

1 Objectif

Analyse factorielle des données mixtes – Comparaison des logiciels Tanagra et R.

Habituellement, on utilise l'analyse en composantes principales (ACP) lorsque toutes les variables actives sont quantitatives, l'analyse des correspondances multiples (ACM ou AFCM) lorsqu'elles sont toutes catégorielles. Mais que faire lorsque nous avons un mix des deux types de variables ?¹

Une stratégie possible consiste à découper en classes les variables quantitatives et de passer par une ACM. Elle est néanmoins peu avantageuse lorsque (1) le nombre d'observations est faible, en effet l'ACM donne des résultats peu stables dans ce cas ; (2) le nombre de variables catégorielles est faible par rapport aux variables quantitatives, en effet le découpage en classes n'est pas anodin, il induit souvent une perte d'information.

Une autre piste consiste à remplacer les variables catégorielles par une série d'indicateurs via un codage disjonctif complet. L'information n'est pas dénaturée car il y a une bijection exacte entre les variables initiales et les variables recodées. Mais il n'est pas très judicieux de traiter directement avec une ACP des données mélangeant des colonnes de valeurs numériques (continues) et des indicateurs. Les dispersions étant différentes, il y a de fortes chances que les résultats soient biaisés.

L'analyse factorielle des données mixtes (AFDM) de Jérôme Pagès (Pagès, 2004) repose sur cette seconde piste, mais elle introduit une subtilité supplémentaire. A l'instar de l'ACP normée où l'on réduit les variables (c'est une forme de recodage) pour uniformiser leurs influences, il propose de substituer au codage 0/1 des variables qualitatives un codage 0/x où « x » est soigneusement calculé à partir des fréquences des modalités. On peut dès lors utiliser un programme usuel d'ACP pour mener l'analyse (Pagès, 2004 ; page 102). Les calculs donc bien maîtrisés. L'interprétation des résultats requiert en revanche un effort supplémentaire puisqu'elle sera différente selon que l'on étudie le rôle d'une variable quantitative ou qualitative.

Dans ce tutoriel, nous montrons la mise en œuvre de l'AFDM avec les logiciels **Tanagra 1.4.46** et **R 2.15.1** (package FactoMineR). Nous mettrons l'accent sur la lecture des résultats. Il faut pouvoir analyser simultanément l'impact des variables quantitatives et qualitatives lors de l'interprétation des facteurs. Les outils graphiques sont très précieux dans cette perspective.

2 Données

Nous travaillons sur le fichier AUTOS2005AFDM.TXT accessible en ligne². Nous disposons de 38 modèles de véhicules décrits par 'q = 12' variables actives : puissance, cylindrée, vitesse, longueur, largeur, hauteur, poids, CO2 et prix sont quantitatives ; origine (France, Europe, Autres), carburant (diesel, essence) et type 4x4 (oui, non) sont qualitatives.

¹ Cette introduction repose sur les idées développées dans l'article de Jérôme Pagès, « Analyse factorielle de données mixtes », Revue de Statistique Appliquée, tome 52, n°4, 2004 ; pages 93-111. Le texte est accessible en ligne : http://archive.numdam.org/ARCHIVE/RSA/RSA_2004__52_4/RSA_2004__52_4_93_0/RSA_2004__52_4_93_0.pdf

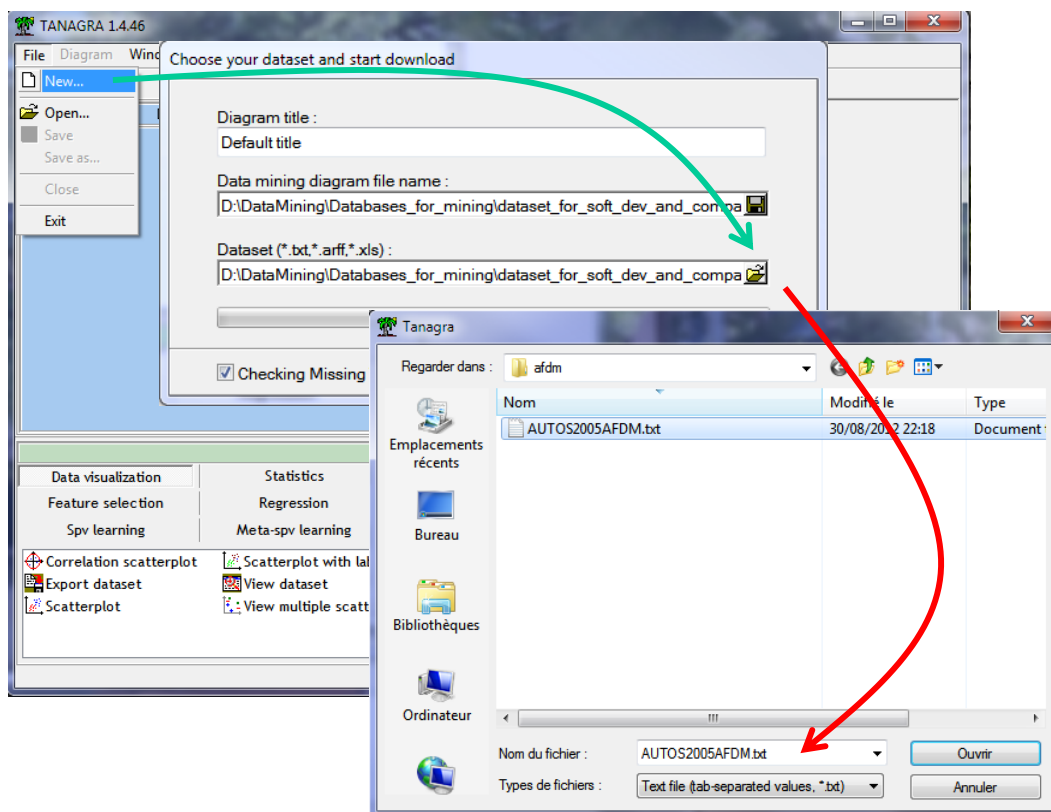
² Cours de Pierre-Louis Gonzalez au CNAM : « [STA101 – Analyse des données : méthodes descriptives](#) ». Certaines variables ont été supprimées, 2 observations manifestement atypiques itou.

Clairement, le découpage en classes des 9 premières variables pour pouvoir passer à l'ACM, outre la difficulté technique que cela représente (choix du nombre d'intervalles, choix des bornes de découpage) et la perte d'information qui en résulte, n'est pas très judicieuse au regard du faible nombre des variables qualitatives de l'étude. L'AFDM s'impose naturellement dans ce contexte.

