Interpretation of hydrogeochemical facies by multivariate statistical methods

GJ van Tonder* and FDI Hodgson
Institute for Groundwater Studies, P.O. Box 339, Bloemfontein 9300, South Africa

Abstract

Hydrochemical facies interpretations are useful tools for determining the chemical histories of groundwater bodies. This paper applies a number of multivariate statistical methods like cluster analysis, principal components analysis, factor analysis and discriminant function analysis to the interpretation of water samples from the Venterstad area, Cape Province. From the analyses it is concluded that three groundwater masses exist. These conclusions are verified by a comparison of the results with actual ^{14}C and tritium age determinations on some of the samples.

Introduction

Hydrogeochemical facies are frequently compared by means of graphical representations such as trilinear diagrams (Piper, 1944), which have contributed greatly to the understanding of trends in water quality. These facies interpretations are useful tools for determining the origins and chemical histories of groundwater masses. Although these facies interpretations are still speculative, a whole new field has been opened and many chemical analyses of groundwater samples are presently performed with the purpose of distinguishing between the various types of groundwater occurrences.

Diagrammatic techniques for the classification and interpretation of hydrogeochemical data are numerous. Each of these serves a purpose. Although the Piper diagram is extensively used throughout the literature, it suffers from a serious drawback, namely that percentages of the various ions are considered and actual concentrations of the dissolved solids are therefore neglected. This obvious shortcoming is overcome by the Durov doubled triangular diagram and also the Schoeller diagram (Zaporosec, 1972). The latter display, however, poses the problem that a limited number of water analyses can be presented in a diagram. Even the Durov diagram suffers specific drawbacks and a limited number of parameters can be considered.

Lawrence and Upchurch (1976, 1982) and Dalton and Upchurch (1978) showed that factor analysis can be used in much the same way as the graphical methods.

Basic principles of the various multivariate statistical methods

Cluster analysis

Cluster analysis can be used to classify variables or objects into one or more homogeneous groups, in such a way that the relationship between the various groups will become evident. Authors such as Hartigan (1972) and Agterberg (1974) have given a comprehensive account of the mathematics and applicability of the technique.

Suppose m characteristics are measured on n objects. Some similarity measure can then be computed between every pair of objects. Coefficients of similarity which can be used, are the correlation coefficient and the standardized Euclidean distance d_{ij}.

$$d_{ij} = \sqrt{\sum_{i=1}^{n}(X_{ik} - X_{jk})^2/n)}$$

where X_{ik} denotes the k-th variable measured on object i and X_{jk} the k-th variable measured on object j, and d_{ij} is the distance between objects i and j. After the completion of this computation, a symmetric n x n matrix of similarities between observations is obtained. Objects with highest mutual similarity are then grouped together in a dendrogram.

Principal component analysis

Principal components are eigenvectors of a variance-covariance or correlation matrix (Morrison, 1976). The eigenvectors of a matrix yield the principal axes of an ellipsoid and the eigenvalues represent the lengths of these axes. Principal component analysis is concerned with the finding of these axes and the measurement of their magnitudes.

Suppose X is a data matrix. The variance-covariance matrix S, is found from

$$S = X' X/n - 1$$

where X' is the transpose of X.

The total variance of the system is the sum of the variances of the original variables and is calculated from

$$\text{Var } X = \text{Trace } S$$

In other words, the total variance is the sum of the diagonal elements in S. The eigenvalues of S are found by expanding the following determinant

$$|S - \mu(I)| = o$$

Note that $\Sigma\mu_i = \text{Trace } S$. From this the matrix A of eigenvectors associated with the corresponding eigenvalues is obtained, whereafter the values of the principal components can be calculated from

$$Z = X.A.$$

Factor analysis

The factor model may be expressed as (Jöreskog *et al.*, 1976)

---

*To whom all correspondence should be addressed.

