1 Topic

Multiple Factor Analysis for Mixed Data with Tanagra and R (FactoMineR package).

Usually, as a factor analysis approach, we use the principal component analysis (PCA) when the active variables are quantitative; the multiple correspondence analysis (MCA) when they are all categorical. But what to do when we have a mix of these two types of variables?[1]

A possible strategy is to discretize the quantitative variables and use the MCA. But this procedure is not recommended if we have a small dataset (a few number of instances), or if the number of qualitative variables is low in comparison with the number of quantitative ones. In addition, the discretization implies a loss of information. The choice of the number of intervals and the calculation of the cut points are not obvious.

Another possible strategy is to replace each qualitative variable by a set of dummy variables (a 0/1 indicator for each category of the variable to recode). Then we use the PCA. This strategy has a drawback. Indeed, because the dispersions of the variables (the quantitative variables and the indicator variables) are not comparable, we will obtain biased results.

The Jérôme Pages' "Multiple Factor Analysis for Mixed Data" (2004) [AFDM in French] relies on this second idea. But it introduces an additional refinement. It uses dummy variables, but instead of the 0/1, it uses the 0/x values, where 'x' is computed from the frequency of the concerned category of the qualitative variable. We can therefore use a standard program for PCA to lead the analysis (Pages, 2004; page 102). The calculation process is thus well controlled. But the interpretation of the results requires a little extra effort since it will be different depending on whether we study the role of a quantitative or qualitative variable.

In this tutorial, we show how to perform an AFDM with **Tanagra 1.4.46** and **R 1.15.1** (FactoMinerR package). We emphasize the reading of the results. We must study simultaneously the influence of quantitative and qualitative variables for the interpretation of the factors.

2 Dataset

We process the AUTOS2005AFDM.TXT data file[2]. We have n = 38 instances, described by 'q = 12' descriptors: puissance (horsepower), cylindrée (engine displacement), vitesse (speed), longueur (length), larger (width), hauteur (height), poids (weight), CO2 et prix (price) are quantitative; origine (origin)(France, Europe, Autres), carburant (fuel-type) (diesel, essence) et type 4x4 (four-by-four)(oui, non) are qualitative. The use of the AFDM seems to prevail because the mixed nature of the variables. We will see in the next sections if we obtain relevant results.

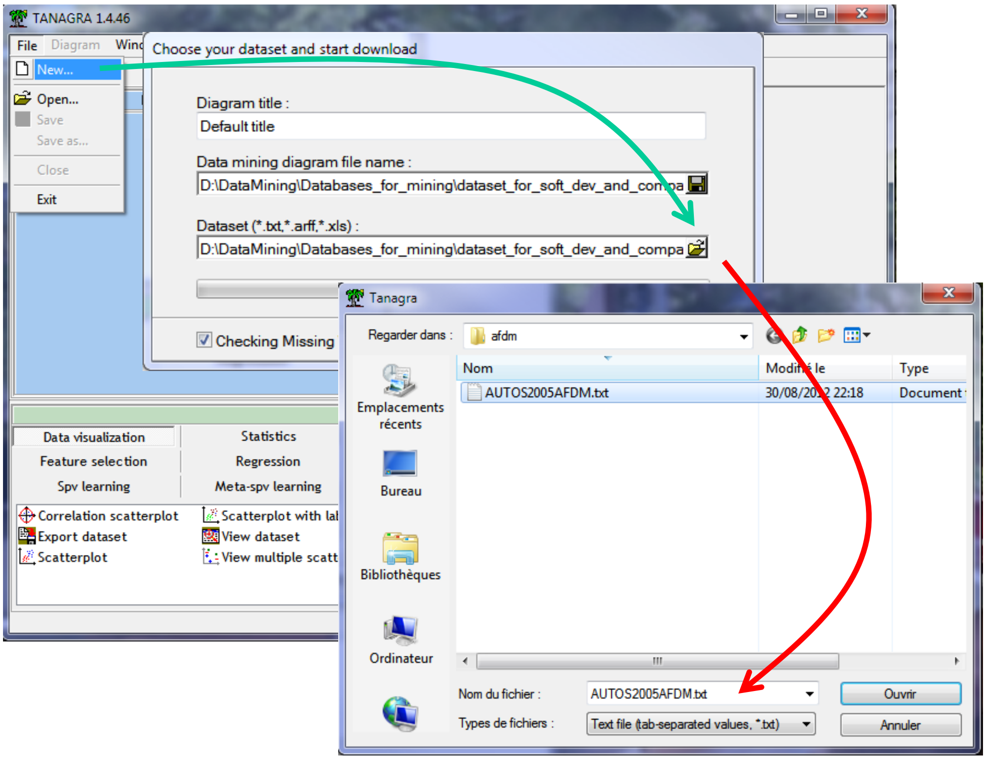
[1] This introductory section is based on the Jérôme Pages' paper « Factor Analysis for Mixed Data » (in French), Revue de Statistique Appliquée, tome 52, n°4, 2004, pages 93-111 ; available online: http://archive.numdam.org/ARCHIVE/RSA/RSA_2004__52_4/RSA_2004__52_4_93_0/RSA_2004__52_4_93_0.pdf

[2] See "STA101 – Analyse des données : méthodes descriptives" (Pierre-Louis Gonzalez). Some variables are removed; two outliers are removed also.

3 AFDM with TANAGRA

3.1 Creating a diagram and importing the data file

After launching Tanagra, we click on the FILE / NEW menu to create a new diagram. We select the AUTOS2005AFDM.TXT file (text file with tab separator). We confirm by clicking OK.



