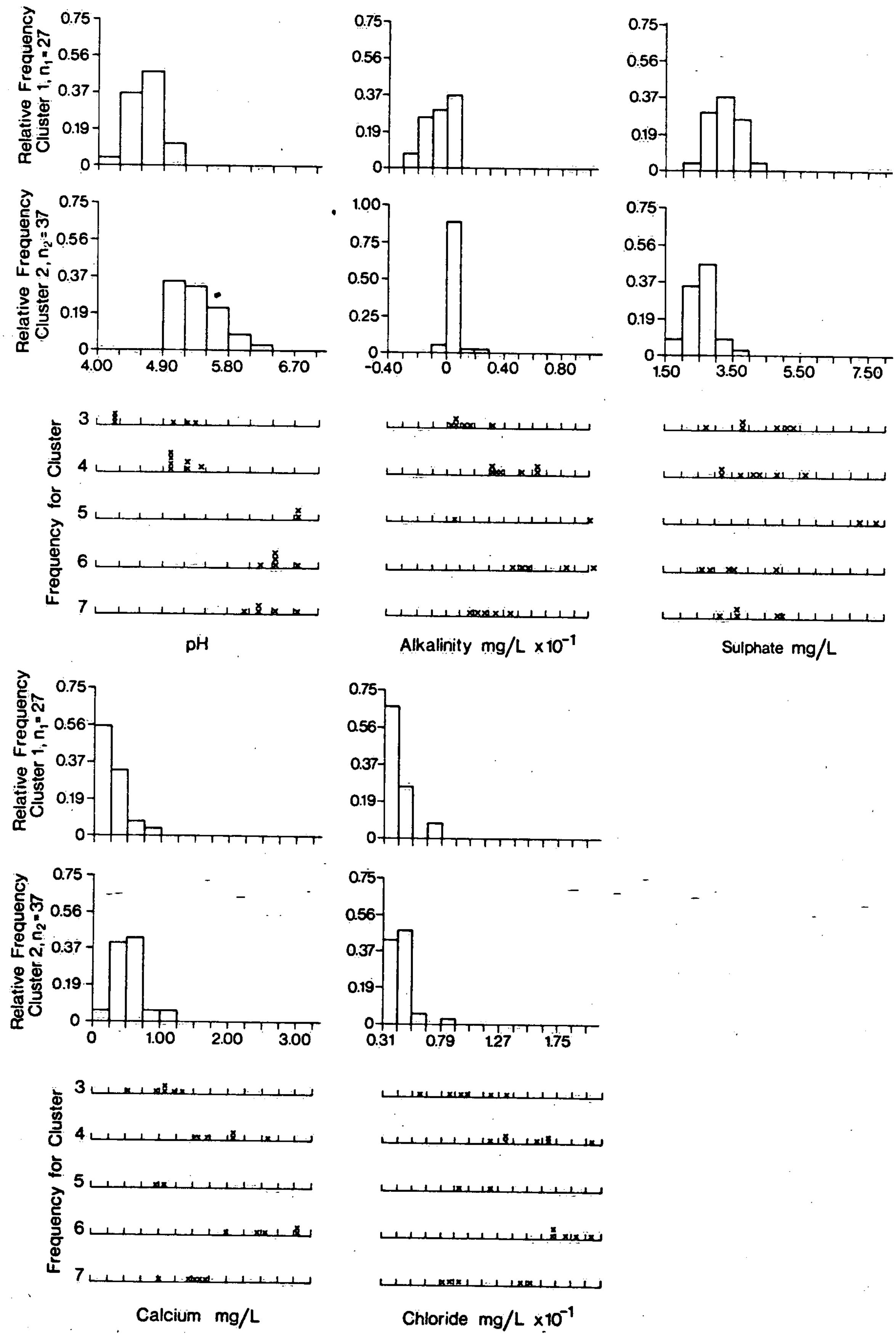

Relative Frequency Cluster 1, n_1=27
Relative Frequency Cluster 2, n_2=37
Frequency for Cluster
pH
Alkalinity mg/L x10^{-1}
Sulphate mg/L
Calcium mg/L
Chloride mg/L x10^{-1}

