1 Topic

Clustering algorithm for mixed data (numeric and categorical attributes), using the latent variables (principal components) from the factor analysis for mixed data.

The aim of cluster analysis is to gather together the instances of a dataset in a set of groups. The instances in the same cluster are similar according a similarity (or dissimilarity) measure. The instances in distinct groups are different. The influence of the used measure, which is often a distance measure, is essential in this process. They are well known when we work on attributes with the same type. The Euclidian distance is often used when we deal with numeric variables; the chi-square distance is more appropriate when we deal with categorical variables. The problem is a lot of more complicated when we deal with a set of mixed data i.e. with both numeric and categorical values. It is admittedly possible to define a measure which handles simultaneously the two kinds of variables, but we have trouble with the weighting problem. We must define a weighting system which balances the influence of the attributes, indeed the results must not depend of the kind of the variables. This is not easy[1].

Previously we have studied the behavior of the factor analysis for mixed data (AFDM in French). This is a generalization of the principal component analysis which can handle both numeric and categorical variables[2]. We can calculate, from a set of mixed variables, components which summarize the information available in the dataset. These components are a new set of numeric attributes. We can use them to perform the clustering analysis based on standard approaches for numeric values.

In this paper, we present a tandem analysis approach for the clustering of mixed data. First, we perform a factor analysis from the original set of variables, both numeric and categorical. Second, we launch the clustering algorithm on the most relevant factor scores. The main advantage is that we can use any type of clustering algorithm for numeric variables in the second phase. We expect also that by selecting a few number of components, we use the relevant information from the dataset, the results are more reliable[3].

We use Tanagra 1.4.49 and R (ade4 package) in this case study.

2 Dataset

The "bank_customer.xls" data file describes the customers of a bank. The variables correspond to their characteristics: age, seniority, etc. SCORE is a supplementary variable. It depicts a score assigned to each customer by the bank advisor. The challenge is to produce a grouping of the customers from their characteristics, and then to comment the obtained categories using the SCORE variable.

Here are the first 5 lines of the file.

[1] Z. Huang, « Clustering large datasets with mixed numeric and categorical values », in Proc. of the First PAKDD, 1997.

[2] « Factor Analysis for Mixed Data », http://data-mining-tutorials.blogspot.fr/2013/03/factor-analysis-for-mixed-data.html

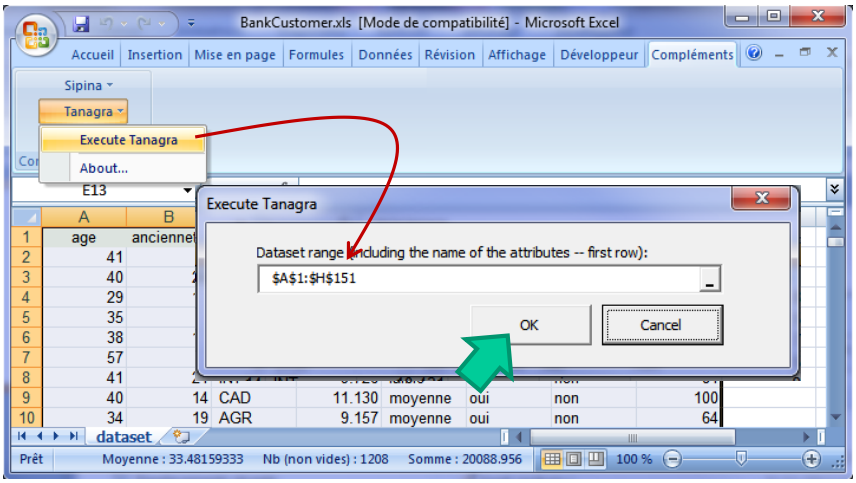
[3] It seems that in some circumstances [see Arabie, P., Hubert, L., 1994. Cluster analysis in marketing research. In: Bagozzi, R.P. (Ed.), Handbook of marketing research. Blackwell, Oxford.], that we cannot detect a priori, a wrong selection of the components can hide the clusters. The graphical representation of the dataset is important to assist the user for this kind of analysis.

age	anciennete	profession	revenu	epargne	carte_bleue	pea	score
41	6	CAD	10.870	moyenne	oui	non	84
40	22	INT	10.035	moyenne	oui	non	51
29	12	OUV	9.087	moyenne	oui	oui	77
35	6	CAD	11.180	moyenne	oui	non	55
38	14	INT	10.431	moyenne	oui	non	87

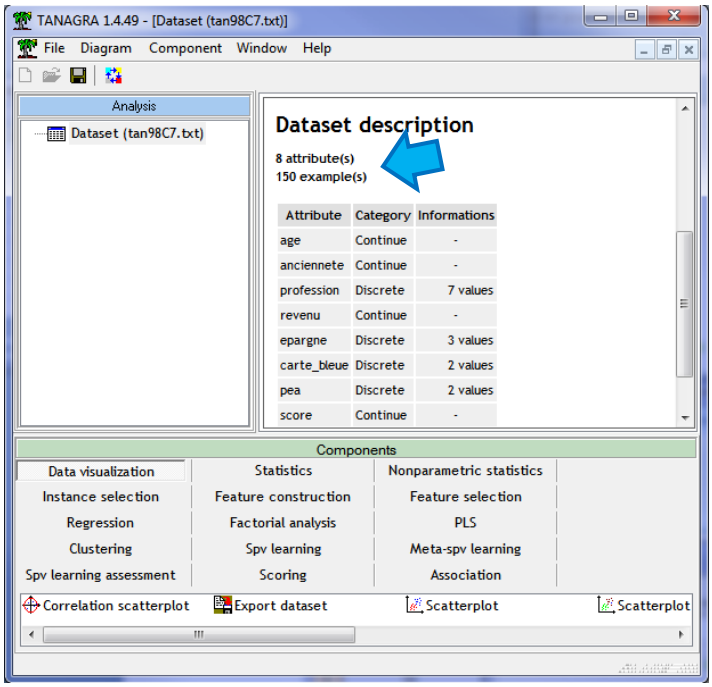
3 Clustering from mixed data with Tanagra

3.1 Importing the dataset

To import "bank_customer.xls", we use the add-in "**tanagra.xla**" which sends the dataset from the Excel spreadsheet to Tanagra[4]. A dialog box enables to check the data range (**$A$1:$H$151**). We confirm by clicking on the OK button.



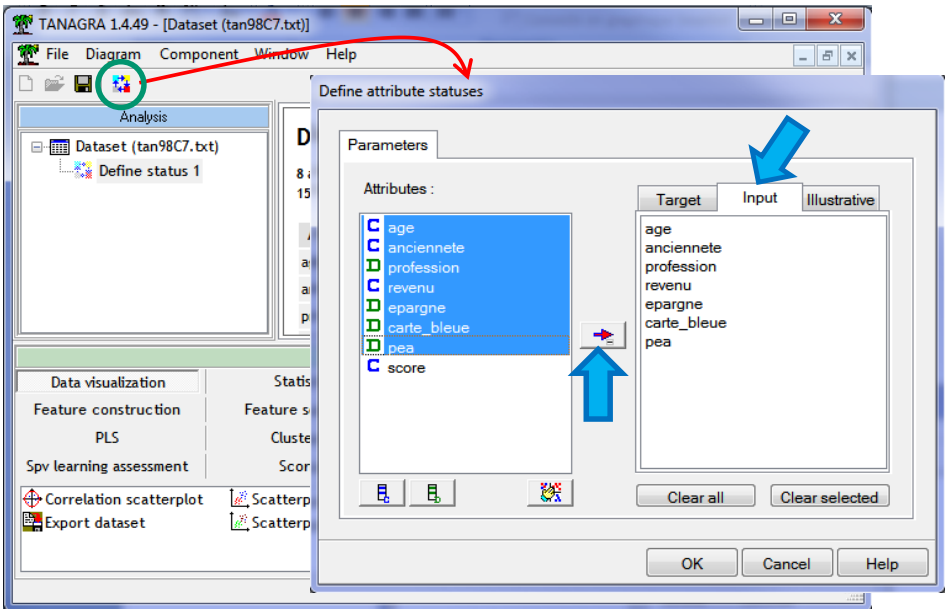
Tanagra is launched. We check that we have **150 instances** and **8 variables**.



