DOCUMENT RESUME

ED 046 970 TM 000 288

AUTHOR Dubin, Robert; Champoux, Joseph E.

TITLE Typology of Empirical Attributes: Dissimilarity Linkage Analysis (DLA).

INSTITUTION California Univ., Irvine.

SPONS AGENCY Office of Naval Research, Washington, D.C. Personnel and Training Research Programs Office.

REPORT NO TR-3

PUB DATE Jun 70

NOTE 34p.

EDRS PRICE EDRS Price MF-$0.65 HC-$3.29

DESCRIPTORS *Classification, *Cluster Analysis, *Cluster Grouping, Matched Groups, *Models, *Statistical Analysis, Taxonomy, *Typology

IDENTIFIERS *Linkage Analysis

ABSTRACT

Dissimilarity Linkage Analysis (DLA) is an extremely simple procedure for developing a typology from empirical attributes that permits the clustering of entities. First the procedure develops a taxonomy of types from empirical attributes possessed by entities in the sample. Second, the procedure assigns entities to one, and only one, type in the taxonomy. This two-step procedure clearly contrasts with many existing clustering techniques that are concerned only with the second step of this two-stage procedure. To develop a taxonomy of attribute types, the method searches for attributes that go together. A statistical test of association is first used to identify all pairs of attributes whose empirical values are significantly associated. Attribute pairs are then linked together to form serpentine clusters, each of which represents an attribute type. The attributes defining each type are not similar. In fact, the method specifically avoids using any criterion of similarity when developing the types. Each entity is then assigned to the type it most closely resembles. An entity may unequivocally fit a type. Or, if an entry does not possess all of the characteristics of a type, it is assigned to the type with which its attribute values best match. Discrete clusters of entities, based on their attribute types, are thus formed. In short, this method moves from types defined by dissimilar attributes, to clusters of similar entities in each type of the taxonomy. (Author)

ED0 46970

U.S. DEPARTMENT OF HEALTH, EDUCATION
& WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED
EXACTLY AS RECEIVED FROM THE PERSON OR
ORGANIZATION ORIGINATING IT. POINTS OF
VIEW OR OPINIONS STATED DO NOT NECES-
SARILY REPRESENT OFFICIAL OFFICE OF EDU-
CATION POSITION OR POLICY

TYPOLOGY OF EMPIRICAL ATTRIBUTES: DISSIMILARITY LINKAGE ANALYSIS (DLA)

ROBERT DUBIN
JOSEPH E. CHAMPOUX

University of California, Irvine

Technical Report 3

June, 1970

INDIVIDUAL-ORGANIZATIONAL LINKAGES

Project Directors

Robert Dubin
Lyman W. Porter

University of California
Irvine, California 92664

Prepared under ONR Contract N00014-69-A-0200-9001

NR Number 151-315

*Distribution of this document is unlimited.
Reproduction in whole or in part is permitted
for any purpose of the United States Government.*

TM 000 288

TYPOLOGY OF EMPIRICAL ATTRIBUTES: DISSIMILARITY LINKAGE ANALYSIS (DLA)

Robert Dubin
and
Joseph E. Champoux

A ubiquitous problem of analysis is to establish categories and types, that taken together constitute a taxonomy of a domain of inquiry (Dubin, 1969; Weber, 1949). Two approaches exist for solving this problem: (1) a theoretical taxonomy is established, *a priori*, in which formal definitions are given for the categories or types composing the taxonomy (e.g. Dubin, 1959 and 1960); or (2) an empirical taxonomy is derived from a body of data (e.g. Dubin and Dubin, 1963 and 1965). In both approaches the taxonomy established must conform to the logical criteria of all classification schemes, namely that it is determinate and exhaustive; and that the categories are mutually exclusive and internally homogeneous.

INTRODUCTION

When a domain is imperfectly or inadequately known the usual approach in scientific inquiry is to derive empirical taxonomies for purposes of adequately describing such domain. The technologies for accomplishing this task have only recently been systematized. This paper explains one very simple technical method for deriving an empirical taxonomy and its integral types.

The approach employed here is unique because of its simplicity. It is also unique because it employs a test of *going together* rather than a test of *similarity* for grouping the attributes that define each type of the derived taxonomy.

An initial dsitinction needs to be made between *category* and *type*. A *category* is a single cell of a matrix. A *type* is an associated set of cells of a matrix. Any *n* by *m* matrix will produce *nm* categories. The same matrix will produce less than *nm* types, for, by the definition of type, at least two cells need to be associated to produce a single type. The economy of a taxonomic system producing two or more types is that the total number of categories of the matrix may be subsumed under a far fewer number of types.

Here is a standard problem faced by a researcher. Starting from hunch, or randon knowledge of a domain, data are collected producing values on an *ad hoc* set of attributes of a sample population presumed to be drawn from the domain of interest (Ashby, 1952; Dubin, 1969, ch. 3). The researcher then asks: "How can I characterize this sample population on the attributes I have measured, with the fewest number of types so that each sample member may be assigned to one and only one type?" Remember, each sample member is measured on all attributes in the set so that the researcher wants to know whether the arbitrary set of attributes utilized, or some subset of this set, can produce a typology consisting of two or more types. If a typology is successfully produced, then the type label can be employed to characterize each sample member, rather than the entire array of his special values measured on all the attributes employed.

