How to export a distance matrix

Most datasets contain variables measured on different scale levels. This creates the problem of how to simultaneously incorporate these variables into one cluster analysis. One way is to compute distinct distance matrices for each group of variables; that is, one distance matrix based on, for example, ordinal variables and another based on nominal variables. To illustrate this, we make use of the SPSS syntax. For example, consider the dataset *cars.sav* from our example application in Chapter 9. Run a hierarchical clustering procedure by clicking Analyze ▸ Classify ▸ Hierarchical Cluster; this opens the dialog box shown in Figure A9.1.

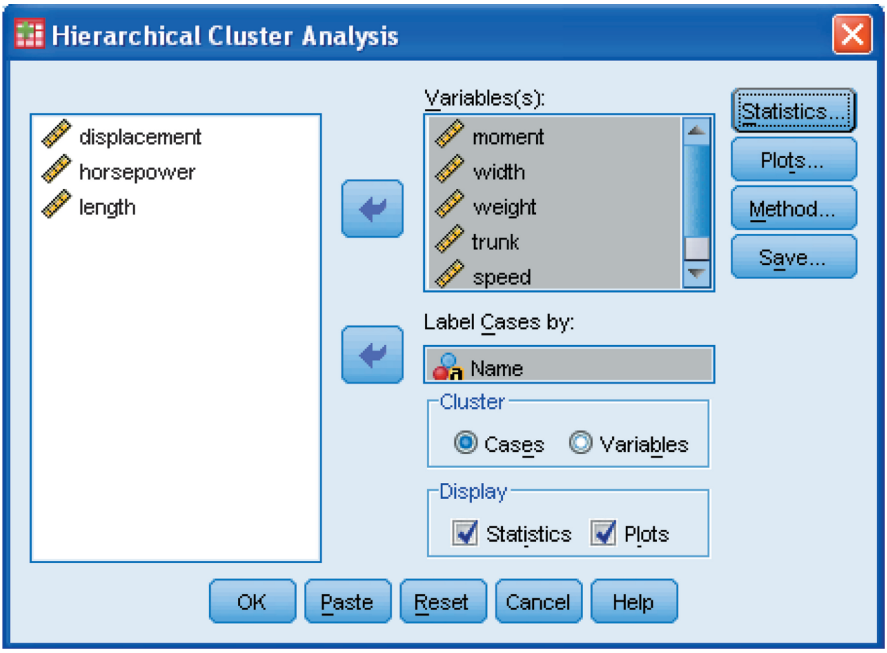


Figure A9.1 Hierarchical cluster analysis dialog box.

Move the variables *moment, width, weight, trunk, speed, and acceleration* into the **Variables** box and specify the name in the box **Label Cases by**. Use the **Number of Clusters** option under **Statistics** to pre-specify that three clusters should be retained from the data. Click on **Method** and specify the cluster method, distance measure, and the type of standardization of the values. In this example, we use the single linkage method (**Nearest neighbor**) based on **Euclidean distances**. Since the variables are measured on different levels (e.g., speed versus weight), make sure you standardize the variables, using, for example, the option **Range -1 to 1 (by variable)** in the **Transform Values** submenu.

Instead of clicking **OK** in the main menu, click on the **PASTE** button. This will open the syntax window shown in Figure A9.2. As SPSS calculates the distances using a separate procedure (which we don't notice when running the clustering procedure using the menu options), there are several syntax commands in the file. It is not too hard to guess that CLUSTER is the command representing the hierarchical clustering procedure. To export the distance matrix (in SPSS it is called the proximity matrix) into a new dataset, simply add the line "\MATRIX OUT (*)." to the syntax code of the CLUSTER command. Make sure that there is a decimal point at the end of this line, as decimal points always mark the end of a syntax command. The asterisk (*) in the parentheses requests SPSS to open a new data window. Alternatively, you could specify a path such as 'c:\distances.sav', which would save the file distances.sav to c:\ (make sure that you use a standard single quote\ inverted comma). Your final syntax code should look like the example in Figure A9.2.

