Subject

Showing the complementarity of the data mining (clustering) and visualization (factorial analysis) methods.

Dataset

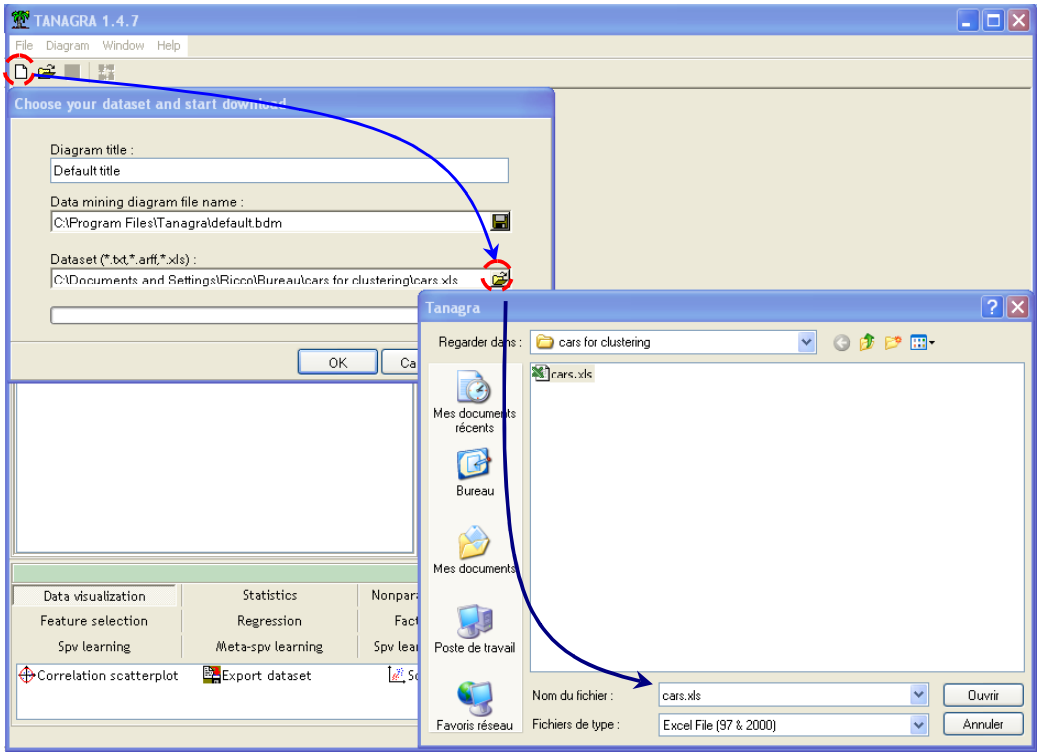
We use the CARS.XLS dataset. It contains 38 vehicles and various descriptors. The goal is to build homogenous clusters of vehicles.

Some results about this dataset are available on the net http://lib.stat.cmu.edu/DASL/Stories/ClusteringCars.html. It seems there are three main clusters in this dataset. We see in this tutorial if we obtain the same result.

HAC with TANAGRA

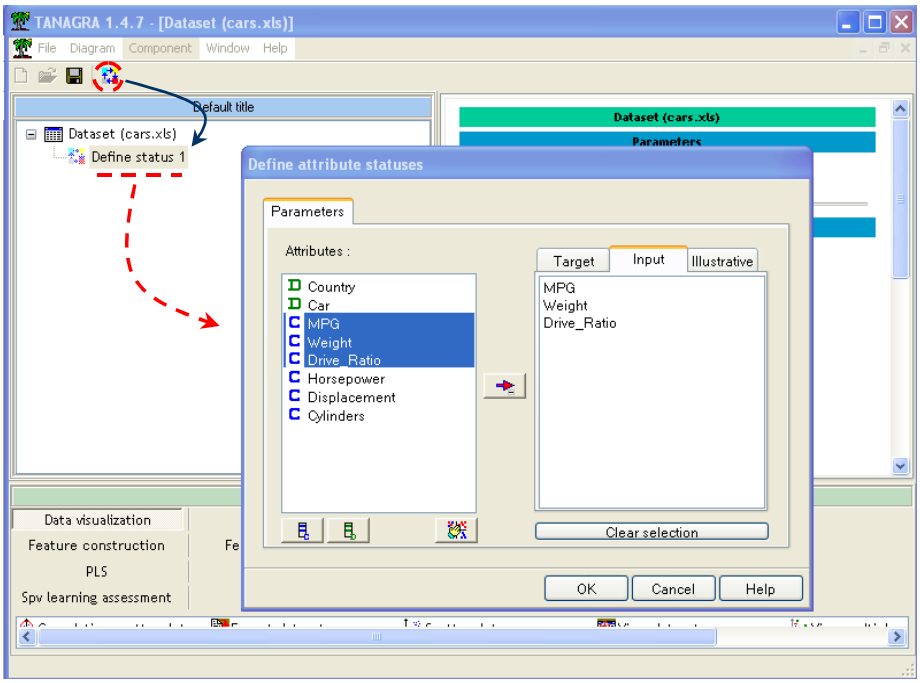
Downloading the dataset

We must build a new diagram and import the dataset. To do that, we click on the FILE/NEW menu and select the file CARS.XLS.