The utility of having typologies is readily revealed in obvious examples from psychology where types of motivation, personality, or inter-personal relations are examined; in sociology where routine concern is with types of social groups, or collective behavior; in political science when focusing on types of governments, or types of governance; and in applied fields like medicine when diagnosing for types of disease.

Typologies always serve to subset a domain. The scientific purpose is to utilize the typology to compare and contrast representatives of two or more types with each other on characteristics other than those employed to derive the typology. In short, any analysis of contrast or relationship employed in research is grounded in a comparison of samples drawn from two or more types within a single domain.

Until recently no systematic attention was paid to the development of theory and technology for solving the problems of producing empirical taxonomies and their integral typologies. We now have such a literature. This paper presents one solution to the problem of producing an empirical typology that derived directly from a research project in which 3200 persons were measured on 124 attributes. We needed to order the attributes so that typologies produced would in turn permit an economical classification of the 3200 individuals.

This solution presented here is a member of the family of techniques found under the rubric *cluster analysis*. At the same time, however, the technique is quite different in purpose and method from conventional clustering schemes. In order to see this contrast, and to provide background information for our description of the technique, we shall briefly describe the nature and scope of cluster analysis.

CLUSTER ANALYSIS

Cluster analytic techniques search out the systematic (or latent) structure of a data matrix (Ball, 1965; Johnson, 1967). These techniques are particularly useful when there is no theoretical scheme or model to guide an analyst through a large matrix of data (Johnson, 1967, p. 241). Further, it would be clearly impossible to expect to "discover," by inspection, the structure of a large data matrix without using a search procedure specifically designed for that purpose.

By *structure* we mean the orderly groupings of data points in the data matrix. Each grouping (or cluster) contains data points that are more like each other than like data points outside of the group (Ball, 1965, p. 535; Bonner, 1964, p. 22). A major contribution of cluster analysis is its ability to reveal such natural groupings. The groups are defined by the data itself; they are not formed by the use of some external criterion of classification (Friedman and Rubin, 1967, p. 1159).

There is no shortage of clustering techniques. Their abundance is almost overwhelming, making the job of selection of a single technique to fit a particular problem exceedingly difficult. Ball (1965) for example, reviewed 27 techniques reported in the literature between 1960 and 1965.

Clustering techniques have seen wide application. The techniques have been extensively applied to problems of classifying plants and animals into types (Rogers and Tanimoto, 1960; Sokal and Sneath, 1963). In psychology, cluster analysis is used to identify types of individuals based on their patterns of responses on psychological tests (McQuitty, 1956). Bonner (1964) has demonstrated the use of cluster analysis in classifying diseases. The United States Navy has employed clustering techniques to solve the problem of developing a coherent occupational classification structure for enlisted personnel (Carr, 1967).

All clustering techniques employ two basic steps in order to define subsets or types of attributes in the matrix of attributes. The first step is the *putting together* of attributes that go together to form clusters. This is commonly done by using measures of association between all attributes taken two at a time in the matrix of attributes. There are many such measures with many names (Helmstader, 1957; Sokal and Sneath, 1963). For nominal measures of values on attributes, nonparametric measures of association such as chi-square may be used. Euclidian distance and the matching coefficients of numerical taxonomy are also suitable for nominal scales. For ordinal measures of values on attributes, the correlation coefficient is widely used, as well as Euclidian distance measures.

In either case the determination of *togetherness* of attributes in a cluster is by means of a measure of association. The higher the value of the measure of association, the more alike are the attributes measured. (For distance measures, the smaller the distance, the more alike are the attributes.) This is a point of view that underlies the philosophy of putting together the attributes that go together. The central point here is that attributes are brought together because they are considered to be similar. In contrast, our approach to this problem is to *link* attributes rather than expect them to come together because they are similar to each other. (See our basic linkage rule, p. 17.)

The second basic step in clustering techniques is the determination of the boundaries between clusters of attributes. When distance measures are used, the boundary is established by determining how far out from a central point (arbitrary or representative) can any attribute be and still be a member of a cluster. For similarity measures, a threshold level of measured association determines cluster membership. When the measured association of an attribute with one or all existing members of a cluster exceeds the threshold value, the attribute is included in the cluster. Otherwise, it is not. In both instances the boundary is arbitrary since the maximum distance and the threshold level of association are arbitrary.

In general, clustering techniques use measures of association to form clusters. The clustering technologies also specify the manner in which *likeness* or *closeness* of data points is to be determined, establishing boundaries between types that permit unequivocal assignment of each data point to one and only one type. With these two basic steps in mind we can see clearly the objective of clustering techniques as defined by Ball (1965):

> The essential characteristics of the techniques ... is the sorting of the set of data patterns into subsets, such that each subset contains data points that are as much 'alike' as possible (p. 535).

Or, as McQuitty (1957) has defined the term *type* in the context of an empirically determined typology:

> A type is here defined as a category of persons of such a nature that everyone in the category is in some way more like some other person in the category than he is like anyone not in the category (p. 213).