Choose your dataset and start download

Diagram title : Default title

Data mining diagram file name : D:\DataMining\Databases_for_mining\dataset_for_soft_dev_and_compa

Dataset (*.txt,*.arff,*.xls) : D:\DataMining\Databases_for_mining\dataset_for_soft_dev_and_compa

☑ Checking Missing

Regarder dans : afdm

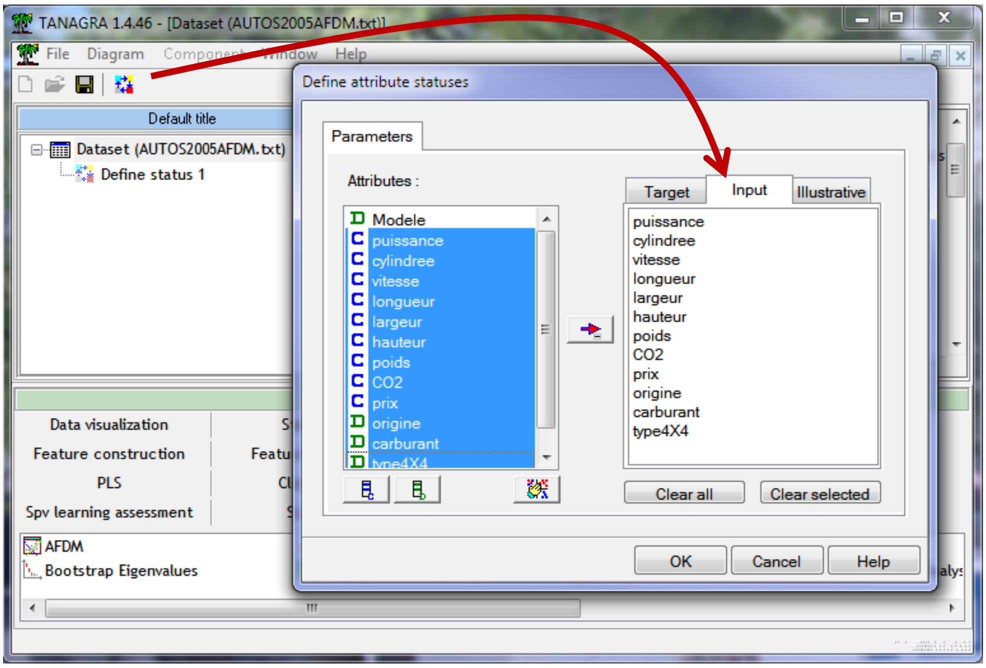
Nom	Modifié le	Type
AUTOS2005AFDM.txt	30/08/2012 22:18	Document

Nom du fichier : AUTOS2005AFDM.txt

Types de fichiers : Text file (tab-separated values, *.txt)

The data is imported. We obtain a description of the variables: model corresponds to the label of observations, POWER ...PRICES are quantitative variables, ORIGIN ... TYPE4X4 are qualitative.

3.2 The AFDM component

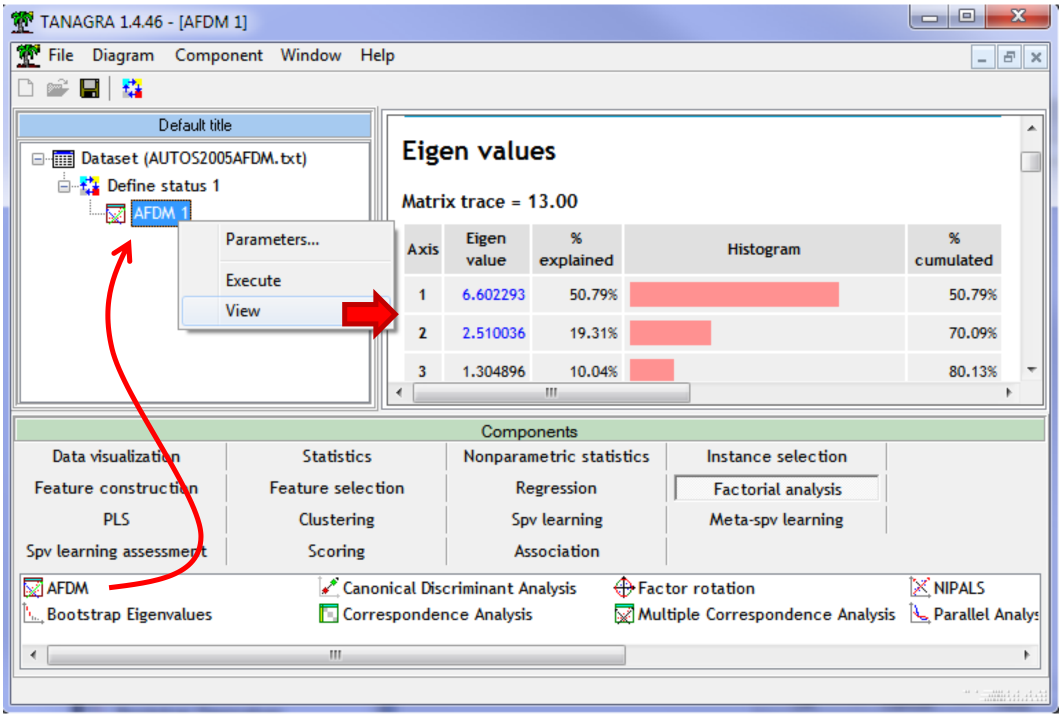


Define attribute statuses

Attributes : D Modele, C puissance, C cylindree, C vitesse, C longueur, C largeur, C hauteur, C poids, C CO2, C prix, D origine, D carburant, D type4X4

Input : puissance, cylindree, vitesse, longueur, largeur, hauteur, poids, CO2, prix, origine, carburant, type4X4

First, we must first specify the role of the variables. We use the DEFINE STATUS component. Except MODEL which is a label to identify each vehicle, we set all variables as INPUT. Then, we add the AFDM component (FACTORIAL ANALYSIS tab). We click on VIEW to launch the calculations and visualize the results.



Eigen values

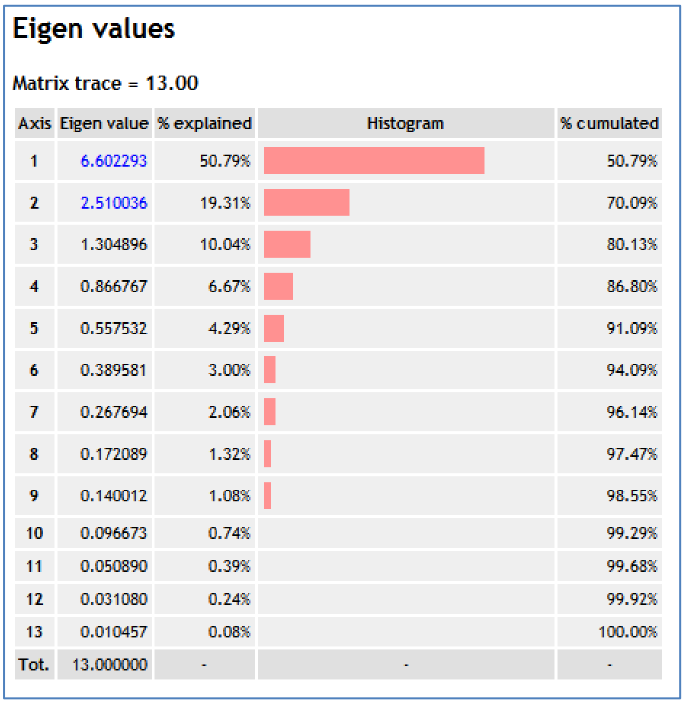
Matrix trace = 13.00

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%

We detail below the various tables produced by the AFDM.

