1 Objectif

Classification automatique sur données mixtes (mélange de variables qualitatives et quantitatives). Utilisation des facteurs de l'analyse factorielle de données mixtes.

La classification automatique ou typologie (clustering en anglais) vise à regrouper les observations en classes : les individus ayant des caractéristiques similaires sont réunis dans la même catégorie ; les individus présentant des caractéristiques dissemblables sont situés dans des catégories distinctes. La notion de proximité est primordiale dans ce processus. Elle est quantifiée différemment selon le type des variables. La distance euclidienne est souvent utilisée (normalisée ou non) lorsqu'elles sont quantitatives, la distance du khi-2 lorsqu'elles sont qualitatives (les individus qui possèdent souvent les mêmes modalités sont réputées proches).

L'affaire se corse lorsque nous sommes en présence d'un mix de variables quantitatives et qualitatives. Certes il est toujours possible de définir une distance prenant en compte simultanément les deux types de variables (ex. la distance HEOM[1]). Mais le problème de la normalisation est posé. Telle ou telle variable ne doit pas avoir une influence exagérée uniquement de par sa nature.

Précédemment, nous avons présenté l'analyse factorielle de données mixtes (AFDM)[2]. Il s'agit de projeter les observations dans un repère factoriel élaboré à partir d'un mélange de variables actives qualitatives et quantitatives. On montre que l'approche est équivalente à l'ACP normée (analyse en composantes principales) lorsque les variables sont toutes quantitatives, à l'ACM (analyse des correspondances multiples) lorsqu'elles sont toutes qualitatives.

Nous proposons dans ce tutoriel de réaliser la classification sur données mixtes en deux étapes : [1] nous procédons tout d'abord à une AFDM pour produire une représentation intermédiaire des données ; [2] puis, nous effectuons une classification ascendante hiérarchique (CAH) à partir des facteurs « représentatifs » de l'AFDM. Cette analyse en deux temps est couramment utilisée même lorsque les variables sont exclusivement quantitatives (on passe par l'ACP) ou qualitatives (ACM). L'idée est de procéder à un nettoyage des données – une sorte de régularisation – en éliminant les derniers facteurs qui correspondent à du bruit spécifique à l'échantillon que nous utilisons c.-à-d. des informations qui correspondent aux fluctuations d'échantillonnage ne reflétant en rien un phénomène réel dans la population. Les résultats sont ainsi plus stables[3].

---

[1] http://axon.cs.byu.edu/~randy/jair/wilson2.html

[2] http://tutoriels-data-mining.blogspot.fr/2013/08/analyse-factorielle-de-donnees-mixtes.html

[3] Cf. Maurice Roux, « Algorithmes de classification », Masson, 1985 ; page 15, « 1.2. Prétraitement par l'analyse factorielle ». **Attention**, cette approche en deux temps, connue sous l'appellation « tandem analysis » dans les références anglo-saxonnes n'est cependant pas la panacée. Les deux étapes étant réalisées de manière indépendante, dans certains cas difficiles à identifier a priori, la projection dans un espace de représentation réduit issu d'une sélection des premiers facteurs d'une analyse factorielle peut masquer les informations permettant de dissocier les groupes. Cf. DeSarbo, W. S., Jedidi, K., Cool, K. And Schendel, D., 1990. Simultaneous Multidimensional Unfolding and Cluster Analysis: An Investigation of Strategic Groups, Marketing Letters, 2 129-146 ; De Soete G., and Carroll, J. D., 1994. K-means Clustering in a Low-dimensional Euclidean Space, in: E.Diday et al. (Eds), New Approaches in Classification and Data Analysis, Springer, Heidelberg, 212-219.

Nous utiliserons les logiciels Tanagra 1.4.49 et R (package ade4) dans ce tutoriel.

2 Données

La base « bank_customer.xls » décrit les clients d'une banque. Les variables actives sont relatives à la situation du client : âge, ancienneté auprès de la banque, profession, logarithme du revenu, niveau d'épargne, détention d'une carte bleue, détention d'un PEA (plan épargne en action). La variable illustrative SCORE correspond à une note attribué par le conseiller clientèle pour évaluer l'intérêt d'un client[4]. Voici les 5 premières lignes du fichier.