Received 25 May 1985.

$$X_j = \sum_{r=1}^{p} A_{rj}F_r + A_jE_j$$

The equation indicates that any variable X_j consists of a linear combination of p common factors plus a unique factor A_jE_j. The F's refer to hypothetical variables called factors, which can be expressed as

$$F = XB \ , \ B = AD_{\mu}^{-1/2}$$

where A is the p x p matrix with columns which are characteristic of vectors of the correlation matrix R, and D_{μ} is a p x p diagonal matrix whose diagonal elements are the characteristic roots of R and $D_{\mu}^{-1/2}$ is the diagonal matrix formed by inverting D_{μ} and taking the square root of its elements (Klovan, 1975).

In order to obtain the best interpretation, the factor loadings (i.e. $AD_{\mu}^{-1/2}$) can be rotated by means of Kaisers' varimax scheme.

Two types of factor analysis, the R-mode and Q-mode exist. The interrelationship between variables is investigated under the R-mode, whereas the interrelationship between samples is studied with the Q-mode.

Discriminant functions

The discriminant function categorizes a sample into one of two or more groups of samples which are defined by prior knowledge to

TABLE 1
CHEMICAL ANALYSIS OF 17 WATER SAMPLES FROM THE VENTERSTAD AREA, CAPE PROVINCE
(AFTER VAN DER LINDE AND HODGSON, 1977)

Sample	Ca^{++}	Mg^{++}	Na^{+}	K^{+}	$CO_3^{=}$	HCO_3^{-}	$SO_4^{=}$	Cl^{-}	Age (years)
1	58	17	60	5	0	311	29	32	0*
2	42	42	46	0	42	238	19	65	16**
3	54	23	19	1	38	220	43	1	
4	61	29	35	2	35	227	61	12	
5	84	47	34	1	38	358	48	32	
6	66	32	48	1	39	294	78	15	
7	3	1	75	0,1	19	73	11	14	
8	21	5	148	2	38	246	61	30	
9	9	1	83	1	16	156	4	16	
10	10	4	67	0	24	134	0	25	1640*
11	8	1	83	0	0	183	14	28	
12	8	1	79	0,4	12	152	5	16	
13	21	2	125	0,7	16	76	35	136	
14	34	4	165	1	5	40	24	284	4120*
15	27	2	151	0,7	0	46	13	252	
16	25	1	140	0,8	5	25	22	200	
17	26	1	131	0,8	3	24	45	160	