3 AFDM avec TANAGRA

3.1 Création d'un diagramme et importation des données

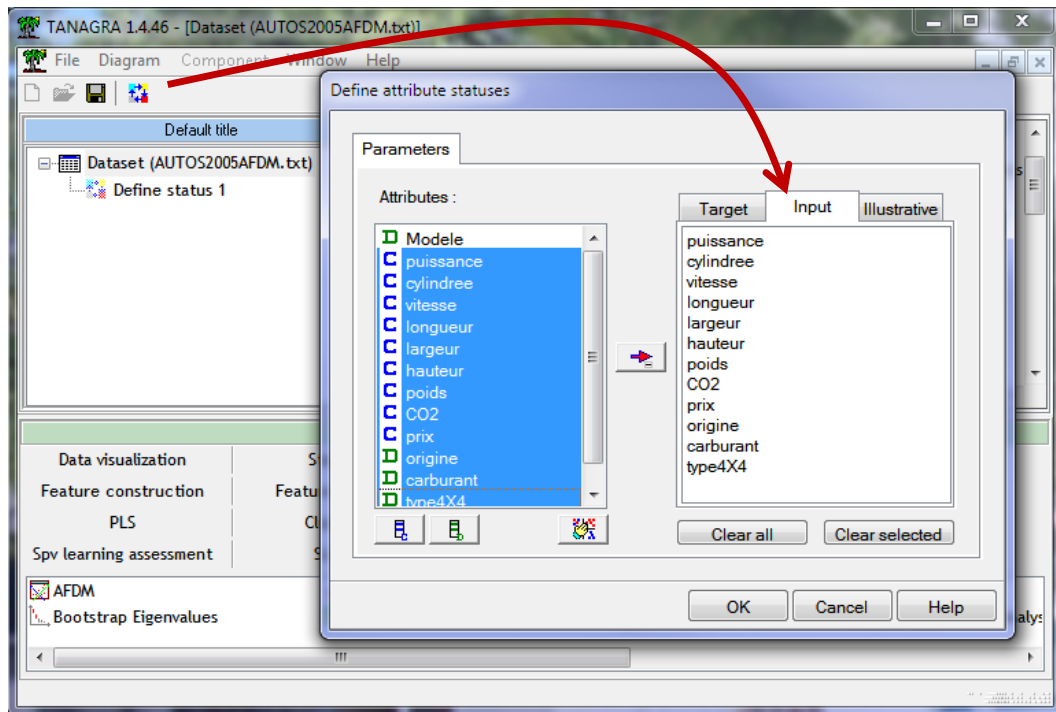
Après avoir démarré Tanagra, nous actionnons le menu FILE / NEW pour créer un nouveau diagramme. Nous sélectionnons le fichier AUTOS2005AFDM.TXT (fichier texte avec séparateur tabulation). Nous validons en cliquant sur OK.



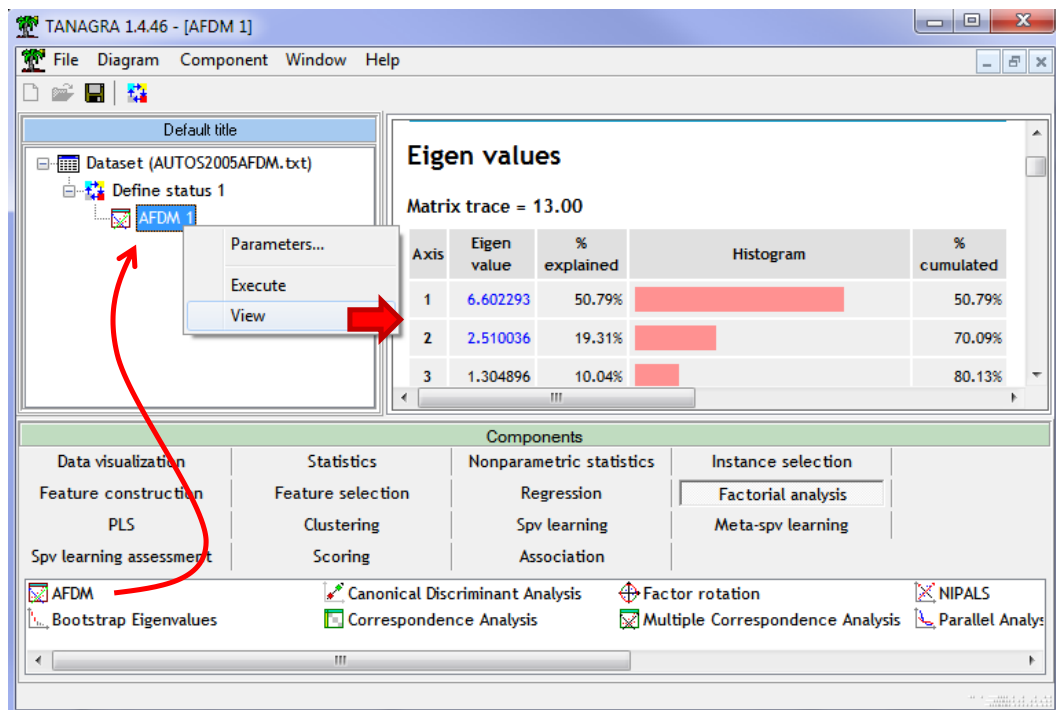
Les données sont importées. Nous obtenons une description des variables : MODELE correspond à l'étiquette des observations, PUISSANCE...PRIX sont les variables quantitatives, ORIGINE...TYPE4X4 sont les qualitatives.

3.2 Le composant AFDM

Pour pouvoir lancer l'AFDM, nous devons dans un premier temps spécifier les variables actives. Nous utilisons le composant DEFINE STATUS. A l'exception de MODELE qui est une étiquette permettant de reconnaître individuellement chaque véhicule, nous plaçons toutes les variables en INPUT.



Nous ajoutons ensuite le composant AFDM (onglet FACTORIAL ANALYSIS). Nous cliquons sur VIEW pour lancer les calculs et afficher les résultats.



Voyons en le détail.

3.3 Interprétation des résultats

3.3.1 Tableau des valeurs propres

Il indique la variance expliquée par les axes. Nous avons 'p = 13' facteurs car nous avons 9 variables quantitatives et 3 qualitatives avec respectivement 3, 2 et 2 modalités. Dans le tableau de données utilisé pour les calculs internes, nous avons 9 + 3 + 2 + 2 = 16 colonnes. Mais, nous avons introduit

artificiellement de la colinéarité. En effet, la somme des indicatrices d'une variable qualitative vaut systématiquement 1. De fait, le nombre de valeurs propres non nulles est en réalité égale à : $9 + [(3-1)+(2-1)+(2-1)] = 13$. Confirmé par les résultats ici, la somme des valeurs propres est bien égale à 13, et les 13 premiers axes expliquent 100% de l'inertie totale.

