Special
Distribution
Publication

1. The Kansas mineral industry . . . 1962; with directory of Kansas mineral producers, by Grace Muilenburg, R. G. Hardy, and Allison Hornbaker, 1963.
2. Economic development for Kansas, mineral and water resources: Report of the Governor's Economic Development Committee, by W. W. Hambleton, and others, 1962.
3. BALGOL program for trend-surface mapping using an IBM 7090 computer, by J. W. Harbaugh, 1963.
4. FORTRAN II program for coefficient of association (Match-Coeff) using an IBM 1620 computer, by R. L. Kaesler, F. W. Preston, and Donald Good, 1963.
5. Activities of State Geological Survey of Kansas, biennium ending June 30, 1963, by Grace Muilenburg, 1963.
6. Summary secondary recovery operations in Kansas during 1962: Report of the Kansas Secondary Recovery Committee, by E. D. Goebel and M. C. Colt, 1964.
7. The Kansas mineral industry . . . 1963; with directory of Kansas mineral producers, by Allison Hornbaker and R. G. Hardy, 1964.
8. Annotated bibliography of the Kansas Precambrian, by D. F. Merriam, 1964.
9. BALGOL programs for calculation of distance coefficients and correlation coefficients using an IBM 7090 computer, by J. W. Harbaugh, 1964.

STATE GEOLOGICAL SURVEY OF KANSAS

W. Clarke Wescoe, M. D., Chancellor of The University and ex officio Director of the Survey

Frank C. Foley, Ph.D., State Geologist and Director

William W. Hambleton, Ph.D., Assoc. State Geologist and Assoc. Director

Raymond C. Moore, Ph.D., Sc.D., Principal Geologist Emeritus

John M. Jewett, Ph.D., Senior Geologist

Doris E. Nodine Zeller, Ph.D., Editor

Grace E. Muilenburg, B.S., Public Information Director

Kenneth J. Badger, Chief Draftsman

Lila M. Watkins, Secretary

Research Divisions	Division Head
Basic Geology	Daniel F. Merriam, Ph.D.
Petrography	Ada Swineford, Ph.D.
Geochemistry	Frederic R. Siegel, Ph.D.
Mineral Resources	Allison L. Hornbaker, M.S.
Oil and Gas	Edwin D. Goebel, M.S.
Ceramics	Norman Plummer, A.B.

Cooperative Studies with United States Geological Survey

Ground-Water Resources . . . . . Robert J. Dingman, B.S., District Geologist

Mineral Fuels . . . . . W. L. Adkison, B.S., geologist in charge

Branch Offices

Geological Survey Well Sample Library, 4150 Monroe, Wichita

Geological Survey Southwest Kansas Field Office, 310 N. 9th Street, Garden City

KANSAS GEOLOGICAL SURVEY COMPUTER PROGRAM
THE UNIVERSITY OF KANSAS, LAWRENCE

PROGRAM ABSTRACT

Title (If subroutine state in title):

Program for calculation of correlation coefficients

Computer: IBM 7090 or 7094 Date: December, 1963

Programming language: BALGOL, also known as SUBALGOL, a dialect of ALGOL-58

Author, organization: John W. Harbaugh, Department of Geology
Stanford University, Stanford, California

Direct inquiries to: Author, or

Name: Daniel F. Merriam Address: State Geological Survey
Lawrence, Kansas

Purpose/description: Calculate Pearson product-moment correlation coefficients on either, or both, a column-by-column basis and row-by-row basis. Has a variety of options, including deletions of specified columns, data transformation for row-by-row calculations, and choice of printing and punching either upper or lower half of correlation matrix.

Mathematical method:

Restrictions, range:

Storage requirements: 32,768 words high-speed memory

Equipment specifications:

Memory 20K 40K 60K 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)

(5) If row-by-row calculations have been performed, and if specified, the transformed data array is printed and punched. In addition, the type of transformation is specified.

(6) A table listing row number, row mean, and standard deviation is printed out. This is followed by printing and punching of either the upper or the lower half of the correlation matrix.

REFERENCES

Harbaugh, J. W., and Demirmen, Ferruh, in press, Application of factor analysis to petrologic variations of Americus Limestone (Lower Permian), Kansas and Oklahoma: Kansas Geol. Survey Bull.

Imbrie, John, and Purdy, E. G., 1962, Classification of modern Bahamian carbonate sediments; in Classification of Carbonate Rocks: Am. Assoc. Petroleum Geologists Memoir 1, p. 253-272.

Krumbein, W. C., and Imbrie, John, 1963, Stratigraphic factor maps: Am. Assoc. Petroleum Geologists Bull., v. 47, no. 4, p. 698-701.

Schwartz, J. T., 1961, Introduction to matrices and vectors: McGraw-Hill Book Co., p. 1-163.

Sokal, R. R., 1961, Distance as a measure of taxonomic similarity: Systematic Zoology, v. 10, no. 2, p. 70-79.

_____, and Sneath, P. H., 1963, Principles of numerical taxonomy: W. H. Freeman & Co., p. 1-359.

data set must be either all integers or all decimal-point numbers, and must agree with the type specified on the first control card (TOP). The data will be read from left to right and will be stored as an array of M columns and N rows. Ordinarily, variables are arranged by columns and observations by rows.

Correlation Coefficient Program Limitations and Operating Times

Array limitations:--The following limitations have been placed in use of the program by the present array dimensions (lines 6 and 7, Table 7). Original data arrays are limited to 200 rows and 20 columns. However, if row-by-row calculations are to be performed, no more than 120 rows may be present in the original data array. The limit of 120 is dictated by the fact that 14,400 elements are occupied by the array, PC (,), which contains the correlation matrix.

Depending on the circumstances, the dimensions of the array PC (,) could be increased if some of the other arrays were reduced in size. For example, if the data are in decimal point form, the Z(200,20) array could be cut to token dimensions, Z(1,1). Similarly, if there are fewer than 20 columns in the original data array, its width can be reduced, thus freeing space in memory for the PC (,) array.

Operating times:--The program compiles in about 10 seconds. Execution time depends on operations performed and dimensions of the data array. Calculation of a 54 x 54 correlation matrix, preceded by data transformation, requires less than 10 seconds with the IBM 7090 computer.

Output From Correlation Coefficient Program

The types of output data from the correlation coefficient program (examples are shown in Tables 11, 12, 13) are listed in order below.

(1) Alphanumeric headings for identification purposes at the tops of some pages.

(2) Statements which identify the type of data being written are printed out at the head of each table.

(3) If specified on the control card, the original data array, without columns that have been specified for deletion, is printed and punched.

(4) If column-by-column calculations have been performed, a table listing the column number, column mean, and column standard deviations of original data is printed. This is followed by printing and punching of either the upper or the lower half of the correlation matrix.

(4) An integer (N) specifying number of rows in the data array.

(5) An integer (OP) specifying whether calculations are to be performed on a column-by-column or row-by-row basis, as follows:

0 Calculate both on a column-by-column basis and then on a row-by-row basis.

1 Calculate on a column-by-column basis only.

2 Calculate on a row-by-row basis only.

(6) An integer (C) specifying whether calculations on a row-by-row basis call for transformation of data, as follows:

0 Make calculations with both transformations and print out two complete sets of means, standard deviations, and correlation coefficients.

1 Transform so as to express data in terms of unit standard deviations per each column.

2 Transform by dividing by highest value in each column.

(7) An integer (CHC) specifying whether upper or lower half of matrix of coefficients is to be printed out, as follows:

7777 Output upper half of matrix with ones in principal diagonal omitted.

1234 Output lower half of matrix with ones of principal diagonal included.

(8) An integer (Q) specifying whether columns in original data array are to be omitted as follows:

4444 Omit one or more columns.

1234 Do not omit any columns.

(9) An integer (DATLIST) specifying whether original and transformed data are to be printed and punched, as follows:

0 Do not print and punch list of data.

1 Print and punch data.

Second program control card:--If columns of original data array are to be deleted (Q = 4444 above), then a series of 0's or 1's must be placed on the third data card. The total number of 0's and 1's must be equal to the number of columns in original data array. Punch 1 if the column is to be retained, 0 if it is to be deleted.

Example:

5 1 1 0 1 0 1 1 1 1 0

This example specifies that of an original data array of 10 columns, columns 3, 5, and 10 are to be deleted. Punch 5 in column one. DO NOT include this card if Q is not equal to 4444 on second data card.

Data Cards:--Data cards follow the control card. Each data card should have 5 punched in column 1. Data values must be separated by at least one space, and any convenient number of data values may be placed on a card, columns 2 through 80 being available. In BALGOL, format input specifications for data ordinarily are not used. The values for a given

```
5$AMERICUS CHEM DATA WITH COLUMNS 5 AND 7 OF ORIG DATA ARRAY DELETED      $
CORRELATION COEFFICIENTS FOR TRANSPOSED, NORMALIZED ARRAY, EACH X VALUE DIVIDED BY HIGHEST
    VALUE FOR THAT VARIABLE

ARRAY CONSISTING OF ORIGINAL DATA DIVIDED BY HIGHESTVALUE IN COLUMN

5   .500   .680   .526   .030   .426   .049
5  1.000   .320   .579   .040   .369   .002
5   .542   .260   .579  1.000   .277   .004
5   .667   .580   .474   .030   .206   .003
5   .708   .520   .684   .120   .241   .003
5   .142   .460   .947   .100   .551   .003
5   .183   .720   .684   .020   .423   .004
5   .250   .620   .684   .060   .409   .004
5   .250   .760   .579   .090   .362   .004
5   .083   .540   .632  1.000   .331   .005
5   .167   .660   .737   .280   .430   .007
5  1.000   .320   .579   .180   .733   .004
5   .292   .580   .789   .040  1.000   .004
5   .250  1.000  1.000   .300   .931   .005
5   .067  1.000   .842   .300   .910  1.000

  ROW NO.        MEAN         STANDARD DEV.

     1            .3685            .2446
     2            .3849            .3382
     3            .4436            .3141
     4            .3265            .2610
     5            .3794            .2736
     6            .3672            .3250
     7            .3391            .2914
     8            .3378            .2586
     9            .3408            .2638
    10            .4318            .3387
    11            .3800            .2593
    12            .4693            .3381
    13            .4508            .3712
    14            .5811            .4070
    15            .6865            .3661

CORRELATION COEFFICIENTS OF UPPER HALF OF MATRIX WITH ONES IN PRINCIPAL DIAGONAL OMITTED

CORRELATION COEFFICIENTS BETWEEN ALL POSSIBLE PAIRS OF ROWS OF ORIGINAL DATA ARRAY

5   .664 -.175  .897  .835  .646  .871  .885  .904 -.085  .713  .609  .708  .757
5   .157  .122  .849  .880  .298  .310  .415  .333 -.310  .184  .920  .350  .197
5  -.487  .008  .178  .032 -.151 -.068 -.098  .773  .155  .147 -.168 -.019 -.706
5   .952  .407  .629  .678  .705 -.151  .481  .686  .393  .443 -.213  .556  .637
5   .717  .671 -.020  .558  .722  .440  .478 -.268  .868  .904  .760  .287  .896
5   .342  .853  .889  .405  .989  .977  .199  .936  .285  .798  .938  .479  .959
5   .206  .939  .389  .817  .931  .379  .231  .911  .281  .705  .896  .381  .519
5  -.249  .052  .357 -.128  .190  .739  .945  .366  .538  .302 -.450  .898  .409
5   .457
```

First program control card:--The second card that accompanies each data set must be a program-control card containing the following information in the order listed below. Note that each number is to be separated from adjacent numbers by at least one blank column on the card.

(1) 5 in column 1.

(2) An integer (TOP) specifying whether data are in integer (2222) or decimal point (4444) form.

(3) An integer (M) specifying number of columns in the data array.