*^{14}C – method of age determination
**Tritium method of age determination

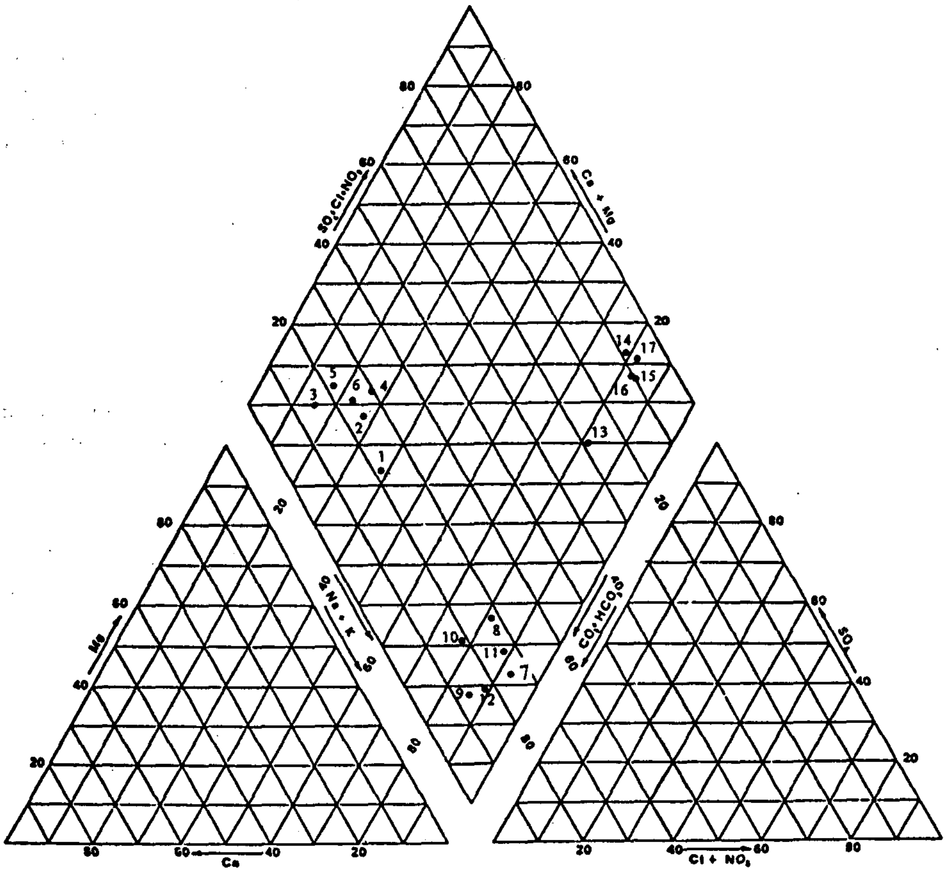

Figure 1
Trilinear diagram showing composition ratios for the water samples from the Venterstad area.

be different. This categorizing is accomplished on the basis of values obtained by measurement of the same set of variables for each sample in all groups. The function is determined from these samples of known populations.

A clear description of the mathematics needed to compute these functions, can be found in Davis (1973).

Application of multivariate statistical methods to the water quality of the Venterstad area, Cape Province

To demonstrate the effectiveness of the various statistical methods for hydrochemical typing, the chemical analyses of 17 groundwater samples from the Venterstad area have been selected. Van der Linde and Hodgson (1977) have stressed that groundwater of a specific composition is indicative of a certain environment and they proved it with the usual diagrammatic techniques for the Venterstad area. The results of the chemical analyses of the 17 groundwater samples are presented in Table 1. A close investigation of the parameter values in this table will indicate that three types of water are present. The obvious difference is the variation in the calcium, magnesium, sodium, bicarbonate, sulphate and chloride values. A plot of the Piper diagram (Figure 1) confirms the existence of three types of water in the Venterstad area. The three types of water have respectively a $Ca/Mg(HCO_3)_2$-, $NaHCO_3$- and NaCl-character.

Cluster analysis

The intention is to cluster the water samples into chemically homogeneous groups. The first step would be to calculate the 17 × 17 matrix of similarities between the samples, using the correlation coefficient as a measure. Thereafter a weighted pair-group averaging can be performed. The linkage table is presented in Table 2.

TABLE 2
LINKAGE OF WATER SAMPLES USING CORRELATION COEFFICIENTS

Sample Number	Pairs	Similarity Level
1	2	1,29
3	4	1,31
9	10	1,24
6	17	1,29
3	6	1,28
14	16	1,26
3	5	1,25
14	15	1,24
1	3	1,23
13	14	1,17
8	9	1,20
1	7	1,16
11	12	1,19
8	11	1,18
1	8	1,26
1	13	0,69

Figure 2 presents the dendrogram of the similarity levels in Table 2. Clearly three major clusters exist. Water samples 1, 2, 3, 4, 5 and 6 form one group, while samples 7, 8, 9, 10, 11 and 12 group together to form the second group and the third group constitutes samples 13, 14, 15, 16 and 17. Group 1 is linked to group 2 at a level of 1,16, while group 2 is linked to group 3 at a level of 0,69. From this it is clear that group 2 resembles group 1 to a much greater extent than it resembles group 3. The conclusions drawn from these calculations confirm the existence of three types of water.

