Objectif

Montrer la complémentarité des méthodes de fouille de données (clustering) et de visualisation (analyse en composantes principales).

Fichier

Nous traitons le fichier CARS.XLS. Il est composé de 38 véhicules décrits par une série de variables. La source des données est un magazine américain, il n'est pas étonnant que la majorité des véhicules le soient (22 sur 38).

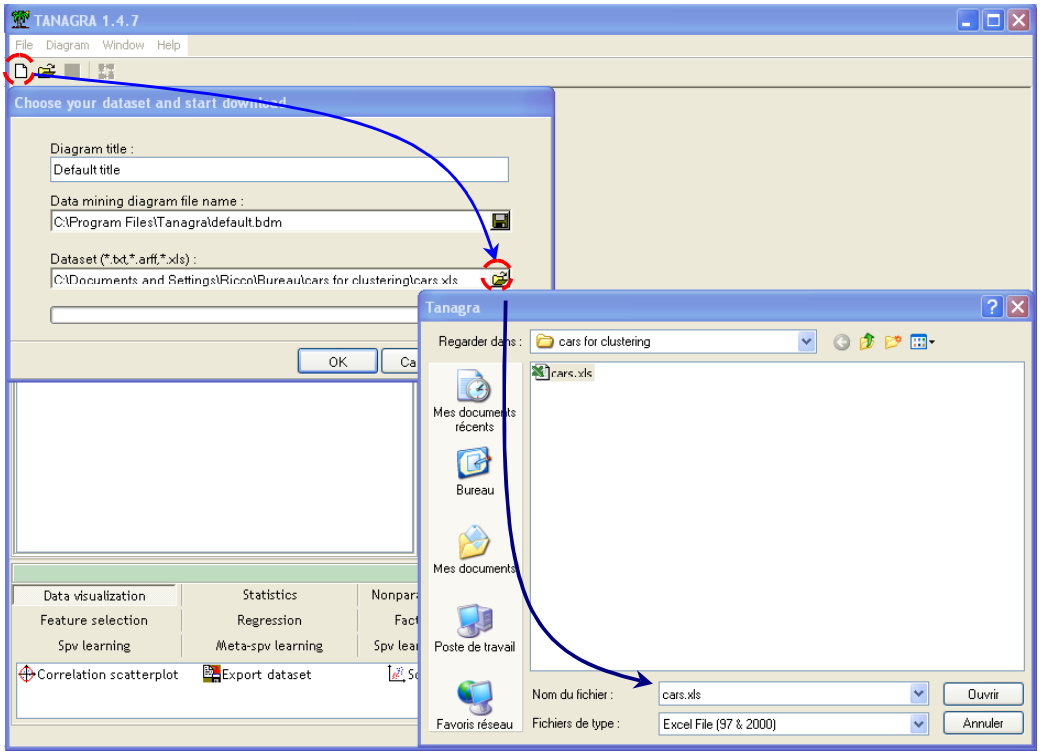
L'objectif est de produire des groupes homogènes de véhicules et de les caractériser.

Nous disposons d'une référence pour ce fichier. Les résultats sont en ligne http://lib.stat.cmu.edu/DASL/Stories/ClusteringCars.html. Les auteurs indiquent un partitionnement en 3 classes. Nous verrons si nous obtenons la même chose.

CAH avec TANAGRA

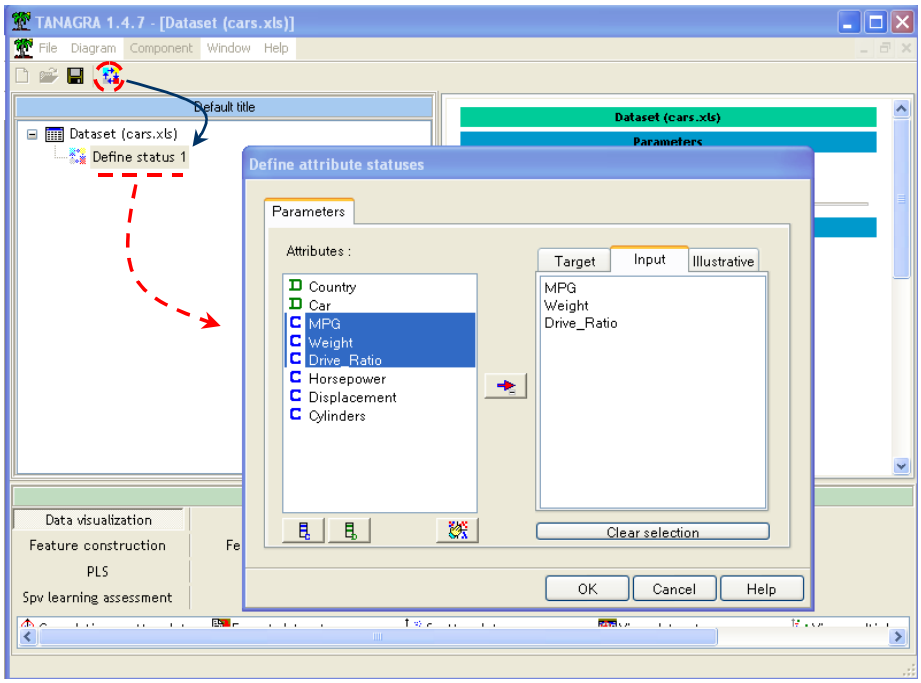
Charger les données

Nous devons dans un premier temps créer un diagramme et charger les données. Pour ce faire, nous cliquons sur le menu FILE/NEW. Nous sélectionnons le fichier CARS.XLS, au format EXCEL.



