# 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 <u>http://lib.stat.cmu.edu/DASL/Stories/ClusteringCars.html</u>. 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<sup>1</sup> 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<sup>2</sup>: 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.

<sup>&</sup>lt;sup>1</sup> http://en.wikipedia.org/wiki/Data\_clustering

<sup>&</sup>lt;sup>2</sup> <u>http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/enHAC\_IRIS.pdf</u>



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<sup>3</sup>.



TANAGRA detects this solution.

<sup>&</sup>lt;sup>3</sup> 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.



Default title 🧹	Rep	port		Dendrogram	
🖃 🏢 Dataset (cars.xls)				HAC 1	^
E 🔁 Define status 1				Parameters	
				Results	
	Cluster	ing resu	ılts		
	Clusters	3			
	Cluster	Description	Size		
	cluster nº 1	c_hac_1	11		
	cluster n°2	c_hac_2	8		
	cluster n°3	c_hac_3	19	-	
	Cluster	rs vs. Ta	rge	t Attributes	~
	<				2

#### 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.

Price Diagram Component Window Help       Define attribute statuses         Default title       Define attribute statuses         Parameters       Atributes :       Target Input Illustrative         Country       MPG       Weight         Define status       Country       Country         Car       Country       MPG         Weight       Drive_Ratio       Drive_Ratio         Drive_Ratio       Displacement       Oisplacement         Cluster_HAC_1       Clear selection         Data visualization       Feature construction       Feature seconom       Negrossion         PLS       Clustering       Spv learning       Meta-spv learning         Spv learning assessment       Scoring       Association       Meta-spv learning         Charter second       Meta-spv learning       Kohonen-SOM	TANAGRA 1.4.7 - [HAC 1]		
Default Nile       Define attribute statuses         Image: Define status 1       Parameters         Image: Define status 1       Attributes :         Image: Define status 2       Attributes :         Image: Define status 1       Attributes :         Image: Define status 2       Attributes :	File Diagram Component Window He	lp.	- E ×
Define attribute statuses         Image: Define status 1         Image: Define sta	D 📽 🖪 🙀		
Image: Section of the status       Parameters         Image: Section of the status       Attributes :       Target Input Illustrative         Image: Section of the status       Imput Illustrative       Imput Illustrative         Image: Section of the status       Imput Illustrative       Imput Illustrative         Image: Section of the status       Imput Illustrative       Imput Illustrative         Image: Section of the status       Imput Illustrative       Imput Illustrative         Image: Section of the status       Imput Illustrative       Imput Illustrative         Image: Section of the status       Imput Illustrative       Imput Illustrative         Image: Section of the status       Imput Illustrative       Imput Illustrative         Image: Section of the status       Imput Illustrative       Imput Illustrative         Image: Section of the status       Imput Illustrative       Imput Illustrative         Image: Section of the status       Imput Illustrative       Imput Illustrative         Image: Section of the status       Image: Section of the status       Image: Section of the status         Image: Section of the status       Image: Section of the status       Image: Section of the status       Image: Section of the status         Image: Section of the status       Image: Section of the status       Image: Section of the status       Imagee	Default title	fine attribute statuses	
Data visualization       OK       Clear selection         Data visualization       OK       Cancel       Help         Feature construction       Feature selection       recorner analysis         PLS       Clustering       Spv learning       Meta-spv learning         Spv learning assessment       Scoring       Association         Image: CT       Image: CTP       Image: CTP       Image: CTP	<ul> <li>□ Dataset (cars.xls)</li> <li>□ 1</li> <li>□ 1</li> <li>□ 1</li> <li>□ 20</li> <li>□ 1</li> <li>□ 20</li> <li>□ 20</li></ul>	Parameters Attributes : Car Car C MPG C Weight C Drive_Ratio C Horsepower C Displacement C Qvlinders C Cluster_HAC_1	Target Input Illustrative Country MPG Weight Drive_Ratio Horsepower Displacement Cylinders
Data visualization     OK     Cancel     Help       Feature construction     Feature second in regression     records analysis       PLS     Clustering     Spv learning     Meta-spv learning       Spv learning assessment     Scoring     Association       Image: CT     Image: CTP     Image: K-Means     Image: Kohonen-SOM	<	見 見   ※	Clear selection
Data visualization     OK     Cancel     Help       Feature construction     Feature selection     Regression     Facconic analysis       PLS     Clustering     Spv learning     Meta-spv learning       Spv learning assessment     Scoring     Association       Image: CT     Image: CTP     Image: CTP     Image: CTP			
Feature construction     Feature section     reconstruction       PLS     Clustering     Spv learning       Spv learning assessment     Scoring     Association       Association     Association	Data visualization		OK Cancel Help
PLS     Clustering     Spv learning     Meta-spv learning       Spv learning assessment     Scoring     Association       Ar, CT     Ar, CTP     Arr HAC     K-Means	Feature construction Feature	selection Regression	ractoriat anatysis
Spv learning assessment Scoring Association 축, CT 축, CTP 채ስHAC 옵.K-Means I Kohonen-SOM	PLS Clust	ering Spv learning	Meta-spv learning
Kan CT An CTP An HAC Kanas Kohonen-SOM Kohonen-SOM	Spv learning assessment Sco	ring Association	
	🕭 ст 🖧 стр	ланас 🤷 к	(-Means 🙀 Kohonen-SOM

Then, we add the GROUP CHARACTERIZATION component.