INPUT attributes

We use the DEFINE STATUS component in order to define the INPUT attributes. We select the following descriptors: MPG, WEIGHT and DRIVE_RATIO.



HAC

We use a hierarchical agglomerative clustering algorithm[1] in this tutorial. We obtain a dendrogram, which shows the level of each aggregation.

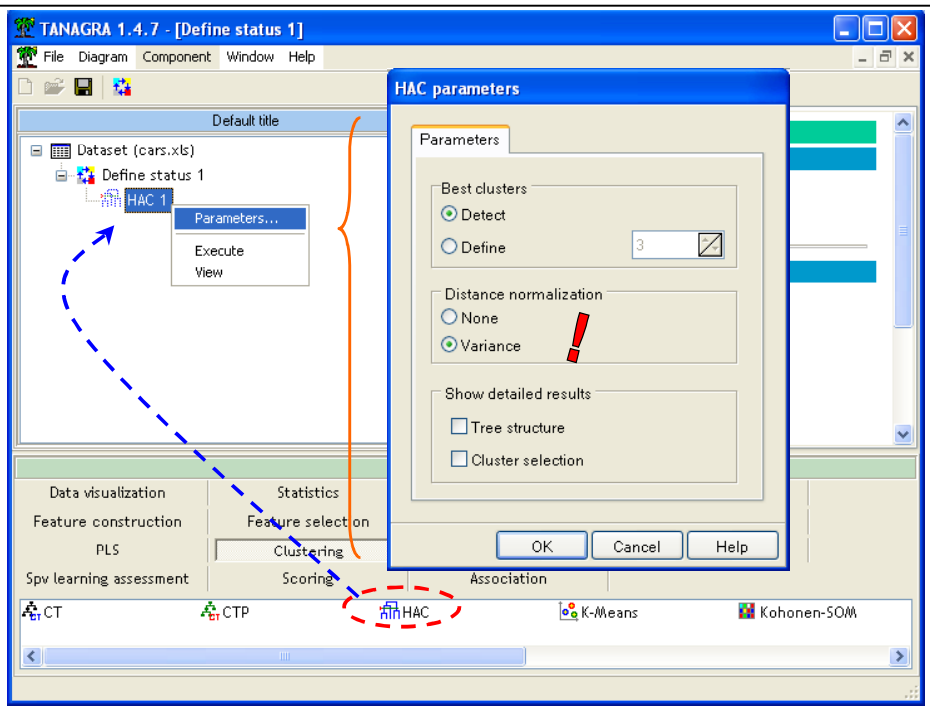
TANAGRA can detect automatically the highest jump in the dendrogram. By default, it proposes the corresponding clusters.

Because we have a few examples in this dataset, the leaves of the tree (dendrogram) can be each example. If we have many examples (thousands…), it is more judicious to use the principle of hybrid clustering[2]: first, low-level clusters are built from fast clustering algorithms such as K-MEANS or SOM; second, HAC starts from these clusters and builds the dendrogram. The computing time is clearly improved while preserving quality of the results.

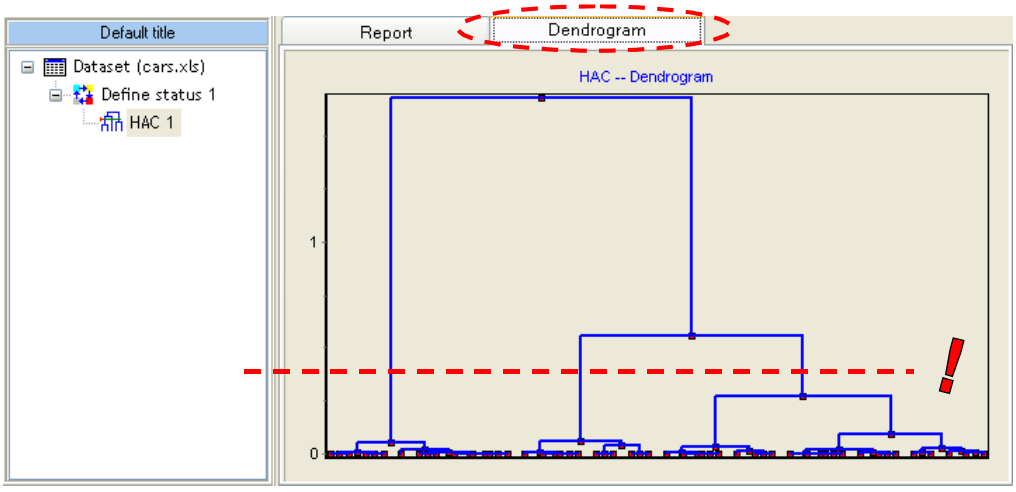
We see below the corresponding stream diagram. We standardize the dataset because they are not expressed in the same units.