Table 13. Continuation of output of correlation coefficient program using second set of input data listed in Table 10. Columns 5 and 7 of original data array have been deleted.

```
5$AMERICUS CHEM DATA WITH COLUMNS 5 AND 7 OF ORIG DATA ARRAY DELETED      $
CORRELATION COEFFICIENTS FOR TRANSPOSED, NORMALIZED ARRAY, EMPLOYING (X - MEAN)/(STD DEV)

DATA IN TERMS OF UNIT STANDARD DEVIATIONS

5    .308    .371 -1.098  -.668  -.321  -.097
5   1.960  -1.328  -.740  -.636  -.546  -.290
5    .446  -1.611  -.740  2.427  -.912  -.278
5    .859  -.101  -1.457  -.668 -1.193  -.283
5    .997  -.384  -.024  -.381 -1.053  -.285
5   -.875  -.667  1.767  -.445   .176  -.282
5   -.738   .560  -.024  -.700  -.331  -.279
5   -.518   .088  -.024  -.572  -.387  -.281
5   -.518   .749  -.740  -.477  -.574  -.279
5  -1.068  -.290  -.382  2.427  -.696  -.277
5   -.793   .277   .334   .130  -.303  -.270
5   1.960  -1.328  -.740  -.189   .898  -.281
5   -.380  -.101   .692  -.636  1.957  -.281
5   -.518  1.882  2.125   .194  1.685  -.274
5  -1.123  1.882  1.051   .194  1.601  3.738

  ROW NO.        MEAN         STANDARD DEV.

     1           -.2508           .5197
     2           -.2633          1.0428
     3           -.1113          1.2955
     4           -.4738           .7608
     5           -.1883           .6141
     6           -.0545           .8779
     7           -.2519           .4388
     8           -.2823           .2430
     9           -.3065           .4913
    10           -.0476          1.1409
    11           -.1040           .3944
    12            .0532          1.0845
    13            .2087           .8838
    14            .8490          1.0765
    15           1.2236          1.4997

CORRELATION COEFFICIENTS OF UPPER HALF OF MATRIX WITH ONES IN PRINCIPAL DIAGONAL OMITTED

CORRELATION COEFFICIENTS BETWEEN ALL POSSIBLE PAIRS OF ROWS OF ORIGINAL DATA ARRAY

5    .356 -.276   .831   .276 -.830   .167   .011   .653 -.410 -.511   .257 -.282 -.346
5    .007   .299   .694   .802 -.356 -.679 -.594 -.406 -.362 -.881   .877 -.249 -.699
5   -.600   .144   .224 -.267 -.792 -.795 -.432   .772 -.147   .284 -.529 -.629 -.474
5    .708 -.810 -.215 -.268   .343 -.226 -.634   .441 -.635 -.687 -.337 -.203 -.334
5   -.182 -.147 -.264 -.465   .438 -.571 -.529 -.599   .208   .411 -.489 -.114   .496
5   -.315   .515   .620   .142   .957   .740 -.276   .635 -.763   .172   .724   .506   .567
5   -.352   .642 -.748   .163   .723   .463 -.044   .315 -.535 -.288   .215   .318   .448
5   -.326 -.430 -.161 -.105 -.918   .017   .691   .299   .133 -.543 -.566   .642   .175
5    .189
```

ROW NO.	MEAN	STANDARD DEV.
1	.4776	.2902
2	.4903	.3533
3	.5106	.3210
4	.4228	.3086
5	.4537	.3144
6	.4701	.3414
7	.4265	.3280
8	.4214	.3094
9	.4173	.3155
10	.5045	.3416
11	.4668	.2926
12	.5836	.3554
13	.5619	.3748
14	.6404	.3677
15	.6286	.3374

CORRELATION COEFFICIENTS OF LOWER HALF OF MATRIX WITH ONES IN PRINCIPAL DIAGONAL INCLUDED

CORRELATION COEFFICIENTS BETWEEN ALL POSSIBLE PAIRS OF ROWS OF ORIGINAL DATA ARRAY

```
5  1.000
5   .783 1.000
5   .208  .355 1.000
5   .918  .883  .361 1.000
5   .833  .871  .462  .959 1.000
5   .774  .520  .297  .618  .672 1.000
5   .881  .531  .240  .787  .786  .888 1.000
5   .878  .599  .325  .822  .844  .895  .991 1.000
5   .865  .534  .304  .827  .820  .793  .980  .977 1.000
5   .261  .046  .834  .252  .321  .474  .459  .481  .493 1.000
5   .814  .464  .436  .707  .730  .919  .959  .956  .932  .668 1.000
5   .707  .895  .264  .673  .623  .500  .392  .441  .345 -.008  .367 1.000
5   .772  .509  .053  .513  .481  .866  .752  .734  .639  .218  .740  .669 1.000
5   .723  .313  .108  .486  .489  .861  .846  .806  .771  .417  .860  .388  .880
5  1.000
5  -.051 -.497 -.605 -.238 -.225  .196  .284  .204  .226 -.153  .176 -.531  .172
5   .338 1.000
```

```
5$AMERICUS CHEM DATA WITH COLUMNS 5 AND 7 OF ORIG DATA ARRAY DELETED      $
INPUT DATA ARRAY WITH COLUMNS DELETED AS SPECIFIED

5   .120   .340   .100   .003   .180   .880
5   .240   .160   .110   .004   .156   .028
5   .130   .130   .110   .100   .117   .080
5   .160   .290   .090   .003   .087   .056
5   .170   .260   .130   .012   .102   .048
5   .034   .230   .180   .010   .233   .060
5   .044   .360   .130   .002   .179   .073
5   .060   .310   .130   .006   .173   .064
5   .060   .380   .110   .009   .153   .073
5   .020   .270   .120   .100   .140   .085
5   .040   .330   .140   .028   .182   .116
5   .240   .160   .110   .018   .310   .066
5   .070   .290   .150   .004   .423   .067
5   .060   .500   .190   .030   .394   .095
5   .016   .500   .160   .030   .385 17.820
```

ROW NO.	MEAN	STANDARD DEV.
1	-.1009	.5372
2	-.0891	.9526
3	-.0867	1.1431
4	-.3549	.7296
5	-.1635	.6910
6	.0165	.7729
7	-.2098	.5102
8	-.2477	.4542
9	-.2928	.6429
10	-.0229	1.0101
11	-.0687	.3752
12	.2300	1.2060
13	.3217	.9549
14	.7004	1.0488
15	.3684	2.0556

CORRELATION COEFFICIENTS OF LOWER HALF OF MATRIX WITH ONES IN PRINCIPAL DIAGONAL INCLUDED

CORRELATION COEFFICIENTS BETWEEN ALL POSSIBLE PAIRS OF ROWS OF ORIGINAL DATA ARRAY

```
5  1.000
5   .443 1.000
5  -.257  .284 1.000
5   .683  .675  .199 1.000
5   .047  .587  .291  .715 1.000
5  -.586 -.279 -.268 -.702 -.197 1.000
5   .008 -.453 -.451  .107  .232  .120 1.000
5  -.154 -.245 -.192  .209  .500  .134  .920 1.000
5   .190 -.265 -.138  .458  .403 -.376  .858  .835 1.000
5  -.367 -.328  .781 -.122 -.067 -.120 -.062  .024  .125 1.000
5  -.407 -.719 -.054 -.352 -.074  .438  .702  .626  .481  .483 1.000
5   .436  .744  .119  .196 -.086 -.152 -.780 -.739 -.696 -.354 -.828 1.000
5   .052 -.108 -.514 -.584 -.696  .486 -.234 -.404 -.566 -.442 -.168  .457 1.000
5  -.278 -.656 -.630 -.746 -.623  .536  .192 -.073 -.172 -.229  .372 -.216  .625
5  1.000
5  -.272 -.579 -.374 -.488 -.517 -.002 -.051 -.210 -.107 -.153 -.054 -.297  .101
5   .393 1.000
```

Table 12.--Continuation of listing of output from correlation coefficient program using first set of input data in Table 10.

```
5$AMERICUS LS CHEM DATA WITH ALL COLS IN ORIGINAL DATA ARRAY INCLUDED     $
CORRELATION COEFFICIENTS FOR TRANSPOSED, NORMALIZED ARRAY, EACH X VALUE DIVIDED BY HIGHEST
     VALUE FOR THAT VARIABLE

ARRAY CONSISTING OF ORIGINAL DATA DIVIDED BY HIGHESTVALUE IN COLUMN

5   .500   .680   .526   .030   .683   .426   .926   .049
5  1.000   .320   .579   .040   .656   .369   .957   .002
5   .542   .260   .579  1.000   .462   .277   .961   .004
5   .667   .580   .474   .030   .456   .206   .967   .003
5   .708   .520   .684   .120   .353   .241  1.000   .003
5   .142   .460   .947   .100   .629   .551   .928   .003
5   .183   .720   .684   .020   .398   .423   .978   .004
5   .250   .620   .684   .060   .359   .409   .986   .004
5   .250   .760   .579   .090   .302   .362   .992   .004
5   .083   .540   .632  1.000   .479   .331   .966   .005
5   .167   .660   .737   .280   .503   .430   .952   .007
5  1.000   .320   .579   .180  1.000   .733   .853   .004
5   .292   .580   .789   .040   .915  1.000   .875   .004
5   .250  1.000  1.000   .300   .776   .931   .860   .005
5   .067  1.000   .842   .300   .341   .910   .568  1.000
```

Table 11.--Output of correlation coefficient program using first set of input data listed in Table 10.

```
5$AMERICUS LS CHEM DATA WITH ALL COLS IN ORIGINAL DATA ARRAY INCLUDED      $
INPUT DATA ARRAY WITH COLUMNS DELETED AS SPECIFIED

5    .120    .340    .100    .003   9.450    .180  48.210    .880
5    .240    .160    .110    .004   9.070    .156  49.840    .028
5    .130    .130    .110    .100   6.390    .117  50.050    .080
5    .160    .290    .090    .003   6.310    .087  50.360    .056
5    .170    .260    .130    .012   4.880    .102  52.060    .048
5    .034    .230    .180    .010   8.700    .233  48.330    .060
5    .044    .360    .130    .002   5.510    .179  50.940    .073
5    .060    .310    .130    .006   4.960    .173  51.320    .064
5    .060    .380    .110    .009   4.170    .153  51.640    .073
5    .020    .270    .120    .100   6.630    .140  50.290    .085
5    .040    .330    .140    .028   6.950    .182  49.540    .116
5    .240    .160    .110    .018  13.830    .310  44.430    .066
5    .070    .290    .150    .004  12.650    .423  45.570    .067
5    .060    .500    .190    .030  10.730    .394  44.790    .095
5    .016    .500    .160    .030   4.720    .385  29.580  17.820

MEANS,STANDARD DEVIATIONS, AND CORRELATION COEFFICIENTS ON A COLUMN BY COLUMN BASIS OF DATA
   ARRAY
COLUMN NO.       MEAN          STANDARD DEV.

     1            .0976            .0726
     2            .3007            .1059
     3            .1307            .0279
     4            .0239            .0313
     5           7.6633           2.8767
     6            .2143            .1067
     7          47.7967           5.4034
     8           1.3074           4.4178

CORRELATION COEFFICIENTS OF LOWER HALF OF MATRIX WITH ONES IN PRINCIPAL DIAGONAL INCLUDED


5  1.000
5  -.633 1.000
5  -.559  .485 1.000
5  -.201 -.215 -.039 1.000
5   .377 -.256  .162 -.136 1.000
5  -.249  .450  .671 -.154  .579 1.000
5   .212 -.481 -.425 -.019 -.150 -.731 1.000
5  -.300  .510  .269  .046 -.267  .426 -.902 1.000



5$AMERICUS LS CHEM DATA WITH ALL COLS IN ORIGINAL DATA ARRAY INCLUDED      $
CORRELATION COEFFICIENTS FOR TRANSPOSED, NORMALIZED ARRAY, EMPLOYING (X - MEAN)/(STD DEV)

DATA IN TERMS OF UNIT STANDARD DEVIATIONS

5    .308    .371 -1.098  -.668    .621  -.321    .076  -.097
5   1.960 -1.328  -.740   -.636    .489  -.546    .378  -.290
5    .446 -1.611  -.740   2.427   -.443  -.912    .417  -.278
5    .859  -.101 -1.457  -.668    -.470 -1.193   .474  -.283
5    .997  -.384  -.024   -.381   -.968 -1.053   .789  -.285
5   -.875  -.667  1.767  -.445    .360   .176    .099  -.282
5   -.738   .560  -.024   -.700   -.749  -.331   .582  -.279
5   -.518   .088  -.024   -.572   -.940  -.387   .652  -.281
5   -.518   .749  -.740   -.477  -1.214  -.574   .711  -.279
5 -1.068  -.290  -.382   2.427   -.359  -.696   .461  -.277
5   -.793   .277   .334    .130   -.248  -.303   .323  -.270
5   1.960 -1.328  -.740   -.189   2.144   .898  -.623  -.281
5   -.380  -.101   .692   -.636   1.733  1.957  -.412  -.281
5   -.518  1.882  2.125    .194   1.066  1.685  -.556  -.274
5 -1.123  1.882  1.051    .194  -1.023  1.601 -3.371  3.738
```

