Benchmarks Online

Skip Navigation Links

Page One

Campus Computing News

Holiday Hours

SPAM Blocking Measures Instituted on UNT Mailhosts

Whitelisting BULK: Mail Using GroupWise Rules

Your GroupWise Archive

An Online Campus Directory

JAWS 5.0 has arrived

New Equipment in the Adaptive Lab to Expand Learning Opportunities

SPSS 12.0 is Available Now from Academic Computing Services

Today's Cartoon

RSS Matters

The Network Connection

Link of the Month

WWW@UNT.EDU

Short Courses

IRC News

Staff Activities

Research and Statistical Support
University of North Texas

RSS Matters

Link to the last RSS article here: Got EBCDIC? Take This PROC and Call Me in the Morning. -- Ed.

Basics of Cluster Analysis

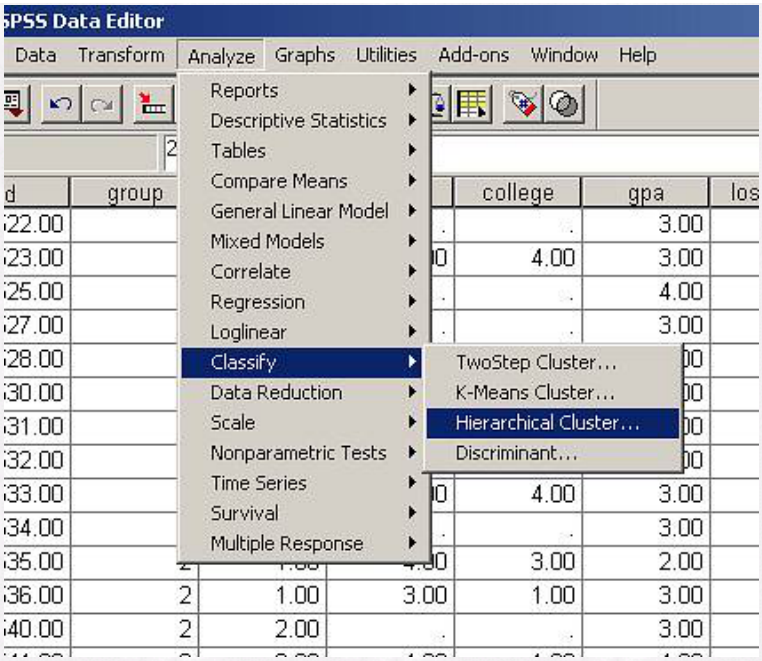
By Mike Clark, Research and Statistical Support Services Consultant

Cluster analysis is a statistical technique used to categorize cases or variables into like groups or 'clusters' with the usual goal to then proceed with other more conventional analyses using the information gained from the cluster analysis. Essentially it is a statistical method of classification and is often performed often when one has but a vague idea of what to expect from the data, and so is in that sense largely exploratory in nature. An example would be the biological classification of animals into kingdom, phylum, class and so forth down to species. The primary goal in cluster analysis is to find some meaningful structure in the data, and when it comes to cluster analysis, there are many options to choose from on how to go about this. This article will discuss approaches available in SPSS with a bit about S-Plus at the end.

Hierarchical cluster analysis

The goal of cluster analysis is to choose cluster membership that will minimize the variability within the clusters, and maximize the differences between clusters. In SPSS, the first step toward conducting a cluster analysis is to Click Analyze/Classify