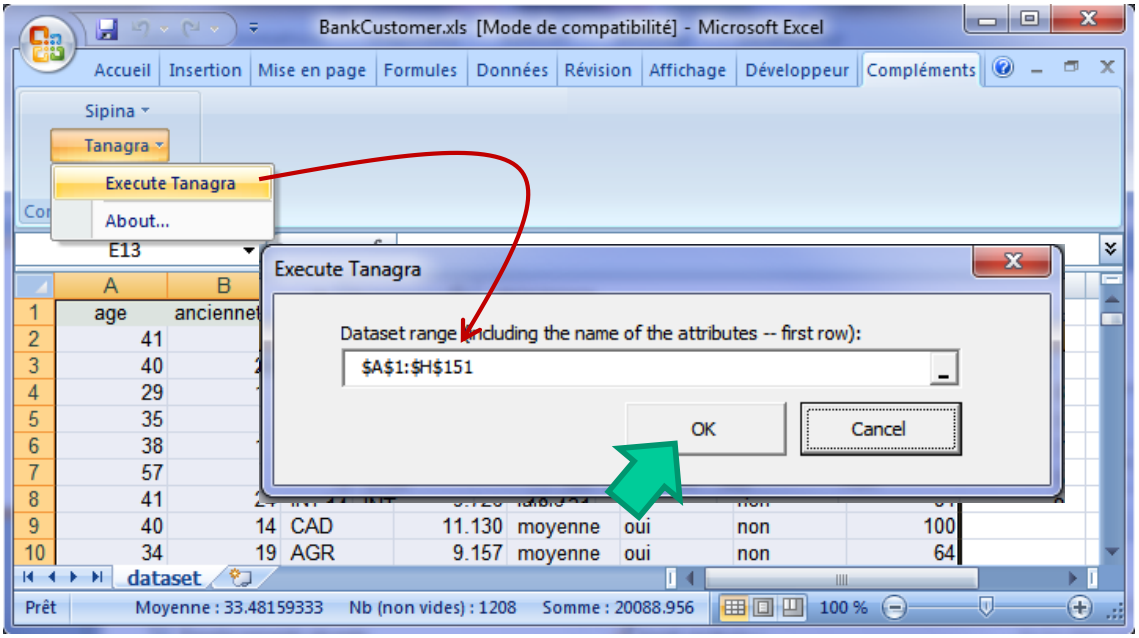
age	ancienneté	profession	revenu	epargne	carte_bleue	pea	score
41	6	CAD	10.870	moyenne	oui	non	84
40	22	INT	10.035	moyenne	oui	non	51
29	12	OUV	9.087	moyenne	oui	oui	77
35	6	CAD	11.180	moyenne	oui	non	55
38	14	INT	10.431	moyenne	oui	non	87

L'enjeu est de produire une typologie des clients à partir de leurs caractéristiques, puis de situer les catégories par rapport à la perception des conseillers clientèles illustrée par la variable SCORE.

3 Typologie sur données mixtes avec Tanagra

3.1 Importation des données

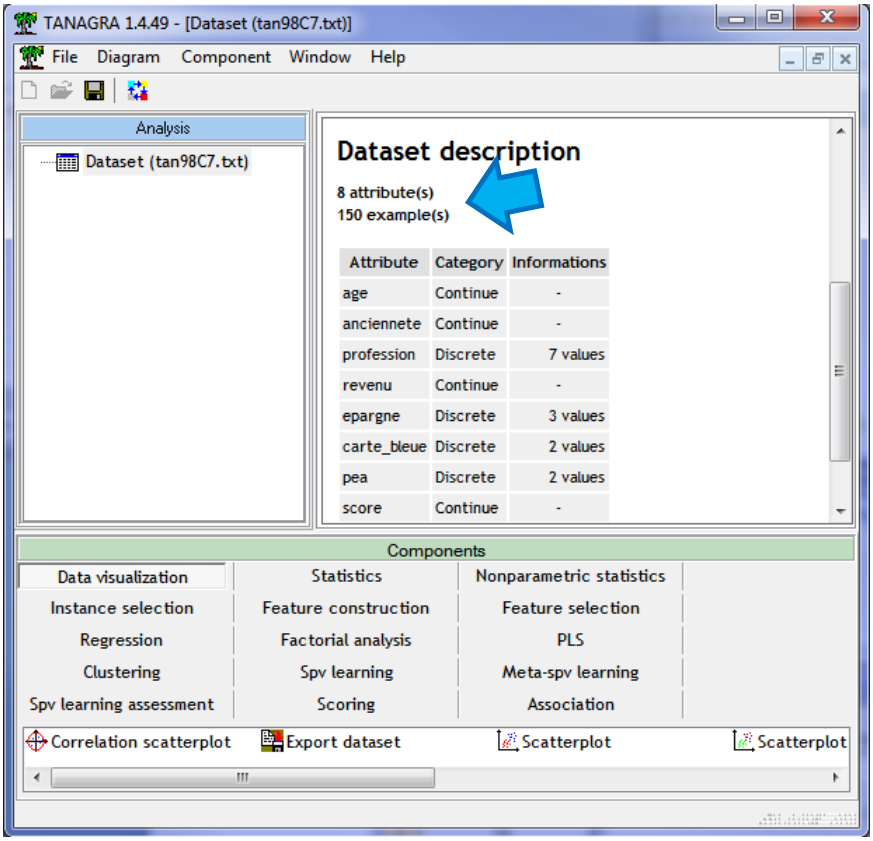
Nous chargeons le fichier dans le tableur Excel. Nous sélectionnons les données puis, via la macro complémentaire « tanagra.xla »[5], nous les envoyons vers Tanagra.