FIGURE 8

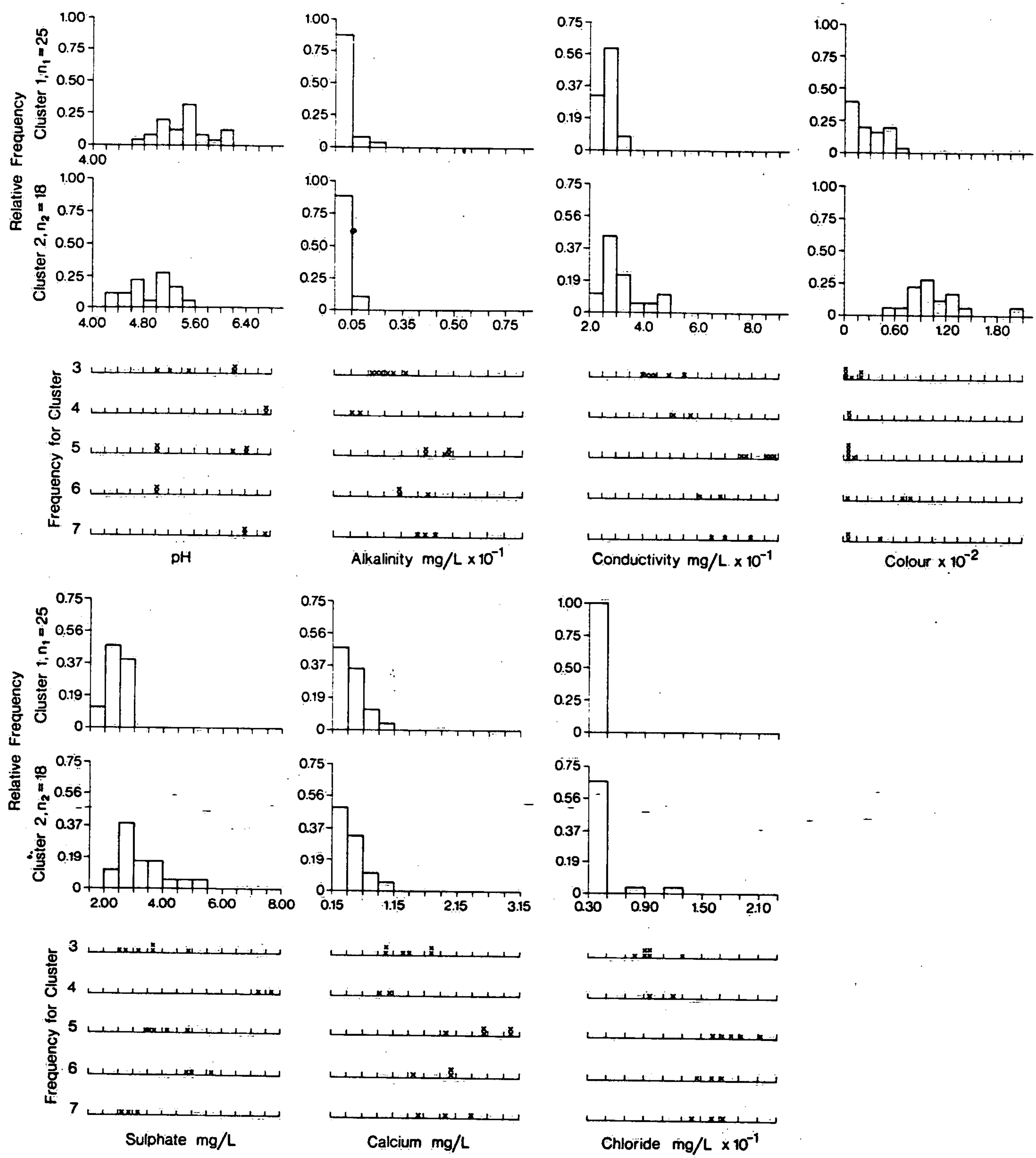

Relative Frequency
Cluster 1, n_1 = 25
Cluster 2, n_2 = 18
Frequency for Cluster
pH
Alkalinity mg/L x 10^{-1}
Conductivity mg/L x 10^{-1}
Colour x 10^{-2}
Sulphate mg/L
Calcium mg/L
Chloride mg/L x 10^{-1}