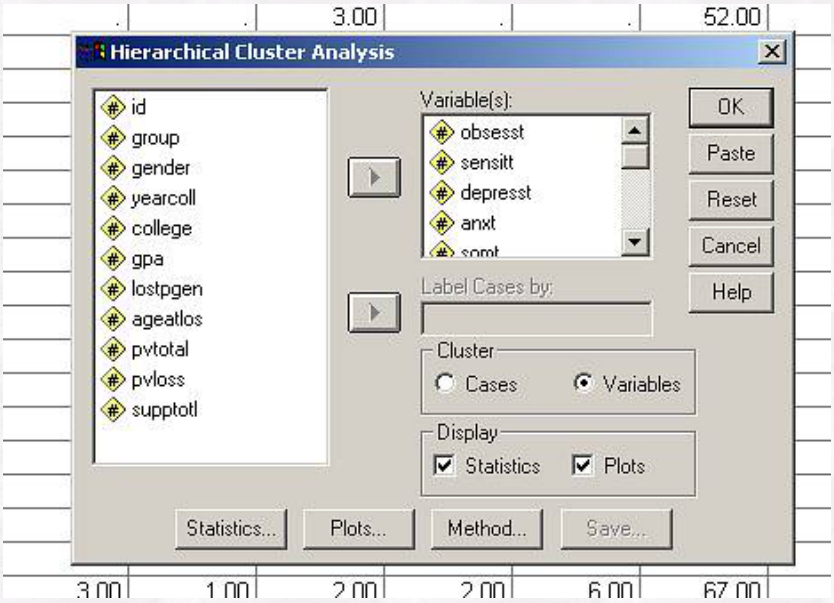
Figure1



Already we have three choices of clustering techniques available (the "Discriminant" refers to discriminant function analysis, which can be thought of as the confirmatory counterpart to the exploratory procedure of cluster analysis). We will start with hierarchical clustering first. This is to be used when one truly has little idea of what to expect with the data. First you'll select which variables are of interest, and then you'll tell SPSS whether it's the cases/individuals you are trying to classify or whether you are looking for structure within the variables (e.g. items of some questionnaire).

Clicking on 'statistics' will bring up a dialog box where we can choose some things to include in our output. The agglomeration schedule will show us at what point various cases become part of a cluster, and this will be different depending on the linkage method (see below) chosen. The proximity matrix will tell us how far apart the cases/variables are from one another and is a good option to choose. Also we can specify a certain number of clusters or range of clusters if we at least have an inkling as to how many groups to expect.

