Mathematical Geology, Vol. 15, No. 5, 1983

Dual Extraction of *R*-Mode and *Q*-Mode Factor Solutions[1]

Di Zhou,[2] Theodore Chang,[3] and John C. Davis[4]

It is mathematically possible to extract both R-*mode and* Q-*mode factors simultaneously* (RQ-*mode factor analysis*) *by invoking the Eckhart-Young theorem. The resulting factors will be expressed in measures determined by the form of the scalings that have been applied to the original data matrix. Unless the measures for both solutions are meaningful for the problem at hand, the factor results may be misleading or uninterpretable. Correspondence analysis uses a symmetrical scaling of both rows and columns to achieve measures of proportional similarity between objects and variables. In the literature, the resulting similarity is a* χ^2 *distance appropriate for analysis of enumerated data, the original application of correspondence analysis. Justification for the use of this measure with interval or ratio data is unconvincing, but a minor modification of the scaling procedure yields the profile similarity, which is an appropriate measure. Symmetrical scaling of rows and columns is unnecessary for* RQ-*mode factor analysis. If the data are scaled so the minor product* W'W *is the correlation matrix, the major product* WW' *is expressed in the Euclidean distances between objects. Therefore,* RQ-*mode factor analysis can be performed so that the* R *mode is a principal components solution and the* Q *mode is a principal coordinates solution. For applications where the magnitudes of differences are important, this approach will yield more interpretable results than will correspondence analysis.*

KEY WORDS: scaling, correspondence analysis, principal components analysis, *RQ*-mode factor analysis.

INTRODUCTION

Correspondence analysis, since its introduction in geology by Teil (1975) and David, Dagbert, and Beauchemin (1977), has been used in a variety of applications, particularly in geochemistry. The method is appealing because it makes possible the simultaneous extraction of both *R*-mode and *Q*-mode factors. The *R*- and *Q*-mode solutions are displayed on the same set of diagrams, greatly facilitating the interpretation of the factors. Unfortunately, few people have yet been concerned about the limitations of the applicability of correspondence

[1] Manuscript received 25 February 1982; revised 17 August 1982.

[2] Wuhan College of Geology, Wuhan, Hubei, People's Republic of China. Present address: Kansas Geological Survey, University of Kansas, Lawrence, Kansas 66044 U.S.A.

[3] Department of Mathematics, University of Kansas, Lawrence, Kansas 66044 U.S.A.

[4] Kansas Geological Survey, University of Kansas, Lawrence, Kansas 66044 U.S.A.

0020-5958/83/1000-0581$03.00/0

analysis, or have sought other techniques that have the same desirable features but which perform better for certain types of problems.

Simultaneous extraction of R- and Q-mode factors, which may be called *RQ-mode factor analysis*, is not unique to correspondence analysis. It can be achieved in many ways, provided that a general constraint on the scaling of the data is satisfied. This paper discusses the general requirements for RQ-mode analysis, reviews the conditions under which various alternative procedures are appropriate, and suggests the consideration of principal components analysis as an RQ-mode procedure.

The term "factor analysis," as used here, includes the many versions of components analysis. "Principal components analysis" refers to the specific procedure proposed first by Hotelling (1933). Varimax and other rotational schemes are not considered, because at present they have not been incorporated into RQ-mode procedures and discussion of their possible use is beyond the scope of this paper. The notation, unless otherwise specified, is given in Table 1, where the definitions of factor loadings and factor scores follow the usage of Klovan and Imbrie (1971). The word "similarity" is used in a broad sense to

Table 1. Mathematical Notation Used in This Paper[a]

X	raw data matrix ($n \times m$)
W	scaled data matrix ($n \times m$)
n	number of objects
m	number of variables
p	number of factors extracted [$p \leqslant \min(n, m)$]
Λ	diagonal matrix of eigenvalues for $W'W$ or WW', listed in descending order
U	unit-orthogonal eigenvector matrix ($m \times p$) of $W'W$
V	unit-orthogonal eigenvector matrix ($n \times p$) of WW'
A^R	R-mode factor loading matrix ($m \times p$), $A^R = U\Lambda^{1/2}$
A^Q	Q-mode factor loading matrix ($n \times p$), $A^Q = V\Lambda^{1/2}$
F^R	R-mode factor score matrix ($n \times p$), $F^R = WA^R\Lambda^{-1}$
F^Q	Q-mode factor score matrix ($m \times p$), $F^Q = W'A^Q\Lambda^{-1}$
$\bar{x}_j$	average value of variable j, $\bar{x}_j = \frac{1}{n}\sum_i x_{ij}$
s_j	standard deviation of variable j, $s_j = \left[\frac{1}{n}\sum_i (x_{ij} - \bar{x}_j)^2\right]^{1/2}$

[a] A lower case character indicates an element of the matrix represented by the corresponding upper case character.

indicate measures of the relationships between variables or objects, including measures of dissimilarity.

THE IMPORTANCE OF SCALING

As pointed out by Miesch (1980), the alternative versions of factor analysis differ primarily in the way in which the data are scaled prior to factoring. Scaling determines the measure of similarity and hence the nature of the factor solution.

This point can be seen more clearly through a geometrical interpretation of factor analysis. In R-mode, the scaled variables can be regarded as points or vectors in object space; the configuration of these points or vectors is a graphical representation of the similarities between variables. Factoring essentially is a rotation of the reference axes under some constraint such as accounting for the maximum variance in the data. After factoring, while the configuration of variable vectors remains unchanged, this configuration usually can be seen more clearly in a reduced dimensionality (Fig. 1). The preservation of variable configuration during R-mode factoring can be expressed algebraically by the equality

$$W'W = A^R A^{R'} \tag{1}$$

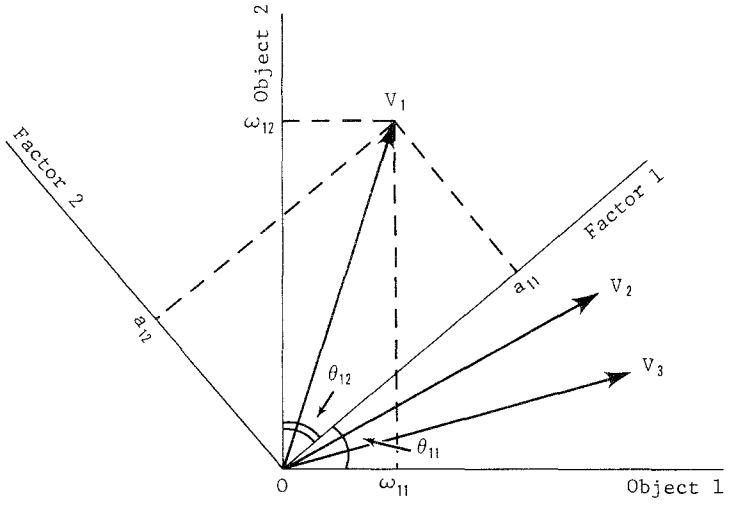


Fig. 1. Schematic representation of R-mode factor analysis. Configuration of scaled variable vectors V_1, V_2, and V_3 remain invariant during factoring. Coordinates of V_1 are ω_{11} and ω_{12} with respect to object axes or a_{11} and a_{12} with respect to factor axes. Direction cosines of factor axes in object space are given by score matrix F^R. For example, $\cos\theta_{11} = f_{11}$ and $\cos\theta_{12} = f_{12}$.

Table 2. Artificial Data Used to Show the Effects of Scaling

Data matrix X				Data matrix Y				
	I	II	III		1	2	3	4
1	4	16	16	I	4	1	2	1
2	1	4	13	II	16	4	8	4
3	2	8	14	III	16	13	14	13
4	1	4	13					

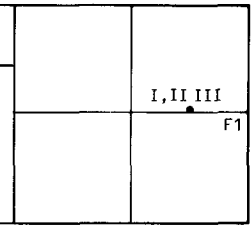
The equality indicates that similarities between variables, represented by the matrix $W'W$, are inherent in the dot product matrix of factor loadings, if no factors have been eliminated. Similar relationships hold for Q-mode factor analysis, in which

$$WW' = A^Q A^{Q'} \tag{2}$$

A. Principal components analysis of matrix X

$$W_{ij} = \frac{x_{ij} - \bar{x}_i}{s_j\sqrt{n}}$$

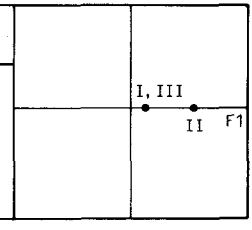
Scaling	Correlation coefficient			A^R F1
	I	II	III	
I	1.0			1.0
II	1.0	1.0		1.0
III	1.0	1.0	1.0	1.0



B. Principal components analysis of matrix X

$$W_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{n}}$$

Scaling	Variance-covariance			A^R F1
	I	II	III	
I	2.0			1.4
II	8.0	32.0		5.7
III	2.0	8.0	2.0	1.4



C. Principal coordinates analysis of matrix Y

$$W_{ij} = \frac{y_{ij} - \bar{y}_j}{s_j\sqrt{n}}$$

Scaling	Euclidean distance of standardized data			A^Q F1	F2
	I	II	III		
I	0.0			-1.3	0.3
II	1.5	0.0		-0.1	-0.5
III	2.7	1.6	0.0	1.4	0.2

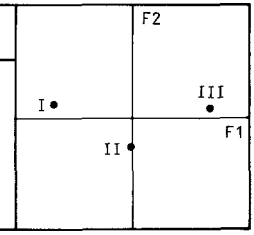


Fig. 2. Factor analysis of artificial data given in Table 2.

