DANIEL F. MERRIAM, Editor

FORTRAN IV PROGRAM FOR Q-MODE CLUSTER ANALYSIS OF NONQUANTITATIVE DATA USING IBM 7090/7094 COMPUTERS

By

G. F. BONHAM-CARTER
Stanford University

COMPUTER CONTRIBUTION 17
State Geological Survey
The University of Kansas, Lawrence
1967

Editor's Remarks

The Kansas Geological Survey is the only geological organization now known to be distributing computer program decks and data decks. Most programs are written in FORTRAN IV, although some are available in ALGOL or FORTRAN II. They are sold for a limited time at nominal cost. Several of the early programs have been withdrawn from circulation or replaced by more versatile versions.

Programs are available for several computers including the Burroughs B5500, CDC 3400, GE 625, IBM 1620, 7040, 7090, 7094, and 360 System. An up-to-date list of decks is available by writing the Editor, COMPUTER CONTRIBUTIONS, Kansas Geological Survey, The University of Kansas, Lawrence, Kansas, 66044, U.S.A.

FORTRAN programs are distributed now on magnetic tape unless requested otherwise. We have initiated this system because (1) it is easier to reproduce error-free programs, (2) mailing is simplified, and (3) costs are minimized. This distribution method will allow us to give better service to users.

Data decks available include a description of more than 2,650 Precambrian wells drilled in the State, and 880 analyses of helium-bearing gases from Kansas natural gas fields. This information is available at a nominal cost; a description of the sets can be obtained by writing the editor.

The program described here, "FORTRAN IV program for Q-mode cluster analysis of nonquantitative data using IBM 7090/7094 computers" by G.F. Bonham-Carter, can be obtained for $15.00 for a limited time.

FORTRAN IV PROGRAM FOR Q-MODE CLUSTER ANALYSIS OF NONQUANTITATIVE DATA USING IBM 7090/7094 COMPUTERS

By

G. F. BONHAM-CARTER

ABSTRACT

CLUST3 is an IBM 7090/7094 FORTRAN IV program for classifying objects into groups based on a large number of nonquantitative characters considered simultaneously. Either Sokal and Michener's coefficient or Jaccard's coefficient may be used as a measure of similarity. The mean expected value of similarity also is calculated. The weighted or unweighted pair-group method may be used for clustering and results may be displayed automatically by setting a plotting option for drawing a dendrogram. A version for the IBM 360/67 also is obtainable from the Kansas Geological Survey.

INTRODUCTION

Cluster analysis is a mathematical method of classification which has been applied in a number of fields, particularly biological taxonomy (Sokal and Sneath, 1963). By this method, a measure of resemblance or similarity is computed between all possible pairs of objects being classified; the objects then are linked progressively to form groups, by the criterion that the average similarity between members of the same group is greater than the average similarity between members of different groups. The technique is particularly useful if each object is described by a large number of variables or attributes, for under these circumstances, subjective grouping 'by eye' becomes difficult.

Despite the usefulness of numerical classification, application is limited in geology due to the difficulty of assigning meaningful numbers to many geological attributes. For instance, properties such as the shape and color of rocks do not fit easily a metric scale of measurement. Furthermore, where measurement techniques do exist, they are laborious and time-consuming, e.g. petrographic modal analysis by point-counting.

This program is designed specifically to perform cluster analyses of data that are essentially qualitative. Observations on each object (geological sample, locality, etc.) are recorded on a two-state nominal scale, e.g. attribute present / attribute absent, character positive / character negative, etc. Semiquantitative measurements made on a scale such as abundant / present / rare/ absent may also be used, after coding into a two-state form, as described below.

By employing cluster analysis to group and simplify these data, geological properties which are difficult to measure quantitatively may be used for multivariate classification. Geological usage of cluster analysis with nonquantitative data includes limestone facies analysis (Klovan, 1964, Bonham-Carter, 1965, 1967) and ecology of Recent Foraminifera and Ostracoda (Kaesler, 1966). For geological applications of cluster analysis involving fully quantitative data, see Purdy (1963), Howd (1964), Behrens (1965), and Parks (1966).

METHOD

General

A number of clustering techniques have been used in various fields, their level of mathematical sophistication differing considerably.

A method described by Friedman and Rubin (1966) and Rubin (1965) starts with an arbitrary classification of n objects into k groups, and procceeds to improve on this initial classification by iteratively moving objects one at a time from group to group, until an 'optimum' classification is reached. Several numerical criteria can be used for optimization. An analogous technique, described by Cassetti (1964), starts with some initial classification and improves it iteratively, using discriminant functions as optimization criteria.

These methods are most suitable for use with fully quantitative data, where similarity measures have an inverse linear relation to Euclidean distances. However, work by Shepherd (1962), Kruskal (1964), and others on multidimensional scaling suggests that pairwise similarity measures, placed in rank order, may be used for mathematically sophisticated analyses of this type after conversion to Euclidean distances by a method involving a minimum of distortion.

The clustering method programmed here is more 'rough-and-ready,' but utilitarian and probably more suitable for the data employed. Rather than starting with an initial classification, which is later adjusted, ungrouped objects progressively are linked, first into pairs, then into progressively larger groups by an iterative process. The method is known as the 'pair-group' method using average linkage, an and is fully discussed by Sokal and Sneath (1963). This program essentially was developed by coding the computational instructions described by Sokal and Sneath (1963) and adding a number of refinements, particularly that of automatically plotting the results in the form of a dendrogram.

The Data Array

Table 1 is an example of a raw data array which could be used for this type of cluster analysis. Note that each row of the array consists of a description of a single object or geological sample in terms of presence or absence of a number of attributes or characters. For this program, the convention 2 = 'character present'; 1 = 'character absent' has been adopted, rather than the usual 1,0. Here, 0 means 'character uncertain' or 'no information' as distinct from absence.

Table 1.-Example of a raw-data array.

CHARACTERS

SAMPLES	Corals	Brachiopods	Foraminifera	Ostracods	Trilobites	Algae
1	2	0	2	1	0	2
2	1	2	2	1	1	0
3	2	0	2	2	2	1
4	2	2	1	1	1	2
5	1	1	1	2	2	2

2 = present
1 = absent
0 = uncertain or no information

Table 2 illustrates how semiquantitative attributes may be coded into two-state form. For an attribute with four states abundant / present / rare / absent, three columns of two-state characters are required for its description (see also Bonham-Carter, 1965).

Measures of Similarity

Two objects described by a number of two-state characters may be compared by aligning the two sequences of digits and counting the frequency of 'matches' and 'mismatches' between them. In Table 3, a 'positive match' indicates a match of 2 with 2, a 'negative match' 1 with 1, and a 'mismatch' 2 with 1. The simplest measure of similarity is the ratio of matches to the number of characters compared, and may be expressed as:

$$S_{SM} = \frac{p + n}{p + n + m}$$

where

S_{SM} = Sokal and Michener's simple matching coefficient,
p = frequency of positive matches,
n = frequency of negative matches,
m = frequency of mismatches.

Another useful matching coefficient is Jaccard's coefficient of association (S_J):

$$S_J = \frac{p}{p + m}$$

This coefficient ignores negative matches, and thus prevents a high degree of similarity from occurring between objects that have a large number of absent characters in common. Both coefficients range between 1.0 (complete similarity) and 0.0 (complete dissimilarity). Either S_{SM} or S_J may be used with this program; little modification would be required for calculating other measures of similarity of which there are many (Sokal and Sneath, 1963).

Table 2.-Coding semiquantitative data into two-state form.

Samples	Brachiopods	Brachiopod Coding		
1	abundant	2	2	2
2	present	2	2	1
3	rare	2	1	1
4	absent	1	1	1
5	uncertain	0	0	0

The program only makes comparisons between objects or samples - the so-called Q-mode approach. R-mode comparisons between variables are not made easily using qualitative data. Contingency coefficients (Siegel, 1956) might be used if their range could be scaled logically between 0 and 1.

Clustering Procedure

At each cycle of the grouping procedure, the similarity matrix (Table 5) is scanned for the largest values in both row and column. Objects corresponding to these values are linked by pairs, and each pair is represented essentially by a single object of intermediate composition in the subsequent cycle.

Similarity values are calculated between groups of objects by taking the arithmetic average of the similarity coefficients between their component members. Two alternative methods are provided as a program option. They are the 'weighted' and 'unweighted' methods of Sokal and Sneath (1963). In the former technique, an equal weight is assigned to each group being linked, irrespective of the number of objects that each contains. In the latter technique, groups are 'unweighted' in so far as equal importance is assigned to each object, giving larger groups more weight than smaller ones.

Dendrogram Representation

The final step in the analysis entails the display and interpretation of the cluster results. Successive linkages can be displayed in the form of a two-dimensional hierarchy diagram in which the degree of similarity is represented along one axis and the objects are listed along the other axis, as in Figure 1. If objects are strongly partioned into groups, each group will appear as a well-defined cluster. More commonly, however, relationships are not so clear, entailing difficult and somewhat subjective decisions as to the level of similarity at which clusters are defined.

OPERATING INSTRUCTIONS

General

Although it is possible to construct dendrograms on the off-line printer, e.g. Parks (1966), an accurate and neatly drawn diagram can be produced by a digital plotter. This cluster program is in two parts. The first (CLUST3) performs all analyses and punches a deck of cards containing the coordinates for plotting a dendrogram. The second program (DNPLOT) is a short routine which uses punched output from CLUST3 as input, and controls the sequence of plotting instructions to a 10-inch CALCOMP digital plotter. The plotting routine has been separated from the main program so that the CALCOMP call statements, which usually vary from one installation to the next, may be modified without upsetting the more complex logic of the main program.

The programs are written in FORTRAN IV, Version 13 and have run on both an IBM 7094 and an IBM 7090. The data array is limited to 130 objects and 100 characters. This size restriction could be overcome by further tape manipulation, allowing a larger number of objects to be handled, but computation time would increase considerably. A full analysis of 130 limestone samples with 40 characters takes about two minutes execution time, both by the weighted and unweighted methods.

Input Instructions for CLUST3

Following the program deck and data identification card (normally $ DATA), these cards are required:

(i) Title Card

Columns

1-72		Title for the job, to be printed at the head of each page of output (12A6).

(ii) Control Card

Columns

3-5	NS	Number of objects; integer less than 131.
8-10	NT	Number of characters; integer less than 101.
15	NDAT	Data printout option. 1 = data matrix printed 0 = option suppressed
20	KOEF	Coefficient option.

Table 3.-Matching 2 samples with 8 characters.

CHARACTERS

Sample A	2	2	1	1	1	1	2	2
Sample B	2	1	1	1	2	2	2	1
Match	PM	MS	NM	NM	MS	MS	PM	MS

where

PM = positive match
NM = negative match
MS = mismatch

Table 4.-Data decks for Test Deck Number 1 and Test Deck Number 2.

```
TEST DECK NUMBER 1.
   25   25    1    0    1    0    1    0    2    2
(25I1,2X,A6)
1111212221201212111221 2102  DDD--1
```

```
TEST DECK NUMBER 1.
   25   25    1    0    1    0    1    0    2    2
(25I1,2X,A6)
11112122212012121112212102  DDD--1
2112222111211121112112212  BBB--5
2221111101121202221221021  CCC--3
1211212221221212111212122  DDD--3
2112121121101112122221001  EEE--2
2022222011211121112111212  BBB--1
2112221112211221012112212  BBB--2
1221112212122212211122111  AAA--4
0221111121221212221221121  CCC--1
2122222112210121112111212  BBB--3
2112121221112122101221221  EEE--1
2222111121021222221221121  CCC--4
2112121221111112121221222  EEE--4
2112121121112112121221222  EEE--5
1211012202212101112212122  DDD--2
1221122212122211001112111  AAA--2
1021112212122212221112110  AAA--1
1211212212222121212112122  DDD--4
2112101121101102121221222  EEE--3
1011212202212121112110112  DDD--5
2112220111211221112101212  BBB--4
1221120202122012221110111  AAA--3
1222211121221212221221221  CCC--2
2222111121021212221221121  CCC--5
1221122222122212211122111  AAA--5
```

A

```
TEST DECK NUMBER 2.
   25   25    1    0    1    0    0    0    2    2
(25I1, 1X, A6)
2221121112122112112122121       1
2111211111112122220201212       2
1121212121111120221121211       3
2111212112121111122111221       4
1112221122121112121212111       5
2112122022211121111212212       6
1211111222111122111001111       7
2211122211111011111211221       8
1021111122221112211111112       9
2121211111121101202111221      10
2210112221122111211211112      11
1211022221120211110111221      12
2111221111111221122211112      13
2112111111012121211111112      14
1112221011122112221211102      15
1111212112121211211222212      16
1121222112111112221121212      17
1112202211212222121221211      18
22122211121202112122121221      19
2112111202111122212121222      20
1111221122111111212211121      21
1221122211201212222121211      22
1122211102111210111111111      23
1122211221221211211022222      24
1212120121110122212112221      25
```

B

		1 = Jaccard's coefficients 0 = Sokal and Michener's coefficients
25	NWT	Cluster method option. 1 = unweighted method 0 = weighted method
30	MATIN	Option for reading in triangular array of similarity coefficients. 1 = matrix to be read in 0 = option suppressed
35	MATOUT	Option for printing out array of similarity coefficients. 1 = matrix printed 0 = option suppressed
40	NOCLUS	Clustering suppression option. 1 = clustering stages suppressed 0 = clustering as normal
45	NDORD	Re-ordered data matrix option. 2 = original data matrix punched and printed in dendrogram order 1 = matrix printed only 0 = option suppressed

Note: The following 4 variables must be supplied if a dendrogram is to be plotted:

50	KDEN	Dendrogram option. 2 = dendrogram coordinates punched and printed 1 = punched only 0 = option suppressed
51-55	XLNG	Length of X-axis along which objects are listed, in inches. Any integer between NS/4 and NS/6. If no value is supplied, integral value of NS/4 is used (F5.0).
56-60	CON	Highest possible value of similarity. Only required if MATIN = 1. Automatically set to 1 if no value is supplied (F5.0).
61-67	BOTVAL	Similarity value to be set at right-hand end of Y, or similarity axis. See Figure 1. Allows dendrogram to be "stretched" if all linkages occur within a narrow range at high levels of similarity. If no value supplied, automatically set to 0.0. If BOTVAL is greater than similarity level of lowest link, BOTVAL is readjusted. Cols. 61-65 Integer part of number. Cols. 66-67 Decimal part of number (no decimal point required).