[1] http://en.wikipedia.org/wiki/Data_clustering

[2] http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/enHAC_IRIS.pdf

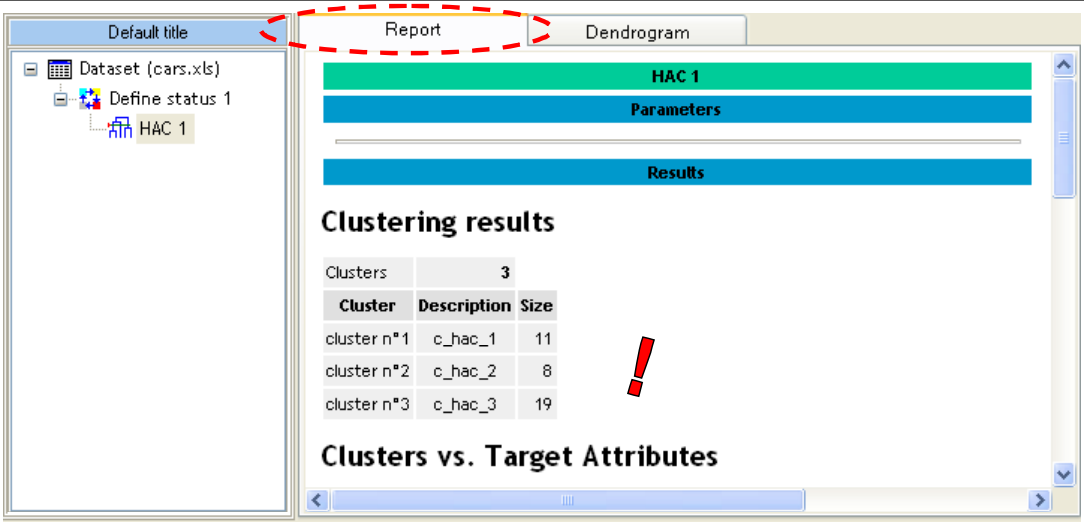


We click on the VIEW menu, we obtain the following results. The dendrogram is now available (1.4.8 version). The clustering in three classes seems to be indeed the most obvious[3].



TANAGRA detects this solution.

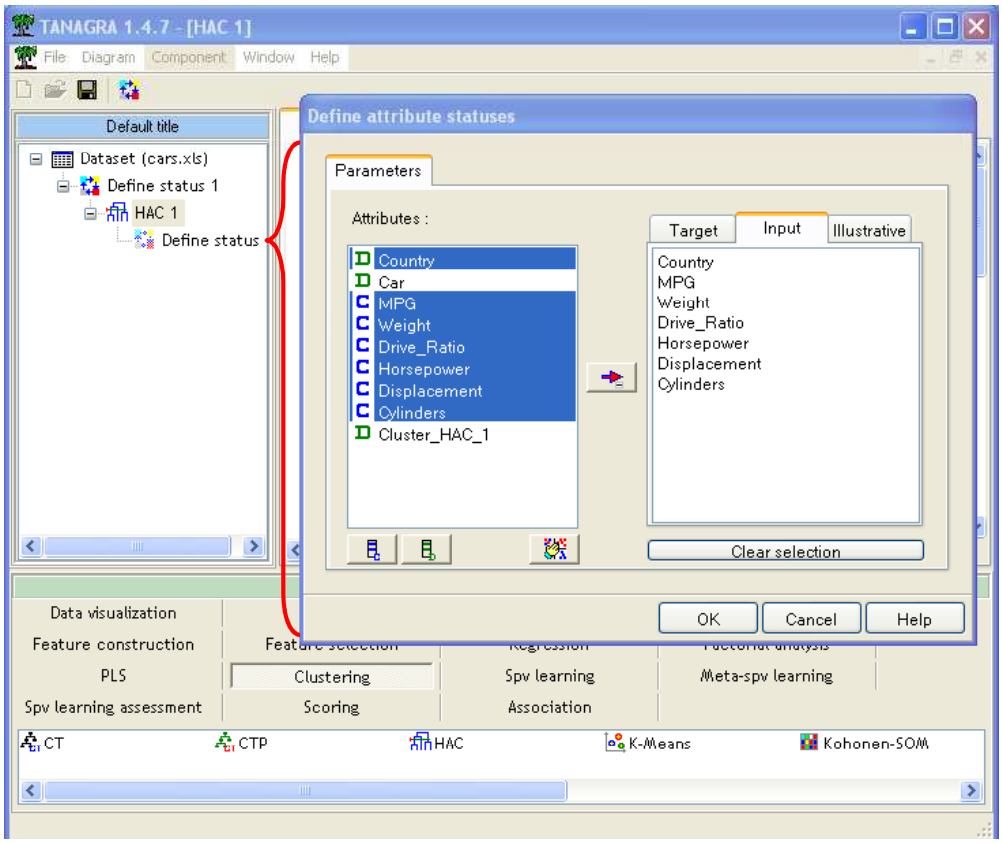
[3] In the majority of the cases, the subdivision in two classes produces the highest jump. It is often an artifact due to the fact it is the first subdivision of the data. For this reason, TANAGRA tries to detect the most important jump only for the subdivisions into 3, 4, etc clusters.