3.3 The reading of the results

3.3.1 The eigenvalues table



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	-	-	-

This table presents the variance explained by each factor. We have 'p=13' factors because there are 9 quantitative variables and 3 qualitative variables with respectively 3, 2 and 2 values. In the internal data table submitted to the PCA program, we have 9 + 3 + 2 + 2 = 16 columns. But we know that the sum of the values of the indicators related to a qualitative variable is a constant. Thus, the number of eigenvalues higher to 0 is 9 + [(3-1)+(2-1)+(2-1)] = 13. The results of the calculations confirm this fact. Indeed, the sum of the eigenvalues is 13. The first 13 factors explain 100% of the total inertia.

The selection of the right number of factor is a difficult problem. Tanagra highlights the first two factors because the corresponding eigenvalues are statistically higher than 1 at the 10% level. The critical value is defined by the following formula (Karlis, Saporta and Spinakis; 2003)[3]:

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

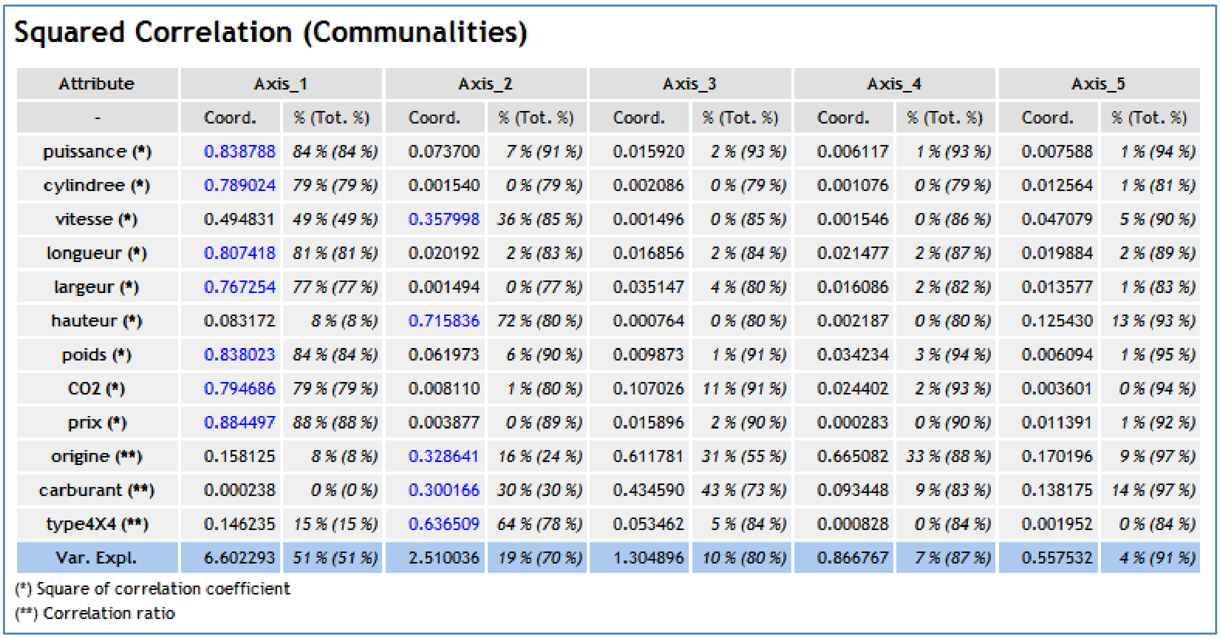
With 'p' is the number of eigenvalues theoretically higher than 0 (p = 13); 'n' is the number of instances (n = 38). For our dataset, we have the following cutoff

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

The first two factors present eigenvalues upper than this critical value (**6.60** and **2.51**). We use them in the interpretation of the results. We observe that they explain 70.09% of the total inertia. This is rather high. Indeed, in contrast to PCA, and like the MCA (multiple correspondence analysis), due to the presence of qualitative variables, the explained variability is often less concentrated on the first factors for the AFDM. This phenomenon is even more pronounced when the proportion of qualitative variables in the database (and the associated number of categories) increases.

This critical value is more restrictive than the usual Kaiser-Guttman rule (cutoff = 1). Its main merit is that it takes into account both 'p' (the maximum number of factors) and 'n' (the sample size). Nevertheless, we must always take cautiously this kind of cutoff value. The analyses of their characteristics help us to determine if the remaining factors must be really discarded.

3.3.2 Variable coordinates



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 – Variable coordinates

[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.

This table describes the influence of variables, whether quantitative or qualitative, in the determination of the factors. For the first case, the values correspond to the square of the correlation coefficient; for the second one, the values correspond to the correlation ratio.

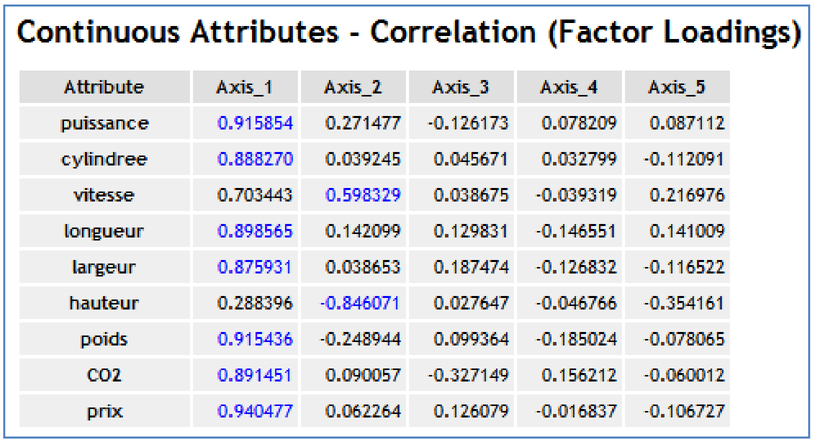
The sum of the values for each row is equal to 1 for the quantitative variables; it is equal to the number of categories minus 1 for the qualitative variable. The percentages are computed according to these totals.

