

Cluster analysis (cases)

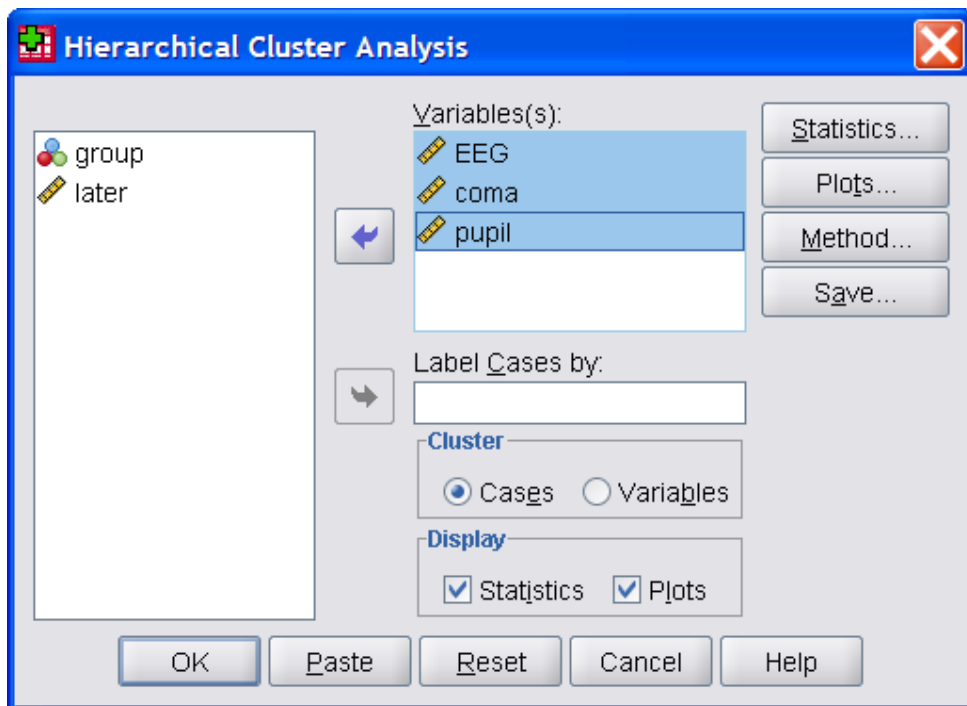
An alternative medical example: the (fabricated) Traumatic Brain

Injury data continued

We already know from our discriminant analysis that there are three groups in the TBI data (GROUP 1: back at work at 6 months, GROUP 2: reasonable recovery but not back at work at 6 months, GROUP 3: dependent or dead at 6 months). This time, we consider just the data on the classification variables, (EEG, COMA and PUPIL), and ignore the group membership data. The relevant data are in med.discriminant.sav on the book website. We will perform a cluster analysis and see whether our three groups emerge. As the scales for the three IVs (EEG, COMA and PUPIL) all happen to be similar, we do not need to standardize the scores for this dataset.

Requesting a cluster analysis in SPSS

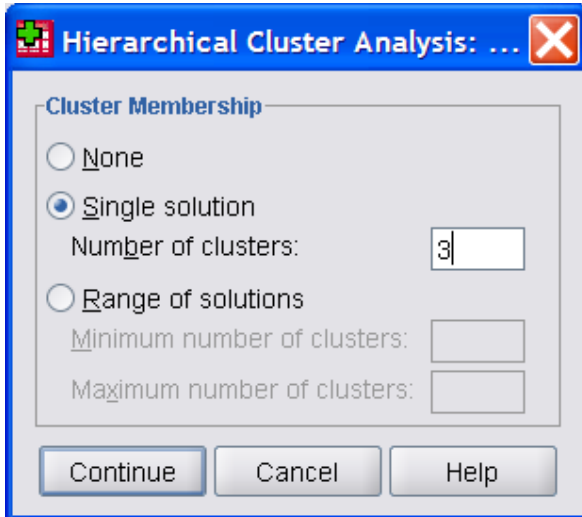
To start the cluster analysis in SPSS we go to **Analyze** on the menu bar, then **Classify**, then **Hierarchical Cluster ...**, and we get the dialog box shown in SPSS Dialog Box 10.1. We enter the variables EEG, COMA and PUPIL into the **Variable(s)** box using the arrow. Make sure the radio button for **Cases** is selected under **Cluster** (click it if not), since we are hoping to form clusters from our cases (not our variables).