2 TANAGRA 1.4.7 - [Gro	up characterization 1]							
Tile Diagram Componen	t Window Help							- 8 ×
D 📽 🔛 👫								
Default	title		Gro	un cha	racteriza	tion 1		^
🖃 🥅 Dataset (cars.xls)				Para	ameters			
😑 🚰 Define status 1								
😑 🚮 HAC 1				R	ecults			
😑 🛟 Define s	tatus 2				o su ti s			- 1
🔲 🚺 Grou	up characterization 1					Description	of "Clus	ter_H
	<b>A</b>	Cluster_H	\C_1-c	:_hac_1	1	Cluster_I	HAC_1=	c_hac
	1	Examples		[28	1.9 %] 11	Examples		
	1	Att - Desc	Test value	Group	Overral	Att - Desc	Test value	Grou
	1	Continuous attribu	ites			Continuous attril	outes	
	1	Displacement	5.4	301.64	177.29	Drive Balliu	2.7	3.5
	1	Weight	5.1	3.79	2.86	Horsepower	0.9	109.4
		Cylinders	5.0	7.45	5.39	Weight	0.4	2.6
1		Horsepower	4.3	130.82	101.74	weight .	0.4	~
		:						>
		Components						
Data visualization	Statistics	Nonparametric s	tatisti	cs	Instar	ice selection		
Feature construction	Feature selection	Regressio	n		Factorial analysis			
PLS	Clustering	Spv learni	ng		Meta	spv learning		
Spv learning assessment	Scoring	Associatio	n				_	
🕁 Bartlott's tost	🔆 Brown - Forsythc's t	ost 💢 Fishor's	test		<.	Group charact	erizatio	
<								>

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.

Clu	ister_HAC_1	Cluster_HAC_1=c_hac_1							
Examples		[ 28	8.9 %] 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 attribute	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	0.0 %] 19					
Att - Desc	Test value	Group	Overral					
Continuous attribute	S							
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.

Defau	ılt title			Des	incipal Component Apple		
			Parameters Generated axis : 10 Results Figen values				
		Axis I	Eigen value %	explained	Histogram	n % cumulate	ted
		Axis I 1	E <b>igen value</b> % 2.358113	explained 78.60%	Histogram	n % cumulate 78.60	<b>ted</b> 60%
		Axis I 1 2	E <b>igen value %</b> 2.358113 0.602792	explained 78.60%	Histogram	m % cumulate 78.60 98.71	<b>ted</b> 60% 70%
		Axis 1 2 3	Eigen value % 2.358113 0.602792 0.039095	explained 78.60% 20.09% 1.30%	Histogram	n % cumulat 78.60 98.70 100.00	<b>ted</b> 60% 70% 00%
		Axis   1 2 3	Eigen value % 2.358113 0.602792 0.039095 Compo	explained 78.60% 20.09% 1.30%	Histogran	m % cumulat 78.60 98.70 100.00	ted 60% 70% 00%
Data visualization	Statistics	Axis I 1 2 3 Nonp:	Eigen value % 2.358113 0.602792 0.039095 Compo arametric stat	explained 78.60% 20.09% 1.30% intents tistics	Histogram	n % cumulate 78.60 98.70 100.00 Feature construction	ted 60% 70% 00%
Data visualization Feature selection	Statistics Regression	Axis I 1 2 3 Nonpa	Eigen value % 2.358113 0.602792 0.039095 Compo arametric stat actorial analy:	explained 78.60% 20.09% 1.30% intents tistics	Histogram	m % cumulat 78.60 98.70 100.00 Feature construction Clustering	ted 60% 70% 00%

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.

🛣 TANAGRA 1.4.7 - [Sca	tterplot 1]	r			
Tile Diagram Component	: Window Help	HAC parameters			- 8 3
Defax Defaxet (carx.vk) Define status 1 Define status	A title araneters xocuto ion 1 lew alysis 1 orrelation scatterplot terplot 1	Parameters Best clusters Detect Define Distance non None Variance Show detaile Tree strut Cluster set	4 malization d results cture election		•
			OK Cancel	Help	*
Data visualization	Statistics	Nonparametric statistics	Instance selection	Feature constr	uction
Feature selection	Regression	Factorial analysis	PLS	Clustering	8
Spv learning	Meta-spv learning	Spv learning assessment	Scoring	Associatio	n
Correlation scatterplot	Export dataset	🛃 Scatterplot	🔛 View dataset	其 View mu	ultiple scatt
<				)	>



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.

C	luster_HAC	_1=c_hac_3		Cluster_HAC_1=c_hac_4			
Examples		[2	21.1 %] 8	Examples		[ 28	3.9 %] 11
Att - Desc	Test value	Group	Overral	Att - Desc	Test value	Group	Overral
Continuous attribute	S			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.