Caractérisation des classes en classification automatique

Variables actives et illustratives quantitatives et qualitatives

Ricco RAKOTOMALALA
Université Lumière Lyon 2

PLAN

1. Position du problème
2. Caractérisation univariée
   a. De la partition
   b. Des groupes
3. Caractérisation multivariée
   a. Pourcentage d'inertie expliquée
   b. Distance entre centres de classes
   c. Couplage avec l'analyse factorielle
   d. Utilisation d'une technique supervisée (ex. analyse discriminante)
4. Conclusion
5. Bibliographie

La classification automatique

Constitution des groupes à partir des caractéristiques de proximité

Classification automatique

Typologie, apprentissage non-supervisé, clustering

Variables « actives », servent à la constitution des groupes. Souvent (mais pas toujours) toutes quantitatives.

Variables « illustratives », ne participent pas à la constitution des groupes, mais permettent d'appuyer l'interprétation.

Modele	puissance	cylindree	vitesse	longueur	largeur	hauteur	poids	CO2	prix	origine	carburant
PANDA	54	1108	150	354	159	154	860	135	8070	Europe	Essence
TWINGO	60	1149	151	344	163	143	840	143	8950	France	Essence
CITRONC2	61	1124	158	367	166	147	932	141	10700	France	Essence
YARIS	65	998	155	364	166	150	880	134	10450	Autres	Essence
FIESTA	68	1399	164	392	168	144	1138	117	14150	Europe	Diesel
CORSA	70	1248	165	384	165	144	1035	127	13590	Europe	Diesel
GOLF	75	1968	163	421	176	149	1217	143	19140	Europe	Diesel
P1007	75	1360	165	374	169	161	1181	153	13600	France	Essence
MUSA	100	1910	179	399	170	169	1275	146	17900	Europe	Diesel
CLIO	100	1461	185	382	164	142	980	113	17600	France	Diesel
AUDIA3	102	1595	185	421	177	143	1205	168	21630	Europe	Essence
MODUS	113	1598	188	380	170	159	1170	163	16950	France	Essence
AVENSIS	115	1995	195	463	176	148	1400	155	26400	Autres	Diesel
P407	136	1997	212	468	182	145	1415	194	23400	France	Essence
CITRONC4	138	1997	207	426	178	146	1381	142	23400	France	Diesel
MERC_A	140	1991	201	384	177	160	1340	141	24550	Europe	Diesel
MONDEO	145	1999	215	474	194	143	1378	189	23100	Europe	Essence
VECTRA	150	1910	217	460	180	146	1428	159	26550	Europe	Diesel
PASSAT	150	1781	221	471	175	147	1360	197	27740	Europe	Essence
VELSATIS	150	2188	200	486	186	158	1735	188	38250	France	Diesel
LAGUNA	165	1998	218	458	178	143	1320	196	25350	France	Essence
MEGANECC	165	1998	225	436	178	141	1415	191	27800	France	Essence
P307CC	180	1997	225	435	176	143	1490	210	28850	France	Essence
P607	204	2721	230	491	184	145	1723	223	40550	France	Diesel
MERC_E	204	3222	243	482	183	146	1735	183	46450	Europe	Diesel
CITRONC5	210	2496	230	475	178	148	1589	238	33000	France	Essence
PTCRUISER	223	2429	200	429	171	154	1595	235	27400	Autres	Essence
MAZDARX8	231	1308	235	443	177	134	1390	284	34000	Autres	Essence
BMW530	231	2979	250	485	185	147	1495	231	46400	Europe	Essence
ALFA 156	250	3179	250	443	175	141	1410	287	40800	Europe	Essence

Objectif de l'étude : Identifier les catégories (groupes) de voitures « similaires » (c.-à-d. qui se ressemblent au regard de leurs propriétés)

Objectif : identifier des groupes d'observations ayant des caractéristiques similaires (ex. comportement d'achats de clients, caractère « polluant » de véhicules, etc.)

On veut que :

(1) Les individus dans un même groupe se ressemblent le plus possible

(2) Les individus dans des groupes différents se démarquent le plus possible

Pourquoi ?

→ Identifier des structures sous-jacentes dans les données

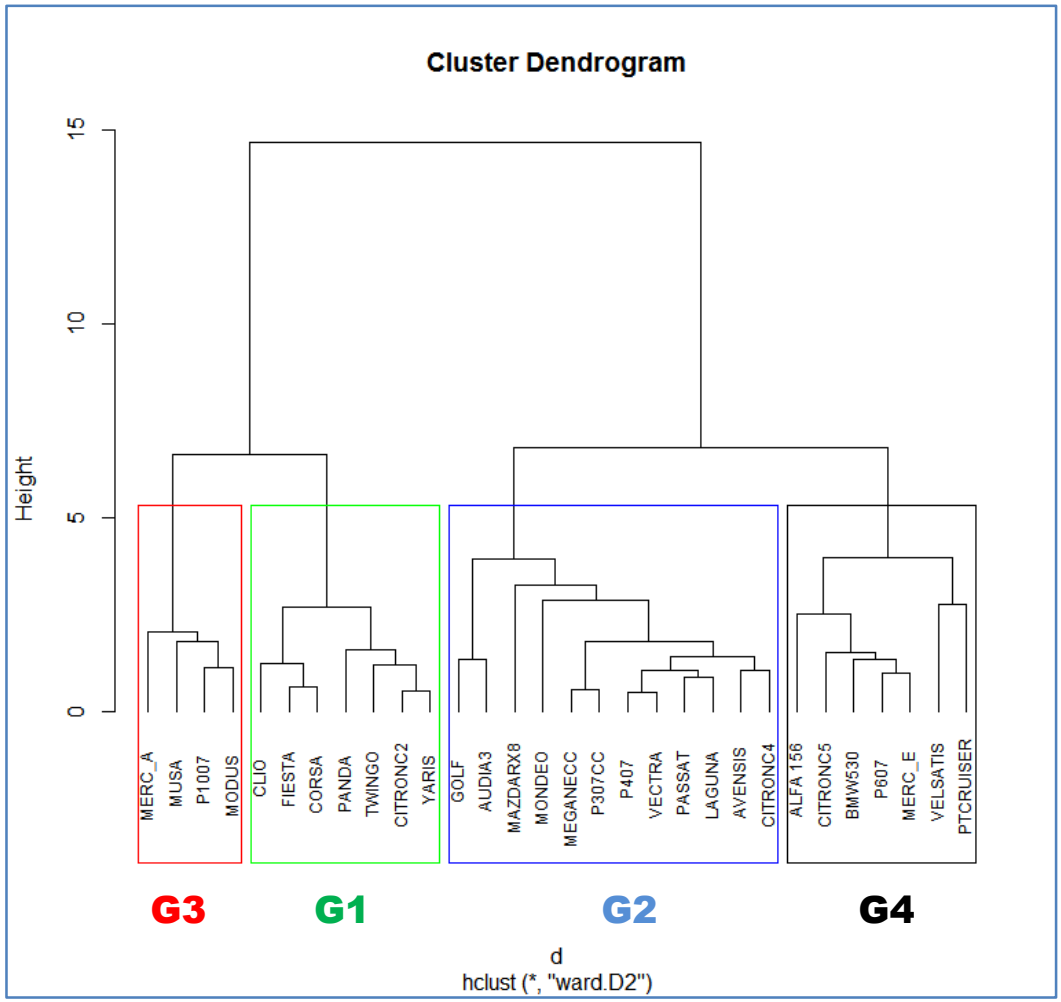
→ Résumer des comportements

→ Affecter de nouveaux individus à des catégories

→ Identifier les cas totalement atypiques

Classification automatique

Interprétation des groupes



Sur quelles informations repose l'interprétation des résultats ?