Figure 2

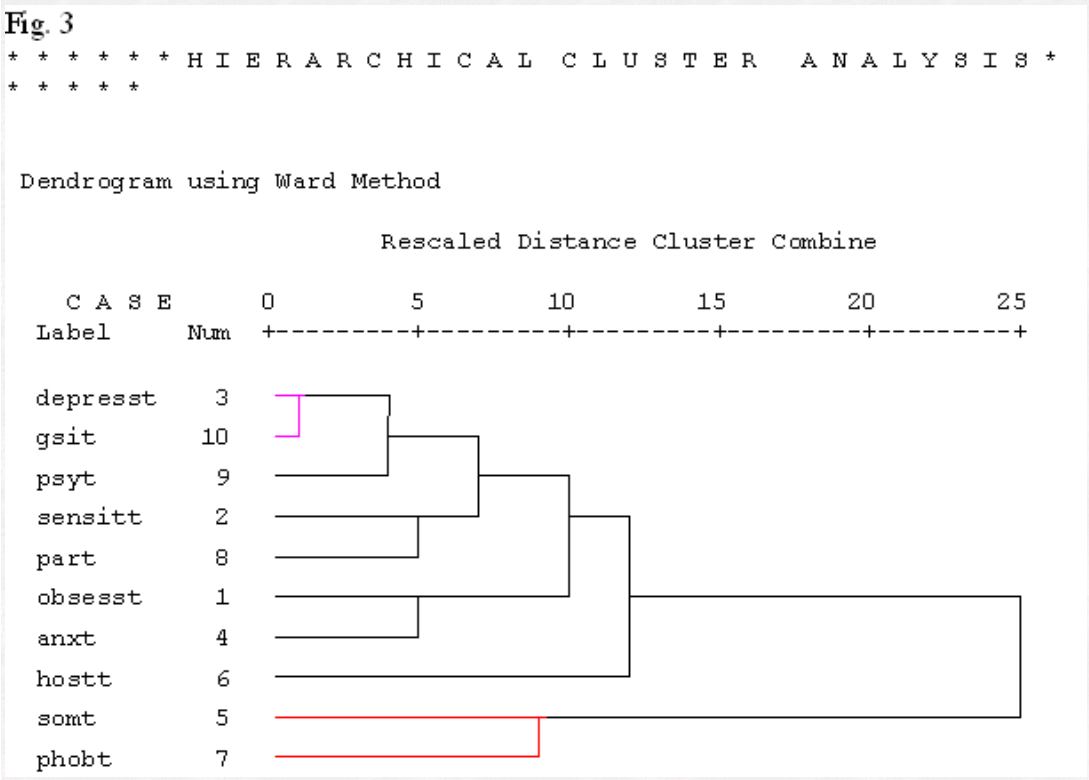


Next, clicking on 'plots' we can choose some options for visual display- the dendrogram or icicle plot. Both of these can be quite unwieldy when large numbers of cases or variables are used, in which case SPSS won't even display the entire dendrogram without an extra step. Suffice it to say, the dendrogram is this branching sort of thing that shows each variable or case as its own individual cluster at one end (left) and then how they are

combined until they are all are eventually a part of one cluster. The length of the branch shows how far apart each case or variable is from the other(s) in its cluster. In figure 3, the depression and gsit variables are vary close to one another (purple) while the somatization and phobia variables (red) are not as alike though still clump together to form their own cluster. If we went with a 2 cluster scenario, the somatization and phobia indices would make up one cluster while the other variables would make up the other cluster.

Figure 3

Fig. 3



The icicle plot below is our other 'visual' representation. I don't know about anyone else but these are just annoying to me and seem a throwback to when we had to do this sort of thing because it *was* about as visual as you could get back then. The way to read them is from bottom to top. Our variables are all nice and separate to start off with (except for depression and gsit which have already been combined), and when they are joined in the middle (i.e. an X fills up the space between them) they have joined a cluster with that neighbor. The look of this will depend on the linkage method chosen.

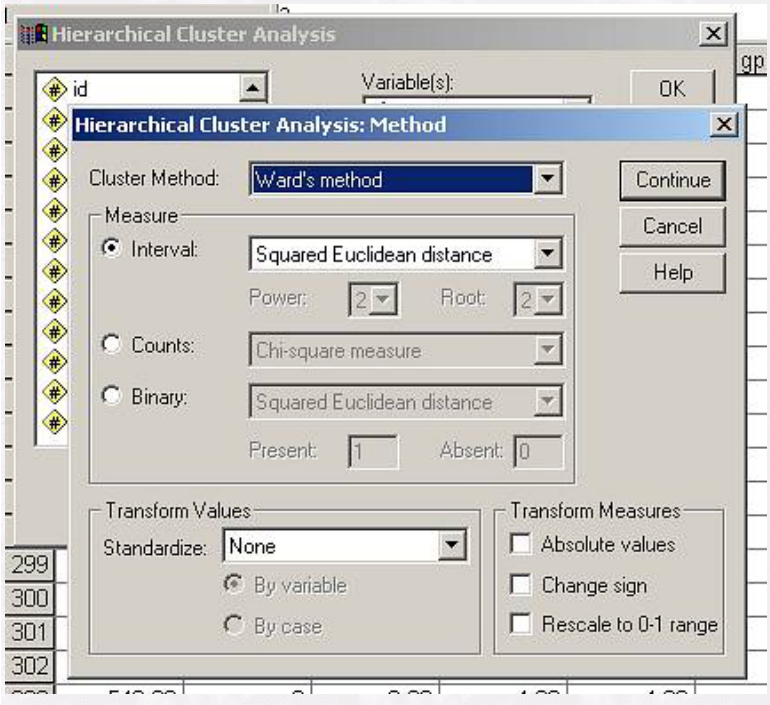
Figure 4

Vertical Icicle

Number of clusters	Case																		
	hostt		part		psyt		gsit		depresst		sensitt		anxt		obsesst		phobt		somt
1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
2	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X	X	X
3	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X		X
4	X		X	X	X	X	X	X	X	X	X	X	X	X	X		X		X
5	X		X	X	X	X	X	X	X	X	X		X	X	X		X		X
6	X		X	X	X	X	X	X	X	X	X		X		X		X		X
7	X		X		X	X	X	X	X	X	X		X		X		X		X
8	X		X		X	X	X	X	X		X		X		X		X		X
9	X		X		X		X	X	X		X		X		X		X		X

