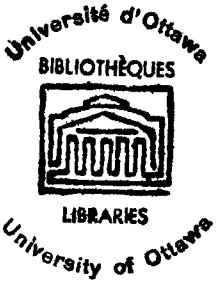
QUANTITATIVE DATA GROUPING TECHNIQUES AS APPLIED TO CRIMINOLOGY

by

Paul S. Maxim



Thesis submitted to the School of Graduate Studies of the University of Ottawa, in partial fulfillment of the requirements for the degree of Master of Arts in Criminology.

ACKNOWLEDGEMENT

The author wishes to express his deepest appreciation to his thesis director, Dr. C.H.S. Jayewardene and to Dr. J. Ciale for their continued guidance and encouragement.

A special note of thanks goes to L. Culumovic for his invaluable assistance with the voluminous programming involved in this project.

Gratitude is also expressed to Mr. M. Nolan, Deputy Director of the Regional Reception Centre, Kingston, Ontario, for his efforts in making available the data for this study.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

I

INTRODUCTION TO DATA GROUPING TECHNIQUES

1. Introduction

Any discussion of data grouping techniques must necessarily cover a wide range of topics and areas of study. In searching the relevant literature the researcher will find himself covering fields as broad as mathematics, biology, geography, and logics and indeed, his own substantive area which in this case is criminology. Along with a range of disciplines, the researcher will also have to cope with a variety of substantive issues and an even wider range in nomenclature. Depending on his background, the researcher will refer to data grouping techniques under the headings of classification, typologies, taxonomy, cluster analysis, and numerous other names. Invariably, these terms refer to similar phenomena: the grouping or clustering of data into discrete homogeneous categories.

This monograph is essentially a discussion of a certain subset of data grouping techniques, namely, quantitative data grouping techniques. These techniques are characterized by a reliance on quantifiable variables for grouping units or phenomena.

This presentation will commence with an overview of quantitative data grouping techniques; presenting some of those most applicable and commonly used in the

field of criminology, and then proceed to outline the generalized logical structure of these techniques. Finally, two techniques which represent different classes of sorting strategies, will be applied to similar data in order to compare their ensuring clusters. From this comparison it is hoped that a clearer understanding of the operation of these techniques will be achieved.

2. Purpose

Typology construction appears to be a rather ubiquitous phenomenon in criminology as well as in all other areas of scientific activity. The most poignant apology one can present for engaging in typology construction is probably contained in Good's[1] list of the purposes of classification. Basically, says Good, typologies are constructed:

i) for mental clarification and communication;

ii) for discovering new fields of research;

iii) for planning an organizational structure or machine;

iv) as a check list; and

v) for fun.

Needless to say, these reasons need not necessarily be viewed as being mutually exclusive. The zeitgeist being as it is, however, requires an emphasis

to be placed on the first four reasons.

Certainly mental clarification and communication may be reason enough to engage in typology construction. The grouping or clustering of data and concepts aids in memory and cognitive formulation. It is easier to communicate and understand a term such as "felon" than it is to list the entire universe of behaviours for which, if one were apprehended and adjudicated by the state, one would be stigmatized as being a criminal. Typologies provide us with a degree of mental economy by grouping phenomena which share certain essential characteristics.

Grouping phenomena into categories with high intraclass homogeneity and interclass heterogeneity may also prove to be of some heuristic value. Questions such as why does this group of individuals have these characteristics in common; what is the significance of having this kind of profile, and why do these individuals fall into this category and not another, soon lead to further research efforts. The results of inductive quantitative data grouping techniques can be compared with hypothetico-deductive groupings; the result of which ought to give rise to questions concerning any similarity or difference between the theoretical and empirical.

Data groupings may also be relevant for individuals attempting to construct or modify social or organizational structures. A cluster analysis of the criminal justice system may lead to an indication of which functional units are most similar and which may be combined to form a more efficient structural model. Similarly, hierarchical, agglomerative cluster methods may be used to determine patterns of social congeniality within prison populations, thereby indicating which individuals when placed together would cause the least disturbance. A certain degree of predictability may be achieved through the knowledge of which variables relate to which others. Knowing that certain phenomena are "similar" is in itself a fairly significant piece of information.

Beyond the traditional aetiological reasons for searching for patterns or consistencies in empirical data, the criminologist is faced with specific applied problems which relate directly to the need for effective grouping techniques. Jails and prisons have, and are continuing, on both humanitarian and pragmatic grounds, to differentiate various "types" of prisoners. It was not so long ago that prison reformers were advocating the separation of juveniles from adults and males from females. These

discriminations were pursued on ethical and ideological bases; to prevent moral impropriety between males and females, and to accentuate the rising belief in the qualitative difference between juvenile waywardness, or delinquency and adult criminality.

Today it is usually pragmatism which dictates the need for discrimination in prison and jail populations. As stated in a recent publication by the United States Bureau of Prisons:

> "The days have gone forever when jails, other than those in large metropolitan areas, dealt almost exclusively with local citizens. Not only were the reputations and backgrounds of these people generally known, but commitment to jail was a conspicuous event. From his own knowledge and information readily available to him, the jailer could quickly size up the situation and determine how best to handle each prisoner. In these days of rapid community growth and mobility, when it is commonplace for people to relocate frequently and travel from one end of the country to the other

> in a few hours, increasing numbers of jail prisoners are committed as total strangers."[2]

It is important in prisons from a treatment point of view, and also from a management perspective, to be able to make predictions from available information as to how certain inmates will behave. Which inmates should go to which treatment programmes? Given physical and budgetary constraints, is it advisable to send all inmates for individual counselling, or only those cases who are in greatest need, or those who will benefit the most? How do we decide which inmates should be placed into maximum, medium or minimum security settings? From a logistical point of view, it is a waste of resources -- human and otherwise -- to maintain maximum security facilities for all prisoners when medium or minimum would suffice. Yet it is essential for society to be able to distinguish those individuals who pose the gravest danger to society and to segregate them, both for their own and society's protection.

Discriminatory treatment of prisoners cannot be carried out irrationally though. The philosophy of Western man is such that certain notions of "justice" must be adhered to, and one of the first principles of

justice is that individuals should not be subjected to arbitrary discrimination. As Benn and Peters indicate in their discussion on egalitarianism,

> "... the principle reiterated in the historic declarations of rights that all men stand equal can be reconciled with the plain fact that we feel it right to treat men differently. Equals should be treated equally, unequals unequally: but the respect in which they are considered unequal must be relevant to the differences in treatment that we propose. And so, until an inequality of attribute or condition has been shown to be relevant, it is improper to make it a basis of distinction. Impartiality involves not merely recognizing similarities, but knowing which differences ought to be ignored."[3]

It is often considered a rare moment when the goals of science and political thought coincide, since in many instances, these two endeavours appear to be at loggerheads. In this case, however, the coincidence exists. Political thought states that similarities ought

to be recognized for justice's sake; the recognition of similarity has long been the cornerstone of scientific enquiry.

3. A Case for Quantitative Grouping Techniques

When speaking of classifications or types in social science literature, it is usual to make the distinction between ideal or constructed types,[4] and empirical types. The concept of the ideal type is usually credited to Max Weber who indicates that,

> an ideal type is formed by the one-sided accentuation of one or more points of view and by the synthesis of a great many diffuse, discrete, more or less present and occasionally absent concrete individual phenomena, which are arranged according to those one-sidedly emphasized viewpoints into a unified analytic construct. In its conceptual purity, this mental construct cannot be found anywhere in reality.[5]

An ideal type is not a stereotype or a polar type, but rather a hypothetical construct. Indeed, such sociological debate centres around whether ideal types are constructs or hypotheses[6]. There are few examples of

what could be termed ideal types in criminological literature, however, Wood[7] suggests that Albert A. Cohen's explanation of delinquency implicitly uses ideal types since it is not actually a general theory of delinquency, but rather "an explanation of lower-class, male, hedonistic (nonutilitarian), gang delinquency"[8]. Similarly, he suggests that Cloward and Ohlin's opportunity theory effectively employs ideal type constructs when it deals with different delinquent subcultures.

The alternative to ideal type constructs are empirical types based upon the cross-classification of variables. In this instance, types result from the categorization of one or more variables. For example, the world may be categorized along a uni-dimensional continuum separating criminals from non-criminals. A multi-dimensional classification scheme may be created by the addition of more variables. Thus, dividing the world into criminal and non-criminal categories and then further subdividing it by sex -- male and female -- presents a fourfold typology or classification. Unlike ideal types, empirical types generally have empirical or observable referents. It may be, however, that certain populations do not encompass representatives of all possible categories, but this is a result of simple

idiosyncracies rather than a theoretical attempt to postulate non-entities.

Although ideal types have great heuristic value in theory construction, their value is limiting in an applied setting because of their lack of empirical referent, and the difficulties they pose when discrete categorization is required. As such, we will deal exclusively with empirical types in this study.

There are two general strategies or approaches which may be used to produce empirical types. They may be constructed inductively, in which case the researcher takes raw data and searches for natural clusters within the data, or they may be constructed deductively, by commencing with a theoretical formulation and postulating probable clusters. These two approaches are far from being mutually exclusive. Instead, they should be viewed as two points in an iterative process which provides theoretical propositions to determine which variables may prove to be most valuable, and empirical observation to confirm or modify those propositions. The days are long gone when social scientists such as George Lundburg[9] could espouse the virtues of a raw empiricism. Today,[10] it is recognized that empirical observation is not made randomly, but is guided by some theoretical principles

although these principles may not be the most explicit. Similarly, experience and observation determine the shape of our theorization.

Quantitative grouping techniques have certain strengths not always shared by qualitative methods.

Some quantitative techniques take advantage of the fact that given n variables in k states, not all k states may 'naturally' occur, that is, the k^n logically possible categories may not be manifested empirically.

For example, if thirty-one personality variables are measured dichotomously, there would be a total of 2^{31} or 2,147,483,648 possible categories. Quantitative techniques thus locate those clusters which exist empirically.

The use of quantitative techniques may also aid the theoretical understanding of phenomena since a grouping technique is just as dependent of its algorithm (or mathematical model) as it is on the actual data to which it is applied. The researcher's selection of one cluster algorithm over another is a reflection of the researcher's expectations as to how the data under study is spatially distributed, and what type or shape of classification system he suspects exists. If one cluster technique is found to be superior over another for some reason, then

the algorithm upon which that technique is based adds to the description, if not the explanation, of the observed groupings within the data. It may be too much to expect that empirical typologies and theoretical classifications will achieve a high degree of isomorphism, however, attempts to investigate the why and wherefore of this incongruence can only lead to a further understanding of the phenomena under investigation. Indeed, one of the major tasks (and one of the most neglected) facing numerical typologies is the development of measures of goodness of fit which will adequately describe observed discrepancies between numerical and theoretical typologies.

Another important property of quantitative grouping techniques is their heuristic value in exploring 'virgin territory'. It may be known that specific variables describe a phenomenon, yet the exact relationship of those variables may be unknown. Cluster techniques, because of their ability to synthesize a vast quantity of data, may give insight into the underlying structure of a phenomenon. Factor analysis is one clustering technique which has been used quite extensively for this purpose.[11]

Numerical data grouping techniques also possess some more direct advantages over other methods. They allow us to integrate fairly large amounts of data from many different sources. For example, criminal typologies

which synthesize many environmental, personal and attributional variables may simply be too cumbersome for non-numerical techniques to handle. The use of quantifiable techniques permits the use of many of our most advanced data handling tools - computers. Such tools make it possible for researchers to test many combinations of variables and many different models or algorithms over large sample sizes with a minimal expenditure of personal and financial resources.

With the effort to quantify concepts usually comes the effort to provide more and better nominal definitions in order to operationalize the concepts under investigation. As several investigators have recently indicated[12] one of the greatest hindrances to advancement in the social sciences is the lack of clear and precise nominal definitions. Because of the fogginess of many concepts, it is possible that much information regarding similarities and groupings of phenomena may already exist within the literature awaiting an adequate method or model of extraction to make them visible.

As with any mathematically based scheme, quantitative data grouping methods, or more properly the algorithms upon which they are based, may be treated as a scientific "model".[13] Indeed, they are very powerful

models due to their very explicit expression through mathematical symbols and relationships. Verbal models often lack the conciseness, clarity and capacity for logical manipulation offered by mathematics. As well, certain types of relations, such as those between variables related in hyperspace, are much easier to conceptualize in mathematical rather than in linguistic modes of presentation.

When presented in highly explicit mathematical logic, a priori assumptions and model limitations often become much more obvious. For example, mathematical measures of resemblance require the term "resemblance" to be operationalized in a very precise manner -- in a relational form much more explicit than most dictionary definitions of the term. Thus, it may be that at an operational level the positions and the number of boundaries between individuals reflect an arbitrary choice of numerical models, but the basis of comparison between those units will be consistent vis-a-vis the model selected. This consistency and clarity serves to make more obvious any similarity (or conversely, dissimilarity) between units, thus saving valuable research time by not necessitating the researcher's wading through reams of illogical relationships.

Often, a valuable by-product of quantitative techniques is the great ease by which they can be graphically illustrated with a high degree of isomorphism between the mathematic model and the graphic conceptualization. This enhances both the researcher's power of communication and the variety of perspectives through which he can observe and present his data.

Blind acceptance of quantitative data grouping, it should perhaps be pointed out, violates the principles of scientific investigation. Because of their explicitness quantitative methods often carry the seeds of their own destruction. When used to their optimum, quantitative methods lay bare the interrelationships between concepts, but this visibility also leaves them vulnerable to being disproven as viable models or techniques. While other systems of logic may find defence behind a hazy wall of interpretation and re-interpretation, quantitative techniques are totally vulnerable to the assessment of their truth-value. It may well be that certain types of data and social phenomena are totally unamenable to quantitative investigation and analysis.