Two types of clusters emerge from a clustering technique depending on the criterion used for admission to the cluster (Cureton, Cureton, and Durfee, 1970; Johnson, 1967; Sokal and Sneath, 1963). The first, called compact clusters, occur when an object is admitted to a cluster only if it has a specified minimum level of association with *all* existing members of the cluster. Here, a completed cluster is said to contain highly similar objects. The second, called serpentine or amoeboid clusters, occur when an object is admitted to a cluster if it has its highest index of association with at least *one* existing member of the cluster. This method of clustering is also called single linkage clustering (Sokal and Sneath, 1963, pp. 180-181). As the name implies, clusters of this type may become elongated and include highly dissimilar items. Figure 1 shows these two typer of clusters. As can be seen from the figure, the end points of the serpentine cluster may indeed be dissimilar.

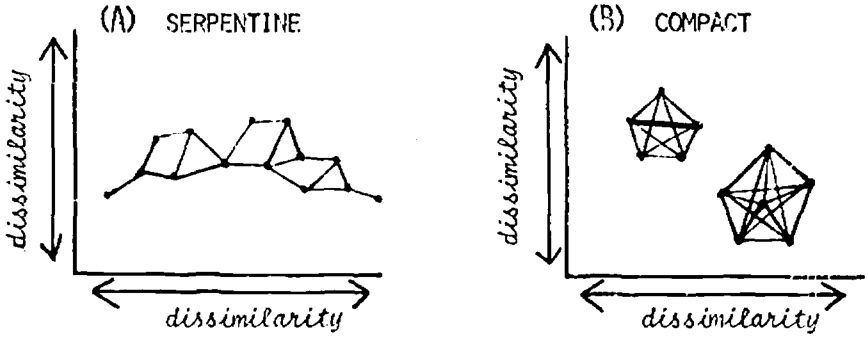

FIGURE 1. Serpentine (A) and Compact (B) Clusters. Any data point in a serpentine is linked to at least one other; all data points in a compact are linked to all others (modified after Sokal and Sneath, 1963, p. 192).

THE METHOD

We start with a distinction between the entity possessing attributes and the bundle of attributes possessed. The entities included in such a problem constitute a sample of "wholes" drawn from a population. These wholes may be a sample of people, a sample of plants, a sample of rocks, a sample of diseases, and, in general, any sample of entities that share common membership in a defined domain. The entities are identical to what Sokal and Sneath (1963, p. 121) call OTU's, Operational Taxonomic Units.

For each member of the sample of entities a set of attributes is measured in identical fashion. The attributes may be determined *a priori* or they may represent an *ad hoc* selection of attributes measured on the sample of entities.

We then emerge with a matrix with individual entities on one axis and attributes on the other axis. Each cell of the matrix contains the measured value of the particular attribute on the given entity. The analytical problem is now to determine how the entities may be grouped or typed in accordance with the values taken by the attributes for each entity.

We solve the analytical problem by first asking whether we can develop groupings of the attributes in the matrix. We want to know whether attributes A, B, C, D...N can be divided into subsets because they *go together* when measured on the entities included in the population sample.

Note carefully that the idea of going together means that the range of values on one attribute is regularly associated with a range of values on another attribute. The going together of two or more attributes does not depend upon the attributes being *like* each other, only that their particular values appear to be systematically related beyond a chance probability. Indeed it is quite clear that the very definition of the attributes included in the analytical problem requires that each attribute be different from all others in some determinant way, for if it is not then it would not be included in the array of attributes chosen for analysis. Thus, our purpose is not to measure similarity, which would only prove, if found, that two or more attributes were redundancies of each other. Our purpose is rather to find out how dissimilar attributes associate with each other because of the regularity with which their respective measured values are associated.

The second part of our analytical problem is then to find a method for assigning each individual entity to one and only one of the types that emerge when we have discovered how the attributes go together. When we have made such assignment of entities to particular types, we are confident that the entities within a type are more like their fellow members on attribute values than any of them are like the members of any other type in the particular taxonomy.

To summarize: (1) we want to be able to group dissimilar attributes into types to form a taxonomy of the types; and (2) we then want to be able to assign each entity on which the attributes have been measured to one and only one type.

The whole purpose of this exercise is to be able to give each entity a type label that specifically and concretely summarizes the values that entity possesses on a determinant number and kind of attributes. We can then use the type label to stand for all the attributes and their associated values that define the particular type. Thus, the type label turns out to be an important and economical analytical tool for then examining the relationship between types and other characteristics of the entities or their environments.

OPERATIONAL STEPS

The objective of the method described below is to develop a taxonomy of types, each of which is composed of attributes that *go together* because their values are associated in the sample. The method is presently designed to examine a data matrix of attributes, each of which is measured on a binary scale. In our specific case, this matrix is 124 attributes by 3200 respondents. This technique can be applied to data matrices of any size. The computational simplicity of the technique permits it to be manually applied to small matrices. Large matrices would have to be handled by a computer. The only limitation on matrix size would then be the storage capacity of the computer.

