4.3.3.1 Hierarchical clustering schemes (HCS)

An hierarchical clustering scheme takes a matrix of dissimilarity measures between a set of objects and represents the objects as being gathered into clusters on the basis of this information. It describes not one clustering but rather (for p points) p different clusterings, referred to as *levels* of a single total hierarchical *scheme*. At the highest level, all the objects are contained in one cluster, at the next highest there are two and so on until, at the lowest level, there are as many clusters as there are points. The defining characteristic of a hierarchical scheme is that once a point is incorporated into a cluster at a lower level it may not 'leave' that cluster at a higher one. Thus the clusters form a hierarchical scheme in the sense that each level is a special case of the next highest. We now consider in some detail the method of hierarchical clustering.

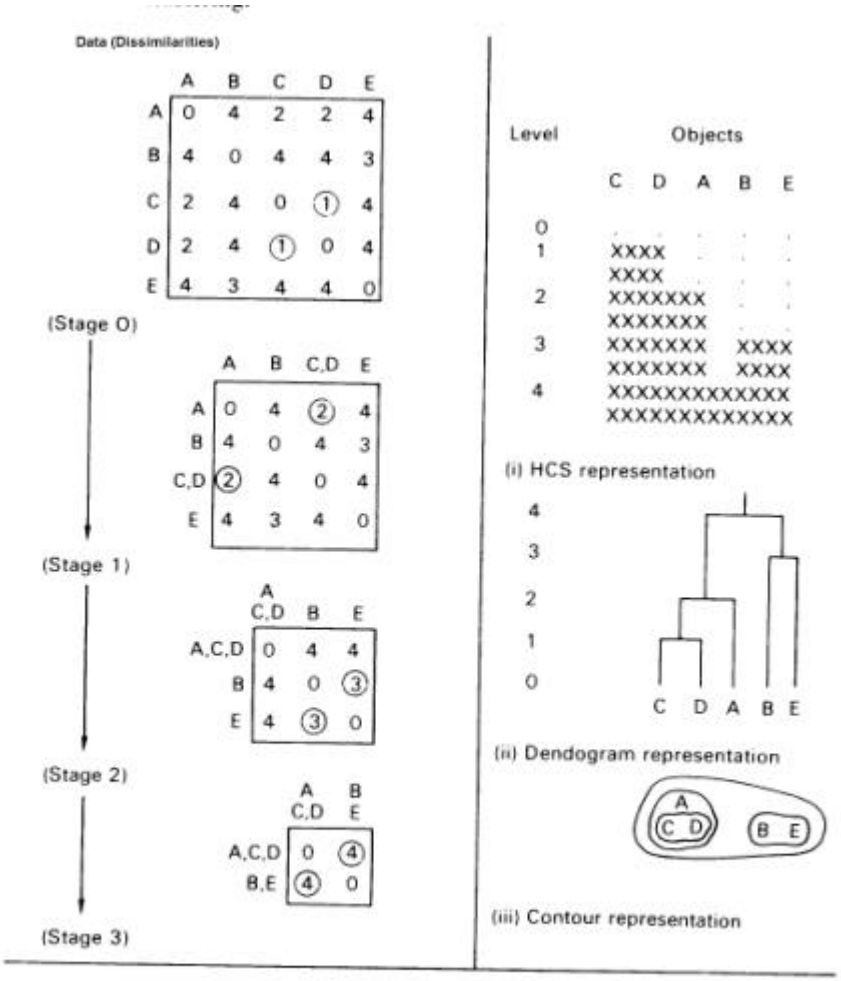

Figure 4.2 *Illustrative example of the HCS procedure and forms of representation*

The Method

Stage 0 The process of clustering begins by inspecting the original data matrix dissimilarities and identifying first the most similar (or least dissimilar) pair of objects *(C* and *D)* and then merging them into a cluster. We now have the closest cluster of two points: *(C, D).*

Stage I Points *C* and *D* are from now on treated as a single *object* and the data matrix is reduced by removing the row and column of *C* and *D* and substituting one representing the cluster *(C, D).* In this example, the dissimilarity between C and each other object is the same as that between *D* and the same object—e.g. δ(C,A) = δ*(D, A)* = 2. (Normally this will not be the case.) The smallest entry in the reduced matrix, the currently most similar link, is now identified, and turns out to be between the cluster *(C, D)* and *A.* So the new structure consists of the dense cluster *(A. C. D)* and a set of unlinked points.

