Interpreting cluster analysis results

With numeric and categorical variables (active and/or illustrative)

Ricco RAKOTOMALALA
Université Lumière Lyon 2

Outline

1. Interpreting the cluster analysis results
2. Univariate characterization
   a. Of the clustering structure
   b. Of the clusters
3. Multivariate characterization
   a. Percentage of explained variance
   b. Distance between the centroids
   c. Combination with factor analysis
   d. Utilization of a supervised approach (e.g. discriminant analysis)
4. Conclusion
5. References

Cluster analysis

Clustering, unsupervised learning

Cluster analysis

Also called: clustering, unsupervised learning, typological analysis

"Active" input variables, used for the creation of the clusters. Often (but not always) numeric variables

"Illustrative" variables. Used only for the interpretation of the clusters. To understand on which characteristics are based the clusters.

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

Goal: Identifying the set of objects with similar characteristics

We want that:

(1) The objects in the same group are more similar to each other

(2) Thant to those in other groups

For what purpose?

→ Identify underlying structures in the data

→ Summarize behaviors or characteristics

→ Assign new individuals to groups

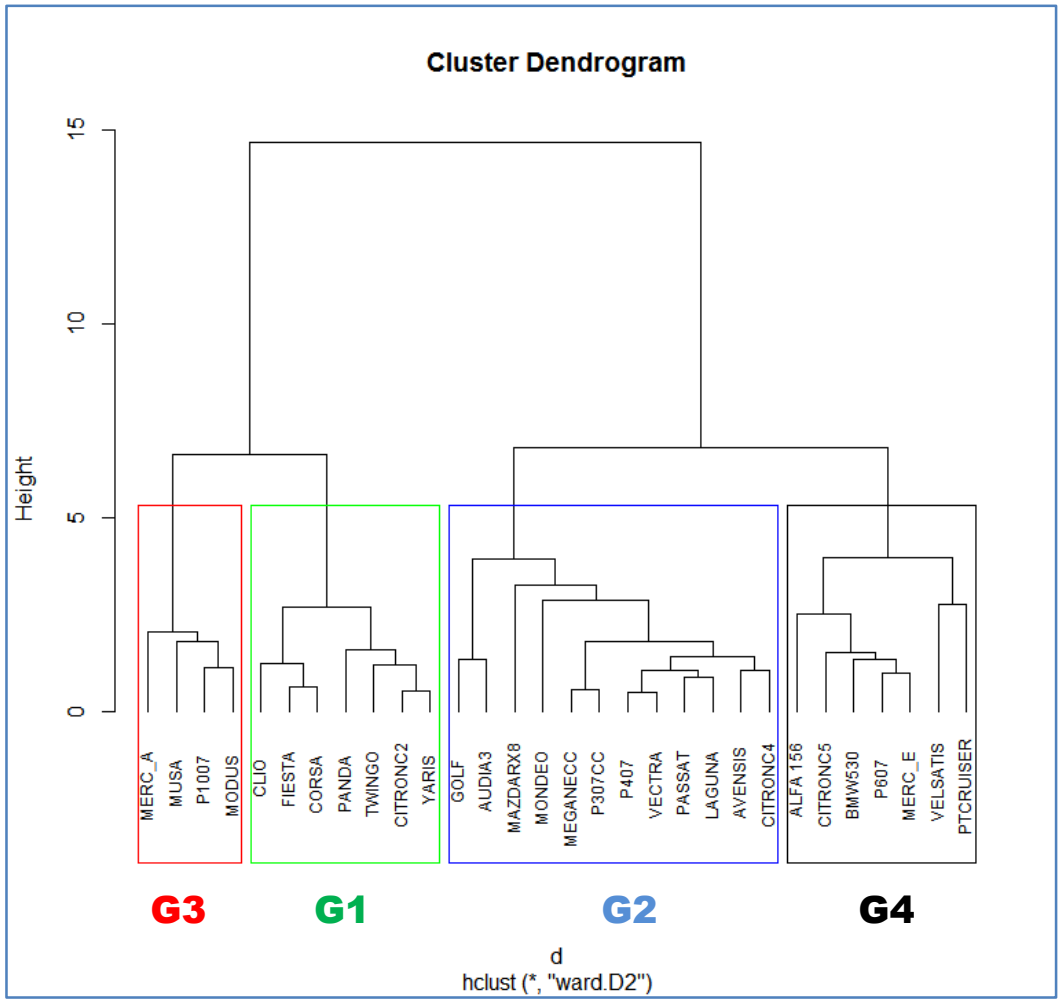
→ Identify totally atypical objects

The aim is to detect the set of "similar" objects, called groups or clusters.

"Similar" should be understood as "which have close characteristics".

Cluster analysis

Interpreting clustering results



On which kind of information are based the results?

To what extent the groups are far from each other?