Eigen values				
Matrix trace = 13.00				
Axis	Eigen value	% explained	Histogram	% cumulated
1	6.602293	50.79%		50.79%
2	2.510036	19.31%		70.09%
3	1.304896	10.04%		80.13%
4	0.866767	6.67%		86.80%
5	0.557532	4.29%		91.09%
6	0.389581	3.00%		94.09%
7	0.267694	2.06%		96.14%
8	0.172089	1.32%		97.47%
9	0.140012	1.08%		98.55%
10	0.096673	0.74%		99.29%
11	0.050890	0.39%		99.68%
12	0.031080	0.24%		99.92%
13	0.010457	0.08%		100.00%
Tot.	13.000000	-	-	-

Le choix du nombre d'axes est une question récurrente en analyse factorielle. Tanagra a mis en surbrillance les deux premiers. Pour un risque approximatif de 10%, il sélectionne les axes portés par une valeur propre supérieure à (Karlis, Saporta et Spinakis ; 2003)³ :

$$seuil = 1 + 1.65 \sqrt{\frac{p-1}{n-1}}$$

Où 'p' est le nombre total de valeurs propres théoriquement non nulles ($p = 13$), et 'n' le nombre d'observations ($n = 38$). En ce qui nous concerne, le seuil est égal à

$$seuil = 1 + 1.65 \sqrt{\frac{13-1}{38-1}} = \mathbf{1.94}$$

Les deux premiers axes (**6.60** et **2.51**) passent haut la main ce seuil. Nous les conservons pour l'interprétation. Nous remarquons qu'ils traduisent 70.09% de l'information disponible. Ce qui est relativement élevé. En effet, contrairement à l'ACP, et à l'instar de l'ACM, du fait de la présence de variables qualitatives, j'ai constaté que l'inertie expliquée est plus dispersée sur les axes en AFDM. Elle le sera d'autant plus que la proportion de variables qualitatives dans la base (et le nombre de modalités associées) augmente.

³ D. Karlis, G. Saporta and A. Spinakis, « A simple rule for the selection of principal components », in Communications in Statistics - Theory and Methods, Vol. 32, n°3, 2003 ; pp. 643-666.

Ce seuil est plus restrictif que la règle usuelle du Kaiser (seuil = 1). Il a le mérite de prendre en compte à la fois la dimension maximale (p) et l'effectif de l'échantillon (n). Malgré tout, il faut être prudent par rapport à une valeur brute, sans nuances. On sait par expérience que la subtilité est de mise en statistique exploratoire. Un résultat est fait pour être interprété. Dans notre cas, peut être que le 3^{ème} axe est intéressant en réalité. Seule l'inspection de ses caractéristiques peut nous le dire.

3.3.2 Tableau des coordonnées

Le tableau des coordonnées décrit l'impact des variables, qu'elles soient quantitatives ou qualitatives, dans la définition des axes. Pour les premières, il s'agit du carré du coefficient de corrélation linéaire ; pour les secondes, les valeurs correspondent au carré du rapport de corrélation.

Squared Correlation (Communalities)										
Attribute	Axis_1		Axis_2		Axis_3		Axis_4		Axis_5	
	Coord.	% (Tot. %)	Coord.	% (Tot. %)	Coord.	% (Tot. %)	Coord.	% (Tot. %)	Coord.	% (Tot. %)
puissance (*)	0.838788	84 % (84 %)	0.073700	7 % (91 %)	0.015920	2 % (93 %)	0.006117	1 % (93 %)	0.007588	1 % (94 %)
cylindree (*)	0.789024	79 % (79 %)	0.001540	0 % (79 %)	0.002086	0 % (79 %)	0.001076	0 % (79 %)	0.012564	1 % (81 %)
vitesse (*)	0.494831	49 % (49 %)	0.357998	36 % (85 %)	0.001496	0 % (85 %)	0.001546	0 % (86 %)	0.047079	5 % (90 %)
longueur (*)	0.807418	81 % (81 %)	0.020192	2 % (83 %)	0.016856	2 % (84 %)	0.021477	2 % (87 %)	0.019884	2 % (89 %)
largeur (*)	0.767254	77 % (77 %)	0.001494	0 % (77 %)	0.035147	4 % (80 %)	0.016086	2 % (82 %)	0.013577	1 % (83 %)
hauteur (*)	0.083172	8 % (8 %)	0.715836	72 % (80 %)	0.000764	0 % (80 %)	0.002187	0 % (80 %)	0.125430	13 % (93 %)
poids (*)	0.838023	84 % (84 %)	0.061973	6 % (90 %)	0.009873	1 % (91 %)	0.034234	3 % (94 %)	0.006094	1 % (95 %)
CO2 (*)	0.794686	79 % (79 %)	0.008110	1 % (80 %)	0.107026	11 % (91 %)	0.024402	2 % (93 %)	0.003601	0 % (94 %)
prix (*)	0.884497	88 % (88 %)	0.003877	0 % (89 %)	0.015896	2 % (90 %)	0.000283	0 % (90 %)	0.011391	1 % (92 %)
origine (**)	0.158125	8 % (8 %)	0.328641	16 % (24 %)	0.611781	31 % (55 %)	0.665082	33 % (88 %)	0.170196	9 % (97 %)
carburant (**)	0.000238	0 % (0 %)	0.300166	30 % (30 %)	0.434590	43 % (73 %)	0.093448	9 % (83 %)	0.138175	14 % (97 %)
type4X4 (**)	0.146235	15 % (15 %)	0.636509	64 % (78 %)	0.053462	5 % (84 %)	0.000828	0 % (84 %)	0.001952	0 % (84 %)
Var. Expl.	6.602293	51 % (51 %)	2.510036	19 % (70 %)	1.304896	10 % (80 %)	0.866767	7 % (87 %)	0.557532	4 % (91 %)

(*) Square of correlation coefficient
(**) Correlation ratio

Figure 1 - Tableau des coordonnées

La somme des valeurs en ligne vaut « 1 » pour les variables quantitatives, « nombre de modalités – 1 » pour les qualitatives. Ainsi, le pourcentage en ligne (%) indique la part d'information de la variable qui est retranscrite par l'axe. Pour la variable puissance par exemple, 84% de l'information qu'elle véhicule est restituée par le premier axe. Il y a très peu de chances qu'elle pèse significativement sur les autres. Le cumul des pourcentages en ligne vaut 1.

La somme des valeurs en colonne est égale à la valeur propre associée à l'axe. Nous pouvons y voir la contribution de la variable dans la définition du facteur. Tanagra n'affiche pas les pourcentages en colonne parce cela revient à détecter les plus fortes valeurs simplement. Ici, on se rend compte que puissance, cylindrée, longueur, largeur, poids, CO2 et prix sont celles qui pèsent le plus sur la définition du premier axe. Nous ne savons pas en revanche quel est le sens des relations entre ces différentes variables. Les tableaux suivants préciseront cette information.

Tanagra utilise la règle suivante pour mettre en surbrillance les valeurs : (1) l'axe doit être significatif c.-à-d. sa valeur propre est supérieur au seuil ci-dessus (section 3.3.1) ; (2) la proportion en ligne doit être supérieur à '1/p' (nombre maximal de facteurs) ; (3) la proportion en colonne doit être supérieure à '1/q' (nombre de variables de l'étude).

3.3.3 Tableau des corrélations

Ce tableau précise le sens des relations entre les variables quantitatives et les facteurs.