(iii) Variable Format Card

Columns	
1-72	Variable format, enclosed in parentheses. The first field must be for reading integers from the data array; the second field for reading an alphanumeric name, to identify the object number, e.g. (3I1, 5X, A6). If MATIN = 1, a floating point format is required, and the alphanumeric names are supplied

on different cards (see below), e.g. (9F8.4).

(iv) Name Card (s)

To be used only if MATIN = 1. Alphanumeric names, 1 for each object in data array, are punched continuously in consecutive fields 6 columns wide, 12 to a card. Number of cards required = ((NS-1)/12)+1. If no names supplied, the correct number of blank cards must be inserted. Note: For normal program operation, these cards are not required; only if MATIN = 1.

(v) Data Array Cards

If MATIN = 0. Objects x characters data array. If objects are rows of this array, each card, or series of cards, contains a single row, followed by an alphanumeric name for the object, read in according to the variable format. Note: Each row begins on a new card.

If MATIN = 1. Objects x objects array of similarity coefficients; lower left triangle only, columnwise. Each column begins on a new card, starting with the element from the principal diagonal, read according to the variable format.

(vi) Multiple Job Processing

Repeat steps (i) through (v), and load decks consecutively.

Input Instructions for DNPLOT

This program is extremely simple to use, as only a single card need be supplied in addition to the deck punched by CLUST3. A description of the punched deck is included here, in case of difficulty.

Following the program deck and identification card:

(i) Quantity Card

Columns		
1-3	NUMDEC	Number of jobs, and consequently dendrograms, to be plotted on this machine pass. Note: This card must be supplied by the program user.

(ii) Title Card (supplied in punched deck)

(iii) Control Card (supplied in punched deck)

Columns		
1-5	NS	As for CLUST3.
6-15	XLNG	As for CLUST3.
16-25	HT	Height of lettering on X-axis.
26-35	CON	As for CLUST3.
36-45	BOTVAL	As for CLUST3.
46-50	NWT	As for CLUST3.
51-55	KOEF	As for CLUST3.

(iv) Sequence Cards (supplied in punched deck)

The names of the objects, placed in the correct order for the dendrogram. Each field is 6 columns wide, 12 to a card. Total number of cards = ((NS-1)/12)+1.

(v) Linkage Cards (supplied in punched deck)

X-Y coordinates in floating point inches of four points to be joined that will form one complete linkage. Total number of cards = NS-1.

(vi) End Card (supplied in punched deck)

Contains 99.9 in columns 2-5, denoting the end of the linkages.

(vii) Stem Card (supplied in punched deck)

X-Y coordinates in floating point inches of two points to be joined that will form the final stem.

(viii) Multiple Job Processing

Repeat steps (ii) through (vi) and load decks consecutively, making sure that NUMDEC is set to the correct number.

Output from CLUST3

The following tables are produced:

(i) Title Page

General information on this job.

(ii) Raw Data Matrix

Printed if NDAT = 1.

(iii) Redundant Characters List

If there are any characters that are invariant for all objects in the data array, these characters are listed and are not used in the subsequent computations.

(iv) Reliability Values

Occasionally, a large number of "holes" occur in the data, i.e. a large number of entries in the data array are coded as 0, meaning no information or uncertain. In order to recognize this weakness in the data, reliability values are output alongside those pairs of objects whose similarity value is based on a relatively small number of character comparisons. For each pair of objects being matched:

R = NC/NT

where

R = value of reliability,

ranging between 0.0 (completely reliable) to 1.0 (totally unreliable).

NC = number of character comparisons where at least one of the pair of objects is coded as 0.

NT = total number of characters. The serial numbers of the pair of objects concerned, their similarity value and their reliability value are listed if R is greater than 0.25.

(v) Expected Value of Association

If a data array is constructed by randomly arranging an equal number of 2's and 1's into rows and columns, the mean expected value (MEV) of S_{SM} is 0.5, and of S_J is 0.33. In arrays where 2's and 1's are randomly arranged, but the frequency of 2's is not the same as the frequency of 1's, there will be a variation in the MEV. Thus the MEV gives an indication of the similarity level around which clustering would be expected to take place if the same number of 2's and 1's were randomly arranged. This is helpful in assessing the significance of the dendrogram results.

The expected matching coefficients for two objects A and B is given by

$$E_{SM} = \frac{(c.d) + (e.f)}{(c.d) + (e.f) + (e.d) + (c.f)}$$

and

$$E_J = \frac{c.d}{(c.d) + (e.d) + (c.f)}$$

where

E_J = expected value of S_J

E_{SM} = expected value of S_{SM}

c = frequency of 2's in object A

d = frequency of 2's in object B

e = frequency of 1's in object A

f = frequency of 1's in object B.

Then

$$MEV = \frac{\sum_{i=1}^{NS} \sum_{j=1}^{NS} (E)_{ij}}{(NS)^2}$$

The occurrence of 0's is assumed to be small and is ignored.

(vi) Similarity Matrix

The lower half of the matrix of similarity coefficients is printed if MATOUT = 1.

(vii) Table of Linkages

The serial numbers of each pair of objects or samples to be linked are listed, with the level of similarity alongside. Groups of objects are labeled by a single member. For example, if samples 1 and 27 are grouped together, this cluster might be labeled by 1 in the subsequent cycles.

(viii) Dendrogram Order

The order of objects for placement along the X-axis of the dendrogram is listed.

(ix) Re-ordered Data Matrix

The original data matrix may be printed or punched in the order specified for the dendrogram. This is very useful for determining the composition of groups and for further analysis of the classification (Bonham-Carter, 1967).

(x) Dendrogram Information

A deck of cards may be punched for input to DNPLOT. In addition the dendrogram coordinates may be printed.

Output from DNPLOT

There is no printed output. Users are advised to ask the CALCOMP operator for output plotted in black ink. Ballpoint plots are normally messy, often spoiled by small blotches.

EXAMPLES

Two Test Problems Using Artifical Data

Two sets of data are listed in Table 4. In Test Deck Number 1, the 25 rows of numbers might correspond to limestone samples and the columns to various attributes such as the presence or absence of various fossil organisms. Test Deck Number 2 might record similar information from a different set of rocks. It is difficult to discern natural groupings of samples from these data arrays 'by eye'; there are too many factors to consider simultaneously.

An extract from the lower half of the similarity matrix (using S_{SM}) from Test Deck Number 1, the mean expected value of association, the tables of linkage values, and the re-ordered data matrices from both test decks are listed in Tables 5, 6, and 7. Dendrograms for both sets of data are shown in Figures 1 and 2.

Clearly, Test Deck Number 1 contains five well-defined groups, each group being homogeneous in composition and strongly differentiated from the other groups. It is noteworthy that the expected value of S_{SM} for this deck is close to 0.5. All groups are defined at a similarity level greater than 0.85.

Samples from Test Deck Number 2, on the other hand, do not show strong groups. Some linkages occur at a fairly high level of similarity, but the majority of links occur in a region around $S_{SM} = 0.5$, which is also the expected level of association. In fact, these data were coded by using a pseudorandom number generator.

Geological Example

In a study of the bottom sediments of an area in the Persian Gulf, Houbolt (1957) collected a large number of grab samples for analysis. He measured the abundance of 20 types of Foraminifera from each sample, recording the results on a semi-quantitative scale (Table 9). Displayed in chart form, the raw data present a complex array of symbols, difficult to study by eye. From this information Houbolt distinguished a number of foraminiferal assemblages, which he subsequently showed to be closely related to the depth of water from which each sample was collected.

As a cluster analysis test, 70 of these samples were coded into two-state form, arbitrarily grouping some of Houbolt's categories together (Tables 9 and 10). The full data array is listed in Table 11, the sample names being Houbolt's original locality numbers, prefixed by an abbreviation for the foram-assemblage name into which each sample was placed originally.

These data were analyzed using Jaccard's coefficients and the weighted method. Execution time was about 80 seconds. The resulting dendrogram is shown in Figure 3. A number of reasonably well-defined groups are present. Certainly the cluster level is much higher than the mean expected value of association ($S_J = 0.1$).

By studying the distribution of the sample name prefixes in the left margin, it can be seen that many of the dendrogram groups correspond quite closely to the original assemblages of Houbolt. As might be expected, there is no exact concurrence.

If the dendrogram clusters are plotted against depth (Figure 4), these groups are seen to be strongly depth-dependent too. Even without Houbolt's conclusions, it would be reasonable to interpret the

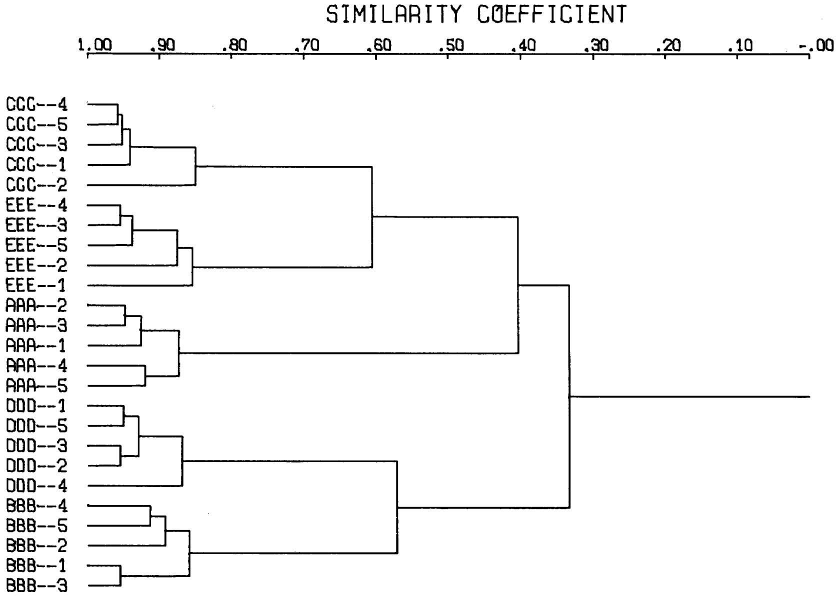

Figure 1.-Dendrogram plot for Test Deck Number 1, Sokal and Michener's coefficients, unweighted pair-group method.