The current method is similar to existing clustering methods in one important respect. It is a linkage type of technique and produces clusters that are serpentine in structure. Its closest relatives in the family of clustering techniques are the single linkage method of Sneath (1957), Johnson's (1967) connectedness method, and the elementary linkage analysis technique developed by McQuitty (1957).

We were confronted with a body of data that consisted of 124 attributes, each describing one feature of the nature of industrial work or its environment. Every respondent was asked to indicate whether each attribute was important to him for any reason. In a paper and pencil instrument the respondent checked any item among the 124 that for him was important. Thus, every attribute had a score of present or absent, the absent score being determined when the respondent failed to check the item. Our problem was then to determine how these attributes, measured in the binary scoring system, could be grouped into types based upon the responses to the same questionnaire by almost 3200 industrial workers.

The method we evolved employs a binary scoring system for determining the value of each attribute. It has the important limitation that it will not generalize to any taxonomy in which one or more of the attributes is measured in a more complex than a binary manner. The reason for this will become apparent below.

In this section we describe the step-by-step procedure together with an illustrative example.

1. *Test Independence of All Pairs of Attributes* - Using the nonparametric chi-square test for two-by-two contingency tables, determine the independence or dependence of *all* pairs of attributes. If the computed chi-square value is significant, at the desired level of significance, the pair of attributes are dependent or related. If the computed chi-square value is not significant, the pair of attributes are not related (Siegal, 1956, pp. 104-111, 199-200). The contingency tables for this test are of the following form (in our empirical problem each attribute was dichotomized into zero and nonzero values; in the general case *any* dichotomization will work):

		ATTRIBUTE A	
		A	$\bar{A}$
ATTRIBUTE B	B	YY	NY
	$\bar{B}$	YN	NN

ERIC Full Text Provided by ERIC

Where Y = attribute has a nonzero value, and N = attribute has a zero value. Hence, the symbols in the four cells are interpretel as follows:

YY - Attributes A and B both have nonzero values.

NY - Attribute A has a zero value; B has a nonzero value.

YN - Attribute A has a nonzero value; B has a zero value.

NN - Attributes A and B both have zero values.

All pairs of attributes for which the relationship is not significant are ignored in the subsequent analysis. The remaining steps of the procedure are applied only to the statistically significantly related attributes. Thus, we normally expect to drop from further analysis all attributes not significantly related to any other. This is not surprising since we may have started with an *ad hoc* collection of attributes and should expect some to prove useless on analysis.

From this point on, the degree of association and the computed chi-square value, are no longer considered. As promised in an earlier section of this paper, the actual clustering of attributes does not use any measure of degree of association in the clustering procedure. The two-by-two table used in the chi-square calculation, however, is retained for use in the next step.

2. *Select Most Probable Kind of Association Between Two Attributes* - For each significant association select the one cell of the two-by-two table with the highest frequency as representing the *most probable* form of the association between the two attributes. Here we make the very simple assumption that the one best way to characterize how

the two attributes *go together* when they are associated beyond a chance probability, is to choose the one cell of the fourfold table having the highest frequency. This is simply another way of saying that if we were to assign probabilities of occurrence to each of the four cells of a fourfold table in which we have established a significant relationship, the cell with the highest frequency would have the highest probability of occurring.

It will be observed that if the relationship is significant in the fourfold table, the frequencies will be asymmetrically distributed in the four cells. The cell with the highest frequency must contain more than one-quarter of the total frequencies, and often will contain a majority. Thus, the rule for selecting the most probable relationship provides a realistic choice.

We now have a label for every pair of the attributes in the problem that has proved to be related beyond a chance probability. This label is the cell designation for the cell with the highest frequency, e.g., YY, NY, YN, or NN.

If we had any more complex relationship than a fourfold table, the most probable form of the relationship would be poorly determined by choosing the cell with the highest frequency. Thus, if one attribute had values measured on it that were trichotomized, a dispersion of frequencies among all six cells of the two-by-three table could mean that the cell with the highest frequency could have almost as few as one-fifth of the total frequencies. (For the relationship to be significant there must be an unequal distribution of frequencies among the cells, hence, one-fifth rather than one-sixth as the probable lower limit of minimum cell frequency.) This would certainly be an inadequate representation of the relationship between the two variables. It is for this reason we have suggested above that our clustering method is limited to attributes measured solely on a dichotomous scale.

3. *Array Pairs of Attributes* - Arrange the remaining significant relationships in a table similar to that shown in Figure 2. The order of the rows and columns is entirely arbitrary. The method does not depend on the order of the entries.

The columns are identified with the individual attributes. The rows are identified as the significant pair relationships among attributes. The only criterion for the construction of this table is that the row and column entries be an exhaustive listing of all attributes and the significantly related pairs of attributes.

For each row of the table there is the designation indicating the two associated attributes. Find the two corresponding columns and enter into these two cells, determined by the intersection of the row with each of these columns, the Y or N symbol derived from the fourfold table measuring the association between the particular pair of attributes. This will be the label derived in Step 2.

The resultant table with all the entries recorded will be comparable to the one shown in Figure 2.