Dans quelle mesure ces groupes sont-ils éloignés les uns des autres ?

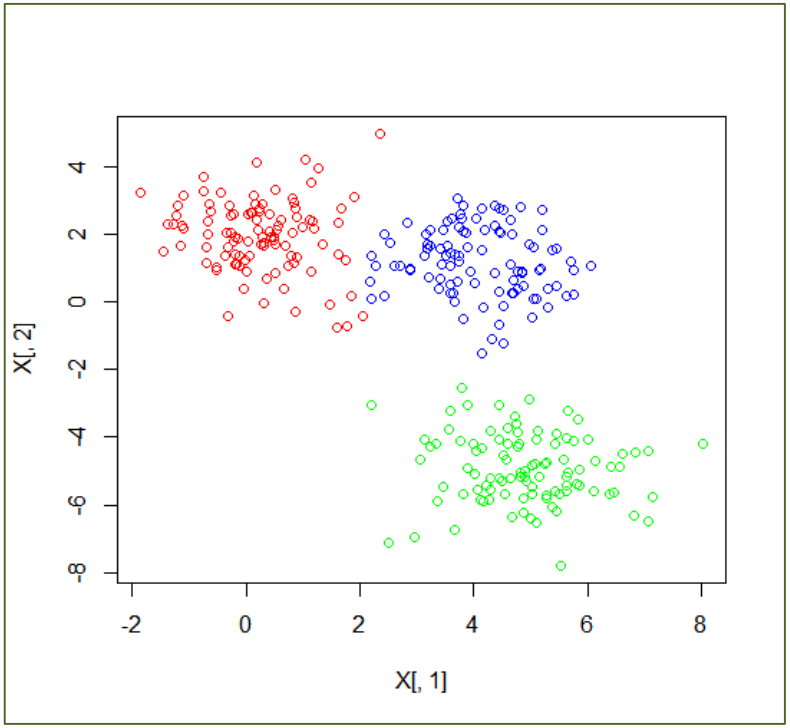
Quelles sont les caractéristiques qui rapprochent les individus du même groupe, et qui différencient les individus appartenant à des groupes distincts ?

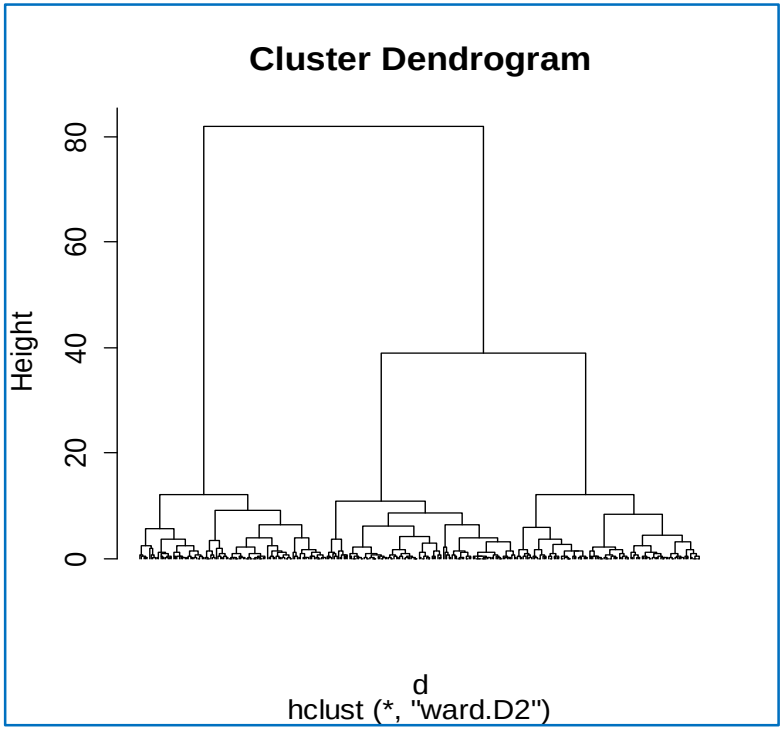
Au regard des variables actives qui ont servi à constituer les groupes.

Mais aussi au regard des variables illustratives qui amènent un autre point de vue sur la constitution des classes.

Classification automatique

Autre exemple dans le plan





Cet exemple permettra de comprendre la nature des calculs réalisés pour caractériser la partition et les groupes.

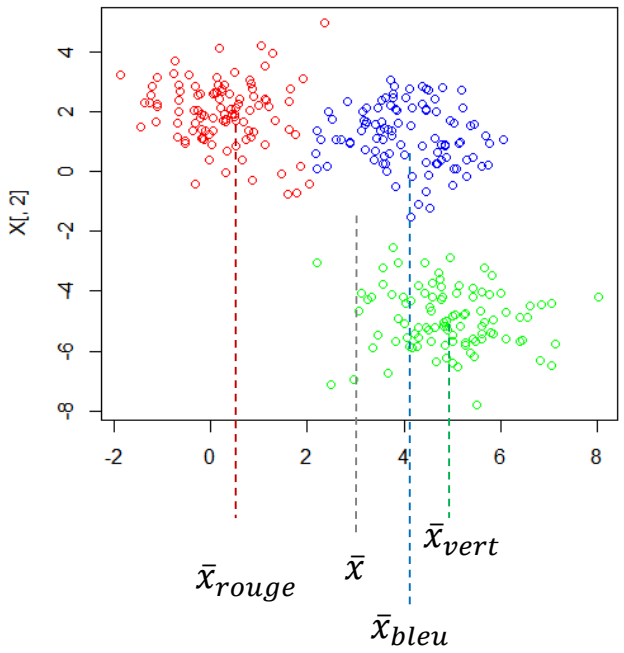
Caractérisation univariée

Interprétation à l'aide des variables prises individuellement

Caractérisation de la partition

Variables quantitatives

Evaluer dans quelle mesure la variable – prise individuellement – « contribue » à la constitution de la partition.



L'idée est de mesurer la dispersion de la variable attribuable à l'appartenance aux groupes

Equation d'analyse de variance

$$\text{Variabilité totale} = \text{Variabilité inter - classes} + \text{Variabilité intra - classe}$$

$$SCT = SCE + SCR$$

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{g=1}^{G} n_g(\bar{x}_g - \bar{x})^2 + \sum_{g=1}^{G}\sum_{i=1}^{n_g}(x_i - \bar{x}_g)^2$$

Le rapport de corrélation η est défini par :

$$\eta^2 = \frac{SCE}{SCT}$$

η^2 indique la proportion de variance de X expliquée par les groupes ($0 \leq \eta^2 \leq 1$). On peut l'interpréter (avec beaucoup de prudence) comme le pouvoir discriminant de la variable.