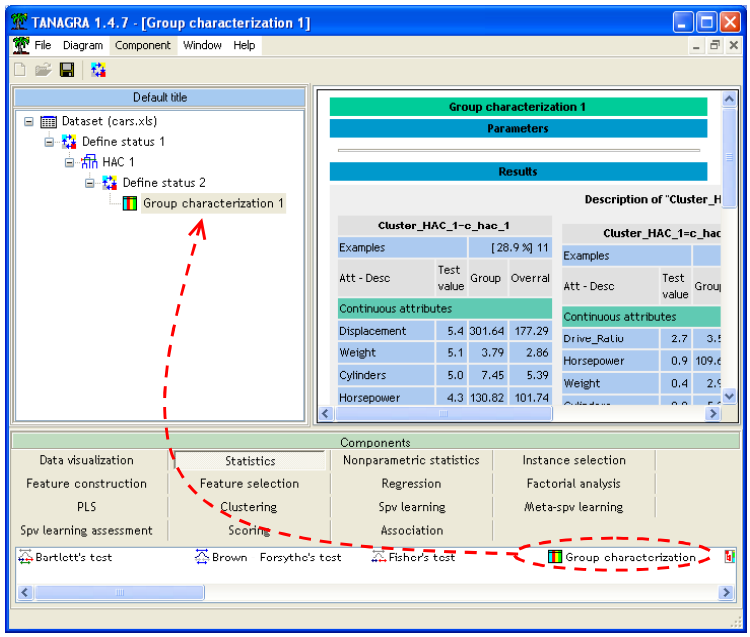
Group characterization

In the following step, we must characterize the clusters. The GROUP CHARACTERIZATION component allows comparing the mean of the continuous variables (the proportion for the values of discrete attribute) in the whole dataset and in the clusters.

To do that, we add the DEFINE STATUS component. We set as TARGET the cluster attribute from the HAC; we set as INPUT the other attributes, except the CAR attribute.



Then, we add the GROUP CHARACTERIZATION component.



To have a better visualization of the results, you can copy them in a spreadsheet.

The first cluster (C_HAC_1) corresponds to the large vehicles. They are heavy, powerful and use much fuel. All cars in this cluster are US vehicles, 50% of US cars are in this cluster.

Cluster_HAC_1=c_hac_1			
Examples			[28.9 %] 11
Att - Desc	Test value	Group	Overral
Continuous attributes			
Displacement	5.4	301.64	177.29
Weight	5.1	3.79	2.86
Cylinders	5	7.45	5.39
Horsepower	4.3	130.82	101.74
MPG	-4.1	17.88	24.76
Drive_Ratio	-4.4	2.5	3.09
Discrete attributes			
Country=U.S.	3.3	[50.0 %] 100.0 %	57.90%
Country=Italy	-0.6	[0.0 %] 0.0 %	2.60%
Country=France	-0.6	[0.0 %] 0.0 %	2.60%
Country=Sweden	-0.9	[0.0 %] 0.0 %	5.30%
Country=Germany	-1.5	[0.0 %] 0.0 %	13.20%
Country=Japan	-1.8	[0.0 %] 0.0 %	18.40%

The second cluster (C_HAC_2) corresponds to the "middle" vehicles. They are mainly European cars. They are moderate horsepower and size. They have a high DRIVE_RATIO. That is a characteristic of European cars where one likes responsive cars.

Cluster_HAC_1=c_hac_2			
Examples		[21.1 %] 8	
Att - Desc	Test value	Group	Overral
Continuous attributes			
Drive_Ratio	2.7	3.53	3.09
Horsepower	0.9	109.63	101.74
Weight	0.4	2.95	2.86
Cylinders	0	5.38	5.39
Displacement	-0.9	152	177.29
MPG	-2.2	20.16	24.76
Discrete attributes			
Country=Sweden	2.8	[100.0 %] 25.0 %	5.30%
Country=France	1.9	[100.0 %] 12.5 %	2.60%
Country=Germany	1.1	[40.0 %] 25.0 %	13.20%
Country=Japan	-0.5	[14.3 %] 12.5 %	18.40%
Country=Italy	-0.5	[0.0 %] 0.0 %	2.60%
Country=U.S.	-2.1	[9.1 %] 25.0 %	57.90%

At last, the third cluster (C_HAC_3) corresponds to "small" cars. These are vehicles of small size, light, not very powerful and have low consumption.

