Cluster Analysis and the BC TRY System

Daniel E. Bailey

Tryon-Bailey Associates

This is an updated copy of a brochure produced by Daniel E. Bailey for Tryon-Bailey Associates, Boulder, Colorado around 1971.

Currently available component programs, in a successor system called TRYSYS, are listed in "*Component Programs of the Tryon System of Cluster & Factor Analysis.*" This document, and other information about TRYSYS, is available from Robert Dean, robertBdean@gmail.com.

The original BCTRY system is documented in Robert C. Tryon and Daniel E. Bailey, *Cluster Analysis,* McGraw-Hill, 1970.

CLUSTER ANALYSIS
AND THE BC TRY SYSTEM

Daniel E. Bailey

Tryon-Bailey Associates, Inc.
and
University of Colorado

This paper is a brief discussion of cluster analysis in multivariate data. The scope of the paper ranges from the analysis of variables through the analysis of objects observed on those variables. In the interest of brevity, only passing mention is made of many of the topics.

Before digital computers were widely available, many of the analytic tools described here could be used only on small sets of data and others were not available at all. The BC TRY System is an implementation of these tools in a system of computer programs. The programs are described in connection with the various cluster analysis procedures discussed here.

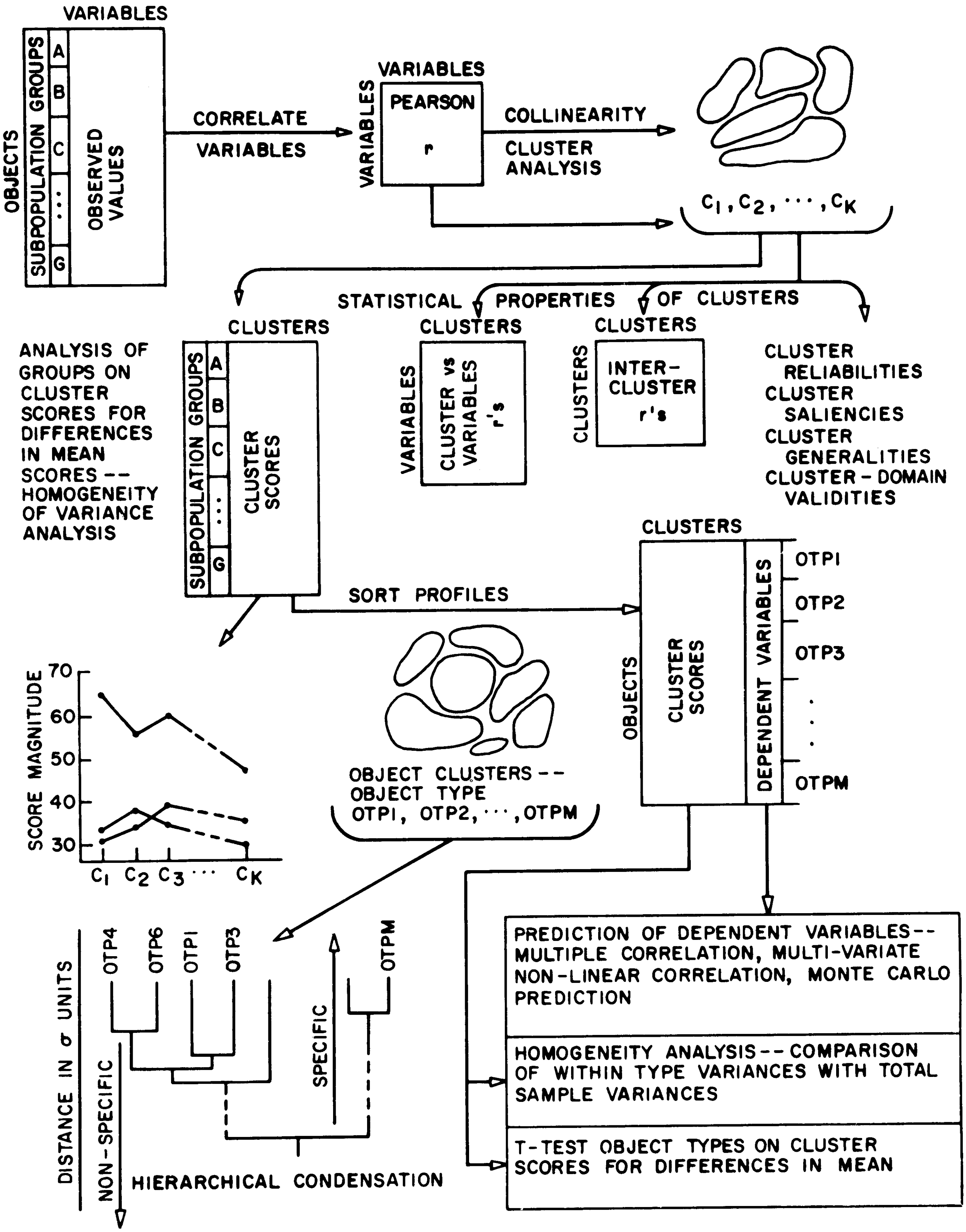
Figure 1 is a schematic presentation of the major stages of cluster analysis. Many details and specialized techniques of cluster analysis are not indicated in the figure. For example, factor analysis can be substituted for collinearity cluster analysis in certain circumstances. Also, the analytic processes can be applied to a number of sets of data and the results of the various separate analyses can be compared using other techniques of cluster analysis. The reader is referred to other literature (see Tryon, R. C. and Bailey, D. E., CLUSTER ANALYSIS, McGraw-Hill, 1970) for more discussion and additional references.

The data amenable to cluster analysis are those of multivariate experiments in which a number of different objects (subjects, respondents, etc.) are each assessed (observed, measured, counted, etc.) on a number of different variables. The objects may be selected from an undifferentiated population or they may be samples from several subpopulations. The variables generally are restricted to numerical representations of ordinal (or dichotomous) characteristics of the objects with certain formal properties (see below).

There are four goals or sub-objectives in cluster analysis of variables: 1) condensation of the variables into basic dimensions that capture all of the general covariation among the variables, 2) selection of homogeneous subsets of variables that are observable representations of the basic dimensions, 3) description of the statistical properties of the dimensions and clusters, and 4) geometric (graphical) description of the cluster structure of the data. Accomplishment of these objectives in the BC TRY System is called V-analysis.

The objectives of multivariate analysis, as usually stated, are centered on the observed variables. The point of view of this paper is that variable analysis, V-analysis, is only a preliminary stage. In much multivariate analysis work, specific individual differences among objects are of primary interest. These are represented only generally in V-analysis. However, cluster analysis of objects, called O-analysis in the BC TRY System,

Figure 1 Schematic Represention of Analysis of One Data set



provides specific information about individual differences.

Before O-analysis can be undertaken, the objects must be described in terms of the basic dimensions developed in the V-analysis of the data. Since the dimensions are not measured directly they must be estimated. The form of this estimation process is a linear composite of the observed variables, weighting the observed variables as a function of the dimensional structure of the V-analysis. The results are called factor or cluster scores, one score for each object on each dimension or cluster.

Once a score has been estimated for an object on each of the dimensions or clusters from V-analysis an object can be represented as a point in score space or by the graph of a score profile. Two objects that have the same score profiles are located at the same point in score space. Also, two objects having the same "pattern" of scores but different score magnitudes (i.e., elevations) are collinear in score space. Because of these and other properties there are a variety of ways to define object clusters. There are two broad classes of object clustering procedures: 1) proximity clustering, which selects object clusters on the basis of small distances between objects in score space (total pattern similarity in the score profiles); and 2) collinearity clustering, which selects object clusters on the basis of proportionality of patterns in the score profiles, not directly depending on profile elevation. The results of an object cluster analysis are simple groupings of the objects, called O-types in the BC TRY System. The clusters have several statistical characteristics in terms of the cluster scores from which the O-analysis is derived.

When the object clusters are treated as single points (mean scores) in the score space, the V-analysis concepts apply to them. Their relationships can be described statistically just as the relationships of variables and clusters in V-analysis can be described. However, treating each object cluster as a sub-sample leads to an analysis of cluster scores similar to the analysis of variance, comparing the within object-cluster variance with the overall variance of the cluster scores. These score homogeneities [1 - (within/total)], can be averaged across the score dimensions to summarize the uniformity of score profiles for an object cluster. Also, the weighted average of the homogeneities for a cluster score dimension, averaging across all object clusters, indicates the generalized coefficient of regression of the cluster scores on the object clusters, the coefficient of curvilinear correlation, eta.