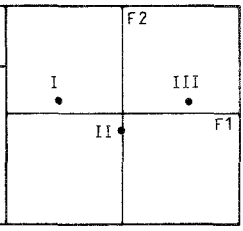
Therefore, the manner in which the data matrix X is scaled in order to produce W determines the nature of the similarity measure in the factor solution.

Two sets of artificial data (Table 2), in which Y is the transpose of X, can be used to show the effects of scaling. The data are constructed so that vector II is derived by multiplying vector I by 4, and vector III is derived by adding 12 to vector I. We are interested in the effects of scaling on the three vectors, I, II, and III; hence, all R-mode analyses are performed on the data matrix X, and all Q-mode analyses on matrix Y.

Figure 2 shows various scaling equations, the resulting similarity matrices, and the corresponding factor solutions. Scalings which involve division only (E and F) result in scale-invariant measures of similarity, and vectors I and II coincide in the factor diagram. Measures of this type, such as the cos θ and profile distance, are proportional similarities and are insensitive to differences in magnitude. Scaling involving subtraction only (B) results in a location-invariant measure of similarity and the coincidence of vectors I and III in the factor diagram. Scaling involving both division and subtraction (A) yields a scale- and

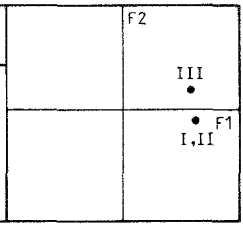
D. Principal coordinates analysis of matrix Y

Scaling	Euclidean distance of raw data			A^Q F1	F2
	I	II	III		
$w_{ij} = \frac{y_{ij} - \bar{y}_j}{\sqrt{n}}$ I	0.0			-8.5	1.7
II	10.0	0.0		0.0	-3.5
III	17.0	10.0	0.0	8.5	1.7



E. Imbrie Q-mode factor analysis of matrix Y

Scaling	Cosine theta			A^Q F1	F2
	I	II	III		
$w_{ij} = \frac{y_{ij}}{\sqrt{\sum_k y_{ik}^2}}$ I	1.0			1.0	-0.2
II	1.0	1.0		1.0	-0.2
III	0.9	0.9	1.0	0.9	0.3



F. Correspondence analysis of matrix X

Scaling	Profile distance			G F1
	I	II	III	
$w_{ij} = \frac{x_{ij}}{\sqrt{\sum_k x_{ik}}\sqrt{\sum_l x_{lj}}}$ I	0.0			-0.3
II	0.0	0.0		-0.3
III	0.5	0.5	0.0	0.2

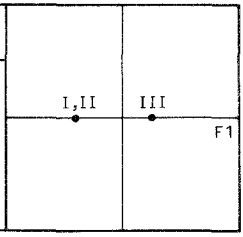


Fig. 2. Continued

location-invariant measure of similarity, such as the correlation coefficient. Vectors I, II, and III all coincide in the resulting factor diagram. In C and D of Fig. 2, scaling is not performed along the three vectors, I, II, and III, but rather along vectors 1, 2, 3, and 4. The resulting similarity measure for vectors I, II, and III is the Euclidean distance and the three are distinct in the factor diagram.

From this artificial example it can be appreciated that a scaling procedure should be selected which provides a measure of similarity relevant to the problem being investigated. In many geological applications the correlation coefficient, and more rarely the covariance, are appropriate measures of the similarity between variables. Therefore, principal components analysis is usually a satisfactory *R*-mode procedure.

Q-mode analysis, however, is more complicated. For some applications, such as the "end member" problem (Jöreskog, Klovan, and Reyment, 1976, p. 86-100), proportional similarities are required because it is the proportions rather than the magnitudes of constituents which indicate the source of an observation. In this situation, the cos θ coefficient provides a good measure and Imbrie's *Q*-mode vector analysis may be preformed. However, in other applications the magnitudes of the constituents are important. For example, in a study of trace elements in rocks, a specimen containing 100 ppm of Cu and 500 ppm of Pb is quite different than another containing 10 ppm of Cu and 50 ppm of Pb in terms of their significance as guides to mineralization, even though the elements occur in the same proportions. In such an application, a distance measure should be used, and principal coordinate analysis is a suitable *Q*-mode procedure.

If a scaling procedure results in $W'W$ (or its specific transformation) being a similarity matrix of variables which is appropriate for the problem at hand, and at the same time WW' (or its specific transformation) is a meaningful measure of similarity between the objects, then *RQ*-mode factor analysis can be achieved. Although it is always computationally possible to calculate a *Q*-mode solution from an *R*-mode solution, and vice versa, by invoking the Eckart-Young theorem (Eckart and Young, 1936; Johnson, 1963), the dual solutions may not be meaningful unless this requirement is met. Then, the *R*-mode and *Q*-mode solutions are related in the following manner

$$W = \overbrace{V}^{F^R}\ \overbrace{\Lambda^{1/2}\ U'}^{A^R} \qquad R\text{-mode solution}$$

$$W = \underbrace{V\ \Lambda^{1/2}}_{A^Q}\ \underbrace{U'}_{F^Q} \qquad Q\text{-mode solution} \qquad (3)$$

CORRESPONDENCE ANALYSIS AS AN *RQ*-MODE PROCEDURE

Correspondence analysis was first proposed by Benzecri (1969) for the factor analysis of contingency tables in which the elements represent frequen-

cies of occurrences. Hill (1974) equated the method to Fisher's canonical analysis and other earlier procedures. This paper focuses on the common geological practice of applying correspondence analysis to measurement data in which entries in the data matrix represent interval- or ratio-scale measurements of variables on objects.

The rationale of correspondence analysis is based on conditional probabilities. The distributional distance, or χ^2 distance, is used as a similarity measure for both rows and columns (Benzecri, 1969; Benzecri *et al.*, 1980; Teil, 1975; David, Dagbert, and Beauchemin, 1977; Jambu, 1980). Such a treatment is reasonable for contingency table data, because a contingency table divided by its gross sum can be regarded as a table of estimates of probabilities. Also, a contingency table is essentially a double classification of one set of objects and hence the symmetrical treatment of rows and columns is logical.

A table of measurements, however, has very different properties. Although a table containing nonnegative measurements can be divided by its gross sum, entries in the resulting transformed table can rarely be regarded as estimates of probabilities. Therefore, a rationale based on conditional probability, such as the one given by David, Dagbert, and Beauchemin (1977), is not necessarily valid. In addition, the rows and columns of a table of measurements are not mutually symmetrical, so there is no reason to presume that a symmetrical similarity measure must be used. Therefore, a symmetrical scaling procedure is not necessary for *RQ*-mode factor analysis of measurement data.

This does not mean that correspondence analysis cannot be applied to measurement tables. It is possible to justify the use of correspondence analysis without invoking conditional probabilities, by the use of the "profile distance" measure (Appendix, I). The profile distance corresponds to a scaling involving division only, and hence is a kind of proportional similarity and is not sensitive to differences in magnitude. The measure is scale invariant but not location invariant, and differs from either the correlation coefficient or the Euclidean distance. In order to determine if correspondence analysis is appropriate for a given application, it is necessary to determine whether the profile distance is a suitable measure of the similarities between variables and between objects in the problem.

