CHAPTER VII-TN 10-Appendix: PROGRAMMES AND CLUSTER ANALYSES
By C. Peebles (some material not relevant in 2006 was removed)

Cluster analysis is a term which covers a multitude of strategies and algorithms designed either to group similar individuals as a function of their attributes or, more rarely, to group a number of related variables scored over a series of individuals. Excluding parametric multivariate techniques such as factor analysis and joint metric space and partial order scalogram analysis, Wishart (1969) has classified cluster-analytic strategies into two basic groups: (1) "natural class" seeking algorithms and (2) optimum solution methods. The first of these major divisions has received attention from only a few workers - (e.g. see Wishart 1669, 1971, 1972). It is the second of Wishart's two categories that has seen the bulk of activity in cluster analysis. This is largely because it offers solutions that are optional in terms defined by the investigator rather than in "natural" terms that are defined by complex methodological and epistemological dimensions of the problem and data. Greig-Smith (1964) provides a discussion of "natural" class problems in numerical classification of biological organisms. In the assignment of individuals to species, the requirements of both proper taxonomic assignment and phylogenitic grouping must be met.

Within the second major group of clustering strategies a further division can be made. As Wishart says:

> *Probably the most common technique is the hierarchic fusion algorithm which has the advantage (although very expensive with large populations) in presentation of the resulting "dendrogram". The construction of "keys" was a requirement of botanical applications that gave rise to the early monothetic divisive techniques, enabling observers to identify plant communities by the presence and absence of certain key species. The third recurring technique improves a given classification by iterative relocation of cluster members so as to optimize some objective measure of overall homogeneity in terms of the similarity between individuals and clusters. Also known as the "X-mean", "transition" and "Euclidian cluster" methods it is economical in computer processor time and appears to find global optimum solutions for most small populations.*

In general, most hierarchic-fusion algorithms have in the past been called polythetic-agglomerative to contrast them with the monothetic-divisive strategies. Polythetic-agglomerative clustering groups individuals on the basis of a measure of similarity or dissimilarity generated from the measures of their attributes. Based on these measures of similarity or dissimilarity between all members, clusters of related individuals are grouped by a "rule" in the algorithm. An almost inexhaustible array of measures of relationship between individuals measured over both binary and continuous attribute states have been used. The methods - or rules - for grouping individuals have been equally large: single, average, and complete linkage, error sum, information gain, and cliques, clumps, and stars to name a few. In each case the measure is either minimized or maximized as a criterion for inclusion of an individual in a group or for the fusion of two groups.

The monothetic-divisive algorithms work in exactly the opposite manner. Instead of being built into more and more inclusive groups a population of individuals is partitioned into more and more homogeneous sub-sets. Williams and Lambert (1959) the originators of monothetic divisive analysis in plant ecology, defined the problem as the subdivision of a population so that all associations disappear; but there will in general be a large number of alternative subdivisions fulfilling this requirement. We therefore propose the concept of efficient subdivision, by which we intend subdivision of that species which, in the two subclasses

resulting, produces the smallest total number of residual significant associations.

They propose χ^2 as the measure and the variable with the highest $\Sigma \chi^2$ as the divisor. That is, if the population is divided into individuals with and individuals without the attribute which shows the highest $\Sigma \chi^2$ the largest number of associations will be eliminated. The resulting sets will be the most homogenous of any possible pair drawn from the initial population.

The use of $\Sigma \chi^2$, however, has several undesirable properties. In the qualitative case there is dissatisfaction with the χ^2 model. It is excessively sensitive to skewness of the underlying distribution; as a result, the simultaneous possession by an individual of two uncommon attributes assumes a quite disproportionate importance (Lance and Williams 1970).

Lance and Williams (1970), Orlocki (1969), McNaughton-Smith (1965) and others have proposed the Information Statistic be used in place of $\Sigma \chi^2$. The Information Statistic asymptotically approaches the χ^2 distribution when the population over which it is computed is (MacNaughton-Smith 1965).

Cluster analysis, whether agglomerative or divisive, monothetic or polythetic, is a data reduction device. The sampling distributions and statistical properties of various techniques have not been worked out. Thus any statistical inference based solely on the results of a cluster analysis is suspect at best. If, however, the results of a cluster analysis fulfill some theoretical expectation or are subjected to further, external statistical testing then cluster analysis can be a very powerful technique. The programs used in this study were part of Wishart's CLUSTAN IA Suite (1969).