[4] Ne nous affolons pas, ce fichier est totalement fictif.

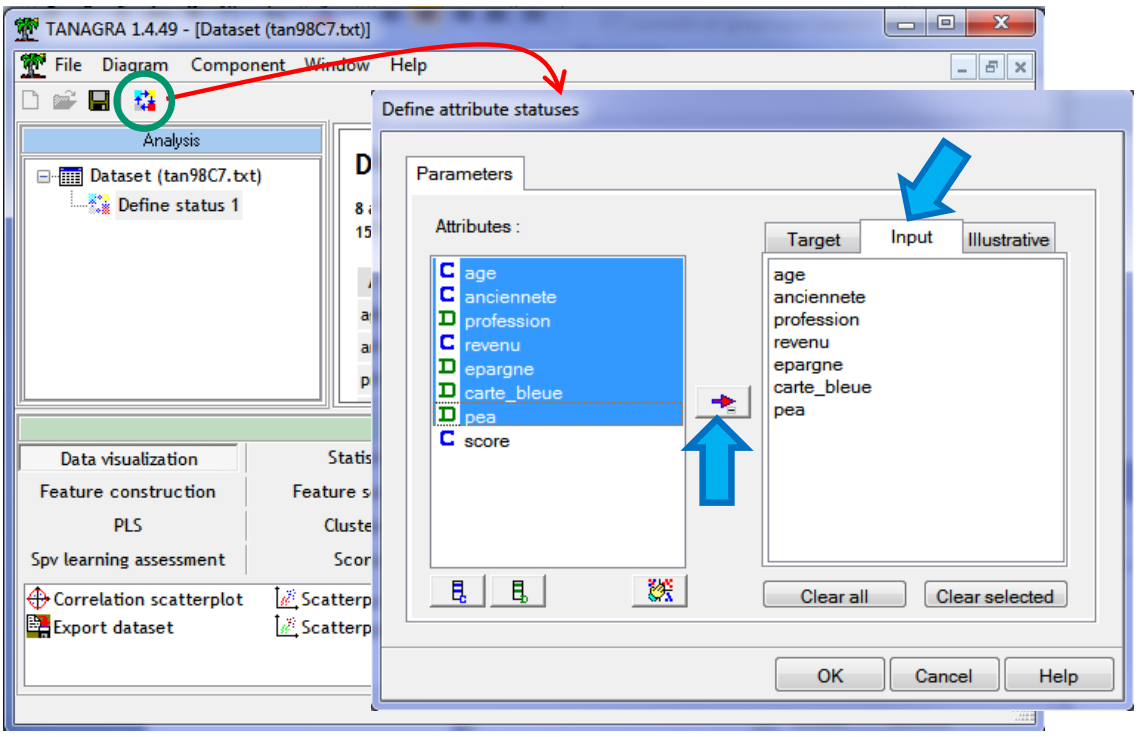
[5] Voir http://tutoriels-data-mining.blogspot.fr/2010/08/ladd-in-tanagra-pour-excel-2007-et-2010.html pour l'installation et l'utilisation de la macro-complémentaire. Ce type de dispositif existe également pour les versions précédentes d'Excel (2003 à 1997, http://tutoriels-data-mining.blogspot.fr/2008/03/importation-fichier-xls-excel-macro.html) ou encore pour le tableur Calc des suites « Libre Office » et « Open Office » (http://tutoriels-data-mining.blogspot.fr/2011/07/tanagra-addon-pour-openoffice-33.html).

Tanagra est démarré, nous vérifions que **150 observations** et **8 attributs** ont bien été importés.