Définir le problème

L'étape suivante consiste à désigner, à l'aide du composant DEFINE STATUS, les variables actives (INPUT) : il s'agit des variables MPG, WEIGHT et DRIVE_RATIO.



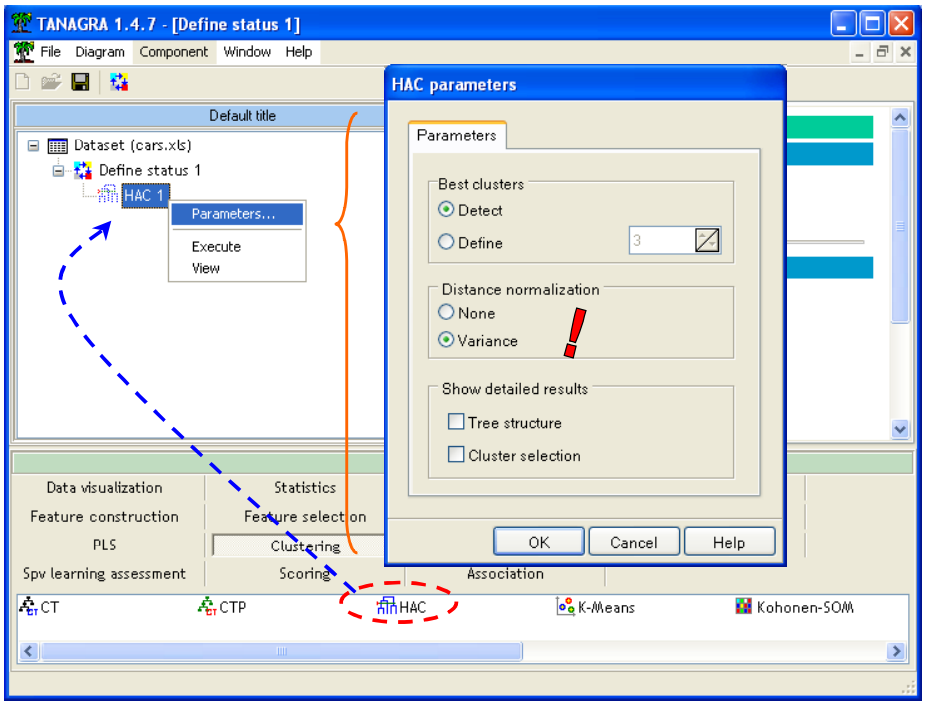
CAH

Nous choisissons la CAH pour la construction de la typologie. Elle affiche les différents niveaux d'agrégation et donne des indications sur le nombre de classes à retenir.

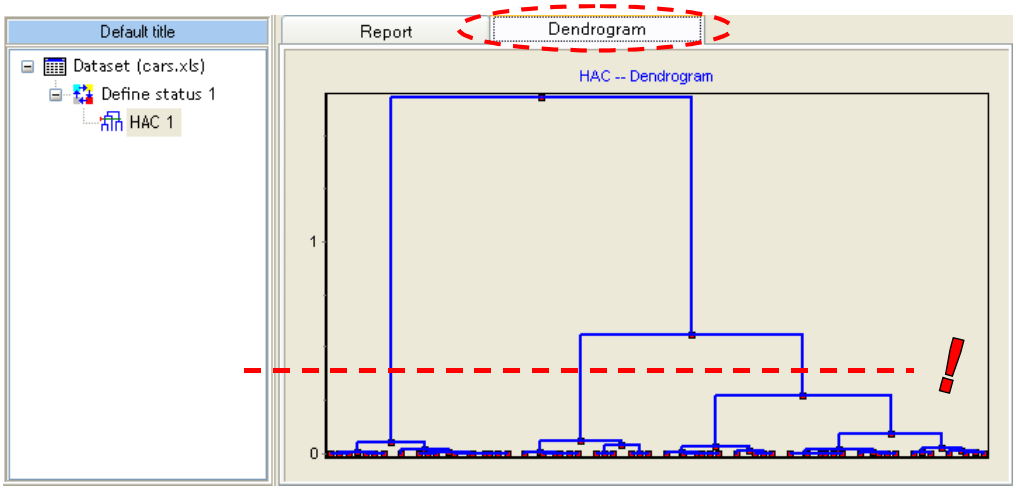
TANAGRA intègre une option, il peut détecter automatiquement le saut le plus élevé dans le dendrogramme. Par défaut, il propose la classification correspondante.

Enfin, les effectifs étant faibles pour cet exemple, nous pouvons prendre comme racine de la hiérarchie les observations du fichier de données. Si l'effectif est élevé, plusieurs milliers d'observations, il est plus judicieux de s'appuyer sur le principe de la classification mixte en définissant un premier partitionnement assez fin (une cinquantaine de classes) avec une méthode pouvant gérer de gros volumes de données (un K-MEANS ou un SOM par exemple), puis de les utiliser comme racine du dendrogramme. Le temps de calcul est nettement amélioré tout en préservant la qualité des résultats.