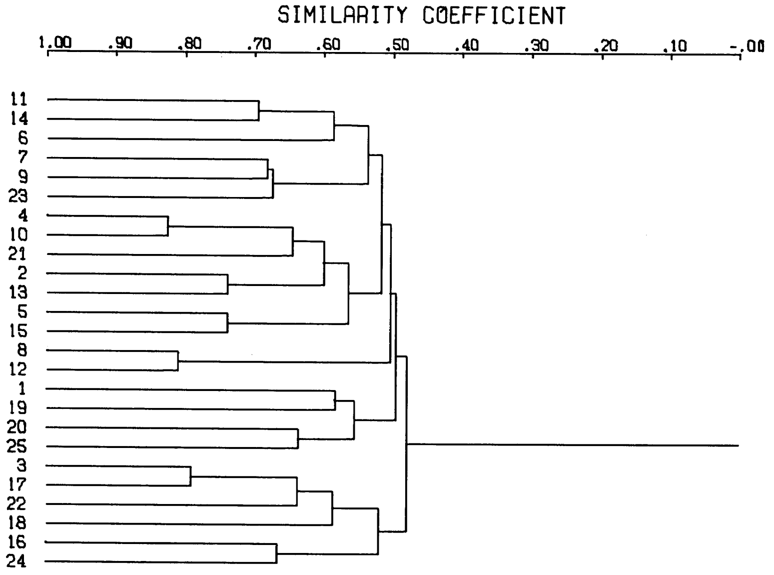

Figure 2.-Dendrogram plot for Test Deck Number 2, Sokal and Michener's coefficients, unweighted pair-group method.

variation in foraminiferal composition of these samples as being primarily depth controlled.

There is no question here as to which classification is 'better'. Classification is an arbitrary device, whether subjective or objective. The original hand sorting of samples into foraminiferal assemblages must have been very difficult and tedious, despite the familiarity with the raw material. Here a person unfamiliar with the raw data (not that this is necessarily desirable!) has simply processed them and in a few seconds computation time has been able to arrive at somewhat similar conclusions. This is a powerful demonstration of the usefulness of cluster analysis as a tool for classification.

Cautionary Note

It must also be stressed, however, that different classifications can be obtained from the same data, using different similarity coefficients or different clustering methods. Furthermore, although the clusters formed at high levels of similarity are normally stable and reasonable, it is sometimes difficult to account for linkages at lower similarity levels. This is probably caused by the sequential nature of the clustering process, coupled with the repeated averaging of similarity values.

It is suggested, therefore, that cluster analysis be used as a data reduction tool, to give computational aid for classification. It may be necessary, for instance, to use several clustering methods on the same data, to decide, maybe arbitrarily, how best to partition the data. This expedient may seem unsatisfactory and to defeat some of the objectives of numerical classification. However, with experience this technique is found to be exceedingly valuable for providing insight into complex multivariate data, as is shown by the above examples.

Table 5.-Similarity matrix for Test Deck Number 1 (in part).

```
       TEST DECK NUMBER 1.
   THE FOLLOWING PAIRS OF SAMPLES HAVE A HIGH PROPORTION OF UNCERTAIN MATCHES
           SAMPLE PAIR     SIMILARITY COEFFICIENT     PROPORTION OF UNCERTAIN MATCHES

             5   22              0.3889                         0.2800
            19   22              0.2778                         0.2800
      MEAN EXPECTED VALUE OF ASSOCIATION =  0.5004


      TEST DECK NUMBER 1.
      MATRIX OF COEFFICIENTS OF ASSOCIATION


    SAMPLES
 NAME      NO.      17     18     19     20     21     22     23     24     25
 AAA--1    17      1.000
 DDD--4    18      0.609  1.000
 EEE--3    19      0.250  0.182  1.000
 DDD--5    20      0.524  0.864  0.211  1.000
 BBB--4    21      0.190  0.435  0.550  0.600  1.000
 AAA--3    22      0.947  0.524  0.278  0.450  0.150  1.000
 CCC--2    23      0.435  0.320  0.636  0.182  0.391  0.476  1.000
 CCC--5    24      0.500  0.250  0.714  0.095  0.318  0.550  0.875  1.000
 AAA--5    25      0.826  0.520  0.273  0.455  0.217  0.905  0.480  0.542  1.000

      TEST DECK NUMBER 1.
      MATRIX OF COEFFICIENTS OF ASSOCIATION


    SAMPLES
 NAME      NO.      1      2      3      4      5      6      7      8      9     10     11     12     13     14     15     16
 DDD--1     1      1.000
 BBB--5     2      0.652  1.000
 CCC--3     3      0.150  0.182  1.000
 DDD--3     4      0.913  0.640  0.182  1.000
 EEE--2     5      0.286  0.500  0.650  0.182  1.000
 BBB--1     6      0.571  0.913  0.250  0.609  0.450  1.000
 BBB--2     7      0.591  0.875  0.238  0.583  0.429  0.773  1.000
 AAA--4     8      0.478  0.240  0.591  0.520  0.273  0.261  0.292  1.000
 CCC--1     9      0.182  0.167  0.952  0.250  0.619  0.227  0.217  0.583  1.000
 BBB--3    10      0.636  0.875  0.190  0.625  0.429  0.955  0.826  0.292  0.174  1.000
 EEE--1    11      0.409  0.458  0.571  0.333  0.810  0.455  0.391  0.333  0.478  0.435  1.000
 CCC--4    12      0.174  0.250  0.952  0.208  0.667  0.318  0.304  0.500  0.913  0.261  0.609  1.000
 EEE--4    13      0.348  0.480  0.591  0.280  0.864  0.478  0.417  0.280  0.542  0.417  0.875  0.583  1.000
 EEE--5    14      0.348  0.480  0.591  0.280  0.864  0.435  0.417  0.280  0.542  0.458  0.875  0.583  0.920  1.000
 DDD--2    15      0.950  0.545  0.238  0.955  0.263  0.500  0.476  0.500  0.333  0.524  0.381  0.238  0.364  0.364  1.000
 AAA--2    16      0.524  0.348  0.450  0.565  0.250  0.381  0.391  0.870  0.455  0.409  0.304  0.364  0.261  0.261  0.550  1.000

      TEST DECK NUMBER 1.
      MATRIX OF COEFFICIENTS OF ASSOCIATION


    SAMPLES
 NAME      NO.      1      2      3      4      5      6      7      8      9     10     11     12     13     14     15     16
 AAA--1    17      0.524  0.261  0.550  0.522  0.250  0.273  0.318  0.913  0.545  0.318  0.273  0.455  0.304  0.304  0.500  0.905
 DDD--4    18      0.826  0.560  0.273  0.920  0.136  0.522  0.542  0.600  0.333  0.542  0.250  0.292  0.200  0.200  0.864  0.609
 EEE--3    19      0.300  0.500  0.700  0.227  0.900  0.450  0.429  0.227  0.619  0.429  0.857  0.714  0.955  0.955  0.300  0.150
 DDD--5    20      0.950  0.682  0.100  0.955  0.211  0.667  0.619  0.500  0.190  0.714  0.333  0.143  0.273  0.273  0.900  0.550
 BBB--4    21      0.524  0.913  0.300  0.522  0.550  0.905  0.909  0.217  0.273  0.864  0.500  0.364  0.522  0.522  0.400  0.286
 AAA--3    22      0.368  0.190  0.579  0.429  0.389  0.263  0.250  0.857  0.600  0.300  0.400  0.500  0.381  0.381  0.421  0.947
 CCC--2    23      0.174  0.280  0.818  0.240  0.591  0.348  0.333  0.480  0.875  0.292  0.500  0.833  0.560  0.560  0.273  0.348
 CCC--5    24      0.130  0.208  0.952  0.167  0.714  0.273  0.261  0.542  0.957  0.217  0.565  0.958  0.625  0.625  0.238  0.409
 AAA--5    25      0.391  0.240  0.545  0.440  0.364  0.261  0.292  0.920  0.583  0.292  0.417  0.500  0.360  0.360  0.455  0.870
```

Table 6.-Linkage values for (A) Test Deck Number 1, (B) Test Deck Number 2.

```
TEST DECK NUMBER 1.

CLUSTERING BY THE UNWEIGHTED PAIR-GROUP METHOD

 SAMPLE NUMBERS     LEVEL OF ASSOCIATION   CYCLE NUMBER
   4      15              0.9545                1
   6      10              0.9545                1
   8      25              0.9200                1
  12      24              0.9583                1
  13      19              0.9545                1
  16      22              0.9474                1
  21       2              0.9130                1
  12       3              0.9524                2
  13      14              0.9373                2
  16      17              0.9261                2
   2       7              0.8920                2
   1      20              0.9500                2
  12       9              0.9406                3
  13       5              0.8758                3
  16       8              0.8734                3
   2       6              0.8592                3
   1       4              0.9294                3
  12      23              0.8504                4
  13      11              0.8542                4
   1      18              0.8683                4
  12      13              0.6060                5
   1       2              0.5712                5
  12      16              0.4033                6
  12       1              0.3331                7    A
```

```
TEST DECK NUMBER 2.

CLUSTERING BY THE UNWEIGHTED PAIR-GROUP METHOD

 SAMPLE NUMBERS     LEVEL OF ASSOCIATION   CYCLE NUMBER
   2      13              0.7391                1
   3      17              0.7917                1
   4      10              0.8261                1
   5      15              0.7391                1
   7       9              0.6818                1
   8      12              0.8095                1
   3      22              0.6377                2
   7      23              0.6742                2
  11      14              0.6957                2
  16      24              0.6667                2
   4      21              0.6461                3
  20      25              0.6364                3
   4       2              0.6004                4
   1      19              0.5833                4
   4       5              0.5647                5
   1      20              0.5553                5
   3      18              0.5864                5
  11       6              0.5870                5
   3      16              0.5198                6
  11       7              0.5369                6
  11       4              0.5167                7
  11       8              0.5030                8
  11       1              0.4955                9
  11       3              0.4804               10    B
```

Table 7.-Re-ordered data matrix for (A) Test Deck Number 1, (B) Test Deck Number 2.

```
TEST DECK NUMBER 1.

PRINTOUT OF ORIGINAL DATA IN DENDROGRAM ORDER

  SAMPLES         CHARACTERS
NAME     NO.

CCC--4   12   2222111121021222221221121
CCC--5   24   2222111121021212221221121
CCC--3    3   2221111101121202221221021
CCC--1    9   0221111121221212221221121
CCC--2   23   1222211121221212221221221
EEE--4   13   2112121221111112121221222
EEE--3   19   2112101121101102121221222
EEE--5   14   2112121121112112121221222
EEE--2    5   2112121121101112122221001
EEE--1   11   2112121221112122101221221
AAA--2   16   1221122212122211001112111
AAA--3   22   1221120202122012221110111
AAA--1   17   1021112212122212221112110
AAA--4    8   1221112212122212211122111
AAA--5   25   1221122222122212211122111
DDD--1    1   1111212212012121112212102
DDD--5   20   1011212202212121112110112
DDD--3    4   1211212212212121112112122
DDD--2   15   1211012202212101112212122
DDD--4   18   1211212212222121212112122
BBB--4   21   2112220111211221112101212
BBB--5    2   2112222111211121112112212
BBB--2    7   2112221112211221012112212
BBB--1    6   2022222011211121112111212
BBB--3   10   2122222112210121112111212    A
```

```
TEST DECK NUMBER 2.

PRINTOUT OF ORIGINAL DATA IN DENDROGRAM ORDER

  SAMPLES         CHARACTERS
NAME     NO.

  11     11   2210112221122111211211112
  14     14   2112111110121212111111112
   6      6   2112122022211121111212212
   7      7   1211111222111122111001111
   9      9   1021111122221112211111112
  23     23   1122211102111210111111111
   4      4   2111212112121111122111221
  10     10   2121211111121101202111221
  21     21   1111221122111111212211121
   2      2   2111211111121222202012l2
  13     13   2111221111111221122211112
   5      5   1112221122121112121212111
  15     15   1112221011122112221211102
   8      8   2211122211111011111211221
  12     12   1211022221120211110111221
   1      1   2221121112122112112122121
  19     19   2212211121202112122121221
  20     20   2112111202111122212121222
  25     25   1212120121110122212112221
   3      3   1121212121111120221121211
  17     17   1121222112111112221121212
  22     22   1221122211201212222121211
  18     18   1112202211212222121221211
  16     16   1111212112121211211222212
  24     24   1122211221221211211022222    B
```