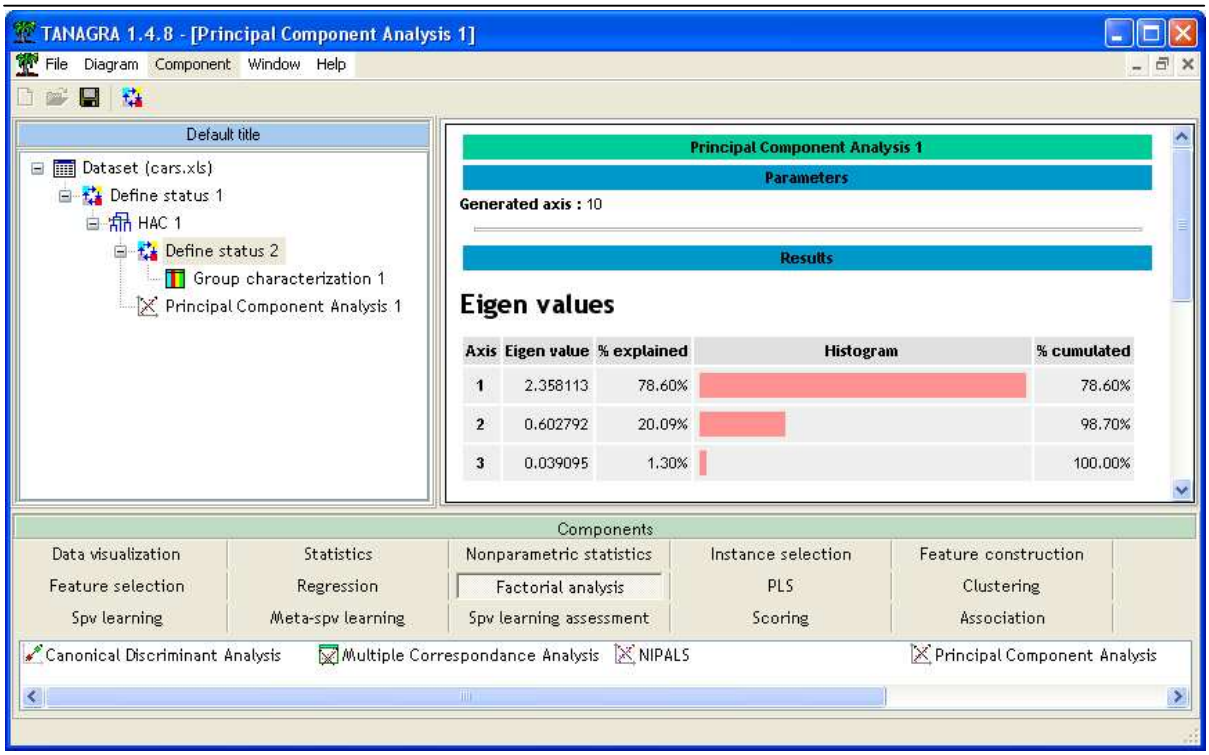
Cluster_HAC_1=c_hac_3			
Examples		[50.0 %] 19	
Att - Desc	Test value	Group	Overral
Continuous attributes			
MPG	5.5	30.68	24.76
Drive_Ratio	1.8	3.25	3.09
Displacement	-4.2	115.95	177.29
Cylinders	-4.5	4.21	5.39
Horsepower	-4.6	81.58	101.74
Weight	-4.9	2.29	2.86
Discrete attributes			
Country=Japan	2.1	[85.7 %] 31.6 %	18.40%
Country=Italy	1	[100.0 %] 5.3 %	2.60%
Country=Germany	0.5	[60.0 %] 15.8 %	13.20%
Country=France	-1	[0.0 %] 0.0 %	2.60%
Country=U.S.	-1.3	[40.9 %] 47.4 %	57.90%
Country=Sweden	-1.4	[0.0 %] 0.0 %	5.30%

According to the results on the website, we obtain three clusters: "large", "middle" and "small" cars.

PCA with TANAGRA

PCA

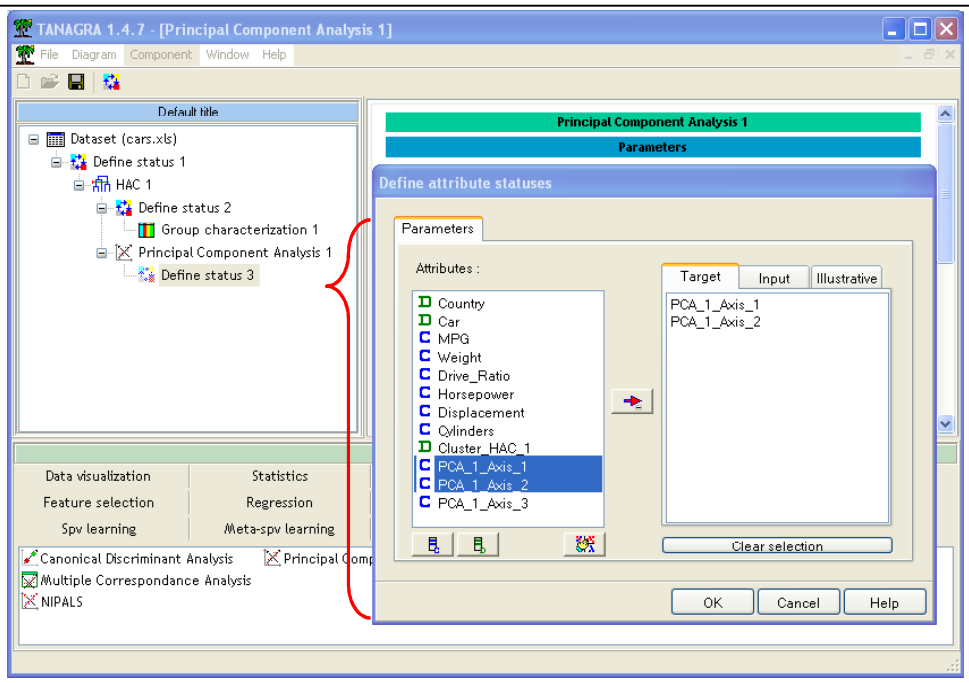
In order to visualize the clusters, we can use factorial analysis. We add a PCA (Principal Component Analysis) component in the diagram.