Attribute	Axis_1	Axis_2	Axis_3	Axis_4	Axis_5
puissance	0.915854	0.271477	-0.126173	0.078209	0.087112
cylindree	0.888270	0.039245	0.045671	0.032799	-0.112091
vitesse	0.703443	0.598329	0.038675	-0.039319	0.216976
longueur	0.898565	0.142099	0.129831	-0.146551	0.141009
largeur	0.875931	0.038653	0.187474	-0.126832	-0.116522
hauteur	0.288396	-0.846071	0.027647	-0.046766	-0.354161
poids	0.915436	-0.248944	0.099364	-0.185024	-0.078065
CO2	0.891451	0.090057	-0.327149	0.156212	-0.060012
prix	0.940477	0.062264	0.126079	-0.016837	-0.106727

Figure 2 - Tableau des corrélations

Nous notons un effet taille fort sur le premier axe : les grosses cylindrées sont puissantes, chères, polluantes, et imposantes (longues et larges). Sur le second axe, sans tenir compte des variables qualitatives, nous notons une opposition entre la hauteur des véhicules et leur vitesse c.-à-d. à 'taille' (au sens du premier axe) égale, les véhicules hauts s'opposent aux rapides. Les variables origine, carburant et type4x4 devraient préciser l'idée. C'est ce que nous verrons dans la section suivante.

3.3.4 Tableau des moyennes conditionnelles

Ce tableau positionne les modalités sur les axes factoriels. Nous avons également une indication sur leurs contributions. Elles dépendent à la fois de l'écartement avec l'origine, de l'effectif et de la valeur propre associée à l'axe. **La somme des contributions des modalités doit être égale à la contribution de la variable.**

Attribute		Axis_1		Axis_2		Axis_3		Axis_4		Axis_5	
		Mean	CTR	Mean	CTR	Mean	CTR	Mean	CTR	Mean	CTR
origine	Autres	1.5877	1.5218	-1.5116	9.5438	-0.8961	12.4089	0.4682	7.6798	0.5140	22.3670
	France	-1.0414	0.8512	0.6577	2.3488	-0.5750	6.6428	-1.0505	50.2475	-0.1547	2.6335
	Europe	-0.1559	0.0220	0.4377	1.2005	1.0957	27.8319	0.5982	18.8041	-0.2086	5.5262
	Tot.	-	2.3950	-	13.0931	-	46.8835	-	76.7314	-	30.5266
carburant	Diesel	0.0441	0.0020	-0.9647	6.6087	0.8370	18.4052	-0.3163	5.9580	0.3085	13.6960
	Essence	-0.0357	0.0016	0.7810	5.3499	-0.6776	14.8994	0.2561	4.8232	-0.2497	11.0873
	Tot.	-	0.0036	-	11.9587	-	33.3046	-	10.7812	-	24.7833
type4X4	oui	2.5243	1.9235	-3.2472	22.0219	-0.6785	3.5579	0.0688	0.0829	0.0848	0.3041
	non	-0.3825	0.2914	0.4920	3.3367	0.1028	0.5391	-0.0104	0.0126	-0.0128	0.0461
	Tot.	-	2.2149	-	25.3586	-	4.0970	-	0.0955	-	0.3502

Figure 3 - Tableau des moyennes conditionnelles

Prenons le cas de la variable TYPE4X4 sur le second axe. Le carré du rapport de corrélation est $COORD_2(TYPE4X4) = 0.636509$ (Figure 1). Sa contribution est donc de $CONTRIB_2(TYPE4X4) = 0.636509 / 2.510036 * 100 = 25.3586$ (Figure 3). Voyons maintenant la modalité OUI. Elle correspond à 5 observations. Sa contribution est obtenue par $CONTRIB_2(TYPE4X4 = OUI) = [5 * (-3.2472 - 0)^2] / [38 * 2.510036^2] * 100 = 22.0219$ (Figure 3).

Complétons la lecture du 2nd axe maintenant. Nous avons vu qu'il était déterminé par une opposition entre la vitesse et la hauteur. Cette seconde caractéristique est surtout le fait des véhicules de type 4X4 (oui), roulant plutôt au gazole (CARBURANT = diesel) et plutôt d'origine «autres». Il s'agit de tout-terrains asiatiques en fait si l'on revient aux données.

Modele	puissance	cylindres	vitesse	longueur	largeur	hauteur	poids	CO2	prix	origine	carburant	type4X4
SANTA_FE	125	1991	172	450	185	173	1757	197	27990	Autres	Diesel	oui
MURANO	234	3498	200	477	188	171	1870	295	44000	Autres	Essence	oui
LANDCRUI	204	4164	170	489	194	185	2495	292	67100	Autres	Diesel	oui
OUTLAND	202	1997	220	455	178	167	1595	237	29990	Autres	Diesel	oui
X-TRAIL	136	2184	180	446	177	168	1520	190	29700	Autres	Diesel	oui

En croisant les variables, chez les 'type 4x4 = oui' (5 véhicules), nous notons une surreprésentation des diesel (4/5) et de l'origine 'Autres' (4/5).

Nombre de type4X4		Étiquettes c		Total général
Étiquettes de lignes		non	oui	
Diesel		39.39%	80.00%	44.74%
Essence		60.61%	20.00%	55.26%
Total général		100.00%	100.00%	100.00%

Nombre de type4X4		Étiquettes c		Total général
Étiquettes de lignes		non	oui	
Autres		15.15%	100.00%	26.32%
Europe		45.45%	0.00%	39.47%
France		39.39%	0.00%	34.21%
Total général		100.00%	100.00%	100.00%

De même, ces véhicules ont tendance à être plus hautes et moins rapides que les autres.

TYPE 4X4			
Étiquette			
non	oui	Total général	
Moyenne de hauteur	148.6	172.8	151.8

TYPE 4X4			
Étiquette			
non	oui	Total général	
Moyenne de vitesse	202.8	188.4	200.9