The sum of the values for each column is equal to the eigenvalue associated to the corresponding factor. We can analyze these values like the influence of the variables for the determination of the factor. For instance, we observe that puissance, cylindree, longueur, largeur, poids, C02 and prix are the most important variables for the first factor. But we do not know the nature of the relations between these variables. The other tables below enable to give responses to this question.

Tanagra uses the following rule to highlight the values: (1) the factor must be significant i.e. its eigenvalue is higher than the critical value defined above (section 3.3.1); (2) the row percentage must be higher than '1 / p ' (maximum number of factors); (3) the column percentage must be upper than '1 / q ' (number of variables of the study).

3.3.3 Correlations table

This table specifies the direction of the relationship between the quantitative variables and the factors.



Continuous Attributes - Correlation (Factor Loadings)

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 – Correlations table

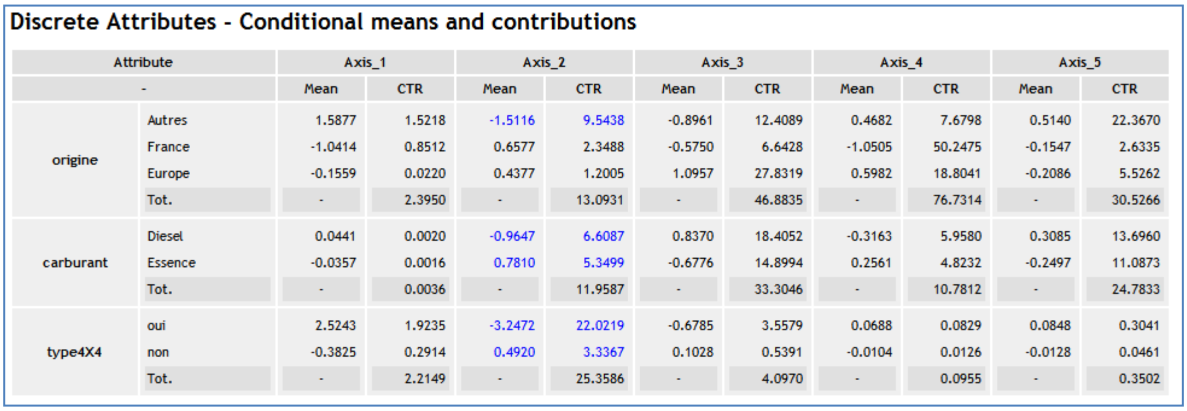
The first factor highlights the positive relations between the length and the width of the cars, their engine size, their price, and their air pollution. For the second factor, we observe a negative relation between their height and their speed. These interpretations will be enhanced by the analysis of the influence of the qualitative variables.

3.3.4 Conditional means table

This table provides the coordinates of categories of the qualitative variables. We have also indications about their contributions for the determination of the factors. The sum of the contributions of the categories of a qualitative variable is equal to the contribution of the variable.

Let us consider the TYPE4X4 attribute on the second factor. The squared of the correlation ratio is $\text{COORD}_2(\text{TYPE4X4}) = 0.636509$ (Figure 1). Its contribution is thus $\text{CONTRIB}_2(\text{TYPE4X4}) = 0.636509$ /

2.510036 * 100 = 25.3586 (Figure 3). The category 'TYPE4X4 = OUI' corresponds to 5 instances. Its contribution is computed as follows: CONTRIB_2(TYPE4X4 = OUI) = [5 * (-3.2472 – **0**)^2] / [38 * 2.510036^2] * 100 = 22.0219 (Figure 3).

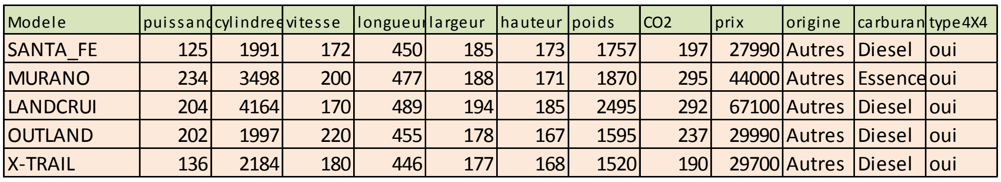


Discrete Attributes - Conditional means and contributions

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

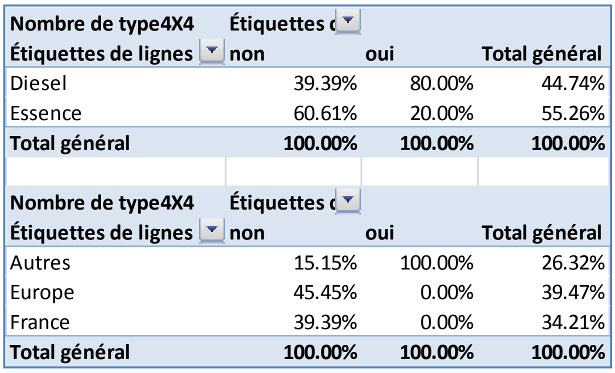
Figure 3 – Conditional means table

We can complete the interpretation of the second factor now. We saw that it was determined by an opposition between speed and height. The second characteristic is especially the result of TYPE4X4 vehicles (yes), which use diesel fuel type (CARBURANT) and with the ORIGIN "other." They correspond to the Asian 4x4 vehicles if we see the data.



Modele	puissanc	cylindree	vitesse	longueu	largeur	hauteur	poids	CO2	prix	origine	carburan	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

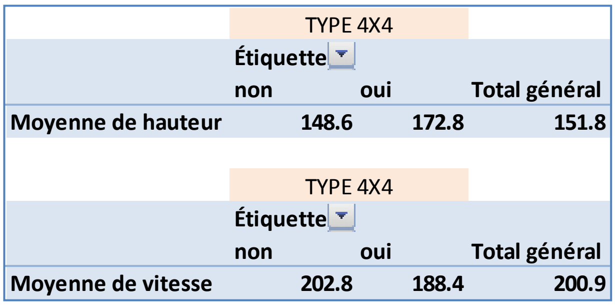
By computing the cross-tabulation between these variables, we observe an over-representation of the fuel-type = diesel (4/5) and the origin = other (4/5) among the 4x4 vehicles (5 cars).



Nombre de type4X4	Étiquettes c		
Étiquettes de lignes	non	oui	Total général
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		
Étiquettes de lignes	non	oui	Total général
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%**

In addition, these vehicles tend to be slower than others, their height is greater.