Caractérisation de la partition

Variables quantitatives – Illustration sur le fichier des « autos »

Moyennes conditionnelles

	G 1	G 3	G 2	G 4	% epl.
poids	952.14	1241.50	1366.58	1611.71	85.8
longueur	369.57	384.25	448.00	470.14	83.0
cylindree	1212.43	1714.75	1878.58	2744.86	81.7
puissance	68.29	107.00	146.00	210.29	73.8
vitesse	161.14	183.25	209.83	229.00	68.2
largeur	164.43	171.50	178.92	180.29	67.8
hauteur	146.29	162.25	144.00	148.43	65.3
prix	11930.00	18250.00	25613.33	38978.57	82.48
CO2	130.00	150.75	185.67	226.43	59.51

La constitution des groupes s'est appuyée avant tout sur le poids, la longueur et la cylindrée (les autres variables contribuent quand même pas mal).

La segmentation se traduit par une différenciation des véhicules par les prix.

Remarque : Ce n'était pas le propos ici, mais on notera une croissance des moyennes conditionnelles dans le sens gauche – droite pour quasiment toutes les variables ($G1 < G3 < G2 < G4$) (après réarrangement). A approfondir dans l'interprétation des groupes.

Caractérisation de la partition

Variables qualitatives – V de Cramer

Une variable qualitative induit également une partition sur les observations. L'idée est de la confronter avec celle issue de la classification automatique.

Un tableau de contingence fait l'affaire.

Nombre de Groupe	Étiquett		
Étiquettes de lig	Diesel	Essence	Total général
G1	3	4	7
G2	4	8	12
G3	2	2	4
G4	3	4	7
Total général	**12**	**18**	**30**

$$v = \sqrt{\frac{0.44}{30 \times \min(4-1, 2-1)}} = 0.1206$$

Manifestement, la partition ne se traduit pas par une différenciation selon le type de carburant.

Le KHI-2 d'indépendance permet de caractériser la liaison

Le v de Cramer est une mesure issue du KHI-2 qui varie entre 0 (absence de liaison) et 1 (liaison parfaite)

$$v = \sqrt{\frac{\chi^2}{n \times \min(G-1, L-1)}}$$

Caractérisation de la partition

Variables qualitatives – Tableaux des profils

Le tableau des profils donne une idée de la nature des groupes.

Nombre de Group	Étiquettes		
Étiquettes de lig	Diesel	Essence	Total général
G1	42.86%	57.14%	100.00%
G2	33.33%	66.67%	100.00%
G3	50.00%	50.00%	100.00%
G4	42.86%	57.14%	100.00%
Total général	**40.00%**	**60.00%**	**100.00%**

*Globalement, 60% des véhicules carburent à l'« essence ». La proportion passe à 66.67% dans le groupe **G2**.*

Nombre de Group	Étiquettes		
Étiquettes de lig	Diesel	Essence	Total général
G1	25.00%	22.22%	23.33%
G2	33.33%	44.44%	40.00%
G3	16.67%	11.11%	13.33%
G4	25.00%	22.22%	23.33%
Total général	**100.00%**	**100.00%**	**100.00%**

*44.44% des véhicules à essence se retrouvent dans le groupe **G2**, lequel pèse pour 40% de la population.*

Cette idée de comparaison de proportions sera approfondie dans l'interprétation des groupes.

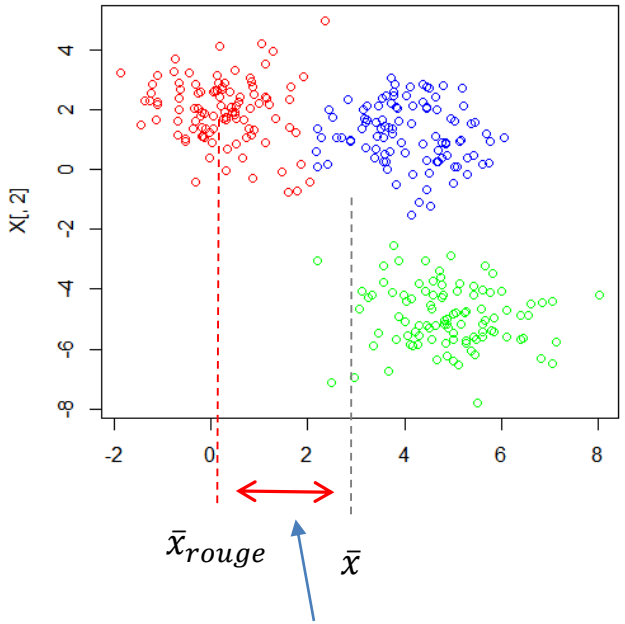
Caractérisation des groupes

Variables quantitatives – Valeur test

Comparaison des moyennes. Moyenne de la variable pour un groupe vs. Moyenne globale de la variable.

Les échantillons sont imbriqués. Au dénominateur, nous avons l'écart type de la moyenne dans le cas d'un tirage sans remise de n_k éléments parmi n.

$$vt = \frac{\bar{x}_g - \bar{x}}{\sqrt{\frac{n - n_g}{n-1} \times \frac{\sigma^2}{n_g}}}$$



L'écart est-il « significatif » ?

- σ^2 est la variance empirique calculée sur l'ensemble de l'échantillon
- n, n_k sont respectivement la taille de l'échantillon global, et celle du groupe « k »

La statistique suit très *approximativement* une loi normale (|vt| > 2, écart significatif à 5%).

Attention, contrairement aux illustratives, un test d'écart n'a pas vraiment de sens pour les variables actives parce qu'elles ont participé à la création du groupe.

Caractérisation des groupes

Variables quantitatives – Valeur test – Exemple

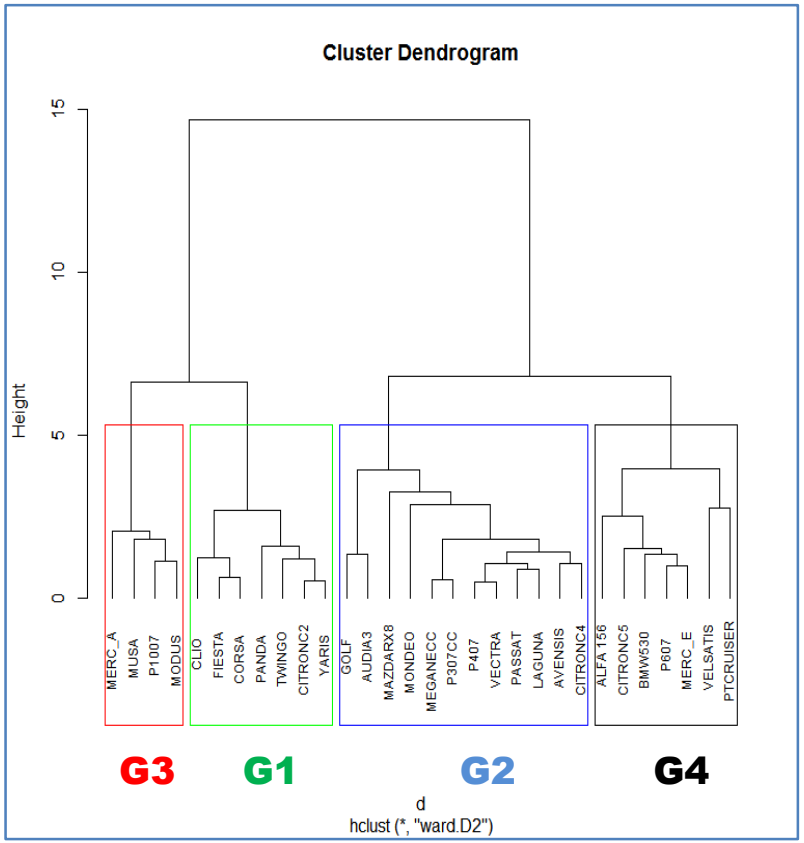
On identifie mieux la nature des groupes.

