Clustering Checklist

Cluster v. Discriminant Analysis: Although both cluster analysis and discriminant analysis classify objects (or cases) into categories, discriminant analysis requires you to know group membership for the cases used to derive the classification rule. The goal of cluster analysis is to identify the actual groups.

Types of Clustering: Hierarchical, K-Means, Two-Step. If you have a large data file (even 1,000 cases is large for clustering) or a mixture of continuous and categorical variables, you should use the SPSS two-step procedure. If you have a small data set and want to easily examine solutions with increasing numbers of clusters, you may want to use hierarchical clustering. If you know how many clusters you want and you have a moderately sized data set, you can use k-means clustering.

Curse of Different Variable Type: If variables are measured on different scales, variables with large values contribute more to the distance measure than variables with small values. Need to standardize data.

Outliers: K-means clustering is very sensitive to outliers, since they will usually be selected as initial cluster centers. This will result in outliers forming clusters with small numbers of cases. Before you start a cluster analysis, screen the data for outliers and remove them from the initial analysis.

Irrelevant Variables: Cluster analysis has no mechanism for differentiating between relevant and irrelevant variables. Therefore the choice of variables included in a cluster analysis must be underpinned by conceptual considerations. This is very important because the clusters formed can be very dependent on the variables included.

Hierarchical Clustering: Agglomerative methods, in which subjects start in their own separate cluster. The two ’closest’ (most similar) clusters are then combined and this is done repeatedly until all subjects are in one cluster. At the end, the optimum number of clusters is then chosen out of all cluster solutions. Divisive methods, in which all subjects start in the same cluster and the above strategy is applied in reverse until every subject is in a separate cluster. Agglomerative methods are used more often than divisive methods, so this handout will concentrate on the former rather than the latter.

Data Types: Interval, ordinal, categorical.

Distance Measures: Euclidean, Manhattan, and Canberra.

Chaining: Clusters end up being long and straggly.

Ward’s method: In this method all possible pairs of clusters are combined and the sum of the squared distances within each cluster is calculated. This is then summed over all clusters. The combination that gives the lowest sum of squares is chosen. This method tends to produce clusters of approximately equal size, which is not always desirable. It is also quite sensitive to outliers. Despite this, it is one of the most popular methods, along with the average linkage method.