1 Hierarchical Cluster Analysis

- This procedure attempts to identify relatively homogeneous groups of cases (or variables) based on selected characteristics, using an algorithm that starts with each case (or variable) in a separate cluster and combines clusters until only one is left. You can analyze raw variables or you can choose from a variety of standardizing transformations. Distance or similarity measures are generated by the Proximities procedure.

- Statistics are displayed at each stage to help you select the best solution. Statistics include agglomeration schedule, distance (or similarity) matrix, and cluster membership for a single solution or a range of solutions. Plots include dendrograms and icicle plots.

    * ***Agglomeration schedule.*** Displays the cases or clusters combined at each stage, the distances between the cases or clusters being combined, and the last cluster level at which a case (or variable) joined the cluster.
    * ***Proximity matrix.*** Gives the distances or similarities between items.
    * ***Cluster Membership.*** Displays the cluster to which each case is assigned at one or more stages in the combination of clusters. Available options are single solution and range of solutions.
    * ***Dendrograms*** can be used to assess the cohesiveness of the clusters formed and can provide information about the appropriate number of clusters to keep.
    * ***Icicle plots*** display information about how cases are combined into clusters at each iteration of the analysis.
      (User can specify a range of clusters to be displayed Orientation allows you to select a vertical or horizontal plot.)

1.1 Data

- The variables can be quantitative, binary, or count data. Scaling of variables is an important issue–differences in scaling may affect your cluster solution(s).

- If your variables have large differences in scaling (for example, one variable is measured in pounds and the other is measured in years), you should consider standardizing them (this can be done automatically by the Hierarchical Cluster Analysis procedure).

1.2 Case Order

If tied distances or similarities exist in the input data or occur among updated clusters during joining, the resulting cluster solution may depend on the order of cases in the file. You may want to obtain several different solutions with cases sorted in different random orders to verify the stability of a given solution.

1.3 Assumptions

The distance or similarity measures used should be appropriate for the data analyzed. Also, you should include all relevant variables in your analysis. Omission of influential variables can result in a misleading solution. Because hierarchical cluster analysis is an exploratory method,

results should be treated as tentative until they are confirmed with an independent sample. *(N.B We will look at cross-validation techniques later in the semester).*

1.4 Implementation

To obtain a Hierarchical Cluster Analysis From the menus choose:

`Analyze > Classify > Hierarchical Cluster...`

If you are clustering cases, select at least one numeric variable. If you are clustering variables, select at least three numeric variables.

2 SPSS Implementation and Output

- Hierarchical Cluster Analysis is implemented by the `classify` option on the `analyse` menu. Three options shall appear. Select `Hierarchical`.
- We performed a hierarchical cluster analysis in SPSS, selecting all the variables (except categorical variables) in the **Variable(s)** box. We can label the cases by a categorical variable.
- We shall further requested the Dendrogram in the output. We changed all variables to z-scores to yield equal metrics and equal weighting, selected the Squared Euclidean distance (the default) method of determining distance between clusters and the **Ward's method** for clustering, and saved a 3-cluster solution as a new variable.

2.1 Proximity matrix

- The output will print distances or similarities computed for any pair of cases. *We will not be covering this in detail here, but it is a major learning outcome for this course.*

2.2 Cluster Membership

- This box allows you to specify a set number of clusters. If you have a hypothesis about how many clusters there are, you can specify a set number of clusters, or create a number of clusters within a range.

2.3 Icicle Plot

- Icicle plots visually represent information on the agglomeration schedule. You can select that all clusters are included in the icicle plot, or restrict it to a range of clusters.
- Also, you can read the plot from bottom up (vertical orientation) or from left to right (horizontal orientation).

2.4 Distance Measure

There are different distance measure choices depending on the level of measurement of the data: interval, count, or binary. For nearly all of this module example,the data were on an interval scale, and the squared euclidean measure will suffice for lab classes, until instructed otherwise.