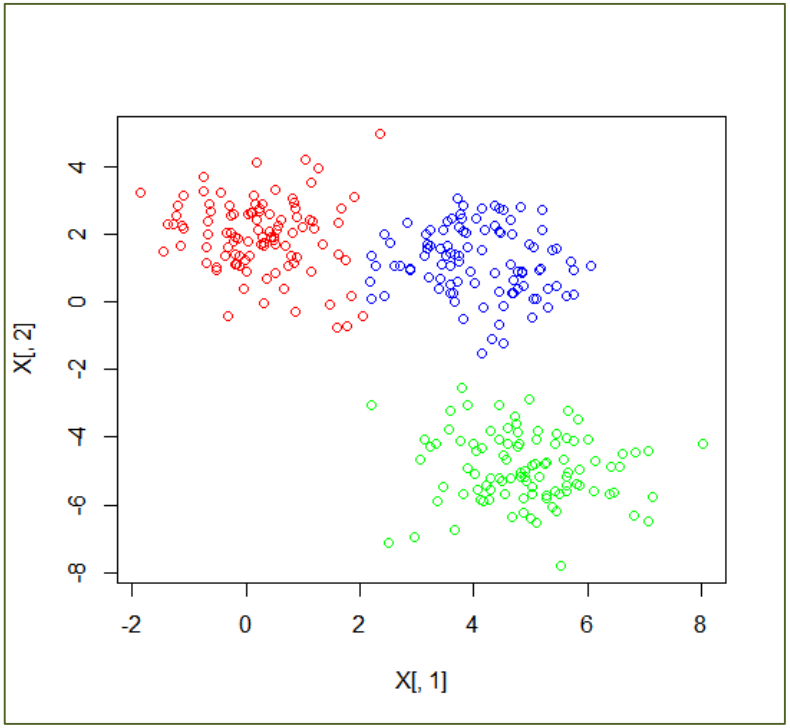
What are the characteristics that share individuals belonging to the same group and differentiate individuals belonging to distinct groups?

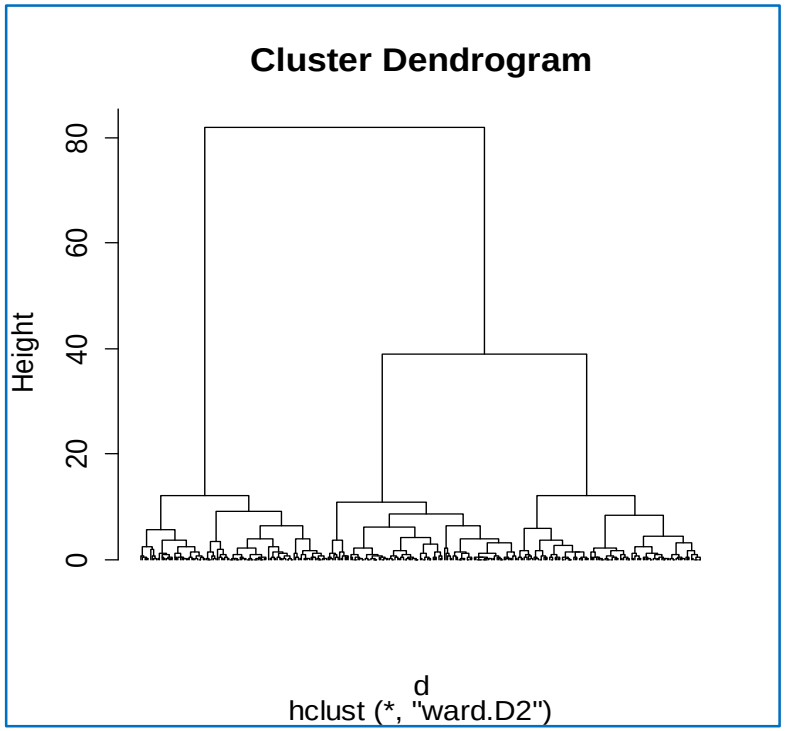
In view of active variables used during the construction of the clusters.

But also regarding the illustrative variables which provide another point of view about the nature of the clusters.

Cluster analysis

An artificial example in a two dimensional representation space





This example will help to understand the nature of the calculations achieved to characterize the clustering structure and the groups.

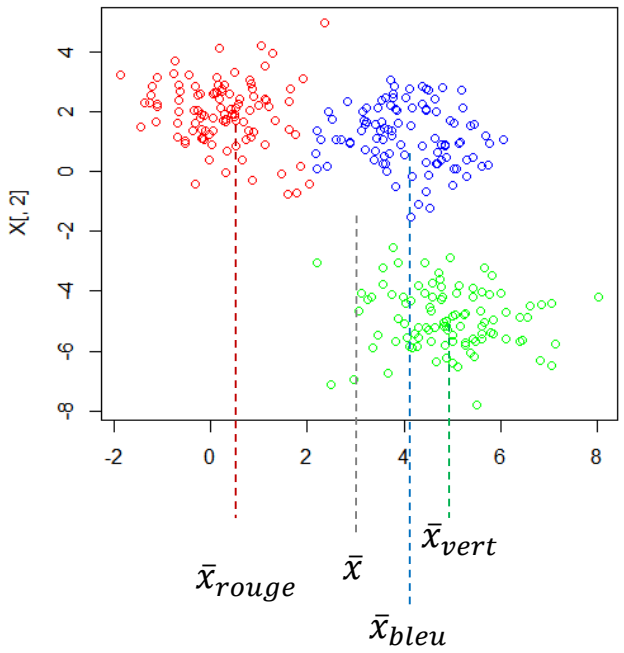
Univariate characterization

Interpretation using the variables taken individually

Characterizing the partition

Quantitative variable

Evaluate the importance of the variables, taken individually, in the construction of the clustering structure



The idea is to measure the proportion of the variance (of the variable) explained by the group membership

Huygens theorem

$$TOTAL.SS = BETWEEN\text{-}CLUSTER.SS + WITHIN-CLUSTER.SS$$

$$T = B + W$$

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

The square of the correlation ratio is defined as follows:

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

η^2 corresponds to the proportion of the variance explained ($0 \leq \eta^2 \leq 1$). We can interpret it, with caution, as the influence of the variable in the clustering structure.