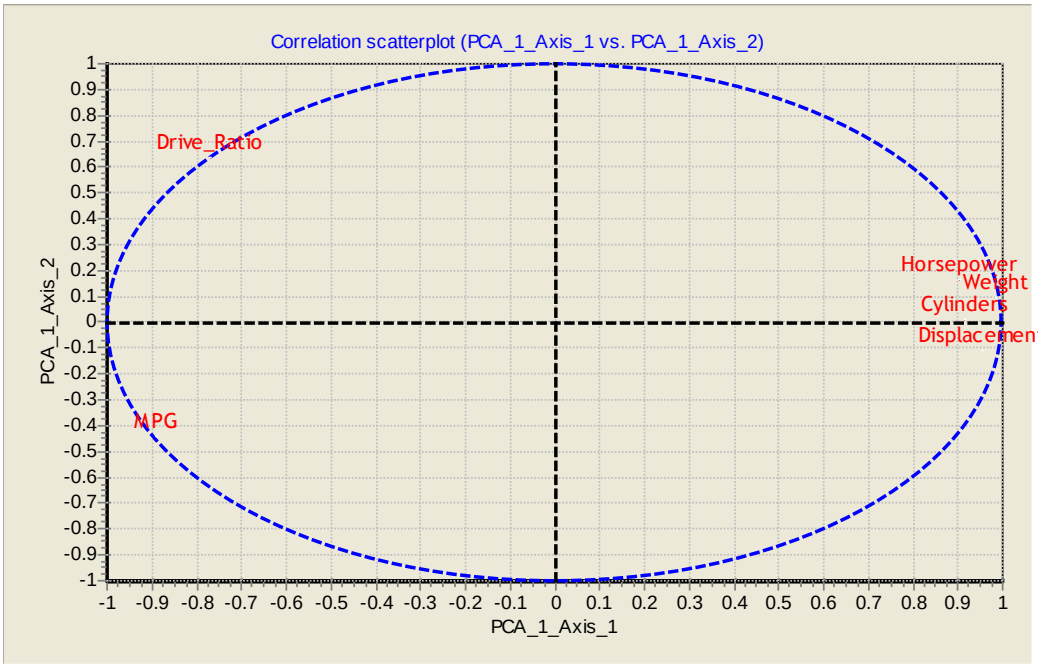
The first two axes summarize 98% of the available information. The scatter plot on these axes will be rather faithful to the position of the examples in the initial space.

Correlation circle

In order to obtain the correlation circle, we add again a DEFINE STATUS component and set as TARGET the first two axes. We set as INPUT the other descriptors: those which make it possible to build the axes (MPG, WEIGHT, DRIVE_RATIO); and those which one uses to interpret the axes (HORSEPOWER, DISPALCEMENT, CYLINDERS).