2.5 SPSS Agglomeration Schedule

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	**1**	**3**	**,002**	**0**	**0**	**17**
2	13	18	2,708	0	0	9
3	12	17	4,979	0	0	14
4	20	21	5,014	0	0	7
5	11	14	8,509	0	0	10
6	5	8	11,725	0	0	8
7	19	20	11,871	0	4	14
8	2	5	13,174	0	6	13
9	13	15	14,317	2	0	12
10	**9**	**11**	**19,833**	**0**	**5**	**15**
11	6	7	22,901	0	0	15
12	10	13	23,880	0	9	16
13	2	4	28,378	8	0	17
14	**12**	**19**	**31,667**	**3**	**7**	**16**
15	6	9	40,470	11	10	18
16	10	12	44,624	12	14	19
17	1	2	47,720	1	13	20
18	6	16	49,963	15	0	19
19	6	10	64,785	18	16	20
20	1	6	115,781	17	19	0

- The procedure followed by cluster analysis at Stage 1 is to cluster the two cases that have the smallest squared Euclidean distance between them.
- Then SPSS will recompute the distance measures between all single cases and clusters (there is only one cluster of two cases after the first step).

- Next, the 2 cases (or clusters) with the smallest distance will be combined, yielding either 2 clusters of 2 cases (with 17 cases unclustered) or one cluster of 3 (with 18 cases unclustered).
- This process continues until all cases are clustered into a single group. For the sake of clarify, we will explain Stages 1, 10, and 14.

Stage 1

- At Stage 1, Case 1 is clustered with Case 3. The squared Euclidean distance between these two cases is 0.002.
- Neither variable has been previously clustered (the two zeros under Cluster 1 and Cluster 2), and the next stage (when the cluster containing Case 1 combines with another case) is Stage 17.
- (*Note that at Stage 17, Case 2 joins the Case-1 cluster.*)

Stage 10

- At Stage 10, Case 9 joins the Case-11 cluster (Case 11 was previously clustered with Case 14 back in Stage 5, thus creating a cluster of 3 cases: Cases 9, 11, and 14).
- The squared Euclidean distance between Case 9 and Case-11 cluster is 19.833.
- Case 9 has not been previously clustered (the zero under Cluster 1), and Case 11 was previously clustered at Stage 5.
- The next stage (when the cluster containing Case 9 clusters) is Stage 15 (when it combines with the Case-6 cluster).

Stage 14

- At Stage 14, the clusters containing Cases 12 and 19 are joined, Case 12 has been previously clustered with Case 17, and Case 19 had been previously clustered with Cases 20 and 21, thus forming a cluster of 5 cases (Cases 12, 17, 19, 20, 21).
- The squared Euclidean distance between the two joined clusters is 31.667.
- Case 12 was previously joined at Stage 3 with Case 17. Case 19 was previously joined at Stage 7 with the Case- 20 cluster.
- The next stage when the Case-12 cluster will combine with another case/cluster is Stage 16 (when it joins with the Case-10 cluster).

The branching-type nature of the Dendrogram allows you to trace backward or forward to any individual case or cluster at any level. It, in addition, gives an idea of how great the distance was between cases or groups that are clustered in a particular step, using a 0 to 25 scale along the top of the chart. While it is difficult to interpret distance in the early clustering phases (the extreme left of the graph), as you move to the right relative distance become more apparent. The bigger the distances before two clusters are joined, the bigger the differences in