G1			
Examples			[23.3 %] 7
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
hauteur	-0.69	146.29 (4.35)	148.00 (7.36)
cylindree	-3.44	1212.43 (166.63)	1903.43 (596.98)
puissance	-3.48	68.29 (14.97)	137.67 (59.27)
vitesse	-3.69	161.14 (12.02)	199.40 (30.77)
longueur	-3.75	369.57 (17.32)	426.37 (44.99)
largeur	-3.95	164.43 (2.88)	174.87 (7.85)
poids	-4.21	952.14 (107.13)	1310.40 (252.82)

G3			
Examples			[13.3 %] 4
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
hauteur	4.09	162.25 (4.57)	148.00 (7.36)
poids	-0.58	1241.50 (80.82)	1310.40 (252.82)
cylindree	-0.67	1714.75 (290.93)	1903.43 (596.98)
largeur	-0.91	171.50 (3.70)	174.87 (7.85)
puissance	-1.09	107.00 (27.07)	137.67 (59.27)
vitesse	-1.11	183.25 (15.15)	199.40 (30.77)
longueur	-1.98	384.25 (10.66)	426.37 (44.99)

G2			
Examples			[40.0 %] 12
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
largeur	2.27	178.92 (5.12)	174.87 (7.85)
longueur	2.11	448.00 (19.90)	426.37 (44.99)
vitesse	1.49	209.83 (20.01)	199.40 (30.77)
poids	0.98	1366.58 (83.34)	1310.40 (252.82)
puissance	0.62	146.00 (39.59)	137.67 (59.27)
cylindree	-0.18	1878.58 (218.08)	1903.43 (596.98)
hauteur	-2.39	144.00 (3.95)	148.00 (7.36)

G4			
Examples			[23.3 %] 7
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
cylindree	4.19	2744.86 (396.51)	1903.43 (596.98)
puissance	3.64	210.29 (31.31)	137.67 (59.27)
poids	3.54	1611.71 (127.73)	1310.40 (252.82)
longueur	2.89	470.14 (24.16)	426.37 (44.99)
vitesse	2.86	229.00 (21.46)	199.40 (30.77)
largeur	2.05	180.29 (5.71)	174.87 (7.85)
hauteur	0.17	148.43 (5.74)	148.00 (7.36)



Calcul étendu aux variables illustratives

G1			
Examples			[23.3 %] 7
Att -	Test value	Group	Overral
Continuous attributes : Mean			
CO2	-3.08	130.00 (11.53)	177.53 (45.81)
prix	-3.5	11930.00 (3349.53)	24557.33 (10711.73)

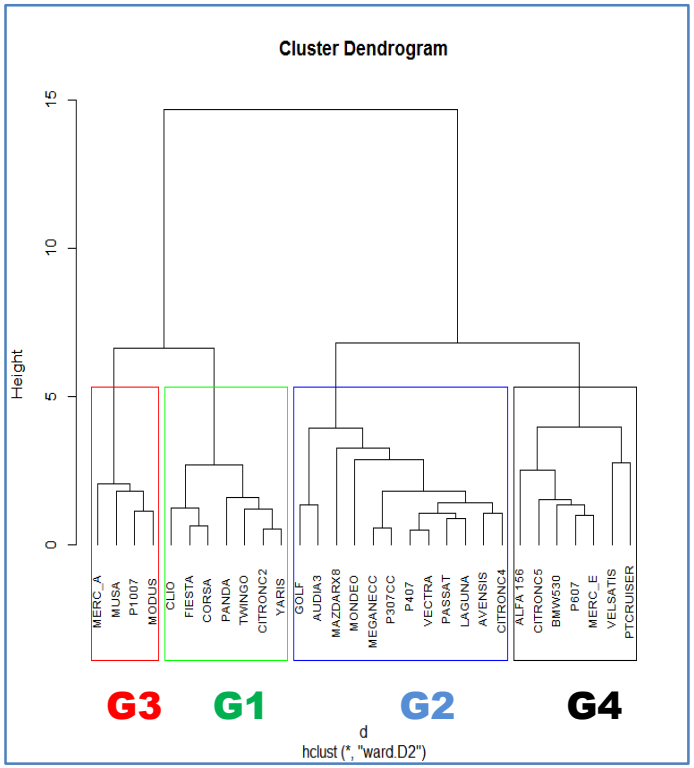
G3			
Examples			[13.3 %] 4
Att -	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
CO2	-1.23	150.75 (9.54)	177.53 (45.81)
prix	-1.24	18250.00 (4587.12)	24557.33 (10711.73)

G2			
Examples			[40.0 %] 12
Att -	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
CO2	0.78	185.67 (38.49)	177.53 (45.81)
prix	0.43	25613.33 (3879.64)	24557.33 (10711.73)

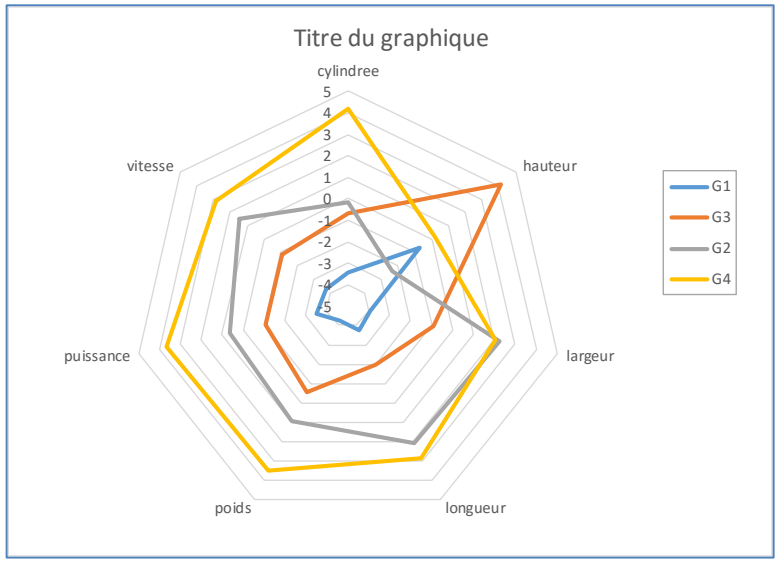
G4			
Examples			[23.3 %] 7
Att -	Test value	Group	Overral
Continuous attributes : Mean			
prix	4	38978.57 (6916.46)	24557.33 (10711.73)
CO2	3.17	226.43 (34.81)	177.53 (45.81)