PRINCIPAL COMPONENTS ANALYSIS AS AN *RQ*-MODE PROCEDURE

Often the similarity between objects is more relevantly expressed by a distance measure than by proportional similarity. Let $H = WW'$ be a matrix of dot products of object vectors in variable space; the Euclidean distance between object points i and l is then determined by

$$d_{il}^2 = h_{ii} + h_{ll} - 2h_{il} \tag{4}$$

When W is formed by

$$w_{ij} = \frac{1}{n^{1/2}}\left(\frac{x_{ij} - \bar{x}_j}{s_j}\right), \qquad \text{for} \quad \begin{matrix} i = 1, 2, \cdots, n \\ j = 1, 2, \cdots, m \end{matrix} \tag{5}$$

it can be shown algebraically that d_{il} is $1/n^{1/2}$ times the Euclidean distance of the standardized data

$$d_{il}^2 = \frac{1}{n}\sum_j \left(\frac{x_{ij} - x_{lj}}{s_j}\right)^2 \tag{6}$$

If

$$w_{ij} = \frac{1}{n^{1/2}}\,(x_{ij} - \bar{x}_j), \qquad \text{for} \quad \begin{matrix} i = 1, 2, \cdots, n \\ j = 1, 2, \cdots, m \end{matrix} \tag{7}$$

d_{il} is $1/n^{1/2}$ times the Euclidean distance measured on the raw data

$$d_{il}^2 = \frac{1}{n}\sum_j (x_{ij} - x_{lj})^2 \tag{8}$$

In both instances, $H = WW'$ is the matrix that Gower (1967) defined for principal coordinates analysis

$$h_{il} = e_{il} - e_{i\cdot} - e_{\cdot l} + e_{\cdot\cdot}, \qquad \text{for} \quad \begin{matrix} i = 1, 2, \cdots, n \\ l = 1, 2, \cdots, n \end{matrix} \tag{9}$$

where

$$e_{\cdot l} = -\tfrac{1}{2}\, d_{il}^2$$

$$e_{i\cdot} = \frac{1}{n}\sum_l e_{il}$$

$$e_{\cdot l} = \frac{1}{n}\sum_i e_{il}, \qquad \text{and}$$

$$e_{\cdot\cdot} = \left(\frac{1}{n}\right)^2 \sum_i \sum_l e_{il} \quad .$$

(cf. Appendix, II). However, it is also true that the scaling in (5) and (7) are the same as those used in principal components analysis. This means that the scaling in (5) can give $W'W$, the correlation matrix of variables, and at the same time $WW' = H$, in which the Euclidean distances of objects are embedded. Therefore, this scaling procedure meets the requirements of RQ-factor analysis. The same conclusions can be drawn for the scaling procedure in (7).

Because of these relationships, principal components analysis can be used as an RQ-mode factor technique. All that is necessary is to add steps for calculating A^Q and F^Q by

$$A^Q = WU, \qquad \text{and} \tag{10}$$

$$F^Q = U \tag{11}$$

Since A^R gives the coordinates of variable points in factor space and A^Q gives the coordinates of object points in factor space, A^R and A^Q can be plotted with respect to the factor axes on the same set of diagrams. The similarities between variables and the similarities between objects are represented by the configurations of their points. The associations between variables and objects can be examined in the following way. Equation (3) can be rewritten as

$$A^Q \Lambda^{-1/2} A^{R'} = W$$

That is,

$$\frac{1}{(\lambda_k)^{1/2}} \sum_k a_{ik}^Q a_{jk}^R = w_{ij}, \qquad \begin{matrix} i = 1, 2, \cdots, n \\ j = 1, 2, \cdots, m \end{matrix} \tag{12}$$

The left-hand side of (12) can be regarded as the dot product of the object vector $\mathbf{a}_i^Q$ and the variable vector $\mathbf{a}_j^R$ with a factor $\lambda_k^{-1/2}$. The magnitude of the product, w_{ij}, is inversely related to the distance between the object point i and the variable point j. Thus, a cluster of object points can be interpreted as being characterized by the nearest variable points.

The duality between principal components analysis and principal coordinates analysis using the Euclidean distance was first pointed out by Gower (1966). Unfortunately, this duality has not been utilized for RQ-mode factor analysis until now. Gabriel (1971) did use a biplot graphic display of A^R and F^R to study the associations of objects to variables. Because $A^Q A^{Q'}$ preserves the similarities of objects and $F^R F^{R'}$ does not, it seems more reasonable to plot A^Q instead of F^R for displaying the interrelationships between objects.

For problems in which the similarities between variables can best be represented by the correlation coefficient or the covariance, and the similarities between objects by the Euclidean distance, principal components analysis is more relevant than correspondence analysis. In the literature, correspondence analysis commonly is applied to petrochemical data. For these data the profile distances between objects are close to their Euclidean distances because the rows of the data matrix sum to a constant. The differences between the results obtained by the two procedures are obscured. If, however, the data have variable row sums or column sums, the two procedures may produce remarkably different results.

APPLICATIONS OF *RQ*-MODE FACTOR PROCEDURES

Two examples are presented to show the differences between correspondence analysis and principal components analysis as an *RQ*-mode factor procedure. The first example was selected because the data have been extensively used to demonstrate a variety of multivariate techniques (Imbrie, 1963; Krumbein and Imbrie, 1963; Manson and Imbrie, 1964; Lee, 1969; Howarth, 1973; David, Dagbert, and Beauchemin, 1977). The second example was chosen to show that principal components analysis yields a more geologically meaningful result than correspondence analysis when the interrelationships between objects are best represented by the Euclidean distance and similarities between variables by the correlation coefficient.

W. T. Fox compiled the thicknesses of lithologic components in an Upper Permian stratigraphic unit as measured in 31 wells in southeastern Colorado and western Kansas, U.S.A. The data include the thicknesses (in feet) of sand, shale, carbonate, and evaporite (Fig. 3). Two derived variables, total thickness, and nonclastic thickness (carbonate plus evaporite), were added by Krumbein and Imbrie (1963).

Correspondence analysis was applied to these data by David, Dagbert, and

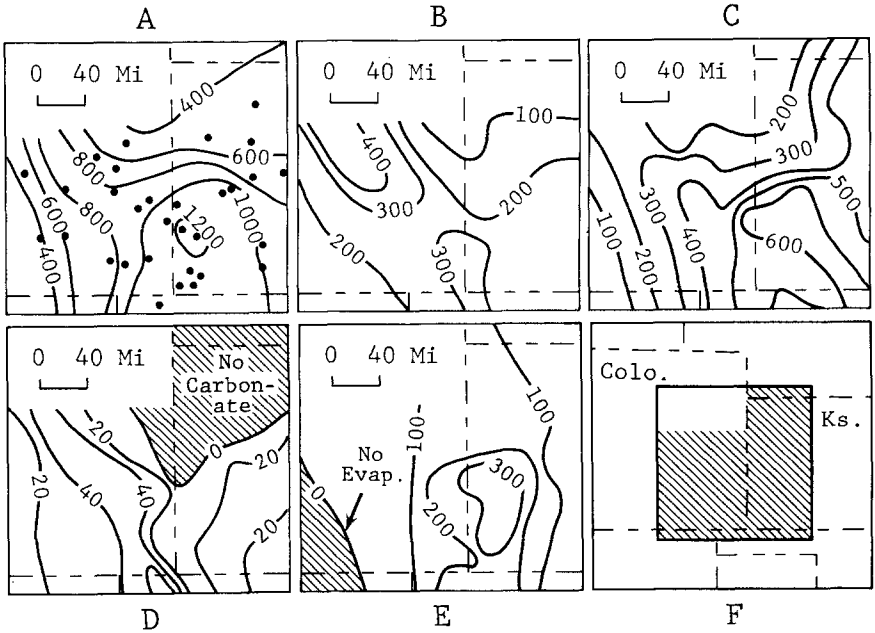


Fig. 3. Isopach and net thickness maps of four-component Upper Permian stratigraphic unit in Colorado and Kansas (from Krumbein, 1962). (A) Total thickness; (B) feet of sandstone; (C) feet of shale; (D) feet of carbonates; (E) feet of evaporites; (F) index map.