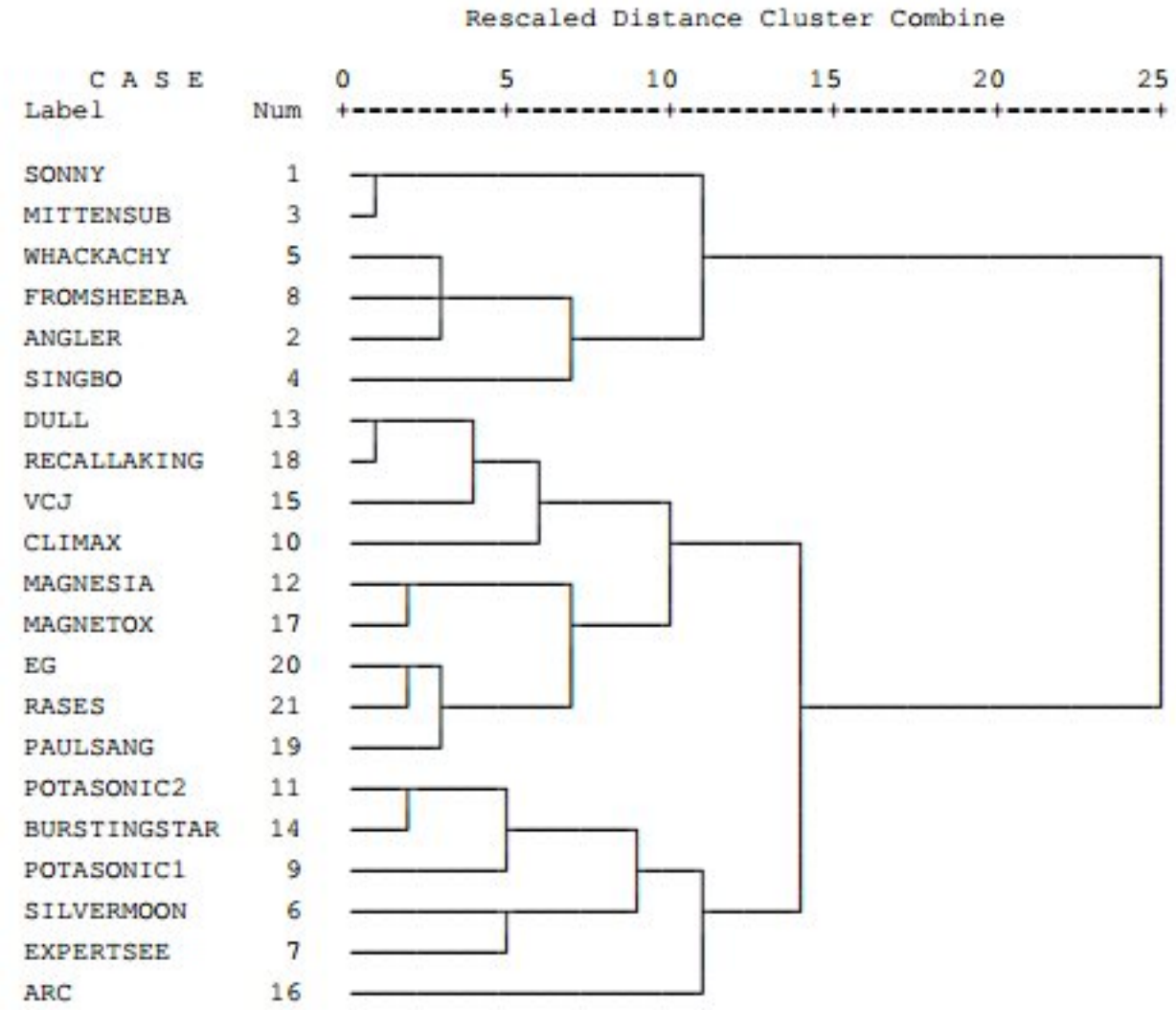

Figure 1: Corresponding Dendrogram

these clusters. To find a membership of a particular cluster simply trace backwards down the branches to the name.

3 Non-Hierarchical Clustering

(Remark: This material provides a dove-tail with the next topic: K-means Clustering)

- This method of clustering is very different from the hierarchical clustering and Ward method, which are applied when there is no prior knowledge of how many clusters there may be or what they are characterized by.
- The k-means clustering approach is used when you already have hypotheses concerning the number of clusters in your cases or variables. For example, you may want to specify exactly three clusters that are to be as distinct as possible.
- This is the type of research question that can be addressed by the k-means clustering algorithm. In general, the k-means method will produce the exact k different clusters demanded of greatest possible distinction. Very often, both the hierarchical and the k-means techniques are used successively.
    - Ward's method is used to get some sense of the possible number of clusters and the way they merge as seen from the dendrogram.
    - Then the clustering is rerun with only a chosen optimum number in which to place all the cases (i.e. k means clustering).
- Non-hierarchical cluster analysis tends to be used when large data sets are involved. It is sometimes preferred because it allows subjects to move from one cluster to another (this is not possible in hierarchical cluster analysis where a subject, once assigned, cannot move to a different cluster).
- Two disadvantages of non-hierarchical cluster analysis are:
    1. it is often diffcult to know how many clusters you are likely to have and therefore the analysis may have to be repeated several times
    2. it can be very sensitive to the choice of initial cluster centres. Again, it may be worth trying di?erent ones to see what impact this has.

3.1 Optimal Number of Clusters

One of the biggest problems with cluster analysis is identifying the optimum number of clusters. As the joining process continues, increasingly dissimilar clusters must be joined. i.e. the classification becomes increasingly artificial. Deciding upon the optimum number of clusters is largely subjective, although looking at a dendrogram would help.