Table 8.-Listing of punched output from CLUST3, input to DNPLOT, (A) Test Deck Number 1, (B) Test Deck Number 2.

```
TEST DECK NUMBER 1.
  25   6.25000   0.15000   1.00000  -0.         1    0
CCC--4CCC--5CCC--3CCC--1CCC--2EEE--4EEE--3EEE--5EEE--2EEE--1AAA--2AAA--3
AAA--1AAA--4AAA--5DDD--1DDD--5DDD--3DDD--2DDD--4BBB--4BBB--5BBB--2BBB--1
BBB--3
  0.25000  1.00000  0.25000  1.37500  0.50000  1.37500  0.50000  1.00000
  1.50000  1.00000  1.50000  1.40909  1.75000  1.40909  1.75000  1.00000
  2.75000  1.00000  2.75000  1.47368  3.00000  1.47368  3.00000  1.00000
  3.50000  1.00000  3.50000  1.72000  3.75000  1.72000  3.75000  1.00000
  4.50000  1.00000  4.50000  1.40909  4.75000  1.40909  4.75000  1.00000
  5.25000  1.00000  5.25000  1.78261  5.50000  1.78261  5.50000  1.00000
  6.00000  1.00000  6.00000  1.40909  6.25000  1.40909  6.25000  1.00000
  0.37500  1.37500  0.37500  1.42857  0.75000  1.42857  0.75000  1.00000
  1.62500  1.40909  1.62500  1.56455  2.00000  1.56455  2.00000  1.00000
  2.87500  1.47368  2.87500  1.66541  3.25000  1.66541  3.25000  1.00000
  4.00000  1.00000  4.00000  1.45000  4.25000  1.45000  4.25000  1.00000
  5.37500  1.78261  5.37500  1.97159  5.75000  1.97159  5.75000  1.00000
  0.56250  1.42857  0.56250  1.53416  1.00000  1.53416  1.00000  1.00000
  1.81250  1.56455  1.81250  2.11818  2.25000  2.11818  2.25000  1.00000
  3.06250  1.66541  3.06250  2.13975  3.62500  2.13975  3.62500  1.72000
  4.12500  1.45000  4.12500  1.63542  4.62500  1.63542  4.62500  1.40909
  5.56250  1.97159  5.56250  2.26712  6.12500  2.26712  6.12500  1.40909
  0.78125  1.53416  0.78125  2.34659  1.25000  2.34659  1.25000  1.00000
  2.03125  2.11818  2.03125  2.31250  2.50000  2.31250  2.50000  1.00000
  4.37500  1.63542  4.37500  2.18494  5.00000  2.18494  5.00000  1.00000
  1.01563  2.34659  1.01563  4.54589  2.26563  4.54589  2.26563  2.31250
  4.68750  2.18494  4.68750  4.85899  5.84375  4.85899  5.84375  2.26712
  1.64063  4.54589  1.64063  6.37051  3.34375  6.37051  3.34375  2.13975
  2.49219  6.37051  2.49219  7.00233  5.26563  7.00233  5.26563  4.85899
 99.90000                                                                A
  3.87891  7.00233  3.87891 10.00000
```

```
TEST DECK NUMBER 2.
  25   6.25000   0.15000   1.00000  -0.         1    0
   11      14       6       7       9      23       4      10      21       2      13       5
   15       8      12       1      19      20      25       3      17      22      18      16
   24
  1.00000  1.00000  1.00000  3.86364  1.25000  3.86364  1.25000  1.00000
  1.75000  1.00000  1.75000  2.56522  2.00000  2.56522  2.00000  1.00000
  2.50000  1.00000  2.50000  3.34783  2.75000  3.34783  2.75000  1.00000
  3.00000  1.00000  3.00000  3.34783  3.25000  3.34783  3.25000  1.00000
  3.50000  1.00000  3.50000  2.71429  3.75000  2.71429  3.75000  1.00000
  5.00000  1.00000  5.00000  2.87500  5.25000  2.87500  5.25000  1.00000
  0.25000  1.00000  0.25000  3.73913  0.50000  3.73913  0.50000  1.00000
  1.12500  3.86364  1.12500  3.93182  1.50000  3.93182  1.50000  1.00000
  5.12500  2.87500  5.12500  4.26087  5.50000  4.26087  5.50000  1.00000
  6.00000  1.00000  6.00000  4.00000  6.25000  4.00000  6.25000  1.00000
  1.87500  2.56522  1.87500  4.18522  2.25000  4.18522  2.25000  1.00000
  4.50000  1.00000  4.50000  4.27273  4.75000  4.27273  4.75000  1.00000
  2.06250  4.18522  2.06250  4.59652  2.62500  4.59652  2.62500  3.34783
  4.00000  1.00000  4.00000  4.75000  4.25000  4.75000  4.25000  1.00000
  0.37500  3.73913  0.37500  4.71739  0.75000  4.71739  0.75000  1.00000
  2.34375  4.59652  2.34375  4.91789  3.12500  4.91789  3.12500  3.34783
  4.12500  4.75000  4.12500  5.00198  4.62500  5.00198  4.62500  4.27273
  5.31250  4.26087  5.31250  4.72283  5.75000  4.72283  5.75000  1.00000
  0.56250  4.71739  0.56250  5.16798  1.31250  5.16798  1.31250  3.93182
  5.53125  4.72283  5.53125  5.32196  6.12500  5.32196  6.12500  4.00000
  0.93750  5.16798  0.93750  5.34984  2.73438  5.34984  2.73438  4.91789
  1.83594  5.34984  1.83594  5.47338  3.62500  5.47338  3.62500  2.71429
  2.73047  5.47338  2.73047  5.54024  4.37500  5.54024  4.37500  5.00198
  3.55273  5.54024  3.55273  5.67634  5.82813  5.67634  5.82813  5.32196
 99.90000                                                                B
  4.69043  5.67634  4.69043 10.00000
```

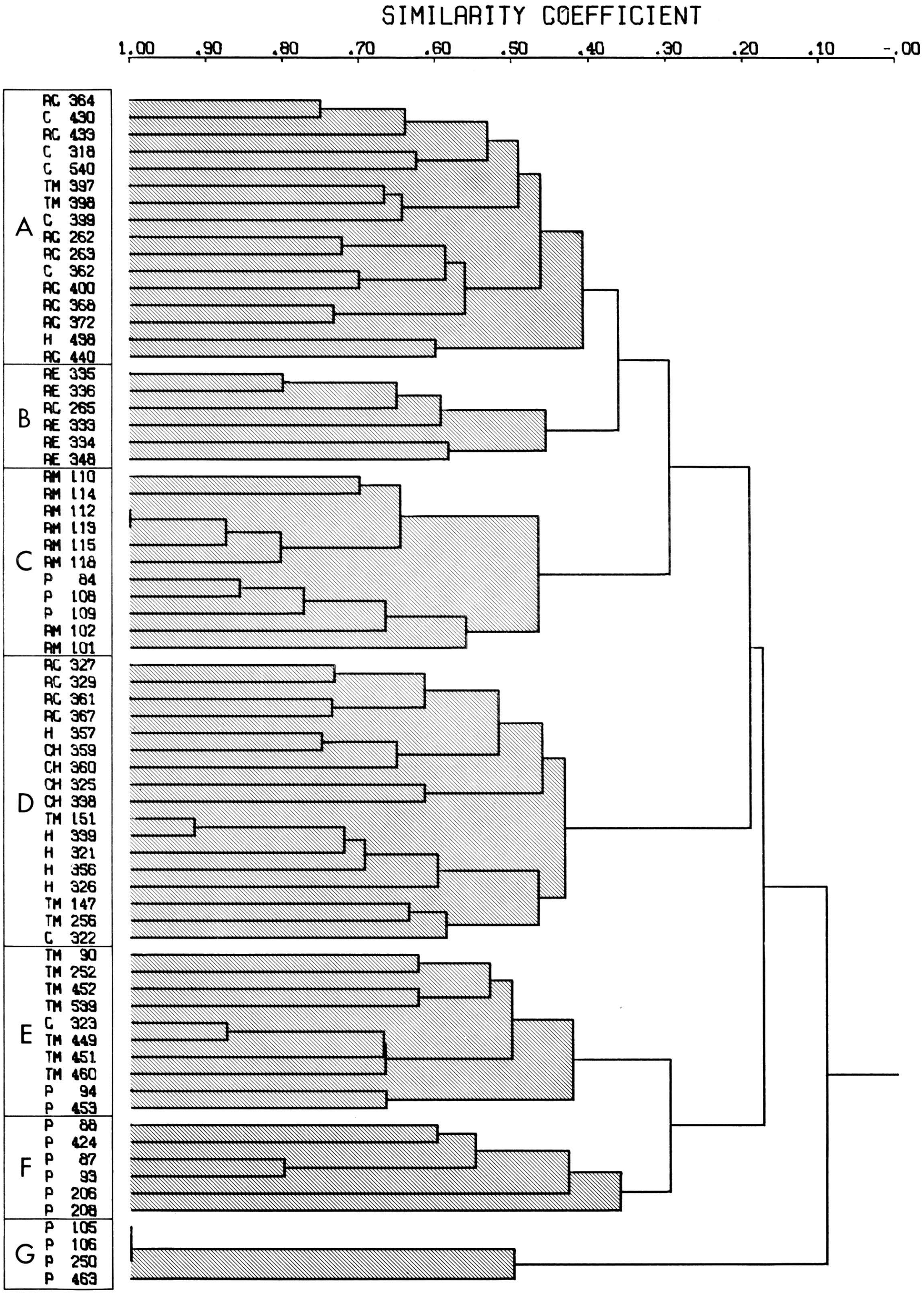

Figure 3.-Dendrogram plot of Houbolt's foram data from Persian Gulf, Jaccard's coefficient of association, weighted pair-group method.

Table 9.-Conversion of Houbolt's semiquantitative scale to two-state code.

Symbol	Name	Description	Coding
##	rock forming	> 1/4 of sample	2 2 2
#	flood	100 specimens - 1/4 sample	
++	abundant	21-100 speciments	2 2 1
+	common	6-20 specimens	
-	rare	2-5 specimens	2 1 1
∘	single	1 specimen	
	none	0 specimen	1 1 1

Table. 10.-Sample RE 333 coded into two-state form.

Textularia	Gaudyrina	Miliolidae	Nodophthalidium	Nonion	Elphidium	Elphidiella	Peneroplidae	Sorites	Bolivina	Reussella	Discorbis	Eponides	Rotalia	Cibicides	Amphistegina	Heterostegina	Het. quatarensis	Operculina	Globigerina	Foraminifera Genus
221	111	221	111	111	111	221	111	111	111	111	111	111	222	111	111	111	111	111	111	Coding
+		+				++							#							Symbol

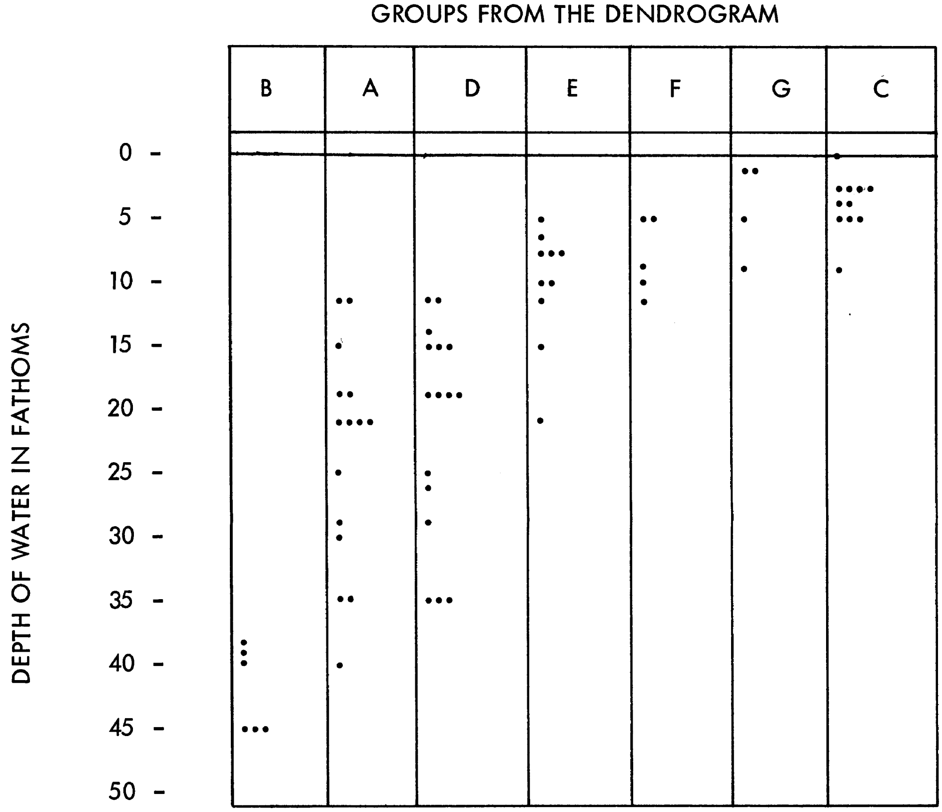

Figure 4.-Diagram showing the distribution of samples from dendrogram groups with depth of water.