	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**

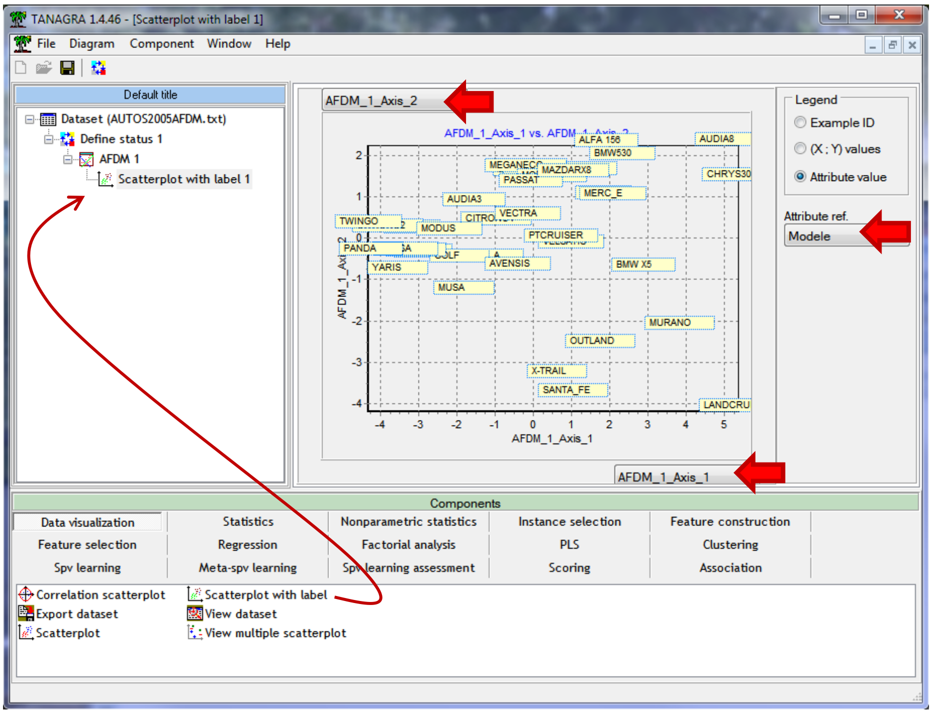
3.3.5 Eigenvectors table

The eigenvectors tables provide the factor scores coefficients. These coefficients enable to deploy the model on a new instance (supplementary instances in the factor analysis terminology) i.e. they enable to compute the coordinates of new instances. We must center and the scale the values of the quantitative variables and the indicators of the qualitative variables.

3.4 Graphical representations

3.4.1 Graphical representation of individuals

When we have labeled instances, the graphical representation is helpful to interpret the factors. We add the SCATTERPLOT WITH LABEL component (DATA VISUALISATION tab) into the diagram. We set the first factor on the abscissa, the second one on the ordinate.

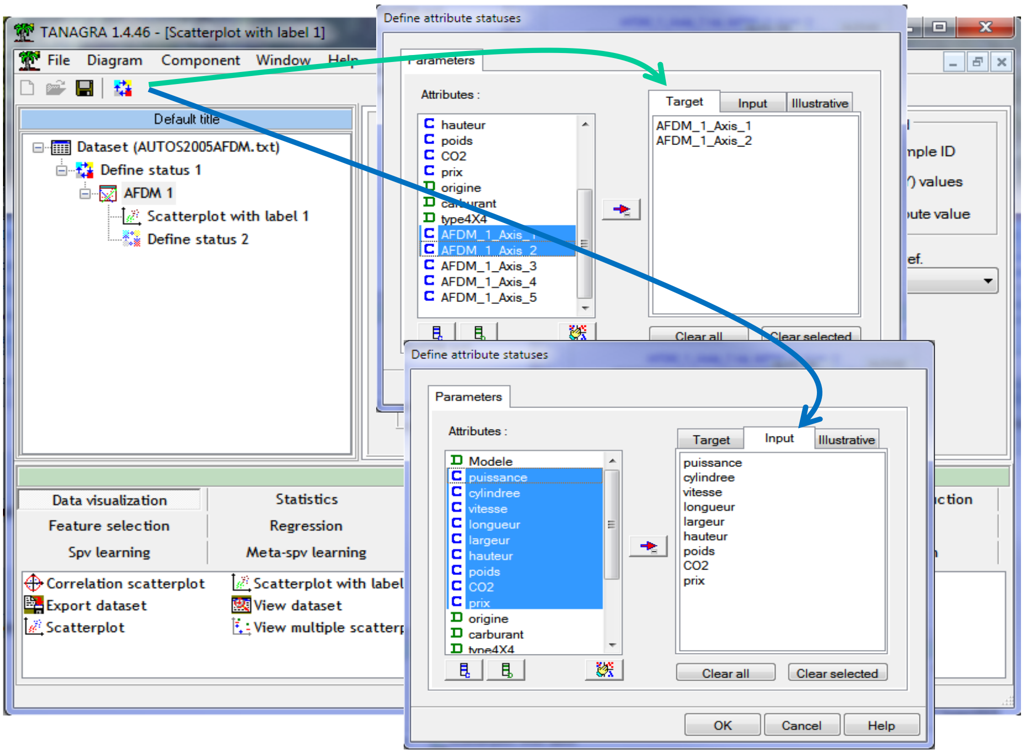


We observe the opposition between big and small cars on the first factor, and the distinctive feature of the Asian all-road cars on the second factor.