It will now be noted that we have recorded all of the significant relationships determined in Step 2 and have produced a matrix having the following general characteristics.

(1) All significant relationships among all possible pairs of attributes are displayed.

RELATED ATTRIBUTE PAIRS	ATTRIBUTES					
	A	B	C	D	. . .	K
AB	Y	Y				
AC	N		Y			
AD	Y			N		
⋮						
AK	N					N
BC		N	Y			
⋮						
CD			Y	Y		
⋮						
CK			N			Y
⋮						

FIGURE 2. Array of Significant Attribute Pair Relationships (Illustration).

(2) Every attribute that remains in the matrix is related significantly to at least one other attribute.

(3) Each pair of significant relationships has one and only one of four possible ways that the attributes are related.

(4) Any single attribute may be related to any or all attributes.

(5) The use to be made of the matrix in the succeeding steps is in no way related to the order of rows and columns of the matrix.

In the procedure just outlined we have discarded the information contained in three of the four cells of each of the fourfold tables in which significance is established between pairs of attributes. We have retained and utilized the information in only one of the four cells. However, where standard measures of association are utilized, as with a correlation coefficient, or, in the case of a fourfold table, a contingency coefficient, we retain even less direct contact with the data of original entry. A contingency coefficient or coefficient of correlation will tell us only the amount of association and its direction. By the simple technique employed here, we are able to retain not only the idea that the two attributes go together but also to indicate specifically the most probable way they go together.

4. *Link Pairs of Attributes to Develop Types* - The basic rule for linking two or more *pairs* of attributes is: *two pairs of attributes are linked, if and only if, an attribute common to each has the same value in both.*

Attributes are linked together into types by performing the steps described below. The steps are described as if the method was to be performed manually. A computer could easily be programmed, of course, to perform the same steps.

1. Read down Column 1 (first attribute) and identify all pairs of attributes for which the value in Column 1 is the same. Thus, in our illustration of Figure 2, AB and AD each have a Y in the A Column. These pairs will, therefore, go together as parts of one type. Similarly, AC and AK will go together in another type because each has an N in Column A.
2. Search the array for the other half of the attribute pairs identified in the first column.
3. Search the columns of the attributes identified in Step 2 and identify any other attribute pairs with which the attribute of that column is associated by the same symbol. Referring again to our illustration, we note that in Column C, BC, and CD each has a Y, as does AC. However, neither AB nor AD shares the respective values of B and D with any other pair.
4. Search the array for the other half of the attribute pairs identified in Step 3.
5. Repeat Steps 2, 3, and 4, moving to the second and succeeding columns until all possible links between attributes have been made. In the illustrative case, CK emerges as the last independent pair.

6. Construct a *type* from these linked attributes by recording the attribute and its value in all pair-wise links. Thus, in our illustration we would obtain the following results:

   Type I: [A(Y)] + [B(Y)] + [D(N)], because [A(Y)] is linked to [B(Y)], and [A(Y)] is linked to [D(N)].

   Type II: [A(N)] + [B(N)] + [C(Y)] + [D(Y)] + [K(N)]

   Type III: [C(N)] + [K(Y)]

   It does not matter where this grouping is initiated in the table. It is most convenient to start in the upper left-hand corner of the table.

7. Terminate procedure when all attribute pairs have been grouped into types.

CHARACTERISTICS AND PROPERTIES OF TYPES

Possibly the most obvious characteristic of the method is its disarming simplicity. Small numbers of attributes can be easily handled manually. Large numbers of attributes may require a computer. In any event, however, the procedure for building the types remains the same.

The method will always yield a unique set of types, each defined in the identical way, regardless of the starting point in a given matrix. We described and illustrated the steps of the linkage procedure in terms of starting in Column 1 of the matrix. This starting point was arbitrary. Any starting point may be used with the same solution emerging.

Measures of association are not used by the method to form the types. This characteristic allows the method to be independent of the differential sensitivities of various association measures.

More importantly, since association measures are not used to form types, we are not tempted to argue that the attributes of a type are more similar to each other than to attributes external to the type. The linkage algorithm is specifically designed to bring together those attributes that go together. It does not link similar attributes.

Thus, we see that the criterion for the formation of a type consists of two elements. First, there must be a significant association between members of attribute pairs included in a type. Second, one member of the attribute pair must share the same symbol with one member of at least one attribute pair already in the type.

An attribute *may* be a member of two types. The symbol denoting its membership in the second type, however, is always the opposite of the symbol denoting its membership in the first type. For example, if attribute A appears in Type I with symbol Y (nonzero value) then, if attribute A appears in Type II, its symbol must be N (zero value). This property can easily be seen by recalling the linkage procedure within a single attribute (column). All Y's in a column are linked together and all N's are linked together. This procedure clearly restricts an attribute to membership in no more than two types and always with opposite symbols.

Given M attributes, the method produces a minimum of one type and a maximum of M/2 types if M is even or (M-1)/2 if M is odd. A single type emerges when all attributes are significantly associated with all other attributes. The maximum number of types occurs when each attribute is significantly associated with only one other attribute *and* each attribute pair is unique.