In multivariate studies objects frequently are observed on two kinds of variables, the several independent variables on which V-analysis is based, and one or more dependent variables serving as target variables or criterion variables in multivariate prediction studies. Multiple linear regression techniques can be shown to be less efficient in prediction, when prediction is good, than a variant of prediction known in the BC TRY System as O-type prediction. Where the dependent variable scores within object clusters are homogeneous, knowledge of which cluster an object matches in cluster score profile immediately provides knowledge about the dependent variable for the object. The best estimate of the dependent variable value for a newly observed object is the mean score on the dependent variable for the object cluster that the new object most closely resembles, i.e., whose cluster score profile is most similar.

The Variables

In selecting variables for a multivariate study several considerations are important. The pertinancy of the variables is the basis for success or failure of a study. If the variables observed are not directly or indirectly involved in a phenomenon, the results of the V-analysis or the O-analysis of the data will not have a bearing on the phenomenon. The happy situation is one in which the variables producing the phenomenon are directly observable or measureable. However, when they are not directly observable, the researcher is required to obtain indirect measurements of the variables. These indirect measures become the variables in V-analysis. If the choice of measurement procedures is appropriate, the dimensions from V-analysis of the measurements will be more nearly directly related to the unobservable variables and the researcher has a means of describing their structure. Also, in such fortunate circumstances, the results of an O-analysis provide a population segmentation that is a truer representation of object types in the actual phenomenal world than if the variables observed were not pertinent to the target phenomenon.

Selecting variables is perhaps the most difficult part of multivariate research. If the researcher is able to obtain direct measures, there is no problem. However, the researcher must be able to define the variables in which he would be interested if he were able to measure them directly. Once the domains of variation are defined, the researcher develops ad hoc indirect measures of the variables, which in turn become the observed variables. Great ingenuity often is required to be able to devise ad hoc measures that approximate the target variables. Multivariate studies often are performed on data gathered without much thought about the relevant domains of variation. Such studies often produce systematic results that reflect some covert model of the phenomenon in the mind of the person who devised the measurement procedures. However, a systematic approach provides more interpretable and useful results in both V-analysis and O-analysis.

It is important, where indirect measurement is necessary, to have multiple measurements of each of the target variables. Because the measurements are surely mixtures of numerous sources of variation, including the target variable, the best way to obtain a more pure estimate of the target variable is to devise numerous measurement procedures having only the target variable in common. Even though the target variable is not very well measured by any of several indirect measures, the composite of the indirect measures will be a better approximation, providing that the measures are not systematically correlated with other sources of variation.

The variables in a multivariate study must be considered to be a sample from a population of variables. It is important to have a balanced representation of the various domains of variation so that each of the domains has a reasonable chance of playing the role in the analysis that it does in the actual phenomenon. If one domain is copiously represented in a study, the importance of that domain as a source of general variation in the study may be more impressive than the domains less well represented in the study. At the other extreme, if a domain of variation is represented by a single variable, and both the domain and variable are independent of other domains and variables in the study, the domain will not appear to be a part of the basic dimensional structure of the phenomenon.

Certain measurement problems are important in any discussion of multivariate analysis. Although correlations can be calculated on any numerical multivariate data, the meaning of the correlations (and subsequent analysis) depends on the meaningfulness of the numerical data in relation to the variables they are intended to represent. An analysis of data may be performed without assurances that the results of the analysis can be interpreted reasonably. When the data are direct measurements there is of course no problem. However, when indirect measurement is required, the data obtained should have certain basic characteristics. The ordering of subjects in the data must be the same as the ordering of objects in the actual properties the data represent. Stronger interpretations and statements of differences in magnitude can be made when the measurements preserve proportionality of differences in the properties measured, i.e. where two equal differences in measurement at different places on the scale of measurement represent two equal differences in the properties measured. The emphasis in multivariate analysis, particularly in V-analysis, is on correlation rather than on variance. As such, the variables dealt with are standardized variables with arbitrary origin (mean) and unit (variance) of measurement. Methods are available to work with covariances and variances directly but they are of no particular advantage unless the measurements have a natural origin and unit of measurement that are important to preserve in the analysis. In O-analysis it is perhaps of greater interest to retain natural origins and units of measurement but the general practice is to use standardized values with arbitrary means and variances.

Qualitative data are at times difficult to deal with, particularly if there is no clear ordering of the qualities as a function of the variable that the qualitative observations represent. If the qualitative distinctions constitute an ordering with respect to the phenomenon being studied, then one or another scaling technique can be used to transform the qualitative data into numerical data. Also, if the qualitative distinction is binary, the substitution of any numerical code for the two categories is usually satisfactory.

Where missing data are common in a study, the correlations forming the basis of V-analysis are based on a variety of sample sizes and a variety of different samples. This variation makes the coefficients of correlation non-comparable and any analysis based on the matrix is of questionable value. In general, the more missing data, the more questionable the results of analysis. Also, with large amounts of missing data it becomes impossible to score objects on the dimensions of V-analysis, and O-analysis becomes tenuous.

Where the distributions of the variables are highly skewed (i.e., non-symmetric) it is difficult to make clear interpretations of the results of V-analysis, and O-analysis can often have pathological properties. For dichotomous variables the skewness is represented in a disproportionate number of objects on one side of the dichotomy -- producing artifactual variation in the correlation matrix as a function of the skew.

Each of the variables included in V-analysis is assumed to be a separate variable, having no functional or mechanical relationship with the other variables. A common error made in multivariate analysis is to include a total score variable as well as the sub-scores from the total. The correlation matrix for such a set of variables causes problems in the methods used to determine the basic dimensions of the data. Inclusion in V-analysis of variables that are based on other variables in the analysis should be avoided. Special techniques in the BC TRY System permit an investigator to hold aside such variables in an initial V-analysis and later (in the same computer run) to integrate the dependent variables back into the V-analysis.

The Objects

The term "object" is used in the BC TRY System rather than subject, respondent, etc. because of the generality of the term and the diversity of applications of the System. The one restriction on defining an object is that it be a separable and distinct unit for which the variables in the study can be assessed. The value of a measurement obtained for one object may not be determined by another object in a direct overlapping way, (i.e., independence of observations is assumed). For example, a study should not include as separate objects both census tracts and voting precincts if they share common boundaries within which the same subpopulation is responsible for generating the values of the variables. Another example of non-distinct and overlapping objects is the situation where a group average on a variable and the scores for the individual members of the group are included in a sample, as distinct objects. Objects, within this general rule, can be any sort of thing . Some of the objects that have been involved in applications of the BC TRY System are persons, census tracts, voting precincts, state delegations to the U. S. Senate, samples of earth from far-flung corners of the world, mosquitos, elm trees, sites in glacial valleys, trials in a learning sequence, jobs in an employment service, etc. Repeated measures on a single subject, known as ipsative measures, are quite legitimate where the objects of interest are the occasions of observation and different subjects are not mixed in calculating the correlations for V-analysis. That is, in an ipsative study the V-analysis is done within an individual subject.

In order for the results of an analysis of data to be pertinent to the population in which the phenomenon being studied is important, the sample of objects must carefully be chosen to represent that population. Representativeness of samples of objects for a target population is as important as representativeness of the sources of variations in the variables included in a study. Often, the various strategies of random sampling, stratified sampling, etc., are used to assure the proper constitution of the sample.

A good rule of thumb in selecting sample size is to have at least twice as many objects in the sample as there are variables in the study. When considerations of representativeness dictate large numbers of subsamples, the overall minimal sample size may be larger. Because of the descriptive nature of multivariate analysis, we are concerned about the adequacy of the statistics as estimates. The larger the sample size, the more adequate the statistics are as estimates of population parameters. Also, certain methods of profile analysis in O-analysis require relatively large samples.

V-analysis

There are many different procedures used in multivariate analysis of variables. However, there are four general goals of V-analysis involved: 1) separation of the observed variables into basic dimensions, 2) selection of clusters of variables, i.e., segmentation of the set of variables into clusters of homogeneous measures of the basic dimensions, 3) calculation of the statistical properties of clusters, and 4) display of the geometric properties of the basic dimensions, clusters, and the observed variables.

Dimensional analysis. The reduction of a set of variables into a set of independent dimensions is widely known in the social and behavioral sciences as "factor analysis." In the more mathematically sophisticated disciplines, procedures similar to factor analysis are referred to by such names as "eigen vector" problems. Regardless of the differences in detail and nomenclature, the procedures are all designed to define a set of dimensions forming a basic variable space in which each of the observed variables is represented as a point (or vector). When a variable is represented as a row and column in a correlation matrix, the only information retained about the variable is contained in its correlations with the other variables. Thus, the end product of a dimensional analysis based on the correlation matrix is a set of basic dimensions within which are preserved the patterns of intercorrelations of each variable with all of the other variables. The variables themselves are not represented, only the portions of the variables having mutually shared intercorrelations are represented.