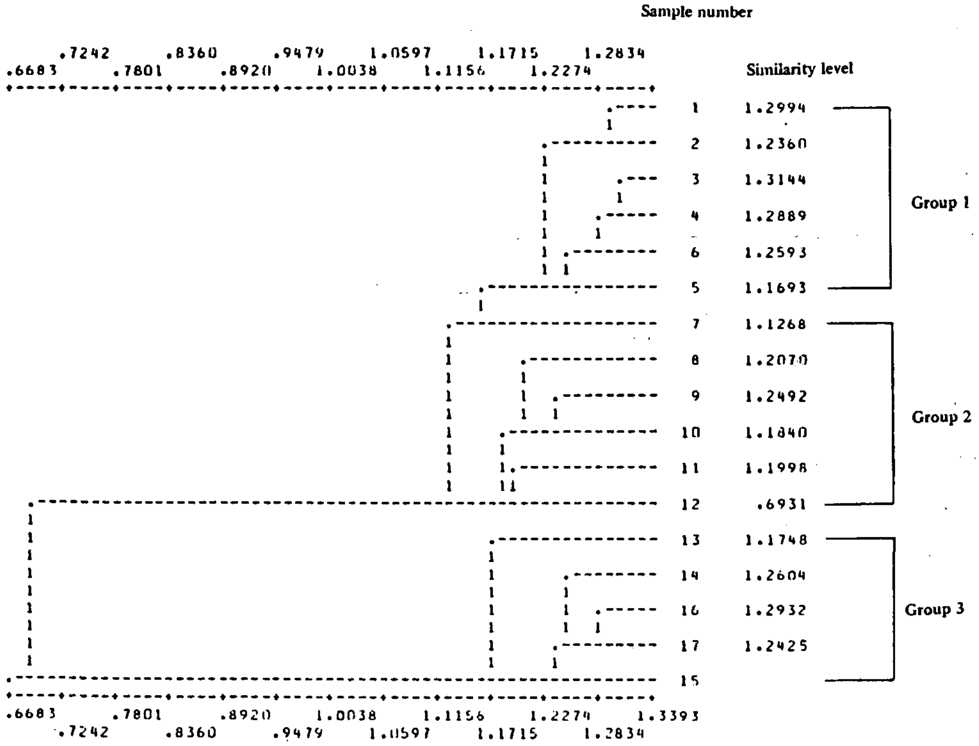

Figure 2
Dendrogram, subdividing the water samples in Table 1, into three major clusters.