Caractérisation des groupes

Variables quantitatives – Valeur test – Exemple



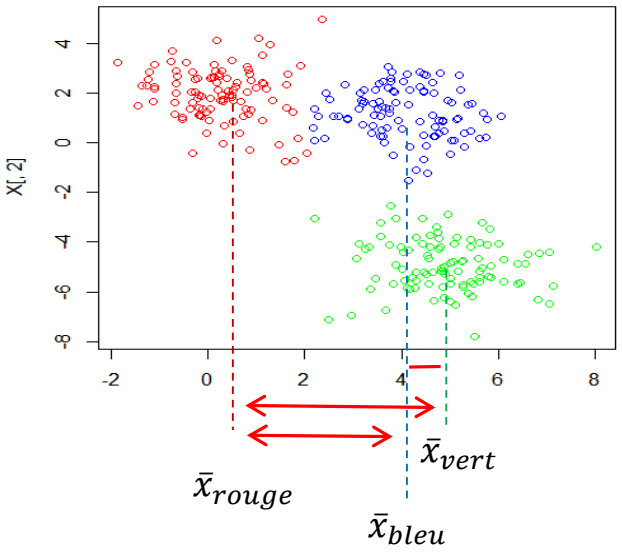
Plus que la valeur calculée des VT, ce sont les disparités et concomitances entre classes qui doivent attirer notre attention.



Il y a 4 classes, mais on se rend compte surtout qu'il y a deux types de « profils » de véhicules dans ce fichier de données. La **hauteur** *joue un rôle essentiel dans cette distinction.*

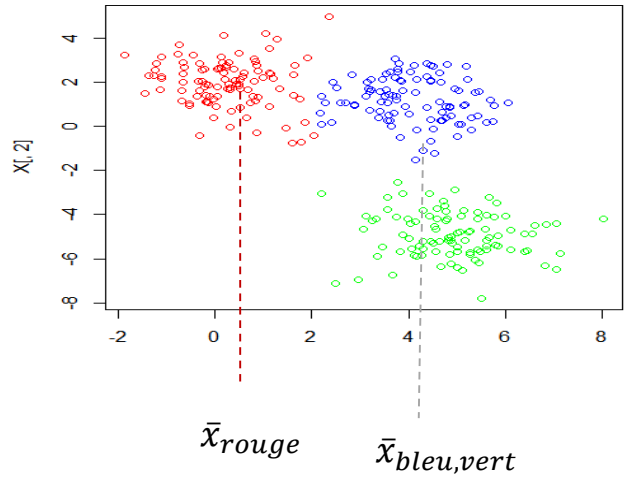
Caractérisation des groupes

Variables quantitatives – Enrichir l'analyse



On peut effectuer une comparaison deux à deux.

Le plus important est de savoir lire correctement les résultats !!!



Ou une comparaison une contre les autres.

Caractérisation des groupes

Un groupe vs. Les autres – Taille d'effet (effect size) de Cohen (1988)

La valeur test est très sensible à la taille de l'échantillon, ex. si tous les effectifs sont multipliés par 100, la VT sera multipliée par $10 = \sqrt{100}$

➔ Tous les écarts deviennent « significatifs ».

$$vt = \frac{\bar{x}_g - \bar{x}}{\sqrt{\frac{n - n_g}{n-1} \times \frac{\sigma^2}{n_g}}} = \sqrt{n_g} \times \frac{\bar{x}_g - \bar{x}}{\sqrt{\frac{n - n_g}{n-1} \times \sigma^2}}$$

La **taille d'effet** permet de dépasser cet écueil en se focalisant sur l'écart standardisé, nonobstant l'effectif des groupes.

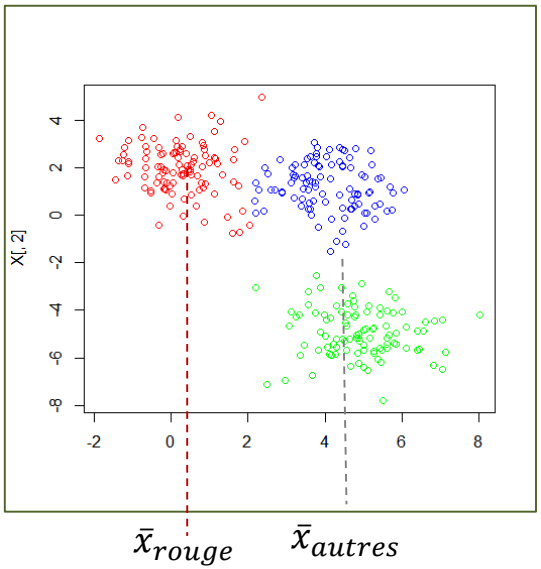
$$es = \frac{\bar{x}_k - \bar{x}_{autres}}{\sigma}$$

- La taille d'effet est insensible à la taille de la base traitée.
- La valeur s'interprète en différences en « écarts-type » (ex. 0.8 ⇔l'écart est équivalent à 0.8 fois l'écart-type). Comparaisons possibles d'une variable à l'autre.
- Quantifier les écarts en probabilités est possible également via les quantiles de la loi normale (cf. page suivante).

Caractérisation des groupes

Un groupe vs. Les autres – Taille d'effet – Illustration et lecture des résultats

Sous hypothèse de normalité des distributions !



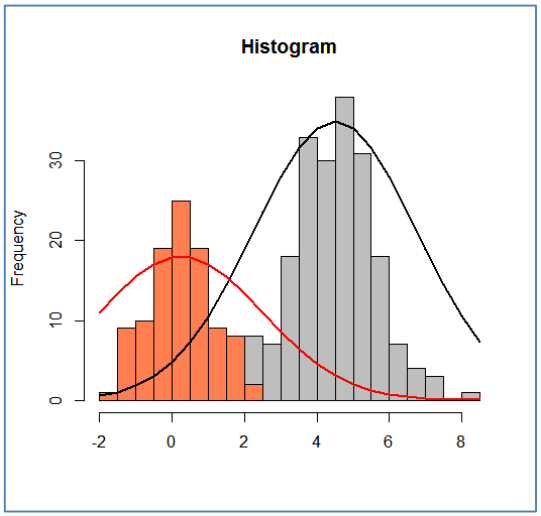
$$es = \frac{\bar{x}_{rouge} - \bar{x}_{autres}}{\sigma} = \frac{0.249 - 4.502}{2.256} = -1.885$$

Plus rigoureusement, on utiliserait l'écart-type intra (pooled) des écarts-type de « rouge » et « autres ».

Φ est la fonction de répartition de la loi normale centrée et réduite.

$$U_3 = \Phi(es) = 0.03$$

Il y a 3% de chances que les valeurs de (« bleu » et « vert ») soient en dessous de la médiane des valeurs de « rouge ». Ou 97% de chances qu'elles soient au dessus.



U2 = Φ(|es|/2) = 0.827. 82.7% des valeurs les plus élevées de « autres » excèdent 82.7% des plus faibles valeurs de « rouge ».