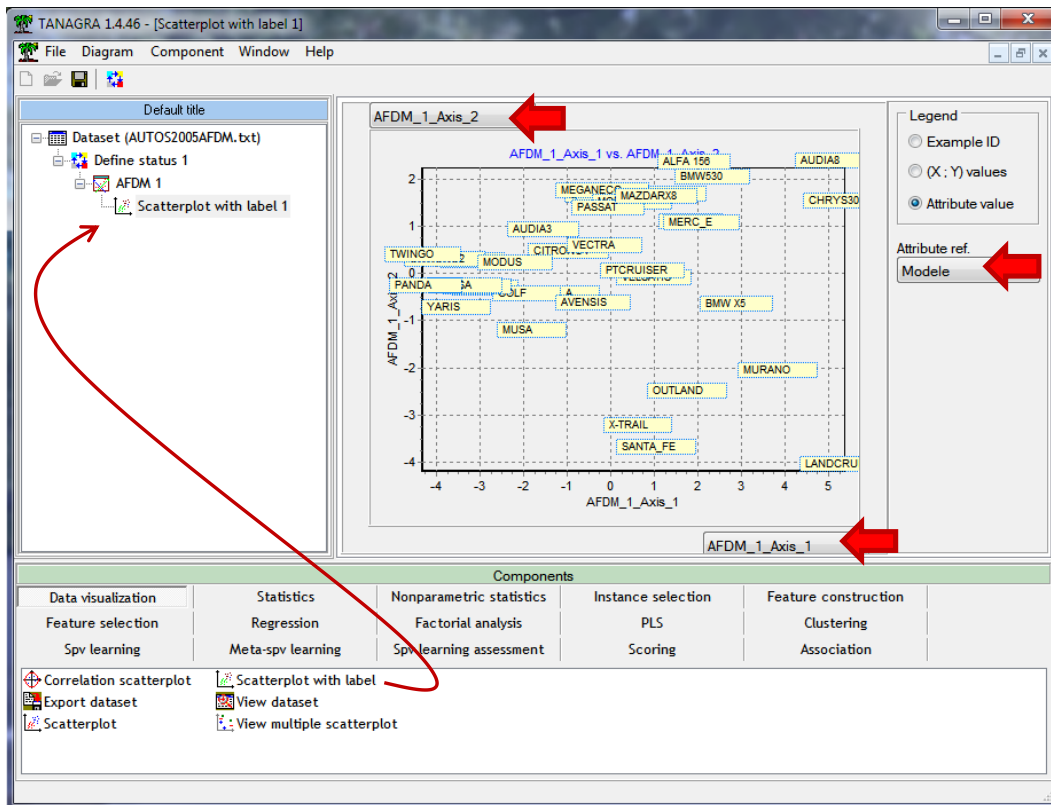
3.3.5 Tableau des vecteurs propres

Le tableau des vecteurs propres est utilisé pour le déploiement c.-à-d. lors de la projection d'un nouvel individu dans le repère factoriel. Il faut centrer ('center') et réduire ('scale') les valeurs prises sur chaque descripteur, puis appliquer les coefficients associés aux axes. L'objectif est de positionner un nouveau venu par rapport aux véhicules existants. Il sera ainsi aisé de déduire ses caractéristiques à partir des observations situées dans son voisinage.

3.4 Représentations graphiques

3.4.1 Projection des individus

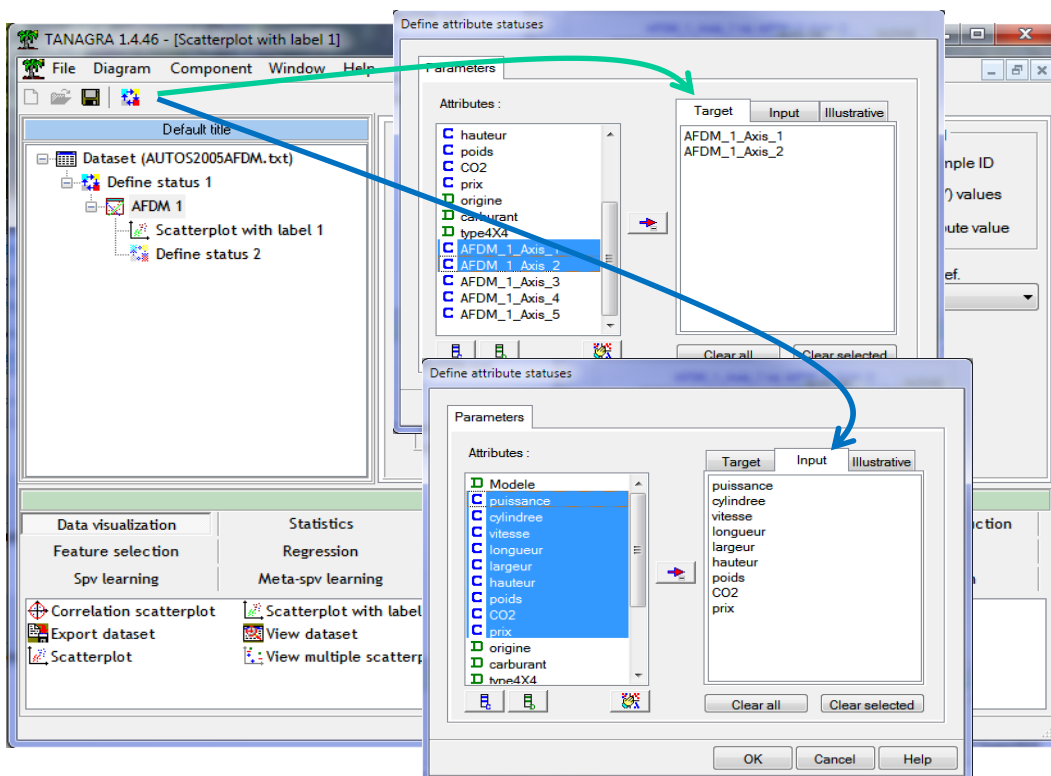
Parlons-en justement des représentations graphiques. Lorsque les observations sont étiquetées, le positionnement des individus dans le repère factoriel est particulièrement intéressant. Nous ajoutons le composant SCATTERPLOT WITH LABEL (onglet DATA VISUALISATION) dans le diagramme. Nous paramétrons le premier axe en abscisse, le second en ordonnée.



Nous illustrons mieux ainsi l'opposition « petit » et « gros » véhicules sur le premier axe, et le positionnement des 4X4 asiatiques sur le second.