Table 11.-Listing of data array and linkage levels for Houbolt's foram problem.

```
HOUBOLT'S FORAM DATA FROM PERSIAN GULF

THE FOLLOWING CHARACTERS ARE NOT USED FOR
CALCULATING COEFFICIENTS OF ASSOCIATION
BECAUSE THEY ARE REDUNDANT
          3  12  24  32  45  57
          6  15  26  33  48  59
          7  18  27  36  51  60
          9  21  30  39  54

     HOUBOLT'S FORAM DATA FROM PERSIAN GULF

     PRINTOUT OF DATA MATRIX

   SAMPLES          CHARACTERS
NAME     NO.

P    84    1  211111221111111221111221111111111111111211111111111111111111
P    87    2  211111221211111111111211111111111111111111111111111111111111
P    88    3  211111221111111111111221111111111111111111111111111111111111
TM   90    4  221111221211111211111211111111111111111111111111111111211111
P    93    5  211111221211111111111211111111111111111211111111111111111111
P    94    6  221111211211111111111111111111111111111111111111111111111111
RM  101    7  111111221111112111112211111111111111111121111111111111111111
RM  102    8  211111221211111221111211111111111111111121111111111111111111
P   105    9  111111211111111111111211111111111111111111111111111111111111
P   106   10  111111211111111111111211111111111111111111111111111111111111
P   108   11  211111221111111221111211111111111111111211111111111111111111
P   109   12  211111221111111211111211111111111111111211111111111111111111
RM  110   13  211111221111111221211111111111111111111121111111111111111111
RM  112   14  111111221111211221111111111111111111111122111111111111111111
RM  113   15  111111221111211221111111111111111111111122111111111111111111
RM  114   16  211111221111221221211111111111111111111122111111111111111111
RM  115   17  211111221111211221111111111111111111111122111111111111111111
RM  118   18  111111221111111221111111111111111111111122111111111111111111
TM  147   19  221111221111111111111111111111111111111111211211211111221111
TM  151   20  221111221111111111111111111111111111111111221221211221211111
P   206   21  111111221211111111111221111111111111111111111111111111211111
P   208   22  111111221211111111111111111111111111111111111111111111111111
P   250   23  111111211111111111111211111111111111111111111111111111111111
TM  252   24  221111221211111111111111111111111111111111111111111111211211
TM  256   25  221111221111111111111111111111111111111111111211221221221111
RC  262   26  221211221111111211221111111111111211211221221211111111111211
RC  263   27  221211221211111211221111211211111211111221221111111111111211
RC  265   28  221111221111111111221111111211111111111121221111111111111111
C   316   29  221111221221111111111211111111111211211111221211111111211111
H   321   30  221111221211111111111111111111111111111111221211211221111111
C   322   31  221111221211111111111111111111111111111111221111211211221111
C   323   32  221111221211111111111111111111111111112112112211111111111111
CH  325   33  111211221111111111111111111111111111112111112212211211211111
H   326   34  221211211211111111111111111111111111111111211221211221111111
RC  327   35  221211221111111111111111111121111111122121221111211111111111
RC  329   36  221211221111111111111111111121111111122121122122121111211111
RE  333   37  221111221111111221111111111111111111111111222111111111111111
RE  334   38  211111221111111111111111111111111211111111221111211111111211
RE  335   39  211111221111111221111111111111111111111111221211111111111211
RE  336   40  211111221111111221111111111111111211111111221221111111111211
CH  338   41  221211221111111111111111111111111111112111221211211211211211
H   339   42  221211221111111111111111111111111111111111221221211221211111
RE  348   43  221111221111111221111111111111111211211221111121111121111211
H   356   44  211111221111111111111111111111111111111111221221221221111111
H   357   45  221111221111111211111111111111211211211211221221221221221111
CH  359   46  221211221211111111111111111111211211211211221221221111221111
CH  360   47  221211221111111211111111111211111111111111221221221221221111
RC  361   48  221221221211111111211111111211111111122122122121122111121211
C   362   49  221221221211111211111111111111211211211211221111111111111211
RC  364   50  221111221221111111111111111111111211111121221111111111111111
RC  367   51  221221221111111111111111111211111111111121221211221111111211
RC  368   52  221211221111111211211111111111211111211221221111111111211111
RC  372   53  221211221211211111211111111111211111211221221111111111111111
TM  397   54  221111221211211111211111111111111211211211211111111111111111
TM  398   55  221111221211211211111111111111111211111211211111111111111211
C   399   56  221111221211211111111211111111111211211211221111211111111211
RC  400   57  221211221221211211111211111221211211211221221111111111111211
P   424   58  211111221111111111111211111111111111111111111112111111111111
C   430   59  221111221221211111111111111111111221111121221111111111111111
RC  433   60  221111221221111221211111111111111211111221221111111111211111
H   438   61  211111221111211111111111111111111111111111221221111111111111
RC  440   62  211111221211111111111111111111111111111111221221111111221111
TM  449   63  221211221211111111111111111111111111112112111111111111111111
TM  451   64  221111221211111111111111111111111111111111111211111111111111
TM  452   65  221111221211211111111111111111111211111111111111111111111111
P   453   66  211111211211111111111111111111111111111111111111111111111111
TM  460   67  221111221211111111111111111111111111111111211111111111111111
P   463   68  211111211111111111111211111111111111111111111111111111111111
TM  539   69  221111221211111111111111111111111211111111111211111111211111
C   540   70  221111221221111211111211111211111211111211211211111111221111
```

```
HOUBOLT'S FORAM DATA FROM PERSIAN GULF

THE FOLLOWING PAIRS OF SAMPLES HAVE A HIGH PROPORTION OF UNCERTAIN MATCHES

      SAMPLE PAIR     SIMILARITY COEFFICIENT     PROPORTION OF UNCERTAIN MATCHES


MEAN EXPECTED VALUE OF ASSOCIATION =  0.1232


           HOUBOLT'S FORAM DATA FROM PERSIAN GULF

           CLUSTERING BY THE WEIGHTED PAIR-GROUP METHOD

SAMPLE NUMBERS        LEVEL OF ASSOCIATION   CYCLE NUMBER
   1       11               0.8571               1
   2        5               0.8000               1
   4       24               0.6250               1
   6       66               0.6667               1
   9       10               1.0000               1
  14       15               1.0000               1
  19       25               0.6364               1
  20       42               0.9167               1
  26       27               0.7222               1
  29       70               0.6250               1
  32       63               0.8750               1
  35       36               0.7333               1
  39       40               0.8000               1
  45       46               0.7500               1
  48       51               0.7368               1
  49       57               0.7000               1
  50       59               0.7500               1
  52       53               0.7333               1
  61       62               0.6000               1
  65       69               0.6250               1
   1       12               0.7738               2
   9       23               1.0000               2
  14       17               0.8750               2
  20       30               0.7212               2
  32       64               0.6696               2
  35       48               0.6155               2
  39       28               0.6515               2
  45       47               0.6526               2
  50       60               0.6384               2
   3       58               0.6000               2
  38       43               0.5833               2
   1        8               0.6667               3
   9       68               0.5000               3
  14       18               0.8036               3
  20       44               0.6939               3
  32       67               0.6682               3
  39       37               0.5932               3
  54       55               0.6667               3
  20       34               0.5994               4
  54       56               0.6429               4
   3        2               0.5500               4
   4       65               0.5312               4
  19       31               0.5874               4
  26       49               0.5867               4
  13       16               0.7000               4
  33       41               0.6154               4
   4       32               0.5023               5
  26       52               0.5613               5
  13       14               0.6465               5
   1        7               0.5613               5
  39       38               0.4570               5
  35       45               0.5193               5
  50       29               0.5307               5
  13        1               0.4670               6
  35       33               0.4622               6
  50       54               0.4906               6
  20       19               0.4673               6
   3       21               0.4286               6
  35       20               0.4332               7
  50       26               0.4618               7
   3       22               0.3625               7
  50       61               0.4074               8
   4        6               0.4228               8
  50       39               0.3628               9
   4        3               0.2976               9
  50       13               0.2964              10
  50       35               0.1918              11
  50        4               0.1759              12
  50        9               0.0930              13
```