Table 10.--Listing of two sets of input data used with correlation coefficient program in calculation of output shown in Tables 11, 12, and 13. Each row in table below is a single card.

```
5$AMERICUS LS CHEM DATA WITH ALL COLS IN ORIGINAL DATA ARRAY INCLUDED      $
5  4444  8  15  0  0  1234  1234  1
5 .120  .34  .10  .003  9.45  .180  48.21  .88
5  .240  .16  .11  .004  9.07  .156  49.84  .028
5  .130  .13  .11  .100  6.39  .117  50.05  .080
5  .160  .29  .09  .003  6.31  .087  50.36  .056
5  .170  .26  .13  .012  4.88  .102  52.06  .048
5  .034  .23  .18  .010  8.70  .233  48.33  .060
5  .044  .36  .13  .002  5.51  .179  50.94  .073
5  .060  .31  .13  .006  4.96  .173  51.32  .064
5  .060  .38  .11  .009  4.17  .153  51.64  .073
5  .020  .27  .12  .100  6.63  .140  50.29  .085
5  .040  .33  .14  .028  6.95  .182  49.54  .116
5  .240  .16  .11  .018  13.83  .310  44.43  .066
5  .070  .29  .15  .004  12.65  .423  45.57  .067
5  .060  .50  .19  .030  10.73  .394  44.79  .095
5  .016  .50  .16  .030  4.72  .385  29.58  17.82
5$AMERICUS CHEM DATA WITH COLUMNS 5 AND 7 OF ORIG DATA ARRAY DELETED      $
5 4444  8  15  2  0  7777  4444  1
5  1 1 1 1 0 1 0 1
5 .120  .34  .10  .003  9.45  .180  48.21  .88
5  .240  .16  .11  .004  9.07  .156  49.84  .028
5  .130  .13  .11  .100  6.39  .117  50.05  .080
5  .160  .29  .09  .003  6.31  .087  50.36  .056
5  .170  .26  .13  .012  4.88  .102  52.06  .048
5  .034  .23  .18  .010  8.70  .233  48.33  .060
5  .044  .36  .13  .002  5.51  .179  50.94  .073
5  .060  .31  .13  .006  4.96  .173  51.32  .064
5  .060  .38  .11  .009  4.17  .153  51.64  .073
5  .020  .27  .12  .100  6.63  .140  50.29  .085
5  .040  .33  .14  .028  6.95  .182  49.54  .116
5  .240  .16  .11  .018  13.83  .310  44.43  .066
5  .070  .29  .15  .004  12.55  .423  45.57  .067
5  .060  .50  .19  .030  10.73  .394  44.79  .095
5  .016  .50  .16  .030  4.72  .385  29.58  17.82
```

```
143         L2..
144         XH = Y(1,J)  $
145         FOR I =(2,1,N) $
146         XH = MAX(Y(I,J),XH)  $
147         FOR I =(1,1,N) $
148         X(I,J) = Y(I,J)/XH $
149         J = J + 1 $
150         UNTIL J EQL  M  + 1   $ GO TO L2 $
151   IF DATLIST EQL 1 $ BEGIN
152   FORMAT HED3(*ARRAY CONSISTING OF ORIGINAL DATA DIVIDED BY HIGHEST*,
153         *VALUE IN COLUMN*,W,W)$  WRITE ($$ HED3)$
154         OUTPUT JOE(FOR I =(1,1,N) $ FOR J =(1,1,M) $ X(I,J)) $
155   WRITE($$ JOE, JILL) $ END$
156        END $
157       QCALC..
158         I = 1 $
159   FORMAT HED4(W,*  ROW NO.        MEAN         STANDARD DEV.*,W,W)$
160   WRITE ($$ HED4)$
161         L3..
162         SUM(I)=SMSQ(I)=0.0$
163         FOR J = (1,1,M) $
164            BEGIN
165               SUM(I) = SUM(I)+ X(I,J) $
166               SMSQ(I) = SMSQ(I) +(X(I,J).X(I,J)) $
167            END $
168         MEAN(I) = SUM(I)/MF$
169         STD(I) = SQRT((SMSQ(I)/(MF )) -(MEAN(I).MEAN(I))) $
170         OUTPUT ODS2( I, MEAN(I), STD(I))$
171         WRITE ($$ ODS2, FMT2) $
172         I = I + 1 $
173         UNTIL I  EQL (N+1) $ GO TO L3 $
174         FOR I =(1,1,N-1) $  FOR  K =(I+1,1,N) $
175            BEGIN
176               P(I,K)=0.0$
177               FOR J =(1,1,M) $ P(I,K) = P(I,K) +(X(I,J).X(K,J)) $
178               P(I,K)=((P(I,K)/(MF ))-(MEAN(I).MEAN(K)))/(STD(I).STD(K))$
179            END $
180         IF CHC EQL 7777 $
181         BEGIN
182         OUTPUT KOR(FOR I =(1,1,N-1) $ FOR K =(I+1,1,N) $ P(I,K) )$
183         WRITE($$ FMT7) $ WRITE ($$ FMT3) $
184         WRITE ($$ KOR, FMT6) $
185         END $
186         IF CHC NEQ 7777 $
187            BEGIN
188               WRITE ($$ FMT8) $ WRITE ($$ FMT3) $
189               FOR J = (1,1,N) $ P(J,J) = 1.0 $
190               FOR I =(2,1,N) $ FOR K =(1,1,I-1) $ P(I,K) = P(K,I) $
191               FOR I =(1,1,N) $
192                  BEGIN
193                  OUTPUT KORL(FOR K =(1,1,I) $ P(I,K)) $
194                     WRITE ($$ KORL , FMT6) $
195                  END $
196            END $
197         IF C EQL 0 $ (C = 3 $ GO LB ) $
198      END $
199   GO START $
200   FINISH $
```

Input To Correlation Coefficient Program

Alphanumeric heading card:--An example input data set is listed in Table 10. The first card of each input data set contains alphabetic and numerical (alphanumeric) information for identification purposes. This information will be reproduced at the top of certain pages (Tables 11, 12, 13) printed by the computer's printer. The card must be punched as follows:

(1) 5 in column 1.

(2) $ in columns 2 and 75.

(3) Any desired combination of letters, numbers, characters and blanks in columns 3 through 74.

```
68        FOR I =(1,1,N) $
69           BEGIN
70              SUM(J) = SUM(J) + X(I,J) $
71              SMSQ(J) = SMSQ(J) + (X(I,J).X(I,J)) $
72           END $
73        MEAN(J) = SUM(J)/NF$
74        STD(J) = SQRT((SMSQ(J)/(NF )) -(MEAN(J).MEAN(J)))  $
75        OUTPUT ODS1( J, MEAN(J), STD(J))$
76        FORMAT FMT2(I5, 2X17.4,W) $
77        WRITE ($$ ODS1, FMT2) $
78        J = J + 1 $
79        UNTIL J EQL M+1 $ GO TO L1 $
80        FOR J =(1,1,M-1) $ FOR K =(J+1,1,M) $
81           BEGIN
82              P(J,K) = 0.0 $
83              FOR I =(1,1,N) $  P(J,K) = P(J,K) +(X(I,J).X(I,K)) $
84              P(J,K) =((P(J,K)/(NF ))-(MEAN(J).MEAN(K)))/(STD(J).STD(K))$
85           END  $
86        IF CHC EQL 7777 $
87        BEGIN
88        OUTPUT COR(FOR J =(1,1,M-1) $ FOR K =(J+1,1,M) $ P(J,K)) $
89        WRITE ($$ FMT7) $
90        WRITE ($$ COR, FMT6) $
91        END $
92        COMMENT IF CHC NEQ 7777, OUTPUT LOWER HALF MATRIX WITH ONES IN
93        PRINCIPAL DIAGONAL $
94        IF CHC NEQ 7777 $
95           BEGIN
96              WRITE ($$ FMT8) $
97              FOR J =(1,1,M)$           P(J,J) = 1.0  $
98              FOR J =(2,1,M) $FOR K =(1,1,J-1) $ P(J,K) = P(K,J) $
99              FOR J= (1,1,M)$
100                BEGIN
101                   OUTPUT CORL(FOR K =(1,1,J) $ P(J,K)) $
102                   WRITE ($$ CORL, FMT6) $
103                END $
104          END $
105     END $
106  IF OP NEQ 1 $
107     BEGIN
108        IF C NEQ 1 $(FOR I =(1,1,N)$FOR J =(1,1,M)$ Y(I,J)= X(I,J)) $
109      IF C NEQ 2 $
110       BEGIN
111        WRITE ($$ ALPHA, FMTA) $
112        FORMAT FMT4(*CORRELATION COEFFICIENTS FOR TRANSPOSED, NORMALIZED*
113           ,* ARRAY, EMPLOYING (X - MEAN)/(STD DEV)*, W,W)$
114        WRITE ($$ FMT4) $
115  IF OP EQL 2$   BEGIN
116     FOR J = (1,1,M) $ BEGIN
117        SUM(J) = SMSQ(J) = 0.0 $
118              FOR I = (1,1,N) $ BEGIN
119                 SUM(J) = SUM(J) + X(I,J) $
120                 SMSQ(J) = SMSQ(J) + (X(I,J).X(I,J)) $   END$
121        MEAN(J) = SUM(J)/NF $
122        STD(J) = SQRT((SMSQ(J)/NF) - (MEAN(J).MEAN(J))) $END$ END$
123        J = 1 $
124        LA..
125        FOR I =(1,1,N) $  X(I,J) = (X(I,J) - MEAN(J))/STD(J) $
126        J = J + 1 $
127        UNTIL J EQL  M+1 $  GO LA $
128  IF DATLIST EQL 1 $ BEGIN
129  FORMAT HED2(*DATA IN TERMS OF UNIT STANDARD DEVIATIONS*,W,W)$
130  WRITE ($$ HED2)$
131        OUTPUT JACK(FOR I =(1,1,N) $ FOR J =(1,1,M)$ X(I,J)) $
132                                    WRITE($$ JACK, JILL) $   END$
133        GO QCALC $         END$
134       LB..
135      IF C NEQ 1 $
136       BEGIN
137  WRITE ($$ ALPHA,FMTA) $
138        FORMAT FMT5(*CORRELATION COEFFICIENTS FOR TRANSPOSED, NORMALIZED*
139           ,* ARRAY, EACH X VALUE DIVIDED BY HIGHEST VALUE FOR THAT*,
140           * VARIABLE*, W,W) $
141        WRITE ($$ FMT5) $
142        J = 1      $
```