References

1. GOOD, I.J. (1965) Categorization of classification mathematics and computer science in biology and medicine, London: H.M.S.O.

2. RICHMOND, M.S. (1971) Classification of jail prisoners, Washington, D.C.: U.S. Bureau of Prisons, Department of Justice.

3. BENN, S.I. & R.S. PETERS (1959) The principles of political thought, New York: The Free Press, p. 131.

4. McKINNEY, J.C. (1966) Constructive typology and social theory, New York: Appleton-Century-Crofts.

5. WEBER, M. (1949) The methodology of the social sciences, translated by E.A. Schils & H.A. Finch, Glencoe, Ill.: Free Press, p. 90.

6. MARTINDALE, D. (1959) "Social theory and the ideal type", International Encyclopedia of The Social Sciences, 1968, New York: Macmillan, 177-186.

7. WOOD, A.L. (1969) "Ideal and empirical typologies for research in deviance and control", Sociology and Social Research, 53, 227-241.

8. WOOD, A.L. (1969) op. cit., 230.

9. LUNDBERG, G.A. (1939) "The postulates of science and their implications for sociology", in M. Natanson (ed.) 1963, Philosophy of the Social Sciences, New York: Random House.

10. GOULDNER, A.W. (1964) "Anti-minotaur: The myth of a value free sociology", in I.L. Horowitz (ed.) 1964, The new sociology, New York: Oxford University Prsss.

11. KERLINGER, F.N. (1973) Foundations of behavioural research, 2nd ed., New York: Holt, Rinehart & Winston, 659-692.

12. LACHENMEYER, C.W. (1971) The language of sociology, New York: Columbia University Press.

13. CHAPANIS,A. (1961) "Men, machines and models", American Psychologist, 16, 113-131.

II

A REVIEW OF QUANTITATIVE TECHNIQUES

One of the first, and certainly one of the best known quantitative data grouping techniques to be applied to criminological data is Williams and Lambert's[1,2,3] Association Analysis. As biologists, the problem Williams and Lambert were initially addressing was how to take a large tract of land, or a large body of heterogeneous data and to break it down into more compact homogeneous clusters which could be described as being more similar to, or different from each other in terms of the total amount of available information. This problem was similar to those facing criminologists attempting to devise typological and classification schemes based on attributes of individual criminals. MacNaughton-Smith[4] was one of the first to suggest that Association Analysis could be applied to personal attributes in an attempt to classify individuals. Wilkins and MacNaughton-Smith[5] applied both Association Analysis and a modification known as Predictive Attribute Analysis[6] to a study of English Borstal Boys. Gottfredson[7] and Ballard [8] performed similar analysis on a sample of California parolees.

Figure 1 illustrates the result of a recent study by Sinclair & Chapman[9] which used Association analysis to examine the relevant characteristics of 1009 men measured on 45 variables. Of the 45 variables

measured, 6 appeared to be most important in defining "natural clusters" which appear in the data.

A descriptive summation of the seven types is presented in Table 1.

Computationally, Association Analysis is rather cumbersome, requiring the aid of an electronic computer. However, it is conceptually quite simple. There are four basic steps to follow in the application of the technique. First, all variables are reduced to nominal dichotomies which are coded on a presence-absence, or 0,1 basis. It is this aspect of the analysis which renders it the label monothetic.

Some variables such as sex, naturally lend themselves to dichotomization; other variables which may normally be measured at the ordinal or higher level must be grouped. This requires a value judgment on behalf of the researcher since there is often no easy manner by which to determine optimal (or theoretically significant) divisions or 'cut-off' points. Varying the cut-off point on one variable may affect the optimum position on other variables.

After dichotomization the data is entered into a square matrix, the rows and columns of which consist of the attribute variables upon which the individuals will

FIGURE 1[10]

RESULTS OF SINCLAIR AND CHAPMAN STUDY WITH

ASSOCIATION ANALYSIS

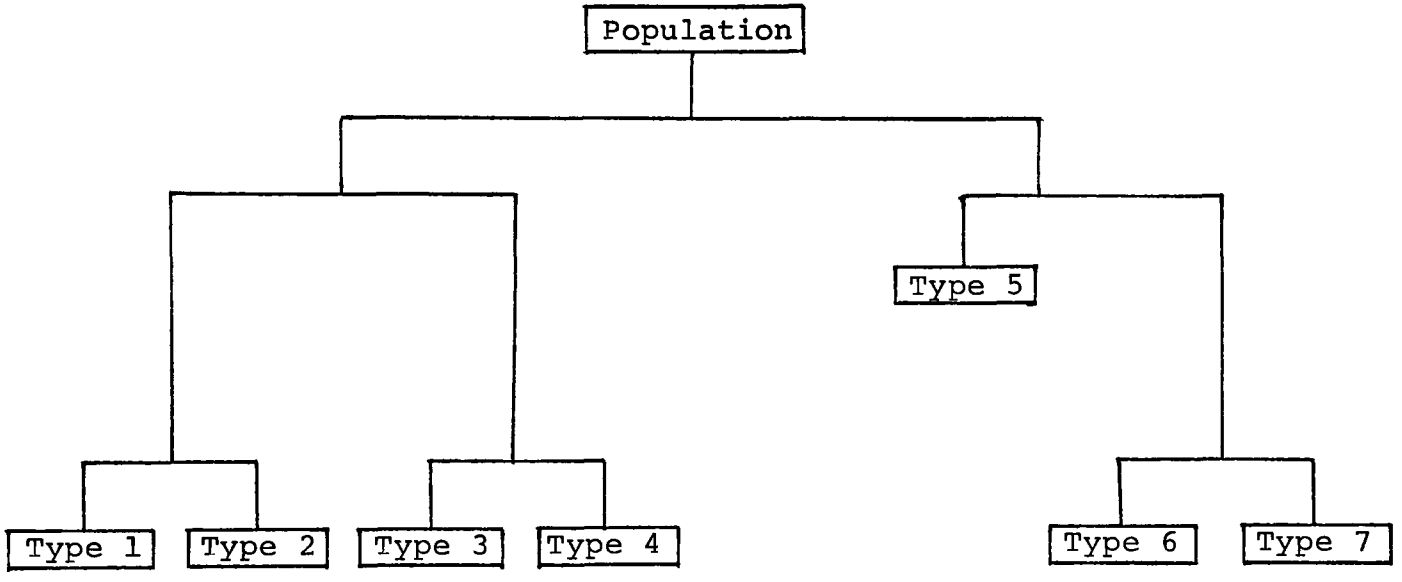


TABLE 1

SINCLAIR AND CHAPMAN TYPOLOGY

	T1	T2	T3	T4	T5	T6	T7
Under 30 yrs. of age	yes	yes	yes	yes	no	no	no
Less than 3 prior conv.	yes	yes	no	no	---	---	---
Under 25 at first conv.	yes	no	yes	no	---	---	---
Planned present offence	---	---	yes	no	---	---	---
Employed as skilled labour	---	---	---	---	yes	no	no
Living at home when arrested	--	---	---	---	---	yes	no

eventually be classified. The chi-square (X^2) statistic is calculated for each pair of variables within the matrix, and then summated down each column. Thus,

$$X_j^2 = X_{j,1}^2 + X_{j,2}^2 + \ldots + X_{j,j-1}^2 + X_{j,j+1}^2 + \ldots + X_{j,m}^2$$

The largest value X_j^2 is chosen as the criterion variable upon which the group is divided. The two ensuing sub-groups are then treated as primary groups and the process is repeated until the X^2 drops below a specified value or "stopping criterion". Since it is theoretically possible to divide the initial group into N sub-groups, some criterion must be provided to indicate the point at which the sub-dividing ought to be terminated. Williams and Lambert[12] call the sub-sets final if all $X_j^2 < 3.84$, which corresponds to the critical value of the X^2 statistic at a 5% significance level, with one degree of freedom. MacNaughton-Smith[13] has advocated a much larger critical value and suggests $\max_j X_j^2 < X^2$, with m-1 degrees of freedom at the .05 level of significance (where m indicates the number of variables remaining within any one group).

Several alternatives have also been brought up in the calculation of the X^2 statistic. Lambert and

Williams[14] have utilized Yate's correction[15] in the calculation of X^2 and have also used the absolute value of the square root of X^2 $\left(\left|\sqrt{X^2_{j_1}}\right| + \left|\sqrt{X^2_{j_2}}\right| + \ldots + \left|\sqrt{X^2_{j_n}}\right|\right)$ to determine criterion variables. MacNaughton-Smith[16] remains a purist, however, and supports only the use of an unadulterated X^2 statistic.

Lance and Williams[17] make an interesting observation, which has made Association Analysis easier to use computationally and also allows for the possibility of its use as a polythetic grouping technique. They noted that the X^2 statistic is N times the correlation coefficient 'r' when the data are coded 0,1. Thus, instead of summating X^2_{ij}, one would summate r_{ij}, and use the largest r_j as the criterion variable for division. It is also interesting to note that X^2 divided by N is the formula for the Phi-coefficient (Ø). Hence, given nominal, dichotomous data, coded 0,1,

$$\frac{X^2}{N} = \phi \doteq r$$

Gower[18] develops this theme and has proposed that multiple correlation be substituted for the summated chi-square or correlation coefficient scores. Using this method, each variable would in turn be used as a dependant variable with the remaining variables included into

the regression equation as independent variables.

Fildes and Gottfredson[19] have used both versions of Association Analysis -- the traditional Williams and Lambert version but substituting ϕ for X^2, and Gower's modified approach which uses multiple regression -- for an analysis of 16,800 parolees. Phi was substituted for the X^2 statistic since not all of the variables were dichotomies, thus rendering this application polythetic. The sample was split in half to provide two sub-samples, thus enabling the authors to gain some insight into the outcome stability of both techniques across samples. A single criterion variable of parole success was used to determine the degree of between group difference. Using a statistic developed by H.R. John, the authors measured the degree to which the various sub-groups differed from the overall population in relation to parole outcome. The John statistic used by Fildes and Gottfredson was defined as:

$$J = \sum_{i=1}^{k} \frac{N_i \left| P_i - P_s \right|}{N P_s Q_s}$$

where,

N = number in sample

N_i = number in 1^{th} of k categories

P_s = success rate of sample

P_i = success rate in 1^{th} category

Q_s = 1 - P_s, or the failure rate of the sample

Fildes and Gottfredson found the Gower technique produced greater sub-group heterogeneity than the traditional Association Analysis based on the statistic ø. The results they obtained are presented in Table 2.

Table 2^{20}

Values of J Statistic

Method	Sample 1	Sample 2
Gower	.193	.190
Association Analysis	.172	.133

The major drawback in using multiple regression in place of X^2 or ø or some other similar statistic is the large number of cases required in order to perform the analysis. The Fildes and Gottfredson study used over 16,000 cases. Computational problems arise with this technique when the number of subjects is initially small, or when the number of subjects in a sub-group becomes small. Not only must one guard against the case of more variables than subjects (hence producing a singular matrix)

but one is soon faced with the problem of identical column vectors. If the data are coded 0,1, for example, it is hoped that a grouping technique would produce homogeneous groups with all 0's or 1's in the columns. This ideal, however, leaves the researcher with redundant information due to similar column vectors and, again, singular matrices.

Monothetic divisive schemes, in general, are open to several criticisms. As Gower has indicated, individuals lacking the particular attribute upon which a specific division is made will be wrongly classified even though they may be similar to other individuals on different attributes. Further, the very act of forcing all of one's variables into dichotomies may greatly distort the nature of the variable and place upon it a property it may not 'naturally' possess. It might also be indicated that the sub-types produced are highly artificial. This criticism may be directed toward any clustering technique, however, since a great deal of value judgment is involved from the initial choice of which variables to select and which technique to choose, to determine a cut-off point. Certainly, the results should only be interpreted in full view of theoretical constraints.

Lange, Stenhouse and Offler[21] in a comparison of Williams and Lambert's Association Analysis with Sneath's 'Q', found the X^2 statistic to be highly unstable in regards to high values of the X^2 statistic. The authors thusly concluded that it would be necessary to abandon those cases with sparse data. Using the split-half technique on an empirical sample with percentage agreement as a measure of similarity, they concluded that overall, both techniques were highly stable in terms of category replication.

The alternative to cluster analysis based on divisive techniques are those based on agglomerative approaches. Instead of starting with one initial group and working toward N sub-groups, one starts with N initial groups, each with a population of size 1, and combines them until only two groups remain. The aim of the exercise, of course, is to maximize the inter-group differences while maximizing intra-group similarity. Unfortunately, the amount of computation required could become quite tedious, for in order to group k variables into N groups, every possible permutation of the k variables must be taken into account over the N groups. Assuming that all variables were coded dichotomously, there would be 2^k possible categories arising from the variable permutations.

With only 10 variables this would lead to 2,048 possible categories and with 20 variables, over 1 million categories would be possible.

A technique devised by Ward[22], known as Ward's Hierarchical Grouping technique, provides a compromise approach in dealing with the problem. Given N individuals, with k attributes, Ward's technique looks at the absolute differences in scores between score values. This technique is considered a compromise over the optimal method of agglomerate grouping since the researcher would not consider all of the possible combinations of k variables over N individuals, but rather would use groupings at prior stages as indicators of optimum groupings. Thus, it reduces n sets to n - 1 mutually exclusive sets by considering the union of all possible $n(n - 1)/2$ pairs.

The Ward technique is based on the computation of a matrix of error terms for each pair of individuals. Those two groups which, then combined, produce the least error value are paired to form a new group. The process continues until only two groups remain. This technique also has the advantage that it provides an indicator of the amount of information lost during the grouping process since information loss is reflected in the error sum of squares. The basic assumption in this model is

that maximum information is available with N groups. Ward gives us the following formula to describe the error sum of squares,

$$SS_E = \sum_{x=i}^{n} X^2_i - \frac{1}{n}\left(\sum_{x=i}^{n} X_i\right)$$

where,

SS_E = error sum of squares
X_i = 1^{th} variable attribute
n = number of individuals in the potential group

Effectively, one produces an error matrix based on the sum of the squared differences between all individuals. That pair of individuals producing the least error potential is then combined to form a new group. Ward has indicated that his method is most effective when $N > 100$.