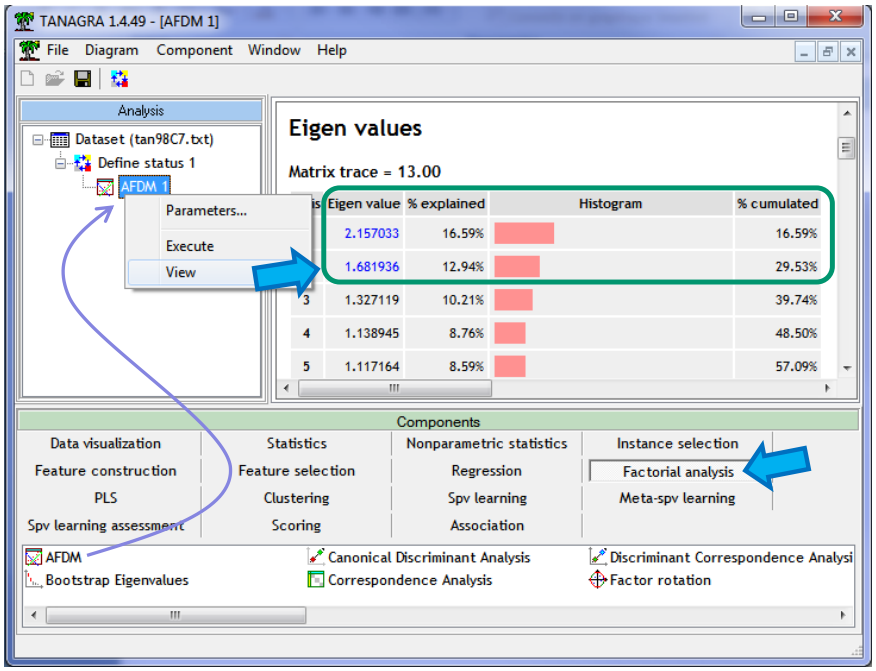
[4] http://data-mining-tutorials.blogspot.fr/2010/08/tanagra-add-in-for-office-2007-and.html; we can use a specific add-on for Open Office and Libre Office.

3.2 Factor Analysis for Mixed Data (AFDM in French)

We select the active variables for the analysis using the DEFINE STATUS tool. We set as INPUT the 7 first variables: "age",..., "pea".



We insert the **AFDM** component (FACTORIAL ANALYSIS tab) into the diagram. We click on the VIEW menu to obtain the results.



The choice of factors to retain is always difficult in factor analysis. This is all the more that we want to use them in subsequent calculations. The quality of the clustering algorithm depends on the number of the components we select.

We select the 2 first components for our dataset. We use only 29.53% of the total variance here. This may seem small. But the aim of the analysis is not to produce an exhaustive view of the data. We want to highlight underlying groups that we can interpret.

To study the results, we show below the "communalities" table that corresponds to the square of the correlation of variables with the factors when they are numeric, to the square of the correlation ratio when they are categorical.

Squared Correlation (Communalities)

Attribute	Axis_1			Axis_2		
-	Coord.	CTR (%)	QLT % (Tot. %)	Coord.	CTR (%)	QLT % (Tot. %)
age (*)	0.031500	1.5 %	3 % (3 %)	0.683161	40.6 %	68 % (71 %)
anciennete (*)	0.064625	3.0 %	6 % (6 %)	0.598183	35.6 %	60 % (66 %)
profession (**)	0.879093	40.8 %	15 % (15 %)	0.301008	17.9 %	5 % (20 %)
revenu (*)	0.922902	42.8 %	92 % (92 %)	0.000083	0.0 %	0 % (92 %)
epargne (**)	0.257906	12.0 %	13 % (13 %)	0.016595	1.0 %	1 % (14 %)
carte_bleue (**)	0.000250	0.0 %	0 % (0 %)	0.024199	1.4 %	2 % (2 %)
pea (**)	0.000757	0.0 %	0 % (0 %)	0.058707	3.5 %	6 % (6 %)
Var. Expl.	2.157033	-	17 % (17 %)	1.681936	-	13 % (30 %)

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

The « Factor loadings » table shows the correlation between the numeric variables and the factors.

Continuous Attributes - Correlation (Factor Loadings)

Attribute	Axis_1	Axis_2	Axis_3	Axis_4	Axis_5
age	0.177481	0.826536	0.031065	0.102036	-0.063550
anciennete	-0.254215	0.773423	-0.021765	0.290430	0.156407
revenu	0.960678	-0.009110	0.143444	0.018713	0.136343

Figure 1 - Correlation between the numeric variables and the components

The « Conditional Means » table shows the conditional means (the mean for each level) for the categorical variables.

Discrete Attributes - Conditional means and contributions

Attribute		Axis_1			Axis_2		
-		Mean	CTR (%)	v.test	Mean	CTR (%)	v.test
profession	CAD	2.2483	28.97	11.268	-0.0056	0.00	-0.032
	INT	-0.9328	3.62	-3.796	-0.0318	0.01	-0.147
	OUV	-1.1116	3.19	-3.412	0.7798	2.58	2.710
	INA	-0.4438	0.37	-1.136	1.4766	6.68	4.281
	AGR	-1.4571	2.13	-2.679	-1.2307	2.50	-2.563
	EMP	-0.7088	1.94	-2.760	-0.9724	6.02	-4.288
	ART	-0.4852	0.54	-1.394	0.1742	0.11	0.566
	Tot.	-	40.75	-	-	17.90	-
epargne	moyenne	-0.1718	0.41	-1.960	0.0798	0.15	1.031
	faible	-1.5638	5.26	-4.332	-0.4869	0.84	-1.528
	elevee	1.0889	6.29	5.179	-0.0139	0.00	-0.075
	Tot.	-	11.96	-	-	0.99	-
carte_bleue	oui	0.0055	0.00	0.193	-0.0479	0.08	-1.899
	non	-0.0979	0.01	-0.193	0.8500	1.36	1.899
	Tot.	-	0.01	-	-	1.44	-
pea	non	-0.0321	0.01	-0.336	-0.2495	1.35	-2.958
	oui	0.0509	0.02	0.336	0.3958	2.14	2.958
	Tot.	-	0.04	-	-	3.49	-

Figure 2 – Conditional means for categorical variables

The « Factor Scores » table provides the factor scores coefficients. They enable to calculate the factor scores of new instances from their characteristics. These are precisely these values that we will use in the clustering process.