Table 9.--Listing of BALGOL statements in correlation coefficient program. Numbers in left column are for reference purposes in this report, and in practice may be placed within columns 73-80 on punched cards. Each BALGOL program card must have the number 2 punched in column 1.

```
 1  COMMENT PROGRAM TO CALCULATE MEANS, STANDARD DEVIATIONS AND CORRELATION
 2  COEFFICIENTS FOR ORIGINAL DATA ARRAY  AND TRANSPOSED, NORMALIZED ARRAY.
 3  HAVE CHOICE OF UPPER OR LOWER HALF OF COR MATRIX WITH OR WITHOUT ONES
 4  J.W. HARBAUGH, GEOLOGY DEPT. STANFORD $
 5  INTEGER   M,N,OP,Q,R,H,I,J,K,T, C , CHC , Z , DATLIST , SP , TOP $
 6  ARRAY    Y(200,20), X(200,20), P(120,120), H(20), SUM(120), STD(120),
 7     SMSQ(120), Z(200,20),      MEAN(120)       $
 8  FORMAT FMT7(*CORRELATION COEFFICIENTS OF UPPER HALF OF MATRIX WITH *,
 9             *ONES IN PRINCIPAL DIAGONAL OMITTED*, W4,W,W),
10          FMT8(*CORRELATION COEFFICIENTS OF LOWER HALF OF MATRIX WITH *,
11             *ONES IN PRINCIPAL DIAGONAL INCLUDED*,W4,W,W),
12          FMT6(*5 *,13X6.3,C),
13          FMT3(*CORRELATION COEFFICIENTS BETWEEN ALL POSSIBLE PAIRS OF *,
14             *ROWS OF ORIGINAL DATA ARRAY*,W,W) $
15  START..
16  INPUT ALPH(A1,A2,A3,A4,A5,A6,A7,A8,A9,A10,A11,A12),
17        HEAD(TOP, M, N, OP, C, CHC, Q,       DATLIST)$
18  READ  ($$ ALPH) $
19  READ  ($$ HEAD) $
20   IF Q EQL 4444 $ BEGIN
21  INPUT COLS(FOR R=(1,1,M) $ H(R))$ READ($$ COLS)$ END$
22  IF TOP EQL 2222 $ BEGIN
23  INPUT DATI(FOR I=(1,1,N)$ FOR T =(1,1,M) $ Z(I,T))$
24  READ ($$DATI)$
25  FOR I = (1,1,N) $  FOR T =(1,1,M)$X(I,T) = Z(I,T) $ END $
26  IF TOP EQL 4444 $ BEGIN
27  INPUT DATD(FOR I= (1,1,N) $ FOR T=(1,1,M) $ X(I,T)      )$
28  READ ($$ DATD)$  END$
29  FORMAT JOHN(*5*,$SP$X7.1,C),  JILL(*5*,$SP$X7.3,C) $
30  OUTPUT ALPHA(A1,A2,A3,A4,A5,A6,A7,A8,A9,A10,A11,A12)$
31  FORMAT FMTA(*5$*,12A6,*$*,C3) $
32  WRITE ($$ ALPHA,FMTA) $
33  IF Q EQL 4444 $
34     BEGIN
35        T = J = R = 1 $
36        JUMP..
37        IF R EQL M+1 $ (M =J-1$ GO ZOOM) $
38        IF H(R) EQL 1 $
39           BEGIN
40              FOR I =(1,1,N)$ X(I,J) =X(I,T) $
41              T = T+1 $ R = R+1 $ J = J+1 $
42              GO JUMP $
43           END $
44        IF H(R) EQL 0 $
45           BEGIN
46              R = R + 1 $
47              T = T + 1 $
48              GO JUMP $
49           END $
50     END $
51  ZOOM..
52  SP = M $ IF SP GTR 11 $ SP = 11 $ MF = M         $ NF = N      $
53  IF DATLIST EQL 1 $ BEGIN
54  FORMAT HED1(*INPUT DATA ARRAY WITH COLUMNS DELETED AS SPECIFIED*,W,W)$
55  WRITE ($$ HED1)$
56  OUTPUT MIKE(FOR I=(1,1,N) $ FOR J =(1,1,M) $ X(I,J))$
57  IF TOP EQL 2222 $ WRITE($$ MIKE, JOHN)$
58  IF TOP EQL 4444 $ WRITE ($$ MIKE, JILL)$     END$
59  IF OP NEQ 2 $
60     BEGIN
61        FORMAT FMT1(*MEANS,STANDARD DEVIATIONS, AND CORRELATION COEFFIC*,
62           *IENTS ON A COLUMN BY COLUMN BASIS OF DATA ARRAY*,W4,*COLUMN*,
63           * NO.       MEAN          STANDARD DEV.*,W4,W,W) $
64        WRITE ($$ FMT1) $
65        J = 1 $
66        L1..
67        SUM(J)=SMSQ(J)=0.0 $
```

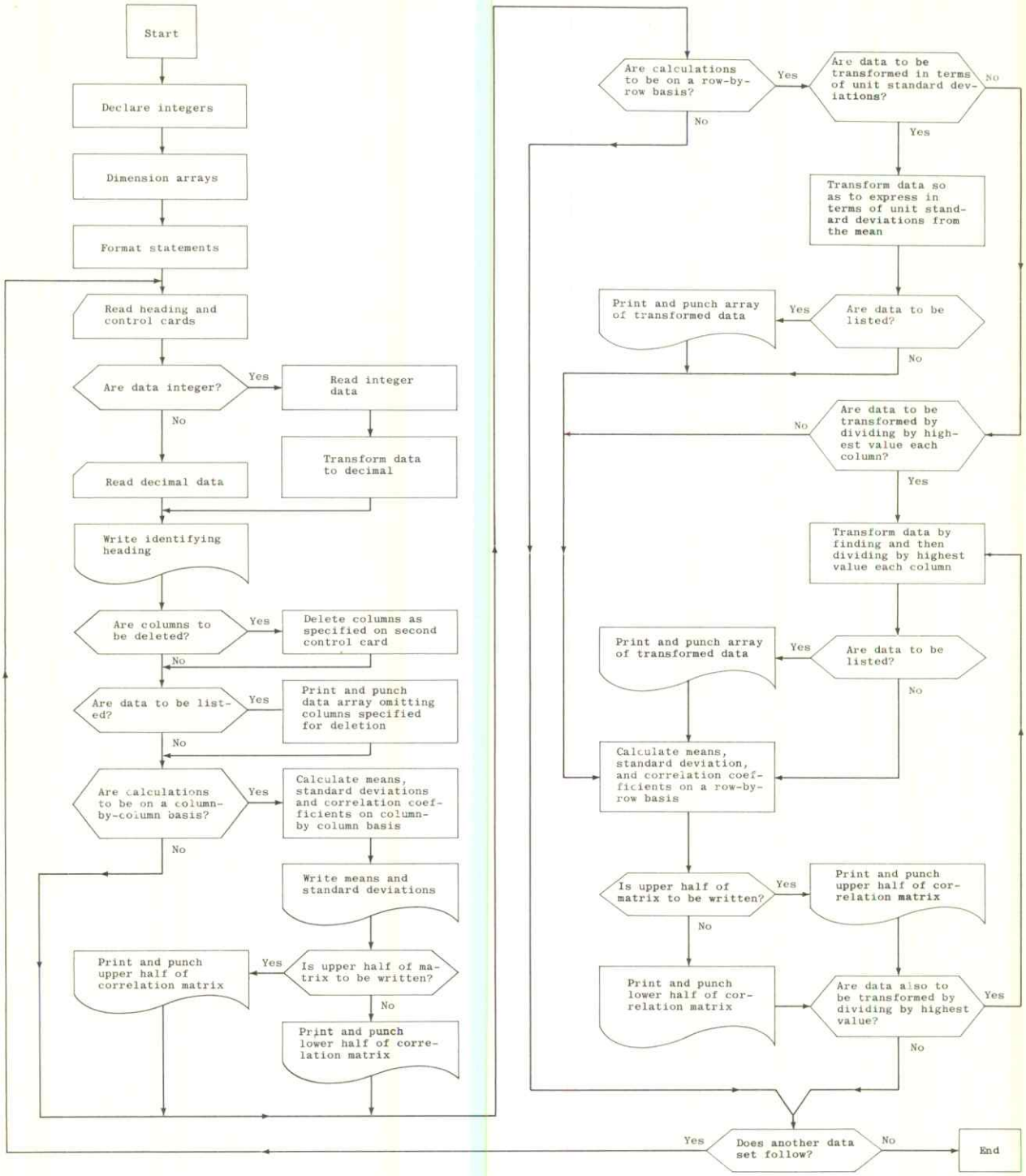


Figure 5. Simplified flow chart of major steps in correlation coefficient program.

(7) If calculations are to be made on a column-by-column basis:

(a) Calculate and print means and standard deviations of each column: lines 64-79.

(b) Calculate correlation coefficients between all possible pairs of columns: lines 80-85.

(c) If specified, punch and print upper half of correlation matrix, omitting ones in principal diagonal: lines 86-91.

(d) Otherwise, punch and print lower half of correlation matrix, filling in ones in principal diagonal: lines 92-104.

(8) If calculations are to be performed on a row-by-row basis:

(a) If specified, transform data so that values in each column are expressed in terms of unit standard deviations from the mean of that column: lines 106-127.

(b) If specified, print and punch transformed data: lines 128-132.

(c) If specified, transform data so that values in each column are divided by highest value in that column: lines 135-150.

(d) If specified, print and punch transformed data: lines 151-156.

(e) Calculate and print means and standard deviations of each row of transformed data: lines 158-173.

(f) Calculate correlation coefficients between all possible pairs of rows: lines 174-179.

(g) If specified, print and punch upper half of correlation matrix, with principal diagonal omitted: lines 180-185.

(h) Otherwise, print and punch lower half of correlation matrix with ones in principal diagonal included: lines 186-196.

(9) If row-by-row calculations are to be made and if calculations are to be made with both types of transformation of original data, C is set to 0, and control is returned to label LB at line 134 so that calculations with divide-by-high-value transformation will be performed after previous calculations with other transformation have been made: line 197.

CORRELATION COEFFICIENT PROGRAM

General

The correlation coefficient program described here provides for calculation of Pearson product-moment correlation coefficients, which are given by the equation

$$r = \frac{\frac{\sum_{i=1}^{n} X_i Y_i}{n} - \overline{XY}}{s_x s_y},$$

where r = correlation coefficient,

X_i = variable X,

Y_i = variable Y,

$\overline{X}$ = arithmetic mean of X values,

$\overline{Y}$ = arithmetic mean of Y values,

s_x = standard deviation of X values,

s_y = standard deviation of Y values,

n = number of values whose subscripts, i, range from 1 to n.

The program provides for various options including: (1) A choice of deleting specified columns in the original data array before calculations are performed. (2) Means, standard deviations, and correlation coefficients may be calculated on either, or both, a column-by-column, or row-by-row basis. (3) For calculations on a row-by-row basis, transformation of the original data is necessary. The data may be transformed so as to be expressed in terms of unit standard deviations from the mean of each column, or transformed by dividing through by the highest value in each column, or both transformations may be used. (4) If desired, the data may be listed, both before and after transformation. (5) Finally, the user has a choice of printing (and punching) either the upper half or the lower half of the matrix of correlation coefficients.

Major Steps In Correlation Coefficient Program

Steps in the program are outlined in the flow chart (Fig. 5) and are listed line by line (card by card) in Table 9. The major steps in the programs are given below:

(1) Integer declarations: line 5.

(2) Dimension arrays: lines 6-7.

(3) Format statements: lines 8-14, 29, 31, 54, 61, 76, 112-113, 129, 138-140, 152-154, 159.

(4) Input and read statements: lines 16-28.

(5) If specified on control cards, delete indicated columns from original data array: lines 33-50.

(6) If specified on control card, print and punch original data array, less columns specified for deletion: lines 53-58.

KANSAS GEOLOGICAL SURVEY COMPUTER PROGRAM
THE UNIVERSITY OF KANSAS, LAWRENCE

PROGRAM ABSTRACT

Title (If subroutine state in title):

Program for calculation of distance coefficients.

Computer: IBM 7090 or 7094 Date: December, 1963

Programming language: BALGOL, also known as SUBALGOL, a dialect of ALGOL-58

Author, organization: John W. Harbaugh, Department of Geology
Stanford University, Stanford, California

Direct inquiries to: Author or to

Name: Daniel F. Merriam Address: State Geological Survey
Lawrence, Kansas

Purpose/description: Calculates distance coefficients on either, or both, a column-by-column basis or row-by-row basis. Has a variety of options, including data transformation and printing and punching of either upper or lower half of matrix.