One of the major limitations of this technique, however, is that all of the attribute variables are considered with equal weight. That is, variables are not rank ordered to any degree with regards to their importance. Veldman[23] has suggested two possible techniques to correct this problem -- the first would be a simple

weighting process based on prior knowledge of the relative significance of the variables. The second technique is to use factor analysis weights (beta-weights) in place of raw data. The weighting process brings forward another problem inherent in analysing within group sums of squares or variance over several variables. That is, those variables with the largest range or variation will have the greatest impact upon the grouping process. The solution however, is to standardize the data beforehand which is easy enough, but standardized data itself negates the benefits of weighting.

If all of the data collected were measured on similar scales, specifically, scales with similar ranges, the problem would be negligible since simple weighting would be acceptable. Unfortunately, empirical reality does not always correspond to a fixed scaling procedure.

Unlike most quantitative grouping techniques, however, research has been performed on the Ward technique in an attempt to assess the accuracy of its grouping. Gross[24], using the Monte-Carlo technique for generating a sample, measured the number of misclassifications produced when samples were drawn from a known population. Gross discovered that when the proportion of misclasifications within the population was small (.02) accurate sample sub-

groupings resulted with a low variance of sample misclassification variance increased sharply unless the sample size was quite large. Gross ran his study under two conditions: the first with two groups of equal distribution ($q_1 = q_2 = .50$), and the second where the two groups were split in a 2:1 ratio ($q_1 = .67$, $q_2 = .33$). Table 3 summarizes Gross's results.

TABLE 3[25]

MEAN & VARIANCE OF SAMPLE MISCLASSIFICATIONS

		n = 100		n = 200	
		.02	.20	.02	.20
$q_1 = q_2$	MEAN	.0525	.2567	.0501	.2376
	VARIANCE	.0305	.1197	.0250	.0317
$q_1 = .67$	MEAN	.0328	.2667	.0319	.2559
$q_2 = .33$	VARIANCE	.0261	.1198	.0134	.0611

References

1. WILLIAMS, W.T. & J.M. Lambert (1959) "Multivariate methods in plant ecology. I. Association-analysis in plant communities", Journal of Ecology, 47, 83-101.

2. WILLIAMS, W.T. & J.M. Lambert (1960) "Multivariate methods in plant ecology. II. The use of an electronic digital computer for association analysis", Journal of Ecology, 48, 689-710.

3. WILLIAMS, W.T. & J.M. Lambert (1961) "Multivariate methods in plant ecology. III. Inverse association-analysis", Journal of Ecology, 49, 364-366.

4. MacNAUGHTON-SMITH, P. (1963) "The classification of individuals by the possession of attributes associated with a criterion", Biometrics, 19, 364-366.

5. WILKINS, L.T. & P. MacNaughton-Smith (1964) "New prediction and classification methods in criminology, Journal of Research in Crime and Delinquency, 1, 19.

6. Ibid.

7. GOTTFREDSON, D.M., K.B. Ballard & L. Lane (1963) Association analysis in a prison sample and prediction of parole performance, Vacaville, Calif.: Institute for the Study of Crime & Delinquency

8. BALLARD, K.B. & D.M. Gottfredson (1964) Predictive attribute analysis and prediction of parole performance, Vacaville, Calif.: Institute for the Study of Crime & Delinquency.

9. SINCLAIR, I. & B. Chapman (1973) "A typological and dimensional study of a sample of prisoners", British Journal of Criminology, 13, 341-353.

10. Ibid.

11. Ibid.

12. WILLIAMS, W.T. & J.M. Lambert (1960) "Multivariate methods in plant ecology", J. Ecol., 48, 689-710.

13. MacNAUGHTON-SMITH, P. (1966) "Some statistical and other numerical techniques for classifying individuals", Home Office Research Unit Report NO. 6, London: H.M.S.O., 22.

14. WILLIAMS, W.T. & J.M. Lambert (1960) "Multivariate methods in plant ecology", J. Ecol., 48, 689-710.

15. BLALOCK, H.M. (1972) Social Statistics, New York: McGraw-Hill.

16. MacNAUGHTON-SMITH, P. (1966) "Some statistical and other numerical techniques for classifying individuals", Home Office Research, p. 23.

17. LANCE, G.N. & Williams, W.T. (1965) "Computer programmes for monothetic classification", Computer Journal, 8, 246-249.

18. GOWER, J.C. (1967) "A comparison of some methods of cluster analysis", Biometrics, 23, 623-37.

19. FILDES, R. & D.M. Gottfredson (1972) "Cluster analysis in a parolee sample", Journal of Research in Crime and Delinquency, 9, 2-11.

20. Ibid.

21. LANGE, R.T., N.S. Stenhouse & C.E. Offler (1965) "Experimental appraisal of certain procedures for the classification of data", Australian Journal of Biological Science, 18, 1189-1205.

22. WARD, J.H. (1963) "Hierarchiacal grouping to optimize an objective function", Journal of the American Statistical Association, 58, 236.

23. VELDMAN, D.J. (1967) Fortran programming for the behavioural sciences, New York: Holt, Rinehart & Winston.

24. GROSS, A.L. (1972) "A Monte-Carlo study of the accuracy of a hierarchical grouping procedure", Multivariate Behavioural Research, 7, 379-389.

25. Ibid.

III

STRUCTURES OF QUANTITATIVE DATA GROUPING TECHNIQUES

1. Variables

i) Meaningful Variables

Perhaps the most difficult question taxonomists have to face at an operational level is which variables ought to be included in an analysis. The prime concern of taxonomists relates to measuring degrees of similarity between his operational units, however, it may be argued that this is an impossible task. An absolute assessment of similarity cannot be made since our operations require an arbitrary selection of variables from an almost infinite offering. We can state though, that it is possible to make valid comparisons of operational units if those comparisons are based upon meaningful variables. By "meaningful" we refer to those variables which are theoretically significant to the pivotal concept. Thus, if a researcher is interested in the social ecology of crime, he will consider such things as regional job markets, income distribution, population density, and so forth. He will most likely not consider such variables as the depth of bedrock in an area or its average mean rainfall. Even the raw empiricist selectively chooses the variables he wishes to study. All scientists start with some a priori assumptions; the task is to provide an explicit and logically coherent rationale for those assumptions.

The rationale behind the choice of variables is extremely important for the choice of certain data grouping procedures. Factor analysis, for example, attempts to describe n variables in terms of k underlying factors ($k < n$), by measuring the degree of covariance between groups of variables. In order to provide interpretable results the variables clustered by factor analysis must possess some underlying logical unity. The various questions on an I.Q. test, for example, illustrate that unity, for while each question is different, they supposedly relate to one of three or four basic factors which make up that concept referred to as intelligence.

ii) Invalid Variables

Variables which do not have more than one state or level of measurement are said to be invariant. Invariant variables have no place in the analysis and may actually be a hindrance by necessitating unnecessary computation and by taking up valuable computer storage.

iii) Correlated Variables

Variables which are known to be highly correlated ought also to be eliminated from the data matrix. They provide only redundant information which adds no further insight to the analysis. If the correlations between most of the variables within the data matrix is known, then the

entire exercise becomes tautologous since the researcher is simply grouping known groups -- the exercise thus becomes trivial. There are also several technical reasons why variables known, or suspected to be correlated, ought to be eliminated from the data matrix. The first, which we have mentioned earlier, is the unnecessary use of computational time and computer storage space. A second reason stems from the fact that singular correlational matrices will be produced which prevent the use of certain techniques such as regression analysis.

Singular matrices occur when column vectors within the data matrix are either identical, or are linear transformations of one another. If, in a data matrix, column 4 is simply a replica of column 1 multiplied by a scalar constant, then for all intents and purposes, column 4 adds no information to the analysis since it is a perfect co-relation of column 1. Singular matrices, or non-independant column vectors as they are also known, may cause computational problems for the researcher, since their non-independance may not be visibly evident. Linear combinations of two or three vectors may provide the information contained in a fourth vector, thus causing the matrix to be singular. Needless to say, the more variables included in an analysis, the greater the

probability this problem will arise, since the probability of producing spurious linear combinations is increased, especially when the data has a limited range, such as in the case of dichotomous data.

iv) Measurement, scaling and coding

Some of the most frustrating problems facing social scientists in operationalizing variables relate to measurement, scaling and coding data.[1,2] Since Stevens'[3] early exhortations, scientists have become increasingly aware of the necessity of relating the correct statistic to the level at which the data is measured. Thus for nominal data such statistics as X^2, Fisher's Exact, and Yules Q, have been developed. Likewise for ordinal level data one has recourse to rank order correlation coefficients, Gamma, and Kendal's Taus. Interval and ratio level data allow the use of the vast multitude of "high powered" parametric statistics as well as the more crude, non-parametric statistics when the data is categorized. The consistent low level of measurement with which criminologists, and indeed most social scientists, are forced to work with has probably been the largest single factor in restricting the use of quantitative data grouping techniques. Researchers in other fields such as biology and geography, may be considered fortunate

in that much more of their raw data is measured at higher levels of measurement.

One particularly useful tactic available to social scientists is the transformation of nominal data with multiple categories into "dummy variables".[4,5,6] Ordinal level data, for example, can be converted into nominal without any loss of information by using more than one variable to code for the present single variable. As an illustration, if the variable "marital status" were coded 1 - single, 2 - married, 3 - divorced, and 4 - other, the following transformation would be used:

Subject	Original Code	Dichotomous Code			
		I	II	III	IV
A	1	1	0	0	0
B	3	0	0	1	0
C	2	0	1	0	0
D	1	1	0	0	0
E	4	0	0	0	1

Single would now be represented by a 1 under variable I, married by a 1 under variable II, divorced and other by similar representations under variables III and IV. This transformation will later be shown to be extremely useful in allowing us to use monothetic classification models on multistate data and permitting the use of many similar

level data.

v) How many variables

If we assume a data matrix M, with n rows and m columns, we can ask questions relating to three dimensions of that matrix. What, if any, is the relationship between the proportion of rows to columns?

Some techniques (for example Ward's H-Group) are known to be optimal when the number of cases (n) is greater than 100. It is difficult, however, to specify a general optimal sample size since both the number of variables and the actual cluster algorithm have to be taken into account. Usually, there are no problems relating to too many cases; it is too few cases which often creates difficulties, although practical considerations such as the amount of storage available in computer facilities must also be taken into account. Sample sizes which are too small may not allow for "natural" population clusters to appear in the analysis especially when there are known deviant or minority cases in the population. Small sample sizes may also create computational difficulties for cluster techniques based on similarity coefficients with significance level cut-off criteria. The researcher may be faced with extremely low or no significance levels on variables which are normally quite highly correlated

in the population.

The number of acceptable columns (m), on the other hand, presents different problems. It is usually too many variables which we have to guard against. The larger the size of m in a data matrix, the less likely significant variables will be allowed to make themselves evident (this is similar to the signal to noise ratio problem faced in signal detection theory). As well, the more variables present, the higher the probability of a singular matrix occurring due to spurious combinations between column vectors. Because of the problem of singular matrices, it is also necessary than $n > m$. Where $n \ll m$ it is invariable that the matrix[7] is singular.

2. Q vs R Groupings

There are two major methods of grouping an n x m data matrix. One can group attributes which are generally defined as columns (m) in the data matrix, or one can group individuals or cases which are customarily represented as rows (n) in the data matrix. The grouping of variables or attributes is referred to a 'R' type clustering, and this is by far the most common strategy invoked. 'Q' type clustering, on the other hand, is the grouping of individual cases across several variables. As will be indicated later,

Q-type clustering imposes greater restraints on the typologist than does the R-type. The distinction between R and Q analysis has its origins in factor analysis, however, more and more literature is appearing in which this distinction and nomenclature is being generalized to all grouping procedures.

Generally, the same algorithm may be used for either Q or R type clustering; the researcher simply transposes his data matrix prior to calculating his correlation or similarity matrix.

3. Similarity Coefficients

Ultimately, the aim of any cluster or grouping technique is to assess the degree of similarity or likeness (or conversely, dissimilarity) between the elements of a given set of variables and to group those elements which are the most alike. In order to accomplish this task, criteria and procedures of assessing similarity must be established. Fortunately the process of determining similarity is not unique to typologies.[8] Analytic statistics is based upon concepts of similarity of groups and degrees of variance both between and within groups. By borrowing some of the techniques already devised to assess likeness for the purposes of hypothesis testing, etc., an initial matrix

of similarity can be established. This matrix of similarity is defined as an n x n or an m x m matrix of similarity coefficients between those elements the researcher wishes to group. In the case of grouping attributes, for example, the matrix of similarity coefficients (m x m) will consist of ½m (m-1) correlation coefficients between all m variables. Only the upper portion of the matrix, to the right of the diagonal, is used since it is assumed that the correlation between m_i and m_j is the same as between m_j and m_i.

Following Sokal and Sneath[9], and Cormak[10], three basic types of similarity coefficients may be listed: (1) coefficients of distance, (ii) association coefficients, and (iii) correlation coefficients.

The choice of which type of similarity coefficient to choose will be based on two basic considerations -- statistical and substantive. Statistically, we will take into consideration such things as the level at which the data is measured. Substantive issues will revolve around theoretical concerns relating to criminology. For example, it may be acceptable to utilize distance coefficients and ecology studies where variables such as age distributions and unit area are compared, however, distance coefficients between personality variables do not offer the same degree of satisfaction, and certainly most investigators would

be more responsive to the use of association coefficients.