I've mentioned linkage method twice so it's about time to explain it, or them, I suppose. After choosing our stats and plots we click on the 'Method' button. The measure option refers to what you want use as your measure of distance between two points (cases or variables). If using count or binary data you'll need to select something from one of those menus. If you are using data that has different scales of measurement you may want to standardize or transform the values for analysis. The cluster (linkage) method is the option you'll choose that determines distances between clusters (rather than the individual cases), and when mini-clusters will combine into a larger cluster.

Figure 5



So which options should I choose? That's up to you really. There is no real standard though some are used more than others by convention and some believe that some particular methods might be better in particular situations. Sometimes you can choose different options and get the same result, sometimes fairly different ones. Use the past literature of studies of the sort your conducting as a guide, or read up on the different methods yourself and decide which one you like best. Use the one that ends up with clusters that adhere more to your theory even. Remember that cluster analysis is inherently exploratory. As long as you use accepted methods your results will be taken in the descriptive light in which they are presented. No one is going to change an entire way of practice or system of belief based on a cluster analysis alone.

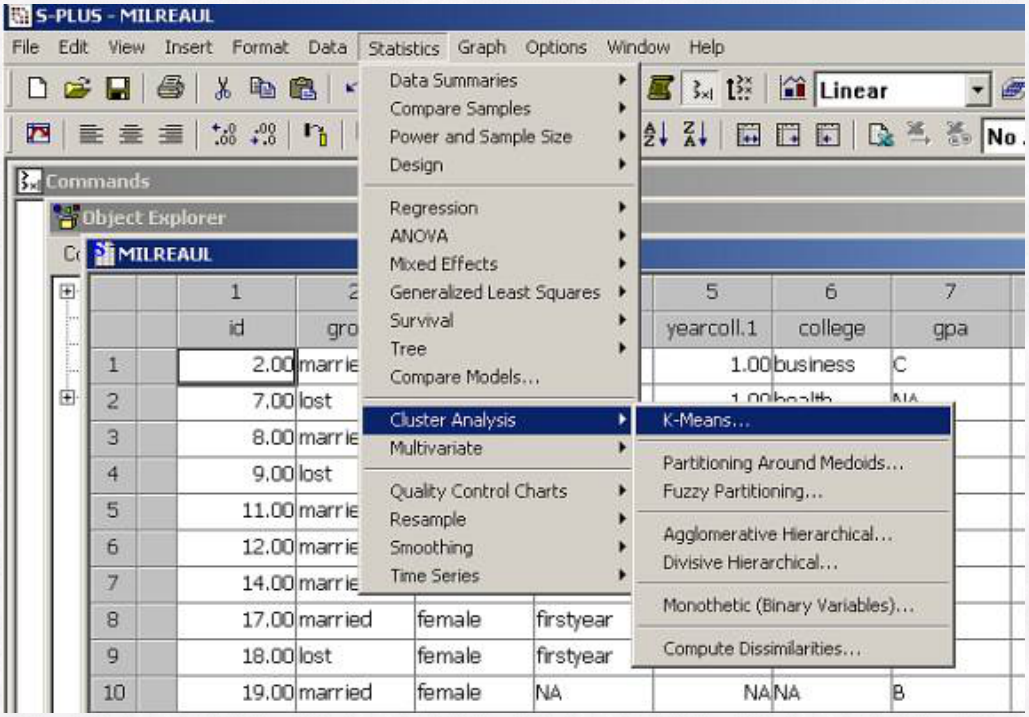
K-means and Two-step cluster analysis

The K-means analysis can be used when you already know how many clusters you are expecting to find. There are less options to choose from (it has basically done the leg work for you by means of using its own procedure) though will give you meaningful output that you won't get in the hierarchical method, such as cluster centers (the mean of a variable for that cluster), distances for a cluster from the other cluster centers, and ANOVA tables that may give information about which variables may be contributing more to the solution. One can save cluster membership and distance from cluster center, and then use the graph procedure to create a boxplot that can point to outliers. The K-means procedure also may be preferable for data containing a large number of cases.