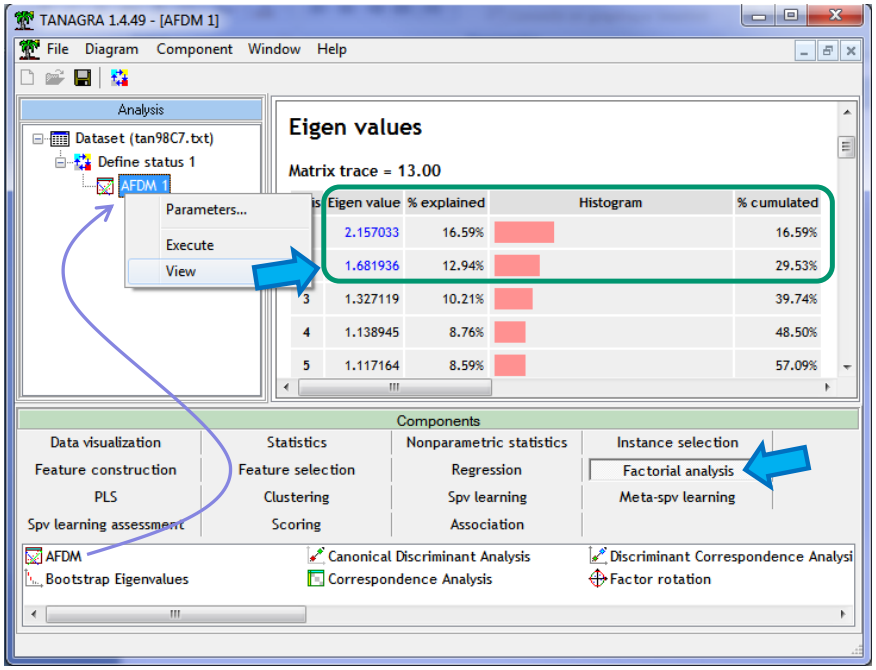


3.2 Analyse factorielle des données mixtes (AFDM)

Nous devons sélectionner les variables actives à l'aide du composant DEFINE STATUS avant de lancer l'AFDM. Nous l'insérons dans le diagramme des traitements à l'aide du raccourci de la barre d'outils. Nous plaçons en INPUT les 7 premières variables (« âge », …, « pea »).



Nous ajoutons le composant **AFDM** (onglet FACTORIAL ANALYSIS). Nous actionnons le menu VIEW pour obtenir les résultats.



Le choix des facteurs à retenir est toujours délicat en analyse factorielle. Il l'est d'autant plus que nous souhaitons les exploiter dans des calculs ultérieurs. De la pertinence de notre choix dépend la qualité de la typologie que nous réaliserons.

Concernant nos données, **nous décidons de conserver les 2 premiers facteurs** en nous fiant au « décrochement » dans l'histogramme des valeurs propres. Certes, nous ne prenons en compte que 29.53% de l'inertie disponible ce faisant. Mais, encore une fois, nous ne souhaitons pas obtenir une vision exhaustive des données mais plutôt nous appuyer sur suffisamment d'informations pour produire une typologie qui soit un tant soit peu pertinente c.-à-d. interprétable[6].

Pour préciser les résultats, nous reprenons ci-dessous le tableau des « communalities » qui correspond au carré de la corrélation des variables avec les facteurs lorsqu'elles sont quantitatives, au carré du rapport de corrélation lorsqu'elles sont qualitatives.

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 %)
ancienneté (*)	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

[6] Ce choix obéit aussi à des visées pédagogiques. Il sera plus aisé de visualiser les groupes dans le premier plan factoriel.

Les « Factor loadings » précisent le sens des relations des variables quantitatives avec les facteurs.

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 - Corrélation des variables quantitatives avec les facteurs

Les « Conditional Means » positionnent les modalités des variables qualitatives sur les facteurs.

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 - Moyennes des modalités sur les facteurs

Les « Factor Scores » fournissent les coefficients des fonctions de projection.

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 - Coefficients des fonctions de projection

Ce dernier tableau est importantissime. En effet, lors du déploiement, il sera nécessaire de calculer ses coordonnées dans le repère factoriel pour pouvoir rattacher un individu à une classe.

Il est possible d'inspecter les coordonnées des individus à l'aide du composant VIEW DATASET que nous insérons dans le diagramme de traitements.