$\boldsymbol{U_1 = \frac{2U_2 - 1}{U_2} = 0.79}$. 79% des deux distributions ne se recouvrent pas (ou 21% des distributions se chevauchent).

D'autres variantes d'interprétations séduisantes existent (ex. CLES 'Common Language Effect Size' de McGraw et Wong, 1992)

Caractérisation des groupes

Variables qualitatives – Valeur test

Basée sur la comparaison des proportions. Proportion dans le groupe vs. Proportion dans la population globale.

Nombre de Grc	Étiquett		
Étiquettes d	Diesel	Essence	Total général
G1	42.86%	57.14%	100.00%
G2	33.33%	66.67%	100.00%
G3	50.00%	50.00%	100.00%
G4	42.86%	57.14%	100.00%
Total général	**40.00%**	**60.00%**	**100.00%**

Nombre de Grc	Étiquett		
Étiquettes d	Diesel	Essence	Total général
G1	3	4	7
G2	4	8	12
G3	2	2	4
G4	3	4	7
Total général	**12**	**18**	**30**

Fréquence du caractère dans le groupe (ex. proportion des voitures à essence parmi les G2 = 66.67%)

Fréquence du caractère dans la population (ex. proportion des voitures à essence = 60%)

$$vt = \sqrt{n_g} \times \frac{p_{l/g} - p_l}{\sqrt{\frac{n - n_g}{n-1} \times p_l \times (1 - p_l)}}$$

$$vt = \sqrt{12} \times \frac{0.6667 - 0.6}{\sqrt{\frac{30-12}{30-1} \times 0.6 \times (1-0.6)}} = 0.5986$$

vt suit une loi normale de manière très approximative, surtout valable pour les variables illustratives. Valeur critique ±2 pour un test bilatéral à 5%

vt est aussi très sensible à la taille de l'échantillon, la notion de taille d'effet peut être aussi utilisée pour les comparaisons de proportions (Cohen, chapitre 6).

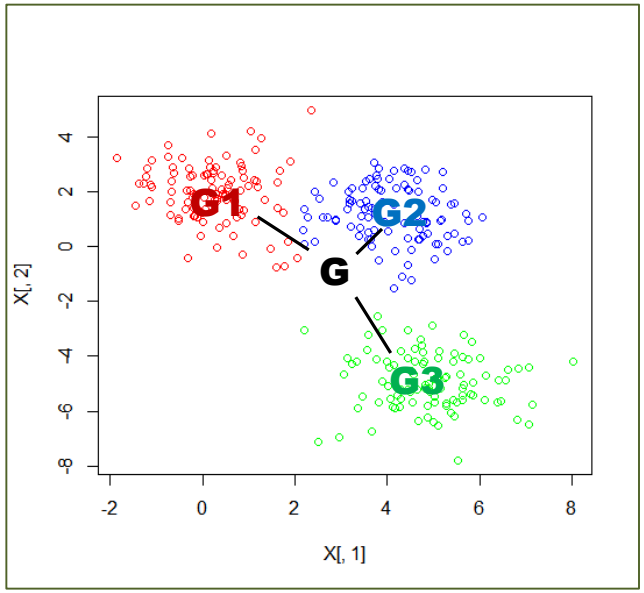
Caractérisation multivariée

Prendre en compte le rôle conjoint des variables (qui ne sont certainement pas indépendantes deux à deux)

Caractérisation de la partition

Pourcentage d'inertie expliquée

$$R^2 = \frac{4116.424}{4695.014} = 0.877$$



Remarque : il faut que les classes soient convexes pour que la mesure ait vraiment un sens c.-à.d. que le barycentre soit bien « au milieu » des points.

Relation fondamentale (Théorème d'Huygens)

$$\text{Inertie totale} = \text{Inertie inter - classes} + \text{Inertie intra - classe}$$

$$T = B + W$$

$$\sum_{i=1}^{n} d^2(i,G) = \sum_{g=1}^{G} n_g d^2(g,G) + \sum_{g=1}^{G}\sum_{i=1}^{n_g} d^2(i,g)$$

Dispersion des barycentres conditionnels autour du barycentre global.

Dispersion à l'intérieur de chaque groupe.

Généralisation multivariée du carré du rapport de corrélation.

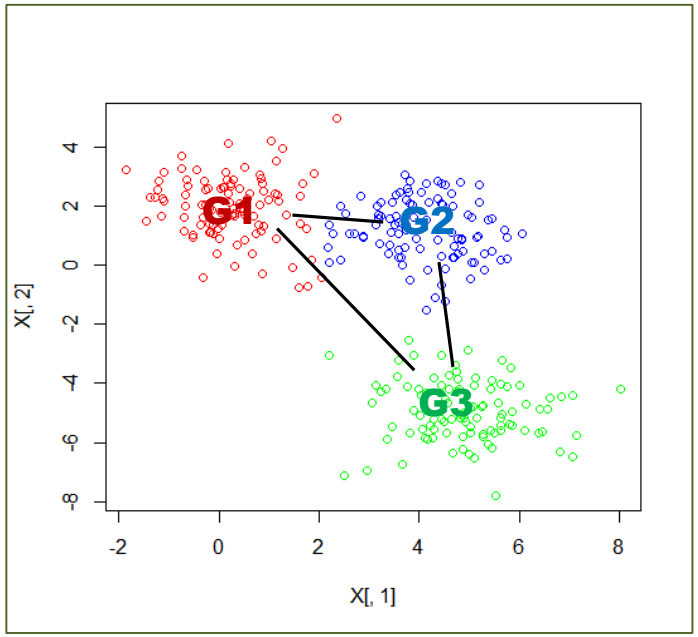
$$R^2 = \frac{B}{T}$$

Pourcentage d'inertie expliquée par la partition.

La valeur en soi est une indication, R^2 permet surtout de comparer des solutions différentes (comportant le même nombre de classes).

Caractérisation des groupes

Evaluer la proximité entre les classes



Distance entre centres de classes (carré de la distance euclidienne ici).

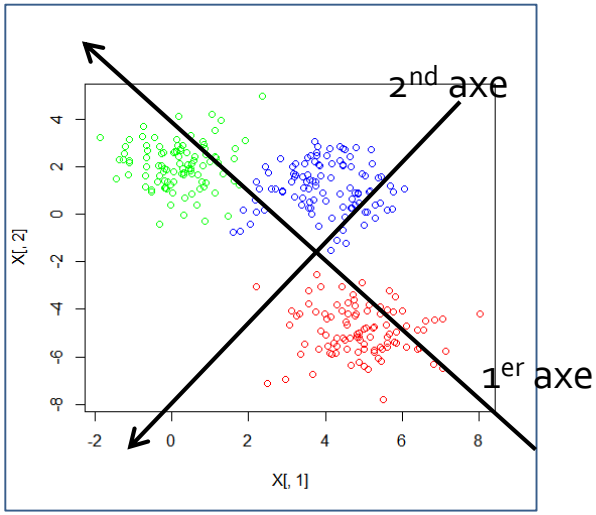
	G1	G2	G3
G1	-	15.28	71.28
G2		-	37.61
G3			-

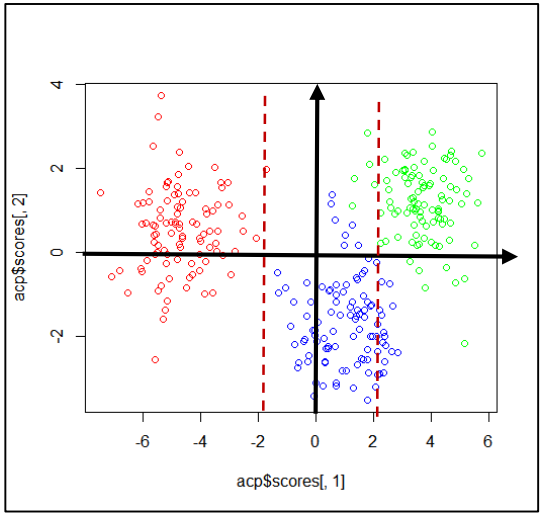
La proximité entre centres de classes doit corroborer les informations proposées entres autres par la caractérisation univariée. Sinon problème.