ASSIGNMENT TO TYPES

The assignment of entities to types of the taxonomy is divided into two steps. (1) The assignment of entities to that type each fits unequivocally. (2) The assignment of entities to that type each "fits" best when the match between entity characteristics and type features is imperfect.

The first step is readily apparent. Each entity is matched against all types to determine whether entity and type characteristics are identical. When they are, the entity is assigned to the matching type. From that point on the type label can be used to identify the entities falling within the type.

The second step requires elaboration, with the decisions leading to the solution of the matching problem being spelled out in detail. The major decision points are to: (a) determine a systematic rationale for treating the deviation of an entity's characteristics from the defining characteristics of the type; (b) establish a rule for assigning the entity to one type; and (c) develop some criterion of the acceptability of the match between the entire taxonomy developed, and the sample of entities from which it is derived.

In determining why a given entity does not exactly match, or perfectly fit into a given type of the taxonomy we first have to return to the original basis for measuring the values on the attributes included in the starting domain. We limit our measures to two values (in our particular example to zero and nonzero values). Therefore, for any given attribute, the entity can have only one of two values on it. A failure of the entity to match the type characteristics must consequently mean that for at least one attribute included among those defining the type, the entity value is opposite that of the type.

In order to assign the entity to a type it will then be necessary to assume that the entity is "in error" to the degree that it does not conform exactly to the characteristics of one type. What meaning can be assigned to the condition of the entity being "in error?" In general, we can consider three possibilities.

(1) The entity is "in error" because it is intrinsically imperfect, defective on the values it possesses for those attributes on which it differs from the type characteristics. In this event, the appropriate decision is to assign the "correct," or type values to the entity attributes. We are here simply assuming that if we remove the intrinsic imperfections in the entity it will then match exactly one of the types.

(2) The value measured on the attribute(s) for which the entity deviates from the type represents a measurement or instrument error. In this event, the appropriate decision is to do exactly what was done in the first instance; change the entity value to conform to a type value on all attributes where they differ. Here the assumption is that we can rectify measurement and instrument errors, in the belief that they are revealed in the process of the research whenever there is a

failure of values measured on a given entity to conform to an empirically derived standard or norm for entities drawn from the same domain.

(3) In the special case where the entities are actively involved in the measuring process (human subjects recording their own attitudes, for example) we can assume that the failure to match the type value is a "response error," having its source in the entity's responding output. Again, the appropriate correction is to change the entity value to that of the type to which it is to be assigned.

In each of the three instances we end up by making the values of the entity conform to the values for the type attributes. This is logical since we are deriving an empirical taxonomy. There will, therefore, be more entities determining the characteristics of each type than there are entities deviating from the type. The weight of correction should favor the group norm over the individual configuration. This point will turn out also to provide the basis for determining the acceptability of the match between the entire taxonomy and the sample of entities on which it is based.

We now turn to the second decision of determining a rule for assigning an entity to one and only one type of the taxonomy. Although it has been specifically noted only in passing, it should be recalled that all entities are measured on all attributes. From the procedure utilized in establishing the typology it is clear that we will discard all attributes that do not bear a statistically significant relationship to at least one other. Obviously then, we would also ignore the discarded attributes when assigning the entity to the closest matching type.

If the rationale is accurate that, for any one or combination of the three reasons just examined, the entity value on an attr:bute is "in error," then we must change the values on the entity to make them conform to the type values. The general rule for making these corrections is: *assign entity to the type requiring minimum changes in the entity's attribute values*. Operationally this means that we assign the entity to the type requiring a minimum change in the number of attributes employed in the type.

Several consequences of this assignment rule need to be examined.

(1) Since the number of attributes entering into the definition of any given type in a taxonomy may be different from the number in all other types, the search procedure for finding the type requiring minimum change in entity values is complicated. To facilitate this search we would start with the type having the fewest defining attributes and count the number of changes in attribute values needed for this and each succeeding type having the same or a greater number of attributes. In the event that there is a tie in the number of attribute values needing changing to assign an entity, the entity should be assigned to the ty,.e having the greatest number of defining attributes. The rationale for this secondary rule is that the more attributes entering into the definition of a type, the more homogeneous is the population of that type (Dubin, 1969, ch. 5). Therefore, we would be utilizing the maximum available information in making the assignment of the entity to the type having the greater number of defining attributes.

(2) The variable number of attributes that may define the several types of a taxonomy differentiates this method from scaling techniques

like Guttman's, for example (Guttman, 1950). In scale analysis all the types are defined by exactly the same array of attributes. Each type is distinguished from all others in scaling by the combination of values that characterize it on the identical set of attributes used throughout the taxonomy.

(3) For each non-fitting entity that must be "corrected" to match a type, there needs to be a decision regarding its ultimate possibility of fitting any type. That is, even if we can find one type to which an entity can be assigned on the basis of minimum changes in values of attributes, does there come a point where the actual number of changes is so great that we can no longer assume that the modified set of scores represents the original individual? We need a rule for determining the limit of changes permitted. For example, if we are matching an entity to a type having only two attributes defining it (the minimum number) then we could make the most divergent entity conform by changing only two values. Suppose the same entity could also match a more complicated type having seven defining attributes by changing values on three attributes. By the rule of minimizing changes the entity should be assigned to the two-attribute type, even though the entity has more attributes (four) on which it exactly matches the complicated type. We, therefore, need a modified rule or rules that make sense of this kind of anomaly.