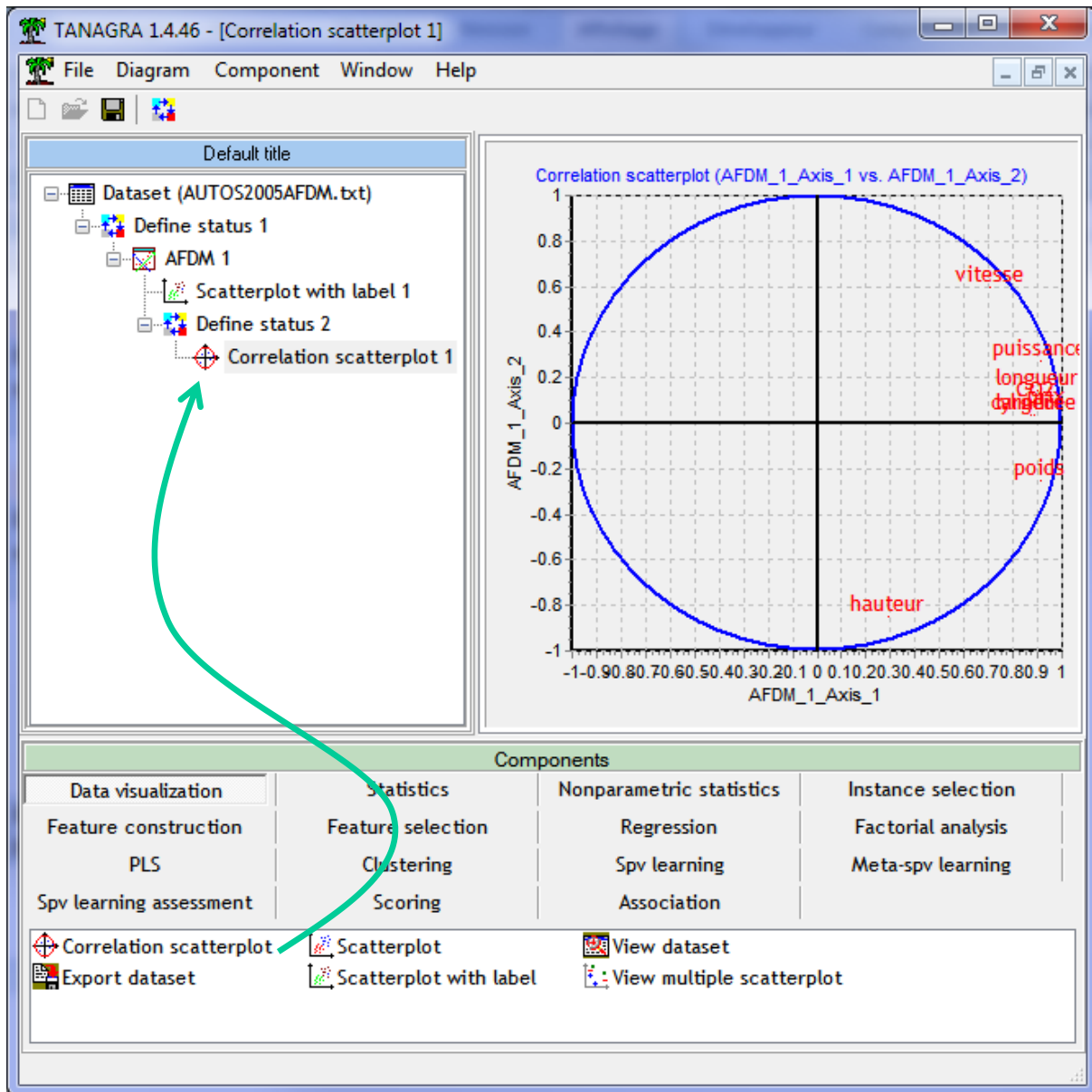
3.4.2 Cercle des corrélations

Par rapport au tableau des corrélations (Figure 2), l'outil « Cercle des corrélations » permet l'introduction des variables illustratives.



Nous ajoutons le composant DEFINE STATUS à la suite de 'AFDM 1'. Nous plaçons en TARGET les deux premiers facteurs, en INPUT les variables quantitatives actives. **Si besoin était, nous aurions pu y placer également des variables illustratives.**

Il nous reste à insérer le composant CORRELATION SCATTERPLOT (onglet DATA VISUALIZATION).



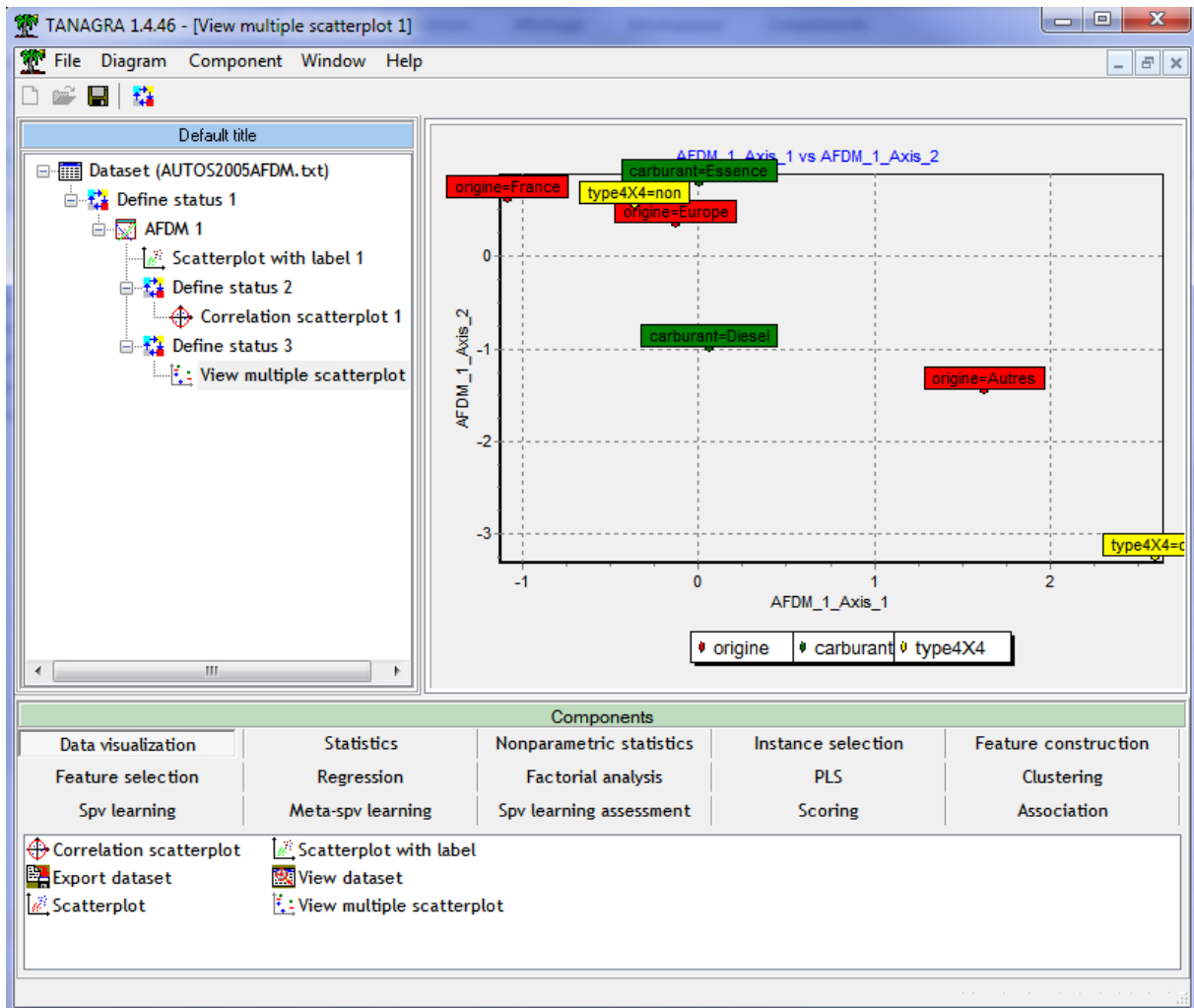
Puisque nous nous cantonnons aux variables actives, ce graphique n'apporte rien de plus par rapport au tableau des corrélations (Figure 2).

3.4.3 Moyennes conditionnelles

De la même manière, par rapport au tableau des moyennes conditionnelles (Figure 3), l'outil graphique de Tanagra autorise l'introduction des variables illustratives. L'interprétation des axes peut en être renforcée.

De nouveau nous insérons le composant DEFINE STATUS dans le diagramme, nous plaçons en TARGET les deux premiers axes, en INPUT les variables qualitatives. Ici également, nous avons la possibilité d'introduire des variables n'ayant pas participé à la construction des axes.

Nous ajoutons alors le composant VIEW MULTIPLE SCATTERPLOT (onglet DATA VISUALIZATION). Nous observons le positionnement des différentes modalités des variables de l'étude.



On constate le rôle particulièrement pesant de la modalité 'TYPE4X4 = OUI' dans notre analyse. On peut se demander légitimement s'il n'est pas opportun de passer cette variable en illustrative, ou encore d'exclure de notre fichier les véhicules correspondants. Nous ne le ferons pas en ce qui nous concerne, mais la question mérite d'être posée. Ici commence réellement le travail de l'analyste...