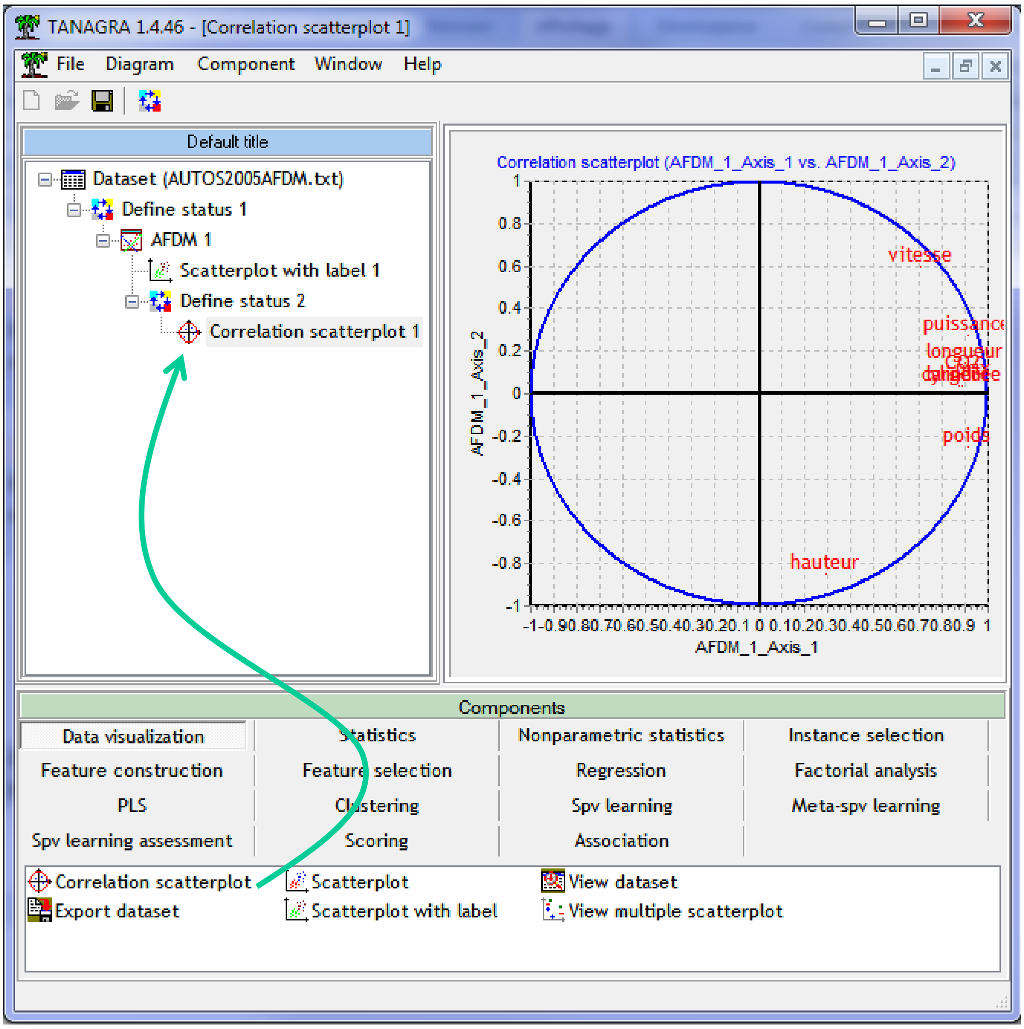
3.4.2 Correlation scatter plot (correlation circle)

This scatter plot enables to represent the correlation of both the active variables (those used during the learning phase) and the supplementary variables (additional variables used to help the interpretation) on the factors.

We insert the DEFINE STATUS component after “AFDM 1” into the diagram. We set the two first factors as TARGET, all the continuous variables as INPUT. Here, we could include also variables which are not used during the calculations of the factors.



Then, we add the CORRELATION SCATTERPLOT tool (DATA VISUALIZATION tab).

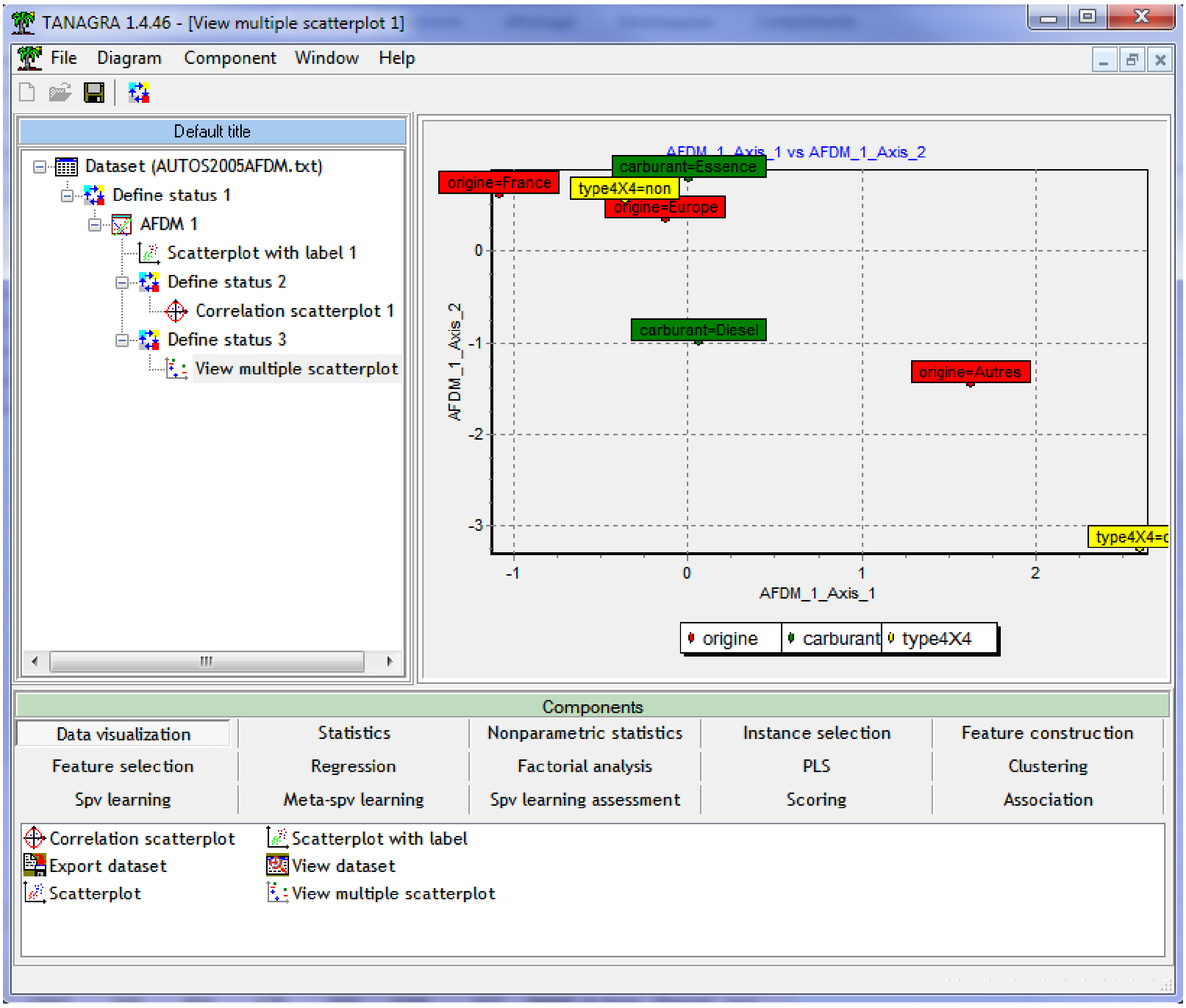


For the active variables (those used for the calculations of the factors), we have the same value as those described in the correlation table (Figure 2).