Characterizing the partition

Quantitative variables – Cars dataset

Conditional means

	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

The formation of the groups is based mainly on weight (*poids*), length (*longueur*) and engine size (*cylindrée*). But the other variables are not negligible (we can suspect that almost all the variables are highly correlated for this dataset).

About the illustrative variables, we observe that the groups correspond mainly a price differentiation.

Note: After a little reorganization, we observe that the conditional means increase from the left to the right (G1 < G3 < G2 < G4). We further examine this issue when we interpret the clusters.

Characterizing the partition

Categorical variables – Cramer's V

A categorical variable leads also to a partition of the dataset. The idea is to study its relationship with the partition defined by the clustering structure.

We use a crosstab (contingency table)

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**

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

Obviously, the clustering structure does not correspond to a differentiation by the fuel-type (carburant).

The chi-squared statistic enables to measure the degree of association.

The Cramer's v is a measure based on the chi-squared statistic with varies between 0 (no association) and 1 (complete association).

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

Characterizing the partition

Rows and columns percentages

The rows and columns percentages provide often an idea about the nature of the groups.

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%**

The overall percentage of the cars that uses "gas" (essence) fuel-type is 60%. This percentage becomes 66.67% in the cluster G2. There is (very slight) an overrepresentation of the "fuel-type = gas" vehicles into this group.

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% of the vehicles "fuel-type = gas" (essence) are present in the cluster G2, which represent 40% of the dataset.

This idea of comparing proportions will be examined in depth for the interpretation of the clusters.

Characterizing the clusters

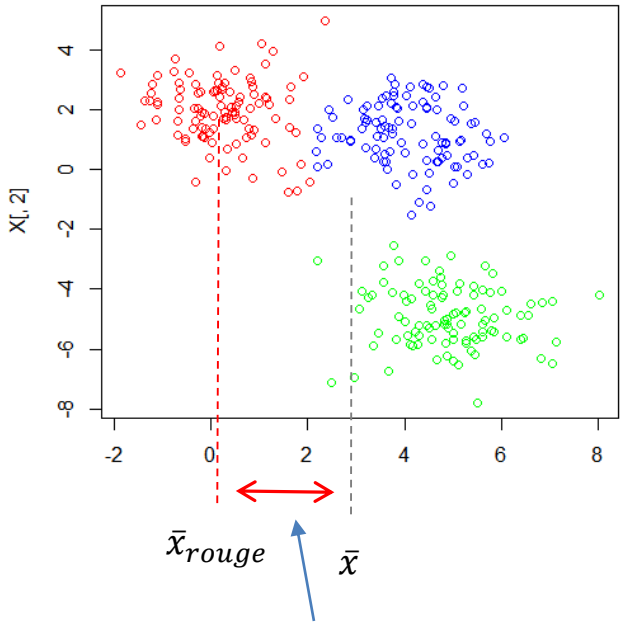
Quantitative variables – V-Test (test value) criterion

Comparison of means. Mean of the variable for the cluster "g" (conditional mean) vs. Overall mean of the variable.

The samples are nested. We see in the denominator the standard error of the mean in the case of a sampling without replacement of n_g *instances among* n.

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

- σ^2 is the empirical variance for the whole sample
- n, n_g are respectively the size of whole sample and the cluster "g"



Is the difference ***significant****?*

The test statistic is distributed approximately as a normal distribution ($|vt| > 2$, critical region at 5% level for a test of significance).

Unlike for illustrative variables, the V-test for test of significance does not really make sense for active variables because they have participated in the creation of the groups. But it can be used for ranking the variables according their influence.

Characterizing the clusters

Quantitative variables – V-test – Example

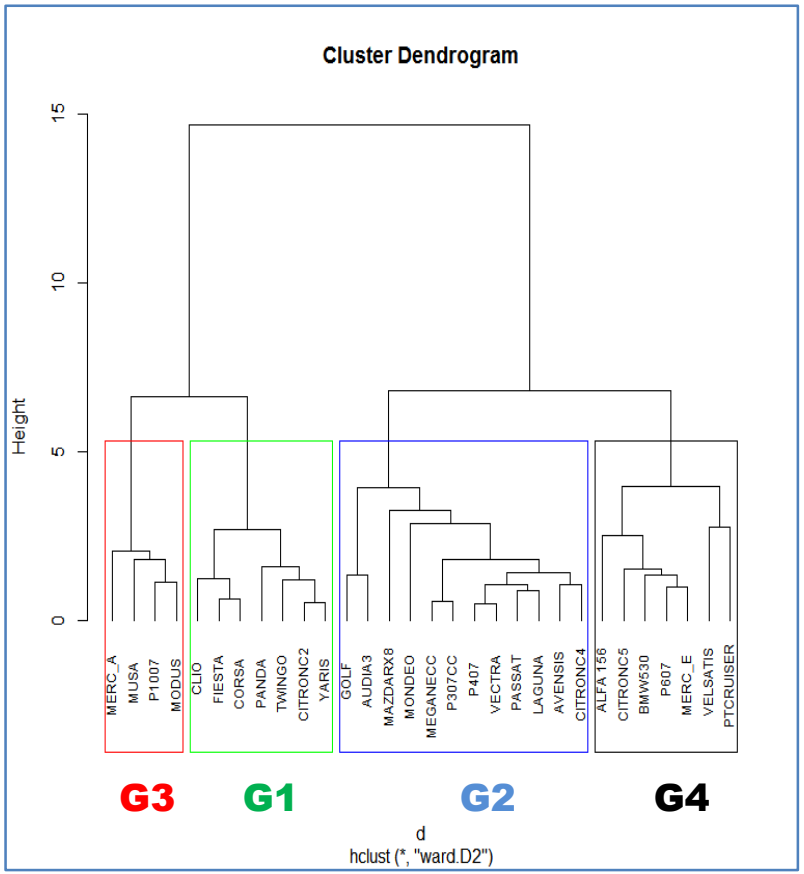
We understand better the nature of the clusters.