Mathematical method:

Restrictions, range:

Storage requirements: 32,768 words of high-speed memory

Equipment specifications:

Memory 20K_____ 40K_____ 60K_____ 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)

Distance Coefficient Program Limitations and Operating Times

Array limitations:--The following limitations have been placed on use of the program by the present array dimensions (lines 3 and 4, Table 6). Original data arrays are limited to 200 rows and 25 columns. However, if row-by-row calculations are to be performed, no more than 130 rows may be present in the original data array. The limit of 130 rows is dictated by the fact that 16,900 words in the computer's memory are occupied by the 130 x 130 array for the variable COR. Depending on circumstances, the dimensions of the COR array could be increased by reducing the dimensions of other arrays. For example, if the data are in decimal point form, the XP(200,25) array could be reduced to token dimensions, XP (1,1). Similarly, if there are fewer than 25 columns in the original data array, its width could be reduced. Conversely, more columns or more rows could be added to the X(,) and XP(,) arrays if the dimensions of the COR(,) array are reduced. If substantially larger arrays are to be handled, major revisions of the program would be necessary so that the distance coefficient matrix could be stored sequentially in a one-dimensional array.

Operating times:--The program is fast, requiring about 8 seconds to compile on the IBM 7090. Execution time varies depending on the dimension of the matrix of coefficients calculated. Calculations for a 54 x 54 matrix containing 1431 coefficients require about 8 seconds with the IBM 7090.

Output From Distance Coefficient Program

Examples of output from the program based on data of Table 7, are shown in Table 8, and are listed in order below:

(1) Alphanumeric heading for identification purposes (both printed and punched).

(2) The statement: GENERAL DISTANCE FUNCTION COEFFICIENTS.

(3) Number of rows and number of columns in original data array.

(4) A statement as to which type of transformation (if any) was made.

(5) A statement as to whether distance coefficients are based on comparisons between columns or between rows.

(6) A statement as to whether distance coefficients are in the upper or the lower half of the matrix.

(7) The matrix of distance coefficients. Coefficients are both printed and punched. A 5 is punched in first column of each card, and the cards may be used as input to other programs, including BALGOL programs and programs in other languages such as FORTRAN. Depending on the specifications of the control card, coefficients based on either, or both, column-by-column or row-by-row calculations are produced.

Table 8.--Example of output from distance coefficient program. Input data used in preparation of example are listed in Table 7. First output data set pertains to example data set shown in Tables 1 to 5. Second data set consists of actual geological data.

```
5$TEST OF DISTANCE COEFFICIENT CALCULATIONS                              $

          GENERAL DISTANCE FUNCTION COEFFICIENTS

NUMBER OF ROWS =   4 ,  NUMBER OF COLUMNS =   3   IN ORIGINAL DATA ARRAY

DATA VALUES TRANSFORMED BY DIVIDING BY HIGHEST VALUE IN COLUMN

DISTANCE COEFFICIENTS BETWEEN ALL POSSIBLE PAIRS OF  COLUMNS OF UNTRANSPOSED DATA ARRAY

DISTANCE FUNCTION COEFFICIENTS LOWER HALF OF MATRIX WITH ONES IN PRINCIPAL DIAGONAL INCLUDED

5  1.000
5   .553 1.000
5   .379  .769 1.000
DATA ARRAY HAS BEEN TRANSPOSED AND COEFFICIENTS CALCULATED BETWEEN ALL POSSIBLE PAIRS OF ROWS
     OF ORIGINAL DATA ARRAY
DISTANCE FUNCTION COEFFICIENTS LOWER HALF OF MATRIX WITH ONES IN PRINCIPAL DIAGONAL INCLUDED

5  1.000
5   .660 1.000
5   .352  .683 1.000
5   .334  .589  .575 1.000


5$AMERICUS LIMESTONE CHEMICAL DATA, KANSAS AND OKLAHOMA                  $

          GENERAL DISTANCE FUNCTION COEFFICIENTS

NUMBER OF ROWS =  15 ,  NUMBER OF COLUMNS =   8   IN ORIGINAL DATA ARRAY

DATA VALUES TRANSFORMED BY DIVIDING BY HIGHEST VALUE IN COLUMN

DISTANCE COEFFICIENTS BETWEEN ALL POSSIBLE PAIRS OF  COLUMNS OF UNTRANSPOSED DATA ARRAY

DISTANCE FUNCTION COEFFICIENTS LOWER HALF OF MATRIX WITH ONES IN PRINCIPAL DIAGONAL INCLUDED

5  1.000
5   .494 1.000
5   .508  .791 1.000
5   .494  .450  .430 1.000
5   .670  .664  .730  .492 1.000
5   .549  .737  .739  .493  .780 1.000
5   .408  .579  .686  .244  .561  .469 1.000
5   .444  .424  .336  .576  .397  .491  .088 1.000
DATA ARRAY HAS BEEN TRANSPOSED AND COEFFICIENTS CALCULATED BETWEEN ALL POSSIBLE PAIRS OF ROWS
     OF ORIGINAL DATA ARRAY
DISTANCE FUNCTION COEFFICIENTS LOWER HALF OF MATRIX WITH ONES IN PRINCIPAL DIAGONAL INCLUDED

5  1.000
5   .779 1.000
5   .613  .616 1.000
5   .866  .821  .633 1.000
5   .822  .822  .665  .906 1.000
5   .782  .659  .605  .711  .733 1.000
5   .837  .663  .591  .791  .787  .835 1.000
5   .841  .693  .623  .816  .822  .840  .953 1.000
5   .830  .667  .613  .816  .807  .781  .934  .932 1.000
5   .613  .520  .808  .593  .614  .648  .643  .657  .657 1.000
5   .822  .661  .670  .765  .776  .864  .896  .899  .875  .735 1.000
5   .725  .812  .582  .684  .679  .633  .591  .611  .580  .499  .614 1.000
5   .747  .632  .517  .631  .629  .790  .714  .707  .669  .548  .730  .702 1.000
5   .697  .551  .521  .591  .605  .746  .709  .691  .675  .593  .743  .596  .802
5  1.000
5   .520  .387  .399  .451  .461  .541  .553  .542  .542  .470  .563  .392  .536
5   .593 1.000
```

Table 7.--Listing of an example set of data cards used as input to distance coefficient program. Output from program using these data is shown in Table 8. First data set is example shown earlier to demonstrate calculation of distance coefficients. Second data set is actual geological example. Each row lists a single data card. Data cards listed are distributed with program decks.

```
5$TEST OF DISTANCE COEFFICIENT CALCULATIONS                              $
5  4  3  1111   3333   7777   8888
5  140    6    116   90  9  420  65  12  770  0  6  558


5$AMERICUS LIMESTONE CHEMICAL DATA, KANSAS AND OKLAHOMA                  $
5 15  8  2222  3333   7777   8888 * AMER CHEM DATA DIST FUNCTION COR CONTROL
5 .120  .34  .10   .003  9.45   .180  48.21   .88
5  .240  .16  .11  .004  9.07    .156   49.84   .028
5  .130    .13   .11    .100   6.39    .117   50.05   .080
5    .160    .29   .09   .003   6.31  .087   50.36   .056
5   .170  .26   .13  .012  4.88   .102   52.06  .048
5  .034  .23  .18  .010  8.70  .233  48.33  .060
5  .044  .36  .13  .002   5.51    .179   50.94    .073
5  .060   .31   .13  .006   4.96   .173   51.32  .064
5   .060  .38   .11   .009  4.17   .153   51.64   .073
5   .020  .27    .12    .100   6.63   .140   50.29   .085
5  .040  .33  .14  .028  6.95  .182  49.54   .116
5  .240  .16  .11  .018  13.83  .310   44.43  .066
5  .070  .29  .15  .004  12.65  .423  45.57  .067
5  .060  .50  .19  .030  10.73  .394  44.79  .095
5  .016  .50   .16  .030  4.72   .385  29.58  17.82
```

Data cards:

(1) 5 in column 1 of each card.

(2) Any convenient number of values may be placed on each card. Program will store data in M columns and N rows. Type (decimal point or integer) of data values must be consistent with TOP above.

(3) Any number of data sets may be used in succession, but cards in each set must follow sequence listed above, including the alphanumeric data card (first data card), control card (second card) and data value cards (third and subsequent cards). Negative values generally may not be used with existing transformation options in program. Ordinarily, variables are arranged in columns and observations in rows.

(2) $ in columns 2 and 75.

(3) Any desired combination of letters, numbers, characters, and blanks in columns 3 through 74.

Program-control card.--The second card that accompanies each data set must be a program control card containing the following information in the order listed below. Note that each number is to be separated from adjacent numbers by at least one blank space on the card.

(1) 5 in column 1.

(2) An integer (N) specifying number of rows in original data array.

(3) An integer (M) specifying number of columns in original data array.

(4) An integer (TOP) specifying whether data values are in integer (1111) or decimal point (2222) form.

(5) An integer (OP) specifying type of transformation of data values desired, as follows:

3333 Divide values in each column by highest value in that column of original data array.

4444 Express data values so that range within each column is from 0 to 1.0, based on the minimum and maximum values present in that column in original data array.

5555 Divide each value by 100.

If an integer other than 3333, 4444, or 5555 is used, distance coefficients will be calculated without transformation of data.

(6) An integer (CHC) specifying whether upper or lower half of matrix of coefficients is to be printed and punched, as follows:

6666 Print and punch upper half of matrix with ones in principal diagonal omitted.

7777 Print and punch lower half of matrix with ones in principal diagonal included.

(7) An integer (TRANS) specifying whether data array is to be transposed so that distance coefficients will be calculated on a row-by-row basis, as follows:

1234 Do not transpose (i.e., calculate coefficients on a column-by-column basis only).

8888 Calculate coefficients on a column-by-column basis, and then transpose so that they are, in turn, calculated on a row-by-row basis.