The Two-step procedure will allow you to use categorical variables to formulate clusters. It is also useful for when you already have one or two possible solutions as far as number of clusters, and gives a choice of statistics that can be used to compare different solutions to determine the best number of clusters to retain. One must have normally distributed continuous variables and categorical variables must have a multinomial distribution.

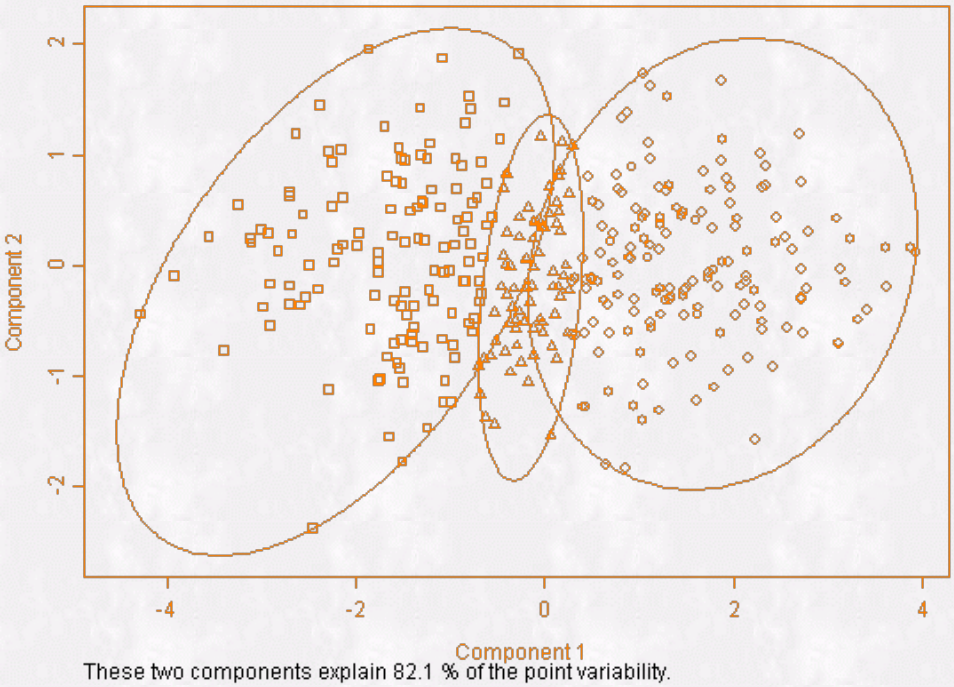
S-plus for cluster analysis

Figure 6



S-plus also provides quite a bit as far as cluster analysis is concerned. It offers a couple things that SPSS does not, e.g. fuzzy partitioning and visual representations of the cluster. With fuzzy partitioning, cases are allowed to have partial membership in more than one cluster (see below), and this might be important information to have depending on your research scenario. The visual representations are very helpful in getting a better feel for the data and spotting potential outliers.

Figure 7



Remember that the purpose of conducting cluster analysis is largely an exploratory one. It is useful for when we'd like to determine whether there are possibly distinct groups based on particular variables for which we have data, though we might also be interested in whether or not the variables themselves might be grouped in some meaningful fashion. The point is we may not be sure what is going on, and the goal would usually be to use the results of the cluster analysis to engage in more structured research later. There are many details to sort out in carrying out such an analysis, and the reader is encouraged to rely heavily on not only the literature on cluster analysis, but also the previous research in their field using such analyses to help them choose the best method for their situation.

References

Bailey, Kenneth (1975). Cluster Analysis. Sociological Methodology, Vol. 6, 59-128.