Eigen vectors - Factor Scores

Attribute	Center	Scale	Axis_1	Axis_2
age	40.553333	9.243763	0.120844	0.637319
anciennete	13.286667	6.735317	-0.173090	0.596365
profession = CAD	0.266667	0.516398	0.538255	-0.001733
profession = INT	0.193333	0.439697	-0.190155	-0.008319
profession = OUV	0.120000	0.346410	-0.178526	0.160603
profession = INA	0.086667	0.294392	-0.060575	0.258460
profession = AGR	0.046667	0.216025	-0.145932	-0.158070
profession = EMP	0.180000	0.424264	-0.139412	-0.245292
profession = ART	0.106667	0.326599	-0.073468	0.033820
revenu	9.886373	0.912768	0.654108	-0.007024
epargne = moyenne	0.653333	0.808290	-0.064361	0.038330
epargne = faible	0.100000	0.316228	-0.229263	-0.091545
epargne = elevee	0.246667	0.496655	0.250721	-0.004093
carte_bleue = oui	0.946667	0.972968	0.002488	-0.027701
carte_bleue = non	0.053333	0.230940	-0.010482	0.116706
pea = non	0.613333	0.783156	-0.011651	-0.116174
pea = oui	0.386667	0.621825	0.014674	0.146315

Figure 3 – Factor scores coefficients

We can visualize the factor coordinates of the individuals by using the VIEW DATASET component.

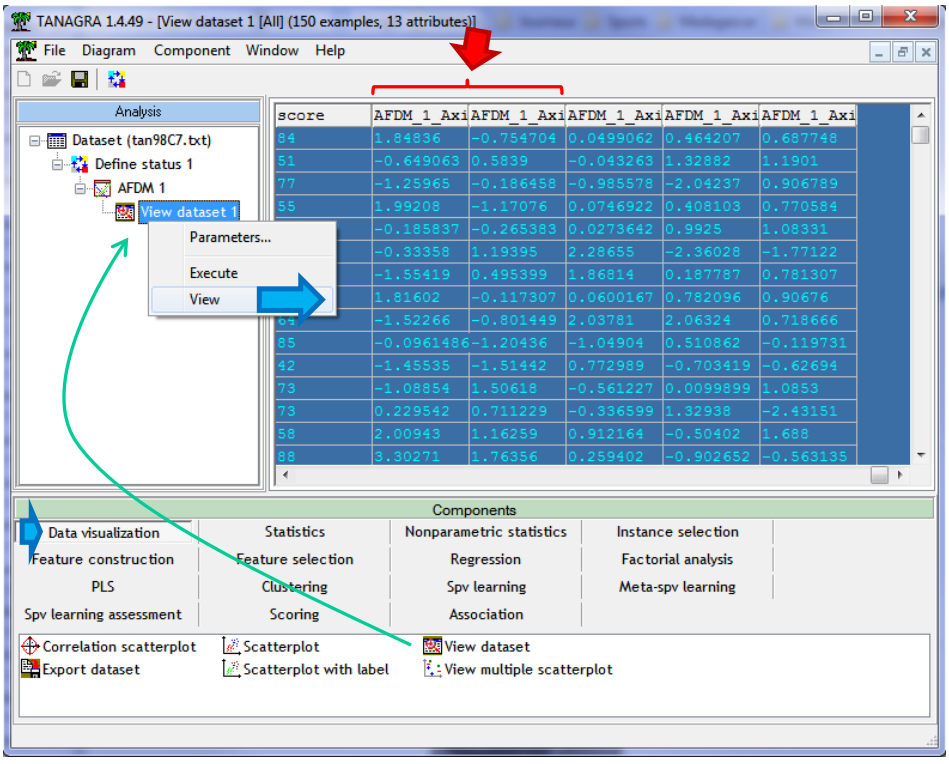
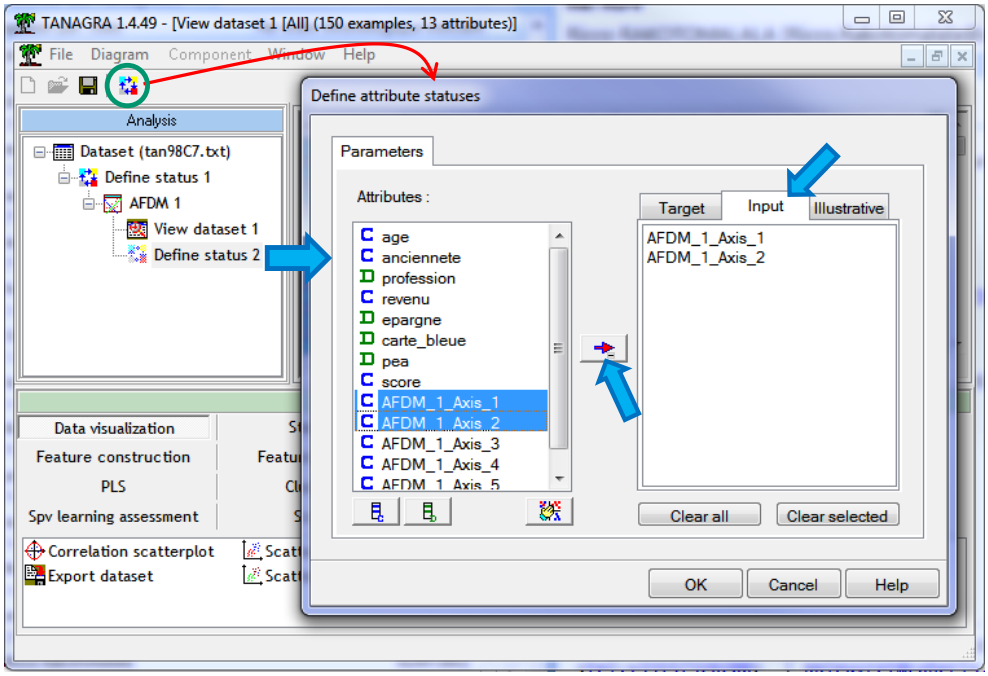


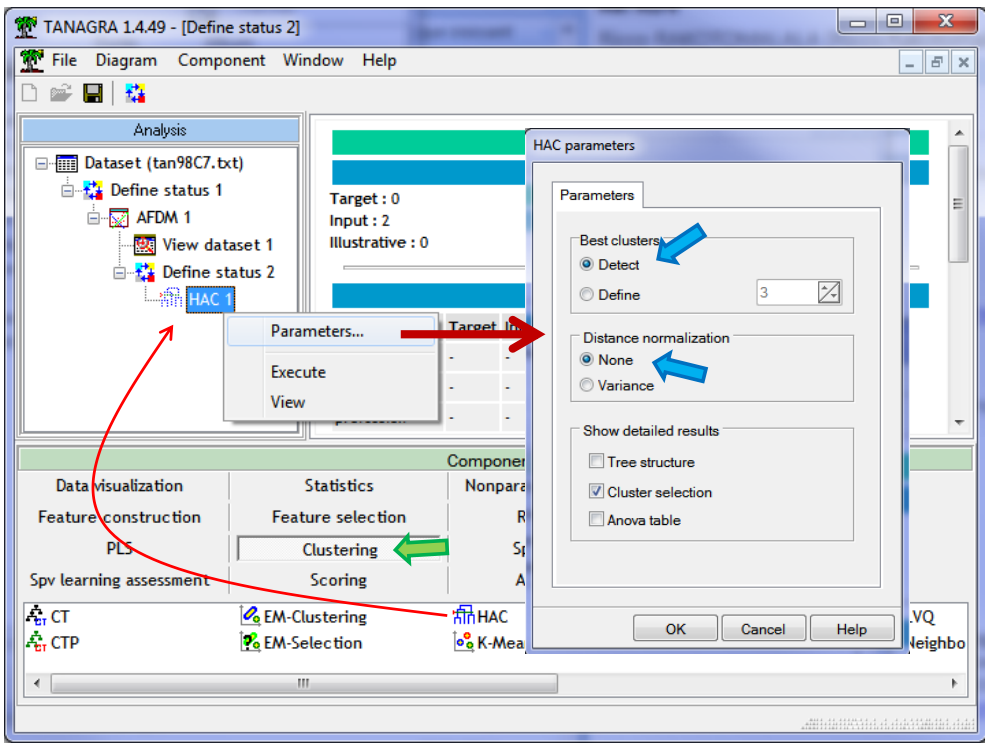
Figure 4 – Factor scores of the individuals for the 5 first components

3.3 HAC from the selected components

We use again the DEFINE STATUS tool to select the components to use in the clustering process.