There are numerous techniques for the calculation of the basic dimensions from a correlation matrix, of which we distinguish two general classes in this paper. The first class is a form of solution of simultaneous equations in which all variables and all basic dimensions are simultaneously involved. This type of technique is generally based on a mathematical model for optimization of certain properties, usually a minimum variance (least squares, etc.) criterion. The specific criteria of optimization vary in the different methods. The result of the process is, however, always an expression of the observed variables as a set of coordinate values on a set of independent (orthogonal) basic dimensions. The basic dimensions are themselves derivative from the intercorrelations of the observed variables in such a way that the coordinate values of an observed variable with the basic dimensions are correlations of the observed variables (taken separately) and a linear composite of all of the observed variables. The primary points of the distinction of these methods and the methods described next are the simultaneous involvement of all observed variables and the optimization criteria one chooses to satisfy in the analysis. This class of techniques is usually called factor analysis.

The second class of calculational techniques to find basic dimensions are those of successive selection of subsets of relatively homogeneous variables in order successively to define the dimensions. The techniques vary within calculational procedures and in the means of identifying the subsets of variables. The technique of key-cluster analysis of the BC TRY System is to select the most general single variable in the entire set of variables and form a cluster about that "pivot" variable by selecting a subset of mutually collinear variables that are collinear with the pivot variable. After each cluster is discovered, the basic dimension corresponding to the cluster is defined as a linear composite (with equal weights) of the variables in the cluster. In this form of cluster analysis, as in the first class of dimensional analysis methods, it is desirable that the basic dimensions be independent. This is insured by the "extraction" of the successive dimensions as they are calculated from the respective clusters. Each successive cluster is based on the intercorrelations of the variables corrected (reduced) in such a way that previous cluster defined dimensions are not represented in the correlation matrix. It may happen that the clusters in the data are not independent, even though the calculation techniques force the basic dimensions to be independent. The orthogonality of the basic dimensions is an artifact of the techniques used and is not a necessary feature of actual data.

The optimization criteria in cluster analysis are not as simple as the optimization criteria of the least squares or minimum variance. Also, the criteria generally are not as clearly satisfied by the techniques used to find basic sets of dimensions by cluster analysis. The optimization criteria most desirable are: a minimum number of clusters (and corresponding basic dimensions) to account for the intercorrelations among the observed variables, which clusters are most nearly independent among all such minimal sets. There is no clear-cut mathematical solution to the cluster selection problem that satisfies these criteria. However, the method of key-cluster analysis, or collinearity cluster analysis, in the BC TRY System employs reasonable ad hoc techniques that have appeared to be satisfactory in numerous applications. The method of key-cluster analysis is described in general terms in the following section.

The results of a factor analysis are at times difficult to interpret. One reason for this difficulty is that the optimization criteria utilized tend to equalize the projections of the variables on the basic dimensions. In order to more closely associate certain of the observed variables with basic dimensions, the basic dimensions are "relocated" by a transformation of the basis dimensions, a "rotation" of the axes of the basis space. The relationships among the variables are not modified, and the relationships among the basis dimensions are not modified, but the relationships between the basis dimensions and the observed variables are modified. The specific form of the transformation varies with the rotational mode. This procedure is not needed in cluster analysis because the basic dimensions are defined by clusters of observed variables in the first place. This is one of the advantages of cluster analysis -- attention from the first is centered on homogeneous subsets of variables and not on hard-to-define composites of all the variables for all dimensions.

Cluster selection. A variety of techniques for the selection of clusters have been proposed. In this paper, the problem is simplified by restricting our attention to a single type of data. We assume that the data are representable in the form of an Euclidean vector space with unit radius (the basic assumption of all factor analysis). If the data are in the form of a correlation matrix (or if a correlation matrix can be calculated reasonably from the data), vector projections, etc., the data satisfy this requirement. Within this restriction, there are two types of clusters: collinearity clusters and proximity clusters. In collinearity clusters the variables of a cluster are all relatively homogeneous with regard to their location in the vector space in terms of direction of the vectors, relative to the origin of the vector space. That is, the vectors of collinearity clusters form a tight bundle of vectors with the ideal cluster being a set of vectors lying precisely on a single line and varying only in length. In proximity clusters the vector length and vector collinearity are not directly relevant. The only relevant condition is that of "nearness" or proximity defined by small Euclidean distances between the vector points (variables). When the vector lengths are close to unity in a proximity cluster it follows that the variables are also relatively collinear. However, when vector lengths are very short, i.e., near the origin, the vectors are all separated by small distances but may be at right angles to each other.

The specific technology of selecting clusters depends on certain features of the data and on the desired results. In V-analysis the minimal number of most independent clusters are selected. In key-cluster analysis, clusters are selected in a sequence, each cluster-defined dimension being "removed" from the data as the cluster is selected: the selection of a given cluster is based on correlations after the influence of all previous clusters has been removed. Two steps are involved: selection of a key or pivot variable, and selection of the remainder of the variables for the cluster. It is desirable to select clusters that have large amounts of general variance and that are composed of variables that are mutually collinear as well as collinear with the pivot variable. In key-cluster analysis, the pivot variable selection procedure insures that the successive clusters are as general as possible given the condition of the correlation matrix. Once the pivot variable has been selected a second variable is added; that variable most collinear with the pivot variable. Successively added variables are those with the highest mean collinearity with the previously selected variables. Overall cluster homogeneity is insured by setting lower limits on the mean collinearity acceptable for a cluster definer. The successive reduction of the values in the correlation matrix as each cluster is defined ensures that the dimensions are independent, although the clusters themselves might not be mutually independent.

Proximity cluster analysis techniques reduce to selecting clusters having small intra-cluster distances. One way to do this is to begin initially with a factor analysis of the correlation matrix to reduce the variables to points in a basic vector space. The distances between variables can then be calculated and a cluster analysis performed from the distances. Techniques that have been developed for this type of analysis depend on selecting a given variable as the first variable of a cluster, adding as the second variable that variable with the smallest distance from the first, etc. In some techniques, distances are averaged as variables are added and in others only the latest variable added is important in seeking the next variable to be added. In the BC TRY System, proximity cluster analysis is performed by a pattern-recognition technique. If two variables are in the same general region of the vector space, they have highly similar patterns of projections on the arbitrary basis dimensions. Also, if the vector space is subdivided into segments depending on relative values on all of the basis dimensions, any two variables falling into one segment will have similar patterns of projections. For example, two variables with the pattern of projections high, high, low, and middle on the four basic dimensions will be located in about the same place in the space, depending on the width of the categories high, low, and middle. Once the categories are defined, a pattern for each variable is easily established and proximity clustering is simply a sorting of the patterns to discover collections of variables with the same patterns. More details on this method are given below as part of the discussion of cluster selection in O-analysis.

<u>Statistical properties of clusters</u>. In V-analysis one of the primary

goals is to express the variables as linear composites of a set of basis dimensions. In factor analysis and in key-cluster analysis of the BC TRY System the dimensions are independent (orthogonal). However, in cluster analysis, the development of the basis dimensions is accompanied by the identification of clusters of variables. The clusters have a substantive existence independent of the orthogonal dimensions. Generally, the clusters are not orthogonal. Also, clusters have a number of interesting statistical properties.

The degree to which each cluster, separately, accounts for the correlation matrix can be calculated. Such measures of saliency indicate the relative importance of the clusters in determining individual differences in the sample. Two such measures are routinely calculated in the Cluster Structure Analysis (CSA) program of the BC TRY System: one based on the actual original correlations accounted for by each cluster, and the other based on a measure of the generality (the communality) of the variables accounted for by each cluster.

General psychometric techniques lead to the definition of the reliability of a cluster composite. Low reliability implies instability of the variation of individual differences in the observed variables. One does not wish to make general interpretations of hypothetical domains of variation that are not reliably measured. Nor does one wish to base an O-analysis on scores for clusters that are not reliable. CSA calculates the reliabilities of each cluster. In addition, CSA seeks out the most likely additions of variables to each cluster and recalculates the reliabilities under the assumption that each of these additional variables is added singly to the cluster and then, sequentially in order of the magnitude of their individual effects on reliability.

Because the clusters themselves are not usually independent, the intercorrelations of clusters are of interest. These correlations are calculated in two ways in the BC TRY component CSA, first for the simplest estimate of the cluster (the simple composite) of the cluster variables (the cluster score), and second for the composites corrected for their unreliability. Hypothetical clusters, perfectly reliable and collinear with the centroids of the observed clusters, are called "domains" in the BC TRY System. Correlations of the cluster composites and the estimated domains indicate the validity with which cluster composites estimate the general domains. The correlations of each of the individual variables in a study with the dimensions defined by these domains is the analog to the matrix of factor coefficients in a factor or cluster analysis, excepting that they are not independent.

Geometric properties of clusters. The basic assumptions in multivariate analysis imply that the data can be displayed geometrically. The correlations of the variables with the basis dimensions provide all that is necessary to graph the geometric properties of the vector space. The BC TRY System program SPAN, for SPherical ANalysis, plots the variables in subsets of three dimensions at a time from the full collection of basis dimensions. The plot is two-dimensional. However, because of certain properties of the projection of a three-dimensional space into a two-dimensional plane, the full three-dimensional picture can be "visualized" from this plot with a little practice. In this way, the actual spatial relationships of variables in clusters and between clusters become intuitively understandable.