A great deal of literature exists which explores the various properties of similarity coefficients. Goodman & Kruskall[11,12,13], Sokal and Sneath[14], Moore and Russell[15], and Cheetham and Hall[16] are good points of entrance to this literature. For the purpose of this study it is sufficient to touch only upon some of the major similarity coefficients and their variations, which are currently the most prevalent in the literature. For more detailed accounts of these coefficients as applied to taxonomy, Sokal and Sneath[17] and Cormack[18] are good references as is the aforementioned literature relating to similarity coefficients per se.

i) Distance Coefficients

Coefficients based upon the Euclidean distance between points tend to have a great deal of intuitive appeal amongst taxonomists. This appeal probably stems from two sources: the ease with which distance can be represented graphically and the familiarity of the concept due to its ubiquity in "lease squares" models of regression analysis. Quite simply, if we have two or more individuals measured on two variables, say X_1 and X_2, the degree of similarity between those individuals is the absolute distance between their coordinates as defined on the plane X_1, X_2. If any two individuals have identical co-

ordinates then the distance is 0, or perfect correlation. Contrariously the maximum degree of dissimilarity would be the greatest distance on the X_1, X_2 plane which would be defined as:

$$d_{max} = \left[(X_{1n})^2 - (X_{2n})^2 \right]^{\frac{1}{2}},$$

Where X_{1n} and X_{2n} are defined as the maximum values measurable for X_1 and X_2. Essentially, this defines the vector which forms one of the diagonals of the plane X_1, X_2. Using simple distance creates problems, however, since both positive and negative values are permissible. Modifications to this procedure have been developed, one of which is to use the absolute value of the distances. Thus, to determine the average distance between n characteristics of a pair of individuals, the formula to be used would be:

$$\frac{1}{n} \sum_{i=1}^{n} \left| X_{ij} - X_{ik} \right|$$

This general formula has been used as long ago as 1909 by Czekanowski[19,20]. Haltenorth[21] used it to distinguish between types of cats and more recently it has been employed by Cain and Harrison[22] and Johnson and Wall[23] in biological taxonomy.

A further modification, and one which is currently quite popular is to summate squared distances, thus the formula:

$$\frac{1}{n} \sum_{i=1}^{n} (X_{ij} - X_{ik})^2$$

As Cormack[24] indicates, these formulae can be represented by a more general representation:

$$\sum_{i=1}^{n} W_i \ (X_{ij} - X_{ik})^2 \qquad \text{with,}$$

$W_i = 1$ for unstandardized data

$W_i = 1/S_k^2$ for data standardized by standard deviation (this is usually denoted in the literature as Δ^2)

$W_i = 1/\max_{j,k} (X_{ij} - X_{ik})^2$ for data standardized by the range

This concept of Euclidian distance can be generalized to cover any number of points within an N dimensional space. The ensuing equation is a derivative of a class of equations referred to in mathematics as Minkowski metrics. The general formula for distance (d or Δ) between two points, i,k is

$$\Delta_{ik} = \left[\sum_{i=1}^{n} \left| X_{ij} - X_{ik} \right|^{\lambda} \right]^{1/\lambda}$$

Lance and Williams[25] present several variations of distance coefficients based on Minkowski metrics. The one which appears to have struck the most responsive chord in the literature is their Canberra metric (d_{can}), which is defined as:

$$d_{can}{}_{(j,k)} = \sum_{i=1}^{n} \left(\frac{|X_{ij} - X_{ik}|}{(X_{ij} + X_{ik})} \right)$$

This metric has the advantage that it is not affected by the entire range of characters. Also, it is responsive to proportional instead of absolute differences between points.

Obviously, extensive elaborations on the theme of distance coefficients are possible, and, as a perusal of the literature will indicate, any comprehensive coverage of these coefficients would fill a separate monograph.

ii) Association Coefficients

As with distance coefficients, the proliferation of association coefficients would appear almost endless. Fortunately, many of these coefficients are highly related on a conceptual level.

Association coefficients are usually applied to nominal and ordinal level data. Data measured at higher levels of measurement can be coded to be made congruent

with association coefficients however, the researcher must recognize that a certain amount of information will be lost through the necessary categorization procedure. It has been the practice of most taxonomists using this process to create nominal dichotomous level data (0,1 or presence-absence) which can be represented by a 2 x 2 contingency table.

		Variable j	
		1	0
Var. i	1	a	b
	0	c	d

a + c b + d n = a + b + c + d

Consistent pairs C = a + d

Inconsistent pairs I = b + c

n = C + I

In this table cells a and d represent consistent pairs, that is, individuals who either possess both attributes or lack both attributes. Cells b and c represent inconsistent or mismatched pairs. These individuals score positively on one attribute but negatively on the other. We will adopt the convention of using 'C' to indicate the consistent pairs a + d, and 'I' to represent the inconsistent cells b + c. The terms consistent and inconsistent

have been adopted as a convention since 0,1 may represent presence-absence, positive and negative, or any other similar dichotomy (high-low, yes-no, etc.)

One of the earliest used association coefficients is that suggested by Jaccard[26] in 1908. The Jaccard coefficient may be defined as:

$$J = \frac{a}{a + I}$$

J equals unity only when I is 0, thus a Jaccard index of 1 indicates no mismatched or inconsistent pairs. The greater the number of inconsistent pairs, however, the closer J will converge to zero. Unfortunately, Jaccard's coefficient considers only positively associated pairs or those on the upper portion of the consistency diagonal. Those pairs which lack specific attributes are not considered in the equation. This association coefficient has been used primarily in biological taxonomy by Sneath[27], although adaptations by Dice[28] and Sørensen[29] which gives more weight to the consistent pairs (cell a) have been developed. Watson, Williams and Lance[30] have used the obverse of the Jaccard coefficient measuring dissimilarity rather than similarity, which they term the "non metric" coefficient. This they define as

$$\frac{I}{(2a + I)} \qquad I = b + c$$

Apparently, the exclusion of negative matches is a contentious issue in numerical taxonomy[31] in biology since in most cases the absence of an attribute gives little information to a taxonomist. A similar situation may arise in criminology when the criminologist is faced with the correlation of small subsamples. If one were attempting to classify inmates in a correctional institution, for example, it would more than likely be found that there are very few murderers and psychopaths in a given population. If the decision of whether to group a diagnosis of psychopathology with murder arose, it would probably not aid to know that most of the population is neither psychopathic nor are they murderers. The hypothetical data shown below illustrates the problem.

		PSYCHOPATHOLOGY		
		Yes	No	
MURDERER	Yes	10	15	25
	No	20	55	75
		30	70	100

Most association coefficients such as the simple matching method would provide a high correlation (in this case .65), implying a high correlation. The Jaccard method on the other hand, would produce a coefficient of

.29, which in this instance, provides a more realistic indicator of association between murder and psychopathology.

The simple matching coefficient illustrated below, takes into account both cells which represent consistencies and calculates the consistencies as a proportion of the total sample.

$$S_m = \frac{C}{n} = \frac{a + d}{a + b + c + d}$$

In so doing the coefficient will have a range from 0 to 1; 1 representing no inconsistencies and 0, of course, a dearth of consistent pairs. With the exception of cases similar to the murder x psychopath example, we can expect coefficients which take both cases of consistency into account to be generally much more useful.

Probably the association coefficient best known by criminologists in reference to typing, is chi square (X^2). Based mainly on the work of Williams and Lambert[32,33,34] and MacNaughton-Smith[35], this statistic has become one of of the taxonomers' mainstays. Although the general formula for X^2 is defined as

$$X^2 = \sum_{i=1}^{n} \frac{(f_o - f_e)}{f_e}$$

where f_o = observed frequency

f_e = expected frequency

n = no. of cells

a special computational formula exists, applicable only in the case of 2 x 2 tables. Using our conventional labeling

$$X^2 = \frac{n\ (ad-bc)^2}{(a+b)\,(c+d)\,(a+c)\,(b+d)}$$

It is quite clear that from the above formula, that if ad equals bc the X^2 value will be 0. Thus a zero X^2 will appear if the number of consistencies and inconsistencies exactly equal each other. MacNaughton-Smith[36] has written a fairly complete expose on the use of the X^2 statistic in reference to typology construction. We would refer the reader to that article for further details, however, it might be added that a few modifications of X^2 have been employed by typologists. As previously mentioned Lance and Williams[37] have made the rather interesting observation that the X^2 statistic is N times the correlation coefficient r when the data are coded 0,1. It is also interesting to note that X^2 divided by N produces the Phi-coefficient (ϕ). Hence, given nominal dichotomous data coded 0,1

$$\frac{X^2}{N} = \phi \doteq r$$

This modification yields an association coefficient which varies from 0 to 1 as opposed to the non-standardized

distribution of X^2. Yet another modification based on X^2 is the contingency coefficient C suggested by Carl Pearson. C is defined as:

$$C = \left(\frac{X^2}{X^2 + N} \right)^{1/2}$$

As with other X^2 related statistics C equals zero when there are no consistent pairs. The upper limit of C varies with the number of rows and columns, however, for 2 x 2 contingency tables $\max_C = .707$. Thus, a 0 to 1 range transformation may be effected by dividing C by .707.

iii) Correlation Coefficients

The Pearson product-moment correlation coefficient has had a history of extensive, albeit rather specialized usage in data grouping techniques. The correlation matrix used in factor analysis from which generalized factors are extracted, is essentially a matrix of Pearson correlation coefficients. Williams and Lambert[38], and Gower[39] have used both the correlation coefficient r and multiple correlation which is an elaboration of the concept in Association Analysis. The basic formula for the Pearson product-moment correlation coefficient (r) is:

$$r = \frac{\sum_{i=1}^{n} (X_{ij} - \bar{X}_j)(X_{ik} - \bar{X}_k)}{\left[\sum_{i=1}^{n} (X_{ij} - \bar{X}_j)^2 \sum_{i=1}^{n} (X_{ik} - \bar{X}_k)^2 \right]^{1/2}}$$

which, for nominal dichotomous data coded 0,1 may be computed as:

$$r_{ij} = \left[\frac{(ad-bc)}{(a+b)(a+c)(c+d)(b+d)} \right]^{\frac{1}{2}}$$

The statistic r is usually used for continuous interval and ratio level data although it is being increasingly utilized with dummy variables (0,1). The correlations range from -1 to +1 which may cause problems in certain kinds of cluster formations as will be indicated later. Most researchers, when using r tend to standardize the rows of the data matrix since r is sensitive to variables being measured on widely varying scales.

4. Sorting Strategies

Once an index of similarity has been selected, there are three basic strategies which may be employed for sorting the data. The first strategy is the divisive procedure. This involves the division of large groups of heterogeneous data into smaller, more homogeneous sub-units. Effectively, this design produces a dendogram with the initial heterogeneous sample at the top and smaller, more homogeneous units at the bottom. Williams and Lambert's association analysis employs this kind of sorting strategy.

Tactically, divisive sorting strategies usually

divide the sample on the basis of one key or critical variable. If this variable is coded dichotomously, the sorting strategy is termed monothetic; if it has more than two categories, it is termed polythetic. The study of Sinclair and Chapman is an illustration of a monothetic division and that of Fildes and Gottfredson a polythetic division. Once a division has been made, the ensuing sub-groups are treated as sample wholes and are themselves re-divided on the basis of some critical or criterion variable. This procedure of dividing sub-groups usually continues until the sub-groups no longer contain a statistically significant eritical variable. At this point they are considered sufficiently homogeneous to be termed "natural clusters".

A second sorting procedure commonly employed in data grouping is the agglomerative technique. While divisive techniques may be considered to produce a descending dendogram, agglomerative techniques produce ascending dendograms. Principally, this strategy commences with N individuals or groups, searches the sample for the two most similar individuals or groups and combines them. This combinatorial procedure continues, searching for the most similar groups as defined by the similarity statistic and combining them until ultimately, all groups have been combined.

The critical issue involved in agglomerative techniques revolves around what the characteristics of the new group should be if two or more individuals are combined with dissimilar profiles. Should the individuals which constitute the group be combined to produce some arithmetic average for the group? Or, should the smallest, or conversely the profile of the highest individual be chosen as representative of the group.

Procedures which characterize the group with some mean value are often referred to as centroid procedures. When the lowest value is chosen, the term nearest neighbour is employed, and when the largest or most deviant value is chosen, the term furthest neighbour is employed.

Several researchers, though mostly biologists, experimented with all three methods and studies by Wishart[40], McQuitty[41], and Sokal and Michner[42] provide examples of clustering techniques using the nearest neighbour, furthest neighbour, and group average procedures respectively. The major difference between the three techniques is that the furthest neighbour procedure produces tight, close-proximity clusters. The nearest neighbour procedure provides more drawn-out clusters and the averaging procedure, produces a distribution somewhat halfway between the two aforementioned.

A third strategy exists for sorting and this may be generally termed the recursive method. The major difference between this sorting strategy and the other two is that recursive methods generally allow individuals to be shifted from one group to another if their profile becomes so dissimilar to their initial groupings that they may be deemed misclassified. Thus, if an individual is assigned to a group early in a sorting procedure, subsequent sortings might lead to the condition where he would best fit into another, more 'similar' grouping.

Generally, recursive sorting procedures have received little attention from researchers since the problems involved in providing a viable algorithm, and the amount of time required to check all individuals with regards to grouping suitability appear to be quite formidable. Some suggestions have been put forward, however, the most viable of which appears to be a proposal to select a number of centroids from the data and to recursively group the data around those points.

Ball and Hall[43] propose such a centroid-based technique which they term ISODATA. Essentially, the ISODATA routine follows seven steps:

1) a "typical" set of response patterns are selected to be used as "cluster points".