3.4.3 Conditional means

Tanagra provides a similar graphical tool for the conditional mean table (Figure 3). Here also, we can incorporate supplementary variables into the graphical representation.

We insert the VIEW MULTIPLE SCATTERPLOT tool (DATA VISUALIZATION tab) into the diagram. We observe the coordinates of the categories of the qualitative variables on the factors. Each point corresponds to the mean of the instances associated to a category on the factors.



We note that the category is very influent on the first two factors. Perhaps it is too influent. It would be more appropriate to set the corresponding variable as a supplementary variable.

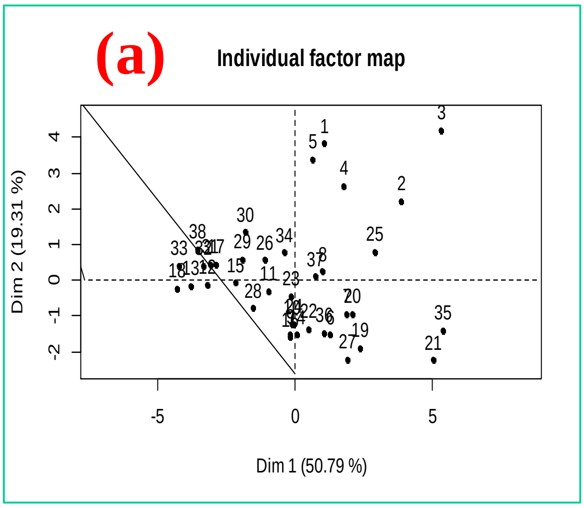
4 AFDM with R (FactoMineR package)

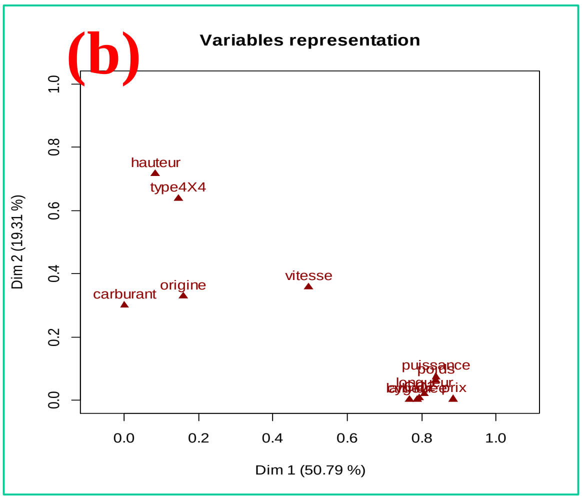
The AFDM approach is available in the FactoMineR[4] package for R. We describe below the main outputs of the tool. The source code below completes the following tasks: loading the data file in the text file format, loading the FactoMineR package, performing the AFDM by asking 5 factors.

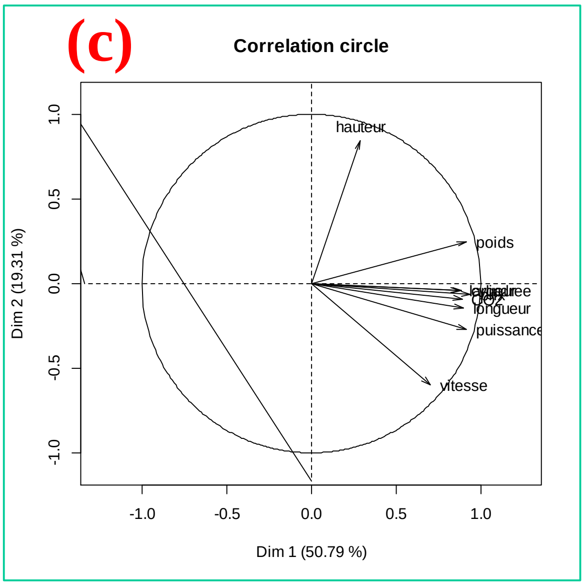
[4] http://cran.r-project.org/web/packages/FactoMineR/index.html

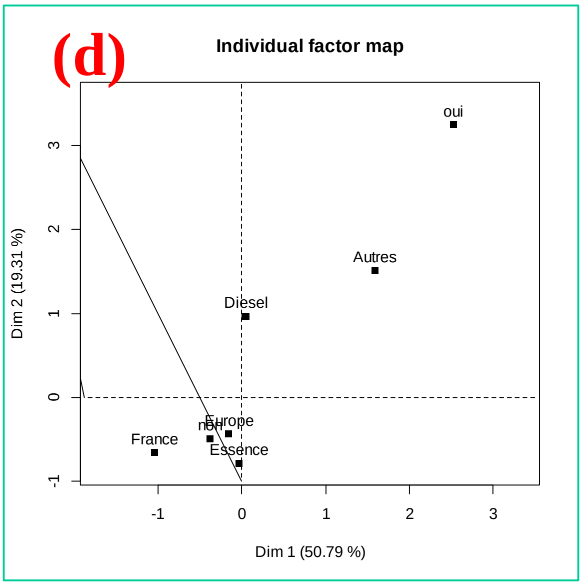
```
rm(list=ls())
#loading the database
autos.data <- read.table(file="AUTOS2005AFDM.txt",row.names=1,header=T,sep="\t")
print(summary(autos.data))
#performing the AFDM
library(FactoMineR)
afdm <- AFDM(autos.data,ncp=5)
```

The command generates several graphical representations: (a) the representations of the instances, (b) the representation of the variables, (c) the correlation circle for the quantitative variables, (d) the conditional means for the qualitative variables.