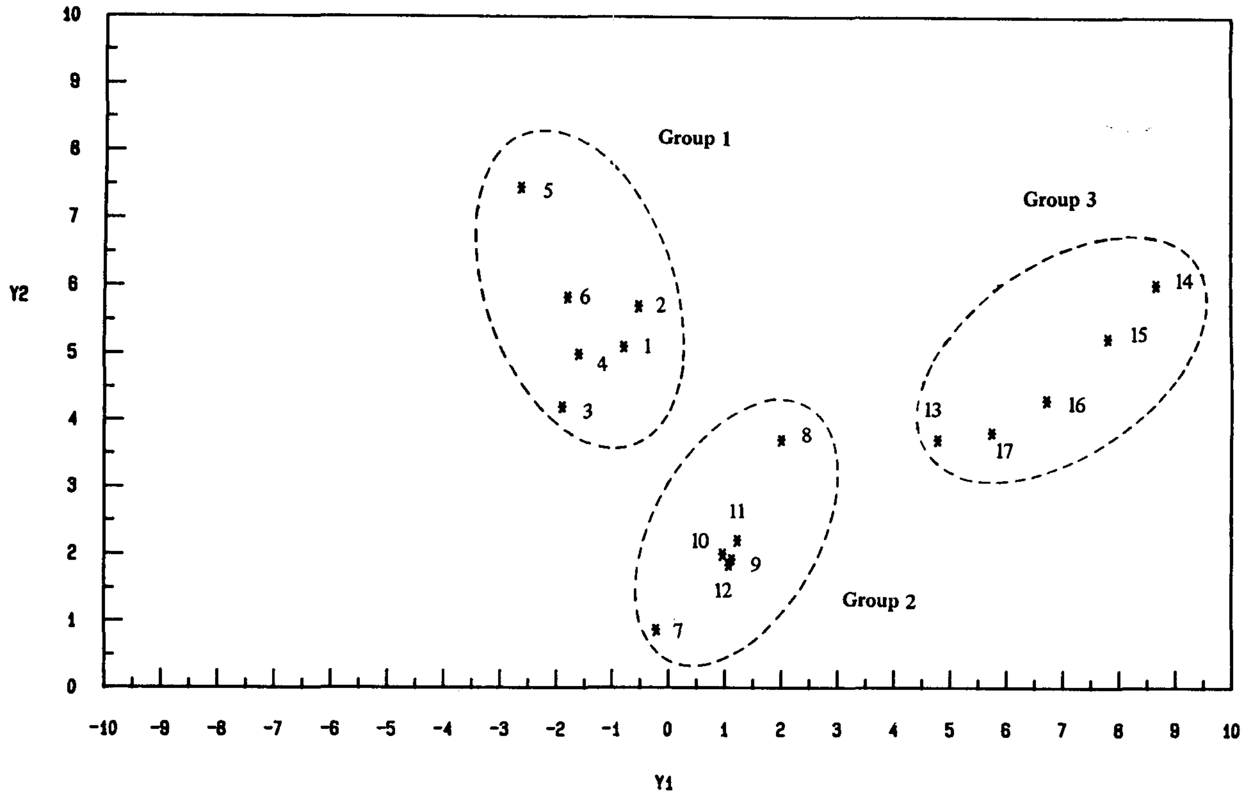

Figure 3
A diagrammatical presentation of the first two principal component scores.

TABLE 3
MATRIX OF PRINCIPAL LOADINGS FOR THE DATA IN TABLE 1

(columns = eigenvectors and rows = variables)

	1	2	3	4	5	6	7	8
1	– ,1458	,5736	– ,0587	– ,0104	– ,6073	– ,2254	– ,4750	– ,0319
2	– ,2380	,5123	– ,1485	– ,2881	,5243	– ,4765	,2700	,0312
3	,5525	,1053	,7532	– ,1876	,0625	– ,2641	– ,0860	– ,0004
4	– ,0015	,0060	,0092	,0243	– ,0323	,0065	– ,0229	,9988
5	– ,0988	,0826	,0204	– ,6183	,2796	,5445	– ,4753	,0088
6	– ,4030	,3812	,5030	,4854	,1693	,4050	,1035	– ,0141
7	– ,0369	,1481	,1045	– ,4600	– ,4797	,2752	,6695	,0073
8	,6658	,4726	– ,3777	,2271	,1200	,3407	,0930	– ,0001

TABLE 4
PRINCIPAL COMPONENT SCORES

(columns = eigenvectors and rows = variables)

	1	2	3	4	5	6	7	8
1	– ,7909	5,1087	3,8777	1,4843	– ,1784	,5302	– ,2089	,0123
2	– ,5311	5,7142	2,2103	– ,1347	1,7432	,4251	– ,0590	– ,0002
3	– 1,9770	4,2273	2,1115	– ,1674	– ,0535	,6855	– ,4642	– ,0307
4	– 1,6075	4,9919	2,5114	– ,4286	– ,1111	,3838	– ,2197	– ,0064
5	– 2,6431	7,4579	3,0376	,3733	,5534	,4715	– ,3162	– ,0485
6	– 1,8019	5,8379	3,4526	– ,2970	– ,0366	,7022	– ,0284	– ,0389
7	– ,2140	,8781	,6383	,1098	,2804	,8952	– ,0536	– ,0122
8	2,0771	3,7216	6,5935	– ,5561	,5082	,8300	– ,2001	– ,0100
9	1,1226	1,9235	3,8181	,2761	,5905	,4078	– ,3948	– ,0143
10	,9635	2,0025	2,9684	,0883	,7325	,5299	– ,4851	– ,0310
11	1,2236	2,2104	3,9248	,7978	,4882	,4822	,0993	– ,0521
12	1,0763	1,8424	3,6570	,3453	,5531	,3702	– ,2880	– ,0327
13	4,7882	3,7003	3,2777	– ,2618	,2558	,5544	– ,2003	– ,0189
14	8,6724	6,0255	2,6216	,3453	,4687	,7900	– ,2689	– ,0221
15	7,8167	5,2343	2,5656	,5623	,5244	,6849	– ,2435	– ,0303
16	6,7196	4,3032	2,6294	– ,0149	,2356	,3786	– ,3019	– ,0215
17	5,7418	3,8145	2,7982	– ,3861	– ,2079	,1712	– ,0446	– ,0197

WATER SA

VOLUME 11
1985

CONTENTS AND AUTHOR INDEX

Supplement to *Water SA* Vol 12 No 1

CONTENTS

VOL 11 • NO 1 • JANUARY 1985

VOL 11 • NO 2 • APRIL 1985

VOL 11 • NO 3 • JULY 1985

VOL 11 • NO 4 • OCTOBER 1985

AUTHOR INDEX

Page numbers in bold refer to sole or first author

Principal components analysis

Table 3 presents the principal axis matrix of the 17 water samples under discussion. The elements of this matrix are the loadings which are used to compute the scores of observations presented in Table 4.