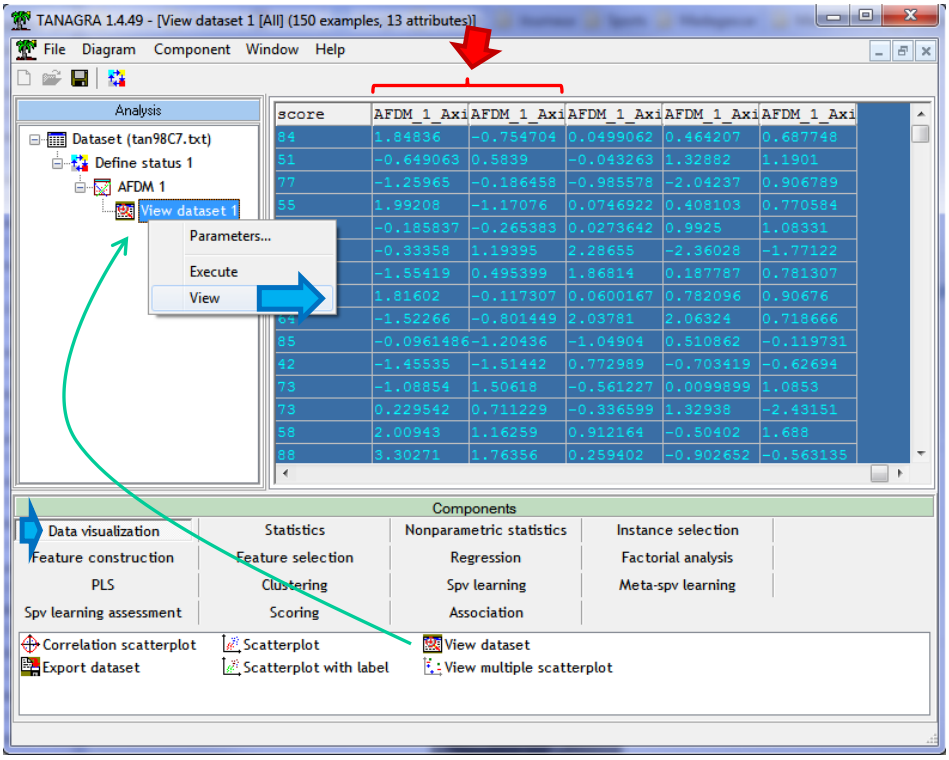
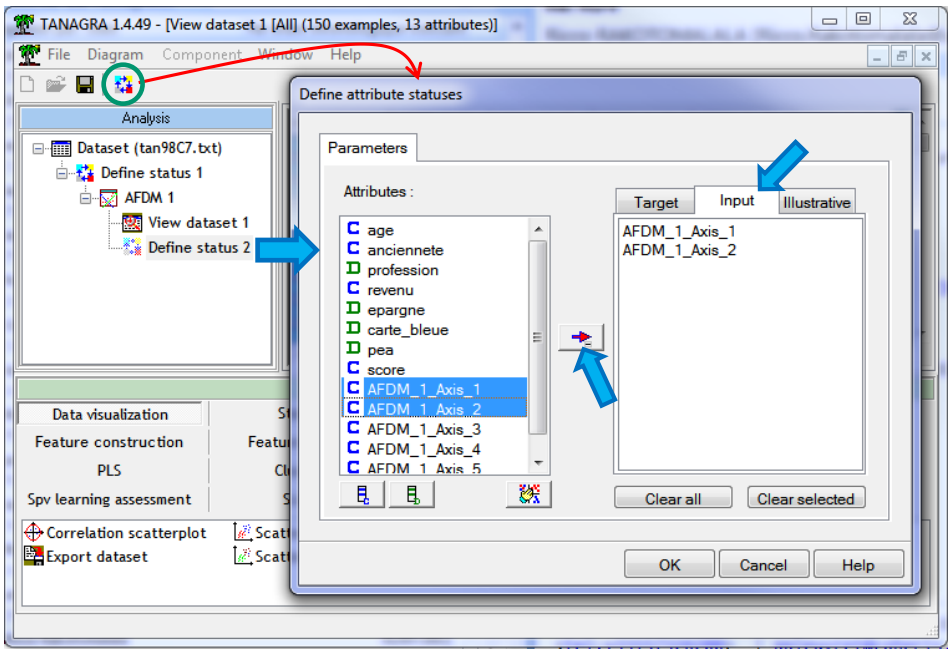