Stage 2 The new reduced matrix consists of the cluster *(A, C, D)* and points *B* and *E.* The smallest entry is now between *B* and *E.* This pair now form a new distinct cluster; the structure at this stage consists simply of the two clusters: *(A, C, D) vs (B, E).* A new reduced matrix is formed.

Stage 3 In this final stage, the remaining two entities - the two clusters *(A, C,* D) and *(B,* E)—are merged, forming the final clustering of all the points.

The process of hierarchical clustering—forming clusters at decreasing levels of compactness—gives considerable insight into the regions of the space. In this case, we can see that the basic contrast is between the (C, D, A) and (B, E) cores of clustering.

As in other areas of data analysis, especially block modelling of social(see White et al. 1976 and Breiger et al. 1975) 'holes', the empty areas, frequently turn out to be quite as significant as the clusters, and both aspects have to be represented in any structural analysis. Empty regions represent two types of significant information: differentiation or dissociation between clusters on the one hand, and/or the significant absence of objects on the other hand, which might mean that certain stimuli have been neglected or overlooked in a study or that no objects actually exist which have a particular combination of attributes.

Clearly, the most significant *clustering* information is contained in the initial stages, and the most significant *dissociation* information is contained in the later stages of a clustering.

In the above example there was no ambiguity in defining the distance between a newly-formed cluster and existing objects (clusters or points), but this will not usually be the case. Consider the simplest case where we have a cluster formed of two points A and B and a third point C. There will be two distances, namely those between A and C, δ(A, C), and between B and C, δ(B, C); and we have to decide:how we are going to use these to define the distance between (A, B) and C, that is δ((A, B), C). If we want the procedure to produce identical clustering schemes when the data are monotonically transformed we cannot take the obvious step of averaging δ(A, C) and δ(B, C). Johnson (1967) suggests two contrasting ways of defining this distance in this instance:

The *maximum* method (otherwise known as the diameter or complete link method) defines the distance δ*((A, B), C)* to be the *maximum of* δ*(A, C)* and δ*(B, C)*.

The alternative *minimum* method (also known as the connectedness or single-link method) conversely defines the distance between the new cluster and the extraneous point to be the *minimum* of the distances between the extraneous point and each of the points in the cluster.

When the data satisfy the ultra-metric inequality (see 6.1.6) and are therefore perfectly representable as an HCS, the two methods produce identical hierarchical clusterings. Otherwise, the two HCSs will differ—often not markedly, but sometimes significantly.

The maximum (diameter) method picks out the largest distance within a cluster as 'the' distance and seeks to minimise the diameter (largest distance between the objects) within a cluster. This tends to produce a fairly small number of compact clusters.

The minimum method, by contrast, selects the smallest distance as 'the' distance and seeks to minimise the largest link needed to produce a chain or connected path between the objects. It tends to produce rather a large number of broken clusters and is often marked by chaining—the continued addition of a single element to a cluster.

In practice, the minimum method is usually to be preferred to the maximum method in exploring the hierarchical structure of a set of data (although both methods should be inspected to see how far the data may legitimately be represented this way.* The chief use of the HCS procedure is to examine not only relatively dense local' structure of highly proximate points (the lower levels of the clustering) but also the open or 'global' structure of spaces which separate or dissociate the clusters (the highest levels).

Hierarchical clustering then, possesses a number of useful characteristics:

it presents not one, but a whole series of linked clusterings of increasing

density, from a ‘clustering’ where each point is a separate cluster to the one where all the points are in a single cluster;

it includes two commonly used types of clustering as special cases and therefore gives the user some idea of how well the data fit the assumptions of the clustering model;

the HCS procedure is non-metric, in the sense that any ordinal rescaling of the data will produce identical results.

***Holman (1972) has shown that a set of errorless data will never perfectly satisfy both the Euclidean distance model and the hierarchical model, but will always satisfy one of the models to some extent.**