SPSS Dialog Box 10.1. Hierarchical cluster analysis dialog box

Click the **Plots** button and then the **Dendrogram** box. The **Plots** dialog box also offers an *icicle plot*, which is another way to illustrate the step by step amalgamation of cases into clusters, but to start with click the **None** radio button to suppress this (see George and Mallery (2005, p. 274) for an example of an icicle plot). Click **Continue** to get back to SPSS Dialog Box 10.1.

The **Statistics** button allows you to decide what to display in the Output Viewer. We can accept the defaults here. If you press the **Save** button you get SPSS Dialog Box 10.2, where we have selected a single solution at the level of three clusters to be saved in the data window. We could instead have chosen to save cluster membership at a range of levels, perhaps from two to four clusters, by clicking the bottom radio button. This could be useful if you have little idea whether there are clusters in your data, or if there are, how many. Click **Continue** to get back to SPSS Dialog Box 10.1.



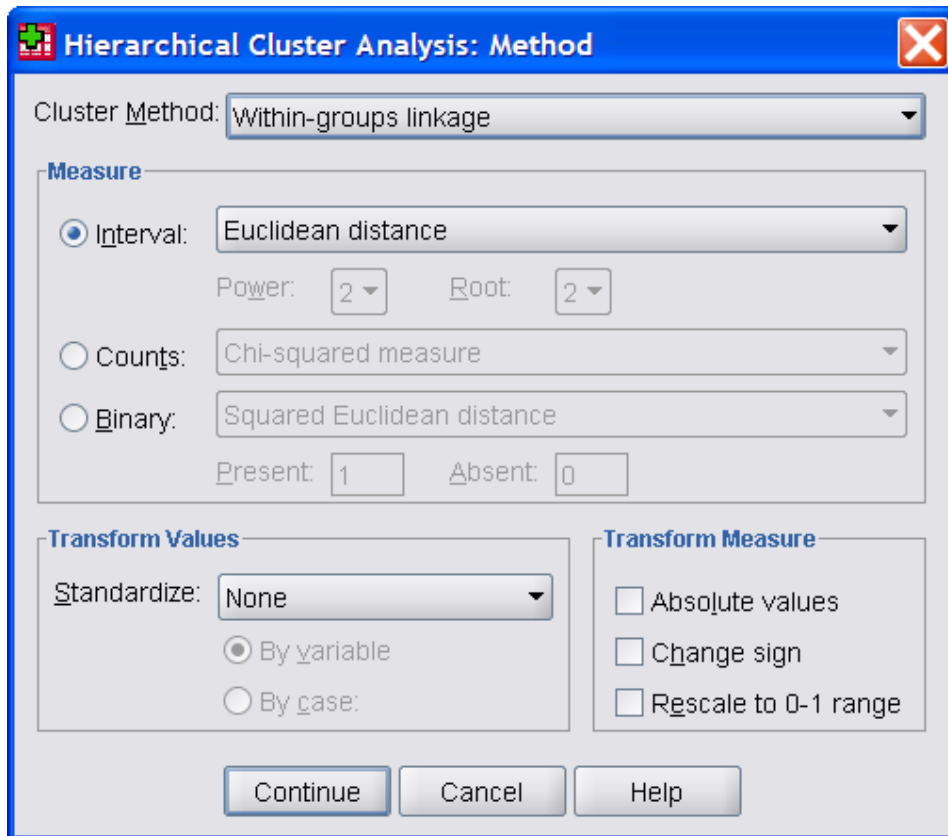
SPSS Dialog Box 10.2. Saving cluster membership in the data window

We have saved the difficult choices to last. Now press the **Method** button and get SPSS Dialog Box 10.3. First look at the radio buttons for **Measure**. Here you need to click the one that describes the kind of data you have. Our data are interval scales, and the arrow by the **Interval** box gives a choice of distance measures available. As discussed above, we will choose **Euclidean distance**. We will not transform the values since all our variables are scores on similar scales.

When we try the method on height and weight data which are measured on quite different scales, then we should standardize our variables to Z scores. To do this, use the arrow to replace **None** by **Z scores** in the **Standardize** box near the bottom of SPSS Dialog Box 10.3. Since we are clustering cases, leave the radio button at **By variable** (the default).

Now use the arrow by the **Cluster Method** box to see the list of available agglomeration methods. You may want to try several of these when you have a new problem and know little about your data. The first two on the list are averaging

methods, putting together cases or clusters in the way that minimizes the average within-group distance or maximizes the average between-group distance. Either of these averaging methods may be useful, as may one of the other averaging methods (median or centroid) or Ward's method. Here we have chosen **Within-groups linkage** as our clustering method.