Voici le diagramme de traitement correspondant, nous paramétrons la méthode de manière à normaliser les données. En effet, étant exprimées dans des unités différentes, il est préférable de ramener les variables dans le même référentiel.

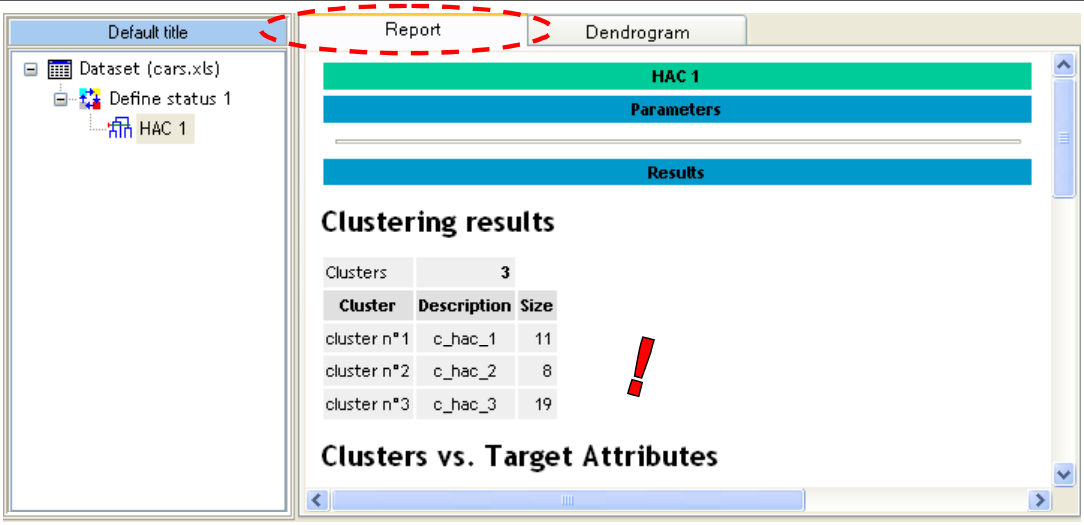


Après exécution (menu VIEW), les résultats s'affichent dans la fenêtre de droite. Avec la version 1.4.8, le dendrogramme est maintenant dessiné. La partition en trois classes semble être effectivement la plus évidente[1].



C'est celle qui a été détectée par TANAGRA.

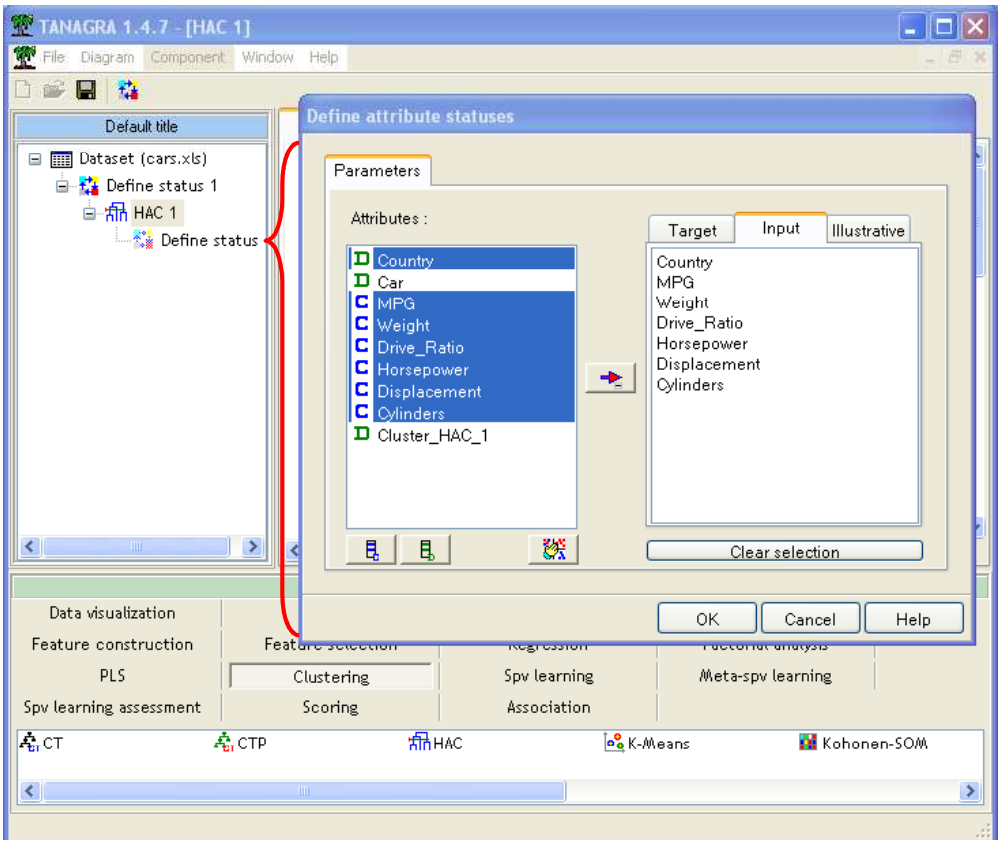
[1] Dans la plupart des cas, la subdivision en deux classes produit le décalage le plus élevé. Il ne faut pas s'y fier. Il s'agit souvent d'un artefact dû au fait que nous introduisons dans ce cas la première subdivision des données. Pour cette raison, TANAGRA essaie de détecter le saut le plus important uniquement pour les subdivisions en 3, 4, etc. classes.