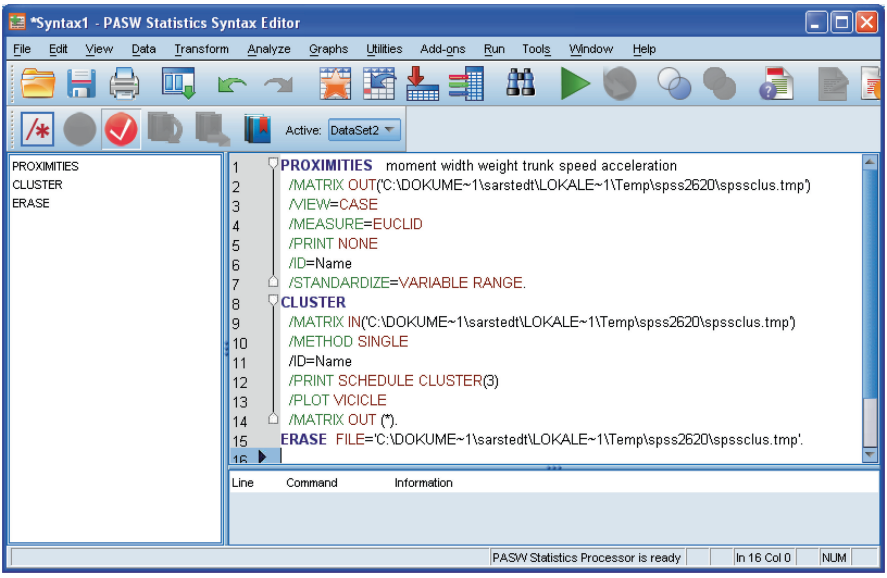
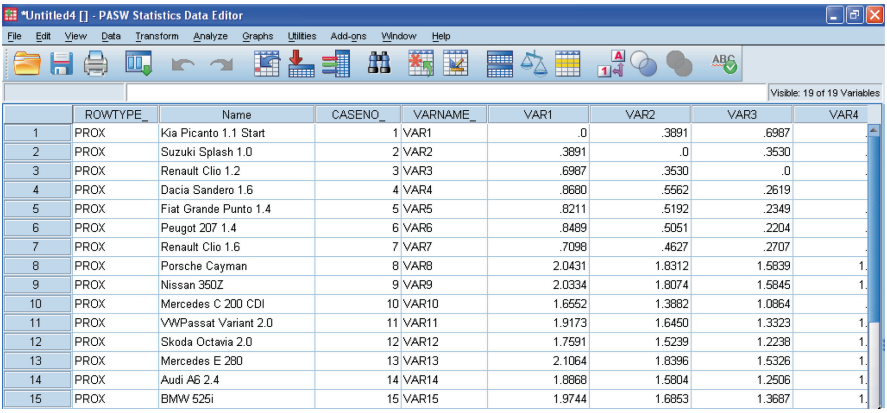


Figure A9.2 Hierarchical cluster analysis syntax code.

Now start the analysis by clicking on Run ▸ All. This will provide you with the SPSS output and will also open a new window containing the distance matrix (Figure A9.3).



	ROWTYPE_	Name	CASENO_	VARNAME_	VAR1	VAR2	VAR3	VAR4
1	PROX	Kia Picanto 1.1 Start	1	VAR1	.0	.3891	.6987	
2	PROX	Suzuki Splash 1.0	2	VAR2	.3891	.0	.3530	
3	PROX	Renault Clio 1.2	3	VAR3	.6987	.3530	.0	
4	PROX	Dacia Sandero 1.6	4	VAR4	.8680	.5562	.2619	
5	PROX	Fiat Grande Punto 1.4	5	VAR5	.8211	.5192	.2349	
6	PROX	Peugot 207 1.4	6	VAR6	.8489	.5051	.2204	
7	PROX	Renault Clio 1.6	7	VAR7	.7098	.4627	.2707	
8	PROX	Porsche Cayman	8	VAR8	2.0431	1.8312	1.5839	1.
9	PROX	Nissan 350Z	9	VAR9	2.0334	1.8074	1.5845	1.
10	PROX	Mercedes C 200 CDI	10	VAR10	1.6552	1.3882	1.0864	
11	PROX	VWPassat Variant 2.0	11	VAR11	1.9173	1.6450	1.3323	1.
12	PROX	Skoda Octavia 2.0	12	VAR12	1.7591	1.5239	1.2238	1.
13	PROX	Mercedes E 280	13	VAR13	2.1064	1.8396	1.5326	1.
14	PROX	Audi A6 2.4	14	VAR14	1.8868	1.5804	1.2506	1.
15	PROX	BMW 525i	15	VAR15	1.9744	1.6853	1.3687	1.

Figure A9.3 Distance matrix in SPSS.

On first sight, the distance matrix might look a bit cryptic, but essentially it has the same form as those that we computed manually in the numerical example. For example, the distance from case 1 (Kia Picanto 1.1 Start) to case 2 (Suzuki Splash 1.0) is 0.3891 units. We can now proceed and save this distance matrix by clicking on File ▸ Save.

We could now use this distance matrix as input for another cluster analysis. In this case, we would have to change the syntax command from MATRIX OUT to MATRIX IN and adjust the expression in the parentheses. For example, the MATRIX IN subcommand "/MATRIX IN ('c:\distances.sav')" would use the file distances.sav located in c:\ to run the cluster analysis. However, when using the MATRIX IN command, you have to make sure that the SPSS file has the right format (i.e. it includes columns labeled *ROWTYPE_*, *Name* etc., as in Figure A9.3).

We could now compute distinct distance matrices of differently scaled variables and compute the (weighted) arithmetic mean of the distances by using, for example, a spreadsheet program such as Microsoft Excel. We could then use this newly computed distance matrix (containing the mean distances) as input for the cluster analysis.