Program listing.

```
$IBJOB
$IBFTC MAIN    NODECK
C.....CLUSTER ANALYSIS PROGRAM - VERSION 3 (CLUST3)
C.....CODED IN FORTRAN IV (VERSION 13), BY G.F.BONHAM-CARTER,
C.....DEPARTMENT OF GEOLOGY, UNIVERSITY OF TORONTO, SPRING 1966.
C.....MODIFIED NOVEMBER 1966, MARCH 1967.
C.....SUITABLE FOR USE WITH IBM 7090 OR 7094 COMPUTERS EQUIPPED WITH
C.....32K CORE STORAGE AND WITH LOGICAL TAPE NUMBERS 1,2, AND 3
C.....AVAILABLE FOR INTERMEDIATE STORAGE. NO DISK STORAGE USED.
C.....PERFORMS EITHER THE WEIGHTED OR THE UNWEIGHTED PAIR-GROUP METHOD
C.....OF CLUSTERING, USING EITHER JACCARD'S OR SOKAL AND MICHENER'S
C.....COEFFICIENTS OF ASSOCIATION AND AVERAGE LINKAGE. Q-MODE ONLY.
C.....CARRIAGE CONTROL CHARACTER 1H$ USED FOR PUNCHING CARDS OFF-LINE
C     ************
C     MAIN PROGRAM
C     ************
      DIMENSION NUM(130,130), FMT(12)
      COMMON TITLE(12), NAME(130)
      COMMON NCYC, NS, NL, NJ, NT, XLNG, CON, NDORD, KDEN, NSHOLD,
     1 BOTVAL, NWT, KOEF, COHOLD
      COMMON COEF(130,130), COSAV(130)
      EQUIVALENCE (NUM, COEF)
   19 READ(5,1) TITLE, NS, NT, NDAT, KOEF, NWT, MATIN, MATOUT,
     1   NOCLUS, NDORD, KDEN, XLNG, CON, BOTVAL, FMT
      NSHOLD=NS
      REWIND 1
      REWIND 2
      REWIND 3
      WRITE(6,2) NS, NT, TITLE
      IF (NWT.EQ.1) WRITE(6,11)
      IF (NWT.EQ.0) WRITE(6,12)
      IF (MATIN.NE.1) GO TO 20
C
C.....READ SAMPLE NAMES AND SKIP DOWN TO 260
      WRITE(6,15)
      KOEF=3
      READ(5,8) (NAME(I), I=1,NS)
      NDORD=0
      GO TO 260
   20 IF (KOEF.EQ.1) WRITE(6,3)
      IF (KOEF.EQ.0) WRITE(6,4)
   40 DO 50 I=1,NS
   50 READ(5,FMT) (NUM(I,J), J=1,NT) ,NAME(I)
      IF (NDAT.EQ.0) GO TO 60
C
C.....PRINT ORIGINAL DATA MATRIX
      WRITE(6,7) TITLE
      WRITE(6,5)
      DO 59 I=1,NS
   59 WRITE(6,6) NAME(I), I, (NUM(I,J), J=1,NT)
   60 WRITE(1) NUM
      REWIND 1
C
C.....CHECK FOR REDUNDANT CHARACTERS
      KK=0
      DO 62 J=1,NT
      DO 61 I=2,NS
      IF (NUM(I,J).NE.NUM(1,J)) GO TO 62
   61 CONTINUE
      KK=KK+1
```

```
      NUM(KK,130)=J
   62 CONTINUE
      IF (KK.GT.0) WRITE(6,16) TITLE, (NUM(I,130),I=1,KK)
C
C.....CONSOLIDATE DATA MATRIX
      NC=NT
      IF (KK.EQ.0) GO TO 68
      DO 66 J=1,NT
      JJ=NT-J+1
C
C.....CHECK TO SEE IF CHARACTER IS IN REDUNDANT LIST
      DO 63 K=1,KK
   63 IF (JJ.EQ.NUM(K,130)) GO TO 64
      GO TO 66
C
C.....MOVE COLUMNS OVER TO FILL GAP
   64 NC=NC-1
      DO 65 JX=JJ,NC
      DO 65 I=1,NS
   65 NUM(I,JX)=NUM(I,JX+1)
   66 CONTINUE
C
C.....FIND FREQUENCY OF 2'S AND 1'S IN EACH SAMPLE  - USED FOR
C.....CALCULATING EXPECTED VALUES OF ASSOCIATION  -  SEE BELOW
   68 DO 71 I=1,NS
      NUM(I,129)=0
   71 NUM(I,128)=0
      DO 77 J=1,NC
      DO 76 I=1,NS
      IF (NUM(I,J)-1) 76,75,74
   74 NUM(I,129)=NUM(I,129)+1
      GO TO 76
   75 NUM(I,128)=NUM(I,128)+1
   76 CONTINUE
   77 CONTINUE
C
C.....CALCULATE COEFFICIENTS OF ASSOCIATION AND STORE ROWISE ON TAPE 2
      ANT=NC
      WRITE(6,7) TITLE
      WRITE(6,13)
      SUM=0.0
      AVEXPT=0.0
      DO 170 I=1,NS
      DO 160 II=I,NS
      ANUMER=0.0
      DENOM=0.0
      REL=0.0
      DO 130 J=1,NC
      K=NUM(I,J) * NUM(II,J) + 1
      IF (KOEF.EQ.1) GO TO 90
      GO TO (100,120,110,120,120),K
   90 GO TO (100,130,110,120,120),K
  100 REL=REL + 1.0
      GO TO 130
  120 ANUMER=ANUMER+1.0
  110 DENOM=DENOM+1.0
  130 CONTINUE
C
C.....FINDING EXPECTED VALUE OF ASSOCIATION
      EXPECT=NUM(I,129)*NUM(II,129)
```

```
      IF (KOEF.EQ.0) EXPECT=EXPECT+FLOAT(NUM(I,128)*NUM(II,128))
      IF (EXPECT.GT.0.0) EXPECT=EXPECT/(EXPECT+
     1 FLOAT(NUM(I,128)*NUM(II,129)+NUM(I,129)*NUM(II,128)))
      AVEXPT=(AVEXPT*SUM+EXPECT)/(SUM+1.0)
      SUM=SUM+1.0
      COSAV(II)=0.0
      IF (DENOM.GT.0.0) COSAV(II)=ANUMER/DENOM
      REL=REL/ANT
      IF (REL.GT.0.25) WRITE(6,14) I,II,COSAV(II), REL
  160 CONTINUE
      WRITE(2)       (COSAV(II), II=I,NS)
  170 CONTINUE
      WRITE(6,17) AVEXPT
C
C.....READ MATRIX OF COEFFICIENTS FROM TAPE 2 AND STORE IN ARRAY COEF
      REWIND 2
      DO 180 I=1,NS
  180 READ(2)       (COEF(I,II), II=I,NS)
      REWIND 2
      GO TO 271
C
C.....READ IN MATRIX OF COEFFICIENTS AS DATA
  260 DO 270 I=1,NS
  270 READ(5,FMT) (COEF(I,II), II=I,NS)
  271 DO 275 I=1,NS
      DO 275 J=I,NS
  275 COEF(J,I)=COEF(I,J)
      IF(MATOUT.EQ.0) GO TO 320
C
C.....WRITE OUT MATRIX OF COEFFICIENTS
      DO 310 KK=1,NS,16
      KKEND=MIN0(NS,KK+15)
      DO 310 JSTART=1,KK,16
      JSTOP=MIN0(NS,JSTART+15)
      WRITE(6,9) TITLE, (J, J=JSTART,JSTOP)
      DO 310 I=KK,KKEND
      IF (KK.EQ.JSTART) JSTOP=I
  310 WRITE(6,10) NAME(I), I, (COEF(I,J), J=JSTART,JSTOP)
  320 IF (NOCLUS.EQ.1) GO TO 19
C
C.....PLACE ZEROS IN PRINCIPAL DIAGONAL
      DO 330 I=1,NS
  330 COEF(I,I)=0.0
      WRITE(6,7) TITLE
      IF (NWT.LE.0) GO TO 350
      WRITE(6,11)
      GO TO 360
  350 WRITE(6,12)
C
C.....BEGIN CLUSTERING - SUBROUTINES LARGE AND RECALC ARE
C.....ALTERNATELY CALLED UNTIL ALL SAMPLES HAVE BEEN GROUPED TOGETHER.
  360 NCYC=0
  370 CALL LARGE
      IF (NS.EQ.2) GO TO 400
      CALL RECALC
      GO TO 370
C
C.....SAMPLES ARE PLACED IN DENDROGRAM ORDER AND COORDINATES
C.....FOR DENDROGRAM CALCULATED.
  400 NS=NSHOLD
```

```
      CALL ORDER
      NS=NSHOLD
      IF (KDEN.GT.0) CALL DENDRO
      GO TO 19
    1 FORMAT(12A6/10I5,2F5.0,F7.2/12A6)
    2 FORMAT(1H1, 5X, 53HCLUSTER PROGRAM BY G.F.BONHAM-CARTER, UNIV OF T
     1ORONTO//  6X, 22HNUMBER OF SAMPLES    =, I4//  6X, 22HNUMBER OF CH
     2ARACTERS =, I4//  6X, 12A6)
    3 FORMAT(1H0,5X,41HUSING JACCARDS COEFFICIENT OF ASSOCIATION)
    4 FORMAT(1H0, 5X, 52HUSING SOKAL AND MICHENERS COEFFICIENT OF ASSOCI
     1ATION)
    5 FORMAT(1H0, 5X, 23HPRINTOUT OF DATA MATRIX/// 4X, 7HSAMPLES, 10X,
     1   10HCHARACTERS/ 1X,12HNAME     NO.//)
    6 FORMAT(1H ,A6,I6,2X,100I1)
    7 FORMAT(1H1, 5X, 12A6)
    8 FORMAT(12A6)
    9 FORMAT(1H1, 4X, 12A6//5X, 37HMATRIX OF COEFFICIENTS OF ASSOCIATION
     1//////4X,7HSAMPLES/1X,12HNAME     NO.,2X,16I7)
   10 FORMAT(1H0,A6,I6,4X,16F7.3)
   11 FORMAT(1H0, 5X, 46HCLUSTERING BY THE UNWEIGHTED PAIR-GROUP METHOD)
   12 FORMAT(1H0,5X, 44HCLUSTERING BY THE WEIGHTED PAIR-GROUP METHOD)
   13 FORMAT(1H0, 74HTHE FOLLOWING PAIRS OF SAMPLES HAVE A HIGH PROPORTI
     1ON OF UNCERTAIN MATCHES// 10X, 11HSAMPLE PAIR, 5X, 22HSIMILARITY C
     2OEFFICIENT, 5X, 31HPROPORTION OF UNCERTAIN MATCHES///)
   14 FORMAT(1H , 8X, 2I5, 13X, F7.4, 20X, F7.4)
   15 FORMAT(1H0, 5X, 49HMATRIX OF SIMILARITY COEFFICIENTS READ IN AS DA
     1TA)
   16 FORMAT(1H1,12A6//1X,81HTHE FOLLOWING CHARACTERS ARE NOT USED FOR C
     1ALCULATING COEFFICIENTS OF ASSOCIATION/1X,26HBECAUSE THEY ARE REDU
     2NDANT/// (5X,I5))
   17 FORMAT(1H0///5X, 37HMEAN EXPECTED VALUE OF ASSOCIATION = , F7.4)
      END
$IBFTC LARG    NODECK
C
C     ****************
      SUBROUTINE LARGE
C     ****************
      DIMENSION INDEX(130)
      COMMON TITLE(12), NAME(130)
      COMMON NCYC, NS, NL, NJ, NT, XLNG, CON, NDORD, KDEN, NSHOLD,
     1 BOTVAL, NWT, KOEF, COHOLD
      COMMON COEF(130,130), COSAV(130), ILIST(65), JLIST(130)
C
C.....SELECTS THOSE PAIRS OF SAMPLES (OR OF GROUPS OF SAMPLES) WITH THE
C.....HIGHEST COEFFICIENTS OF ASSOCIATION FOR LINKAGE IN A PARTICULAR
C.....CLUSTER CYCLE
      IF (NCYC.GT.0) GO TO 10
      WRITE(6,1)
C
C.....THE LOWEST SAMPLE NUMBER IN EACH CLUSTER IS HELD IN INDEX()
      DO 8 I=1,NS
    8 INDEX(I)=I
      KDUM=9999
   10 NCYC=NCYC+1
      NL=0
      IF (NS.GT.2) GO TO 11
C
C.....FOLLOWING STATEMENTS   (DOWN TO 11) EXECUTED ONLY DURING FINAL
C.....CLUSTER CYCLE
      NL=1
```

```
      NJ=1
      COHOLD=COEF(NL,NS) - CON/10.0
      WRITE(3)      NL, NJ
      WRITE(6,7) INDEX(1), INDEX(2), COEF(1,2), NCYC
      WRITE(3)      NL, NS, COEF(NL,NS)
      WRITE(3)      KDUM
      RETURN
   11 DO 12 I=1,NS
      IF (I.LE.65) ILIST(I)=0.0
   12 JLIST(I)=0.0
C
C.....FIND COEFFICIENTS LARGEST IN BOTH ROW AND COLUMN
      DO 50 I=1,NS
      BIG=COEF(I,1)
      JSAVE=1
      DO 30 J=1,NS
      IF (BIG.GE.COEF(I,J)) GO TO 30
      BIG=COEF(I,J)
      JSAVE=J
   30 CONTINUE
      DO 40 J=1,NS
      IF (BIG.LT.COEF(J,JSAVE)) GO TO 50
   40 CONTINUE
      IF (NL.EQ.0) GO TO 42
      DO 41 K=1,NL
   41 IF (JSAVE.EQ.ILIST(K).OR.JSAVE.EQ.JLIST(K).OR.I.EQ.ILIST(K).
     1   OR.I.EQ.JLIST(K).OR.I.EQ.JSAVE) GO TO 50
   42 NL=NL+1
C
C.....STORE INDICES OF CHOSEN COEFFICIENT IN ILIST(NL) AND JLIST(NL)
      ILIST(NL)=I
      JLIST(NL)=JSAVE
   50 CONTINUE
   51 NJ=NL
C
C.....STORE INDICES OF UNGROUPED SAMPLES IN JLIST(NJ)
      DO 60 I=1,NS
      DO 55 J=1,NL
   55 IF(I.EQ.ILIST(J).OR.I.EQ.JLIST(J)) GO TO 60
      NJ=NJ+1
      JLIST(NJ)=I
   60 CONTINUE
C
C.....OUTPUT INFORMATION FOR INTERMEDIATE STORAGE AND PRINTED OUTPUT
      WRITE(3)      NL, NJ
      DO 68 J=1,NL
      IJ=ILIST(J)
      JJ=JLIST(J)
      WRITE(6,7) INDEX(IJ), INDEX(JJ), COEF(IJ,JJ), NCYC
      COSAV(J)=COEF(IJ,JJ)
   68 CONTINUE
      WRITE(3)      (ILIST(J), JLIST(J), COSAV(J), J=1,NL)
   69 IF (NL.NE.NJ) GO TO 70
      WRITE(3)      KDUM
      GO TO 71
   70 NLP1=NL+1
      WRITE(3)      (JLIST(I), I=NLP1, NJ)
C
C.....DETERMINE LOWEST SAMPLE NUMBER FROM EACH CLUSTER FOR USE AS LABEL
   71 DO 80 I=1,NJ
```

```
      JI=JLIST(I)
      IF (I.GT.NL) GO TO 79
      II=ILIST(I)
      JI=MIN0(II,JI)
   79 COSAV(I)=INDEX(JI)
   80 CONTINUE
      DO 90 I=1,NS
   90 INDEX(I)=COSAV(I)
      RETURN
    1 FORMAT(1H0, 11X, 14HSAMPLE NUMBERS, 7X, 20HLEVEL OF ASSOCIATION,
     1  3X, 12HCYCLE NUMBER)
    7 FORMAT(1H , 12X, I4, 5X, I4, 12X, F9.4, 12X, I5)
      END
$IBFTC RCALC   NODECK
C
C     *****************
      SUBROUTINE RECALC
C     *****************
      DIMENSION SIZ(130), DUM(130)
      COMMON TITLE(12), NAME(130)
      COMMON NCYC, NS, NL, NJ, NT, XLNG, CON, NDORD, KDEN, NSHOLD,
     1 BOTVAL, NWT, KOEF, COHOLD
      COMMON COEF(130,130), COSAV(130), ILIST(65), JLIST(130)
C
C.....CALCULATES NEW VALUES OF ASSOCIATION BETWEEN ALL COMBINATIONS OF
C.....PAIRED AND UNPAIRED SAMPLES (OR OF GROUPS OF SAMPLES) USING
C.....ARITHMETIC AVERAGES
      IF (NCYC.NE.1) GO TO 2
C
C.....SET SIZ() TO 1 FOR EACH SAMPLE. EACH GROUP STARTS AT SIZE 1.