9999 Calculate transposed array only, (i.e., on a row-by-row basis only).

```
 84 WRITE ($$ FT99) $
 85 MF = M $
 86 FOR I =(1,1,N) $ XSM(I) = 0.0 $
 87 FOR I =(1,1,N) $ FOR J=(1,1,M) $ XSM(I)=XSM(I) +(X(I,J).X(I,J)) $
 88 FOR K =(1,1,N-1) $FOR L=(K+1,1,N) $  COR(K,L) = 0.0 $
 89 FOR K =(1,1,N-1) $FOR L=(K+1,1,N) $ FOR J =(1,1,M) $
 90 COR(K,L) = COR(K,L) + (X(K,J).X(L,J)) $
 91 FOR K =(1,1,N-1) $FOR L=(K+1,1,N) $
 92 COR(K,L) = 1.0 - (SQRT((XSM(K) +  XSM(L) -(2.0.COR (K,L)))/MF)    )$
 93 GO BELOW $
 94 END$
 95 IF TRANS NEQ 9999 $ WRITE ($$ FT10) $
 96 NF = N $
 97 FOR J =(1,1,M) $ XSM(J) = 0.0 $
 98 FOR J =(1,1,M) $ FOR I=(1,1,N) $ XSM(J)=XSM(J) +(X(I,J).X(I,J)) $
 99 FOR K =(1,1,M-1) $FOR L=(K+1,1,M) $  COR(K,L) = 0.0 $
100 FOR K =(1,1,M-1) $FOR L=(K+1,1,M) $ FOR I =(1,1,N) $
101 COR(K,L) = COR(K,L) + (X(I,K).X(I,L)) $
102 FOR K =(1,1,M-1) $FOR L=(K+1,1,M) $
103 COR(K,L) = 1.0 - (SQRT((XSM(K) +  XSM(L) -(2.0.COR (K,L)))/NF)        )$
104 BELOW..
105 COMMENT PRINT AND PUNCH UPPER HALF OF  COEFFICIENT MATRIX WITH ONES
106  IN PRINCIPAL DIAGONAL OMITTED $
107 IF CHC EQL 6666 $
108 BEGIN
109 WRITE ($$ FMT2) $
110 IF TRANS EQL 9999 $ BEGIN
111 OUTPUT OWL(FOR K =(1,1,N-1) $ FOR L =(K+1,1,N) $ COR(K,L)) $
112 WRITE ($$ OWL, FMT5) $              END$
113 IF TRANS NEQ 9999 $ BEGIN
114 OUTPUT OTT(FOR K =(1,1,M-1) $ FOR L =(K+1,1,M) $ COR(K,L)) $
115 WRITE ($$ OTT, FMT5) $    END$     END$
116 COMMENT PREPARE COEFFICIENTS FOR LOWER HALF OF MATRIX WITH ONES IN
117 PRINCIPAL DIAGONAL INCLUDED $
118 IF CHC EQL 7777 $
119 BEGIN
120 IF TRANS  EQL 9999 $ BEGIN
121 COMMENT FILL PRINCIPAL DIAGONAL WITH ONES $
122 FOR K=(1,1,N)$COR(K,K) = 1.0 $
123 COMMENT FILL LOWER HALF OF MATRIX BY ASSIGNMENT $
124 FOR K =(2,1,N) $ FOR L =(1,1,K-1) $ COR(K,L) = COR(L,K) $
125 WRITE ($$ FMT6) $
126 FOR K =(1,1,N) $
127    BEGIN
128       OUTPUT TTT(FOR L =(1,1,K) $ COR(K,L))$
129       WRITE($$ TTT, FMT5) $
130    END $
131 END$
132 IF TRANS NEQ 9999 $ BEGIN
133 COMMENT FILL PRINCIPAL DIAGONAL WITH ONES $
134 FOR K=(1,1,M)$COR(K,K) = 1.0 $
135 COMMENT FILL LOWER HALF OF MATRIX BY ASSIGNMENT $
136 FOR K =(2,1,M) $ FOR L =(1,1,K-1) $ COR(K,L) = COR(L,K) $
137       WRITE ($$ FMT6) $
138 FOR K =(1,1,M) $       BEGIN
139       OUTPUT OUT(FOR L =(1,1,K) $ COR(K,L))$
140       WRITE($$ OUT, FMT5) $
141    END$         END$  END$
142 IF TRANS EQL 8888 $ ( TRANS  =  9999 $ GO TRANSPOSE ) $
143 GO START $ FINISH $
```

Input to Distance Coefficient Program

Alphanumeric heading card.--The first card of each data set (example data sets are shown in Table 7) contains alphabetical and numerical (alphanumeric) information for identification purposes. This information will be reproduced at the top of certain pages (Table 8) printed by the computer's printer. The card must be punched as follows:

(1) 5 in column 1.

```
 9        FT33(*DATA VALUES TRANSFORMED BY DIVIDING BY HIGHEST VALUE IN *,
10             *COLUMN*,W,W),
11        FT44(*DATA VALUES TRANSFORMED SO THAT MIN TO MAX RANGE = 0.0 *,
12             *TO 1.0 IN EACH COLUMN*,W,W),
13        FT55(*DATA VALUES TRANSFORMED BY DIVIDING BY 100.0 *,W,W),
14        FT99(*DATA ARRAY HAS BEEN TRANSPOSED AND COEFFICIENTS CALCULATE*
15        ,*D BETWEEN ALL POSSIBLE PAIRS OF ROWS OF ORIGINAL DATA ARRAY*,
16       W,W)$
17 FORMAT FMT5(*5 *,13X6.3,C  ),   FMT6(*DISTANCE FUNCTION COEFFICIENTS *,
18 *LOWER HALF OF MATRIX WITH ONES IN PRINCIPAL DIAGONAL INCLUDED*,W,W) $
19 FORMAT FTMN(*NUMBER OF ROWS = *,I3,* ,  NUMBER OF COLUMNS = *,I3,
20       *   IN ORIGINAL DATA ARRAY *,W,W) $
21 FORMAT FT10(*DISTANCE COEFFICIENTS BETWEEN ALL POSSIBLE PAIRS OF *,
22 * COLUMNS OF UNTRANSPOSED DATA ARRAY*,W,W) $
23 START..
24 INPUT ALPH(A1,A2,A3,A4,A5,A6,A7,A8,A9,A10,A11,A12),
25       CONTROL(N,M,TOP, OP, CHC, TRANS) $
26 READ ($$ ALPH) $  READ  ($$ CONTROL) $
27 IF TOP EQL 1111 $
28    BEGIN
29       INPUT DATI(FOR I =(1,1,N) $ FOR J =(1,1,M)$ XP(I,J)) $
30       READ ($$ DATI) $
31       FOR I =(1,1,N) $ FOR J =(1,1,M) $ X(I,J) = XP(I,J)   $
32    END $
33 IF TOP EQL 2222 $
34    BEGIN
35       INPUT DATD(FOR I =(1,1,N) $ FOR J =(1,1,M) $ X(I,J)) $
36       READ ($$ DATD) $
37    END $
38 OUTPUT ALPHA(A1,A2,A3,A4,A5,A6,A7,A8,A9,A10,A11,A12) $
39 WRITE ($$ ALPHA,FMA ) $
40 WRITE ($$ FMT1) $
41 COMMENT TRANSFORM VALUES BY DIVIDING BY HIGHEST VALUE IN COLUMN $
42 OUTPUT RWCL(N, M) $  WRITE( $$  RWCL, FT1N) $
43 IF OP EQL 3333 $
44 BEGIN
45 WRITE ($$ FT33) $
46 FOR J =(1,1,M) $
47    BEGIN
48       XH(J) = X(1,J) $
49       FOR I =(2,1,N) $
50       XH(J) = MAX(XH(J),X(I,J))$
51       FOR I =(1,1,N) $
52       X(I,J) = X(I,J)/XH(J) $
53    END $
54 END $
55 COMMENT TRANSFORM DATA VALUES SO THAT MIN TO MAX RANGE = 0.0 TO 1.0  $
56 IF OP EQL 4444 $
57 BEGIN
58 WRITE ($$ FT44) $
59 FOR J =(1,1,M) $
60    BEGIN
61       XH(J) = X(1,J) $
62       FOR I =(2,1,N) $
63       XH(J) = MAX(XH(J),X(I,J)) $
64       XL(J) = X(1,J) $
65       FOR I =(2,1,N) $
66       XL(J) = MIN(XL(J),X(I,J)) $
67       RGE(J) = XH(J) -  XL(J) $  XLRG(J) = XL(J)/RGE(J) $
68       FOR I =(1,1,N) $
69       X(I,J) = X(I,J)/RGE(J) - XLRG(J)  $
70    END $
71 END $
72 COMMENT          TRANSFORM VALUES BY DIVIDING BY 100.0 $
73 IF OP EQL 5555 $
74 BEGIN
75 WRITE ($$ FT55) $
76    FOR J =(1,1,M) $ FOR I =(1,1,N) $ X(I,J) =  X(I,J)/100.0 $
77 END $
78 TRANSPOSE..
79 COMMENT THIS SECTION WILL TRANSPOSE DATA ARRAY THAT HAS BEEN PREVIOUSLY
80 TRANSFORMED IF TRANS HAS BEEN SET TO 9999 AFTER BEING READ IN AS 8888
81 OR HAVING BEEN READ IN AS 9999  $
82 IF TRANS EQL 9999 $
83 BEGIN
```

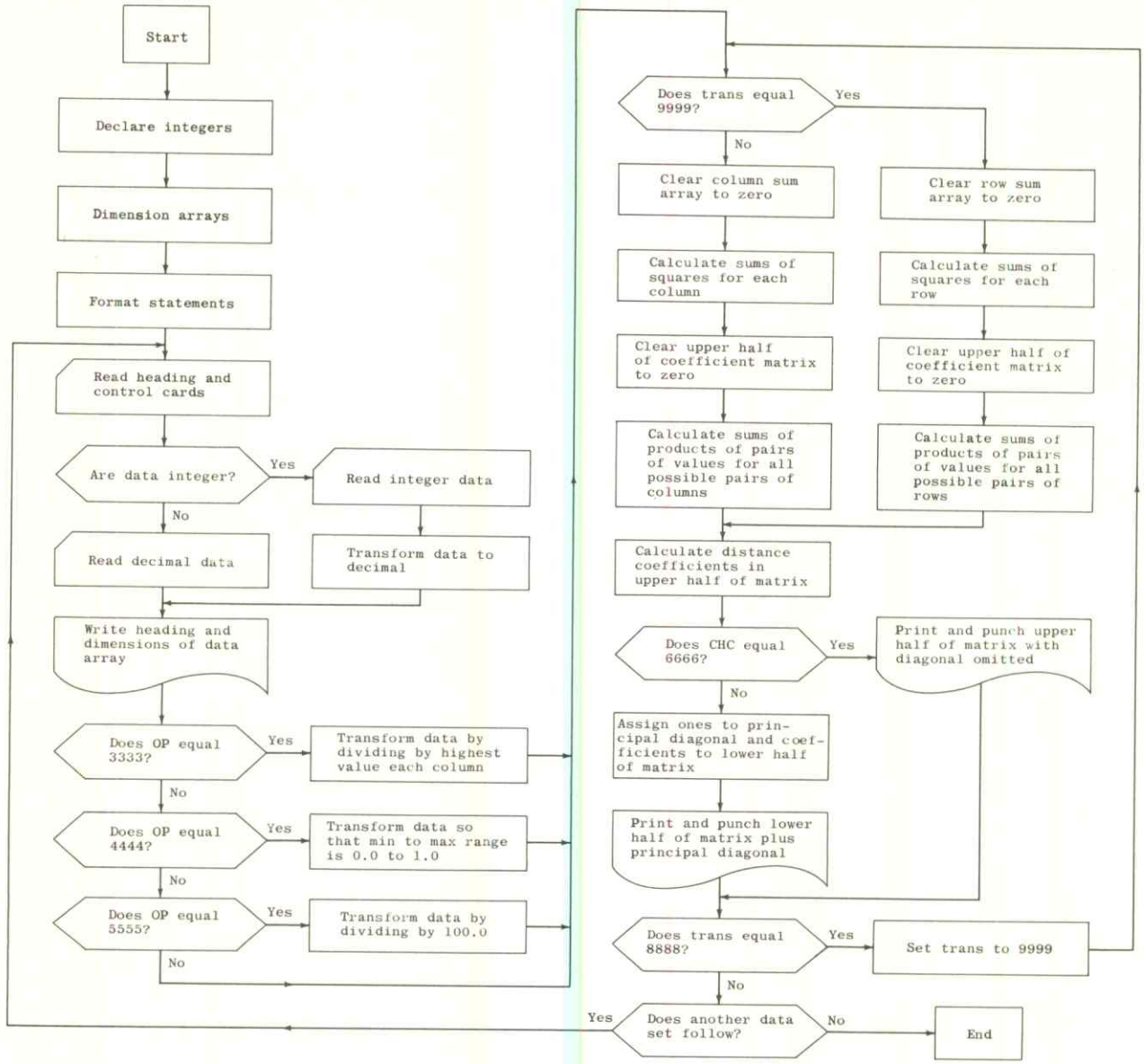


Figure 4. Simplified flow chart of major steps in distance coefficient program.

Table 6.--Listing of BALGOL statements in distance coefficient program. Each line represents one punched card of program. Numbers in left column are for reference purposes in this report, and in practice are placed within columns 73-80 on punched cards. Each program card must have 2 punched in column 1.

```
1 COMMENT PROGRAM 22, CALCULATION OF DISTANCE COEFFICIENTS, J.W.HARBAUGH$
2 INTEGER N,M, TOP, OP, L, K, XP(), I, J, CHC, TRANS, TP $
3 ARRAY  XP(200,25), X(200,25 ), XH(25), XL(25), RGE(25), XLRG(25),
4        XSM(130), COR(130,130) $
5 FORMAT FMA(*5$*,12A6,*$*,C3,W),
6         FMT1(B10,*GENERAL DISTANCE FUNCTION COEFFICIENTS*,W,W),
7         FMT2(*UPPER HALF OF COEFFICIENT MATRIX WITH ONES IN PRINCIPAL*,
8         * DIAGONAL OMITTED*,W,W),
```

(5) If specified in control card, transform data values by dividing by highest value in each column: lines 41-54.