The first modification of the assignment rule is: *no entity may be assigned to a type if the number of changes in attribute values is greater than one-half the number of attributes defining the type.*

As a special case of this rule to cover the instance of a type defined by an odd number of attributes, the following secondary rule is established: *no entity may be assigned to a type with an odd number of defining attributes if the number of changes in the attribute values of the entity exceeds (n/2 + 1)*, where n = number of attributes defining the type. Both of these secondary rules are necessary. The first limit takes care of the problem created by the fact that all entities can be fitted into a single two-attribute type by changing values on a maximum of two attributes. The second limit resolves, conservatively, the indeterminacy about the meaning of "one-half" when there is a model total number of attributes.

Another situation that will be encountered is one where no assignment can be made because the number of changes in entity attribute values exceeds the permissible limits. In that instance the entity is excluded from the sample as an entity that does not belong to the domain from which the sample is drawn. However, since the typology was derived from the data of the total sample, including the now-to-be-discarded entities, we reach an impasse. The most direct solution to this dilemma is to re-analyze the remaining sample after all non-fitting entities have been removed by going back and producing a new typology by the method here described. The new typology will differ relatively little from the old in probable content of the typology. Nevertheless, it is desirable to undertake this re-analysis since it insures that the typology ultimately used will accord with the population sample of entities upon which it is based.

The final decision point has to do with the match between values demanded by the entire typology and the values measured on the total sample population. The limits are clearly established in the procedure utilized for making an assignment of individual entities to their respective types.

If all entities have their attribute values exactly matching the attribute values of the types to which each is assigned, then there is no deviation between sample and typology. If all entities have the maximum number of changes in attribute values permitted by the assignment rules then the number changes in sample attribute values is the sum, over all types, of the permissible number of changes for each type. It will be recalled that no entity may be assigned to a type if the number of changes in its attribute values exceed one-half the number of attributes defining the type (or one-half plus one in the case of an odd number of attributes).

In a real situation we would not expect either of these two extremes to be realized. The individual researcher, who is more knowledgeable than any one else about the domain of his inquiry, must then determine what is to him an acceptable level of overall fit between the typology this method produces and the values of the attributes actually measured in the sample. Thus, for example, the researcher may discover that a failure to match the type attribute values may be observed differentially among the types of the taxonomy. This information may be far more important to the researcher than any measure of general agreement between entity values and typology values.

Put another way, there may emerge a hierarchy of types in which some types are far more completely matched by the sample entities than other types. In this event, the researcher would have more confidence in the types producing the greatest match with the empirical reality and might then concentrate his attention on improving the definition of the types where the match is poor.

We are, therefore, suggesting that rather than focus upon the total match between typology and sample of entities the researcher will find it more profitable to attempt improvement in the definition of individual types least representative of sample members.

The purpose of the typology, after all, is to provide an objective, shorthand way of labelling entities. Ability to improve any single label or type is a net gain toward achieving this objective. We, therefore, recommend that the researcher be more concerned with this issue than to try to develop some single measure (like the coefficient of reproducibility utilized in Guttman's scale analysis) that will measure the general correspondence between the typology and the sample.

SUMMARY

Dissimilarity Linkage Analysis (DLA) is an extremely simple procedure for developing a typology from empirical attributes that permits the clustering of entities. First, the procedure develops a taxonomy of types from empirical attributes possessed by entities in the sample. Second, the procedure assigns entities to one, and only one, type in the taxonomy. This two-step procedure clearly contrasts with many existing clustering techniques that are concerned only with the second step of our two-stage procedure (Ball and Hall, 1967; Sawrey, Keller, and Conger, 1960; Sokal and Sneath, 1963).

To develop a taxonomy of attribute types, the method searches for attributes that go together. A statistical test of association is first used to identify all pairs of attributes whose empirical values are significantly associated. Attribute pairs are then *linked* together to form serpentine clusters, each of which represents an attribute type. The attributes defining each type are *not* similar. In fact, the method specifically avoids using any criterion of similarity when developing the types.

Each entity is then assigned to the type it most closely resembles. An entity may unequivocally fit a type. Or, if an entity does not possess all of the characteristics of a type, it is assigned to the type with which its attribute values best match.

We thus form discrete clusters of entities based on their attribute types. In short, our method moves from types defined by dissimilar attributes, to clusters of similar entities in each type of the taxonomy.

REFERENCES

Ashby, W. R. *Design for a brain*. London: Chapman and Hall, 1952.

Ball, G. H. Data analysis in the social sciences: What about the details? *American Federation of Information Processing Societies Conference Proceeding. Fall Joint Computer Conference*. Washington: Spartan Books, 1965. Pp. 533-560.

Ball, G. H., & Hall, D. J. A clustering technique for summarizing multivariate data. *Behavioral Science*, 1967, *12*, 153-155.