We select the AFDM_1_AXIS_1 and AFDM_1_AXIS_2 columns. Then, we insert the HAC tool into the diagram ("Hierarchical Agglomerative Clustering", CLUSTERING tab). We set the following parameters:



We do not standardize the variables used in the clustering process. Thus, each component influences the results according their weights. Tanagra tries to detect automatically the right number of groups. It draws on the merging height measured at each step of the process. We launch the calculations by clicking on the VIEW menu.

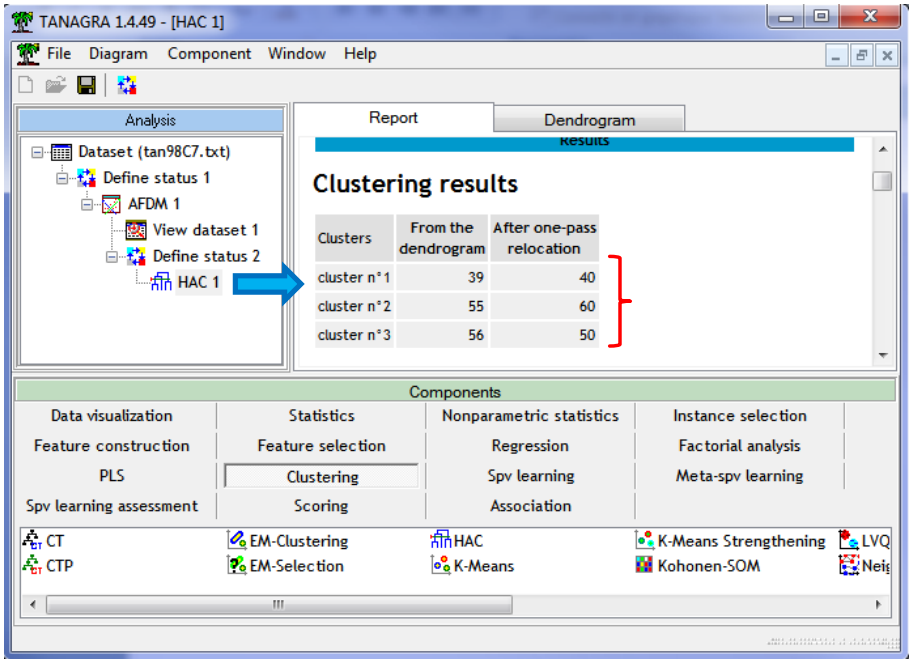
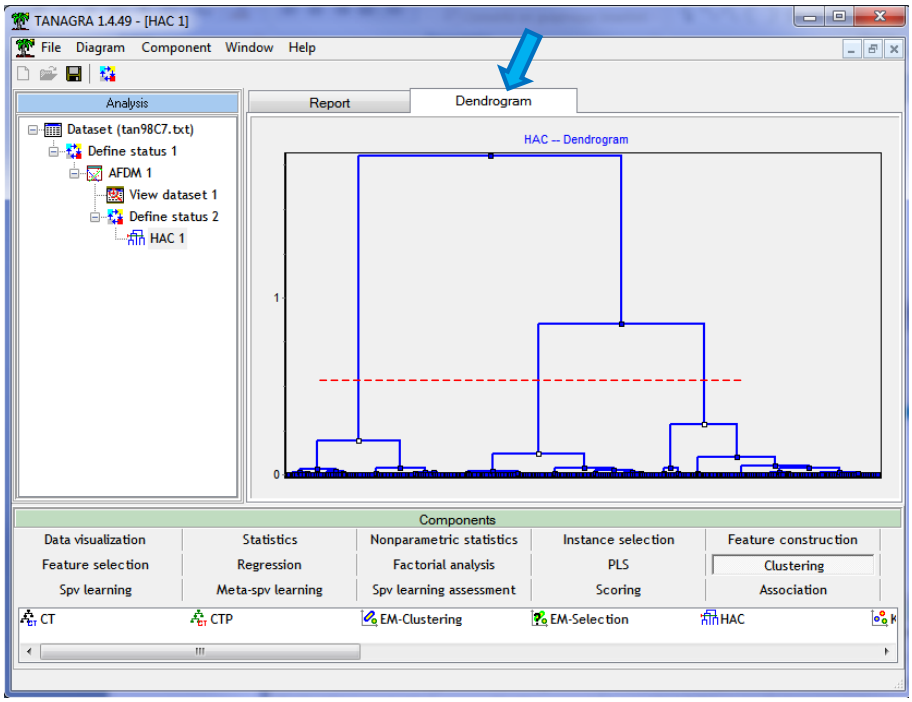


Figure 5 – Results of the HAC algorithm - Clusters' size before and after the relocation process

Tanagra provides 3 groups with respectively: 40, 60 and 50 instances. The cluster sizes are different from those observed in the dendrogram because Tanagra, from the 1.4.48 version[5], performs a last pass on the data in order to assign individuals to the group for which the centroid is the closest. The objective is to obtain more compact groups, the initial partition being constrained by the hierarchical structure of the search for solutions.

The grouping in 3 clusters seems the more relevant solution according the dendrogram (we put aside the solution in 2 groups which correspond [almost] always to the highest merging level).



[5] http://data-mining-tutorials.blogspot.fr/2012/12/tanagra-version-1448.html

In the lower part of the report, Tanagra shows the cluster centroids. We will use this information when we want to assign a new instance to a group.

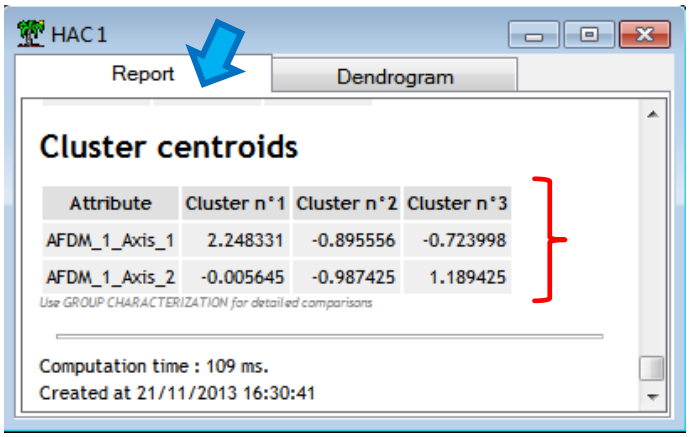


Figure 6 – Cluster centroids

3.4 Visualizing the groups in the factorial map

To check the quality of the solution, we visualize the groups in the first factorial map defined by the two first factors. We insert the SCATTERPLOT tool (DATA VISUALIZATION tab) into the diagram. We set AFDM_1_AXIS_1 on the horizontal axis, AFDM_1_AXIS_2 on the vertical axis. We colorize the points according to the group membership (the CLUSTER_HAC_1 provided by the HAC tool).