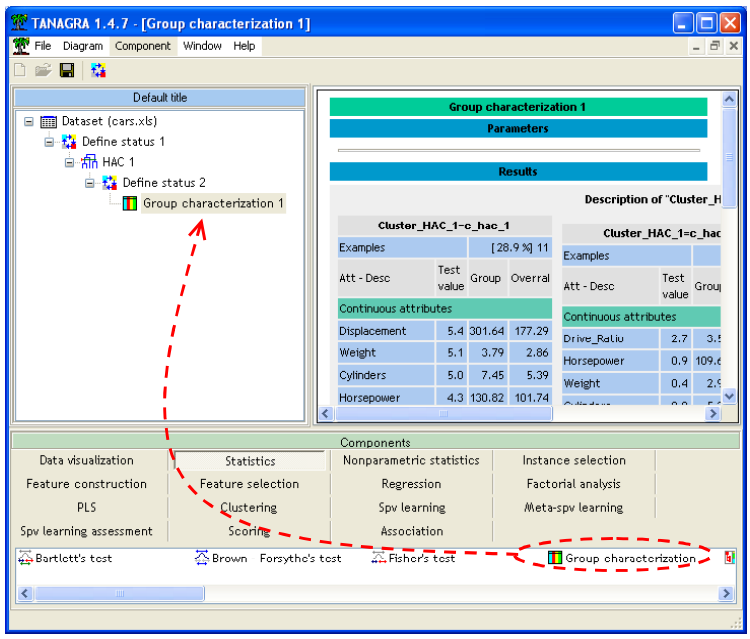
Caractérisation des groupes

L'étape suivante consiste à caractériser ces groupes. L'outil GROUP CHARACTERIZATION est le plus approprié pour cela, il permet de comparer les indicateurs (moyenne ou proportion) marginaux et conditionnellement aux groupes.

Pour ce faire, nous insérons dans un premier temps un nouveau composant DEFINE STATUS. Nous plaçons en TARGET, la classification produite par la CAH ; en INPUT, les autres variables à l'exception de la variable CAR qui indique le modèle du véhicule.



Dans un second temps, nous ajoutons le composant GROUP CHARACTERIZATION.



Pour avoir une vue complète des résultats, vous avez la possibilité de les copier dans un tableur. Il est alors possible d'accéder à des options de présentation plus performantes.

Le premier groupe (C_HAC_1) est composé de véhicules volumineux, lourds, puissants et consommant beaucoup[2]. Ce groupe est composé à 100% de véhicules US, 50% des véhicules US se trouvent dans ce groupe.

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%

[2] MPG est une norme anglo-saxonne qui indique le nombre de miles que l'on peut parcourir avec un gallon de carburant. Plus le chiffre est faible, plus le véhicule consomme.

Le second groupe (C_HAC_2) est essentiellement composé de véhicules européens ou japonais. Il intègre la totalité des véhicules suédois et français. Il s'agit de voitures de taille et de puissance moyenne. Avec une petite particularité, le DRIVE_RATIO (rapport de boîte) est plus élevé, ce qui est caractéristique de l'Europe où l'on apprécie plus les voitures nerveuses.

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%

Enfin, le troisième groupe (C_HAC_3) est composé de voitures américaines pour moitié ; japonaises, italiennes et allemandes pour le reste. Il s'agit de véhicules de petite taille, légères, peu puissantes et consommant peu. Nous pouvons les voir comme le groupe des petites voitures.