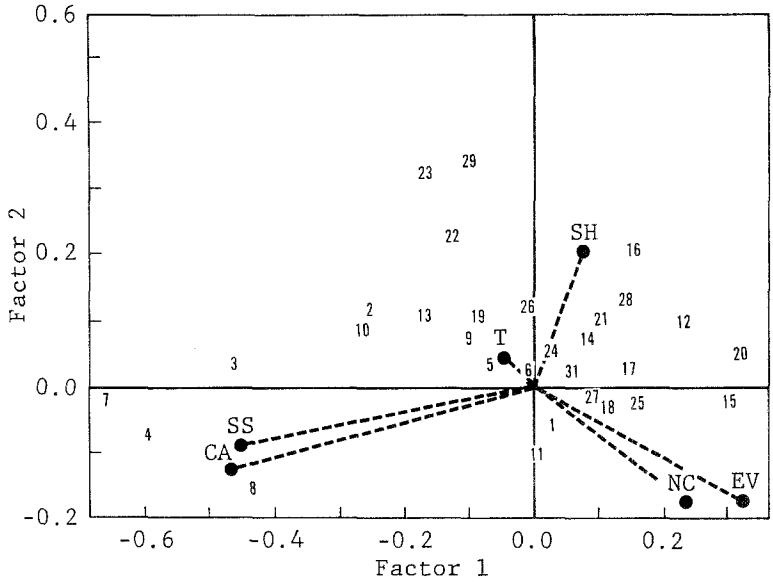


Fig. 4. Correspondence factors 1 and 2 from Fox data (after David, Dagbert, and Beauchemin, 1977). Numbers have been changed to correspond with Table 3.

Beauchemin (1977). Only three factors are extracted because the rank of the data matrix is reduced by one by the effect of closure. Their results are shown in Figs. 4 and 5. *RQ*-mode principal components analysis was applied to these same data, with the results shown in Table 3 and Figs. 6 through 8.

The two methods yield different configurations of variables (thicknesses of lithologies) and of objects (wells). Among the variables, the total thickness shows no association with any other variable in correspondence analysis, but does show strong correlations with shale, nonclastic, and evaporite thickness in principal components analysis. These results are similar to the relationship between I and III in the artificial example (Fig. 2A, F), because total thickness is the sum of four other variables. The relationships between other variables are similar in the two procedures, if the closure effect in correspondence analysis is taken into account.

The differences in the configuration of objects are more notable. In principal components analysis, the objects can be visually divided into seven clusters in the diagram of factors 1 and 2 (Fig. 6), which explains 83.4% of the total variance. The clusters do not appear in the plot of factors 3 and 4 (Fig. 7). Groups I to IV are differentiated along factor 1 by a decrease of shale and evaporite thicknesses. Group VII has the thinnest total section of sediments and especially of sand and carbonate. Group V is characterized by the thickest sand and carbonate. Group VI, containing only well 31, is unique as it has the thick-

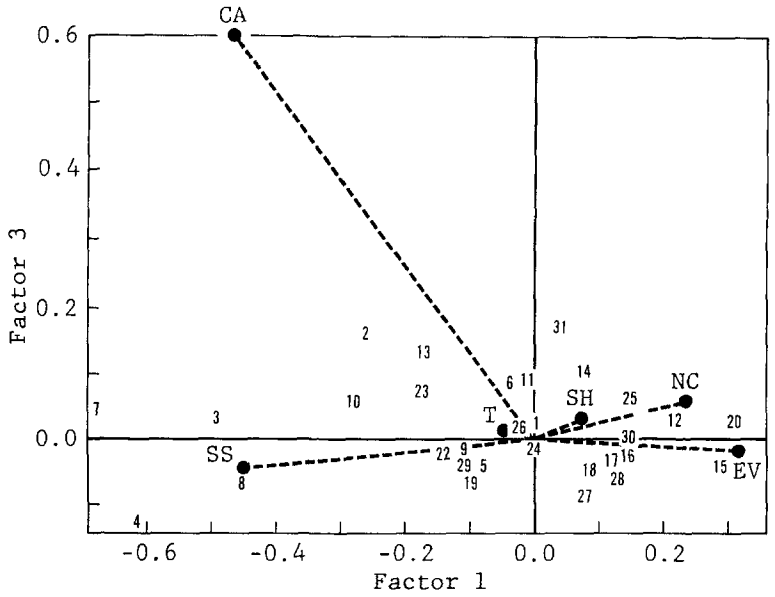


Fig. 5. Correspondence factors 1 and 3 from Fox data (after David, Dagbert and Beauchemin, 1977). Numbers have been changed to correspond with Table 3.

Table 3. *RQ*-Mode Principal Components Analysis of Fox Data. (A) Matrix of Correlations Between Variables (*T* = total thickness; *SS* = thickness of sandstone; *NC* = thickness of nonclastics; *CA* = thickness of carbonates; *EV* = thickness of evaporites). (B) Eigenvalues of Correlation Matrix (λ = eigenvalue; % = percent of trace; Σ% = cumulative percent of trace). (C) *R*-Mode Factor Loadings. (D) *Q*-Mode Factor Loadings.

(A) Correlation Matrix

	T	*SS*	*SH*	*NC*	*CA*	*EV*
T	1.00	–	–	–	–	–
SS	0.24	1.00	–	–	–	–
SH	0.89	-0.12	1.00	–	–	–
NC	0.84	-0.03	0.69	1.00	–	–
CA	0.15	0.46	-0.05	0.06	1.00	–
EV	0.82	-0.11	0.70	0.99	-0.10	1.00

(B) Eigenvalues

	Factors			
	1	2	3	4
λ	3.47	1.54	0.56	0.44
%	57.76	25.66	9.33	7.26
Σ%	57.76	83.42	92.74	100.00

Table 3. Continued

(C) R-Mode Loadings

	Factors			
	1	2	3	4
T	0.96	0.22	−0.13	−0.16
SS	−0.00	0.87	−0.49	0.08
SH	0.87	−0.10	−0.01	−0.47
NC	0.95	−0.01	0.10	0.30
CA	0.02	0.84	0.54	−0.04
EV	0.94	−0.14	0.02	0.30

(D) Q-Mode Loadings

	Factors			
	1	2	3	4
1	0.03	0.04	0.01	0.14
2	−0.10	0.44	0.16	−0.12
3	−0.27	0.49	−0.15	−0.05
4	−0.57	0.00	−0.13	0.11
5	0.10	0.13	−0.19	0.02
6	0.08	0.19	0.12	0.02
7	−0.63	0.01	0.05	0.08
8	−0.40	0.18	−0.17	0.16
9	−0.37	−0.09	−0.02	0.03
10	−0.35	0.06	0.04	0.01
11	0.11	0.23	0.12	0.16
12	0.45	−0.13	0.08	−0.02
13	−0.22	0.12	0.16	−0.03
14	−0.36	−0.26	0.20	0.08
15	0.51	−0.28	−0.06	0.16
16	0.45	−0.14	−0.06	−0.20
17	0.47	−0.04	−0.14	0.03
18	0.39	0.02	−0.16	0.11
19	0.10	0.12	−0.26	−0.07
20	0.57	−0.26	0.11	0.05
21	0.21	−0.04	0.02	−0.04
22	−0.22	−0.10	−0.06	−0.09
23	−0.21	0.02	0.06	−0.21
24	0.18	0.05	−0.10	−0.00
25	0.41	0.07	0.10	0.11
26	0.07	0.05	−0.02	−0.07
27	−0.29	−0.32	−0.02	0.14
28	−0.26	−0.37	0.03	0.05
29	0.02	−0.05	−0.19	−0.30
30	−0.35	−0.51	0.13	−0.22
31	0.33	0.39	0.34	−0.03

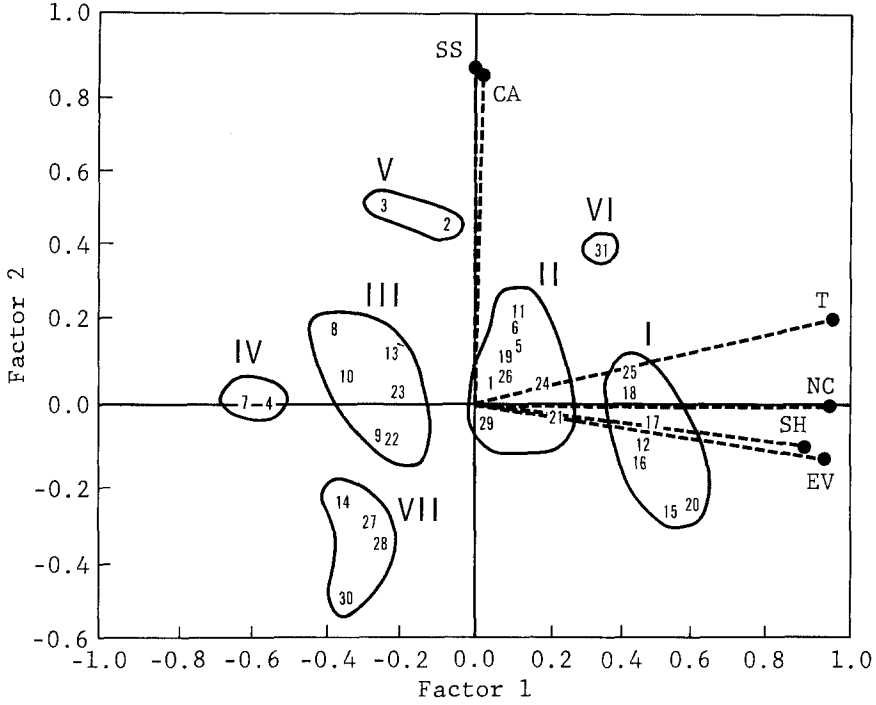


Fig. 6. Principal components analysis *RQ*-mode factors 1 and 2, from Fox data.

est carbonate and at the same time a very thick shale and evaporite sequence. These groups have a somewhat regular spatial distribution (Fig. 8), which is similar to the distribution (Fig. 9) found by Howarth (1973) using nonlinear mapping. The configuration obtained by correspondence analysis (Figs. 4 and 5) is different from that in Fig. 6, but similar to the configuration produced by Imbrie's *Q*-mode factor analysis (Fig. 10).