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)



The calculations are extended to illustrative variables.

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)

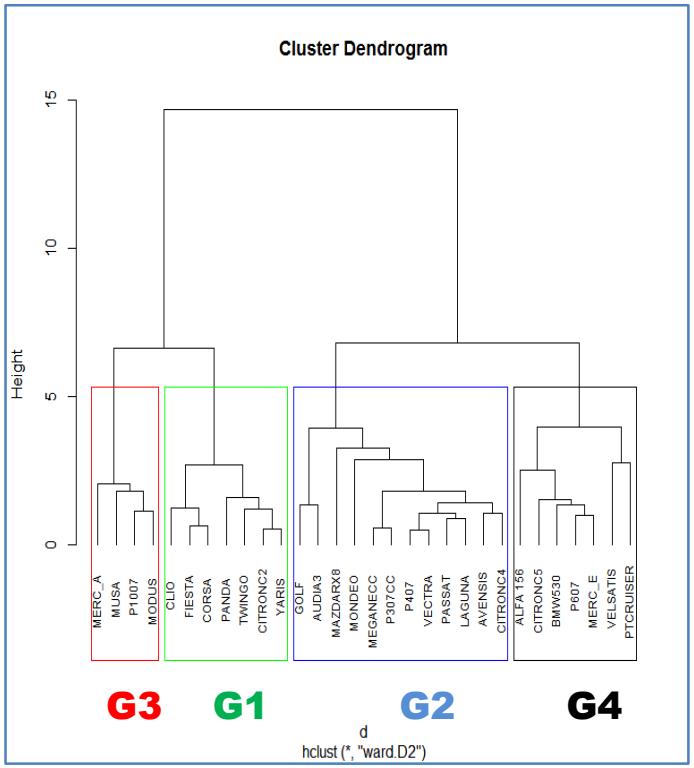
G3			
Examples			[13.3 %] 4
Att -	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
CO2	-1.23	150.75 (9.54)	177.53 (45.81)
prix	-1.24	18250.00 (4587.12)	24557.33 (10711.73)

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)

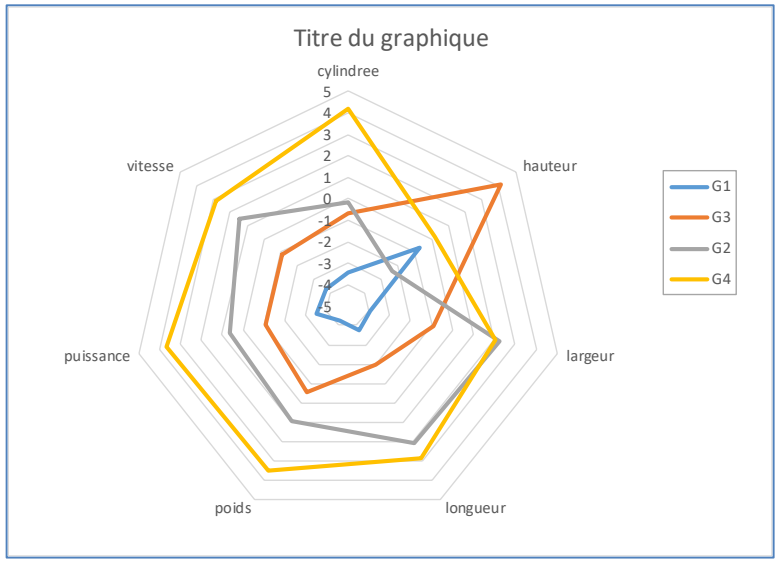
G4			
Examples			[23.3 %] 7
Att -	Test value	Group	Overral
Continuous attributes : Mean			
prix	4	38978.57 (6916.46)	24557.33 (10711.73)
CO2	3.17	226.43 (34.81)	177.53 (45.81)

Characterizing the clusters

Quantitative variables – V-test – Example



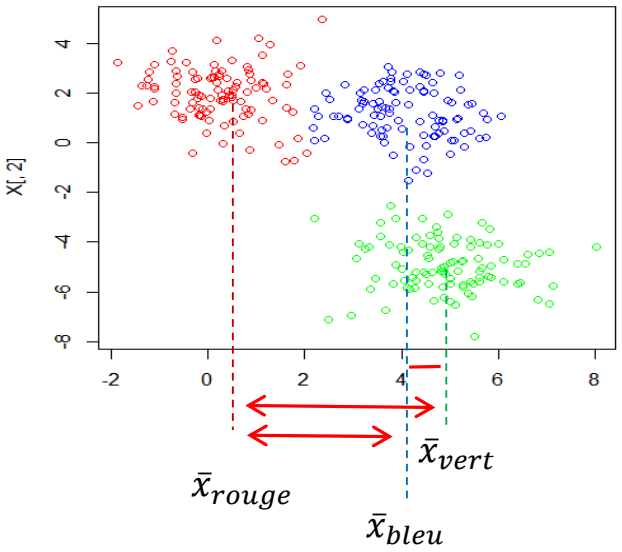
Instead of the computed value of V-TEST, it is the discrepancy and similarity between groups that must draw our attention.