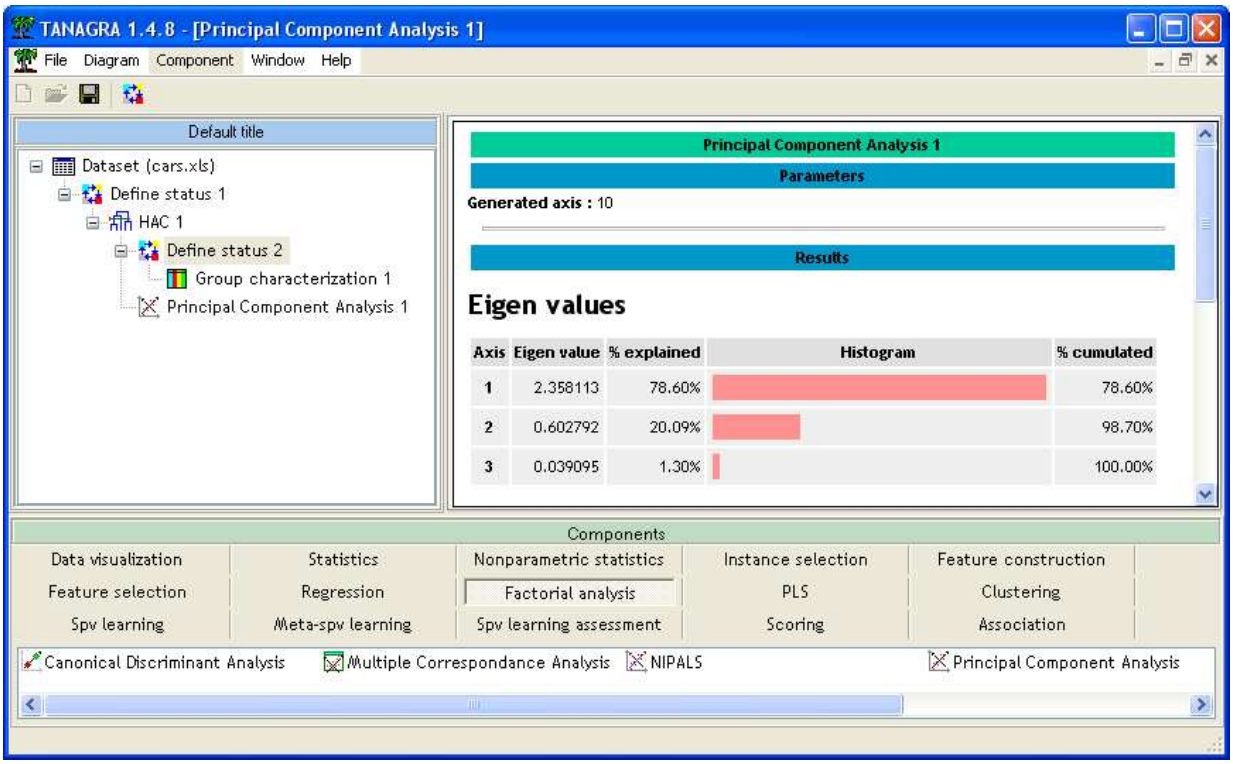
Cluster_HAC_1=c_hac_3			
Examples		[50.0 %] 19	
Att - Desc	Test value	Group	Overral
Continuous attributes			
MPG	5.5	30.68	24.76
Drive_Ratio	1.8	3.25	3.09
Displacement	-4.2	115.95	177.29
Cylinders	-4.5	4.21	5.39
Horsepower	-4.6	81.58	101.74
Weight	-4.9	2.29	2.86
Discrete attributes			
Country=Japan	2.1	[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%

Au final, nous obtenons bien trois classes de véhicules : les « grosses » voitures, essentiellement américaines ; les voitures « moyennes » qui sont plutôt européennes ; et les « petites » voitures qui n'ont pas de nationalité attitrée.

ACP avec TANAGRA

ACP

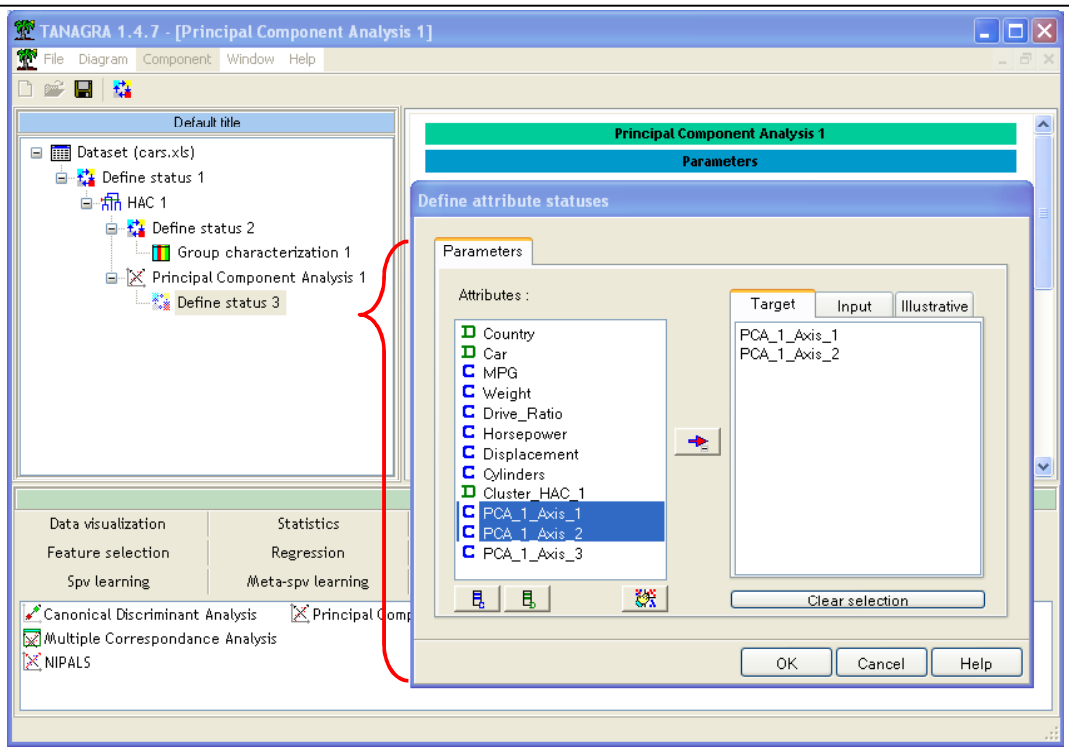
Pour visualiser les groupes et mieux les situer les uns par rapport aux autres, nous allons les projeter dans le premier plan factoriel. Nous ajoutons donc un composant PCA (Principal Component Analysis) dans notre diagramme.