Describing Individual Objects in Terms of Clusters

Having identified factors or clusters, one wishes to describe in terms of the factors or clusters each of the individual objects observed in the sample. If we have an object at hand, what would it score on the clusters or factors? The solution to this problem in factor analysis (factor estimation) traditionally involves solution of systems of simultaneous linear equations. The simpler procedures of cluster scoring produce results correlated nearly 1.0 with the factor estimation results whenever the clusters are reliable and highly collinear. A cluster score is simply the sum of standard scores of the object on the variables in the cluster. For convenience, the composite is usually re-scaled to some conventional mean and standard deviation (the same practice is used in factor estimation). If we are dealing with subjects in the sample from which the correlation matrix was obtained, the basic data are already available in the computer at the end of the process of V-analysis. If a new object is to be scored on the clusters from a previous study, all that is needed is the object's scores on the cluster variables, the mean and standard deviation from the sample used in the cluster analysis, the standard deviation of the sum of standard scores in the cluster, and the desired mean and standard deviation for the cluster scores. Everything is provided by the calculations in the original analysis, excepting the new object's scores.

Cluster Analysis of Objects

The primary motivation for V-analysis is the development of basic dimensions with which to describe the general sources of individual differences in the objects one observes. The results of V-analysis may suggest the meaning of the basic dimensions. However, the utility of the dimensions of V-analysis, or their pertinence to fundamental phenomena, cannot be ascertained readily from the results of V-analysis alone. Rather, the manner in which the objects are segregated into clusters, and the manner in which these object clusters relate to other (criterion or dependent) variables, determines the significance of the results of the V-analysis. The evaluation of the utility of the cluster scores for O-analysis can be made partially from the results of the cluster analysis of objects. However, the substantive significance of V-analysis or O-analysis can be determined only on evidence that the O-analysis leads to an otherwise unachievable understanding of the individual differences of objects on some criterion or dependent variable. The techniques for making such an evaluation are described below as "Typological Prediction".

Cluster selection in O-analysis. The two types of O-analysis parallel the two types of V-analysis: proximity clustering, and collinearity clustering. In proximity clustering the goal is to find groups of objects having small inter-object distances, i.e., objects in the same neighborhood of the multivariate space defined by the cluster or factor scores. Numerous techniques exist for scanning a matrix of distances in order to determine the proximity clusters. One technique is to define an initial cluster as the two objects in the entire sample having the smallest distance. The third object added to it is that object with the smallest distance from either one of the first two objects or the average of the first two objects. Objects are successively added to the cluster until the distance of the next candidate for inclusion exceeds some criterion value. These methods vary in terms of the objects compared on each successive addition, the entire set of previously selected objects or the most recently added object, etc.

A pattern-recognition approach to proximity clustering is implemented in the BC TRY System. An initial set of neighborhoods or sectors of the score space are established under control of the user of the program OTYPE. The observed subjects are sorted into these neighborhoods, by matching the pattern of cluster scores defining the neighborhood with the pattern of scores of the observed subjects. In the OTYPE program only those neighborhoods with one or more objects are ever actually defined. Those sectors (neighborhoods) with multiple objects in them are retained and the objects in those sectors are called core object clusters. The centroid for each core object cluster is calculated and the clustering of the objects is discarded. A new clustering of objects is obtained by defining a cluster for each of the core centroids. The distances between an object and each centroid are calculated. The smallest distance determines the new assignment. Each object is assigned to the core object centroid from which it has the smallest distance even though the objects themselves are shifted from one core object cluster to another in the process. Two clusters defined in this way that are separated by a distance smaller than a user-specified criterion are merged (condensed) into a single cluster, and the process is continued with the reduced number of object core clusters. A core cluster is discarded if it is robbed of all its members by their reassignment to other core clusters. When the memberships of the object clusters do not change in any reassignment cycle, the process is said to have converged, and the final clusters are called O-types in the BC TRY System. Because of the built-in requirement that objects be assigned to clusters from which they have smallest distances, and a rejection of objects when the smallest distances are large, the final clusters are well-defined proximity clusters, residing in neighborhoods with boundary limits fixed by parameters set by the user of the program.

In proximity clustering the number of objects involved must be relatively large in order to use the pattern identification procedures for the OTYPE program, particularly when the number of basic dimensions is large. The number of sectors is a geometric function of the number of dimensions. Consequently, for a large number of dimensions the number of sectors is large. For a given number of sectors, the probability that two or more objects will be in the same sector is a direct function of the number of subjects -- the more subjects the larger the probability that two or more subjects will fall into a given sector. For small numbers of subjects and a large number of dimensions, the OTYPE program sometimes fails, and a technique called EUCO analysis in the BC TRY System is used. In EUCO analysis the pattern of distances of a pair of objects compared with the entire remainder of the sample is taken as an indication of neighborhood similarity. The distances of all pairs of objects are intercorrelated and a standard key-cluster analysis or factor analysis is performed, identifying clusters of objects with similar patterns of distances in the distance matrix.

The second major type of O-analysis is collinearity cluster analysis. In this type of analysis, the distance between objects is less a matter of concern than is the proportionality of score profiles without regard to the elevation (overall magnitude) of the scores in the profiles. In the psychometic literature this type of analysis is sometimes called Q-analysis. One approach is simply to take the transpose of the original score matrix (reversing the role of variables and objects) suitably standardized, and perform an ordinary V-analysis on the correlation matrix resulting from correlating the columns of the transposed matrix. This technique includes all of the specific variation in the scores that are used to determine the inter-object correlations, a feature that is perhaps objectionable. On the other hand, if the score matrix used is the cluster or factor score matrix instead of the original scores, only the general sources of variation among the subjects are retained and the analysis is free of influences of specific sources of variation in the inter-object correlation matrix. A more interesting approach is to treat the cluster or factor scores as vectors in score space and to calculate the inter-object cosines directly rather than the inter-object correlations. This approach is based on an assumption that the objects are perfectly (no error variation) represented by the cluster scores. The cosines are of course direct measures of the degree of collinearity of two sets of scores, without respect to the score magnitudes. A cluster analysis of this matrix of cosines by V-analysis techniques produces the most independent set of object clusters. Other object clusters (falling between or among the most independent set of clusters) can be detected by inspecting a geometric representation of the objects by procedures like those of the SPAN program of the BC TRY System.

Homogeneity analysis of O-types. Once a collection of object clusters is identified, it is important to study the homogeneity of scores within the object clusters. A score dimension does not differentiate the O-types if the O-types are as heterogeneous as the sample of objects at large on a score dimension. For proximity clusters, the scores of the objects in a cluster should be approximately all the same value. Any variation from object to object within proximity clusters should be random variation and the inter-correlations between the cluster scores within the O-type should be essentially zero. For collinearity clusters the scores for any two objects, across all the basic dimensions, should be proportional. As a consequence, the inter-correlations of the scores within a collinearity O-type should all approach 1.0 if the clusters are well defined. The actual values in a collinearity cluster on a given basic dimension may have a wide range. Score dimensions defined by clusters for which the cluster reliability is not satisfactory will tend to have lower degrees of cluster homogeneity than clusters with high reliability.

An additional homogeneity analysis concept is useful in proximity clusters. Each of the basis dimensions influences the assortment of the observed objects into the O-type clusters. However, the degree to which the various dimensions are involved may vary. If the O-types are conceptualized as a set of categories in a generalized regression problem, and if the cluster score dimension is taken as the dependent variable in the regression, then the total variance of the cluster score can be partitioned into a part associated with variation within object clusters and a part associated with differences in mean values of object clusters. The ratio of the variance component associated with the differences in mean values to the total variance is the correlation ratio, eta. Alternatively, this correlation ratio can be defined in terms of the weighted sum of the homogeneities of the O-type clusters. The statistic eta is an expression of the general, non-linear, regression of the cluster score dimension on the O-type classification. If the value of eta is near 1.0, and if one knows the O-type to which an object belongs, the object's score on the dimension can be stated with a high degree of precision simply by stating the mean value of scores within the O-type on that dimension. Conversely, the mean homogeneity for each O-type, averaging across all the basis score dimensions, indicates the average degree that membership in an O-type specifies ranges of scores over all cluster score dimensions.