*There are 4 classes, but we realize that there are mainly two types of vehicles in the dataset ({G1, G3} vs. {G2, G4}). The height (**hauteur**) plays a major role in the distinction of the clusters.*

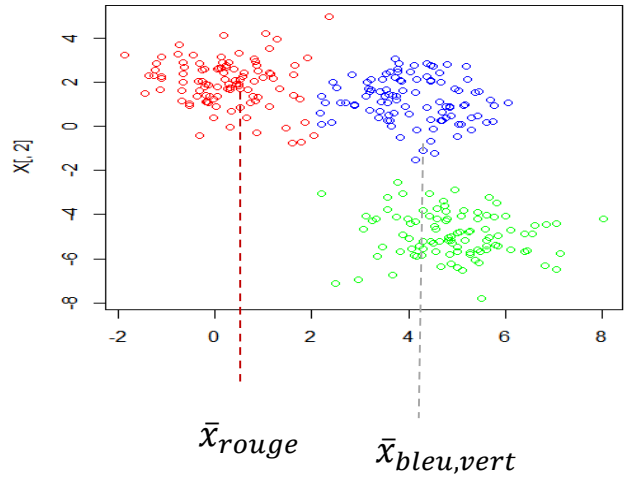
Characterizing the clusters

Quantitative variables – Supplement the analysis



we can make pairwise comparisons.

The most important thing is to know how to read properly the results!!!



Or the comparison of one cluster vs. the others.

Characterizing the clusters

One group vs. the others – Effect size (Cohen, 1988)

The V-Test is highly sensitive to the sample size.

E.g. If the sample size is multiplied by 100, the V-Test is multiplied by $10 = \sqrt{100}$

➔ All the differences become "significant"

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

The effect size notion allows to overcome this drawback. It is focused on the standardized difference, disregarding the sample size.

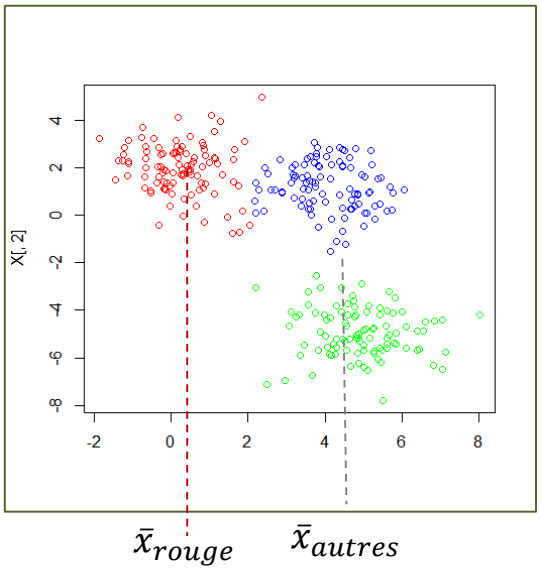
$$es = \frac{\bar{x}_g - \bar{x}_{\text{others}}}{\sigma}$$

- The effect size is insensitive to the sample size.
- The value can be read as a difference in terms of standard deviation (e.g. 0.8 ⇔ the difference corresponds to 0.8 times the standard error). It makes possible a comparison between different variables.
- Interpreting the effect size as difference between probabilities is also possible (using the quantile of normal distribution).

Characterizing the clusters

One group vs. the others – Effect size – Interpreting the results

Under the assumption of normal distribution



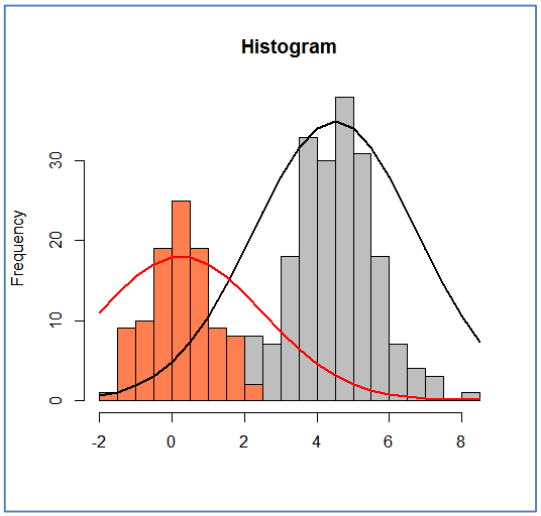
$$es = \frac{\bar{x}_{rouge} - \bar{x}_{autres}}{\sigma} = \frac{0.249 - 4.502}{2.256} = -1.885$$

More strictly, we would use the pooled standard deviation.

Φ cumulative distribution function (cdf) of the standardized normal distribution

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

There are 3% of chance that the values of the "others" groups are lower than the median of the "red" group.



U2 = Φ(|es|/2) = 0.827. 82.7% of the highest values of "others" are higher than 82.7% of the lowest values of "red".

$\boldsymbol{U_1 = \frac{2U_2 - 1}{U_2} = 0.79}$. 79% of the two distributions are not overlapped.

Other kinds of interpretation are possible (e.g. CLES 'Common Language Effect Size' of McGraw & Wong, 1992)

Characterizing the clusters

Categorical variables – V-Test

Based on the comparison of proportions.
Proportion of a category in the studied group vs. its proportion in the whole sample.

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**

Frequency of the category into the group of interest (e.g. proportion of 'fuel-type: gas' among G2 = 66.67%)

Frequency of the category into the whole sample (e.g. proportion of 'fuel-type: gas' = 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 is distributed approximately as a normal distribution. It is especially true for the illustrative variables. Critical value ±2 for a two-sided significance test at 5% level

vt is very sensitive to the sample size. The effect size notion can be used also for the comparison of proportions (Cohen, chapter 6).