(6) If specified on control card, transform data values so that minimum to maximum range of transformed values is 0.0 to 1.0: lines 55-71.

(7) If specified on control card, transform data values by dividing by 100.0: lines 72-77.

(8) If specified, calculate distance coefficients between all possible pairs of rows, as follows:

  (a) Clear row sum array to zero: line 86.

  (b) Calculate sum of squares for each row: line 87.

  (c) Clear upper half of coefficient matrix to zero: line 88.

  (d) Calculate sums of products of pairs of values for all possible pairs of rows: lines 89-90.

  (e) Calculate distance coefficients between all possible pairs of rows: lines 91-92.

(9) If specified, calculate distance coefficients between all possible pairs of columns, as follows:

  (a) Clear column sum array to zero: line 97.

  (b) Calculate sum of squares for each column: line 98.

  (c) Clear upper half of coefficient matrix to zero: line 99.

  (d) Calculate sums of products of pairs of values for all possible pairs of columns: lines 100-101.

  (e) Calculate distance coefficients between all possible pairs of columns: lines 102-103.

(10) If specified on control card, punch and print upper half of distance coefficient matrix with ones in principal diagonal omitted: lines 107-115.

(11) If specified on control card, punch and print lower half of coefficient matrix with ones in principal diagonal: lines 118-141.

(12) If specified on control card, previously untransposed data array will be transposed and coefficients calculated on a row-by-row basis of original data array. This is done by setting TRANS to 9999 and going to label TRANSPOSE in line 78: line 142.

(13) Read in next data set consisting of heading card, control card, and data cards. By going back to initial input statements, successive sets of data will be read and coefficients calculated until all data sets have been processed: lines 143, 23.

We may visualize the relationships between Well 1 and Well 2 in terms of the transformed values of the unit cube formed by Tests 1, 2, and 3 (Fig. 3). On the other hand, we cannot portray graphically the distance between Tests 1 and 2 in terms of measurements at the four wells because four axes at right angles to each other (thus four dimensions) are needed.

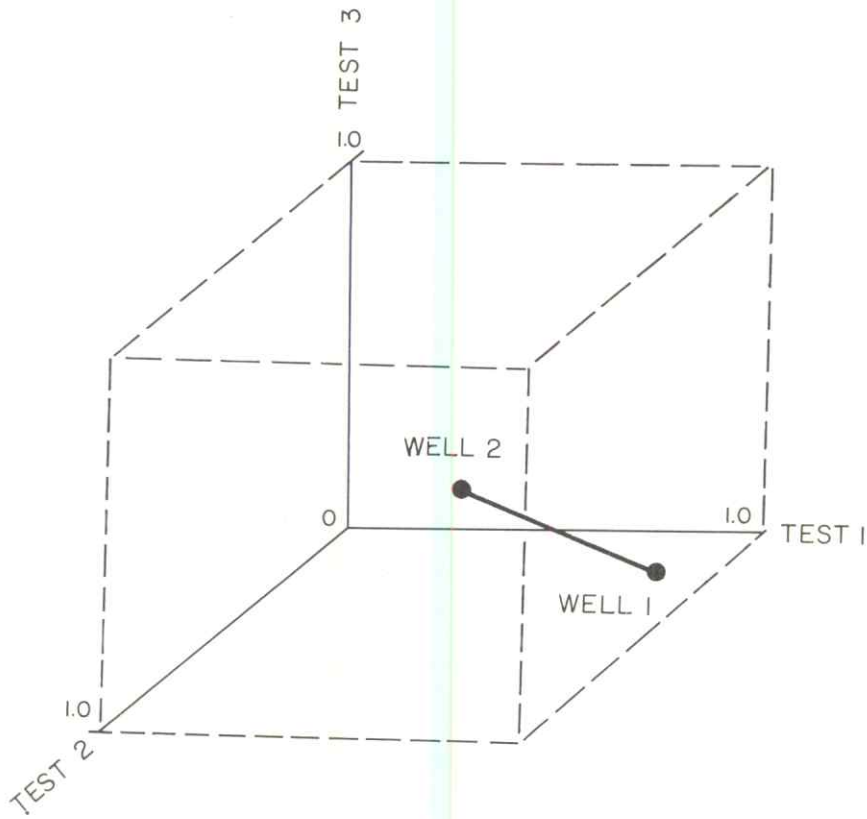


Figure 3. Distance between hypothetical Wells 1 and 2 given by transformed values (Table 1) of Tests 1, 2, and 3.

Steps In Distance Coefficient Program

Steps in the program are outlined in the flow chart (Fig. 4) and are listed in detail in Table 6. Each line (card) of the program is numbered, permitting parts of the program to be referred to by line numbers. Major steps in the program are listed below:

(1) Integer declarations: line 2

(2) Dimension arrays: lines 3-4.

(3) Format statements: lines 5-22.

(4) Input and read statements: lines 23-37.

$$d_r = 1.0 - \sqrt{\frac{1.62 - 2.88 + 2.06}{4}}$$

$$d_r = 1.0 - \sqrt{\frac{0.80}{4}}$$

$$d_r = 1.0 - .447 = .553$$

Calculating the distance coefficient between three pairs of tests in Table 1, we obtain the R-mode coefficient matrix (Table 3).

Table 3.--R-mode distance coefficient matrix between tests of Table 1.

	Test 1	Test 2	Test 3
Test 1	1.00	.55	.38
Test 2		1.00	.77
Test 3			1.00

Calculation of the Q-mode distance coefficient between Well 1 and Well 2 is illustrated in Table 4, and the Q-mode distance coefficients between all pairs of wells are listed in Table 5.

Table 4.--Calculations to show steps in calculating Q-mode distance coefficients, d_q, between transformed values at Wells 1 and 2 of Table 1.

	$(\text{Well 1})^2$	-	2x(Well 1)x(Well 2)	+	$(\text{Well 2})^2$
Test 1	1.00		-1.28		.41
Test 2	.25		- .75		.56
Test 3	.02		- .16		.29
$\sum$	1.27		-2.19		1.26

$$d_q = 1.0 - \sqrt{\frac{1.27 - 2.19 + 1.26}{3}}$$

$$d_q = 1.0 - \sqrt{\frac{.34}{3}}$$

$$d_q = 1.0 - .34 = .66$$

Table 5.--Q-mode distance coefficient matrix between wells of Table 1.

	Well 1	Well 2	Well 3	Well 4
Well 1	1.00	.66	.35	.33
Well 2		1.00	.68	.59
Well 3			1.00	.58
Well 4				1.00

square root of this value from 1, we obtain a coefficient ranging between 0 and 1. For example, in a cube the maximum modified distance is $\sqrt{\frac{3}{3}}$, which in turn yields a distance coefficient of $d = 1 - \sqrt{\frac{3}{3}} = 0$. The greater the modified distance value, the smaller the value of the distance coefficient, and vice versa.

Examples of calculation in R and Q modes:--It may be desirable to calculate distance coefficients as a measure of the degree of similarity between both variables and samples. For example, a psychologist who had given a series of psychological tests to school children might wish to measure the degree of similarity between tests (R mode), or conversely, the degree of similarity between children (Q-mode) on the basis of the test scores. Both R and Q modes may be computed from the same data.

It may be helpful to illustrate the R and Q modes with the data of Table 1, which contains the results of three tests made at each of four wells.

Table 1.--Hypothetical well data illustrating a linear transformation of raw data in calculation of distance coefficients.

	Original data			Data transformed by dividing by highest value		
	Test 1	Test 2	Test 3	Test 1	Test 2	Test 3
Well 1	140	6	116	1.00	.50	.15
Well 2	90	9	420	.64	.75	.54
Well 3	65	12	770	.46	1.00	1.00
Well 4	0	6	558	.00	.50	.72

Steps in calculating the distance coefficient between Test 1 and Test 2 (R mode) of Table 1 are outlined in Table 2.

Table 2.--Steps in calculating R-mode distance coefficients, d_r, between transformed values of Tests 1 and 2 of Table 1.

	$(\text{Test 1})^2$	-	2x(Test 1)x(Test 2)	+	$(\text{Test 2})^2$
Well 1	1.00		-1.00		.25
Well 2	.41		- .96		.56
Well 3	.21		- .92		1.00
Well 4	.00		.00		.25
$\sum$	1.62		-2.88		2.06

To yield a coefficient that will satisfy these requirements, it is necessary (1) to calculate the sums of the squared differences within a multidimensional "hypercube" whose sides have unit length, (2) to divide by the number of variables to obtain the mean square, and (3) to subtract the square root of the mean square from 1. Expressed as an equation, we obtain

$$d = 1 - \sqrt{\frac{\sum_{i=1}^{n} (X_{i1} - X_{i2})^2}{n}} \quad ,$$

where

d = distance coefficient,

n = number of variables,

X_i = variables specified by subscript, i, with the provision that values of X_i must not exceed 1.

Transformations of the original data are generally needed to insure that X_i does not exceed one. If the data values are all positive, a simple linear transformation is to divide by the highest value observed for each variable. This transforms the data values so that the highest value is 1 and that all data values may be "contained" within a multidimensional "cube" whose sides are of unit length. Other transformations, which may be linear or nonlinear, can be used so that the lowest value of a given variable is set to 0.0 and the highest to 1.0, with intermediate values ranging from 0.0 to 1.0. If the data are in percentage form, and no value exceeds 100 percent, a suitable linear transformation is to divide by 100. In the program described here, a choice of three linear transformations is provided. Nonlinear transformations could be made readily by making minor changes in the program. In theory, distance coefficients may be calculated with either positive or negative data values, but the computer program, as written, provides for transformations which require positive data values. If negative values are present, it is suggested that a constant be added so that all values become positive.

If coefficient values are not to exceed 1.0, the calculated distance between two points in the multidimensional "cube" must not exceed 1.0. For example, in a three-dimensional cube whose sides are of unit length, the maximum distance between two points within the cube is the diagonal of the cube, whose length is

$$\sqrt{1^2 + 1^2 + 1^2} = \sqrt{3}$$

Similarly, within a 4-dimensional "cube", the maximum distance between two points is $\sqrt{4}$. Therefore, if we divide the sum of the squared distances measured along the axes, by the number of variables or dimensions, the maximum value for distance is 1. If we subtract the

Similarly, in the four-variable (four-dimensional case), the distance between two points is

$$D = \sqrt{(W_1 - W_2)^2 + (X_1 - X_2)^2 + (Y_1 - Y_2)^2 + (Z_1 - Z_2)^2} \ .$$

For the distance between two points in n-dimensional space, we may generalize by writing

$$D = \sqrt{\sum_{i=1}^{n} (X_{i1} - X_{i2})^2} \ ,$$

where i ranges from 1 to n, and we obtain the square root of the sum of the squared difference in each dimension for n dimensions. For computational purposes (Sokal, 1961), it is easier to express the equation as

$$D = \sqrt{\sum_{i=1}^{n} (X_{i1}^2 - 2X_{i1} X_{i2} + X_{i2}^2)} \ .$$

Adapting distance as a coefficient:--In calculating a coefficient that is a measure of similarity, it is convenient to limit its absolute value to the range 0.0 to 1.0. Furthermore, if the coefficient is to be analogous to widely used measures of correlation, such as the product-moment correlation coefficient, it should indicate maximum similarity when its absolute value approaches 1.0, and minimum similarity when it approaches zero.