      DO 1 I=1,NS
    1 SIZ(I)=1.0
    2 REWIND 2
      DO 50 I=1,NJ
      IJ=JLIST(I)
      IF (I.LE.NL) II=ILIST(I)
      DO 40 J=I,NJ
      IF (J.NE.I) GO TO 5
      COSAV(J)=0.0
      GO TO 40
    5 JJ=JLIST(J)
      IF (J.LE.NL) JI=ILIST(J)
      IF (J.GT.NL.AND.I.GT.NL) GO TO 30
      IF (J.GT.NL.AND.I.LE.NL) GO TO 10
      IF (J.LE.NL.AND.I.GT.NL) GO TO 20
      COSAV(J)=(COEF(II,JI)*SIZ(II)*SIZ(JI)+COEF(II,JJ)*SIZ(II)*SIZ(JJ)+
     1COEF(IJ,JI)*SIZ(IJ)*SIZ(JI)+COEF(IJ,JJ)*SIZ(IJ)*SIZ(JJ))/
     2(SIZ(II)*SIZ(JI)+SIZ(II)*SIZ(JJ)+SIZ(IJ)*SIZ(JI)+SIZ(IJ)*SIZ(JJ))
      GO TO 40
   10 COSAV(J)=(COEF(II,JJ)*SIZ(II)*SIZ(JJ)+COEF(IJ,JJ)*SIZ(IJ)*SIZ(JJ))
     1/(SIZ(II)*SIZ(JJ)+SIZ(IJ)*SIZ(JJ))
      GO TO 40
   20 COSAV(J)=(COEF(IJ,JI)*SIZ(IJ)*SIZ(JI)+COEF(IJ,JJ)*SIZ(IJ)*SIZ(JJ))
     1/(SIZ(IJ)*SIZ(JI)+SIZ(IJ)*SIZ(JJ))
      GO TO 40
   30 COSAV(J)=COEF(IJ,JJ)
   40 CONTINUE
      WRITE(2)      (COSAV(J), J=I, NJ)
   50 CONTINUE
      REWIND 2
```

```
C
C.....OVERWRITE COEF(,) WITH ARRAY OF NEW COEFFICIENTS
      DO 60 I=1,NJ
   60 READ(2)      (COEF(I,J), J=I,NJ)
      REWIND 2
C
C.....COMPLETE LOWER (SYMMETRICAL) HALF OF ARRAY
      DO 70 I=1,NJ
      DO 70 J=I,NJ
   70 COEF(J,I)=COEF(I,J)
      NS=NJ
      IF (NWT.EQ.0) RETURN
C
C.....FOR THE UNWEIGHTED METHOD, CALCULATE THE SIZE OF EACH GROUP AND
C....STORE IN DUM()
      DO 80 I=1,NS
      II=ILIST(I)
      IJ=JLIST(I)
      IF (I.GT.NL) GO TO 75
      DUM(I)=SIZ(II)+SIZ(IJ)
      GO TO 80
   75 DUM(I)=SIZ(IJ)
   80 CONTINUE
C
C.....OVERWRITE SIZ() WITH DUM()
      DO 90 I=1,NS
   90 SIZ(I)=DUM(I)
      RETURN
      END
$IBFTC ORDR    NODECK
C
C     ****************
      SUBROUTINE ORDER
C     ****************
      DIMENSION DUM(65)
      COMMON TITLE(12), NAME(130)
      COMMON NCYC, NS, NL, NJ, NT, XLNG, CON, NDORD, KDEN, NSHOLD,
     1 BOTVAL, NWT, KOEF, COHOLD
      COMMON MLIST(130), NLIST(65), NNLIST(130), NUM(130,130)
      EQUIVALENCE(DUM,NUM)
C
C.....PUTS SAMPLES IN CORRECT ORDER FOR THE DENDROGRAM
C.....WORKS BACKWARDS FROM THE LAST CYCLE TO THE FIRST CYCLE.
      KDUM=9999
      NSAM=NS
      MLIST(1)=1
      NC=NCYC
   10 IF (NC) 60,60,20
C
C.....FOR EACH CYCLE, THE INDICES OF THOSE GROUPS LINKED DURING THAT
C.....CYCLE ARE READ FROM TAPE 3, THE REVERSED ORDER BEING ACHIEVED
C.....BY BACKSPACING
   20 DO 21 I=1,3
   21 BACKSPACE 3
      READ(3)      NL, NJ
      READ(3)      (NLIST(I), NNLIST(I), DUM(I), I=1,NL)
      IF (NL.NE.NJ) GO TO 22
      READ(3)      KDUM
      GO TO 23
   22 NLP1= NL+1
```

```
      READ(3)      (NNLIST(I), I=NLP1, NJ)
   23 DO 24 I=1,3
   24 BACKSPACE 3
      NS=NJ+NL
      NLAST=NJ
C
C.....FOR EACH CYCLE THE INDICES ARE ARRANGED IN DENDROGRAM ORDER
C.....IN MLIST()
   30 INDEX=MLIST(NLAST)
      MLIST(NS)= NNLIST(INDEX)
      NS=NS-1
      IF (INDEX-NL) 40,40,50
   40 MLIST(NS)=NLIST(INDEX)
      NS=NS-1
   50 NLAST=NLAST-1
      IF (NLAST.GT.0) GO TO 30
      NC=NC-1
      GO TO 10
C
C.....THE FINAL MLIST() IS PRINTED
   60 WRITE(6,1) TITLE
      DO 65 K=1,NSAM
      I=MLIST(K)
   65 WRITE(6,8) NAME(I),I
      IF (NDORD.EQ.0) RETURN
C
C.....PRINT ORIGINAL DATA MATRIX IN DENDROGRAM ORDER
      REWIND 1
      READ(1)      NUM
      WRITE(6,7) TITLE
      WRITE(6,2)
      DO 70 K=1,NSAM
      I=MLIST(K)
      IF (NDORD.EQ.2) WRITE(6,3) NAME(I),(NUM(I,J),J=1,NT)
   70 WRITE(6,6) NAME(I),I, (NUM(I,J), J=1,NT)
      RETURN
    1 FORMAT(1H1,5X,12A6//6X,36HORDER OF SAMPLES FOR DENDROGRAM PLOT//)
    2 FORMAT(1H0, 5X, 45HPRINTOUT OF ORIGINAL DATA IN DENDROGRAM ORDER//
     1 /4X, 7HSAMPLES,10X, 10HCHARACTERS/1X, 12HNAME     NO.//)
    3 FORMAT(1H$, A6, 3X, 63I1, 8X, (9X, 63I1, 8X))
    6 FORMAT(1H , A6, I6, 3X, 100I1)
    7 FORMAT(1H1, 5X, 12A6)
    8 FORMAT(1H ,A6,I6)
      END
$IBFTC DNDRO   NODECK
C
C     *****************
      SUBROUTINE DENDRO
C     *****************
      COMMON TITLE(12), NAME(130)
      COMMON NCYC, NS, NL, NJ, NT, XLNG, CON, NDORD, KDEN, NSHOLD,
     1 BOTVAL, NWT, KOEF, COHOLD
      COMMON MLIST(130), NLIST(65), NNLIST(130), Y(2,130), X(130),
     1  U(4), V(4), NAMDEN(130)
C
C.....CALCULATES X-Y COORDINATES OF POINTS TO BE LINKED FOR FORMING
C....A DENDROGRAM.
      NSAM=NS
      IF (BOTVAL.GT.COHOLD) BOTVAL=COHOLD
      IF (CON.EQ.0.0) CON=1.0
```

```
      IF (XLNG.GT.(FLOAT(NS)/4.0).OR.XLNG.LT.(FLOAT(NS)/7.0))XLNG=
     1FLOAT(NS)/4.0
      WRITE(6,8) TITLE
      WRITE(6,2) XLNG, CON, BOTVAL
      IF (KDEN.EQ.2) WRITE(6,3) (I, I=1,4)
      XDV=FLOAT(NS)/XLNG
      HT=0.6/XDV
      DO 10 I=1, NS
   10 X(I)=FLOAT(I)/XDV
C
C.....PUNCH TITLE, CONTROLS AND NAMES IN DENDROGRAM ORDER
      WRITE(6,9) TITLE
      WRITE(6,5) NS, XLNG, HT, CON, BOTVAL, NWT, KOEF
      DO 20 I=1,NS
      K=MLIST(I)
   20 NAMDEN(I)=NAME(K)
      WRITE (6,6) (NAMDEN(K), K=1,NS)
      DO 30 I=1,130
   30 Y(1,I)=1.0
      REWIND 3
      DO 150 N=1,NCYC
C
C.....FOR EACH CLUSTER CYCLE, READ INFORMATION REGARDING LINKAGES FROM 3
      READ(3)       NL, NJ
      READ(3)       (NLIST(I), NNLIST(I), Y(2,I), I=1, NL)
      IF (NL.NE.NJ) GO TO 48
      READ(3)       KDUM
      GO TO 49
   48 NLP1=NL+1
      READ(3)       (NNLIST(I), I=NLP1, NJ)
C
C.....SCALE VALUES OF ASSOCIATION ACCORDING TO SPECIFIED RANGE
   49 DO 50 I=1,NL
   50 Y(2,I)= (((CON-Y(2,I))/(CON-BOTVAL) *9.0) + 1.0)
      K=0
      L=1
   60 K=K+1
      IF (K-NS) 70,70,130
   70 NM=MLIST(K)
C
C.....COORDINATES OF FOUR POINTS TO BE JOINED BY CONTINUOUS LINE,
C.....SO FORMING A LINKAGE, ARE CALCULATED
      DO 80 J=1,NJ
      IF (J.LE.NL.AND.NLIST(J).EQ.NM) GO TO 100
      IF (NNLIST(J).EQ.NM) GO TO 90
   80 CONTINUE
   90 IF (J.LE.NL) GO TO 100
      Y(2,J)= Y(1,NM)
      MLIST(L)=J
      X(L)=X(K)
      L=L+1
      GO TO 60
  100 NNM=MLIST(K+1)
  120 U(1)=X(K)
      U(2)=X(K)
      U(3)=X(K+1)
      U(4)=X(K+1)
      V(1)=Y(1,NM)
      V(2)=Y(2,J)
      V(3)=Y(2,J)
```

```
      V(4)=Y(1,NNM)
      IF (KDEN.EQ.2) WRITE(6,4) (U(I), V(I), I=1,4)
C
C.....COORDINATES PUNCHED ON CARDS
      WRITE(6,7) (U(I), V(I), I=1,4)
      MLIST(L)=J
      X(L)=(X(K) + X(K+1)) / 2.0
      K=K+1
      L=L+1
      GO TO 60
  130 NS=NJ
      DO 140 J=1,NSAM
  140 Y(1,J)= Y(2,J)
  150 CONTINUE
      VAL= 99.9
      WRITE(6,7) VAL
C
C.....COORDINATES OF FINAL STEM
      V(1)=V(2)
      V(2)=10.0
      U(1)= (U(2) + U(3)) / 2.0
      U(2)= U(1)
      WRITE(6,7) (U(I), V(I), I=1,2)
      IF (KDEN.EQ.2)  WRITE(6,4) (U(I), V(I), I=1,2)
      RETURN
    2 FORMAT(1H0, 37HINFORMATION REGARDING DENDROGRAM PLOT//
     1   1X, 18HLENGTH OF X-AXIS =, F6.0, 2X, 6HINCHES//,
     2   1X, 23HY-AXIS PLOTTED BETWEEN , F5.1, 1X, 4HAND , F5.1////)
    3 FORMAT(1H ,65HCOORDINATES IN INCHES OF POINTS TO BE JOINED BY A CO
     1NTINUOUS LINE/1X, 38HEACH ROW OF FOUR POINTS MUST BE LINKED/
     21X,  46HORIGIN AT THE TOP LEFT-HAND CORNER OF THE PAGE//
     34X, 4(8X, I2, 13X)/ 4X, 4(4X, 1HX,9X,1HY,8X)/)
    4 FORMAT(1H , 4X, 4(F7.2, 3X, F7.2, 6X))
    5 FORMAT(1H$, I5, 4F10.5, 2I5)
    6 FORMAT(1H$, 12A6)
    7 FORMAT(1H$,8F9.5)
    8 FORMAT(1H1, 12A6)
    9 FORMAT(1H$, 12A6)
      END
```

```
$IBJOB
$IBFTC DNPLT   NODECK
C.....DENDROGRAM PLOTTING ROUTINE (DNPLOT) FOR USE WITH 10 INCH CALCOMP
C.....PLOTTER. PROGRAM COMPATIBLE WITH FORTRAN IV PLOTTING ROUTINES IN
C.....USE AT STANFORD UNIVERSITY - SEE STANFORD UNIV. COMP. CENTER
C.....LIBRARY PROGRAM NUMBER 120, FEBRUARY 1966.
C.....THIS VERSION CODED BY G.BONHAM-CARTER NOVEMBER 1966.
C.....ACCEPTS INPUT PUNCHED BY CLUST3
      DIMENSION X(130),   TAG(130), U(4), V(4), TITLE(12)
C.....SPECIFIES THE USE OF .001 INCH RESOLUTION PLOTTER
      CALL PLOTSZ(100.0)
C.....READ NUMBER OF DECKS TO BE RUN ON THIS JOB
      READ(5,1) NUMDEC
C.....INITIALISATION AND IDENTIFICATION
      CALL PLOTS
      DO 70 ND=1,NUMDEC
      RND=ND
C.....READ CONTROLS AND SAMPLE ORDER
      READ(5,5) TITLE
      READ(5,2) NS, XLNG, HT, CON, BOTVAL, NWT, KOEF
      READ(5,3)  (TAG(I), I=1,NS)
C.....PLOT TITLE AND MOVE PLOTTING ORIGIN FORWARD
      CALL SYMBOL(0.0,0.0,0.2,23HDENDROGRAM PLOT NUMBER ,90.0,23)
      CALL NUMBER(0.0,4.0,0.2,RND, 90.0, -1)
      CALL SYMBOL(0.5,0.0,0.2,TITLE,90.0,48)
      IF (KOEF.EQ.3)CALL SYMBOL(1.0,0.0,0.2,39HSIMILARITY COEFFICIENTS R
     1EAD IN AS DATA,90.0,39)
      IF (KOEF.EQ.1) CALL SYMBOL(1.0,0.0,0.2,36HJACCARDS COEFFICIENTS OF
     1 ASSOCIATION,90.0,36)
      IF (KOEF.EQ.0) CALL SYMBOL(1.0,0.0,0.2,32HSOKAL AND MICHENERS COEF
     1FICIENTS,90.0,32)
      IF (NWT.EQ.1) CALL SYMBOL(1.5,0.0,0.2,28HUNWEIGHTED PAIR-GROUP MET
     1HOD,90.0,28)
      IF (NWT.EQ.0) CALL SYMBOL(1.5,0.0,0.2,26HWEIGHTED PAIR-GROUP METHO
     1D,90.0,26)
      CALL PLOT(4.0, 0.0, -3)
C.....PLOT SAMPLE NUMBERS ALONG X-AXIS  (LONG AXIS OF PAGE)
      DO 20 K=1,NS
      A=K
      A=(A*HT/0.6) + (HT/2.0)
      C=TAG(K)
   20 CALL SYMBOL(A, 0.0, HT, C, 90.0, 6)
C.....DRAW AND LABEL Y-AXIS (HEIGHT OF PAGE) WITH SCALE OF SIMILARITY
      U(1)=-(HT*2.5)
      U(2)=U(1)
      V(1)=1.0
      V(2)=10.0
C.....DRAW THE AXIS
      CALL LINE(U, V, 2, 1, 1)
      U(2)=U(1) + 0.05
      UO=U(1) - 0.05
      DUMB=CON
      DO 40 K=1,11
      V(2)=V(1)
C.....DRAW THE TIC MARKS
      CALL LINE(U, V, 2, 1, 1)
      VO=V(1) - 0.12
C.....LABEL THE INTERVALS
      CALL NUMBER(UO, VO, 0.13,DUMB, 90.0, 2)
      DUMB=DUMB - ((CON-BOTVAL)/10.0)
```

```
   40 V(1)=V(1) + 0.9
      UO=U(1) - 0.4
C.....LABEL THE AXIS
      CALL SYMBOL(UO, 4.0, 0.20, 22HSIMILARITY COEFFICIENT, 90.0, 22)
C.....READ COORDINATES OF POINTS TO BE LINKED
   50 READ(5,4) (U(I), V(I), I=1,4)
      IF (U(1).EQ. 99.9) GO TO 60
C.....JOIN 4 POINTS HELD IN U() AND V() ARRAYS WITH A CONTINUOUS LINE
      CALL LINE(U, V, 4, 1, 1)
      GO TO 50
   60 READ(5,4) (U(I), V(I), I=1,2)
C.....JOIN 2 POINTS FORMING FINAL STEM
      CALL LINE(U, V, 2, 1,1)
C.....MOVE PLOTTING ORIGIN FORWARD
      XLNG=XLNG+8.0
      CALL PLOT(XLNG, 0.0, -3)
   70 CONTINUE
C.....TERMINATE THIS PLOT
      CALL PLOTE
      STOP
    1 FORMAT(I3)
    2 FORMAT(I5, 4F10.5, 2I5)
    3 FORMAT(12A6)
    4 FORMAT(8F9.5)
    5 FORMAT(12A6)
      END
```