Multivariate characterization

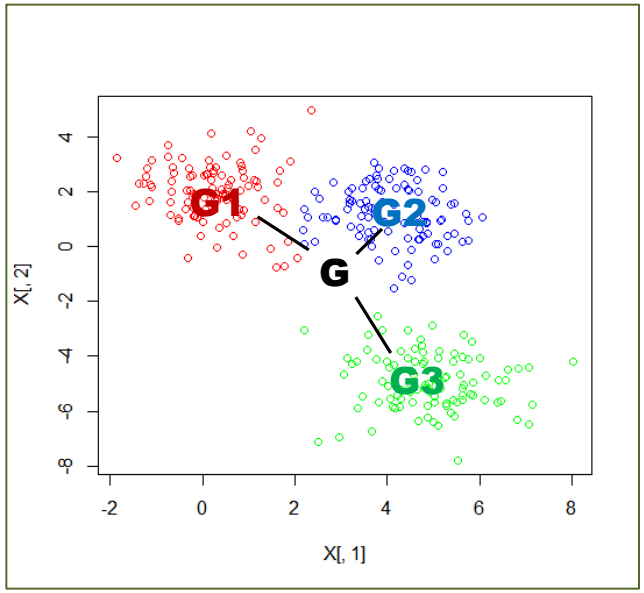
Take into account the interaction between the variables

(which are sometimes highly correlated)

Characterizing the partition

Percentage of variance explained

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



Note: For ensuring that the measure is valid, the clusters must have a convex shape i.e. the centroids are approximately at the center of the clusters.

Huygens theorem

$$\text{Total.SS} = \text{Between - Cluster.SS} + \text{Within - cluster.SS}$$

$$T = B + W$$

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

Dispersion of the conditional centroids in relation to the overall centroid.

Dispersion inside each cluster.

Multivariate generalization of the square of the correlation ratio.

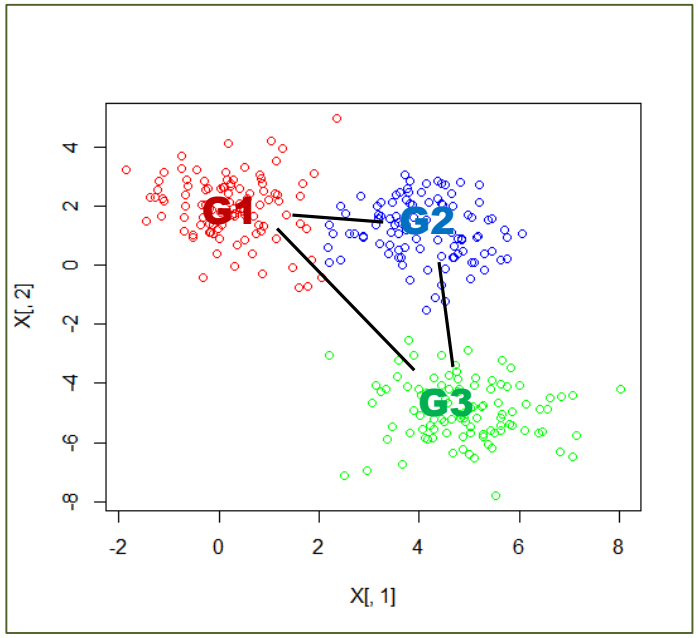
$$R^2 = \frac{B}{T}$$

Proportion of variance explained

The R^2 criterion allows to compare the efficiency of various clustering structures, only if they have the same number of clusters.

Characterizing the partition

Evaluating the proximity between the clusters



Distance between the centroids (Squared Euclidean distance for this example)

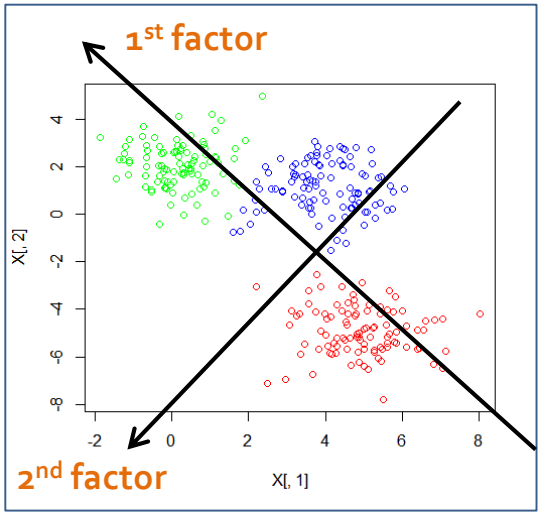
	G1	G2	G3
G1	-	15.28	71.28
G2		-	37.61
G3			-

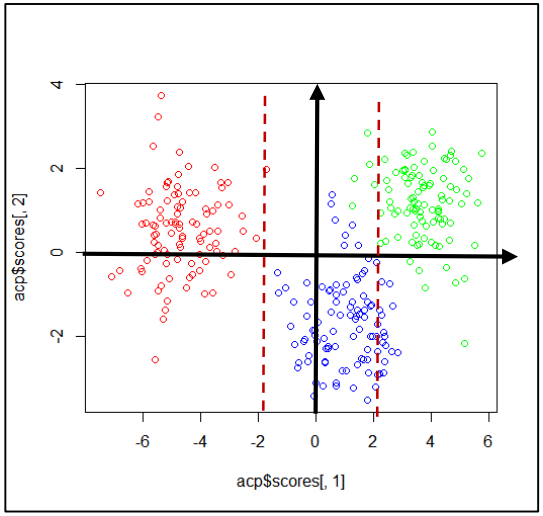
The closeness between the centroids must confirm the results provided by the other approaches, especially the univariate approach. If not, there are issues that need deeper analysis.