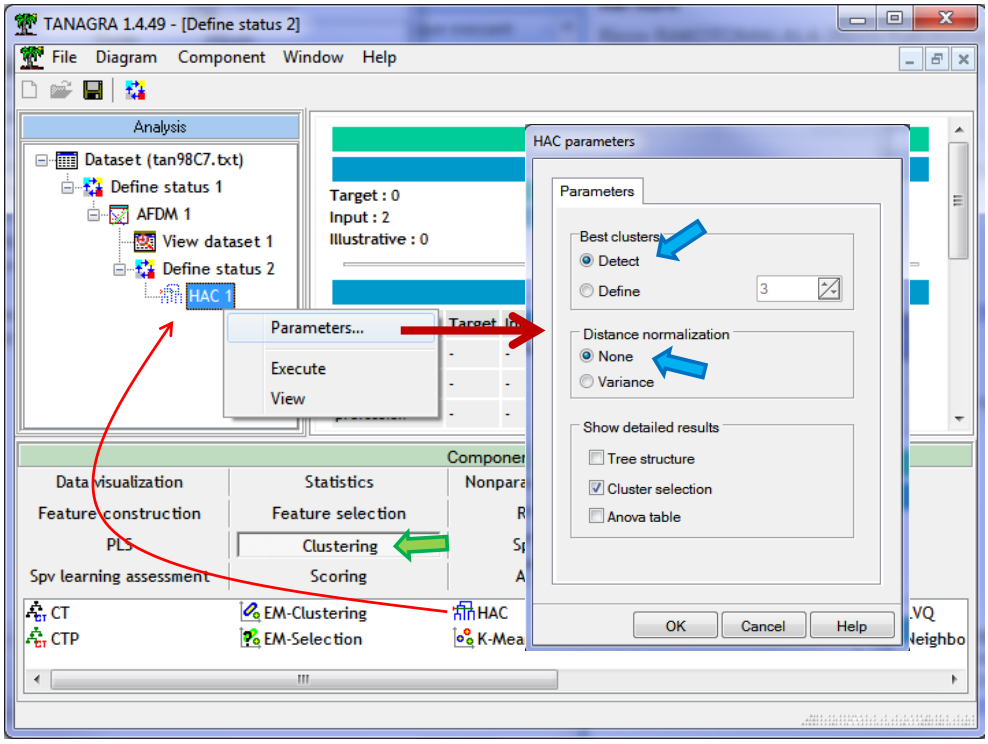
Figure 4 - Coordonnées des individus sur les 5 premiers facteurs

3.3 CAH à partir des facteurs pertinents

Nous insérons de nouveau le composant DEFINE STATUS pour spécifier les facteurs de l'AFDM à utiliser pour la classification ascendante hiérarchique.



Nous sélectionnons les 2 premiers AFDM_1_AXIS_1 et AFDM_1_AXIS_2. Nous ajoutons le composant HAC (« Hierarchical Agglomerative Clustering », onglet CLUSTERING), avec les paramètres :



Nous demandons à utiliser une distance non normalisée afin que les facteurs pèsent selon la part d'inertie rapportée par l'AFDM. Le nombre de clusters est détecté automatiquement (Tanagra recherche l'écart le plus élevé entre deux paliers consécutifs du dendrogramme, en ignorant la partition triviale en 2 classes). Nous lançons les calculs en actionnant le menu contextuel VIEW.