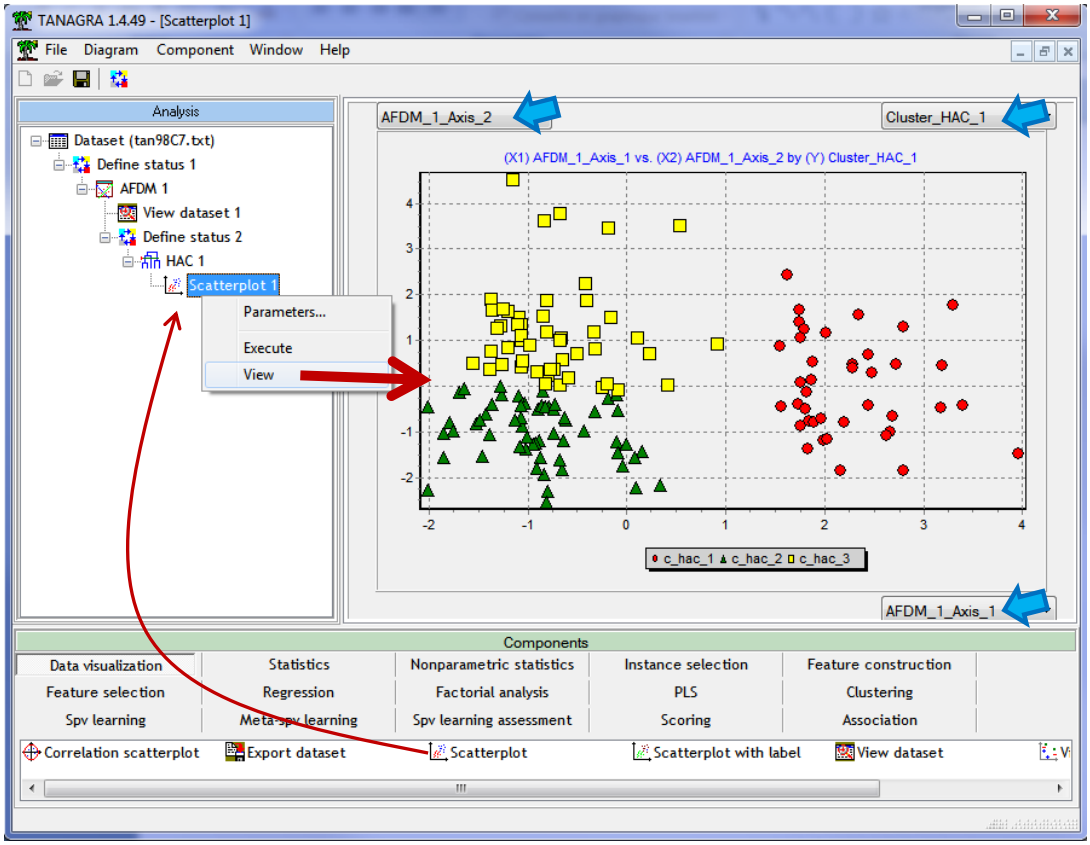


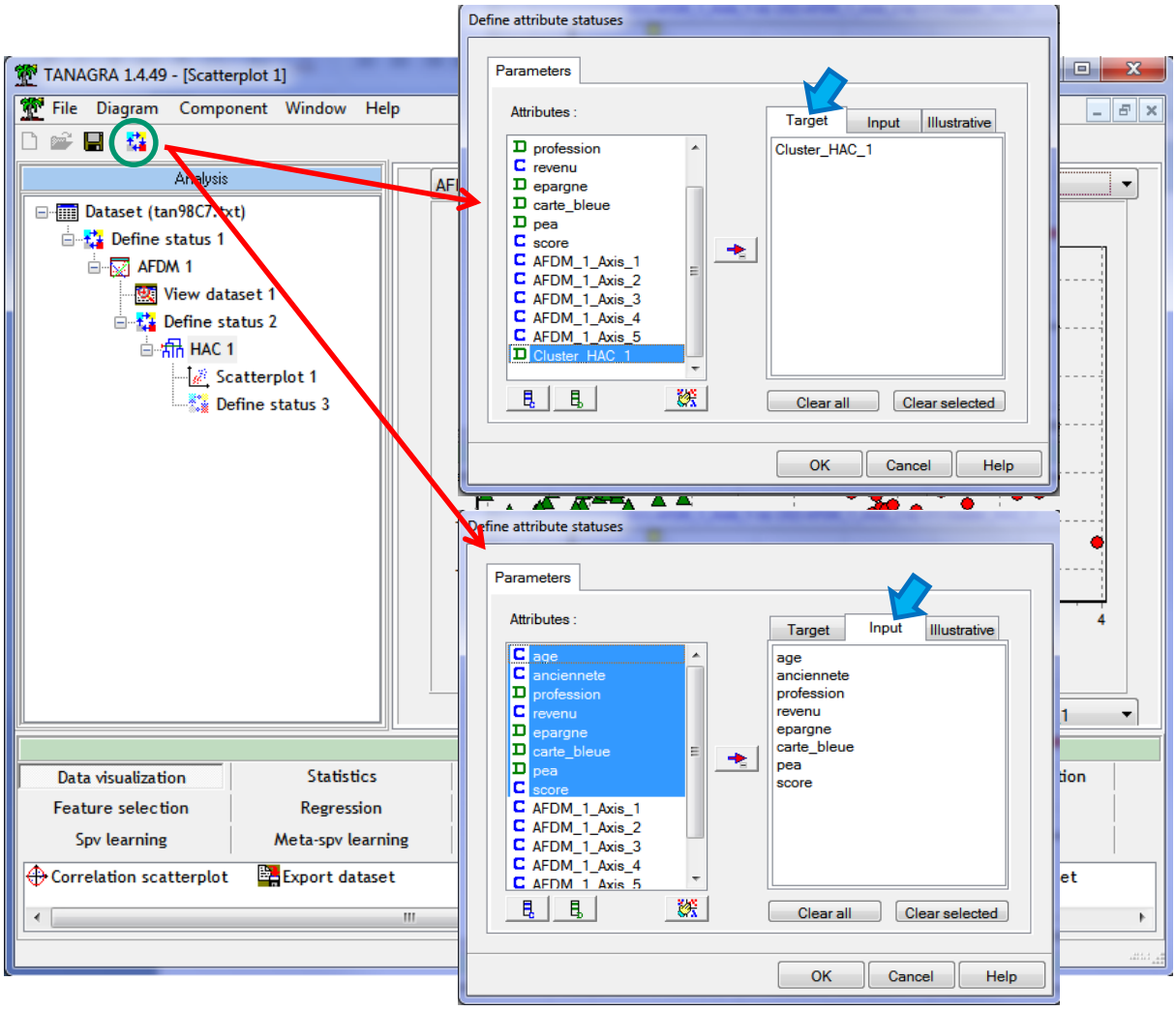
Figure 7 – Visualizing the groups into the first factorial map

We observe clearly the 3 groups highlighted by the clustering algorithm. The first component allows to separate the first cluster (C_HAC_1), the second allows to distinguish the second (C_HAC_2) and the

third (C_HAC_3) clusters. The groups are perfectly separated - there is no overlapping between classes - in the first factorial map. This is quite normal because we had used these first two factors for the clustering.