2) points lying closest to the centres are grouped by selecting the shortest Euclidean distance (Ball and Hall suggest one of the Minkowski metrics -- minimum squared deviations from the centre).
3) if the within group variance exceeds a specified value, 0_E, for any group, that group is split in two.
4) the patterns are regrouped using the new cluster points, and the average response pattern found.
5) the distance 'd' is computed between all pairs of average response patterns.
6) groups which have a combined variance less than 0_C are combined.
7) the entire process is iterated.

The principal reason for investigating recursive sorting strategies lies in the assumption that a significant number of cases may be misclassified at early points in the classification procedure. A recursive sorting strategy would allow for "classification updating" at later points in time.

Studies by Cochran & Hopins, Lange, Stenhouse & Offler, and Gross, appear to indicate that the problem of

misclassification may not be so great as to warrant the programming expense of recursive techniques. While the actual degree of misclassification is technique specific, evidence from the above studies indicates that our present techniques are fairly accurate when dealing with adequate sample sizes and a few states. In general, the probability of misclassification is inversely proportionate to sample size, and proportionate to the number of multivariate states possible.

Thus, it would appear that the value of developing recursive sorting strategies lies in the inability of other methods' abilities to handle small samples and samples with many variables.

5. Characteristics of a Good Typology

Any criteria for distinguishing good typologies and typological schema must address themselves to the purpose for which the typology is to be constructed. Ultimately, the success or acceptability of a typology rests on how well it meets the purpose for which it is constructed. Certain basic characteristics or criteria can be established, however, by which typological schema can be evaluated independently of the goals toward which they are directed. Literature by Deutsch[44], Driver[45],

Fisher[46], and Hempel[47] is perhaps the most pertinent in this area.

Basically, there are two major sets of criteria through which any typology may be assessed. There are macroscopic criteria which are concerned with the philosophical and economic efficiency of typologies, and microscopic concerns which relate to specific logical, structural and functional aspects of typologies.

On a macroscopic level, we may ask what is the purpose of pivotal concept behind the typology? Is, for example, the typology addressing itself to crimes, criminals, or criminality? It is imperative that the purpose of a typology be explicitly stated for two reasons: first, it is the definition goal the typologist is attempting to achieve and second, a precise knowledge of the pivotal concept helps the typologist determine whether the variables he is measuring are pertinent to the task he is pursuing.

A second question we might ask of any typology is whether it is being constructed primarily as a descriptive or a predictive tool. Some authorities would deny that description can or ought to be separated from prediction. As an example, Tyriakian states,

> "...(a) typological goes beyond sheer description by simplifying the ordering of

> the elements of a population, and the known relevant traits of that population, into distinct groupings; in this capacity a typological classification creates order out of the potential chaos of discrete, discontinuous or heterogeneous observations. But in so codifying phenomena, it also permits the observer to seek and predict relationships between phenomena that do not seem to be connected in any obvious way. This is good because a good typology is not a collection of undifferentiated entities, but is composed of a cluster of traits which do in reality 'hang together'."[48]

This argument may be valid, however, it does seem valuable to make a distinction. The differentiation should perhaps not lie strictly between description and prediction, but rather between known predictions (or perhaps more properly 'relationships') and new predictions. Certainly, most individuals who would use a descriptive typology would do so because they can be assured that if an individual fits into a certain category, he will possess certain determined attributes. Functionally, this purpose is quite different from that of making new predictions.

Under these circumstances, the typology would be valued for its heuristic qualities.

On a more structural level, we can ask how good is a typology's organizing power? That is, how many categories does a given typology have, and is the categorized data grouped in such a fashion that both within group homogeneity and between group heterogeneity are maximal? If other grouping systems can perform this task more effectively, then the technique presently in use should be reconsidered. Similarly, the range of the typology should be examined. Ideally, a good typology provides predictions concerning the realtionships between many variables. In general terms, the more variables a typology encompasses and speaks to, the greater the number of situations in which it will be applicable.

From an economic standpoint, typologies may be measured on two 'cost factors'. First, we should ask, what is the cost involved in choosing this system over another? More specifically, we may question whether the information provided by a classification system warrants the man-hours involved in data collection and collation. From a theoretical perspective some typologies are considered to have higher social costs than they provide benefits. One prime example is the "labelling theory"

effect of classifying people as criminals or juvenile delinquents. It could be argued that the labelling process creates such a stigma for the labelled person that the stigmatization process pushes a person into deviancy, thereby creating a self-fulfilling prophecy.

The second economic question we may direct toward a classification schema and one that is somewhat akin to the last question, is what are the social and psychological costs incurred through misclassification? Obviously an ideal classification schema should be so designed that it is completely unambiguous, however- unforseen data perturbations and human error may cause some cases to be wrongly classified. Misclassifying a person as a murderer may, in some jurisdictions, lead to the ultimate personal cost of extinction. Similarly, the misclassification of a dangerous individual may lead to grave consequences for society at large.

These economic cost factors are especially important for practitioners whose task it is to apply social policy.

On a microscopic level we are more concerned with the logical structure of a typology. First of all, are the categories of a typology mutually exclusive or do they overlap? If categories overlap, then how does one decide

in which cluster a subject belongs? When categories overlap it might be questioned whether these categories might not better be viewed as a continuum rather than as discrete entities.

Second, is the typology exhaustive? Does it account for all cases, or are a large proportion of cases unclassifiable or relegated to some nebulous "other" category? Needless to say, the more occurrances a typology can account for, the greater is its value to potential users. A point somewhat akin to this second question is the typology's level of performance in including new data. Typologies with a high degree of universality are generally more preferable to those which are sample specific. This criterion closely follows the general tenet of science which prefers "general laws" to "specific rules".

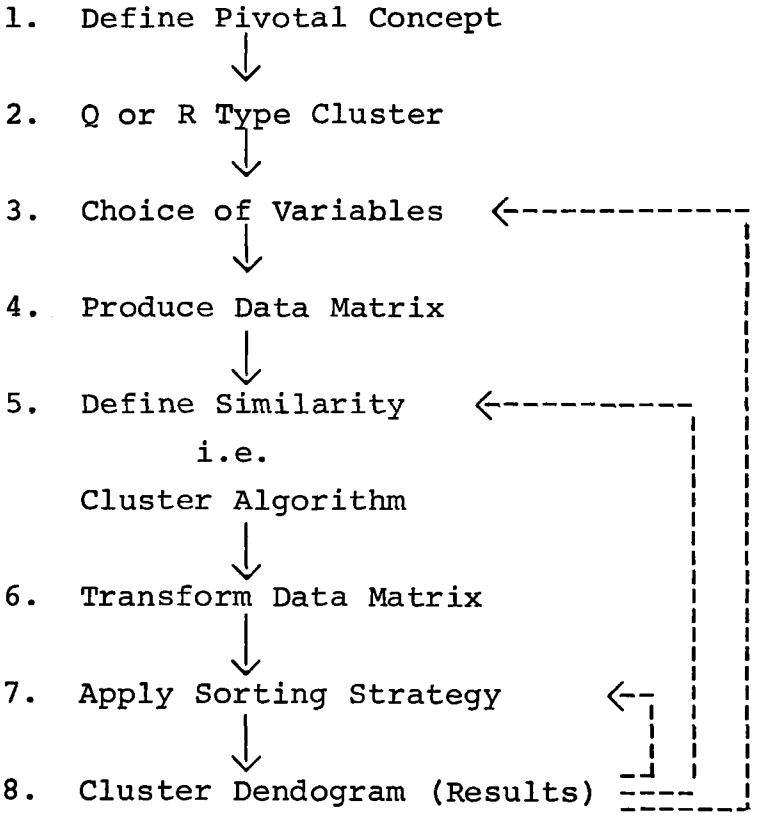
6. Conclusion

Having examined the individual components of quantitative data grouping techniques, it is now essential to see how they fit together to form a unified whole. Figure 2 is a graphic illustration of the logical steps involved in the application of a quantitative data grouping method.

The first step involves the selection and

FIGURE 2

COMPONENTS OF DATA GROUPING TECHNIQUES



definition of a pivotal concept. The pivotal concept is essentially what the ensuing typology is about. In criminology, the pivotal concept may be such things as crimes, criminals, or institutions in which cases one would be clustering types of behaviour, people or jails and prisons. The explication of the pivotal concept is crucial since it dictates the kinds of variables which will be measured at the data gathering stage.

The next step, which is usually performed in conjunction with the definition of the pivotal concept, is to decide whether a 'Q; or an 'R' type grouping is required. Essentially, this is a decision as to whether it is units (people or objects) or variables which are to be clustered. If, for example, the pivotal concept is defined as criminals, then the researcher must state whether or not he is interested in grouping criminals per se , or whether he is looking for commonalities in the variables which characterize those individuals. Grouping units is referred to in 'Q' type grouping, and the grouping of variables is 'R' type.

Once the decision is made as to the pivotal concept, the researcher then decides which variables he will require in his investigation. The choice of which variables to choose rests mainly upon theoretical

considerations, since the researcher must choose from an almost infinite offering. Thus, it is the researcher's theoretical perspective which guides him to those variables which ought to potentially be the most fruitful.

After deciding on which variables are to be measured, the researcher proceeds to the fourth step which is the actual data collection and the compilation of a data matrix. Normally, in a 'Q' type analysis, rows represent units and columns represent variables, however, for 'R' type analysis, this matrix may be inverted. As has been pointed out earlier, care must be taken to ensure that the number of rows exceeds the number of columns, regardless of the type of analysis to be performed.

The fifth step in utilizing a quantitative data grouping technique is to provide a definition of similarity. Operationally, this involves the selection of a cluster similarity algorithm. As has been indicated, the researcher is faced with a great number of existant algorithms to choose from as well as being at liberty to construct his own.

Effectively, the similarity algorithm and the sorting strategy are the two most important aspects of any cluster technique since they form the core upon which

the technique is based.

Having selected a definition of similarity, the researcher proceeds to transform his raw data matrix into a matrix of similarity coefficients, employing the similarity algorithm. Again, if the correlations are between subjects, the procedure is referred to as an 'Q'-type technique and if the correlations are between variables, it is a 'R'-type technique.

Once the raw data has been transformed into a matrix of similarity coefficients, the decision must be made as to how like units are to be grouped. This, then involves the seventh step -- the selection of a sorting strategy. Next to the similarity coefficient, the sorting strategy is probably the most important component of a data grouping technique. Not only can a sorting strategy decide group membership, but the spacial distribution, hence, group cohesiveness is greatly effected by that strategy.

The final phase in employing a quantitative data grouping technique is to present the results. This usually involves the construction of a tree or dendogram through which the results are graphically presented.

Thus, we have outlined the eight basic steps involved in applying a quantitative data grouping technique

to a set of data. Needless to say the interpretation of the significance of the ensuing clusters remains with the researcher, and rests heavily upon his knowledge of his data, and his theoretical acumen. As with most scientific endeavours, several value judgments are required by the researcher at various steps in the application of the cluster procedure. Should the results he obtains be questionable, the researcher may wish to return to his cluster procedure and modify certain components.

Again, referring to Figure 2, the researcher may wish to re-examine three of the more critical areas of the technique with which he has been working. He may refer back to his sorting strategy if he feels that the size and shape of his distributions are awry; he may return to his cluster algorithm if he feels that "similar" units are not being properly adjoined, or he may question the relevance of the typology as a whole and return to his data, and his choice of variables.

There are certainly no immutable rules for the application of cluster techniques. Their value and efficacy rests heavily upon the researcher's feeling for his data, theoretical strength, and experience in the application of various techniques. The continued usage of quantitative techniques, along with evaluative studies can assist the

researcher, however, in enabling him to make more informed judgments in both the application of clustering techniques and in the selection of individual components within those techniques.

References

1. BLALOCK, H.M. (1972) Social statistics, New York: McGraw - Hill.

2. KERLINGER, F.N. (1972) Foundations of behavioural research, 2nd ed., New York: Holt, Rinehart & Winston.

3. STEVENS, S.S. (1946) "On the theory of scales of measurement", Science, 684, 677-680.

4. SUITS, D. (1957) "Use of dummy variables in regress ion equations", Journal of the American Statistical Association, 52, 548-551.

5. COX, D.R. (1970) The analysis of binary data, London: Methuen.

6. BLALOCK (1972) op. cit., 489-502.

7. Van De GEER, J.P. (1971) Introduction to multivariate analysis for the social sciences, San Francisco: Freeman

8. GOODMAN, L.A. & W.H. Kruskal (1963) "Measures of association for cross classification III: Approximate sampling theory", Journal of the American Statistical Association, 58, 310-364.

9. SNEATH, P.H.A. & R.R. Sokal (1973) Numerical Taxonomy, San Francisco: W.H. Freeman & Co.

10. CORMACK, R.M. (1971) "A review of classification", Journal of the Royal Statistical Association, A131, 321-367.

11. GOODMAN, L.A. & W.H. Kruskal (1954) "Measures of association for cross classifications", Journal of the American Statistical Association, 49, 732-764.

12. GOODMAN, L.A. & W.H. Kruskal (1959) "Measures of association for cross classification II: Further discussion and references", Journal of the American Statistical Association, 54, 123-163.

13. GOODMAN, L.A. & W.H. Kruskal (1963) "Measures of association for cross classification III: Approximate sampling theory", Journal of the American Statistical Association, 58, 310-364.

14. SOKAL, R.R. & P.H.A. Sneath (1963) Principles of numerical taxonomy, London: W.H. Freeman & Co.

15. MOORE, A.J. & J.S. Russell (1967) "Comparison of coefficients and grouping procedures in numerical

analysis of soil trace element data", Geoderma, 1, 134-158.

16. CHEETHAM, A.H. & J.E. Hazel (1969) "Binary (presence-absence) similarity coefficients", Journal of Paleontology, 43, 1130-1136.