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)

Characterizing the clusters

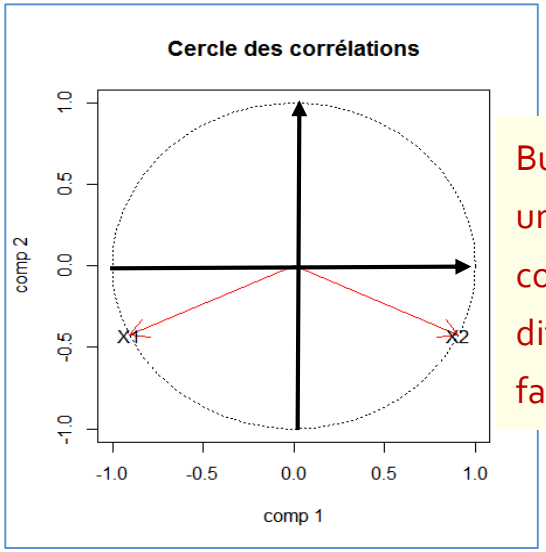
In combination with factor analysis





We observe that the clusters are almost perfectly separable on the first factor.

A factor analysis (principal component analysis - PCA - here since all the active variables are numeric) allows to obtain a synthetic view of the data, ideally in a two dimensional representation space.

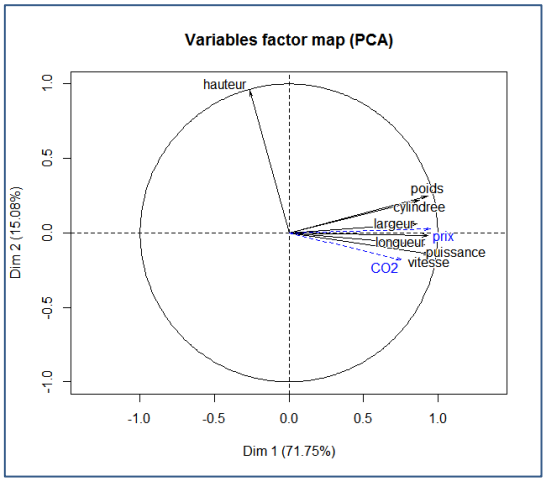


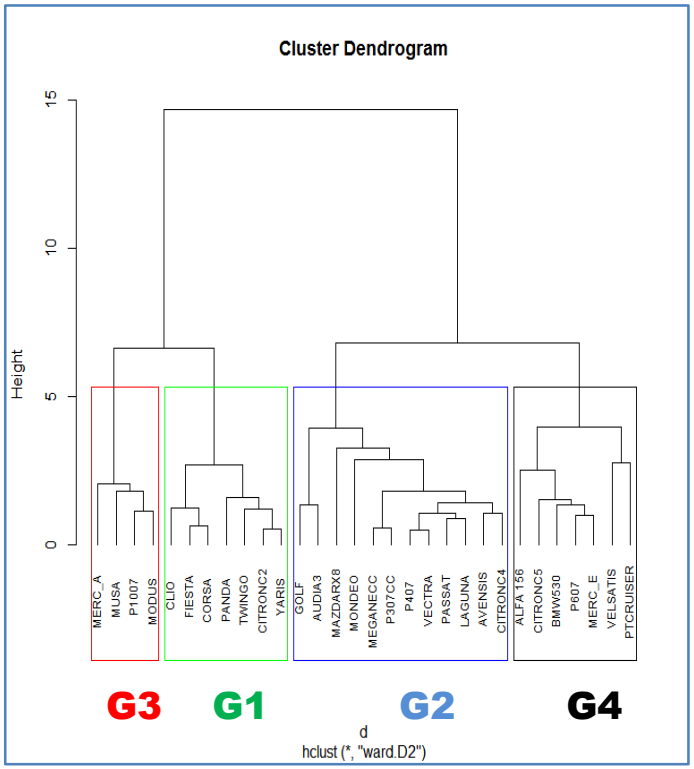
But to the difficulty to understand the clusters comes in addition the difficulty to interpret the factor analysis results.