SPSS Dialog Box 10.3. Choosing the distance measure and agglomeration method

Click **Continue** and **OK** to get the cluster analysis.

Understanding the output

First in the output is a table showing whether any cases have been omitted because of missing data. All ours have been included and we do not reproduce this table. Then comes the Agglomeration Schedule, shown in SPSS Output 10.3. Here you can see (in the Cluster Combined columns) that the first two cases to be joined in a cluster are 28

and 43. This cluster will now be known as number 28 (i.e., its lowest numbered case). Over to the right in the Next Stage column we see that this new cluster will be joined to something else at stage 4. Look down to stage 4 in the left-hand column and see that cluster 28 (cases 28 and 43) now joins case 9. This is now a three case cluster and will be referred to as Cluster 9. The Next Stage column tells us the next join is at stage 7, where we see that case 40 joins the cluster. The Next Stage column tells us the next join is at stage 15, where we see that cluster 19 joins. In the Stage Cluster First Appears (Cluster 2) column, you can see that cluster 19 first appeared at stage 2. Look back to stage 2, and in the Cluster Combined columns you see cases 19 and 42 forming cluster 19, then at stage 9 case 32 joins them. In this way the history of all the clusters can be followed. We have bolded the cases mentioned in this paragraph.

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	28	43	.000	0	0	4
2	19	42	.000	0	0	9
3	38	41	.000	0	0	8
4	9	28	.000	0	1	7
5	5	20	.000	0	0	10
6	10	17	.000	0	0	11
7	9	40	.500	4	0	15
8	33	38	.667	0	3	12
9	19	32	.667	2	0	15
10	5	26	.667	5	0	13
11	10	23	.667	6	0	14
12	33	35	.902	8	0	36
13	5	15	.902	10	0	28
14	1	10	.902	0	11	34
15	9	19	.908	7	9	32
16	37	54	1.000	0	0	47
17	24	53	1.000	0	0	44
18	47	51	1.000	0	0	42
19	44	50	1.000	0	0	29
20	46	48	1.000	0	0	37
21	39	45	1.000	0	0	38
22	27	36	1.000	0	0	30
23	16	31	1.000	0	0	41
24	4	21	1.000	0	0	31
25	11	18	1.000	0	0	34
26	2	14	1.000	0	0	44
27	3	8	1.000	0	0	39
28	5	7	1.024	13	0	33
29	44	49	1.138	19	0	35
30	12	27	1.138	0	22	43
31	4	6	1.138	24	0	40
32	9	34	1.172	15	0	36
33	5	22	1.203	28	0	45
34	1	11	1.218	14	25	40
35	29	44	1.305	0	29	48
36	9	33	1.372	32	12	41
37	46	52	1.382	20	0	42
38	30	39	1.382	0	21	43
39	3	13	1.382	27	0	45
40	1	4	1.478	34	31	46
41	9	16	1.598	36	23	47
42	46	47	1.598	37	18	52
43	12	30	1.619	30	38	48
44	2	24	1.639	26	17	49
45	3	5	1.702	39	33	49
46	1	25	1.729	40	0	51
47	9	37	1.785	41	16	50
48	12	29	1.856	43	35	50
49	2	3	2.323	44	45	51
50	9	12	2.439	47	48	52
51	1	2	2.724	46	49	53
52	9	46	3.039	50	42	53
53	1	9	3.565	51	52	0