G1				G2				G3			
Examples		[32.7 %] 98		Examples		[34.0 %] 102		Examples		[33.3 %] 100	
Att - Desc	Test value	Group	Overral	Att - Desc	Test value	Group	Overral	Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)				Continuous attributes : Mean (StdDev)				Continuous attributes : Mean (StdDev)			
X2	9.78	2.05 (0.97)	-0.59 (3.26)	X2	6.54	1.13 (1.03)	-0.59 (3.26)	X1	10.12	4.92 (1.06)	3.06 (2.26)
X1	-15.32	0.18 (0.82)	3.06 (2.26)	X1	5.1	3.98 (1.00)	3.06 (2.26)	X2	-16.3	-4.93 (1.01)	-0.59 (3.26)

Caractérisation des groupes

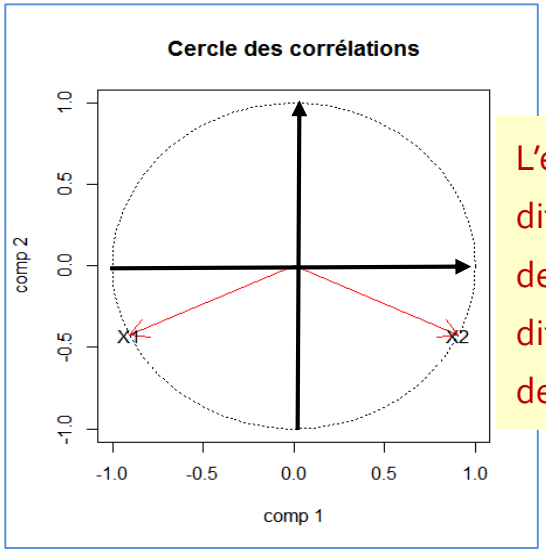
Couplage avec une analyse factorielle





On se rend compte que sur le premier axe, on distingue quasi-parfaitement les classes.

Une analyse factorielle (ACP ici puisque toutes les variables actives sont quantitatives) permet d'obtenir une vue synthétique des données, idéalement dans le plan.

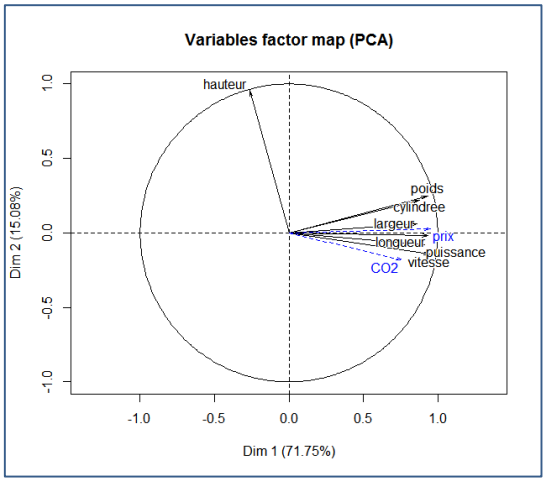


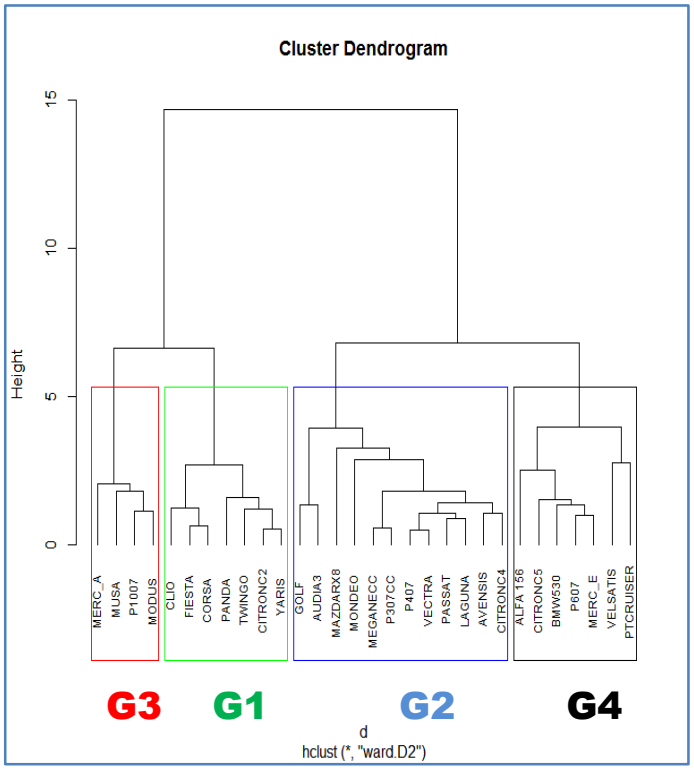
L'ennui est qu'à la difficulté d'interprétation des classes s'ajoute la difficulté d'interprétation des axes factoriels.