Characterizing the clusters

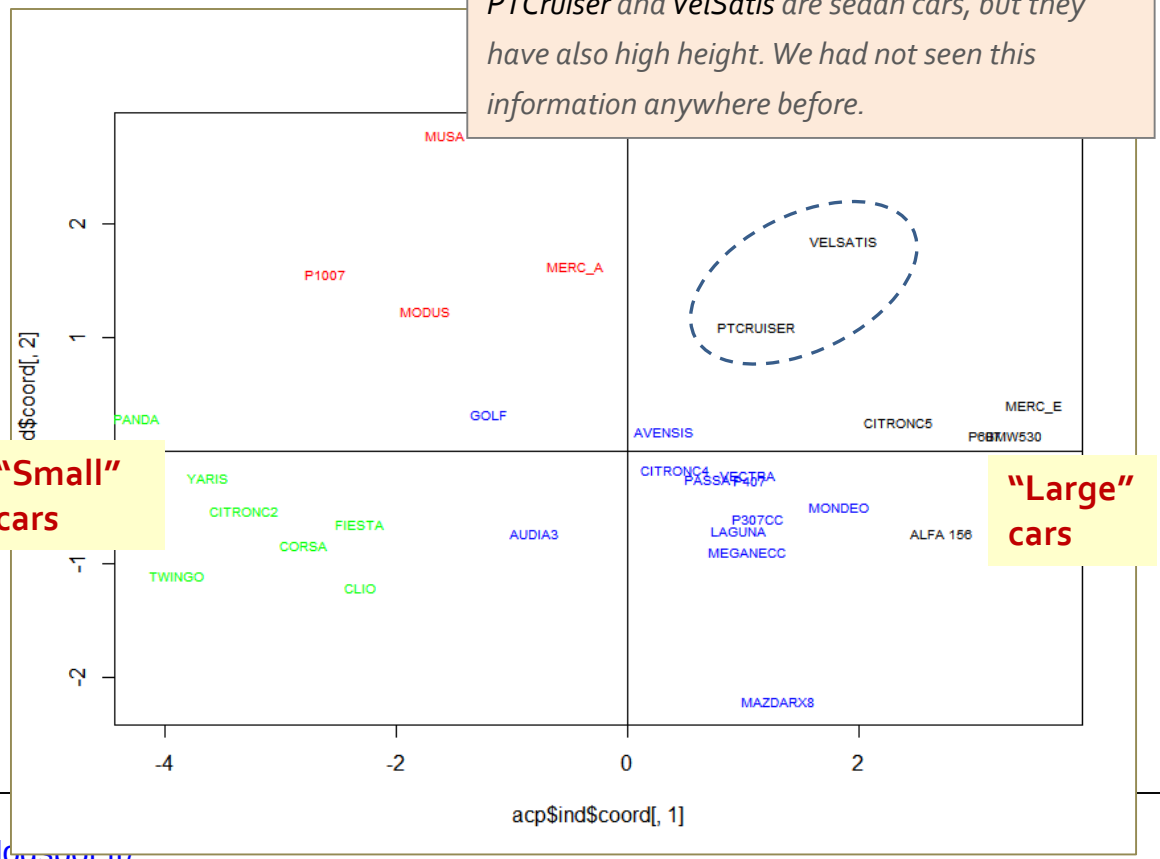
Principal component analysis - Cars dataset

The first factor is dominated by the size of the cars (large cars have big engine, etc.). The 2nd factor is based on the height (hauteur) of cars. We have 86.83% of the information on this first two-dimensional representation space (71.75 + 15.08).





PTCruiser and *VelSatis* are sedan cars, but they have also high height. We had not seen this information anywhere before.



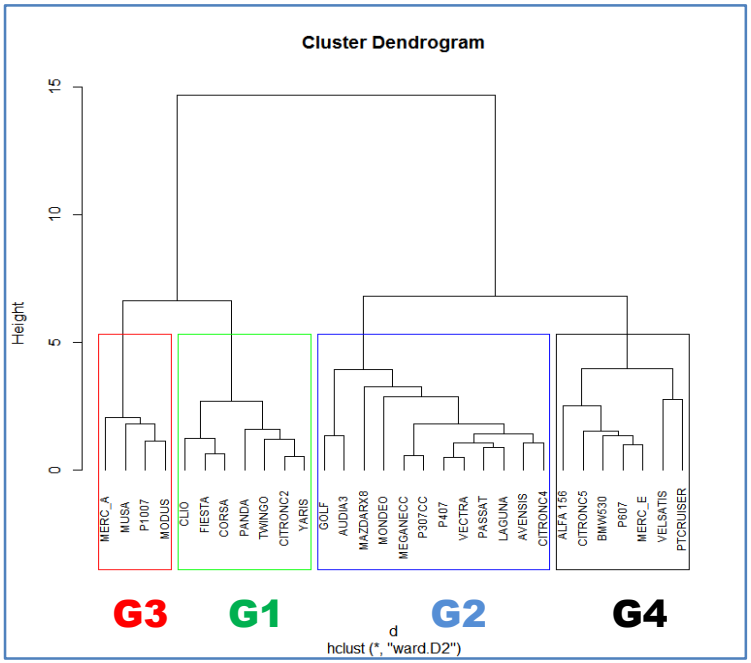
"Small" cars

"Large" cars

Characterizing the clusters

Using supervised approach – E.g. Discriminant Analysis

We predict the clusters membership using a supervised learning algorithm. We have an overall point of view about the influence of the variables.



1st step: we are lucky (because the clusters provided by the K-Means algorithm are convex), we have a perfect discrimination. The discriminant analysis allows to recreate perfectly the clusters for our dataset.

Predicted clusters (linear discriminant analysis - LDA)

Observed clusters	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

2nd step: interpretation of the LDA 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		

These results seem consistent with the previous analysis. Comforting!

On the other hand, this is a very strange result. The speed (vitesse) seems to influence differently the clusters. We know that it is not true in the light of the PCA conducted previously.

Why these variables are not significant?

To the difficulty of recreating exactly the clusters is added the weakness of the supervised method. In this example, clearly, the coefficients of some variables are distorted by the multicollinearity.

Conclusion

- Interpreting the clustering results is a vital step in cluster analysis.

- Univariate approaches have the advantage of the simplicity. But they do not take into account the joint effect of the variables.
- Multivariate methods offer a more global view but the results are not always easy to understand.
- In practice, we have to combine the two approaches to avoid missing out important information.

- The approaches based on comparisons of means and centroids are relevant only if the clusters have convex shape.

References

Books

(FR) 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.

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

Tanagra Tutorials

"Understanding the 'test value' criterion", May 2009.

"Cluster analysis with R – HAC and K-Means", July 2017.

"Cluster analysis with Python – HAC and K-Means", July 2017.