The second example is taken from a study by Sherman, Bunker, and Bush (1971) in the Berea area of Virginia, U.S.A., where a small, highly radioactive quartz monzonite pluton intrudes chlorite-actinolite schist and is overlapped by coastal plain sand and gravel deposits. A total of 22 auger samples were taken along a profile across all formations and analyzed for U, Th, and K. The objective of the original study was to relate the concentrations of these elements to airborne radiometric measurements made along the profile.

The measurements of U, Th, and K concentrations and airborne radiometry listed in Table 4 and shown in Fig. 11 were analyzed by principal components analysis and correspondence analysis. The solution from principal components analysis (Table 5, Fig. 12) shows positive correlations among all of the four variables which collectively form the first factor, accounting for 85% of the total

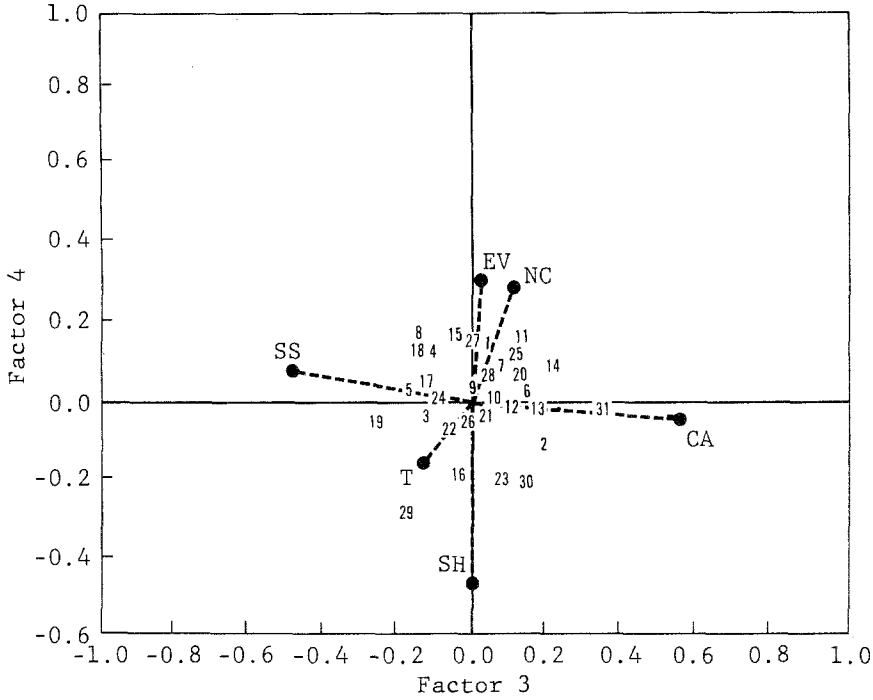


Fig. 7. Principal components analysis *RQ*-mode factors 3 and 4, from Fox data.

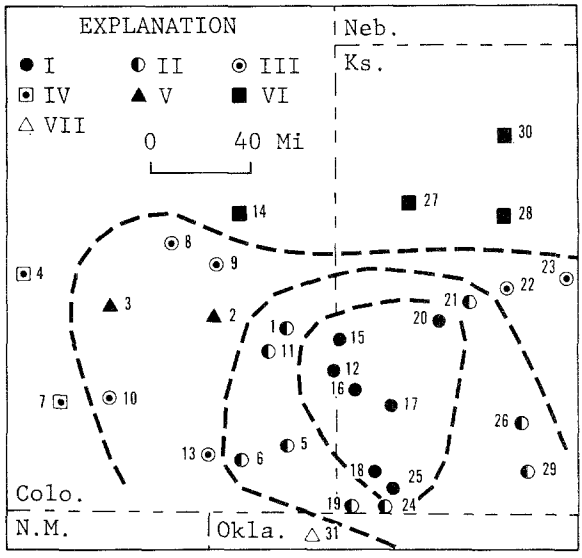
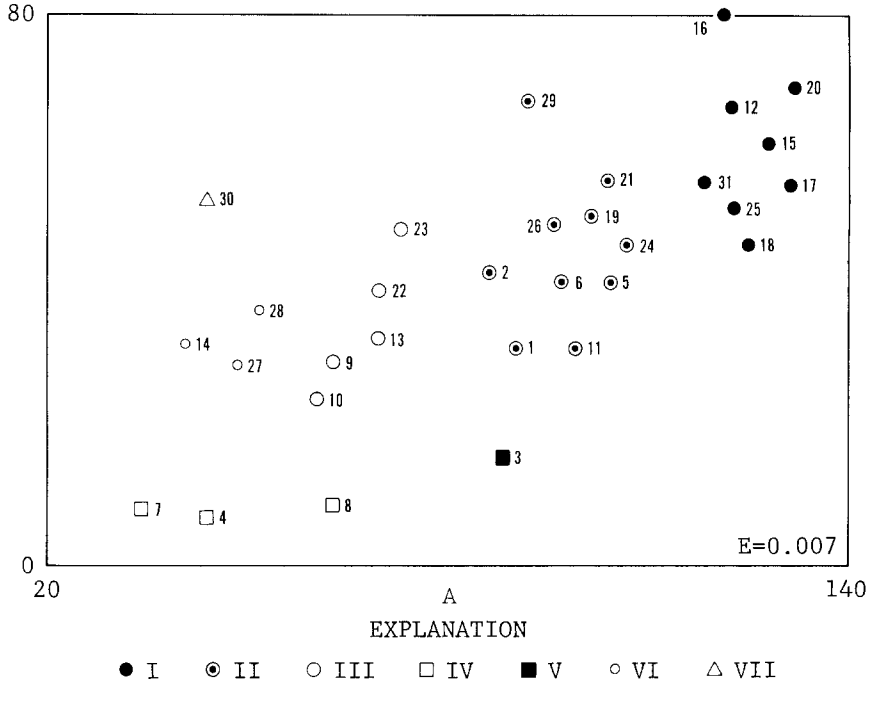
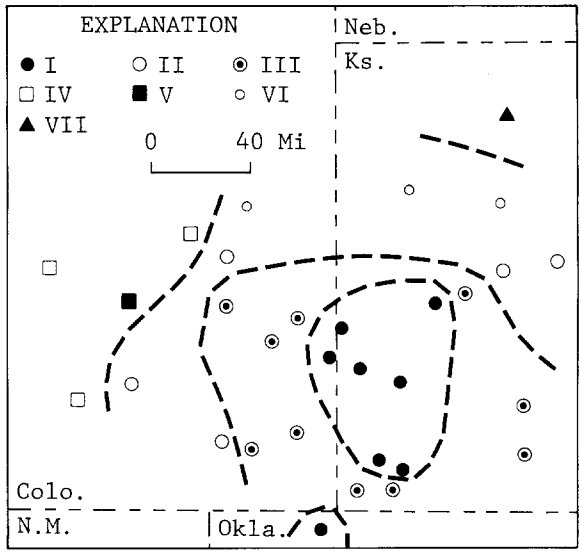


Fig. 8. Spatial distribution of groups distinguished in Fig. 6. Base map taken from Krumbein (1962).





B

Fig. 9. (A) Nonlinear mapping of Fox data. Original dimensionality is 6. (B) Spatial distribution of groups differentiated by nonlinear mapping. (From Howarth, 1973. Base map from Manson and Imbrie, 1964.)