Displaying O-types in vector spaces. It is useful to map the O-types into the vector space of scores. In order to do this, several idealized objects are "invented" to represent the O-types and the score dimensions. The idealized objects representing the O-types are invented simply by finding the mean score of the objects in each of the O-types. The mean score profile for an O-type is treated as an object. In order to represent a score dimension several idealized score patterns are created, such that they vary only on that score dimension and have the mean value on all the other dimensions. Thus, we invent objects each having mean values on the second, third, etc., dimensions and systematically varying on the first dimension, perhaps in steps of one-half standard deviations. If there are three dimensions with means of 50 and standard deviations of 10, some of the idealized objects representing the first dimension would be (35, 50, 50), (40, 50, 50), (45, 50, 50), (55, 50, 50), (60, 50, 50), and (65, 50, 50). Each dimension is thus represented by a set of idealized objects. The sets of idealized objects representing the dimensions and the O-types are combined into a single "sample" of objects and an O-analysis is performed using the methods of the BC TRY program EUCO. This program converts the inter-object distance matrix into a correlation matrix and a key-cluster analysis is performed. Following the cluster analysis, the geometric structure of the O-types is displayed by the print-out of the program SPAN, showing both the O-types as points and the idealized objects representing the score dimensions as points (nicely lined up showing the location of the dimensions in the O-type space). An interesting application of this technique is to introduce into the analysis ideal profile types derived from theory. For example, a theory of personality will perhaps specify profiles for critical personality types or a product field will have ideal customer types specifiable as a profile of scores.

Hierarchical analysis. In many object analyses the number of object clusters may be too large for the number of distinctions that one wants to make in an application of the results of the analysis. Or, the specificity of the aggregation into O-types may be too great. In these circumstances one wishes to condense various O-types into combined groups. The condensation most reasonable is that involving the most similar O-types. An obvious definition of similarity in this context is the distance separating the O-types in the score space. These distances are easily calculated from the mean scores of the O-types. The number of O-types can be reduced by combining the two O-types with the smallest separation (smallest distance). The resultant, more general, O-type replaces the two O-types combined. This process is repeated, combining the two O-types with the least separation, until all of the O-types are finally pooled into one group composed of all of the objects in O-types. The successive stages of reduction are hierarchical. The hierarchy can be described as a branching process like a geneological tree. At the top are the specific O-types, at the bottom is the entire set of objects, and in the middle are intermediate stages of the combination of O-types. Inspection of the mean score profiles of the various combinations of O-types developed in the hierarchy can guide the researcher in selecting a point in the hierarchical condensation for a final set of O-types. A graph of the hierarchical tree is easily developed from the distances between the O-types as they are condensed in the hierarchy.

Typological Prediction

Regression techniques in prediction are well known. One or more dependent variables are of importance in an application but difficult or impossible to observe in the actual application situation. One or more other variables that are easily observable prior to actually engaging in the application are correlated with the dependent variables. A preliminary, or standardization study is performed in order to determine the regression of the dependent variable on the independent variables. The regression equation gives estimates of the dependent variable from knowledge of the independent variables. Where there is a single independent variable the prediction is termed "univariate." Where there are two or more independent variables the techniques of multiple correlation and multiple regression are used, and the prediction is "multivariate." In the social and behavioral sciences, univariate prediction is considered reasonable if 25% of the variance of the dependent variable can be predicted. In multivariate prediction the prediction is generally better, on the order of 70% in successful applications. Both of these techniques are used in the BC TRY System to study prediction from cluster scores. However, a third technique offers more specificity and discriminative accuracy: O-type, or typological prediction. If an object is identified by O-type membership the best estimate of its score profile is the mean score profile for the objects in its O-type. Expanding this to a dependent variable, the best estimate of the value that an object has on a dependent variable is the mean value of the dependent variable for the O-type in which the object is a member. Should an object, not included in the standardization sample, have cluster scores matching the mean scores of a given O-type, the best estimate of that object's dependent variable is the mean value of the dependent variable in the given O-type. This estimate is accurate insofar as the variability of the dependent variable within the O-type is smaller than the variability in the object population. There are several advantages to this method of prediction. First, the homogeneity for certain of the O-types may approach 1.00, corresponding to zero error of prediction for the objects with that O-type score profile. Second, the O-types with less homogeneous scores on the dependent variable are explicitly identified. Prediction for objects with those O-type score profiles can be viewed with greater caution. The regression model for prediction gives a mean accuracy of prediction based on the assumption that the variances in the dependent variable are equally homogeneous over the entire domain of the independent variables. In typological prediction one takes advantage of greater homogeneity in certain segments of the population and is alert to problems with smaller homogeneities in other segments of the population. A third, but less critical, advantage is that only the variables in clusters must be observed to make the prediction. Goodness of fit of an object to the O-types is directly observable from these cluster scores. The fourth advantage, and perhaps the most important, is that no particular mathematical model is required for this type of prediction. The prediction procedure is not based on a linear regression, a quadratic, or any other simplistic form. Rather, the configuration of cluster score patterns is directly the determiner of the prediction without benefit of, or obfuscation by, a specific mathematical model.

The component 4CAST is a useful feature of the BC TRY System for testing the efficacy of prediction without resorting to questionable statistical assumptions. 4CAST is a sampling program that develops the sampling distribution of a variable for samples the size of the various O-types in an O-analysis. For an O-type of a given size, the program draws random samples repeatedly from the full sample and plots the distribution of mean values and homogeneity values for the many samples, comparing the mean and homogeneity on the observed O-type with the distribution resulting from the repeated sampling. The relative frequency of means that are smaller, or larger, than the observed O-type mean can be interpreted as the significance probability of the departure of the O-type mean from the mean of the full sample. Thus, an O-type with a score average approximately equal to the average in the full sample, but with a significantly greater degree of homogeneity, might be as important a finding as one with a mean significantly departing from the full sample mean. The techniques used in 4CAST are modern Monte Carlo techniques, making no assumptions regarding the "underlying" distribution of the score dimensions.

The BC TRY System of Computer Programs

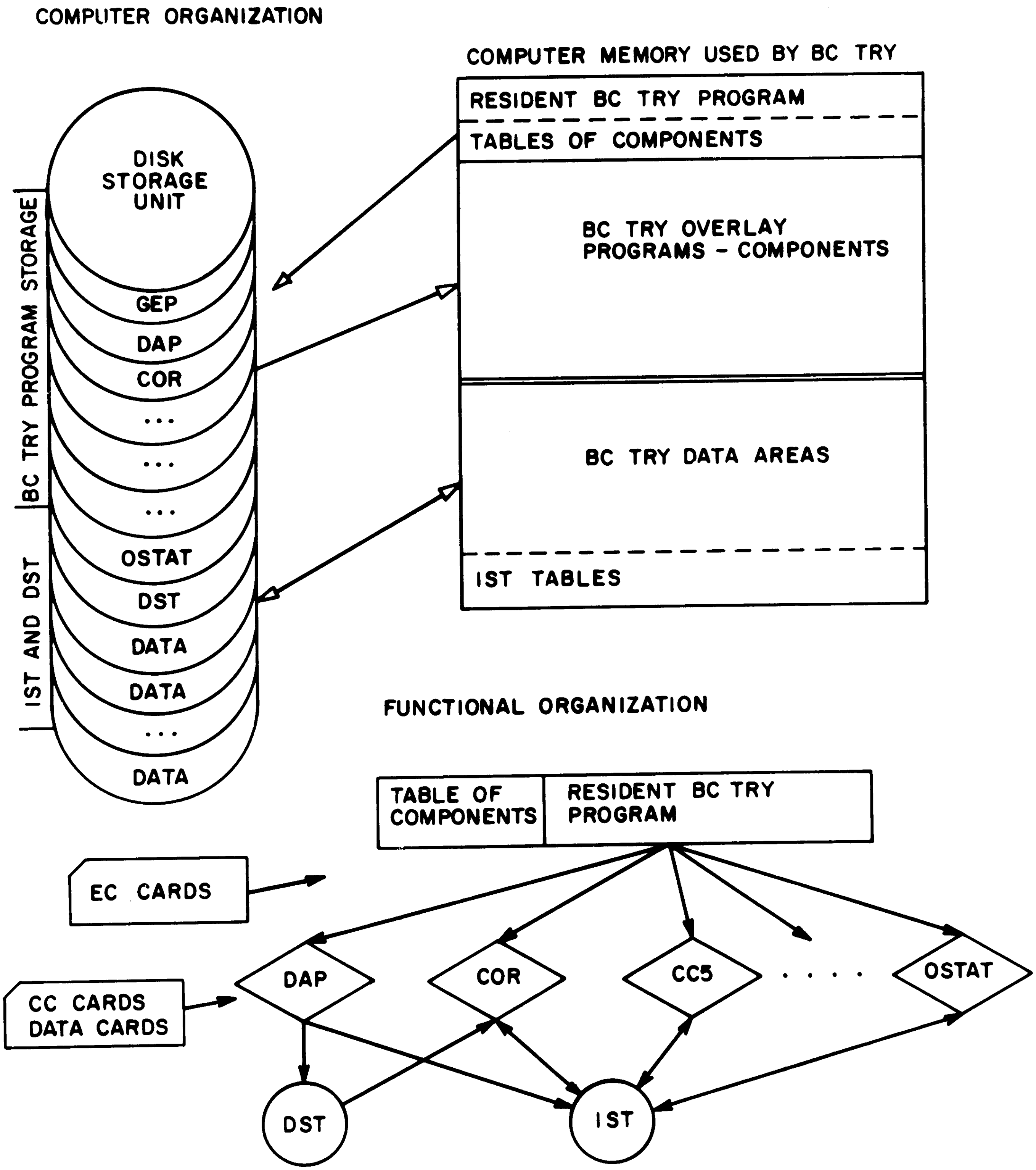
The BC TRY System of computer programs is designed to perform all of the types of analyses discussed above, and a good many more. The modularity and flexibility of the System makes innovative use of the multitudinous options quite easy. The System has been used in a wide variety of substantive problems, encompassing studies of attitudes, chemical engineering, genetics, hydrology, zoology, glaciology, medicine, soil morphology, marketing, job specifications, education, social structure, physical anthropology, economics, aptitudes, etc.