17. SNEATH & Sokal (1963) op. cit.

18. CHEETHAM & Hazel (1969) op. cit.

19. Referenced in Sneath & Sokal (1973) op. cit.

20. Referenced in Sneath & Sokal (1973) op. cit.

21. Referenced in Sneath & Sokal (1973) op. cit.

22. CAIN, A.J. & G.A. Harrison (1958) "An analysis of the taxonomist's judgement of affinity", Proceedings of the Zoological Society of London, 131, 85-98.

23. Referenced in Sneath & Sokal (1973) op. cit.

24. CORMACK (1971) op. cit., 325.

25. LANCE, G.M. & W.T. Williams (1967) "A general theory of classification strategies I: Hierarchical systems", Computer Journal, 9, 373-380.

26. Referenced in Sneath & Sokal (1973) op. cit.

27. SNEATH, P.H.A. (1957) "Some thoughts on bacterial classification", Journal of General Microbiology, 33, 184-200.

28. DICE, L.R. (1945) "Measures of the amount of ecological association between species", Ecology, 26, 297-302.

29. Referenced in Sneath & Sokal (1973) op. cit.

30. WATSON, L., W.T. Williams & G.N. Lance (1967) "A mixed-data numerical approach to angiosperm taxonomy: The classification of Ericales", Proceedings of the Linnean Society of London, 178, 25-35.

31. SNEATH & Sokal (1973) op. cit., 132.

32. WILLIAMS, W.T. & J.M. Lambert (1959) "Multivariate methods in plant ecology. I. Association-analysis in plant communities", Journal of Ecology, 47, 83-101.

33. WILLIAMS, W.T. & J.M. Lambert (1960) "Multivariate methods in plant ecology. II. The use of an

electronic digital computer for association analysis", Journal of Ecology, 48, 689-710.

34. WILLIAMS, W.T. & J.M. Lambert (1961) "Multivariate methods in plant ecology. III. Inverse association-analysis", Journal of Ecology, 49, 717-729.

35. MacNAUGHTON-SMITH, P. (1963) "The classification of individuals by the possession of attributes associated with a criterion", Biometrics, 19, 364-366.

36. MacNAUGHTON-SMITH, P. (1966) "Some statistical and other numerical techniques for classifying individuals", Home Office Research Report No. 6, London: H.M.S.O.

37. LANCE, G.N. & W.T. Williams (1965) "Computer programmes for monothetic classification", Computer-Journal, 8, 246-249.

38. WILLIAMS & Lambert (1960) op. cit.

39. GOWER, J.C. (1967) "A comparison of some methods of cluster analysis", Biometrics, 23, 623-637.

40. WISHART, D. (1969)" Mode analysis, a generalization of nearest neighbour which reduces chaining effects",

A.J. Cole (ed.) Numerical taxonomy, 1969, London: Academic Press.

41. McQUITTY, L.L. (1967) "A novel application of the coefficient of correlation in the isolation of both typal and dimensional constructs", Education and Psychological Measurement, 27, 591-599.

42. SOKAL, R.R. & C.D. Michner (1967) "The effects of different numerical techniques on the phenetic classification of bees of the Hoplites complex (Megachilidae)", Proceedings of the Linnaean Society of London, 178, 59-74.

43. BALL, G.H. & D.J. Hall (1967) "A clustering technique for summarizing multivariate data", Behavioural Science, 12, 153-55.

44. DEUTSCH, K.W. (1966) "On theories, taxonomies and models as communication codes for organizing information", Behavioural Science, 11, 1-17.

45. DRIVER, E.A. (1968) "A critique of typologies in criminology", Sociological Quarterly, 9, 356-373.

46. FISCHER, W.D. (1969) Clustering and aggregation in economics, Baltimore: Johns Hopkins Press.

47. HEMPEL, C.G. (1952) 1969, Fundamentals of concept formation in empirical science, Chicago: University of Chicago Press.

48. TIRYAKIAN, E.A. (1968) "Typological classification", International Encyclopedia of the Social Sciences, 1968, New York: Macmillan, 185.

IV

METHODOLOGY

1. Purpose

In this study we will be comparing two quantitative data grouping techniques. The first procedure -- association analysis -- was chosen because it best represented a divisive grouping technique using a summated distance function (the chi-square statistic) as a criterion for similarity. The other factor weighing heavily in the decision to use association analysis was its relatively high degree of popularity amongst criminological investigators. As our review of the criminological literature indicated, association analysis has been used as if it were the exclusive data grouping technique available to criminology.

The second procedure employed in the analysis was Ward's hierarchical grouping technique. This is an agglomerative technique which also uses a general distance function as its criterion for similarity. The H-group procedure was chosen more for its conceptual simplicity and because it is perhaps the best representative of an agglomerative technique rather than for its popularity amongst criminological investigators.

By comparing the outcome clusters produced by both techniques, we will effectively be looking at the validity of the data grouping procedures. A high degree

of similarity between corresponding clusters produced by the two techniques will allow us to say with some certainty that the ensuing groupings represent natural clusters within the data.

A low level of correspondence, however, would lead us to question our basic assuptions. It may be that one or both of our techniques are incapable of extracting existant clusters. Further, it may be questionable as to whether 'natural clusters' in fact exist within the data.

A high degree of correspondence between techniques does not in itself negate these two latter possibilities however, since the correlation may result from either a chance distribution of the data, or an artifact within the grouping procedures. A high correlation though does give greater credence to accepting the grouping techniques' outcomes.

2. Techniques Used in This Study

Association analysis as indicated earlier, is a divisive grouping technique. The chi-square statistic is used as a measurement of similarity (or more appropriately in this instance, dissimilarity), in order to provide a criterion for dividing groups. The variables upon which

the sample is measured are entered into a square matrix which consists of the chi-square values obtained for each pair of variables. The X^2 values are summated down each column to provide a total X^2 value for each variable in relation to all other variables. Since the X^2 statistic is a distance function, the highest X^2 value obtained by summating the matrix columns ought to indicate which variable is most dissimilar to all other variables (that is, has the greatest summated Euclidian distance).

The sample is divided into two subgroups on the basis of the variable most dissimilar to all other variables. After the initial division, the procedure is repeated on each subgroup. This process continues until, for any given group, the summated column X^2 value drops below a specific criterion value. In this study, the criterion level will be set at the .05 level of significance with m-1 degrees of freedom, where m represents the number of variables in each remaining group.

The agglomerative technique recommended by Ward also uses a distance function as its criterion for similarity. The Ward[1] technique finds the difference between the values of two subjects on each variable, squares it, then summates the values across each variable. The distance between the groups is then made proportionate

to the number of individuals within each group. Initially this will be two. Thus, the initial similarity function for the H-group technique may be presented as:

$$M_{ij} = \frac{\sum_{k=1}^{m}\left(X_{ik} - x_{jk}\right)^2}{2}, \quad i = 1 \ldots n, \; j = 1 \ldots n.$$

M represents the matrix of distances between all individuals or groups.

Those pairs with the least distance or error value are then combined, one pair at a time to form a group. If the error value is 0.0, then the pairs are identical and consequently form a totally homogeneous group. Once a pair of individuals is combined, however, a new error value must be calculated between the new group and all other groups.

In doing so, Ward makes use of the more generalized formula for the distance, or error, between groups:

$$E(n) = \sum_{i=1}^{p}\left[\sum_{k=1}^{n}\sum_{j=1}^{n_k} x_{ijk}^2 - \sum_{k=1}^{n}\frac{1}{n_k}\left(\sum_{j=1}^{n_k} x_{ijk}\right)^2\right]$$

The association analysis and the H-group technique have several similarities in common. They both use a straight distance function (one of the Minkowski metrics: see Appendix A) as a criterion for similarity. The advantage of this procedure over correlation co-efficients is

that the results of the relationships between several variables can be treated additively.

The acceptance of an additive distance function on the part of the researcher naturally requires a formidable value judgment. How, one might ask, is it possible to summate the Euclidian distances between sets of points when the theoretical meanings of the differences between those points may be quite different? The answer is not very straightforward and partially involves the fact that more preferable methods do not appear to exist at this time. Ideally, one would like to use an association coefficient such as Gamma or Yule's Q as a criterion for similarity. With these coefficients, there is no need to incorporate the concept of distance; instead, their guiding principle is a ratio between the consistencies ((0,0), (1,1)) to inconsistencies ((1,0), (0,1)). With one variable, these coefficients are ideal, however, a method has yet to be devised which would allow us to summate or incorporate the results of a series of correlations between several variables. Fortunately though, experience has shown us that distance functions are acceptable, and that in general they produce sampling distributions similar to the X^2 distribution, or in large samples, similar to the normal distribution.

Furthermore, one of the effects of choosing a monothetic schema is to 'standardize' the variables such that the conceptual differences between variable scores are reduced. Some authors[2] have even suggested that the conceptual arbitrariness of summated distances can be further overcome by adding weights to certain variables. The present author cannot accept this suggestion, however, since to do so would assume that the researcher had some prior knowledge of the relative importance of various variables. If a data grouping technique is being used for heuristic purposes, then this assumption of prior knowledge of the relative importance negates the heuristic purpose of the technique. If relative variable weights were known, a simple contingency table incorporating the most important variables could be used to group subjects in this case, and the whole process employed by statistical data grouping techniques could be circumvented.

3. Sample

The data for this study were gathered as part of a larger project in classification at the Regional Reception Centre, Kingston, Ontario, during the summer of 1974. An initial sample of 187 cases was drawn from the current classification files. This was further supplemented with 230 cases drawn at random from the non-active

files of the institution. The information on the inmate's file was obtained from intake reports, psychological examinations and classification officer's reports. Twenty-three variables pertinent to inmate classification were drawn for this study. These variables were primarily selected on the basis of a lack of missing responses and the logical exclusivity of one variable from another. Of the total sample of 417 cases, 400 were retained for this study; the other 17 cases represented missing data.

The data were transformed and coded dichotomously using a 0,1 format. Table 4 indicates the variables used.

TABLE 4

List of Variables Used in Analysis

Variable No.	Variable Name
1	Federal recidivist
2	Diagnosed psychopathology
3	Use of accomplices
4	Level of education less than gr. 8
5	Alcohol problem
6	Drug problem
7	Diagnosed physical ailment
8	Prior criminal record of any kind
9	Use of aliases
10	Age less than or equal to 23
11	Not married or living common law
12	Length of sentence less than 5 yrs.
13	Prior escape history
14	Charged with arson
15	sex offence
16	Murder
17	Parole violation only
18	property offence
19	other off. against person
20	offence under N.C.A.
21	unlawfully at large
22	provincial statute
23	parole violation plus other offence

4. Procedure

Both data grouping techniques were applied to the same data. The association analysis was executed and allowed to proceed until the criterion 5 per cent (.05) level of significance was reached. This, effectively, produced the optimum number of 'natural' clusterings within the data. The agglomerative technique produced an array of groupings, ranging from n to 2, with n = 300. As previously indicated, it is very difficult to determine what is optimal for an agglomerative technique regarding the number of groups produced. Thus, the association analysis was performed first in order to define an optimal number of groups. This number was then chosen as a stopping point for the H-group technique.

When the data were grouped, the group number and membership were obtained for each case. A procedure described by Rand[3] was used to measure the similarity of the two techniques. This procedure consists of looking at all combinations of pairs of subjects. Effectively, the proportion of pairs of subjects appearing together in a group is calculated. For example, if there are five individuals a, b, c, d, and e, who are divided into two groups by two techniques A and B, matrices can be drawn which indicate which pairs of individuals appear together in the same groups.

FIGURE 3

PAIRS PRODUCED BY GROUPING HYPOTHETICAL DATA

TECHNIQUE 'A'

	a	b	c	d	e
a		1	1	0	0
b			1	0	0
c				0	0
d					1
e					

PAIRS PRODUCED BY GROUPING

TECHNIQUE 'B'

	a	b	c	d	e
a		1	0	0	1
b			0	0	1
c				1	0
d					0
e					

SUMMARY

Technique	ab	ac	ad	ae	bc	bd	be	cd	ce	de
A	1	1	0	0	1	0	0	0	0	1
B	1	0	0	1	0	0	1	1	0	0
Consistencies	1		1			1			1	

These matrices may then be compared to see if the subject pairs appear together when subjected to each grouping technique. In our small example we will assume grouping technique A produces two groups $\{(a,b,c,)(d,e)\}$ and technique B the groups $\{(a,b,e)(c,d)\}$. A '1' within the matrix (see Figure 3) signifies that the elements of the pair appear together in the same group; a '0' indicates that they do not. The two matrices showing the presence of pairs are then superimposed on each other. The contiguous cells are then examined and the number of contiguous cells with the same pairing result (either a '1' or a '0') are calculated and found as a proportion of the total number of pairs. In our example there are 4 similar cells (either 1,1 or 0,0) out of 10 possible pairings, thus the proportion of similarities is 0.4 $(\frac{4}{10})$. In more formal terms, given two sets of groupings Y, Y' we can calculate the correlation c using the following computational formula indicated by Rand:

$$c(Y,Y') = \left[\binom{N}{2} - \left[\tfrac{1}{2}\left(\sum_{i=1}^{n}\left(\sum_{j=1}^{n} n_{ij}\right)^2 + \sum_{j=1}^{n}\left(\sum_{i=1}^{n} n_{ij}\right)^2\right) - \sum_{i=1}^{n}\sum_{j=1}^{n} n_{ij}^2\right]\right] \Big/ \binom{N}{2}$$

where Y,Y' have the same N points, and n_{ij} is the number of points in the i^{th} cluster of Y and the j^{th} cluster of Y'.

References

1. WARD, J.H. (1963) "Hierarchical grouping to optimize an objective function", Journal of the American Statistical Association, 58, 236.