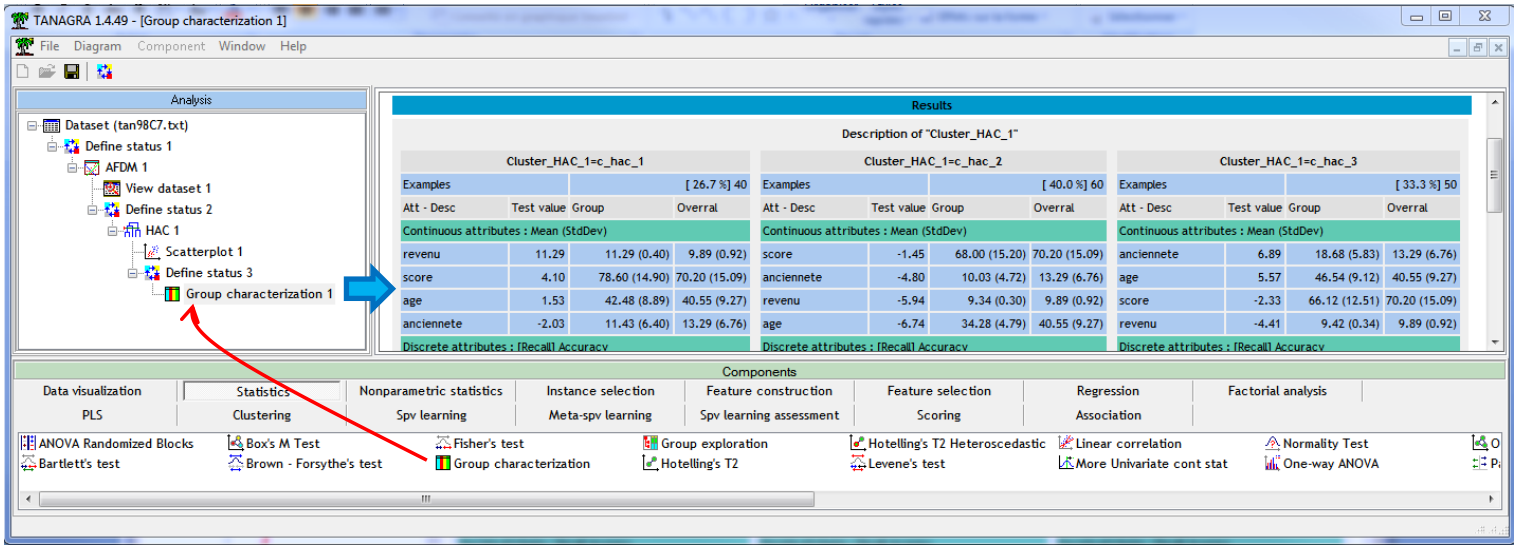
3.5 Description of groups – Active and supplementary variables

Now, we want to understand the distinctive features of the clusters, with the variables used during the clustering process, but also with the additional variable SCORE which describes the appreciation of the bank advisor. We insert again the DEFINE STATUS tool into the diagram. We set as TARGET the variable which associates each individual to a cluster. We set as INPUT all the variables, including the SCORE variable.



Then we insert the GROUP CHARACTERIZATION tool (STATISTICS tab) into the diagram. In this table, some descriptive statistics (mean, frequency, etc.) are computed on the whole dataset, and within each group. The importance of the differences is highlighted with the TEST VALUE indicator[6]. It enables to detect the characteristics that differentiate the group.

[6] http://data-mining-tutorials.blogspot.fr/2009/05/understanding-test-value-criterion.html



We give the detail of the table provided by this tool below.

Description of "Cluster_HAC_1"												
Cluster_HAC_1=c_hac_1				Cluster_HAC_1=c_hac_2				Cluster_HAC_1=c_hac_3			
Examples		[26.7 %] 40		Examples		[40.0 %] 60		Examples		[33.3 %] 50	
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)			
revenu	11.29	11.29 (0.40)	9.89 (0.92)	score	-1.45	68.00 (15.20)	70.20 (15.09)	anciennete	6.89	18.68 (5.83)	13.29 (6.76)
score	4.1	78.60 (14.90)	70.20 (15.09)	anciennete	-4.8	10.03 (4.72)	13.29 (6.76)	age	5.57	46.54 (9.12)	40.55 (9.27)
age	1.53	42.48 (8.89)	40.55 (9.27)	revenu	-5.94	9.34 (0.30)	9.89 (0.92)	score	-2.33	66.12 (12.51)	70.20
anciennete	-2.03	11.43 (6.40)	13.29 (6.76)	age	-6.74	34.28 (4.79)	40.55 (9.27)	revenu	-4.41	9.42 (0.34)	9.89 (0.92)
Discrete attributes : [Recall] Accuracy				Discrete attributes : [Recall] Accuracy				Discrete attributes : [Recall] Accuracy			
profession= CAD	12.21	[100.0 %] 100.0 %	26.70%	profession= EMP	4.41	[77.8 %] 35.0 %	18.00%	profession= INA	4.09	[84.6 %] 22.0 %	8.70%
epargne= elevee	1.76	[37.8 %] 35.0 %	24.70%	profession= AGR	3.31	[100.0 %] 11.7 %	4.70%	profession= OUV	2.66	[61.1 %] 22.0 %	12.00%
pea=oui	0.58	[29.3 %] 42.5 %	38.70%	epargne= faible	2.77	[73.3 %] 18.3 %	10.00%	profession= ART	2.05	[56.3 %] 18.0 %	10.70%
carte_bleue=oui	0.11	[26.8 %] 95.0 %	94.70%	pea=non	2.11	[46.7 %] 71.7 %	61.30%	pea=oui	1.65	[41.4 %] 48.0 %	38.70%
epargne= moyenne	-0.05	[26.5 %] 65.0 %	65.30%	profession= INT	1.85	[55.2 %] 26.7 %	19.30%	profession= INT	1.46	[44.8 %] 26.0 %	19.30%
carte_bleue=non	-0.11	[25.0 %] 5.0 %	5.30%	carte_bleue=oui	0.89	[40.8 %] 96.7 %	94.70%	carte_bleue=non	1.02	[50.0 %] 8.0 %	5.30%
pea=non	-0.58	[25.0 %] 57.5 %	61.30%	profession= ART	0.32	[43.8 %] 11.7 %	10.70%	epargne= elevee	0.27	[35.1 %] 26.0 %	24.70%
profession= AGR	-1.63	[0.0 %] 0.0 %	4.70%	epargne= moyenne	-0.07	[39.8 %] 65.0 %	65.30%	epargne= moyenne	0.12	[33.7 %] 66.0 %	65.30%
profession= INA	-2.27	[0.0 %] 0.0 %	8.70%	profession= OUV	-0.1	[38.9 %] 11.7 %	12.00%	epargne= faible	-0.58	[26.7 %] 8.0 %	10.00%
epargne= faible	-2.45	[0.0 %] 0.0 %	10.00%	carte_bleue=non	-0.89	[25.0 %] 3.3 %	5.30%	carte_bleue=oui	-1.02	[32.4 %] 92.0 %	94.70%
profession= ART	-2.54	[0.0 %] 0.0 %	10.70%	epargne= elevee	-1.85	[27.0 %] 16.7 %	24.70%	profession= EMP	-1.35	[22.2 %] 12.0 %	18.00%
profession= OUV	-2.72	[0.0 %] 0.0 %	12.00%	profession= INA	-1.89	[15.4 %] 3.3 %	8.70%	pea=non	-1.65	[28.3 %] 52.0 %	61.30%
profession= EMP	-3.45	[0.0 %] 0.0 %	18.00%	pea=oui	-2.11	[29.3 %] 28.3 %	38.70%	profession= AGR	-1.91	[0.0 %] 0.0 %	4.70%
profession= INT	-3.6	[0.0 %] 0.0 %	19.30%	profession= CAD	-6.01	[0.0 %] 0.0 %	26.70%	profession= CAD	-5.2	[0.0 %] 0.0 %	26.70%

Figure 8 – Characterization of the clusters - Comparison of the global and conditional means and frequencies

We detail below the description of the first cluster.