This section of this paper is a discussion of the basic design of the System, the executive operation of the System, the data sharing facilities of the System and the primary statistical programs of the System. The discussion is not technical. Rather, this section is a general exposition of the System without development of the technical computer aspects of the System. Figure 2 is a schematic representation of the computer and operational structures of the System.

Basic design characteristics. The BC TRY System was designed and implemented with three primary criteria -- general utility in multivariate analysis (both V-analysis and O-analysis), ease of use, and accuracy. These criteria are satisfied in a number of ways, some of which are explicitly covered here.

The System is modular. The basic building blocks of multivariate analysis are incorporated into the System in independently callable segments. The independence of the segments is limited only to the input requirements of the various types of analyses. However, there are numerous paths for the input of data for each segment of the System, matching the configuration of almost any desired data analysis. Indeed, the data sharing capability of the System is one of the unique elements of the System among computer program packages. Input and output of data between stages of an analysis are accomplished easily, providing the user with a way of introducing new data at intermediate stages and a way of outputting (cn punch cards) results of intermediate calculations of an analysis. The modularity of the System is reflected in the manner in which the various stages of analysis are initiated and controlled. The basic building blocks of the System consist of primary

Figure 2 Structure of the BC TRY System of Computer Programs



components that are executed by use of an executive control, EC, command or "verb" of the System. These verbs play the role of identifying the primary type of operation to be performed. The modifiers of the verbs are component control, CC, commands, in the form of punched cards containing option control parameters. The EC commands are of two sorts, data sharing commands and commands that control statistical, analytical and logical problems. A run of the BC TRY System consists of a sequence of EC commands (each on a separate punch card), CC command cards, and data cards. There are 31 standard EC commands in the System, most of which have a large number of modifiers in the form of CC commands. For example, in V-analysis there are eight primary executive control commands: CC5 for key-cluster analysis, DVP for determination of initial diagonal values in the correlation matrix, FALS for least squares methods of factor analysis, GYRO for rotational methods, etc. In the FALS EC command there are three primary modifiers, for the three methods of least squares factoring available: principal axes, augmented factoring, and canonical factor analysis. Each of the primary modifiers in FALS has two separate sets of parameters that determine such things as how the dimensionality of the solution is to be determined, and how convergence in factor iteration is to be determined. The range of possible analyses involving FALS is extensive when used in concert with other EC commands that provide a manipulation of the basic data that FALS addresses in its calculation. For example, different forms of factor analyses are given when the diagonal values program DVP, executed before FALS, is used with different modifiers providing various types of values. Repeated use of an EC command, preceded by various other EC commands, can be used to provide methodological studies in data analysis. A string of BC TRY EC and CC commands represents completely and accurately a complex and highly individualized multivariate procedure. Because of the richness of the EC and the CC command structure, and because of the data sharing aspects of the System, the user virtually interacts with his data analysis, exercising control over the detail and broad strategy of the analysis.

The flexibility of analysis using the BC TRY System is illustrated in the following example. Imagine data in which a general factor influences the intercorrelation of all of the observed variables (say an instrument factor, or a factor representing systematic context variation across the observational procedures). We wish to "extract" this variation as a general factor and then perform a cluster analysis on the resulting "corrected" correlation matrix. The corrected correlation matrix is of interest in itself and we want a card deck containing the matrix. The sequence of EC commands that would be used are as follows:

START — initializes the computer and data storage facilities

DAP — inputs the raw data, variable names, and object names and provides simple descriptive statistics for the raw data

DPRINT — lists the data as they actually appear to the programs

COR2 — calculates the correlation matrix from the raw data

GIVE produces a card deck containing the status of the analysis at this point

DVP provides diagonal values for the correlation matrix (communalities)

FALS under modifiers used for this analysis, this program calculates a single principal axis factor (the first Eigen vector)

FAST calculates the correlation matrix orthogonal to the first Eigen vector and replaces the original correlation matrix with this reduced matrix

GIST punches specified information; the corrected correlation matrix in this example

DVP calculates new estimates of communality to be used in the cluster analysis

CC5 performs a full key-cluster analysis

GIVE produces a card deck containing the status of the analysis at this point, including data necessary to re-start at a later time.

A simple cluster analysis of the correlation matrix, without the removal of the first principal axis would be performed by the following sequence of EC commands: START, DAP, DPRINT, COR2, GIVE, DVP, CC5. The GIVE command serves to punch the status of the data in the analysis at the point the command is used. The punch cards provided by GIVE can be read by using the command TAKE, re-establishing the status of the analysis as it was when the GIVE was used. Consequently, the following analysis picks up where the previous analysis has just completed calculating the correlation matrix: START, TAKE, DVP, CC5. Now, imagine that we wanted the first analysis described above, but we wanted in addition to have a cluster analysis of the matrix before the first principal axis is extracted, and that we did not want to have the correlation matrix punched out: START, TAKE, DVP, CC5, FALS, FAST, DVP, CC5, GIVE. To perform a varimax rotated principal axis analysis and a key-cluster analysis, using the communality estimates provided by the full principal axis analysis, the following sequence of EC commands would be used: START, TAKE, DVP, FALS, CC5. The print-outs of the procedures appear in the order of the use of the executive commands.

The System is truly eclectic in its selection of procedures implemented as standard procedures and options within the EC commands. In addition, the System contains a subsystem of 35 matrix and vector operators and special control commands to manipulate the data sharing facilities of the BC TRY System. With this special subsystem, any calculation that can be expressed in terms of vector, matrix, scalar, and elemental algebraic operations can be incorporated into the BC TRY System without additional programming or modification of the System.

Use of the System is streamlined in standard analyses by the use of "default" options. In order to avoid requiring a user of the System to punch the CC modifier parameters in routine and standard use of the System, failure to specify parameters results in the substitution of the standard (default) parameters. A few constants associated with a data set are always needed, such as the number of variables, and the number of observations. However, a standard deck of commands, modifiers, and data, has few specially punched cards.

In using programs such as the BC TRY System, errors of certain kinds in data decks, machine failures, control card errors, etc. will cause abberations in an analysis. In the BC TRY System, most such abberations cause automatic punching of the status of the analysis on cards for later re-start of the analysis from the point just previous to the error point. This feature also is callable directly by the EC command GIVE.

The DPRINT EC command causes a print-out of the actual data that are input to the computer, in order to verify the correctness of the data as they are actually entered into the analysis. Variable format facilities of the System are easy to use, but the use of format statements is tricky at times and the user must be able to verify that his format and his data are actually in parallel before he can be sure that the analysis is performed on data that he intended. Use of DPRINT provides this capability as a standard feature.

The outputs of the component programs of the BC TRY System are labelled and annotated for easy interpretation. Together with the full documentation of use of the System, this output clarity makes the System easy to use. For standard uses of the System, an Abridged User's Manual provides a brief and concise guide. For the less ordinary uses, the full User's Manual provides a great deal more documentation on the use and interpretation of output of the System than is found in other packages of programs.

Meticulous care has been observed in the development of the System to control the accuracy of computation. Extensive test cases have been run with cross-checking and parallel hand calculation to insure accuracy of the results.

Executive and sub-monitor operations. Modern computers are operated under the control of a computer program called a monitor, an executive system, or an operating system. These programs coordinate the various languages and jobs executed by the computer, including assignment of auxiliary storage (peripheral) devices to a user and recovery from failures in a user's program. In the BC TRY System a central program called the General Executive Program (GEP) plays the role of a sub-monitor under the operating system of the computer. To the operating system of the computer GEP appears to be a simple user's program. However, to the component programs in the BC TRY System GEP serves the functions of an operating system. Only a few of the functions that GEP performs are mentioned here. GEP has as its primary function the selection of proper component programs from the System on signal (punched on cards) from user. When the order of cards is not appropriate, GEP generates messages to the user and performs certain recovery tasks before relinquishing control to the computer operating system. Second, GEP retains records of the data generated by the programs used in the sequence of a computer run on BC TRY. Improper sequencing of programs in BC TRY can result in information not being available to a program, and GEP signals this to the user and terminates the run with procedures making it possible for the user to recover from the point of the error. The recovery procedures that BC TRY initiates on discovery of a computer fault make it possible to re-initiate a run at a later time without beginning all over again. This and other supervisory functions of GEP makes for greater ease in use of the System and in more error-free operation of the System. Also, the design of GEP, although too technical for this discussion, makes the data sharing features and the modularity of the System possible.