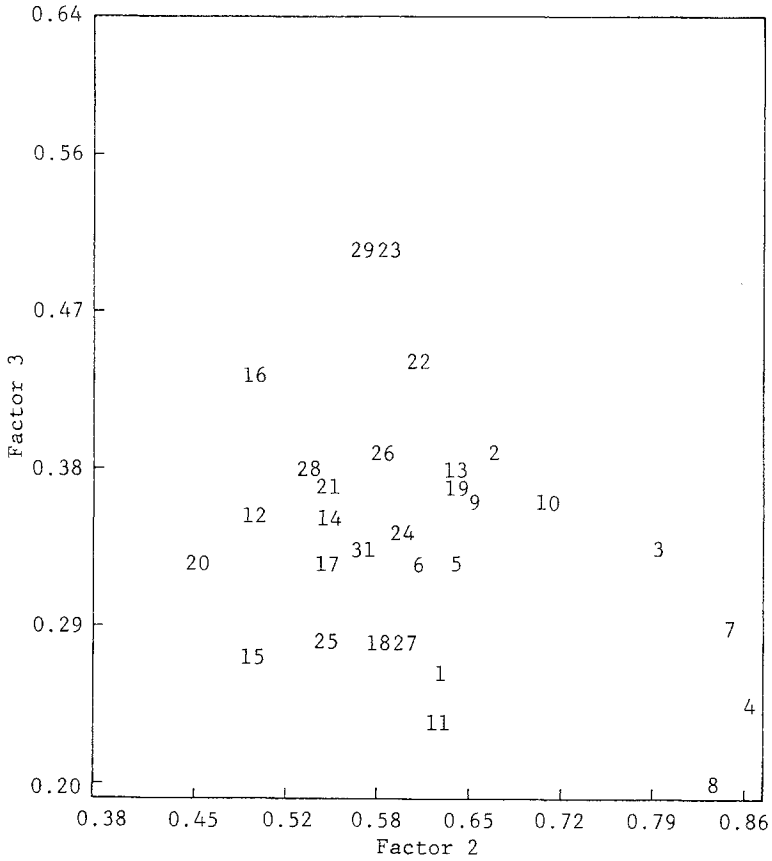


Fig. 10. Q-mode factors 2 and 3 of Fox data (after Manson and Imbrie, 1964).

variation. Samples from different bedrock types are distributed in distinct areas along the first factor, in agreement with their distinct airborne measured radioactivity and concentrations of U, Th, and K. Note that sample 5 was taken from the contact zone between quartz monzonite and schist and is located in a middle position, reflecting a mixture of the two sources.

Correspondence analysis yields a different pattern (Table 6, Fig. 13). The variable of airborne radiometry appears at the center of the diagram and is not associated with the compositional variables. Samples from different bedrocks tend to be clustered together. These results are inconsistent with the geologic interpretation of the relationships between the variables and between the samples in this data set. The indiscriminant clustering of sample points is due to the fact that the distinctions between the three types of bedrock are primarily expressed as differences in magnitudes of the measurements, to which correspondence analysis is not sensitive.

Table 4. Measurements on Quartz Monzonite Pluton, Berea, Virginia[a]

	A	B	C	D
No.	AERO	U	TH	K
1	240	0.63	2.05	0.13
2	360	2.18	5.31	0.31
3	420	2.26	5.61	0.34
4	500	1.71	6.44	0.70
5	580	2.38	7.99	1.73
6	700	3.83	8.32	4.26
7	600	3.79	9.46	1.53
8	650	4.09	14.71	3.11
9	770	4.21	12.00	1.90
10	930	4.72	12.78	2.92
11	1020	6.24	16.31	2.29
12	1000	5.24	14.51	1.88
13	1000	4.73	15.79	4.64
14	1040	4.67	10.30	4.17
15	1150	5.08	13.11	3.97
16	1000	5.27	13.40	4.36
17	960	5.61	10.31	2.05
18	420	2.33	6.83	0.47
19	370	2.64	9.88	0.58
20	400	2.29	6.02	0.34
21	480	2.32	6.14	0.32
22	730	5.94	12.86	1.35

[a](A) Airborne radiometric measurements in counts per second; (B) uranium in parts per million; (C) thorium in ppm; (D) potassium in percent.

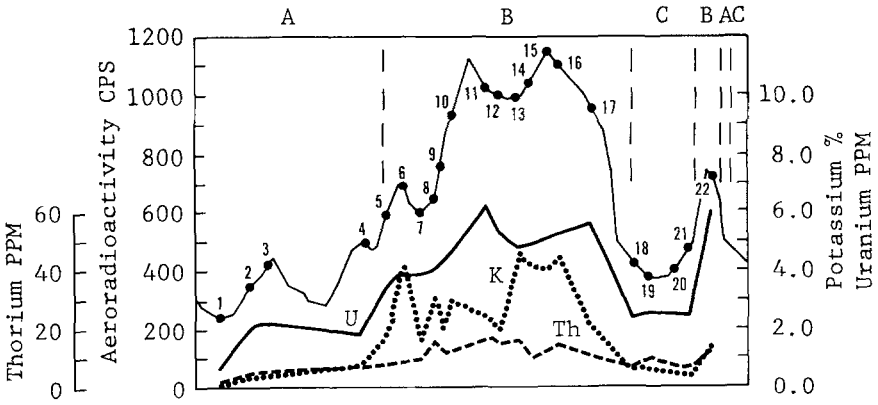


Fig. 11. Profile showing uranium, thorium, and potassium concentrations in soil samples and airborne radiometric intensity in the Berea area, Salem Church quadrangle, Virginia. (A) Chlorite-actinolite schist; (B) quartz monzonite; (C) sand and gravel. (From Sherman, Bunker, and Bush, 1971.)

Table 5. *RQ*-Mode Principal Components Analysis of Berea Data. (A) Matrix of Correlations Between Variables (AERO = airborne radiometric measurement; U = uranium content; TH = thorium content; K = potassium content). (B) Eigenvalues of Correlation Matrix (λ = eigenvalue; % = percent of trace; Σ% = cumulative percent of trace). (C) *RQ*-Mode Factor Loadings. (D) *Q*-Mode Factor Loadings.

(A) Correlation Matrix

	AERO	U	TH	K
AERO	1.00			
U	0.89	1.00		
TH	0.82	0.89	1.00	
K	0.82	0.67	0.69	1.00

(B) Eigenvalues

	Factors			
	1	2	3	4
λ	3.39	0.39	0.15	0.06
%	84.81	9.76	3.87	1.55
Σ%	84.81	94.57	98.45	100.00

(C) *R*-Mode Loadings

	Factors			
	1	2	3	4
AERO	0.96	0.05	−0.22	−0.16
U	0.94	−0.27	−0.13	0.17
TH	0.92	−0.25	0.28	−0.07
K	0.86	0.51	0.09	0.07

(D) *Q*-Mode Loadings

	Factors			
	1	2	3	4
1	−0.75	0.12	−0.03	−0.01
2	−0.49	−0.02	−0.02	0.03
3	−0.45	−0.02	−0.04	0.00
4	−0.41	0.04	−0.00	−0.08
5	−0.22	0.09	0.03	−0.05
6	0.11	0.29	−0.00	0.12
7	−0.08	−0.05	0.00	0.04
8	0.21	0.00	0.23	0.03
9	0.12	−0.08	0.02	−0.03
10	0.31	0.00	−0.01	−0.03
11	0.51	−0.23	0.00	−0.01

Table 5. Continued

(D) Q-Mode Loadings				
	Factors			
	1	2	3	4
12	0.35	−0.18	−0.03	−0.09
13	0.53	0.14	0.14	−0.04
14	0.36	0.21	−0.11	−0.00
15	0.50	0.11	−0.08	−0.07
16	0.49	0.13	0.01	0.03
17	0.26	−0.09	−0.19	0.04
18	−0.40	−0.04	0.01	−0.00
19	−0.31	−0.11	0.15	0.01
20	−0.44	−0.03	−0.01	0.01
21	−0.41	−0.03	−0.05	−0.03
22	0.21	−0.26	−0.03	0.12