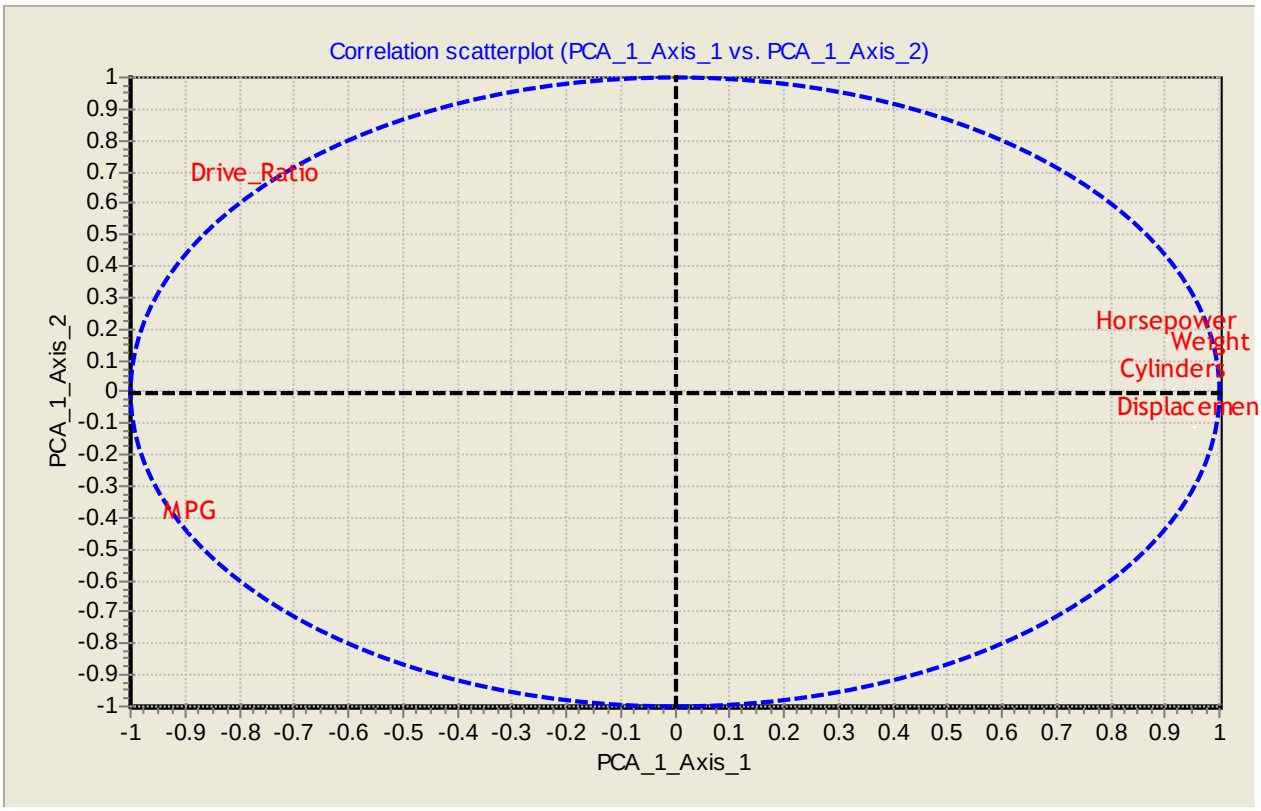
Les deux premiers axes résument 98% de l'information. Cela laisse à penser que nous obtiendrons une représentation satisfaisante des proximités entre les observations.

Cercle de corrélation

Pour obtenir le cercle de corrélation, il faut rajouter un composant DEFINE STATUS et placer en TARGET les 2 axes factoriels, puis en INPUT les variables de l'analyse. Ce dispositif permet d'intégrer dans la même représentation des variables actives (MPG, WEIGHT, DRIVE_RATIO) et les variables illustratives (HORSEPOWER, DISPLACEMENT, CYLINDERS).