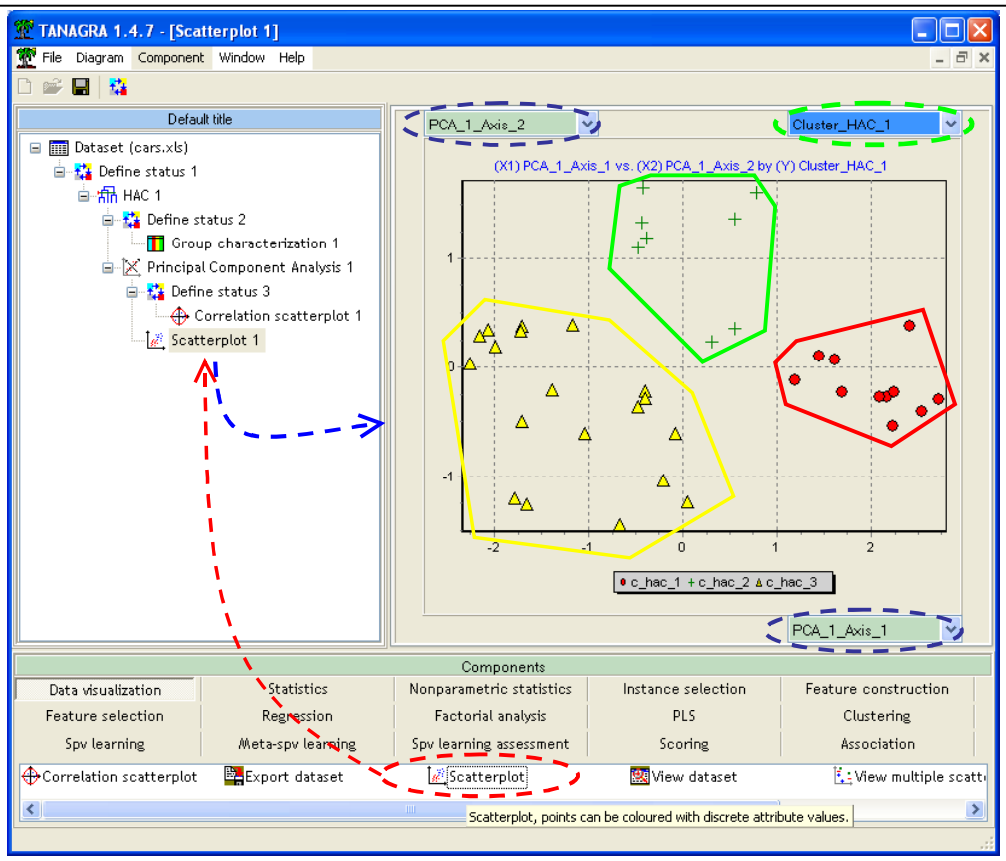


Then, we insert a CORRELATION SCATTERPLOT in order to visualize the correlation circle.



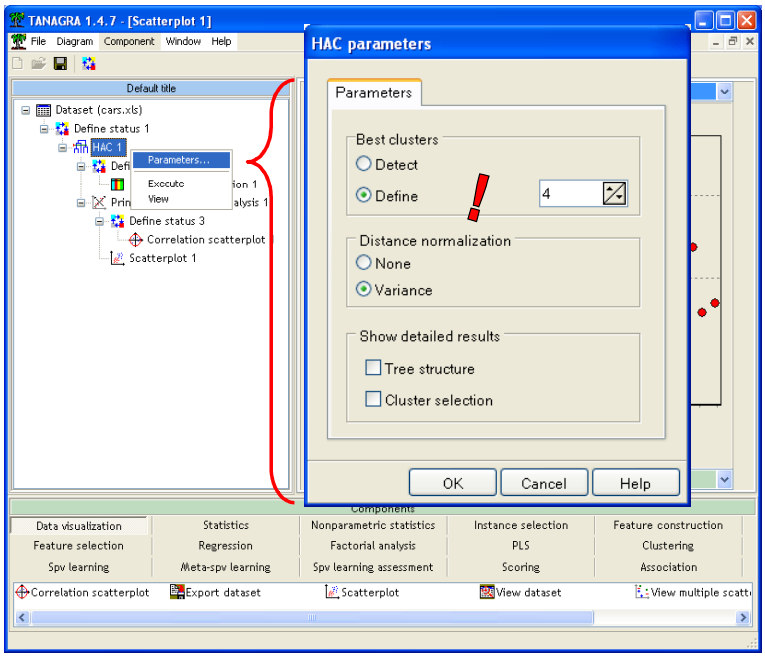
Plotting the observations

In the next step, we want to plot the examples in the new projection space. The goal is to obtain a better visualization of the clusters. We add a SCATTERPLOT component in the diagram. We distinguish the three clusters well.