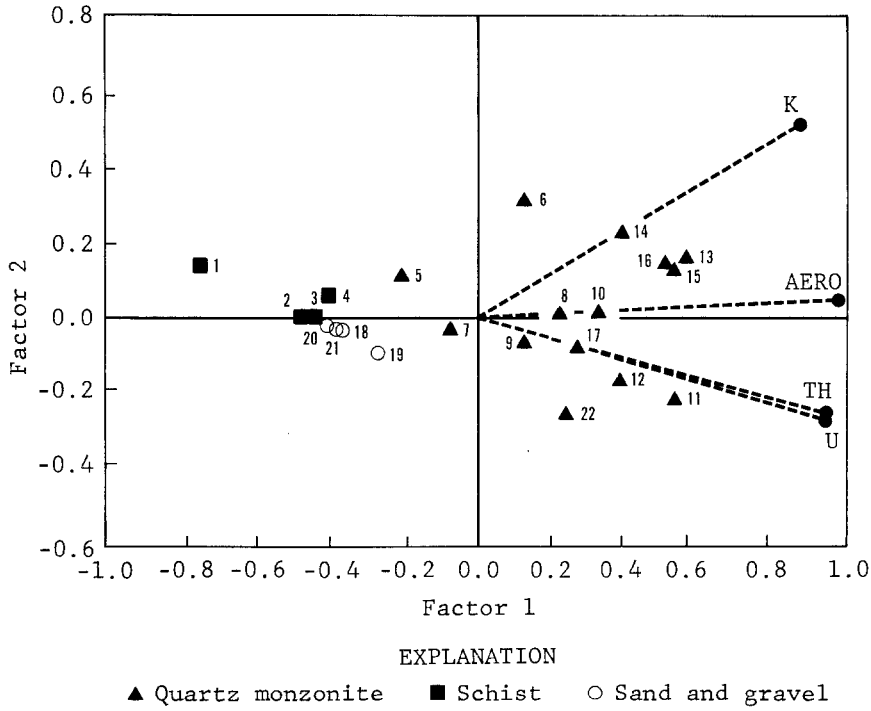


Fig. 12. RQ-mode principal components analysis of Berea area data.

Table 6. Correspondence Analysis of Berea Data (A) Eigenvalues of Profile Distance Similarity Matrix (λ = eigenvalue; % = percent of trace; Σ% = cumulative percent of trace). (B) Correspondence Analysis *G* Matrix (*R* Mode). (C) Correspondence Analysis *F* Matrix (*Q* Mode).

(A) Eigenvalues

	Factors		
	1	2	3
λ	0.0009	0.0007	0.0001
%	52.84	40.86	6.30
Σ%	52.84	93.70	100.00

(B) *G* Matrix

	Factors		
	1	2	3
AERO	−0.00	−0.00	−0.00
U	0.13	0.02	0.13
TH	0.22	0.07	−0.03
K	−0.20	0.46	0.01

(C) F Matrix

	Factors		
	1	2	3
1	−0.04	−0.06	−0.02
2	0.02	−0.03	0.01
3	0.01	−0.04	0.00
4	−0.01	−0.03	−0.02
5	−0.01	0.00	−0.01
6	−0.04	0.05	0.01
7	0.02	0.00	0.01
8	0.05	0.05	−0.01
9	0.01	−0.00	−0.00
10	−0.00	0.00	−0.00
11	0.02	−0.00	0.00
12	0.01	−0.02	−0.00
13	−0.00	0.03	−0.01
14	−0.04	0.01	0.00
15	−0.03	0.00	−0.00
16	−0.02	0.02	0.00
17	−0.02	−0.02	0.01
18	0.03	−0.02	−0.00
19	0.10	0.01	−0.01
20	0.02	−0.03	0.00
21	0.00	−0.04	−0.00
22	0.04	−0.01	0.03

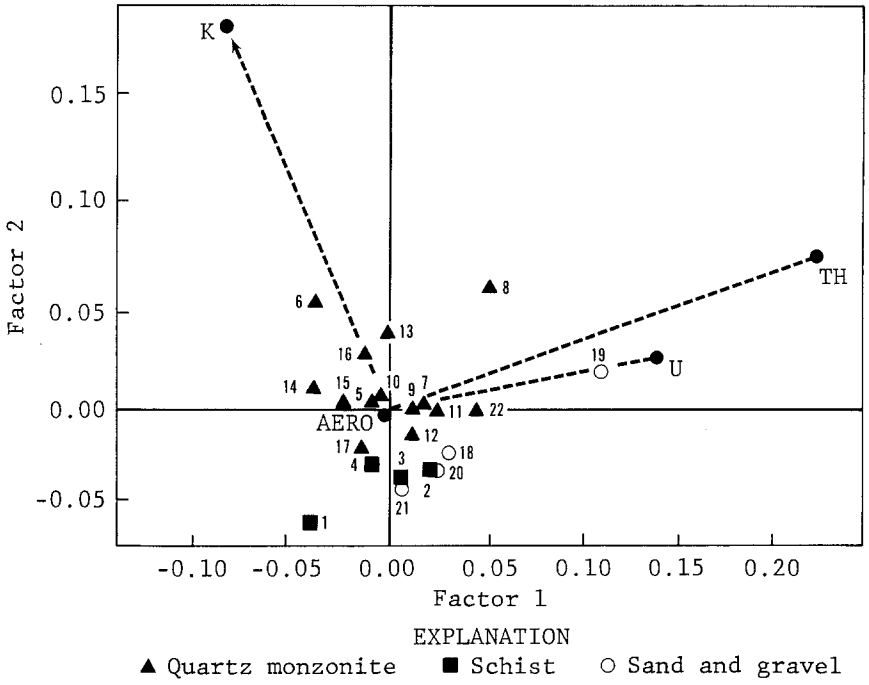


Fig. 13. Correspondence analysis of Berea area data.

CONCLUSIONS

In factor analysis, the scaling performed prior to factoring is of critical importance, because it determines the measure of similarity and hence the nature of the factor solution. The scaling method selected should yield a similarity matrix which is appropriate for the data being analyzed. In *RQ*-mode procedures, it is necessary and sufficient that the scaling simultaneously provides $W'W$ and WW' (or their specific transformations) which are meaningful measures of similarities between variables and between objects, respectively. Row-column symmetrical scaling is not required, unless the rows and columns in the data matrix are inherently symmetrical.

Correspondence analysis, as an *RQ*-mode factor procedure, was originally designed for analyzing contingency table data. Although the procedure can be applied to tables of measurement data, care should be taken because the resulting measure of proportional similarity, the profile distance, may not suitably express the similarities between variables and between objects.

Scaling in principal components analysis results in $W'W$ being a correlation or covariance matrix of variables and WW' being the Euclidean distances between objects. Therefore, principal components analysis can be used as an *RQ*-

mode procedure and applied to the problems where the similarities between variables are better measured by the correlation coefficient or covariance, and the similarities between objects are better measured by Euclidean distance. Such problems are common in geology and geochemistry, so principal components analysis should have broad applications as an *RQ*-mode procedure.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the comments and suggestions of Dr. Geoffrey Hill and Dr. A. T. Miesch. The senior author is supported jointly by the Ministry of Education, People's Republic of China, and the Kansas Geological Survey, who also provided computer time for this research.

APPENDIX

I. A Rationale for Correspondence Analysis When Applied to Measurement Data

Suppose the data set X consists of nonnegative measurements of m variables on n objects. A *profile distance* between variable j and k is defined as

$$d_{jk}^2 = \sum_i \left(\frac{x_{ij}}{x_{\cdot j}} - \frac{x_{ik}}{x_{\cdot k}} \right)^2 \Big/ x_{i\cdot}$$

where

$$x_{\cdot j} = \sum_i x_{ij}$$

and

$$x_{i\cdot} = \sum_j x_{ij}$$

Similarly, the profile distance between object i and l is defined by

$$d_{il}^{*2} = \sum_j \left(\frac{x_{ij}}{x_{i\cdot}} - \frac{x_{lj}}{x_{l\cdot}} \right)^2 \Big/ x_{\cdot j}$$

Let the jth variable $(j = 1, 2, \ldots, m)$ be represented by a point in object space with coordinates

$$b_{ij} = \frac{x_{ij}}{(x_{i\cdot})^{1/2} x_{\cdot j}}, \qquad \text{for} \quad i = 1, 2, \ldots, n$$

The Euclidean distances between these new variable points are equal to the profile distances between corresponding original variables. $B'B$ represents the con-

figuration of these new variable points. Similarly, the ith object ($i = 1, 2, \ldots, n$) is represented by a point in the variable space with coordinates

$$c_{ij} = \frac{x_{ij}}{x_{i\cdot}(x_{\cdot j})^{1/2}}, \qquad \text{for} \quad j = 1, 2, \ldots, n$$

The configuration of object points is represented by CC'.