- The mean of REVENU (income) in the whole dataset is 9.89, the standard deviation is 0.92. In the first group, the mean and the standard deviation become respectively 11.29 and 0.40. To appreciate the gap between the means, we calculate the TEST VALUE which is similar to the Student's t statistic (this is not the true Student's t-test because the samples are not independent here). For a 0.05 significance level, the gap is statistically significant if it is lower than -2 or upper than +2 (approximately). We observe that people in this group have a higher income than the whole population (TEST VALUE = 11.29).
- The mean of SCORE is 70.20 for the whole population; it is equal to 78.60 in this group. The difference is also significant (TEST VALUE = 4.1). That means that the customer advisor has a positive opinion - on average - of the individuals in this group. This is not really surprising. The banker is interested by the people who have high income.
- About the categorical variables, we observe that the proportion of the executive people (Profession = CAD) is 26.7%. In this group, we have only executive people (proportion = 100%). In addition, we observe also that all the executive people in the whole dataset are gathered in this group (recall = 100%). This overrepresentation is highlighted by a high TEST VALUE = 12.21 (it compares the proportions in the case of categorical variables).

We can summarize the main characteristics of each cluster as follows.

Cluster	Characteristics
Group 1	This group corresponds to the people with high incomes [INCOME], which interest the bank [SCORE]. These are customers fairly recent [ANCIENNETE is SENIORITY] which are executives [PROFESSION = CAD], with a slightly higher savings in average [SAVINGS = HIGH]. In short, these are the customers with high potential, to whom perhaps the bank can promote new products.
Group 2	Those are recent young customers who do not really interest banker (TEST VALUE of SCORE = -1.45). Employees (Profession = EMP) and farmers (AGR) are overrepresented. They do not have much savings. In short, they have a little potential for the banker.
Group 3	Those are the traditional customers (old, high seniority) that do not interest at all the banker (TEST VALUE of SCORE = -2.33). They have low incomes, but they are about average for the savings.

Note: We can get a similar interpretation by studying directly the results tables of the factor analysis of mixed data. But the reading requires a better experience of this kind of approach[7].

[7] http://data-mining-tutorials.blogspot.fr/2013/03/factor-analysis-for-mixed-data.html

3.6 Classifying a new instance

We want to associate a new individual ω with one of the groups. Here are its observed values on the active variables.

age	ancienne te	profession	revenu	epargne	carte_bleue	pea
55	22	INT	10.035	moyenne	oui	non

Step 1: Calculating the factor scores. We use the factor scores coefficients (Figure 3) provided by the AFDM (factor analysis for mixed data) to compute the coordinates of the instance into the first factorial map. We use the parameters (mean and scale) for the standardization of the value before applying the coefficients. For the categorical attributes, we use the corresponding dummy variable (e.g. for "pea = non", we set 1 for the corresponding dummy variable, 0 for the other ones).

We detail below the calculation for the first component.

$$F_1 = 0.120844 \times \frac{55 - 40.553333}{9.243763} - 0.173090 \times \frac{22 - 13.286667}{6.735317} + 0.538255 \times \frac{0 - 0.266667}{0.516398}$$
$$- 0.190155 \times \frac{1 - 0.193333}{0.439697} - 0.178526 \times \frac{0 - 0.12}{0.34641} + \cdots - 0.011651 \times \frac{1 - 0613333}{0.783156}$$
$$+ 0.014674 \times \frac{0 - 0386667}{0621825}$$
$$= -0.453$$

We obtain the coordinates of the instance into the first factorial map (F_1: -0.453, F_2: 1.618). It seems that this new individual belongs rather to the third cluster when we observe the scatter plot (Figure 9).

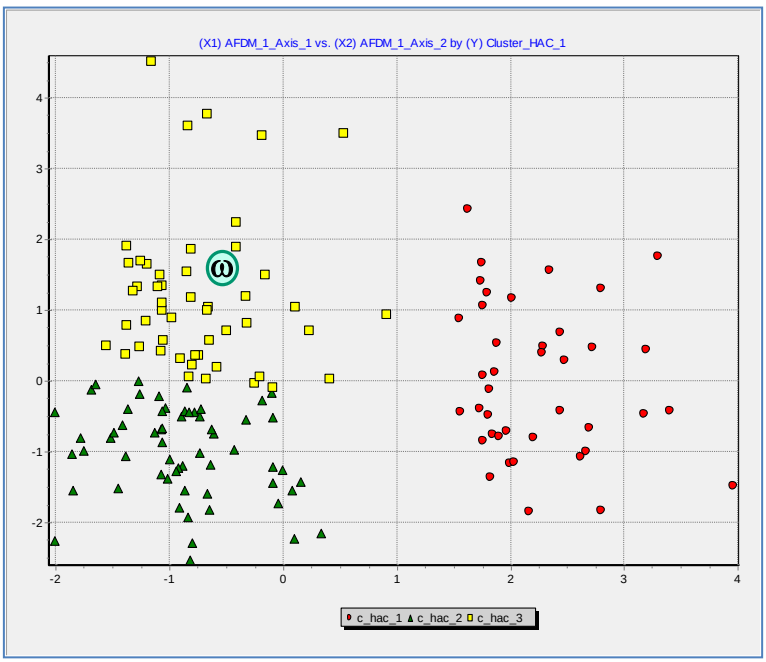


Figure 9 - Positioning the new instance within the existing instances

Step 2: Computing the distance to the clusters' centroid G_k. We use the cluster centroids table (Figure 6) to detect the closest centroid. Below, we calculate the Euclidian distance between the new instance and the centroid of the first cluster.

$$d^2(G_1) = (-0.453 - 2.248331)^2 + (1.618 + 0.005645)^2 = 9.934$$

So, we obtain 3 distance values: $d^2(G_1) = 9.934$; $d^2(G_2) = 6.985$; $d^2(G_3) = 0.257$. Clearly, the centroid of the 3rd group is the closest (Figure 10). This additional individual may be associated to this cluster. This is not surprising in the light of its characteristics: age and seniority are substantially higher than the average; this is a traditional customer of the Bank.