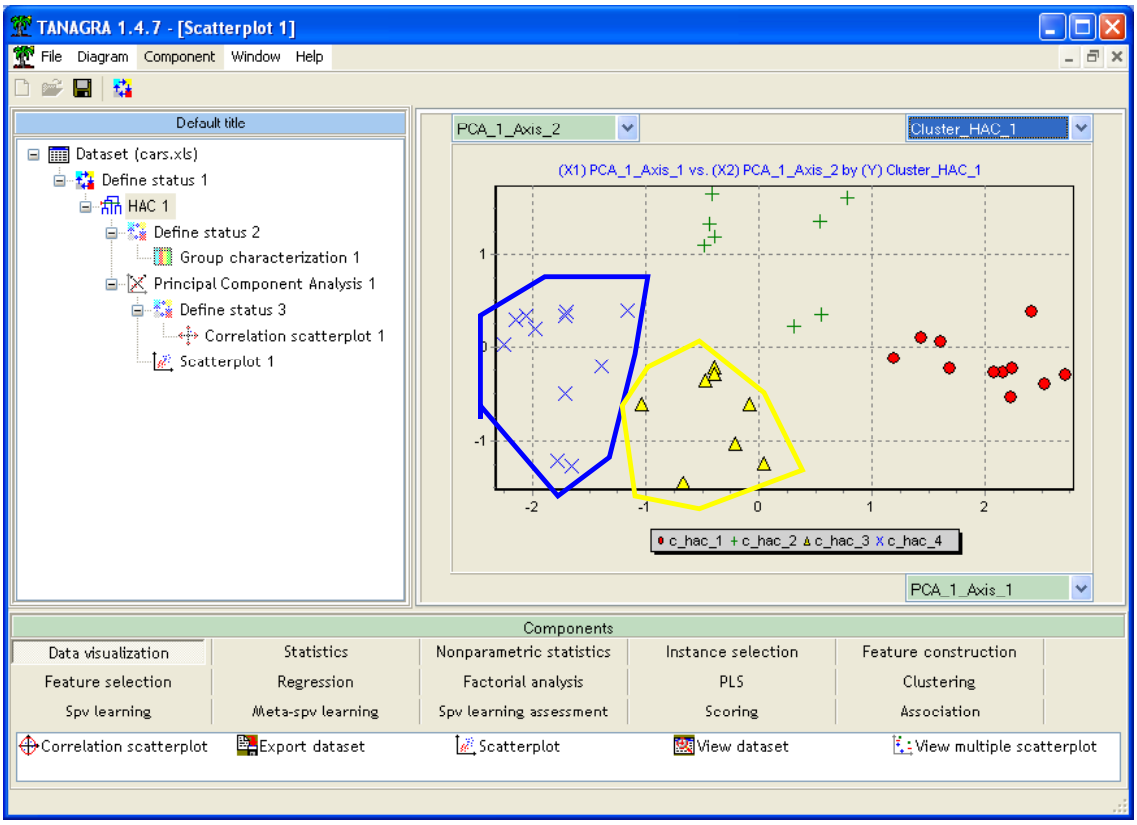


Modifying the number of clusters

We note that the "small cars" cluster (C_HAC_3), which contains 19 vehicles, is rather wide. One can ask if a subdivision of this cluster is more judicious. We modify the parameter of the HAC and set to 4 the number of clusters.



We click on the VIEW menu of the SCATTERPLOT node.



The clustering into 4 groups, which is not obvious in the dendrogram, seems to be more relevant in the scatter plot. The next step is the interpretation of these new clusters.

We see again the GROUP CHARACTERIZATION node. The first two clusters are not modified. The third group was subdivided into two sub-groups with 8 and 11 examples.

Cluster_HAC_1=c_hac_3				Cluster_HAC_1=c_hac_4			
Examples		[21.1 %] 8		Examples		[28.9 %] 11	
Att - Desc	Test value	Group	Overral	Att - Desc	Test value	Group	Overral
Continuous attributes				Continuous attributes			
MPG	1.8	28.55	24.76	MPG	4.4	32.23	24.76
Horsepower	-1.1	92.63	101.74	Drive_Ratio	3.2	3.52	3.09
Displacement	-1.2	142.63	177.29	Cylinders	-3.4	4	5.39
Drive_Ratio	-1.3	2.88	3.09	Displacement	-3.5	96.55	177.29
Weight	-1.4	2.56	2.86	Horsepower	-4.1	73.55	101.74
Cylinders	-1.8	4.5	5.39	Weight	-4.2	2.09	2.86
Discrete attributes				Discrete attributes			
Country=U.S.	1.1	[27.3 %] 75.0 %	57.90%	Country=Japan	1.8	[57.1 %] 36.4 %	18.40%
Country=Japan	0.5	[28.6 %] 25.0 %	18.40%	Country=Germany	1.6	[60.0 %] 27.3 %	13.20%
Country=Italy	-0.5	[0.0 %] 0.0 %	2.60%	Country=Italy	1.6	[100.0 %] 9.1 %	2.60%
Country=France	-0.5	[0.0 %] 0.0 %	2.60%	Country=France	-0.6	[0.0 %] 0.0 %	2.60%
Country=Sweden	-0.7	[0.0 %] 0.0 %	5.30%	Country=Sweden	-0.9	[0.0 %] 0.0 %	5.30%
Country=Germany	-1.2	[0.0 %] 0.0 %	13.20%	Country=U.S.	-2.4	[13.6 %] 27.3 %	57.90%

The new C_HAC_3 cluster corresponds to the "middle" cars. It should be opposed to the second group (C_HAC_2) to better understanding it. They represent together "middle" cars. But in the first case (C_HAC_2), it corresponds to European "middle" cars; in the second case, it corresponds to US "middle" cars. Differentiation relies primarily on the DRIVE_RATIO attribute.

The fourth cluster is the real the "small" cars with low consumption, low horsepower and low weight.

Finally, there are 4 clusters in this dataset: the "large cars"; the "European middle cars" ; "the US middle cars"; and the "small cars".

This example illustrates well that there is not a monolithic approach of the data mining. Several points of view should be adopted. By skillfully combining the visualization and the knowledge discovery technique, we can reach fine information and to thus better exploration the data.