Because $B'B \neq C'C$ and $BB' \neq CC'$, neither B nor C can be used directly in RQ-mode factor analysis. A matrix W is defined as

$$w_{ij} = \frac{x_{ij}}{(x_{i\cdot})^{1/2}(x_{\cdot j})^{1/2}}, \qquad \text{for} \quad \begin{matrix} i = 1, 2, \ldots, n \\ j = 1, 2, \ldots, m \end{matrix}$$

which is related to B by

$$w_{ij} = b_{ij}(x_{\cdot j})^{1/2}$$

and related to C by

$$w_{ij} = c_{ij}(x_{i\cdot})^{1/2}$$

Factoring W yields A^R and A^Q. Because $A^R A^{R'} = W'W \neq B'B$ and $A^Q A^{Q'} = WW' \neq CC'$, matrices G and F are formed by the transformation

$$\begin{aligned} g_{jk} &= a^R_{jk}/(x_{\cdot j})^{1/2}, \\ f_{ik} &= a^Q_{ik}/(x_{i\cdot})^{1/2} \end{aligned} \qquad \text{for} \quad \begin{matrix} i = 1, 2, \ldots, n \\ j = 1, 2, \ldots, m \\ k = 1, 2, \ldots, p \end{matrix}$$

Then $GG' = B'B$ represents the configuration of variable points and $FF' = CC'$ represents the configuration of object points. In other words, G and F give, respectively, the coordinates of variable points and object points in the factor space. The Euclidean distances between these points are equal to the profile distances between corresponding variables or objects.

The relation between G and F is given by

$$\sum_k f_{ik} g_{jk} \Big/ (\lambda_k)^{1/2} = \frac{W_{ij}}{(x_{i\cdot})^{1/2}(x_{\cdot j})^{1/2}}$$

which is similar to the relationship between A^R and A^Q in principal components analysis (Eq. 11). In this manner, the procedure of correspondence analysis, with a minor modification, can be applied to measurement data.

II. Verification of the Equality $H = WW'$

When W is defined by (7), H is defined by (9), and d_{il} is defined by (8). This can be verified in the following steps.

$$\sum_j w_{ij} w_{lj} = \frac{1}{n} \sum_j \left(x_{ij} - \frac{1}{n} \sum_u x_{uj} \right) \left(x_{lj} - \frac{1}{n} \sum_u x_{uj} \right)$$

$$= \frac{1}{n} \sum_j x_{ij} x_{lj} - \frac{1}{n^2} \sum_u \sum_j x_{ij} x_{uj} - \frac{1}{n^2} \sum_u \sum_j x_{uj} x_{lj} + \frac{1}{n^3} \sum_j \left(\sum_u x_{uj} \right)^2$$

$$e_{il} = -\frac{1}{2} d_{il}^2 = -\frac{1}{2n} \sum_j (x_{ij} - x_{lj})^2 = -\frac{1}{2n} \sum_j x_{ij}^2 - \frac{1}{2n} \sum_j x_{lj}^2 + \frac{1}{n} \sum_j x_{ij} x_{lj}$$

$$e_{i\cdot} = \frac{1}{n} \sum_l e_{il} = -\frac{1}{2n} \sum_j x_{ij}^2 - \frac{1}{2n^2} \sum_l \sum_j x_{lj}^2 + \frac{1}{n^2} \sum_l \sum_j x_{ij} x_{lj}$$

$$e_{\cdot l} = \frac{1}{n} \sum_i e_{il} = -\frac{1}{2n^2} \sum_i \sum_j x_{ij}^2 - \frac{1}{2n} \sum_j x_{lj}^2 + \frac{1}{n^2} \sum_i \sum_j x_{ij} x_{lj}$$

$$e_{\cdot\cdot} = \frac{1}{n^2} \sum_i \sum_l e_{il} = -\frac{1}{2n^2} \sum_i \sum_j x_{ij}^2 - \frac{1}{2n^2} \sum_l \sum_j x_{lj}^2 + \frac{1}{n^3} \sum_j \left(\sum_u x_{uj} \right)^2$$

Hence, $h_{il} = e_{il} - e_{i\cdot} - e_{\cdot l} + e_{\cdot\cdot} = \Sigma_j w_{ij} w_{lj}$, or $H = WW'$.

A similar verification holds when W, H, and d_{il} are defined by (5), (9), and (6), respectively.

REFERENCES

Benzecri, Jean-Paul, 1969, Statistical analysis as a tool to make patterns emerge from data, *in* S. Watanabe (Ed.) Methodologies of pattern recognition: Academic Press, New York, p. 35–74.

Benzecri, Jean-Paul, F. Benzécri, A. Birou, S. Blumenthal, A. De Bœck, J-P. Bordet, G. Cancelier, P. Cazes, F. da Costa Nicolau, M. Danech-Pajouh, R. Delprat, M. Demonet, B. Escoffier, A. Forcade, Fr. Friant, Y. Grelet, D. Kalogéropoulos, L. Lebart, M.-O. Lebeaux, P. Leroy, J.-F. Marcotorchino, T. Moussa, F. Mutombo, Ch. Nora, A. Prost, A. Rezvani, J. Robert, Ch. Rosenzveig, M. Roux, P. Solety, S. Stépan, N. Tabard, N. Tabet, G. Thauront, M. de Virville, and Y. Vuillaume, 1980, L'Analyse des données, Vol. 2, L'Analyse des Correspondances: Dunod, Paris.

David, M., Dagbert, M., and Beauchemin, Y., 1977, Statistical analysis in geology: Correspondence analysis method: Quart. Colorado Sch. Min., v. 72, no. 1, 57 p.

Eckart, C. and Young, B., 1936, The approximation of one matrix by another of lower rank: Psychometrika, v. 1, no. 3, p. 211–218.

Gabriel, K. R., 1971, The biplot graphic display of matrices with application to principal component analysis: Biometrika, v. 58, no. 3, p. 453–467.

Gower, J. C., 1966, Some distance properties of latent root and vector methods used in multivariate analysis: Biometrika, v. 53, no. 3, 4, p. 325–338.

Gower, J. C., 1967, Multivariate analysis and multidimensional geometry: The Statistician, v. 17, no. 1, p. 13–18.

Hill, M. O., 1974, Correspondence analysis: A neglected multivariate method: Jour. Roy. Stat. Soc., Ser. C: Appl. Stat., v. 23, no. 3, p. 340–354.

Hotelling, H., 1933, Analysis of a complex of statistical variables into principal components: Jour. Educ. Psych., v. 24, p. 417–441, 498–520.

Howarth, R. J., 1973, Preliminary assessment of a nonlinear mapping algorithm in a geological context: Math. Geol., v. 5, no. 1, p. 39–57.

Imbrie, J., 1963, Factor and vector analysis program for analyzing geologic data: Technical Report No. 6, Office of Naval Research, Geography Branch, Northwestern University, 83 p.

Jambu, M., 1980, Cluster analysis for data analysis, 1. Methods: Unpublished manuscript, 328 p.

Johnson, R. M., 1963, On a theorem stated by Eckart and Young: Psychometrika, v. 1, no. 3, p. 259–263.

Jöreskog, K. G., Klovan, J. E., and Reyment, R. A., 1976, Geological factor analysis, methods in geomathematics, 1: Elsevier Scientific Publishing Company, Amsterdam, 178 p.

Klovan, J. E., and Imbrie, J., 1971, An algorithm and FORTRAN IV program for large-scale Q-mode factor analysis and calculation of factor scores: *Math. Geol.*, v. 3, no. 1, p. 61–77.

Krumbein, W. C., 1962, Open and closed number systems in stratigraphic mapping: Bull. Amer. Assoc. Pet. Geol., v. 46, p. 2229–2245.

Krumbein, W. C. and Imbrie, J., 1963, Stratigraphic factor maps: Bull. Amer. Assoc. Pet. Geol., v. 47, p. 698–701.

Lee, P. J., 1969, FORTRAN IV programs for canonical correlation and canonical trend surface analysis: Kansas Geol. Surv. Comput. Contrib., v. 32, 46 p.

Manson, V. and Imbrie, J., 1964, FORTRAN program for factor and vector analysis of geologic data using an IBM 7090 or 7094/1401 computer system: Kansas Geol. Surv. Spec. Distrib. Publ. 13, 46 p.

Miesch, A. T., 1980, Scaling variables and interpretation of eigenvalues in principal component analysis of geologic data: Jour. Math. Geol., v. 12, no. 6, p. 523–538.

Sherman, K. N., Bunker, C. M., and Bush, C. A., 1971, Correlation of uranium, thorium, and potassium with aeroradioactivity in the Berea area, Virginia: Econ. Geol., v. 66, p. 302–308.

Teil, H., 1975, Correspondence factor analysis: An outline of its method: Math. Geol., v. 7, no. 1, p. 3–12.