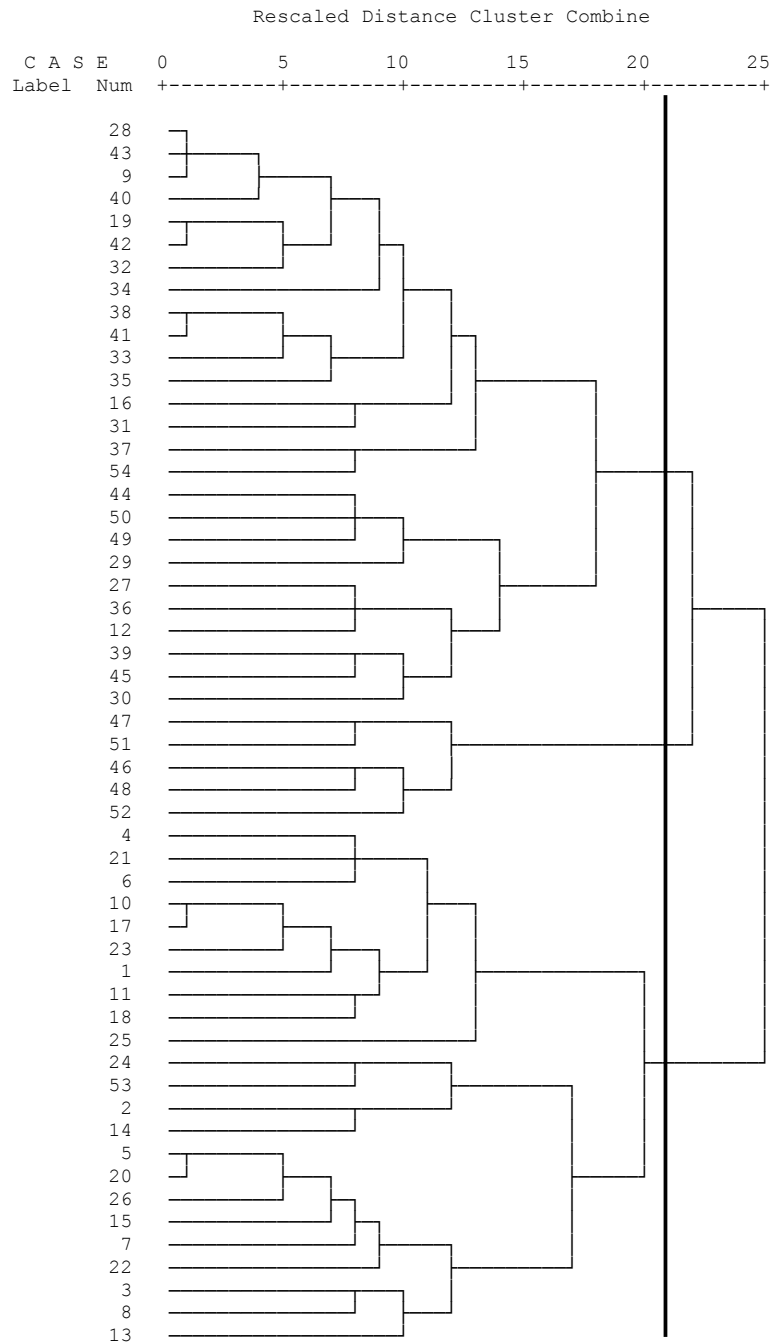
SPSS Output 10.3. Agglomeration of cases and clusters

Now look at the dendrogram, which appears next. When looking at it in the SPSS output window, you may have to click on it, then pull down the bottom of the frame to get the whole diagram into view. It is shown in SPSS Output 10.4. At the top of the dendrogram we see cases 28 and 43 and 9 joining, then being joined by case 40. Cases 19 and 42 join just below and are joined by case 32. These two clusters then join to form a seven case cluster. At the bottom of the diagram you can see cases 3 and 8 forming a cluster which is joined by case 13. In the dendrogram we have added a vertical line at a distance just above 20, where it crosses just three branches: perhaps we can see here that our cases fall into three groups. However, two of the groups combine only a little further to the right. We now compare the clusters with our original groups based on level of recovery discussed in the previous chapter. Since we saved the cluster membership at the level of three clusters, we shall find a new column in the data window displaying cluster membership at this level. The 23 cases labelled as cluster 1 are those at the bottom of the dendrogram, those labelled as cluster 3 are the 26 cases at the top and the small cluster of five cases in the middle is labelled cluster 2. It is easy to get a cross tabulation of GROUP and cluster membership (use **Analyze, Descriptive statistics, Crosstabs**). If you do this you find that all 18 GROUP 2 cases are in cluster 2, but so are four each of GROUP 1 and GROUP 3. Cluster 1 consists of the other 22 GROUP 1 cases and one case from GROUP 3. Cluster 3 has only five cases, all from GROUP 3 (you can see this small cluster in the middle of the dendrogram). So we see quite a close correspondence between the clusters and our groups.

If we repeat the cluster analysis but save cluster membership at the level of just two clusters we find that cluster 2 contains GROUP 2, all but one case from GROUP 3, and

four of the GROUP 1 cases. Cluster 1 contains the other 22 GROUP 1 cases and the remaining GROUP 3 case. We used just two recovery groups (combining GROUP 2 with GROUP 3) so that we tried to predict only whether a patient would return to work within six months. Our two clusters correspond quite well with these two groups, 22 of the 26 who got back to work going into cluster 1, and 27 of the 28 who did not get back to work going into cluster 2. So our cluster analysis does support our original decision to consider two or three outcome groupings for our patients.