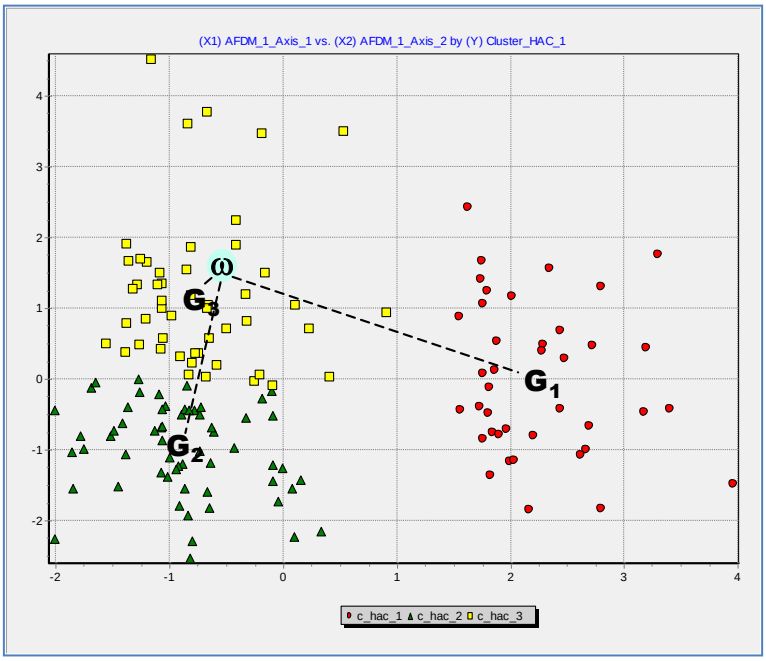


Figure 10 - Positioning the new instance within the centroids

4 Clustering for mixed data using R

The factor analysis for mixed data is available in several R packages. Below, we perform the same analysis using the **dudi.mix()** procedure (ade4 package) for the factor analysis, and using the well-known **hclust()** procedure (stats package) for the clustering analysis. Here is the R program.

```
#loading the data file using the xlsx package
library(xlsx)
bank <- read.xlsx(file="BankCustomer.xls",sheetIndex=1,header=T)
#descriptive statistics
summary(bank)
#active variables
bank.active <- bank[,1:7]
#loading the ade4 package
library(ade4)
#AFDM: Factorial Analysis for Mixed Data
#we select the 2 first components
bank.afdm <- dudi.mix(bank.active,scannf=F,nf=2)
#displaying the factor scores for the 5 first instances
print(head(bank.afdm$li,5))
#euclidian distance between instances
dist.afdm <- dist(bank.afdm$li[,1:2],method="euclidian")
#square of the distance for the Ward's method
```

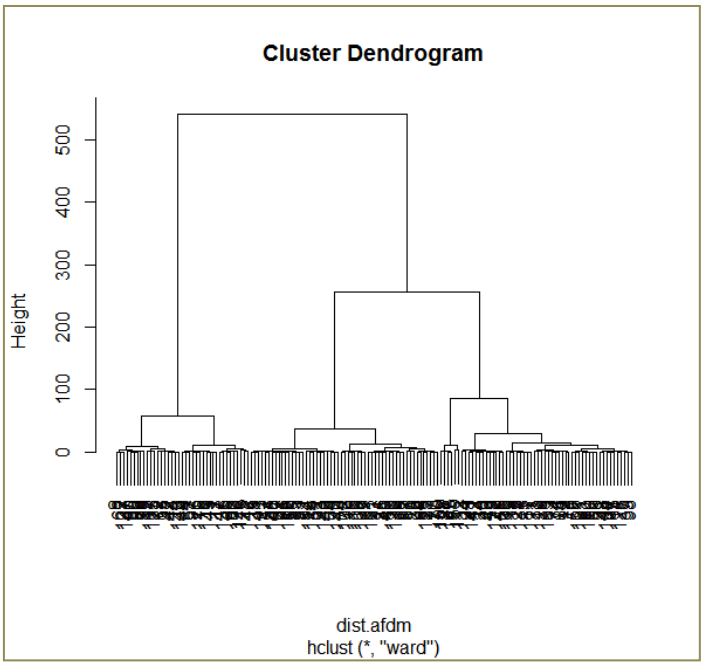
```
#voir http://en.wikipedia.org/wiki/Ward's_method
dist.afdm <- dist.afdm^2
#hierarchical agglomerative clustering from
#the square distance matrix
bank.tree <- hclust(dist.afdm,method="ward")
plot(bank.tree)
#cutting the dendrogram: k = 3 clusters
bank.clusters <- cutree(bank.tree,k=3)
#counting the instances into each cluster
table(bank.clusters)
#first factorial map
#colouring the points according to the cluster membership
plot(bank.afdm$li[,1],bank.afdm$li[,2],col=c("red","yellow","green")[bank.clusters])
#calculating the mean of the SCORE variable within each cluster
print(aggregate(x=bank$score,by=list(bank.clusters),FUN=mean))
```

Below, we detail the results at each step.

Factor scores of individuals. We observe below the factor scores for the 5 first individuals. Because the tools are based on the same underlying algorithm, we obtain the same values as Tanagra (Figure 4). The sign is different for the 2nd component, but the proximity between the instances is the same.

```
> print(head(bank.afdm$li,5))
       Axis1      Axis2
1  1.8483643  0.7547039
2 -0.6490631 -0.5838999
3 -1.2596491  0.1864582
4  1.9920786  1.1707646
5 -0.1858365  0.2653832
```

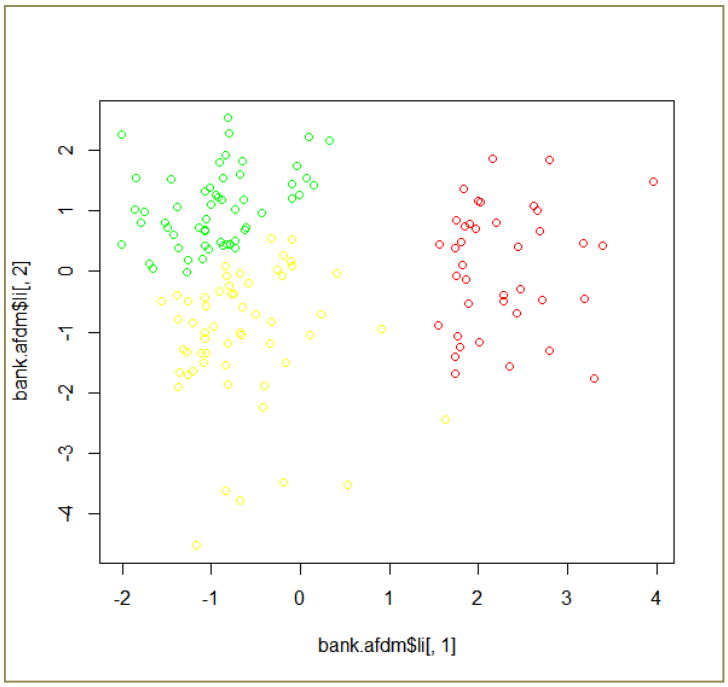
Dendrogram. The **hclust()** procedure uses the square of the distance matrix for the Ward's approach. Obviously, the splitting into three groups is relevant.



Cluster sizes. The cluster sizes are the same as Tanagra (Figure 5), before the relocation process of this last one.

```
> table(bank.clusters)
bank.clusters
 1  2  3
39 56 55
```

Visualizing the groups. The first group is well separated to the others on the first component. The second component enables to distinguish the 2nd and the 3rd groups.



Mean of the SCORE variable according to the groups. The results are consistent with those of Tanagra (Figure 8). The SCORE assigned by the bank advisor is really different according to the groups. They are consistent (with Tanagra) but slightly different because R does not reprocess the results by assigning each instance to the cluster with the closest centroid.

```
> print(aggregate(x=bank$score,by=list(bank.clusters),FUN=mean))
  Group.1        x
1       1 78.64103
2       2 66.62500
3       3 67.85455
```

5 Conclusion

Dealing a dataset with mixed variables is a usual circumstance in real studies. In this paper, we show how to perform a clustering process in that situation. The approach is based on the tandem analysis scheme: first, we perform a factor analysis for mixed data to compute the factor scores of the individuals; second, we use theses coordinates to perform a standard clustering algorithm. The results are relevant. In addition, we can use the solution for detecting the cluster that we can associate to a new unseen instance.