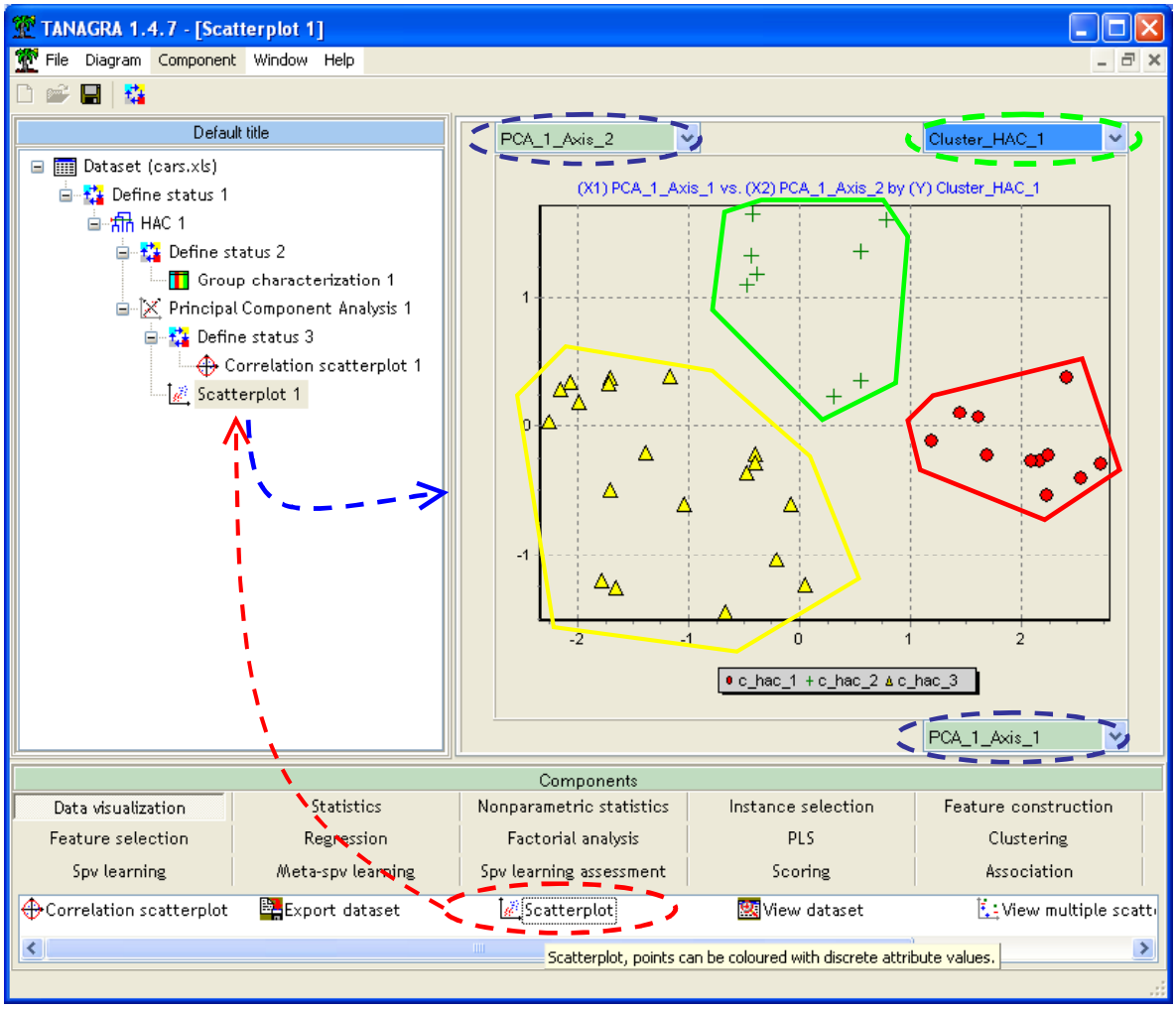
Il reste alors à placer le composant CORRELATION SCATTERPLOT pour obtenir le cercle de corrélation. Cet outil est souvent utilisé pour interpréter les axes factoriels.



Projection des individus

Plus intéressant dans notre cas, nous voulons projeter nos individus sur le premier plan factoriel et surtout les étiqueter selon leur groupe d'appartenance définie par la classification.

Nous plaçons le composant SCATTERPLOT dans notre diagramme et nous le paramétrons de manière adéquate.

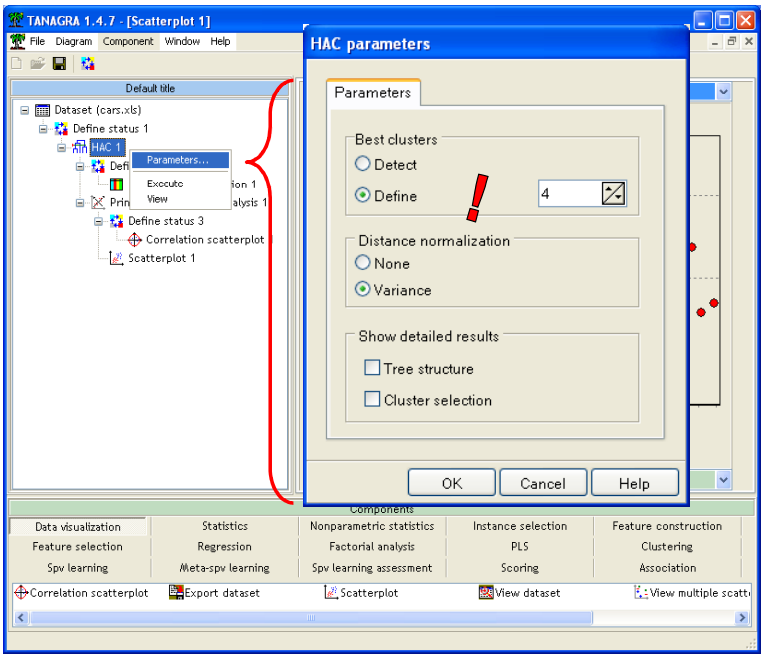


Les trois groupes se démarquent assez nettement.