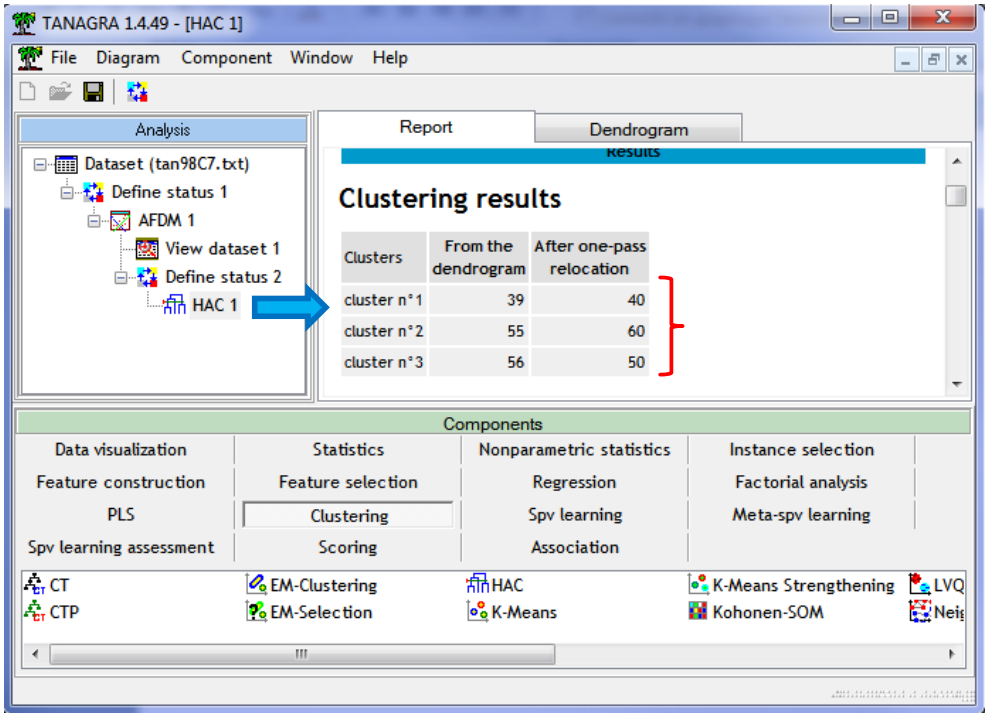
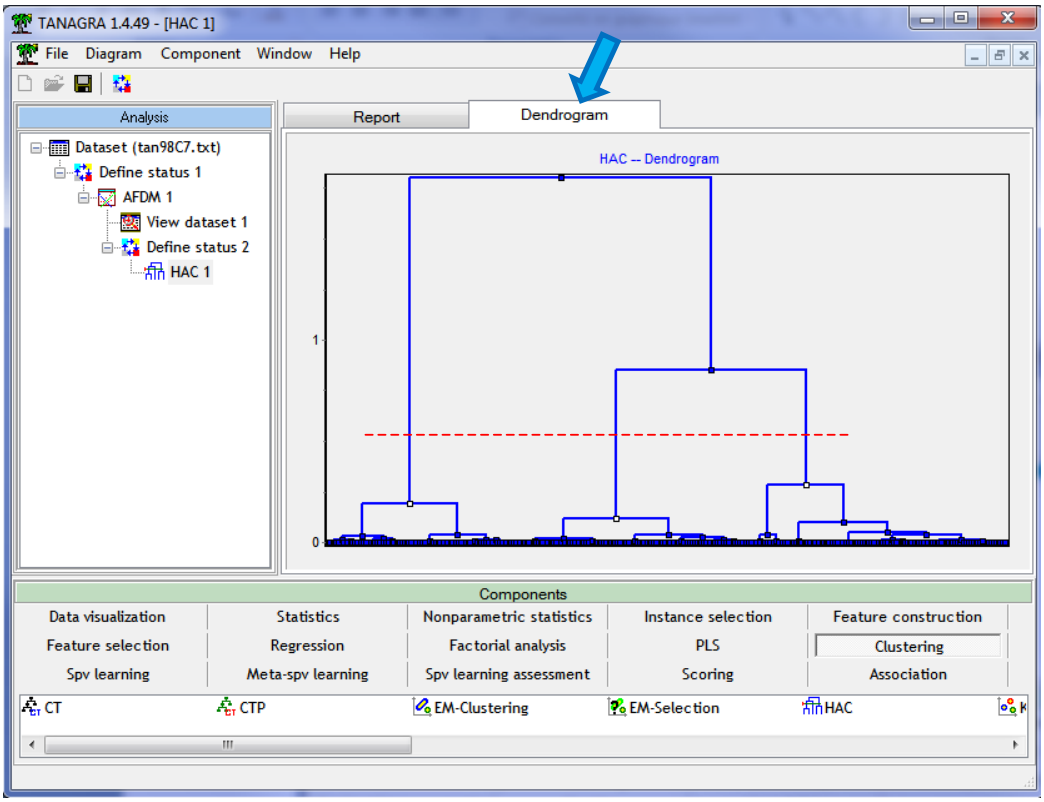


Figure 5 - Résultats de la CAH - Effectifs par classe avant et après réallocation

Tanagra produit 3 classes d'effectifs respectifs : 40, 60 et 50 individus. Les effectifs sont différents de ceux observés dans le dendrogramme car Tanagra, depuis la version 1.4.48, effectue une dernière passe sur les données pour affecter les individus aux barycentres de classes qui leur sont le plus proches[7]. L'objectif est d'obtenir des groupes avec une meilleure cohésion, la partition initiale étant contrainte par la structure hiérarchique de la recherche des solutions.

Le dendrogramme montre que la partition en 3 classes est la plus évidente (si l'on met de côté la subdivision en 2 groupes qui correspond quasiment toujours au saut le plus élevé dans l'arbre).