Bonner, R. E. On some clustering techniques. *IBM Journal of Research and Development*, 1964, *8*(1), 22-32.

Carr, M. J. *The SAMOA method of determining technical, organizational, and communicational dimensions of task clusters*. San Diego, California: U. S. Naval Personnel Research Activity, 1967.

Cureton, E. E., Cureton, L. W., & Durfee, R. C. A method of cluster analysis. *Multivariate Behavior Research*, 1970, *5*, 101-116.

Dubin, R. Deviant behavior and social structure: Continuities in social theory. *American Sociological Review*, 1959, *24*, 147-164.

Dubin, R. Parsons' actor: Continuities in social theory. *American Sociological Review*, 1960, *25*, 457-466.

Dubin, R. *Theory building*. New York: Free Press, 1969.

Dubin, R., & Dubin, E. R. The authority inception period in socialization. *Child Development*, 1963, *34*, 885-898.

Dubin, R., & Dubin, E. R. Children's social perceptions. *Child Development*, 1965, *36*, 809-838.

Friedman, H. P., & Rubin, J. On some invariant criteria for grouping data. *Journal of the American Statistical Association*, 1967, *62*, 1159-1178.

Guttman, L. Chapters 2, 3, 6, 8, and 9. In S. Stouffer (Ed.), *Measurement and prediction*. Princeton, New Jersey: Princeton University Press, 1950.

Helmstadter, G. C. An empirical comparison of methods for estimating profile similarity. *Educational and Psychological Measurement*, 1957, *17*, 71-82.

Johnson, S. C. Hierarchical clustering schemes. Psychometrika, 1967, 32, 241-254.

McQuitty, L. L. Agreement analysis: Classifying persons by predominant patterns of responses. The British Journal of Statistical Psychology, 1956, 9, 5-16.

McQuitty, L. L. Elementary linkage analysis for isolating orthogonal and oblique types and typal relevancies. Educational and Psychological Measurement, 1957, 17, 207-229.

Rogers, D. J., & Tanimoto, T. T. A computer program for classifying plants. Science, 1960, 132, 1115-1118.

Sawrey, W. L., Keller, L., & Conger, J. J. An objective method of grouping profiles by distance functions and its relation to factor analysis. Educational and Psychological Measurement, 1960, 20, 651-673.

Siegal, S. Nonparametric statistics for the behavioral sciences. New York: McGraw-Hill, 1956.

Sneath, P. H. A. The application of computers to taxonomy. The Journal of General Microbiology, 1957, 17, 201-226.

Sokal, R. R., & Sneath, P. H. A. Principles of numerical taxonomy. San Francisco: W. H. Freeman & Company, 1963.

Weber, M. The methodology of the social sciences. Glencoe, Ill.: The Free Press, 1949.

UNCLASSIFIED
Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)
University of California
Graduate School of Administration
Irvine, California

2a. REPORT SECURITY CLASSIFICATION
UNCLASSIFIED

2b. GROUP

3. REPORT TITLE
TYPOLOGY OF EMPIRICAL ATTRIBUTES: DISSIMILARITY LINKAGE ANALYSIS (DLA)

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)
Scientific Technical Report #3

5. AUTHOR(S) (First name, middle initial, last name)
Robert Dubin
Joseph E. Champoux

6. REPORT DATE
15 June 1970

7a. TOTAL NO. OF PAGES
31

7b. NO. OF REFS
23

8a. CONTRACT OR GRANT NO.
N00014-69-A-0200-9001 NR151-315

b. PROJECT NO.

c.

d.

9a. ORIGINATOR'S REPORT NUMBER(S)
Technical Report #3

9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
NONE

10. DISTRIBUTION STATEMENT
Distribution of this document is unlimited. Reproduction in whole or in part is permitted for any purpose of the United States Government.

11. SUPPLEMENTARY NOTES

12. SPONSORING MILITARY ACTIVITY
Personnel and Training Research Programs Office, Office of Naval Research

13. ABSTRACT

Dissimilarity Linkage Analysis (DLA) is an extremely simple procedure for developing a typology from empirical attributes that permits the clustering of entities. First the procedure develops a taxonomy of types from empirical attributes possessed by entities in the sample. Second, the procedure assigns entities to one, and only one, type in the taxonomy. This two-step procedure clearly contrasts with many existing clustering techniques that are concerned only with the second step of our two-stage procedure.

To develop a taxonomy of attribute types, the method searches for attributes that go together. A statistical test of association is first used to identify all pairs of attributes whose empirical values are significantly associated. Attribute pairs are then linked together to form serpentine clusters, each of which represents an attribute type. The attributes defining each type are not similar. In fact, the method specifically avoids using any criterion of similarity when developing the types.

Each entity is then assigned to the type it most closely resembles. An entity may unequivocally fit a type. Or, if an entity does not possess all of the characteristics of a type, it is assigned to the type with which its attribute values best match.

We thus form discrete clusters of entities based on their attribute types. In short our method moves from types defined by dissimilar attributes, to clusters of similar entities in each type of the taxonomy.

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Typology						
Cluster Analysis						
Empirical Typology						
Linkage Analysis						