4 AFDM avec R (package FactoMineR)

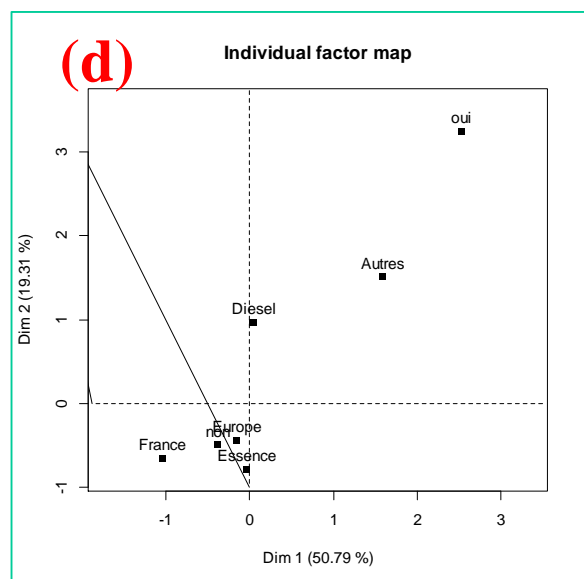
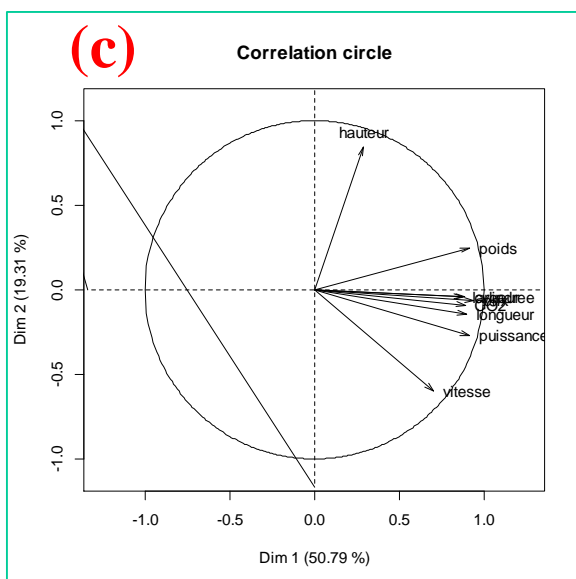
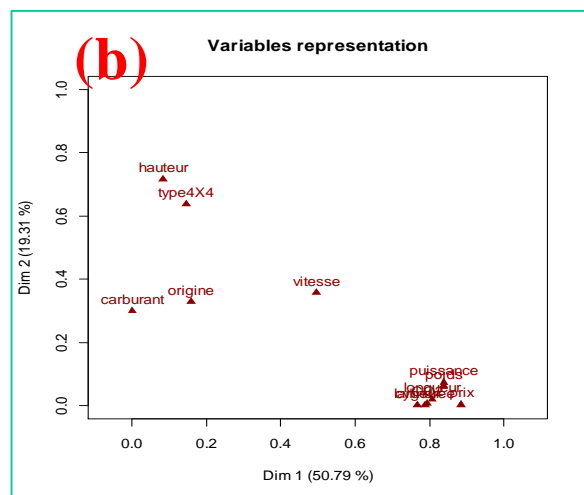
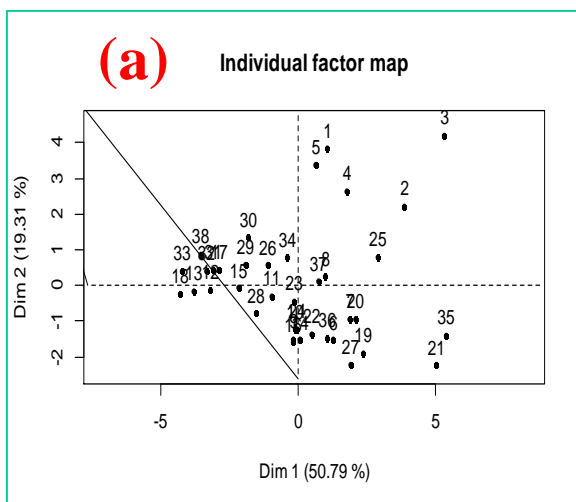
La méthode AFDM est disponible dans le package FactoMineR⁴. Sa mise en œuvre est très simple. L'accès aux résultats est un peu plus compliqué en revanche. Nous décrivons succinctement dans cette section les principales sorties de l'outil.

⁴ <http://cran.r-project.org/web/packages/FactoMineR/index.html>

Le programme ci-dessous effectue les tâches suivantes : charger le fichier texte, en désignant la première colonne comme étiquette des observations ; charger la librairie 'FactoMineR'⁵ ; lancer l'AFDM en lui passant l'ensemble de données et en demandant la création de 5 facteurs.

```
rm(list=ls())
#loading the database
autos.data <- read.table(file="AUTOS2005AFDM.txt", row.names=1, header=T, sep="\t")
print(summary(autos.data))
#performing the AFDM
library(FactoMineR)
afdm <- AFDM(autos.data, ncp=5)
```

R génère automatiquement une série de cartes représentant les individus et les variables dans le premier plan factoriel : (a) projection des individus, (b) coordonnées des variables, (c) cercle des corrélations pour les variables quantitatives, (d) moyennes conditionnelles pour les qualitatives.



⁵ Vous devez l'installer au préalable si le package n'est pas présent sur votre machine. Cf. « [Installation et gestion des packages sous R](#) ».

Par rapport à Tanagra, nous retrouvons bien les mêmes résultats. Certes, le second axe est inversé. Mais les positions relatives des objets sont les mêmes. C'est ce qui importe en analyse factorielle.

La fonction **print()** nous informe que les tableaux de résultats sont intégrés dans des champs de l'objet généré par la procédure AFDM.