FIGURE 9

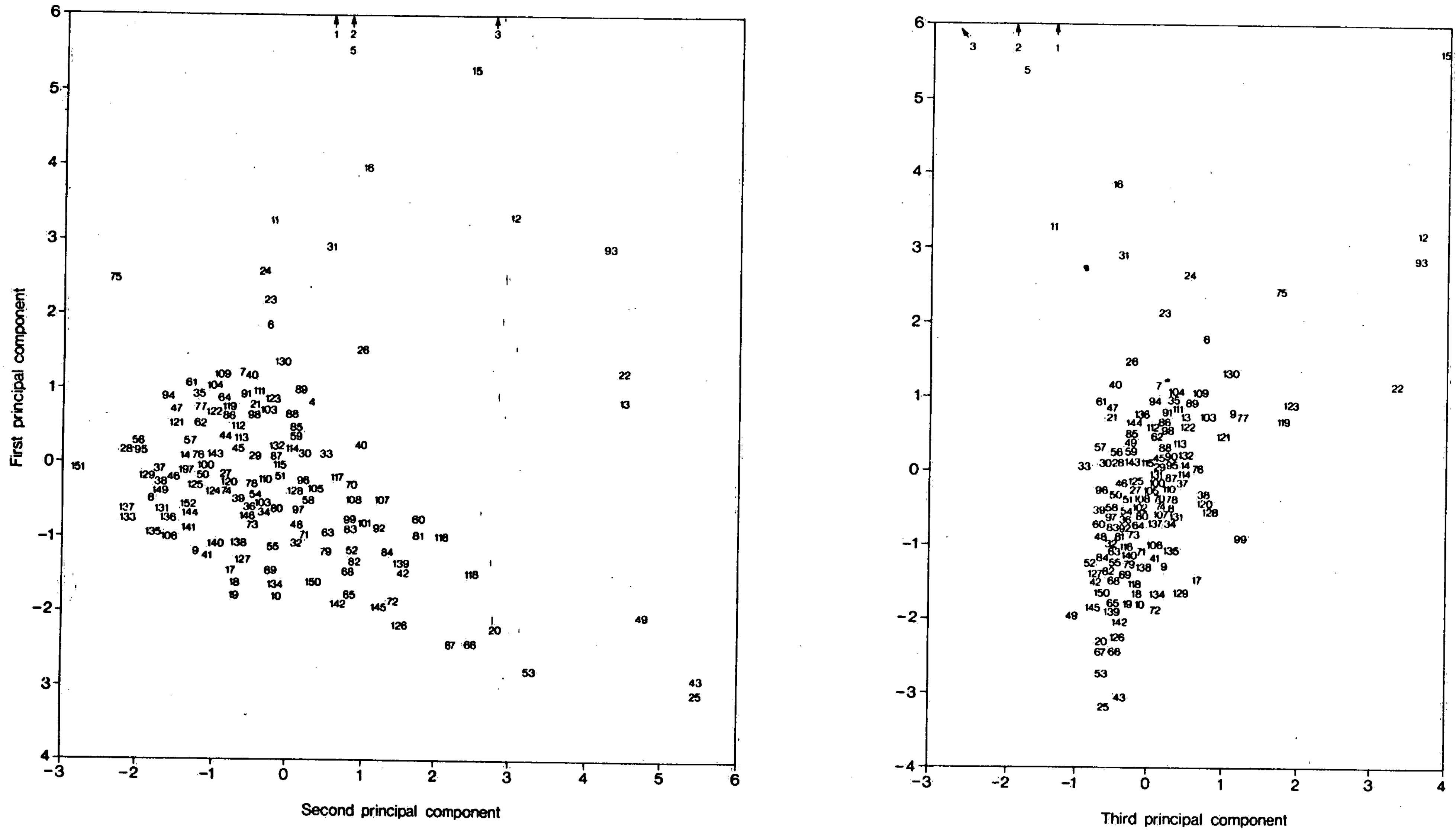

First principal component
