QUICLSTR: A FORTRAN program for hierarchical cluster analysis with large numbers of subjects

PHILIP A. BELL

ACTS Computing Corporation, Southfield, Michigan 48076

and

JOHN L. KOREY

*Department of Political Science
California State Polytechnic University
Pomona, California 91768*

QUICLSTR is a program for hierarchical cluster analysis. It is especially useful for problems involving large numbers of subjects. It can handle most such problems efficiently even on a fairly small computer, and does not require auxiliary storage of data during execution (cf. Levinson & Funk, 1973).

The program uses squared Euclidean distance as the measure of dissimilarity between groups, and clusters according to Ward's (1963) error sum of squares method. At the user's option, data will be converted to standard (z) scores prior to clustering. Input to the program consists of (1) program parameters, read from card, and (2) raw data, read from card, tape, or disk. Output includes, for each step in the clustering process, (1) the error sum of squares, (2) the increase in this error over the previous step, and (3) the acceleration of the increase of error. In addition, for as many steps as the user requests, the members of each group are listed.

The primary limitation of the program is that when the number of subjects clustered is large, the number of variables employed must be correspondingly small. This limitation can

be overcome by factor analyzing variables and clustering on factor scores. If the number of both subjects and variables is relatively small, a program such as that proposed by Veldman (1967) is probably preferable to QUICLSTR.

Computer. The program has been run on both the CDC 3150 and the CDC 3300.

Language. The program uses FORTRAN IV. It has been run under both ANSI (i.e., USASI) and MS FORTRAN compilers.

Restrictions. Restrictions will vary with the amount of core storage actually available for program and data at the user's installation. On the CDC 3150, with 26K words of available storage, the limitations are $N \times M + 5N =$ about 9900. The CDC 3150 version of the program is presently dimensioned for 650 subjects and 10 variables.

Program dimensions may be adjusted simply by altering the DIMENSION and COMMON statements at the beginning of the program.

Acknowledgements. The program was developed at the University of Florida and at the California State University and Colleges, and computer time was made available by these institutions.

Availability. A source listing, write-up, and sample run may be obtained free of charge by writing John L. Korey, Department of Political Science, California State Polytechnic University, Pomona, California 91768.

REFERENCES

- Levinsohn, J. R., & Funk, S. G. CLUSTER: A hierarchical clustering program for large data sets ($n > 100$). *Behavior Research Methods and Instrumentation*, 1973, 5, 432-433.
Veldman, D. J. *FORTRAN programming for the behavioral sciences*. New York: Holt, Rinehart, and Winston, 1967, Pp. 308-317.
Ward, J. H., Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* 1963, 58, 236-244.

CORD: A program to compute the directional correlation coefficient d

JOHN E. ETTLIE

*Department of Management
University of Illinois, Chicago Circle Campus
Chicago, Illinois 60680*

The coefficient of directional correlation, d, was introduced by Strahan (1971). It is a nonparametric statistic that is used to describe the degree of association of two variables over time. Time-series and longitudinal data are considered candidates for analysis by this method. "The term directional correlation is given to the tendency for two variables to change values in the same direction from one sampling occasion to the next," Strahan (1971, p. 211). The d coefficient is computed by assigning a plus (+) to same direction changes, a minus (-) to opposite direction changes, and a zero (0) to unchanged variables in adjacent time periods of data regardless of the magnitude of these changes. The d coefficient is defined by the following equation (Strahan, 1971, p. 211).

$$d = \frac{p - m}{p + m} \quad (1)$$

In this equation p and m represent the accumulated plus and minus signs in the data comparisons.

CORD is a computer program written in FORTRAN (WATFIV) which calculates the d coefficient for time-series

data. The program makes paired comparisons of adjacent time period data cells of two variables and computes the d coefficient according to Equation 1. CORD is presently limited to 99 variables and 99 observations for each variable, but dimension statements can easily be modified by the user to increase or decrease the capacity of the program, subject to storage restrictions of the user's computer installation.

Input. Data input includes the number of variables, the number of observations for each variable, and a specification of the format by which the data is to be read in. The remainder of the input is the time-series data punched in the format specified by the user.

Output. The output is divided into three segments. First, a listing of each d coefficient with a specification of each variable pair. Second, a systemic matrix of these d coefficients. (Presently limited to a 13 x 13 matrix, due to print limitations.) Third, a listing of the time-series data for each variable according to the order of input of observations on the original data cards.

Computer. This program was run on an IBM 370. Computing services were provided by the University of Illinois, Chicago Circle Campus computer center.

Availability. The User's Manual can be obtained at no charge from John Ettlie, Department of Management, University of Illinois, Chicago Circle Campus, Box 4348, Chicago, Illinois 60680.

REFERENCES

- Strahan, R. F. A coefficient of directional correlation for time-series analyses. *Psychological Bulletin*, 1971, 76, 211-214.