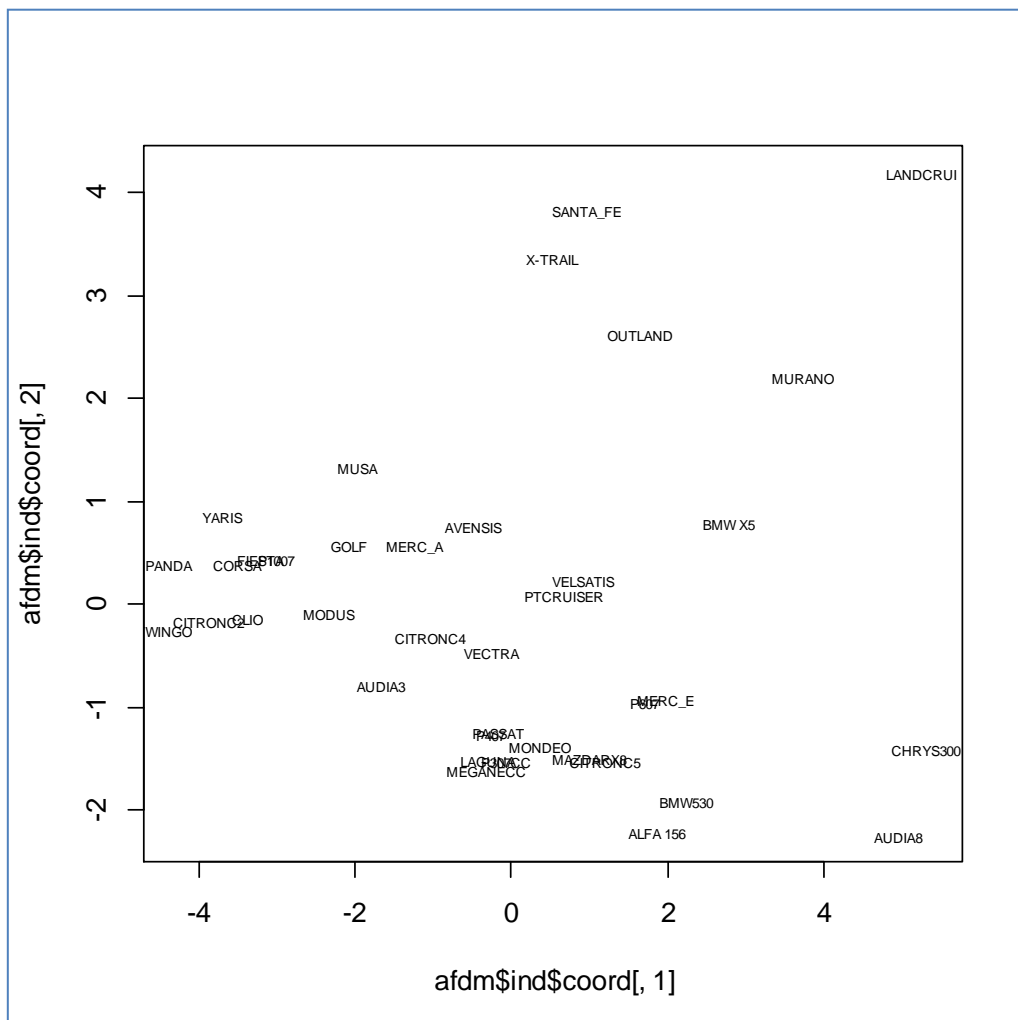
```
R Console
> print(afdm)
*The results are available in the following objects:

  name      description
1 "$eig"    "eigenvalues and inertia"
2 "$var"    "Results for the variables"
3 "$ind"    "results for the individuals"
4 "$quali.var" "Results for the qualitative variables"
5 "$quanti.var" "Results for the quantitative variables"
>
```

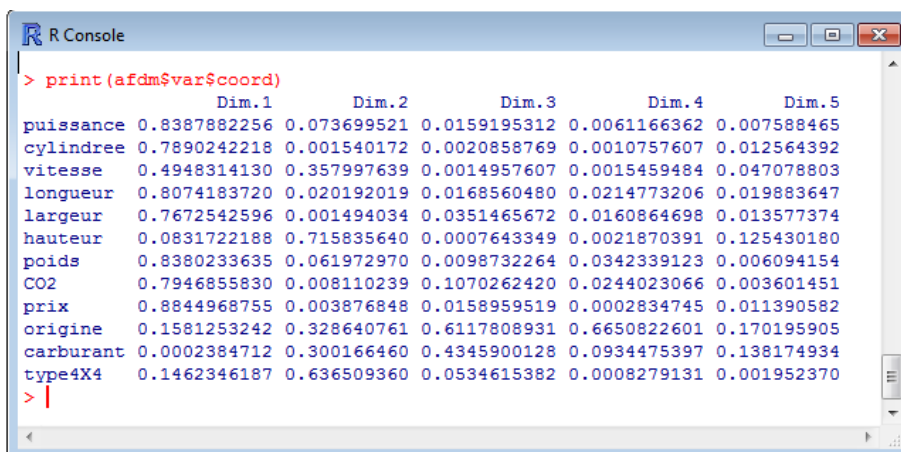
Pour projeter les individus avec leur étiquette dans le premier plan factoriel, nous ferions :

```
plot(afdm$ind$coord[,1],afdm$ind$coord[,2],type="n")
text(afdm$ind$coord[,1],afdm$ind$coord[,2],labels=rownames(autos.data),cex=0.5)
```

Au second axe près (inversé), nous retrouvons le graphique de Tanagra (section 3.4.1).



Pour obtenir, le tableau des coordonnées des variables (cf. Figure 1), nous utilisons l'instruction :



```
> print(afdm$var$coord)
      Dim.1      Dim.2      Dim.3      Dim.4      Dim.5
puissance 0.8387882256 0.073699521 0.0159195312 0.0061166362 0.007588465
cylindree 0.7890242218 0.001540172 0.0020858769 0.0010757607 0.012564392
vitesse   0.4948314130 0.357997639 0.0014957607 0.0015459484 0.047078803
longueur  0.8074183720 0.020192019 0.0168560480 0.0214773206 0.019883647
largeur   0.7672542596 0.001494034 0.0351465672 0.0160864698 0.013577374
hauteur   0.0831722188 0.715835640 0.0007643349 0.0021870391 0.125430180
poids     0.8380233635 0.061972970 0.0098732264 0.0342339123 0.006094154
CO2       0.7946855830 0.008110239 0.1070262420 0.0244023066 0.003601451
prix      0.8844968755 0.003876848 0.0158959519 0.0002834745 0.011390582
origine   0.1581253242 0.328640761 0.6117808931 0.6650822601 0.170195905
carburant 0.0002384712 0.300166460 0.4345900128 0.0934475397 0.138174934
type4X4   0.1462346187 0.636509360 0.0534615382 0.0008279131 0.001952370
```

Il est ainsi possible d'analyser en détail les résultats de l'AFDM.

5 Conclusion

L'analyse factorielle des données mixtes (AFDM) est très peu présente dans les ouvrages, pourtant pléthoriques en Français, qui traitent de l'analyse de données. Quasiment tous s'en tiennent au sempiternels ACP, AFC et ACM. C'est étonnant parce que l'AFDM répond à un vrai besoin. Elle permet de résoudre une nouvelle classe de problème que l'on ne peut pas traiter *directement* avec les techniques usuelles.

Je l'ai découverte un peu par hasard suite à mes pérégrinations sur le web (Merci le projet **NUMDAM** - <http://www.numdam.org/>). Je me suis rendu compte qu'elle était facile à comprendre, que la lecture des résultats n'induisait pas de difficultés insurmontables, et qu'elle était facile à implémenter. J'ai donc décidé de l'intégrer dans Tanagra. Par la suite, j'ai multiplié les études sur d'autres fichiers pour cerner l'intérêt de l'approche. A chaque fois, la faculté de mixer les variables quantitatives et qualitatives permettait d'enrichir les analyses.

Enfin, l'aptitude à appliquer une technique factorielle sur des données mixtes élargit le champ d'action des autres méthodes de data mining. Il devient ainsi possible de procéder à une classification sur un tableau de données mélangeant les variables qualitatives et quantitatives. La démarche serait la suivante : on procède à une AFDM sur l'ensemble des variables, puis on réalise une classification automatique (une CAH ou un arbre de classification) sur les axes factoriels pertinents. En laissant de côté les derniers facteurs non significatifs, on procède à une forme de nettoyage des données en écartant le « bruit » attribuable aux fluctuations d'échantillonnage.