If a transformation of the form $Y_1 = a_1X_1 + a_2X_2 + \ldots + a_8X_8$, where the a's are the element of the first eigenvector in Table 3, is taken, a new set of data which will have a variance of 13,06 (the value of the first eigenvalue) can be created. A similar transform on the second column will yield a series of data points having a variance equal to the value of the second eigenvalue, i.e. 3,26.

If the first two columns of Table 4 are plotted on a $x-y$ plane, a pattern as indicated in Figure 3 is obtained. From studying this figure, it becomes clear that three types of water exist.

Factor analysis

Q-mode factor analysis

The rotated factor matrix having 17 rows and 2 columns for the water samples under discussion is given in Table 5. Only two factors have been extracted, because the first two eigenvalues account for 91,8 per cent of the total variance. Figure 4 shows a plot of factor one against factor two. The existence of three groups of groundwater can again be seen from this figure.

TABLE 5
ROTATED FACTOR MATRIX

(columns = factors and rows = observations)

	1	2
1	,0911	,3600
2	,8717	,3465
3	,9810	,0817
4	,9553	,1895
5	,9607	,1416
6	,9610	,2243
7	,6211	,6594
8	,6259	,7076
9	,6187	,6978
10	,6603	,6997
11	,6252	,7004
12	,6180	,6983
13	,2911	,9515
14	,1435	,9497
15	,1407	,9571
16	,1414	,9694
17	,1754	,9661

R-mode factor analysis

Three factors will be retained because three eigenvalues greater than one exist. Table 6 shows the rotated factor matrix. From this it can be seen that the loadings on variables calcium, magnesium and sulphate are high on factor 1 and low on the other two factors. This could lead to the interpretation that the "new" variable, factor 1, accounts for calcium, magnesium and sulphate; whereas factor 2 accounts for sodium and chloride and factor 3 for potassium.

The columns of Table 7 show the distribution of the three new variables at the 17 location points.

Discriminant functions

For this representation, randomly selected water samples from various areas in South Africa were taken and classified in four groups with stepwise discriminant function analysis. This resulted in the following equations:

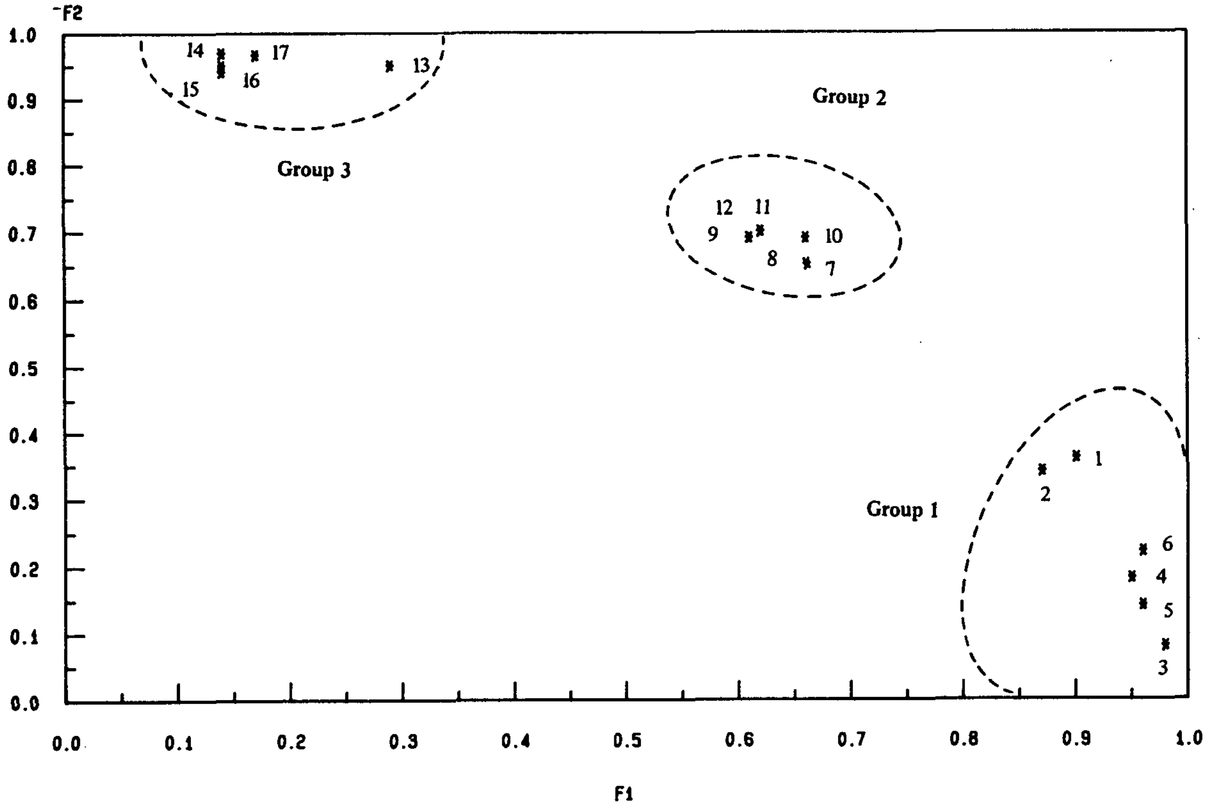

Figure 4
Plot of factor 1 against factor 2 of information in Table 5.

$Ca/Mg(HCO_3)_2$-water

$= 0{,}70604\ (Ca + Mg) + 0{,}28934\ (Cl + SO_4) - 32{,}1272$

$NaHCO_3$-water

$= 0{,}25984\ (Ca + Mg) + 0{,}26131\ (Cl + SO_4) - 8{,}40822$

NaCl-water

$= 0{,}31169\ (Ca + Mg) + 0{,}66534\ (Cl + SO_4) - 31{,}21053$

$Ca/Mg(SO_4/Cl_2)$-water

$= 0{,}82848\ (Ca + Mg) + 0{,}77201\ (Cl + SO_4) - 62{,}4263$

The (Ca + Mg) and ($Cl + SO_4$)values of a water sample (in percentage meq/ℓ) of the cations and anions can now be substituted in each of the four equations. The sample is then classified into the group which yields the highest score. Table 8 presents the discriminant scores for the 17 water samples for the Venterstad area. A water sample belongs to the type of water which yields the highest score in each row. From this table it is evident that samples 1 to 6 have a $Ca/Mg(HCO_3)_2$-character, samples 7 to 12 a $NaHCO_3$-character and samples 13 to 17 a NaCl-character.

Conclusions

Interpretation of the chemical analyses of water samples from the Venterstad area shows that multivariate statistical methods can be used in much the same way as the trilinear diagram. The added advantage of these methods is that they not only depend on the major ion distribution, as in the case with the trilinear diagrams, but could include other ions and non-chemical data. Fewer or more variables can be considered than in the trilinear method.

TABLE 6
MATRIX OF ROTATED FACTOR LOADINGS

	(columns = factors; rows = observations) 1	2	3
Ca^{++}	,8776	– ,1347	– ,3370
Mg^{++}	,8281	– ,4296	– ,0102
Na^+	– ,1847	,8957	– ,0626
K^+	,2017	– ,0311	– ,9582
$CO_3^=$	,6848	– ,5315	,3283
HCO_3^-	,5499	– ,6743	– ,3649
$SO_4^=$	,8339	,0318	– ,1913
Cl^-	– ,0409	,9520	,0675

TABLE 7
A VARIMAX ROTATED FACTOR SCORE MATRIX SHOWING THE DISTRIBUTION OF THE 3 FACTORS ALONG THE 17 WATER SAMPLE LOCALITIES