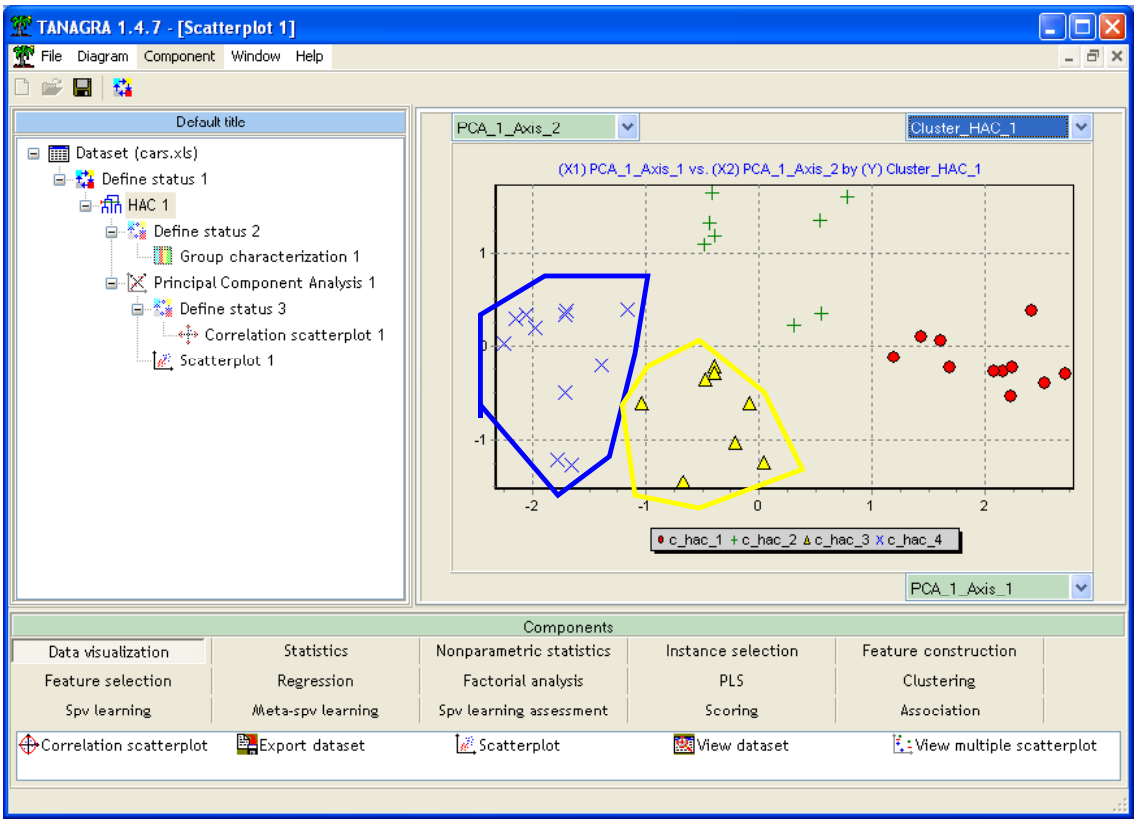
Approfondissement des résultats

Nous remarquons cependant un certain étalement du groupe des « petites » voitures composé de 19 observations. Nous pouvons nous demander si la subdivision de ce groupe n'est pas plus judicieuse.

Nous tentons l'expérience en paramétrant différemment le composant HAC. Nous spécifions explicitement cette fois-ci le nombre de classes à obtenir, nous le fixons à 4.



En exécutant de nouveau le diagramme (menu VIEW du SCATTERPLOT), nous retrouvons notre nuage de points. Nous constatons que le nouveau groupe qui s'est formé provient effectivement de la subdivision du groupe des petites voitures.



La subdivision en 4 groupes, moins évidente dans le dendrogramme, semble plus affirmée ici. Reste à voir ce qui distingue ces deux sous-classes des « petites » voitures.

Nous revenons sur le composant GROUP CHARACTERIZATION. Les deux premiers groupes ne sont pas modifiés. Le troisième groupe de 19 observations laisse maintenant la place à 2 groupes de 8 et 11 observations.

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%

Le nouveau troisième groupe est représentatif du véhicule moyen. Il tient une place quasi-centrale dans le plan factoriel. Il faut surtout l'opposer au second groupe pour le comprendre. Le groupe 2 (C_HAC_2) et le nouveau groupe 3 (C_HAC_3) sont tous les deux des représentants de véhicules moyens. Dans le premier cas, il s'agit de voitures moyennes européennes ; dans le second cas, il s'agit essentiellement de voitures US.

Le quatrième groupe en revanche se démarque par une consommation très basse, les véhicules sont très légers et très peu puissants. C'est en réalité le vrai groupe des « petites » voitures.

Au final, nous pouvons considérer qu'il y a 4 groupes de véhicules dans ce fichier : les « grosses » voitures ; les voitures « moyennes » ; les voitures moyennes à l'Européenne, qui se démarquent essentiellement ici par un rapport de boîte plus élevé ; et enfin les « petites » voitures.

Cet exemple illustre bien combien il n'y a pas d'approche monolithique du traitement de données. Il faut adopter plusieurs points de vue. En combinant adroitement les techniques de visualisation et les techniques de fouille, nous pouvons accéder à des informations fines et ainsi mieux explorer les données.