Second principal component
Third principal component

FIGURE 10

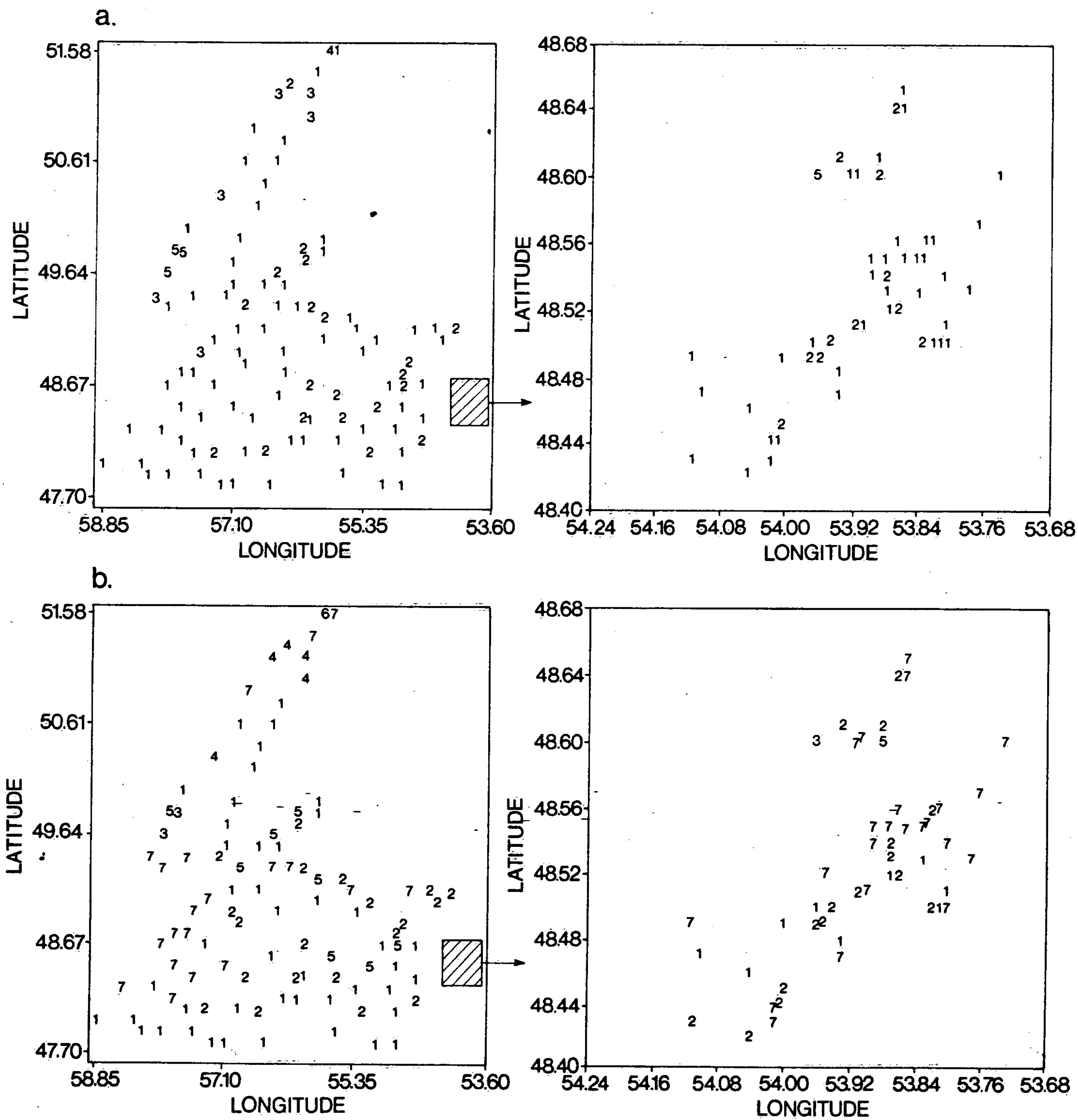
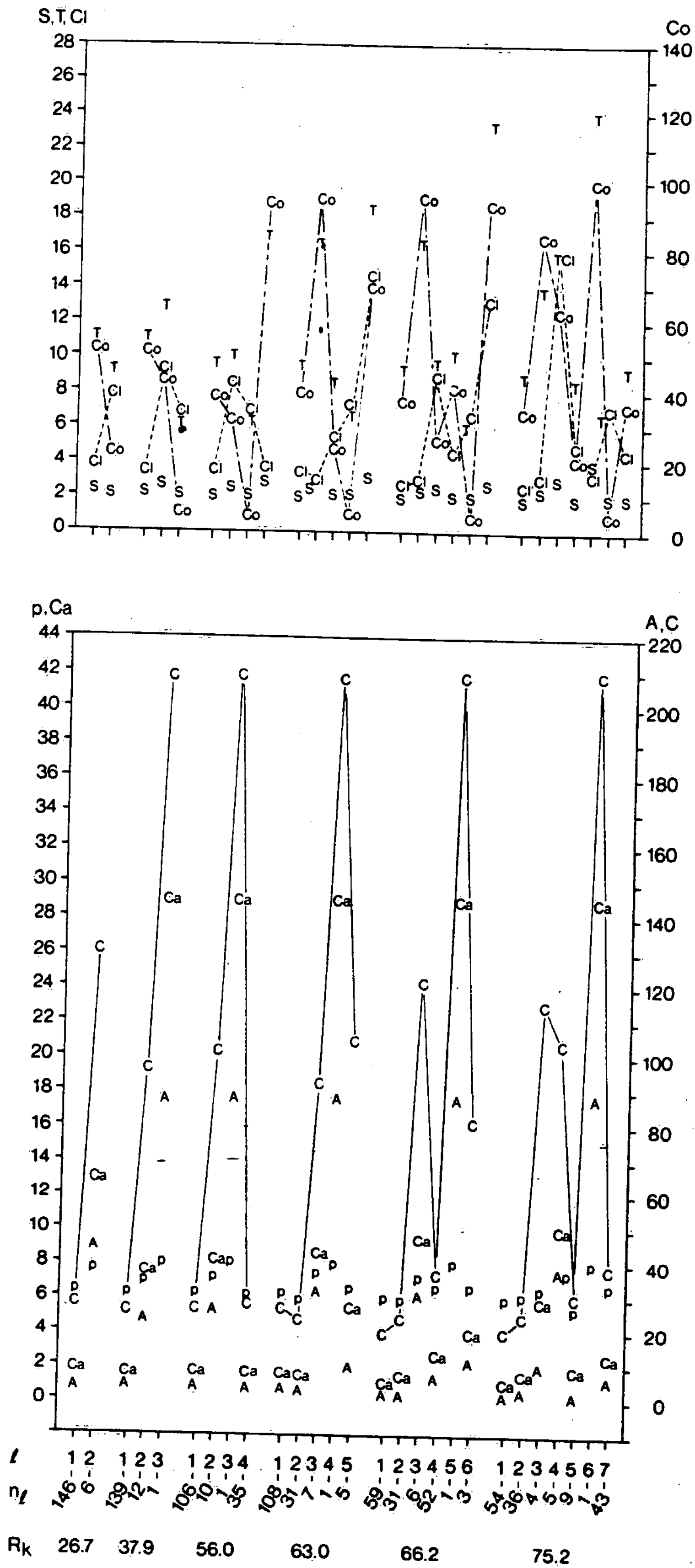