2. SNEATH, P.H.A. & R.R. Sokal (1973) Numerical Taxonomy, San Francisco: W.H. Freeman & Co.

3. RAND, W.M. (1971) "Objective criteria for the evaluation of clustering methods", Journal of the American Statistical Association, 66, 846-850.

V

RESULTS

1. Comparison between Methods

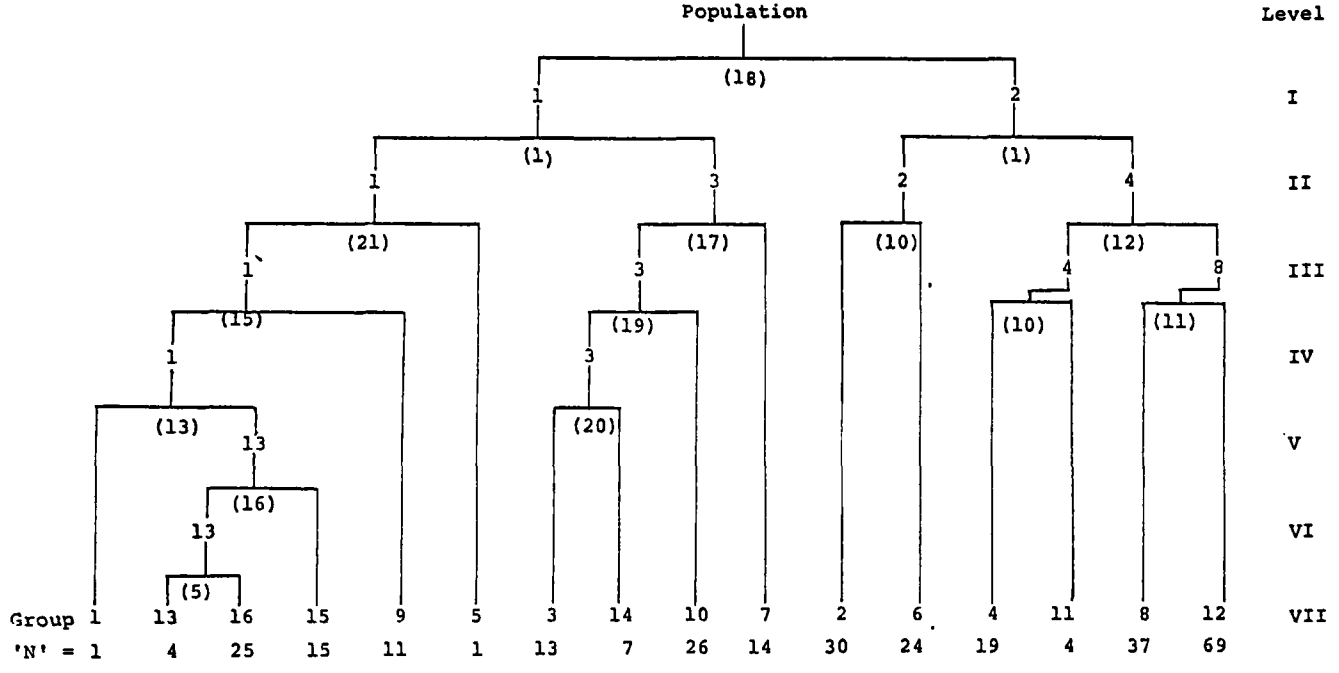
Because the H-group programme requires an n x n intercorrelation matrix in order to determine best fitting groups, the maximum amount of core available in the University of Ottawa's I.B.M. 360 computer was soon surpassed. An attempt was made to re-write the H-group programme using external direct-access disc storage, however, this method was found to be extremely inefficient. There were two operations which took a great deal of time: the first being the searching operation, wherein the groups with the lowest error potential were located, and second, the alteration of the error potential matrix to indicate the new error potentials between groups after two groups had been combined. After five minutes of C.P.U. time, only 13 groups had been collapsed starting with an initial 400 cases. On this basis, it was projected that the entire operation of collapsing from an initial 400 to two groups would take some six hours of computation time. The investigator made the decision that on a cost-benefit basis, it was much more feasible to reduce the number of cases than to utilize the re-designed programme. Using a modification of the routine proposed by Veldman[1], it was found that an initial 300 cases could be collapsed to a minimum of two in five minutes and forty-three seconds.

The decision was made that the incremental amount of information available by including the extra 100 cases was not worth the costs involved. As well, the investigator wished to stay within possible working parameters which would be faced by field workers attempting to use these methods. It is highly unlikely that many field researchers would be able to afford 6 hours of C.P.U. time, however, the time and core space constraints on a sample of 300 cases are well within the reach of most investigators.

The association analysis was performed first, and Figure 4 indicates the results obtained. Using a stopping criterion of a .05 level of significance with the degrees of freedom of m-1 groups, it was found that 16 was the optimal number of groups within the data. As was previously decided, the association analysis was to be compared with the H-group technique by comparing the groups produced at each level of the association analysis dendogram. As Figure 4 indicates, a series of seven levels was obtained. Level 1 produced the maximum of two groups, level 2 the maximum of four, and level 3, the maximum of eight. With the stopping of groups 2, 5, 6, and 7 at level 3 only twelve groups were produced at level four and further stoppages due to the criterion rule produced

FIGURE 4

ASSOCIATION ANALYSIS DENDOGRAM



Note: Figures in brackets represent critical variables (division points)

only fourteen groups at level 5, 15 groups at level 6, and 16 groups at level 7. Thus, seven points of comparison were used which utilized 2, 4, 8, 12, 14, 15, and 16 groups as produced by the association analysis.

The H-group analysis was then performed and the group memberships from the products of the corresponding 2, 4, 8, 12, 14, 15, and 16 groups were obtained.

Table 5 presents the characteristics of the 16 groupings as produced by the H-group procedure. Since information from all 23 variables is used by the cluster algorithm to define the clusters, the distribution of all variables is shown and not just the 13 variables which define the association analysis clusters.

As can be seen from Table 5, the H-group technique does not produce discrete types. Consequently, it is difficult to determine which factors are the most important in defining the clusters, and certainly it is almost impossible to allocate new cases to the produced clusters.

An intergroup comparison, using the previously described Rand technique was performed on the membership of corresponding groups. All possible pairs of individuals were checked to determine whether or not the pairs appeared together in a group or whether they were in dissimilar

TABLE 6

CHARACTERISTICS OF H-GROUP CLUSTERS*

Variable	Group 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	19.41	3.53	2.94	1.76	1.18	2.35	11.18	2.94	12.49	10.00	7.65	0.0	4.12	2.94	8.24	8.82
2	2.70	2.70	8.10	2.70	10.80	0.0	2.70	0.0	0.0	0.0	5.41	2.70	13.51	18.92	2.70	27.03
3	6.87	8.40	22.14	10.69	0.76	7.63	5.34	3.82	8.40	7.63	4.58	3.82	0.0	3.82	0.76	5.34
4	11.15	3.48	6.97	3.48	3.48	0.0	4.18	1.39	1.05	4.18	0.35	0.0	1.05	3.83	2.09	7.67
5	10.36	10.88	11.92	8.29	6.21	4.66	2.07	2.59	11.91	8.80	8.29	5.18	2.07	2.59	3.10	1.03
6	12.06	3.01	14.07	0.0	7.53	1.00	8.54	2.51	5.02	7.03	12.56	5.02	1.00	7.53	6.03	7.03
7	20.00	5.00	5.00	10.00	5.00	10.00	10.00	10.00	10.00	0.0	0.0	0.0	0.0	5.00	0.0	10.00
8	12.54	7.38	10.33	5.90	3.69	2.58	8.11	3.32	8.48	7.01	9.22	0.0	2.58	5.16	5.16	8.48
9	3.63	5.45	0.0	3.63	0.0	0.0	3.63	7.27	7.27	1.81	21.81	20.00	0.0	7.27	5.45	9.09
10	100.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
11	11.17	5.31	14.36	8.51	4.25	3.72	9.57	4.78	2.65	6.91	6.38	1.59	2.12	1.59	5.31	11.70
12	14.66	6.22	12.00	4.44	1.33	3.11	8.00	3.11	8.00	5.77	10.66	2.22	0.0	4.44	5.77	10.22
13	12.18	8.40	12.18	6.72	7.14	3.78	7.14	0.0	6.30	2.94	9.24	5.04	1.68	5.04	5.04	7.14
14	0.0	0.0	0.0	0.0	0.0	0.0	33.33	0.0	0.0	0.0	0.0	66.67	0.0	0.0	0.0	0.0
15	4.00	0.0	4.00	0.0	0.0	24.00	0.0	0.0	0.0	0.0	4.00	4.00	4.00	48.00	0.0	8.00
16	0.0	4.34	0.0	13.04	73.91	0.0	0.0	0.0	0.0	0.0	0.0	4.34	4.34	0.0	0.0	0.0
17	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.00	0.0
18	18.57	2.18	14.20	8.74	0.0	0.0	2.72	3.83	12.02	8.74	13.66	1.09	0.54	1.03	0.0	12.02
19	2.70	3.60	21.62	4.50	0.90	6.30	19.81	6.30	9.90	15.31	0.90	0.0	3.60	1.80	0.0	2.70
20	3.33	63.63	0.0	6.06	0.0	0.0	0.0	0.0	3.33	3.33	0.0	21.21	0.0	0.0	0.0	0.0
21	3.57	0.0	7.14	10.71	0.0	0.0	7.14	32.14	0.0	17.85	10.71	0.0	10.71	0.0	0.0	0.0
22	5.00	0.0	10.00	5.00	5.00	10.00	0.0	10.00	10.00	25.00	10.00	0.0	0.0	10.00	0.0	0.0
23	55.81	2.32	9.30	0.0	0.0	0.0	2.32	0.0	4.65	30.23	27.90	2.32	13.95	2.32	0.0	4.65

*Note: Figures are row percentages

groups. This analysis was performed on the results of both the H-group and the association analysis techniques. Similar pairings ((a,b), (a,c), (a,d), ..., (b,c), (b,d), ..., (c,d), (c,e), ... etc.) were compared to see if both techniques produced consistent results. That is, whether the pairings were consistently together or apart, or whether one technique paired individuals in one group and the other technique in a second. The total number of consistencies of pairs matched within the same groups and not within similar groups, was calculated across the two techniques and found as a proportion of the total number of possible pairings (n(n-1)/2). Table 6 indicates the results obtained by comparing consistencies of pair matchings.

As the results show, the correlation of similarity between the two groups ranges from .538 for two groups, to .869 for sixteen groups. Conversely, we may interpret the results as indicating that the proportion of mismatched pairs was .131 when the optimal 16 groups were used for comparison. Similarly, a proportion of .462 pairs were mismatched when the techniques provided only two groups. This compares with a theoretical maximum of .536 mismatchings for two groups.

As the number of groupings increases towardsthe

TABLE 6

COMPARISON BETWEEN ASSOCIATION ANALYSIS AND H-GROUP CLUSTERS

Dendogram Level	Number of Groupings	Proportion of consistencies of pairing distributions	Proportion of Mismatches
7	16	.869	.131
6	15	.864	.136
5	14	.854	.146
4	12	.851	.149
3	8	.753	.247
2	4	.653	.347
1	2	.538	.462

optimum number as prescribed by the association analysis stopping rule, the similarity of group membership produced by the two techniques converges. At maximal convergence, there is a 13.1% level of disagreement between the groupings produced by the two techniques.

2. Discussion

The major difference in sorting strategies used by the H-group and association analysis procedure is that the association analysis uses only the most statistically significant piece of information at each divisive step. The H-group procedure, however, combines groups on the basis of all of the information available between each pair of groups at every step. It would appear from the results obtained, that there is a convergence at about the optimal grouping level as prescribed by the association analysis technique, between the information used by the H-group technique and that utilized by the association analysis. The most critical information, it seems, is contained in those few variables upon which the decision for division in the association analysis routine is based. These same variables apparently are the most influential in determining the degree of difference between groups in the H-group process.

Under these conditions it would then be most probable to assume that the proportion of mismatches (.131) at the criterion level is due largely to the irreversibility of making wrong decisions at earlier stages of either the agglomeration routine in the H-group procedure or the divisive routine of the association analysis. It is remarkable, however, that given the rigidity of both procedures in not being able to re-allocate mismatched individuals to 'better' groups later in the agglomeration/division procedure, that such a high degree of correlation existed. This is even more remarkable when we consider the fact that the total proportion of possible mismatches increases with the number of groups available. Thus, the total number of theoretically possible mismatches is much greater for 16 groups than it would be for 2 groups.

Given these constraints, then, we can be fairly confident in the reliability and validity of the groupings existent in our sample data, and conversely, this confidence may be extended to the use of both techniques for deducing 'natural' clusters.

Since we now have two procedures which appear to produce similar results, the question naturally begs as to which procedure is preferable for actual field

research. The first critical question one must ask in order to make this decision would seem to relate to sample size. For large samples of 300 and over, it would appear that the computational efficiency offered by the association analysis is preferable. With 300 cases, the association analysis was performed in 1 minute and 23 seconds of actual computation time. Furthermore, computer core usage is minimal, falling well under 100 k bits. For large sample sizes, it is very difficult to provide a routine for the H-group procedure which meets with acceptable levels of core storage or, if written to incorporate external storage devices, acceptable levels of computation time. With less than 300 cases, these restrictions do not apply to the H-group procedure since efficient routines in terms of storage space and C.P.U. time are available. For small samples then, other factors must be considered in choosing one routine over another.

One major benefit of a divisive routine such as the association analysis is that it is much easier to determine the optimal number of 'natural' clusters. Agglomerative techniques do not usually incorporate this feature, since it is much more difficult to provide a stopping rule. The Ward H-group technique though, is unique in that it provides an error potential to indicate the amount of information

loss occurring due to grouping.

The H-group procedure does have an advantage in dealing with very small samples in that it will group subjects when there is a high variable to case (m:n) ratio, and also, simply too few cases for divisive techniques incorporating such statistics as X^2.