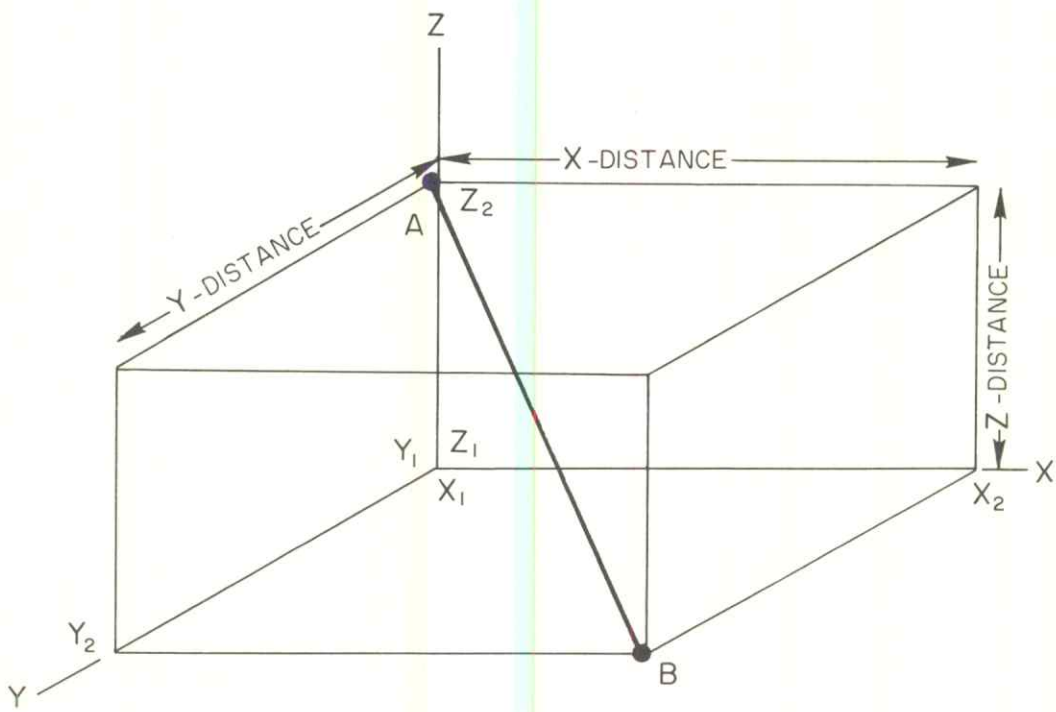


Figure 2. Distance from A to B in three-dimensional case is equal to square root of sum of squared X-distance, Y-distance, and Z-distance.

DISTANCE COEFFICIENT PROGRAM

General

Distance coefficients are based on the concept that a quantitative measure of the degree of similarity between two variables or two samples is provided by the distance that separates them in a rectangular coordinate system; the shorter the distance, the greater the degree of similarity, and vice versa. The use of distance as a measure of degree of similarity in taxonomy has been emphasized by Sokal (1961) and Sokal and Sneath (1963, p. 143-151, 300-301). The distance coefficient, as defined in this report, is a modification of the distance coefficient described by Sokal (1961).

Computation of distance:--The distance, D, between two points whose locations are specified in a two-variable cartesian coordinate system is given by a simple adaptation of the Pythagorean theorem, and is computed by the formula

$$D = \sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2} ,$$

where X_1, X_2, Y_1, and Y_2 are the coordinate values of the two points and D is distance (Fig. 1).

The Pythagorean theorem is equally valid in three or more dimensional space. A clear discussion of the proof is given by Schwartz (1961, p. 100-102). For three variables (Fig. 2), X, Y, and Z, the distance between two points is given by the equation

$$D = \sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2 + (Z_1 - Z_2)^2} ,$$

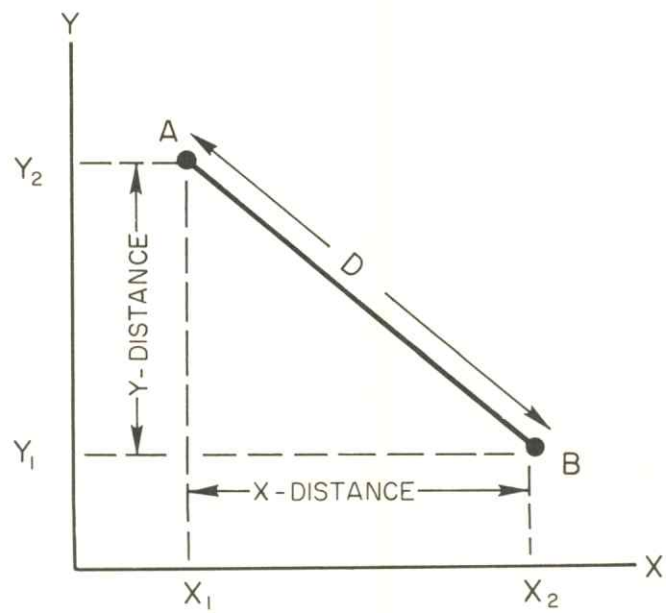


Figure 1. Two-variable or two-dimensional case in which distance, D, is equal to square root of sum of squared X-distance and Y-distance.

INTRODUCTION

The two computer programs described here have been developed for use with geological data, but they should have broad application. The programs complement each other; both programs calculate coefficients which are measures of similarity. The first program described employs distance as a measure of similarity. Because distance has not been extensively used as a measure of similarity, the theory of distance coefficients is discussed below. The second program calculates Pearson product-moment correlation coefficients, which are used widely and are well known. Although programs for calculation of correlation coefficients will be found in most computation center libraries, the program described here has certain extra features useful in geological applications.

Both programs may be used to prepare input data for factor analysis, a statistical technique which is beginning to be widely used in geology (Imbrie and Purdy, 1962; Krumbein and Imbrie, 1963). Some geological applications of factor analysis have used correlation coefficients, but it now appears that distance coefficients have important advantages in factor analysis in which comparisons between samples are made (as contrasted to comparisons between variables). A geological comparison of the results of factor analysis based on both correlation coefficients and distance coefficients is provided by Harbaugh and Demirmen (in press).

Both programs are written in a dialect of ALGOL known as SUBALGOL, or simply as BALGOL. The programs are usable on IBM 7090 and 7094 computers, and if the array dimensions are reduced and certain other details of the programs modified slightly, they can be used with the Burroughs 220 computer.

BALGOL card decks of the programs may be obtained for $5.00 each from the Kansas Geological Survey for a limited time. Trial data sets listed in Tables 7 and 10 will be supplied with the program decks so that the programs can be tested and output compared with the examples of output listed in Tables 8, 11, 12, and 13.

Programs written in BALGOL may be used with IBM 7090 or 7094 computers only in conjunction with BALGOL system tapes. The 7090 BALGOL system of tapes may be obtained by sending four magnetic tapes to the Computation Center, Stanford University, Stanford, California. The systems will be recorded on the tapes and returned. Tapes that are 1200 feet long are adequate. As an alternative, the Kansas Geological Survey will distribute binary decks of both programs for $5.00 each. Changes cannot be made in programs in binary form, and there is no guarantee that binary programs will work properly on different 7090 or 7094 computers. However, if they work, the necessity of obtaining the BALGOL tape systems is avoided.

A manual describing BALGOL and entitled "Burroughs Algebraic Compiler, Revised Edition" (1963), may be obtained from the Burroughs Corporation, Detroit 32, Michigan. "An Introduction to BALGOL" (1961), by R. V. Oakford and J. M. Gere, is published by the Wadsworth Publishing Company and also provides an introduction to BALGOL.

COMPUTER CONTRIBUTIONS

This publication reports additional computer programs of interest to earth scientists and is the third to be published by the Kansas Geological Survey. The first two programs, "BALGOL program for trend-surface mapping using an IBM 7090 computer," by J.W. Harbaugh, and "FORTRAN II program for coefficient of association (Match-Coeff) using an IBM 1620 computer," by R.L. Kaesler, F.W. Preston, and Donald Good, appeared in the Survey's special distribution series and were so well received it was decided to make other programs available. Inasmuch as the programs are timely, a special editorial procedure has been set up for handling these manuscripts so that they can be made readily available as soon as possible.

The purpose of publishing computer programs was adequately summarized by J.W. Harbaugh in the preface of the first publication and is reproduced here.

> "This special publication describes a computer program that will be useful to geologists in Kansas and elsewhere; it is the first of a series of publications of the Kansas Geological Survey in which the objective is to present details of computer programs that should be of general usefulness to geologists and petroleum engineers. The computer revolution is fast sweeping through the petroleum industry, but many small companies and independent operators do not have employees on their staffs who are skilled in computer applications. It is the intention of the Kansas Geological Survey to provide some assistance in computer applications in solving geological and petroleum engineering problems."

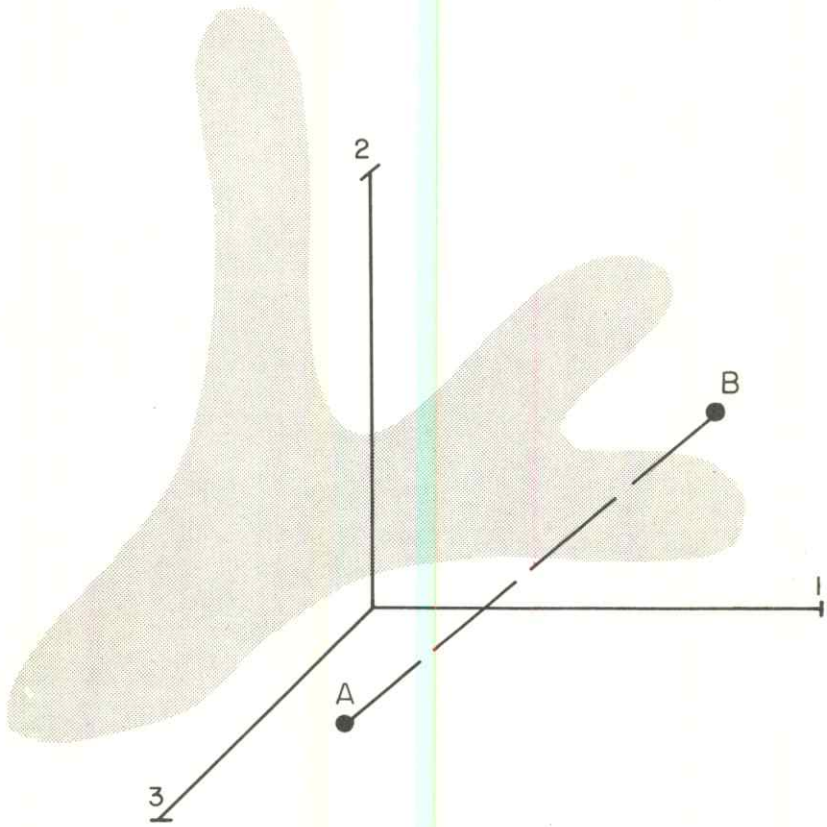
Most of the computer programs published in this series were developed to solve a particular geological or engineering problem, and for many of these problems results of the research will also be published.

The role of the State Geological Survey in relation to the challenge of the "new" era of computers has been stated well by W.W. Hambleton.

> "The Geological Survey is faced with major changes as a result of the computer revolution in the mineral industries, particularly the oil industry. For many years the Survey has maintained a central file of oil well information which has served the petroleum industry in Kansas. In the next few years, this function of the Survey will be partly usurped by automated, computer controlled well information centers that are currently being established by the oil industry. Research tasks are also being redefined as a result of the advent of computers. Major areas of study are being bypassed and others are becoming significant. If the Survey is to continue its role as a service to industry and the state, it must adapt itself to the computer revolution. In response to this new force in science and technology, the Survey is embarking on a plan to gradually adapt its operations to make use of computers where possible. . ."

Of course, part of the role of the Survey is making available results of research and that is the reason for this series of papers on computer applications to the earth sciences. Any comments or suggestions should be addressed to the editor.

Balgol Programs for Calculation of Distance Coefficients and Correlation Coefficients Using an IBM 7090 Computer



By
John W. Harbaugh
Stanford University

SPECIAL DISTRIBUTION PUBLICATION 9

State Geological Survey
The University of Kansas, Lawrence, Kansas
1964