1	– ,3964	– ,5857	– 3,4887
2	,8772	– ,6853	1,2951
3	,6875	– ,8899	,2250
4	1,1382	– ,5251	– ,3408
5	1,7752	– ,7207	,0034
6	1,7289	– ,3753	,1085
7	– 1,2163	– 1,0562	,9126
8	,5064	,2156	– ,3895
9	– 1,1932	– ,6443	– ,1890
10	– ,9730	– ,7300	,8533
11	– 1,1761	– ,4510	,1447
12	– 1,2048	– ,6291	,2380
13	– ,0936	,8325	,3074
14	,1698	1,9971	– ,0385
15	– ,3098	1,6059	,0641
16	– ,2775	1,3436	,1882
17	– ,0375	1,2978	,1063

TABLE 8
DISCRIMINANT FUNCTIONS SCORES RESULTED FROM THE FOUR EQUATIONS. A WATER SAMPLE BELONGS TO THE TYPE OF WATER WHICH YIELDS THE HIGHEST DISCRIMINANT SCORE IN EACH ROW. THESE VALUES ARE UNDERLINED

Sample	$Ca/Mg(HCO_3)_2$	$NaHCO_3$	NaCl	$Ca/Mg(SO_4/Cl_2)$
1	$\underline{17{,}5}$	13,4	2,9	5,8
2	$\underline{28{,}3}$	18,4	11,3	21,3
3	$\underline{32{,}0}$	17,6	5,6	19,8
4	$\underline{29{,}7}$	18,2	9,4	21,0
5	$\underline{33{,}4}$	18,9	9,0	23,7
6	$\underline{27{,}1}$	17,2	8,4	18,0
7	– 20,0	$\underline{-1{,}7}$	– 12,1	– 37,1
8	– 10,9	$\underline{3{,}8}$	– 6,4	– 25,1
9	– 18,8	$\underline{-1{,}2}$	– 17,4	– 40,4
10	– 10,9	$\underline{2{,}3}$	– 11,6	– 29,3
11	– 16,1	$\underline{1{,}5}$	– 9,9	– 32,7
12	– 18,7	$\underline{-1{,}0}$	– 16,6	– 39,7
13	1,5	15,1	$\underline{22{,}3}$	8,2
14	9,7	21,1	$\underline{36{,}3}$	26,1
15	7,3	20,1	$\underline{35{,}1}$	22,9
16	6,9	20,1	$\underline{35{,}1}$	24,5
17	8,1	20,6	$\underline{35{,}8}$	24,5

References

AGTERBERG, F.P. (1974) *Geomathematics.* Elsevier Sc. Publishing Company, New York.

DALTON, M.G. and UPCHURCH, S.B. (1978) Interpretation of hydrochemical facies by factor analysis. *Ground Water* **16**(4).

DAVIS, J.C. (1973) *Statistics and data analysis in geology.* John Wiley and Sons, New York.

HARTIGAN, J.A. (1972) *Clustering algorithms.* John Wiley and Sons, New York.

JÖRESKOG, K.G., KLOVAN, J.E. and REYMENT, R.A. (1976) *Geological factor analysis.* Elsevier Sc. Publishing Company, New York.

KLOVAN, J.E. (1975) R- and Q-mode factor analysis. In R.B. McCam (Ed.) *Concepts in Geostatistics.* Springer-Verlag, New York.

LAWRENCE, F.W. and UPCHURCH, S.B. (1976) Identification of geochemical patterns in ground water by numerical analysis. Proc. 12th Ann. Symp., Advances in Ground Water Hydrology, Amer. Water Resour. Assn.

LAWRENCE, F.W. and UPCHURCH, S.B. (1982) Identification of recharge areas using geochemical factor analysis. *Ground Water* **20**(6).

MORRISON, D.F. (1976) Multivariate statistical methods. 2nd Edition. McGraw-Hill, Inc., New York.

PIPER, A.M. (1944) A graphic procedure in the geochemical interpretation of water analyses. *Trans. Amer. Geophys. Un.* **25**.

VAN DER LINDE, P.J. and HODGSON, F.D.I. (1977) Die gebruik van hidrochemiese tegnieke in grondwatereksplorasie in die Venterstad-omgewing, Kaapprovinsie. *Water SA* **3**(3) 166-171.

ZAPOROSEC, A. (1972) Graphical interpretation of water-quality data. *Ground Water* **10**(2).