Dendrogram using Average Linkage (Within Group)



SPSS Output 10.4. Dendrogram for 54 TBI patients

Comparing distance measures and linkage methods

We repeated our analysis of the TBI data using the Between-groups linkage method.

At the level of three clusters, membership was just as shown above for Within-groups

linkage, except that three of the group 1s were clustered with the group 3s, and also the group 3 case that was formerly clustered with the group 1s is now with the group 3s.

To use the Centroid or Ward methods we need to use squared Euclidean distance. We tried this and found that Centroid linkage reproduced our three clusters from Within-groups linkage except for putting two of the 23 cases from cluster 1 in with the small cluster of five cases. Ward linkage (with squared Euclidean distance) produced exactly the same result as Between-group linkage (with Euclidean distance).

Using Manhattan (Block) as the distance measure and Within-group linkage gave almost the same result at the level of three clusters as we obtained with Euclidean distance, just two cases from cluster 1 in with the small cluster of five cases. The similarity of the results obtained using several methods gives us some comfort that when the data is structured in clusters, as ours is here, then the choice of distance measure and linkage method may not affect our results very much. Experimenting with these changes can help you to see whether your results seem fairly robust or are very dependent on the choices you make for the distance measure and linkage method.

Selection of cases

In addition to the problems of choosing variables on which to base distance measures, choosing distance measures and deciding which clustering method to use, there is the usual problem of deciding how many cases are needed and selecting a suitable sample. Unlike factor analysis, it is the cases that are usually grouped on the basis of some variables that identify characteristics of the cases, and we do not necessarily

need a large number of cases to identify clusters. If the data do have some structure and there are subsets to be found, then the choice of cases and even the number of them may not be important as long as all subsets are represented.

In order to see the effect of reducing the number of cases selected, we made an extra variable called `SELECT` and added it to the data on the three test scores. We gave alternate cases the values of 1 and 2, starting with 1, on `SELECT`. Then we tried using **Data** and **Select cases** to select just those with `SELECT = 1`, and applied the hierarchical cluster analysis just to this half of the dataset (using Euclidean distance and Within-groups linkage). At the level of three clusters, all but one of the cases in this half of the dataset fell into the same clusters as they did when the full dataset was used. The same was true when we repeated the analysis with `SELECT = 2`. It is easy for you to try this experiment for yourself. You could of course do the same sort of thing with a real dataset, and perhaps have added confidence in any clusters you identify if you can still find them using just part of the dataset.

Using discriminant analysis to confirm findings

One way to investigate possible clusters further is to see whether a discriminant analysis can classify the cases successfully into the clusters. To try this, all you need to do is save the cluster membership in the data window for the number of clusters you think is appropriate, then use the cluster membership as the grouping variable in a discriminant analysis. We did this for our TBI data: we saved cluster membership for three clusters, then used this as the grouping variable for a discriminant analysis. We used a stepwise method to add the variables `EEG`, `COMA` and `PUPIL`. All three variables were entered and all cases were correctly classified into their clusters.

So, looking at the dendrogram in SPSS Output 10.4, we might think our cases fall into three clusters, and discriminant analysis can classify all of them correctly using our variables. However, four clusters, or perhaps some different number, may also be possible. We considered the possibility that discriminant analysis would classify the cases into four clusters.

Still looking at the dendrogram in SPSS Output 10.4, you find that by placing a vertical ruler carefully you can find a position where it crosses four branches. At this distance level, the 23 cases that form cluster 1 at the bottom of the dendrogram, form two clusters. How would discriminant analysis perform on these four clusters? To try it, all you need do is save the cluster membership in the data window for four clusters, and use this as the grouping variable for a discriminant analysis. We did this, again using a stepwise method to enter the variables EEG, COMA and PUPIL. All but one case was correctly classified (see SPSS Output 10.5). So, discriminant analysis can classify the cases into four clusters quite successfully, though not quite as well as into three.

Classification Results^a

		Average Lin...	Predicted Group Membership				Total
			1	2	3	4	
Original	Count	1	10	0	0	0	10
		2	0	13	0	0	13
		3	0	0	25	1	26
		4	0	0	0	5	5
%		1	100.0	.0	.0	.0	100.0
		2	.0	100.0	.0	.0	100.0
		3	.0	.0	96.2	3.8	100.0
		4	.0	.0	.0	100.0	100.0

a. 98.1% of original grouped cases correctly classified.

SPSS Output 10.5. Discrimination into four clusters using the variables EEG, COMA and PUPIL