REFERENCES

Behrens, E.W., 1965, Environment reconstruction for a part of the Glen Rose Limestone, Central Texas: Sedimentology, v. 4, p. 65-111.

Bonham-Carter, G.F., 1965, A numerical method of classification using qualitative and semi-quantitative data, as applied to the facies analysis of limestones: Canadian Petroleum Geology Bull., v. 13, p. 482-502.

Bonham-Carter, G.F., 1967, An example of the analysis of semi-quantitative petrographic data: Panel Discussion No. 5., Proceedings of Seventh World Petroleum Congress, Mexico City.

Cassetti, E., 1964, Classificatory and regional analysis by discriminant iterations: Office of Naval Research, Geography Branch, Tech. Rept. 12, ONR Task No. 389-135, 95 p.

Friedman, H.P., and Rubin, J., 1966, On some invariant criteria for grouping data: Invited Paper presented orally to Biometric Session on Cluster Analysis, Statistical Meetings, Brookhaven.

Houbolt, J.J.H.C., 1957, Surface sediments of the Persian Gulf near the Qatar Peninsula: Mouton and Co., The Hague, 113 p.

Howd, F.H., 1964, The taxonomy program- a computer technique for classifying geologic data: Colorado Scho. Mines Quart., v. 59, no. 4, p. 207-222.

Kaesler, R.L., 1966, Quantitative re-evaluation of ecology and distribution of Recent Foraminifera and Ostracoda of Todos Santos Bay, Baja, California: Univ. Kansas Paleontological Contr., Paper 10, 50 p.

Klovan, J.E., 1964, Facies analysis of the Redwater reef complex, Alberta, Canada: Canadian Petroleum Geology Bull., v. 12, p. 1-100.

Kruskal, J., 1964, Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis: Psychometrika, v. 29, p. 1-28.

Parks, J.M., 1966, Cluster analysis applied to multivariate geologic problems: Jour. Geology, v. 74, no. 5, pt. 2, p. 703-715.

Purdy, E.G., 1963, Recent calcium carbonate facies of the Great Bahama Bank. 1. Petrography and reaction groups: Jour. Geology, v. 71,

Rubin, J., 1965, Optimal taxonomy program, (7090-IBM-0026): IBM, Program Information Dept., Hawthorne, New York.

Shephard, R.N., 1962, The analysis of proximities: Multidimensional scaling with an unknown distance function I: Psychometrika, v. 27, p. 125-140.

Siegel, S., 1956, Nonparametric statistics for the behavioral sciences: McGraw Hill Co., New York, 312 p.

Sokal, R.R., and Sneath, P.H.A., 1963, Principles of numerical taxonomy: W.H. Freeman Co., San Francisco, 359 p.

KANSAS GEOLOGICAL SURVEY COMPUTER PROGRAM
THE UNIVERSITY OF KANSAS, LAWRENCE

PROGRAM ABSTRACT

Title (If subroutine state in title): CLUST3

FORTRAN IV program for Q-mode cluster analysis of nonquantitative data using IBM 7090/7094 computers. Version for the IBM 360/67 also available.

Computer: IBM 7090/7094 Date: March, 1967

Programming language: FORTRAN IV, Version 13

Author, organization: G.F. Bonham-Carter, Stanford University
Program developed while at University of Toronto

Direct inquiries to: Author or

Name: Daniel F. Merriam Address: Kansas Geological Survey, Univ. of Kansas
Lawrence, Kansas 66044

Purpose/description: Classification of objects (geological samples, etc.) on the basis of a large number of qualitative and/or semiquantitative attributes.

Mathematical method: Option of Jaccard's or Sokal and Minchener's matching coefficients; weighted or unweighted pair-group method using arithmetic averages.

Restrictions, range: Limited to 130 objects, 100 attributes

Storage requirements: Three scratch tapes for intermediate storage

Equipment specifications: Memory 20K___ 40K___ 60K___ 32 K___

Automatic divide: Yes___ No___ Indirect addressing Yes___ No___

Other special features required___

Additional remarks (include at author's discretion: fixed/float, relocatability; optional: running time, approximate number of times run successfully, programming hours) Punches output for use with DNPLOT, a routine for plotting a dendrogram with 10" CALCOMP plotter. Running time: 130 objects, 25 attributes - about 2 mins by either weighted or unweighted methods.

KANSAS GEOLOGICAL SURVEY COMPUTER PROGRAM
THE UNIVERSITY OF KANSAS, LAWRENCE

PROGRAM ABSTRACT

Title (If subroutine state in title): DNPLOT

FORTRAN IV program for drawing dendrograms using IBM 7090/7094 computers in conjunction with a 10" CALCOMP plotter.

Computer: 7090/7094, CALCOMP plotter Date: March, 1967

Programming language: FORTRAN IV, Version 13

Author, organization: G.F. Bonham-Carter, Stanford University
Original program developed while at University of Toronto

Direct inquiries to: Author or

Name: Daniel F. Merriam Address: Kansas Geological Survey, Univ. of Kansas
Lawrence, Kansas 66044

Purpose/description: Plotting dendrograms, using coordinates punched by the cluster analysis program CLUST3.

Mathematical method: Self-explanatory

Restrictions, range:

Storage requirements:

Equipment specifications: Memory 20K___ 40K___ 60K___ K___

Automatic divide: Yes___ No___ Indirect addressing Yes___ No___

Other special features required 10" CALCOMP plotter

Additional remarks (include at author's discretion: fixed/float, relocatability; optional: running time, approximate number of times run successfully, programming hours) CALCOMP routines are those in use at Stanford Computation Center, and are described in Library Program No. 120, Stanford Computation Center. Probably only minor modifications will be required at different installations.

COMPUTER CONTRIBUTIONS

Kansas Geological Survey
University of Kansas
Lawrence, Kansas

Computer Contribution

1.	Mathematical simulation of marine sedimentation with IBM 7090/7094 computers, by J.W. Harbaugh, 1966	$1.00
2.	A generalized two-dimensional regression procedure, by J.R. Dempsey, 1966	$0.50
3.	FORTRAN IV and MAP program for computation and plotting of trend surfaces for degrees 1 through 6, by Mont O'Leary, R.H. Lippert, and O.T. Spitz, 1966	$0.75
4.	FORTRAN II program for multivariate discriminant analysis using an IBM 1620 computer, by J.C. Davis and R.J. Sampson, 1966	$0.50
5.	FORTRAN IV program using double Fourier series for surface fitting of irregularly spaced data, by W.R. James, 1966.	$0.75
6.	FORTRAN IV program for estimation of cladistic relationships using the IBM 7040, by R.L. Bartcher, 1966	$1.00
7.	Computer applications in the earth sciences: Colloquium on classification procedures, edited by D.F. Merriam, 1966	$1.00
8.	Prediction of the performance of a solution gas drive reservoir by Muskat's Equation, by Apolonio Baca, 1967.	$1.00
9.	FORTRAN IV program for mathematical simulation of marine sedimentation with IBM 7040 or 7094 computers, by J.W. Harbaugh and W.J. Wahlstedt, 1967.	$1.00
10.	Three-dimensional response surface program in FORTRAN II for the IBM 1620 computer, by R.J. Sampson and J.C. Davis, 1967	$0.75
11.	FORTRAN IV program for vector trend analyses of directional data, by W.T. Fox, 1967	$1.00
12.	Computer applications in the earth sciences: Colloquium on trend analysis, edited by D.F. Merriam and N.C. Cocke, 1967	$1.00
13.	FORTRAN IV computer programs for Markov chain experiments in geology, by W.C. Krumbein, 1967	$1.00
14.	FORTRAN IV programs to determine surface roughness in topography for the CDC 3400 computer, by R.D. Hobson, 1967	$1.00
15.	FORTRAN II program for progressive linear fit of surfaces on a quadratic base using an IBM 1620 computer, by A.J. Cole, C. Jordan, and D.F. Merriam, 1967	$1.00
16.	FORTRAN IV program for the GE 625 to compute the power spectrum of geological surfaces, by J.E. Esler and F.W. Preston, 1967.	$0.75
17.	FORTRAN IV program for Q-mode cluster analysis of nonquantitative data using IBM 7090/7094 computers, by G.F. Bonham-Carter, 1967	$1.00