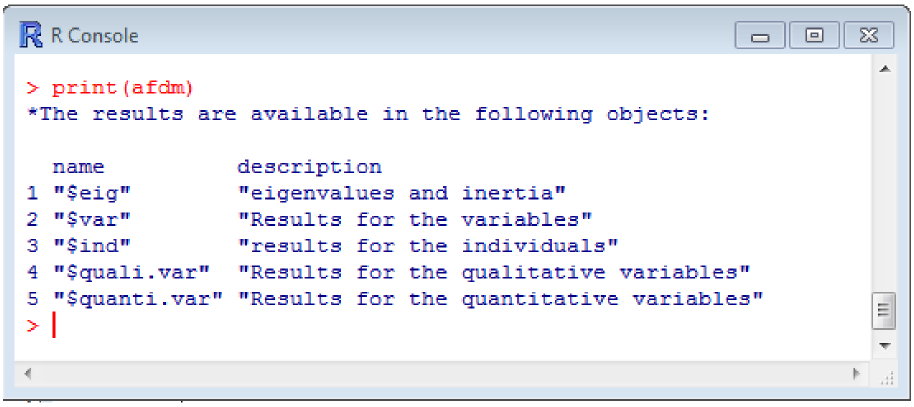








We have the same results as Tanagra of course. The second factor is reversed, but the relative coordinates of individuals (or the variables/categories) are the same. This is the most important in factor analysis. The **print()** command shows the properties related to the object provided by the AFDM procedure.



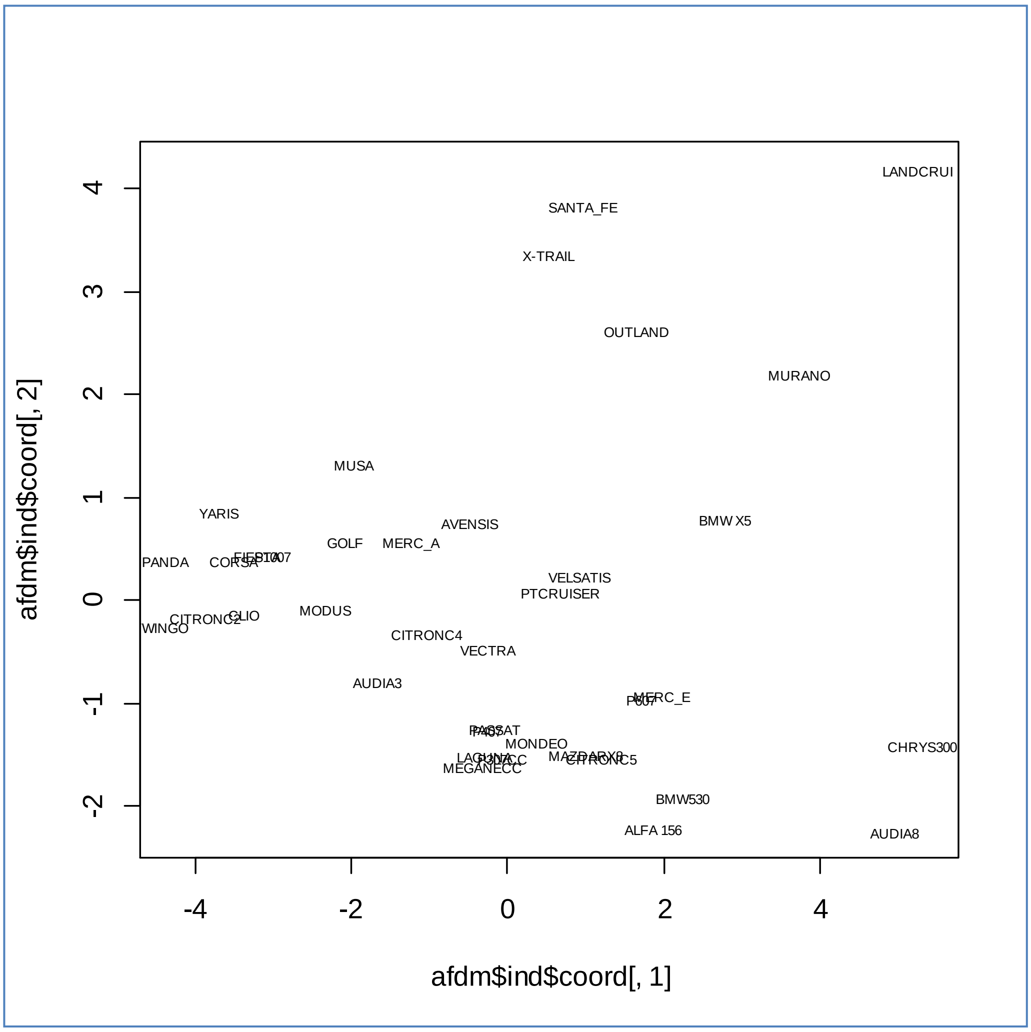
```
> print(afdm)
*The results are available in the following objects:

  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"
> 
```

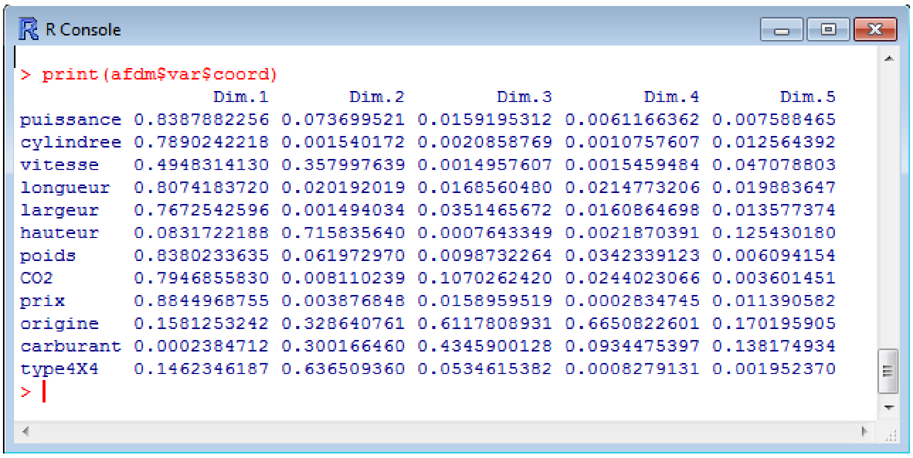
To show the instances with their labels on the first two components, we use the following commands:

```
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)
```

We obtain:



To obtain the coordinates of the variables (see Figure 1), we use the following command:



```
> 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
> 
```

Thus, we can explore in details the results provided by the AFDM procedure.

5 Conclusion

The factorial analysis of mixed data (AFDM) is largely absent in the books describing the exploratory data analysis techniques (in the French books at least). This is surprising because the AFDM responds to a real need. It allows solving a new class of problem that cannot handle directly with usual factor analysis approaches.

In this tutorial, we show that this approach can be performed easily with Tanagar or R. The reading of the results needs a very little extra effort. We must simply adapt the interpretation of the factors, according we consider the influence of the quantitative variables (correlations) or the influence of qualitative variables (conditional means).