a.
b.
LATITUDE
LONGITUDE
51.58
50.61
49.64
48.67
47.70
58.85
57.10
55.35
53.60
48.68
48.64
48.60
48.56
48.52
48.48
48.44
48.40
54.24
54.16
54.08
54.00
53.92
53.84
53.76
53.68

FIGURE 11



LEGEND

p - pH
A - Alkalinity mg/l
C - Conductivity µS/cm
T - Total organic carbon mg/l
S - Sulfate mg/l
Ca - Calcium mg/l
Cl - Chloride mg/l
Co - Colour

FIGURE 12

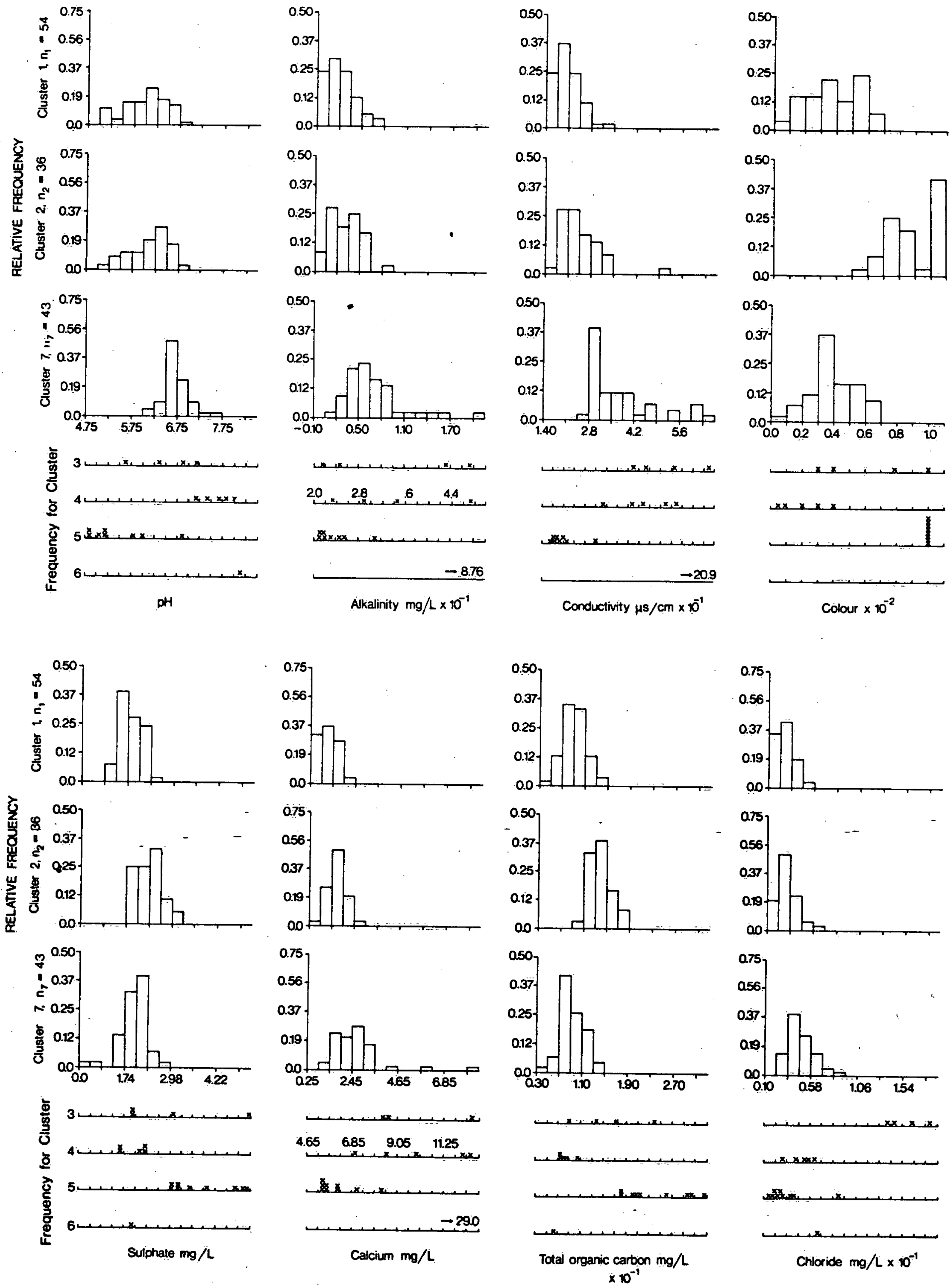

RELATIVE FREQUENCY
Cluster 1, n1 = 54
Cluster 2, n2 = 36
Cluster 7, n7 = 43
Frequency for Cluster
pH
Alkalinity mg/L x 10^-1
Conductivity µs/cm x 10^-1
Colour x 10^-2
Sulphate mg/L
Calcium mg/L
Total organic carbon mg/L x 10^-1
Chloride mg/L x 10^-1

FIGURE 13