Caractérisation des groupes

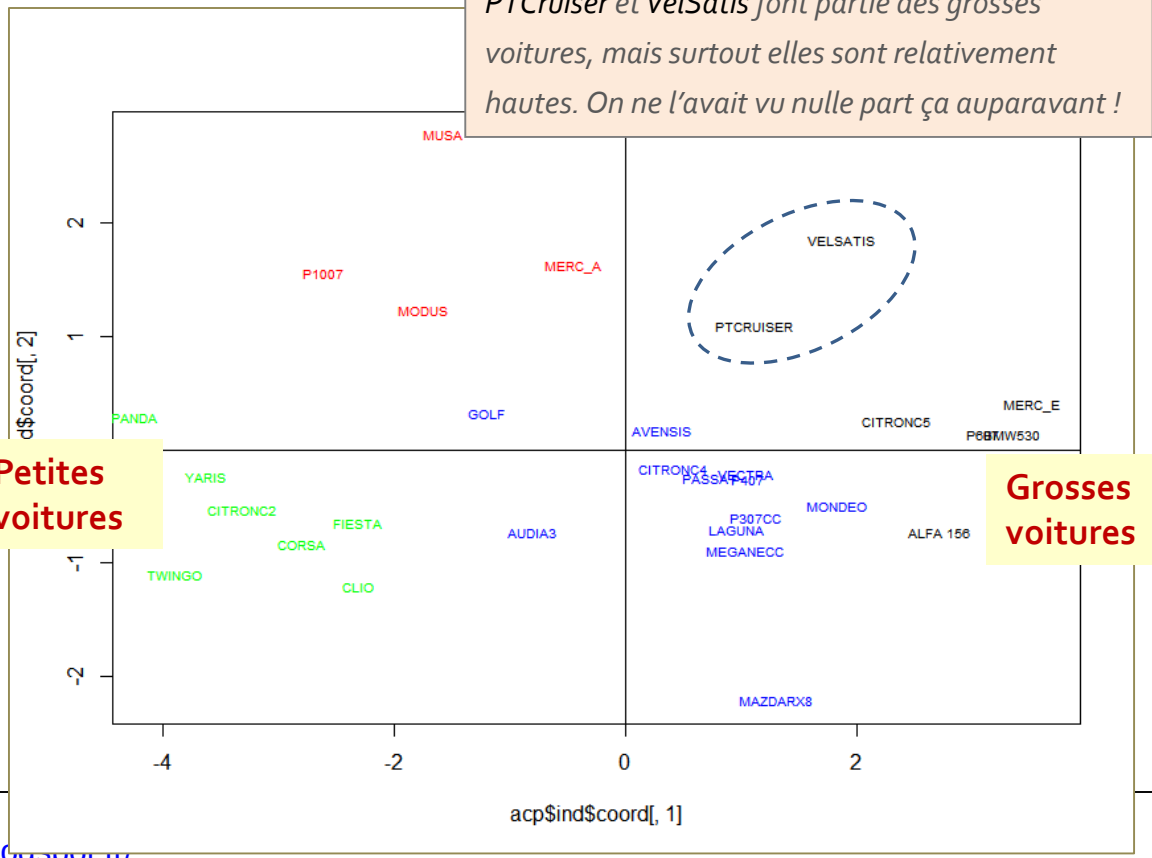
Couplage avec l'ACP – Données voitures

Le 1er axe est dominé par l'effet de quasiment toutes les variables (effet taille). Le 2nd est porté par la variable « hauteur ». On dispose de 86.83% de l'information dans le 1er plan factoriel (71.75 + 15.08).





PTCruiser et *VelSatis* *font partie des grosses voitures, mais surtout elles sont relativement hautes. On ne l'avait vu nulle part ça auparavant !*



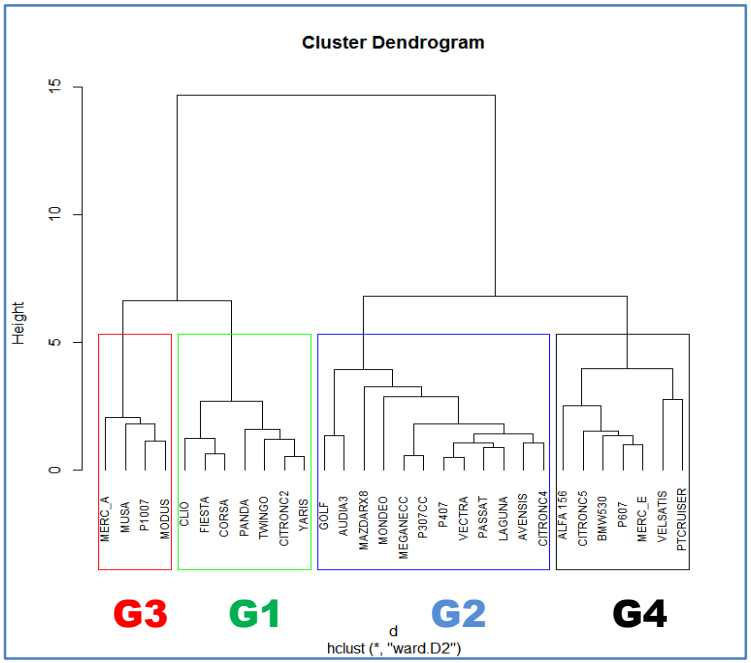
Petites voitures

Grosses voitures

Caractérisation des groupes

Utilisation des méthodes supervisées – Ex. Analyse discriminante

Prédire les groupes à l'aide d'une méthode supervisée, en profiter pour en extraire une interprétation (via les coefficients de l'analyse discriminante par ex.). On dispose directement d'une vue globale de l'influence des variables.



1ère étape : on a de la chance (parce que les classes sont convexes), la discrimination est parfaite, l'AD reproduit fidèlement la constitution des classes.

Classes prédites (analyse discriminante)

Classes observées	G3	G1	G2	G4	Total
G3	4	0	0	0	4
G1	0	7	0	0	7
G2	0	0	12	0	12
G4	0	0	0	7	7
Total	4	7	12	7	30

2ème étape : interprétation des coefficients

	Classification functions				Statistical Evaluation	
Attribute	G1	G3	G2	G4	F(3,20)	p-value
puissance	0.688092	0.803565	1.003939	1.42447	8.37255	0.001
cylindree	-0.033094	-0.027915	-0.019473	0.004058	8.19762	0.001
vitesse	3.101157	3.33956	2.577176	1.850096	9.84801	0.000
longueur	-1.618533	-1.87907	-1.383281	-1.205849	6.94318	0.002
largeur	*12.833058*	*13.640492*	*13.2026*	*13.311159*	1.21494	0.330
hauteur	19.56544	21.647641	19.706549	20.206701	16.09182	0.000
poids	*-0.145374*	*-0.122067*	*-0.130198*	*-0.118567*	0.43201	0.732
constant	-2372.594203	-2816.106674	-2527.437401	-2689.157002		

Semblent cohérents avec les analyses précédentes. Bien !

Super étrange comme résultat.

Pourquoi sont non significatifs ?

A la difficulté de reproduire exactement la partition s'ajoute les fragilités de la méthode supervisée. Dans cet exemple, clairement, des problèmes de colinéarité faussent les coefficients de certaines variables.

Conclusion

- Interpréter les classes est une étape incontournable de la classification automatique.

- Les méthodes univariées ont l'avantage de la simplicité mais ne tiennent pas compte de l'effet conjoint des variables.
- Les méthodes multivariées proposent une vue plus globale mais ne sont pas toujours faciles à appréhender.
- En pratique, il faut s'appuyer sur les deux approches pour éviter de passer à côté d'informations importantes.

- Les techniques basées sur des comparaisons de moyennes et de barycentres ne tiennent la route que si les classes sont convexes (nuages de points conditionnels relativement ovoïdes).

Bibliographie

Ouvrages

Chandon J.L., Pinson S., « Analyse typologique – Théorie et applications », Masson, 1981.

Cohen J., « Statistical Power Analysis for the Behavorial Science », 2nd Ed., Psychology Press, 1988.

Gan G., Ma C., Wu J., « Data Clustering – Theory, Algorithms and Applications », SIAM, 2007.

L. Lebart, A. Morineau, M. Piron, « Statistique exploratoire multidimensionnelle », Dunod, 2000.

Tutoriels

« Classification automatique sous R », octobre 2015.

« Classification automatique sous Python », mars 2016.

« Interpréter la valeur test », avril 2008.