It is conceivable that some modification of Fisher's exact test, or a similar statistic could be used as an alternative for the divisive procedure, but it is highly unlikely that this could compete with the efficiency of an agglomerative technique.

A major problem still remains with the H-group procedure, however, and that is, it is difficult to categorically assign new cases to an existent group. In order to include a new datum or case, the entire H-group procedure must be re-executed. This is not so, however, for the association analysis technique.

References

1. VELDMAN, D.J. (1967) Fortran programming for the behavioural sciences, New York: Holt, Rinehart & Winston.

APPENDIX I

The following is a list of some of the more prevalent similarity coefficients to be found in the literature pertinent to quantitative data grouping techniques.

1. Distance Coefficients

i) Euclidean Distance

$$\sum_{i=1}^{n}(x_{ij} - x_{ik})^2$$

ii) City-block Metric

$$\sum_{i=1}^{n}|x_{ij} - x_{ik}|$$

iii) Minkowski Metrics (General Formula)

$$\left[\sum_{i=1}^{n}|x_{ij} - x_{ik}|^{\lambda}\right]^{1/\lambda}$$

iv) Canberra Metric

$$\sum_{i=1}^{n}\frac{|x_{ij} - x_{ik}|}{(x_{ij} + x_{ik})}$$

v) Coefficient of Divergence

$$\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_{ij} - x_{ik}}{x_{ij} + x_{ik}}\right)^2\right]^{1/2}$$

vi) Profile Similarity Index

$$\frac{2X^2_{.5n} - nd^2_{jk}}{2X^2_{.5n} + nd^2_{jk}}, \quad \text{where } d_{jk} = \sqrt{\Delta_{jk}/n}\,,$$

$$\text{and} \quad \Delta_{jk} = \left[\sum_{i=1}^{n}(x_{ij} - x_{ik})\right]^{1/2}$$

vii) Coefficient of Nearness

$$\frac{(2n)^{1/2} - \Delta}{(2n)^{1/2} + \Delta}$$

2. Association Coefficients

i) Jaccard Coefficient

$$\frac{a}{a + b + c}$$

ii) Weighted Jaccard

$$\frac{2a}{2a + b + c}$$

iii) Simple Matching

$$\frac{a + d}{a + b + c + d}$$

iv) Dissimilarity Coefficient

$$\frac{b + c}{2a + b + c}$$

v) Chi-square (X^2)

$$\frac{n(ad - bc)^2}{(a+b)\ (c+d)\ (a+c)\ (b+d)}$$

vi) Phi (ϕ)

$$\frac{X^2}{N}$$

vii) Contingency Coefficient (C)

$$\left(\frac{X^2}{X^2 + N}\right)^{1/2}$$

viii) Yule's Q

$$\frac{(ad - bc)}{(ad + bc)}$$

3. Correlation Coefficients

i) Pearson Product Moment Correlation

$$\frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)\quad(x_{ik} - \bar{x}_k)}{\left[\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2 \sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2\right]^{1/2}}$$

ii) Angular Separation

$$\frac{\sum_{i=1}^{n} x_{ij}\, x_{ik}}{\left[\sum_{i=1}^{n} x_{ij}^2 \sum_{i=1}^{n} x_{ik}^2\right]^{1/2}}$$

Bibliography

BALL, G.H. & D.J. Hall (1967) "A clustering technique for summarizing multivariate data", Behavioural Science, 12, 153-155.

BALLARD, K.B. & D.M. Gottfredson (1964) Predictive attribute analysis and prediction of parole performance, Vacaville, Calif.: Institute for the Study of Crime and Delinquency.

BENN, S.I. & R.S. Peters (1959) The principles of political thought, New York: The Free Press.

BLALOCK, H.M. (1972) Social statistics, New York: McGraw-Hill.

CAIN, A.J. & G.A. Harrison (1958) "An analysis of the taxonomists' judgement of affinity", Proceeding of the Zoological Society of London, 131, 85-98.

CATTELL, R.B. (1965) "Factor analysis: An introduction to essentials. I. The purpose and underlying models", Biometrics, 21, 190-215.

_____. (1965) "Factor analysis: An introduction to essentials. II. The role of factor analysis in research", Biometrics, 21, 405-435.

CHAPANIS, A. (1961) "Men, machines and models", American Psychologist, 16, 113-131.

CHEETHAM, A.H. & J.E. Hazel (1969) "Binary (presence-absence) similarity coefficients", in Journal of Paleontology, 43, 1130-1136.

COCHRAN, G. & C.E. Hopkins (1961) "Some classification problems with multivariate quantitative data", Biometrics, 17, 10-32.

COLE, A.J. (ed) (1969) Numerical taxonomy, London: Academic Press.

CORMACK, R.M. (1971) "A review of classification", Journal of the Royal Statistical Association, A132, 321-367.

COX, D.R. (1970) The analysis of binary data, London: Methuen.

DRIVER, E.A. (1968) "A critique of typologies in criminology", Sociological Quarterly, 9, 356-373.

DEUTSCH, K.W. (1966) "On theories, taxonomies and models as communication codes for organizing information", Behavioural Science, 11, 1-17.

DICE, L.R. (1945) "Measures of the amount of ecological association between species", Ecology, 26, 297-302.

EVERITT, B.S. (1972) "Cluster Analysis", British Journal of Psychiatry, 120, 143.

FILDES, R. & D.M. Gottfredson (1972) "Cluster analysis in a Parolee Sample", Journal of Research in Crime and Delinquency, 9, 2-11.

FISHER, W.D. (1969) Clustering and aggregation in economics, Baltimore: Johns Hopkins Press.

GEER, J.P. Van de (1971) Introduction to multivariate analysis for the social sciences, San Francisco: W.H. Freeman.

GOOD, I.J. (1965) Categorization of classification mathematics and computer science in biology and medicine, London: H.M.S.O.

GOODALL, D.W. (1953) "Objective methods for the classification of vegetation I: The use of positive interspecific correlation", Australian Journal of Botany, 1, 39-63.

_____. (1964) "A probabilistic similarity index", Nature, 203, 1098.

GOODMAN, L.A. & W.H. Kruskal (1954) "Measures of association for cross-classifications", Journal of the American Statistical Association, 49, 732-64.

_____. (1959) "Measures of association for cross-classification II: Further discussion and references", Journal of the Americal Statistical Association, 54, 123-163.

_____. (1963) "Measures of association for cross-classification III: Approximate sampling theory", Journal of the American Statistical Association, 58, 310-364.

GOULDNER, A.W. (1964) "Anti-minotaur: The myth of a value free sociology", in Horowitz, I.L. (ed.) The new sociology, 1964, New York: The Free Press.

GOWER, J.C. (1967) "A comparison of some methods of cluster analysis", Biometrics, 23, 623-637.

_____. (1971) "A general coefficient of similarity and some of its properties", in Biometrics, 27, 857-871.

GROSS, A.L. (1972) "A Monte-Carlo study of the accuracy of a hierarchical grouping procedure", Multivariate Behavioural Research, 7, 379-389.

GRUYGIER, T. (1964) "Treatment variables in non-linear prediction", in Johnson, N., L. Savitz, & M.E. Wolfgang (eds.), The Sociology of Punishment and Corrections, 1967, New York: John Wiley & Sons.

HEMPEL, C.G. (1952) 1969. Fundamentals of concept formation in empirical science, Chicago: University of Chicago Press.

JOHNSON, S.C. (1967) "Hierarchical clustering schemes", Psychometrika, 32, 241-254.

JARDINE, N. & R. Sibson (1971) Mathematical taxonomy, London: Wiley.

KERLINGER, F.N. (1973) Foundations of behavioural research, 2nd ed., New York: Holt, Rinehart & Winston.

LACHENMEYER, C.W. (1971) The language of sociology, New York: Columbia University Press.

LANCE, G.N. & W.T. Williams (1965) "Computer programmes for monothetic classification", Computer Journal, 8, 246-249.

_____. (1967) "A general theory of classification strategies I: Hierarchical systems", Computer Journal, 9, 373-380.

_____. (1968) "A general theory of classificatory sorting strategies II: Clustering systems", Computer Journal, 10, 271-277.

LANGE, R.T., N.S. Stenhouse & C.E. Offler (1965) "Experimental appraisal of certain procedures for the classification of data", Australian Journal of Biological Science, 18, 1189-1205.

LUNDBERG, G.A. (1939) "The postulates of science and their implications for sociology" in Natanson, M. (ed.) Philosophy of the social sciences, 1963, New York: Random House.

MacNAUGHTON-SMITH, P. (1963) "The classification of individuals by the possession of attributes associated with a criterion", Biometrics, 19, 364-366.

_____. (1966) "Some statistical and other numerical techniques for classifying individuals", Home Office Research Unit Report No. 6, London: H.M.S.O.

MacNAUGHTON-SMITH, P., W.T. Williams, M.B. Dale & L.G. Mockett (1964) "Dissimilarity analysis: A new technique of hierarchical subdivision", Nature, 202, 1034-1035.

MARTINDALE, D. (1959) "Sociological theory and the ideal type" in Gross, L. Symposium on sociological theory, New York: Row, Peterson & Co.

McKINNEY, J.C. (1966) Constructive typology and social theory, New York: Appleton-Century-Crofts.

MOORE, A.W. & J.S. Russell (1967) "Comparison of coefficients and grouping procedures in numerical analysis of soil trace element data", Geoderma, 1, 139-158.

RAND, W.M. (1971) "Objective criteria for the evaluation of clustering methods", Journal of the American Statistical Association, 66, 846-850.

RICHMOND, M.S. (1971) Classification of jail prisoners, Washington, D.C.: U.S. Bureau of Prisons.

SINCLAIR, I. & B. Chapman (1973) "A typological and dimensional study of a sample of prisoners", British Journal of Criminology, 13, 341-353.

SNEATH, P.H.A. (1957) "Some thoughts on bacterial classification", Journal of General Microbiology, 33, 184-200.

SNEATH, P.H.A. & R.R. Sokal (1973) Numerical taxonomy, San Francisco: W.H. Freeman.

SOKAL, R.R. & P.H.A. Sneath (1963) Principles of numerical taxonomy, London: W.H. Freeman.

SPENCE, N.A. & P.J. Taylor (1970) "Quantitative methods in regional taxonomy", Progress in Geography, 2, 1-64.

STEVENS, S.S. (1946) "On the theory of scales of measurement", Science, 684, 677-680.

SUITS, D. (1957) "Dummy variables in regression equations", Journal of the American Statistical Association, 52, 548-551.

TIRYAKIAN, E.A. (1968) "Typological classification", International Encyclopedia of the Social Sciences, 1968, New York: Macmillan, 177-186.

VELDMAN, D.J. (1967) Fortran programming for the behavioural science, New York: Holt, Rinehart & Winston.

WARD, J.H. (1963) "Hierarchical grouping to optimize an objective function", Journal of the American Statistical Association, 58, 236-244.

WATSON, L., W.T. Williams & G.N. Lance (1967) "A mixed-data numerical approach to angiosperm taxonomy: The classification of Ericales", Proceedings of the Linnaean Society of London, 178, 25-35.

WEBER, M. (1949) The methodology of the social sciences, trans. by Schils, E.A. & H.A. Finch, Glencoe, Ill.: Free Press

WILKINS, L.T. & P. MacNaughton-Smith (1964) "New prediction and classification methods in criminology", Journal of Research in Crime and Delinquency, 1, 19.

WILLIAMS, W.T. & M.B. Dale (1965) "Fundamental problems in numerical taxonomy", Advances in Botanical Research, 2, 35-68,

WILLIAMS, W.T., M.B. Dale & P. MacNaughton-Smith (1964) "An objective method of weighting in similarity analysis", Nature, 201, 426.

WILLIAMS, W.T. & J.M. Lambert (1959) "Multivariat methods in plant ecology I: Association analysis in plant communities", Journal of Ecology, 47, 83-101.

_____. (1960) "Multivariate methods in plant ecology II: The use of an electronic digital computer for association analysis", Journal of Ecology, 48, 689-710.

_____. (1961) "Multivariate methods in plant ecology III: Inverse association analysis", Journal of Ecology, 49, 717-729.

WISHART, D. (1969) "An algorithm for hierarchical classification", Biometrics, 25, 165-170.

WOOD, A.L. (1969) "Ideal and empirical typologies for research in deviance and control", Sociology and Social Research, 53, 227-241.

SUMMARY

This monograph is a discussion of the application of quantitative data grouping techniques in the field of Criminology.

The first part of this paper provides an introduction to quantitative grouping techniques. A review of literature illustrating the application of these procedures both within criminology and allied fields follows. An overview of the logical structure of these techniques is then presented along with a generalized model which may be used by researchers wishing to apply and modify existing data grouping procedures.

The literature reviews indicate that existing quantitative grouping procedures are based on three "families" of similarity algorithms -- distance coefficients, association coefficients and regression coefficients -- and three sorting strategies -- agglomerative, divisive and iterative.

The second part of this monograph reports a comparative study which was carried out on two fairly well known techniques -- association analysis and Ward's Hierarchical Grouping Procedure. Both of these techniques were applied to a sample of data (N=300) obtained from inmate classification files at a Regional Reception Centre. The data represented the inmate's offence types, and social background.

It was discovered that there was a convergance in the similarity of group membership produced by each procedure as the optimal number of natural clusters or groups was approached. When a comparative procedure devised by Rand was applied to the groupings produced by the two techniques, the proportion of similar pairings within groups ranged from .538 for two groups produced by each procedure to .869 for 16 groups produced by each procedure.

Given these results, it was concluded that the techniques were sensitive to similar perturbations in the data, thus providing similar clusters. Further, the "naturalness" or validity of the obtained clusters was strengthened because of the similarity of results obtained by the two procedures.