Data sharing in the BC TRY System. A very large number of sets of intermediate statistics in the form of matrices, vectors and lists are generated in the process of performing a cluster analysis. Among these are the correlation matrix, the list of variable names, the list of object names, the diagonal values, the means and standard deviations of variables, lists of cluster memberships, the matrix of factor coefficients, the final residual matrix, the cluster scores, lists of O-type memberships, mean scores of O-types, etc. The intermediate data, as well as the original data, are needed by various components of the System in their functions. Because of the limited memory capacity of the computer it is not possible to store in memory all of the initial data, the intermediate statistics and the component programs of the System. Consequently, peripheral storage devices (magnetic tape or disk) attached to the computer are utilized. Effective use of these peripheral storage facilities is simple in programs smaller than the BC TRY System. However, since runs with the System may involve assmany as a dozen or more components, special provisions are made for communications between components in order to share data on the peripheral devices of the computer. Each vector, matrix, or list treated in this way is called a file. Each file has a name and certain parameters indicating its size and format.

The principal element in the data sharing system is a table of file names, file parameters, and the locations of the files on the peripheral computer devices. The table is maintained in memory by GEP and hence is available for each of the successively called component programs. As a component program needs data that are stored on the peripheral device, it interrogates the table to determine the location and parameters of the data file. As a component program generates data that are to be stored on the peripheral device, it makes entries in the file table and writes the data on the peripheral device. Since there is a record in the table of the information on the peripheral device, no direct communication between components is necessary but the components are able to share data throughout the System.

In addition to access to the data files by the statistical analysis component programs of the BC TRY System, a number of component programs are included in the System to provide the user with direct and convenient access to the data files. Original data from card decks are stored on the Data Storage Tape, DST, (which physically may be a portion of a magnetic disk) by the DAta Processor component DAP2. The intermediate and final results of the execution of component programs of the System result in the storage of data files on the Intermediate Storage Tape, IST, (which physically may be a portion of a magnetic disk). The component GIST (for Generate Intermediate Storage Tape), called by a EC card, is designed to permit a user to enter data files into the IST or to have data files from the IST punched onto cards (in keypunch form). The CC commands and data cards following the GIST EC command determine which files are affected. Thus, a data file that ordinarily is generated by a component program of BC TRY may be deposited on the IST by a user without executing the component program. For example, if a researcher has a correlation matrix but not the raw data he may enter the matrix onto the IST for analysis by the cluster or factor analysis programs. Also, if the researcher wants the matrix of factor coefficients punched on cards, GIST can be used to obtain the desired cards.

The component SMIS (Symbolic Matrix Interpretive System) also can read cards, read files from IST, and write files on the IST. The general matrix and vector calculation power of SMIS permits application of transformations to data read from cards or from IST files and subsequent storage of the transformed data on IST files accessable to other components in the System. For example, there is no provision for using the squared multiple correlation estimate of communalities in the diagonal values program of the System. However, a simple sequence of CC commands for SMIS finds the squared multiple correlations and stores them on the IST as the diagonal values file.

The entire contents of the IST can be punched out on cards (in compact binary form) by the user of the System by using the EC command GIVE. All of the files of the IST are thus preserved. This process is reversible by using the EC command TAKE, which reads a card deck produced by the EC command GIVE and restores the IST to the form that existed at the time the GIVE was executed. Thus, in one run on the System, at a stage where the user wants to be able to re-start the analysis in another run, a GIVE EC command would be included in the control deck. At a later date, the user would include the deck produced by the GIVE in his control and data deck, as the data for a TAKE EC command.

Statistical, analytical and logical programs. This section presents a brief description of the component programs of the BC TRY System that perform statistical, analytical, and logical processing of data in a cluster or factor analysis.

1) Data processor program, DAP. DAP prepares the data storage tape, DST, from a deck of cards containing the raw score matrix. This file is then used by other component programs that operate on the score matrix, such as the correlation program. DAP permits the user to identify variables and objects by names (if not specified, the program assigns appropriate names such as V001, V002, etc.), to reorder the variables and to reflect the variables. It incorporates an extensive system of error checking, accomodates missing data, and computes means and standard deviations.

2) Comment, COMMENT. The user may wish to insert comments in the printout before or after execution of certain components. Unlimited numbers of cards are read and immediately output by the COMMENT EC command.

3) Data verification, DPRINT. Data read by DAP onto the DST can be verified by printing them immediately after DAP and then proof-reading after the computer run. The DPRINT EC command accomplishes this.

4) Correlation, COR2. This component calculates the means, standard deviations and intercorrelations of variables for the data stored on the DST.

5) Correlation, COR3. For data with missing observations on some subjects the calculation of correlations is more complicated because of the

necessity of matching observations in order to calculate means, standard deviations, and cross-products for each pair of variables. The component COR3 performs this task.

6) Reorder and delete, REDE. This component prints the correlation matrix with specified rows and columns reordered or deleted.

7) List missing data statistics, RLIST. Although COR3 provides a printout of all values computed (correlation, matched sample size, matched means, matched standard deviations, etc.) they are printed matrix at a time. The component RLIST prints all such values together for each cell in the matrix.

8) Suppression of variables, SLEP1. This component suppresses variables in such a way that they can be reinstated at a later point in an analysis. The component SLEP2 reinstate the sleeper variables and projects them into the factor solution. This is desirable in two situations: where there are functionally dependent variables (linear dependencies) that would interfere with factoring; and where there are criterion variables that the user wants to deal with as a predicted variable and hence should not be a part of the factoring.

9) Reinstatement of sleeper variables, SLEP2. After factoring, SLEP2 is used to reactivate the suppressed variables and to calculate all factor statistics on all variables.

10) Diagonal values program, DVP. DVP computes or selects a set of values to be inserted as the diagonal elements of the correlation matrix. The values usually desired are the estimates of the communalities of the variables, although other types of values such as reliability coefficients or 1's can be substituted.

11) Key-cluster analysis, CC5. This general program of independent dimensional analysis (factoring) permits a selection of a subset of variables on each dimension, and a variety of other statistics related to factoring. The analysis starts with predetermined diagonal values in the correlation matrix, and re-computes diagonal values until convergence or for a specified number of times. In standard form, it selects the most independent sets of defining variables that are collinear, cumulates partial communalities as factoring proceeds, and terminates factoring on the dimension at which the sum of partial communalities matches the initial sum of communalities. The program is usually followed by the programs CSA and SPAN. Options on control cards permit a variety of other forms of multidimensional analysis: centroid factoring, square root factoring, bifactor analysis, and many forms of object cluster and inverse factor analysis.

12) Non-communality key-cluster analysis, NC2. The component NC2 parallels CC5 in most respects, with the exception that no communalities are involved in the factoring procedures and factoring is not reiterated.

13) Least squares factoring, FALS. This component performs four types of factor analysis: principal components, principal axes, canonical factoring, and augmented factoring. Accurate and efficient methods, based on the work of Housholder, Wilkenson and Ortega, are used in the matrix calculus of this component.

14) Residual and reproduced correlations, FAST. FAST is an independent component that provides printouts of the matrix of correlations precisely contained in the factor matrix or the matrix of residual correlations after the factors are removed from the original correlations. The matrix is substituted for the original correlation matrix on the IST by FAST for subsequent analyses.

15) Cluster structure analysis, CSA. This program provides a complete statistical description of the correlational properties of oblique cluster domains. The subset of variables that defines each cluster is usually selected by a key-cluster factoring component, but may also be derived on other grounds (e.g., theoretical) and input without the benefit of a cluster factoring. CSA determines the degree of generality of each oblique cluster dimension, ignoring the other clusters; provides specific data on how to increase the reliability of each cluster by including additional definers; and computes the mean intercorrelations of the definers of each cluster with other variables unifactorially allocated to it.

16) Non-communality cluster structure analysis, NCSA. This component parallels CSA with the exception that communalities are not defined in the analysis.

17) Sperhical analysis, SPAN. Graphical representations of a cluster or factor structure are produced by the component program SPAN. This component prints out graphs of the augmented structure, three axes at a time, with only the most salient three-spaces being printed.

18) Rotation of factor matrices, GYRO. This component provides alternative methods of rotating the dimensions derived by factoring. Both the quartimax and the varimax rotational models are available in the component.

19) Sampling and merging, BIGNV. Strictly speaking, BIGNV is not a component program, but involves several component programs of BC TRY (SAMPLER, PICK, PRY, MERGER). The component permits a user of the System to perform a V-analysis on very large numbers of variables by means of a sampling procedure and to obtain a condensed result by merging the analyses of the samples.

20) Factor and cluster scoring, FACS. The component FACS permits forming linear composites of the observed variables under several different conditions. The composites may be simple cluster composites, the sums of standardized scores of variables in a cluster, for each cluster; factor estimates by a full orthogonal regression model, or by a limited orthogonal regression model; factor estimates by an oblique regression model; user determined linear transformations with user supplied weights. This program is capable of scoring subjects with missing data.

21) Scatter diagram, RSCAT. RSCAT is a general program that can be used to compute all (or a selected subset of) the scatter diagrams within blocks of variables. Each scatter diagram provides not only the plot, but also the Pearson product-moment correlation coefficient, the curvilinear correlation coefficient for both regressions, marginal frequencies, percentile ranks, standard scores, means, standard deviations, and other statistical properties of the bivariate distributions.

22) Collinearity O-analysis, EUCO. This program sets up the data

necessary to perform a collinearity cluster analysis of objects. Input to the program is the matrix of cluster or factor scores for a sample of objects. Output is a distance matrix that is treated as raw data for a collinearity analysis; the distances are correlated by the component COR2 and the techniques of V-analysis are applied to the resultant correlation matrix.

23) Proximity O-analysis, OTYPE. The component OTYPE uses pattern recognition techniques to select object clusters, iterates cluster condensation and reassignments of objects to convergence on O-types, and performs a hierarchical condensation analysis of the O-types.

24) O-type statistics, OSTAT. This program provides a re-printing of the O-type membership lists in a form more convenient for some uses, calculates the O-type means, standard deviations, and homogeneities for each O-type on each cluster dimension.

25) Comparative analysis data, COMP1. Comparative analysis of the results of several cluster or factor analyses can be performed to make higher order analyses of the factors and clusters. The factor matrices for each of the individual cluster or factor analyses are combined, inter-factor collinearities are computed and the matrix of collinearities is factored. COMP1 causes the factor matrix for a given analysis to be punched on cards appropriate for COMP2.

26) Comparative analysis collinearities, COMP2. The component program COMP2 reads the card decks produced in the analyses using COMP1. Several methods of calculating collinearities are available in COMP2. The matrix of similarities produced by COMP2 is stored on IST as a correlation matrix, available for cluster or factor analysis.

27) Comparative similarity rotation, SIMRO by way of SYDA and SYRN. The combination of SYDA and SYRN work together much like COMP1 and COMP2. SYDA outputs card decks that are used for a comparative analysis using SYRN. Each analysis to be included in a similarity or comparative analysis of factors includes an EC command for SYDA. The decks output by SYDA are combined and read by SYRN, which rotates the factor structures of the various analyses into a structure of highest possible fit (least squares criterion) for the entire set of factor structures. SYRN provides a measurement of the goodness of fit of the rotated factor matrix of each individual analysis with that of the population matrix determined by SYRN.

28) Typological prediction, 4CAST. The component program 4CAST is used to discover the degree to which an outside criterion variable, not a dimension definer in a cluster analysis, can be predicted by the O-type structures. It consists of a Monte Carlo sampling procedure that simulates a random assignment of objects to clusters of the same size as those discovered by OTYPE or EUCO analysis. The mean, standard deviations and homogeneities of the simulations are accumulated into sampling distributions that are used to make statistical significance judgements about the actually observed means, homogeneities, and standard deviations of the O-types.

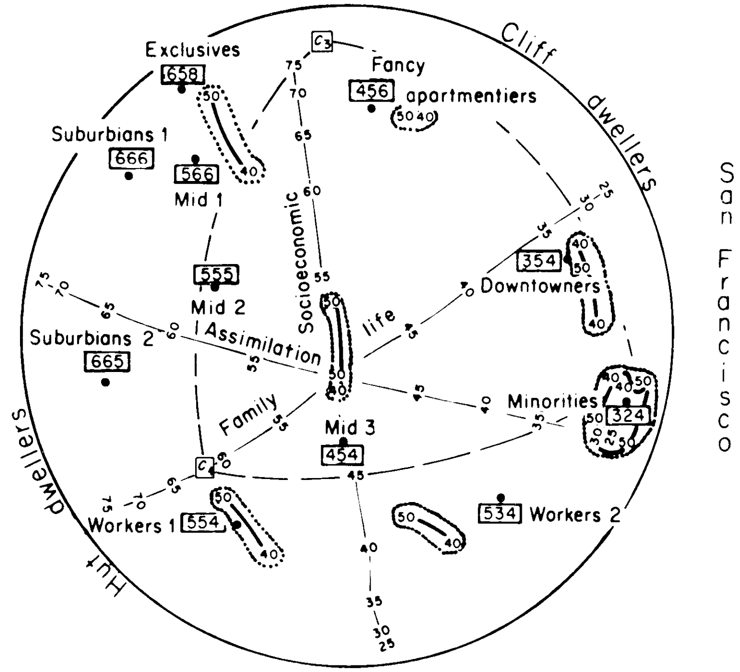
29) Symbolic matrix interpretive system, SMIS. This component performs general matrix algebraic operations and provides the user with almost unlimited capacity for matrix and vector operations with access to the IST. Each operation defined under SMIS is initiated by control cards. By combining

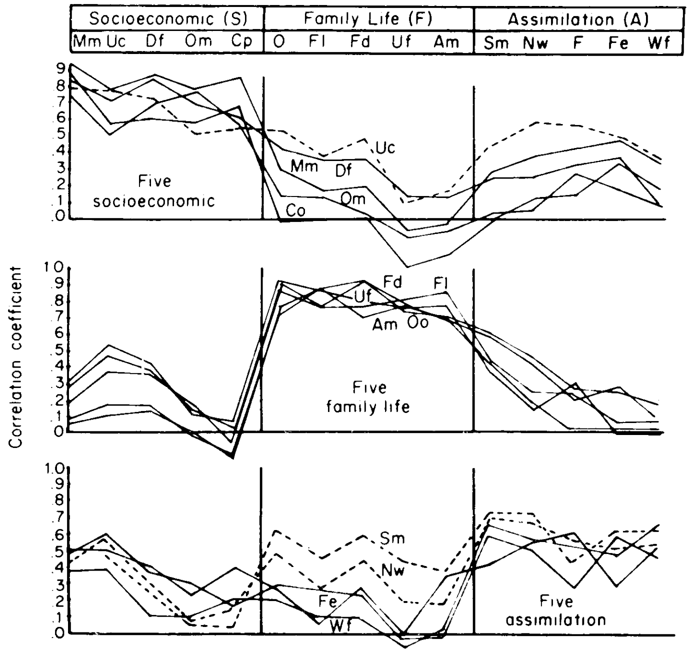
the operations provided in SMIS, the user can perform virtually any calculation desired even though the calculation is not an integral part of the other components of the BC TRY System.

Implementation of the BC TRY System

The BC TRY System initially was written for the IBM 7090 under the Fortran Monitor System (FMS). Subsequently, it has been converted to operate on the CDC 6400 under the SCOPE operating system. Conversion to other computer systems is anticipated.

CLUSTER ANALYSIS
and the
BC TRY SYSTEM





TBA

Tryon-Bailey Associates, Inc.

728 10th St., Boulder, Colorado 80302

TRYON-BAILEY ASSOCIATES, INC.

SPECIALIZING IN:

:::Systems analysis

:::Software development

:::Computer programming

:::Statistical analysis

:::Data analysis problems of any magnitude

:::Research design in social and behavioral science applications

Our experience extends into a number of fields

Applied Social Research

- Personality
- Social Area Analysis
- Voter Characteristics
- Population Analysis
- Attitudes
- Demography
- Epidemiology

Personnel Management

- Aptitude Structures
- Job Satisfaction
- Job Classification

Marketing Research

- Product Appeal
- Consumer Characteristics
- Population Segmentation

Operations Research

- Data Base Analysis
- Information Storage and Retrieval
- Systems Analysis

Custom Programming

- Statistical Analysis of All Kinds
- Urban Modeling and Data Analysis
- Multi-user Information and Record Sharing Systems
- Interactive Programming Language

Educational Systems

- Management Information Systems
- Modeling Educational Systems

THE BC TRY SYSTEM IS OFFERED EXCLUSIVELY BY TRYON-BAILEY ASSOCIATES, INC.

For prompt, efficient, imaginative, and accurate service in your data analysis problems, systems analysis, programming development, - - -

write to: Tryon-Bailey Associates, Inc.
728 10th Street
Boulder, Colorado 80302