Dans la partie basse du rapport, Tanagra fournit les coordonnées des centres de classes. Cette information est importante pour le déploiement. En effet, on s'appuiera sur ces éléments pour rattacher les individus supplémentaires aux catégories.

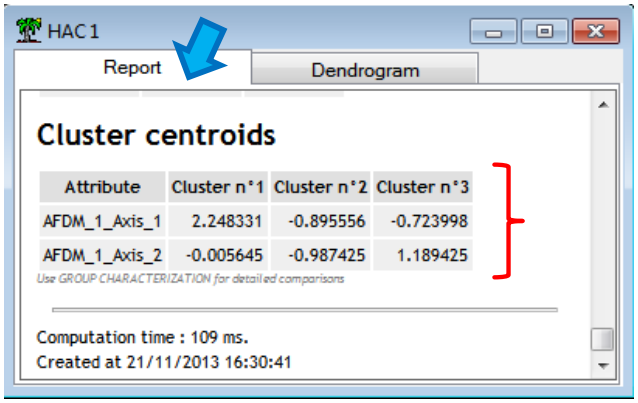


Figure 6 - Coordonnées des centres de classes

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

3.4 Positionnement des classes dans le plan factoriel

Pour mieux apprécier la qualité de la partition, nous visualisons les groupes dans le premier plan factoriel. Nous utilisons le composant SCATTERPLOT (onglet DATA VISUALIZATION). Nous plaçons AFDM_1_AXIS_1 en abscisse et AFDM_1_AXIS_2 en ordonnée. Nous illustrons les points à l'aide de la variable CLUSTER_HAC_1 générée automatiquement par le composant HAC.

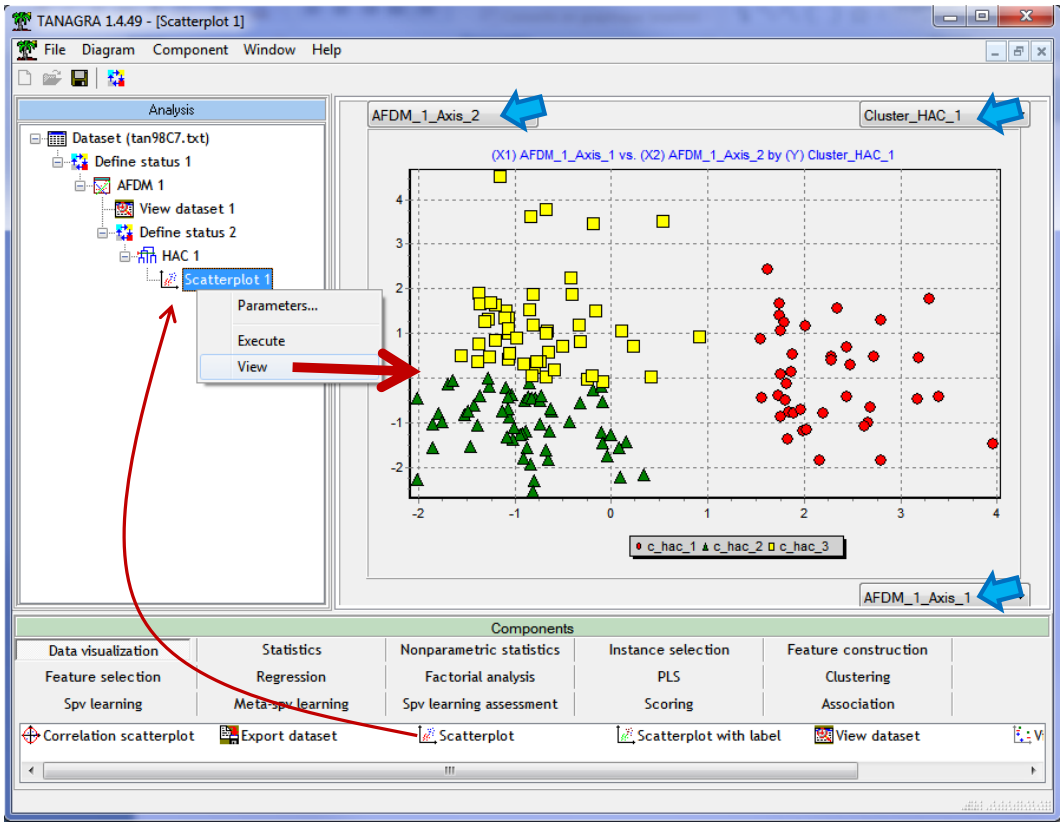


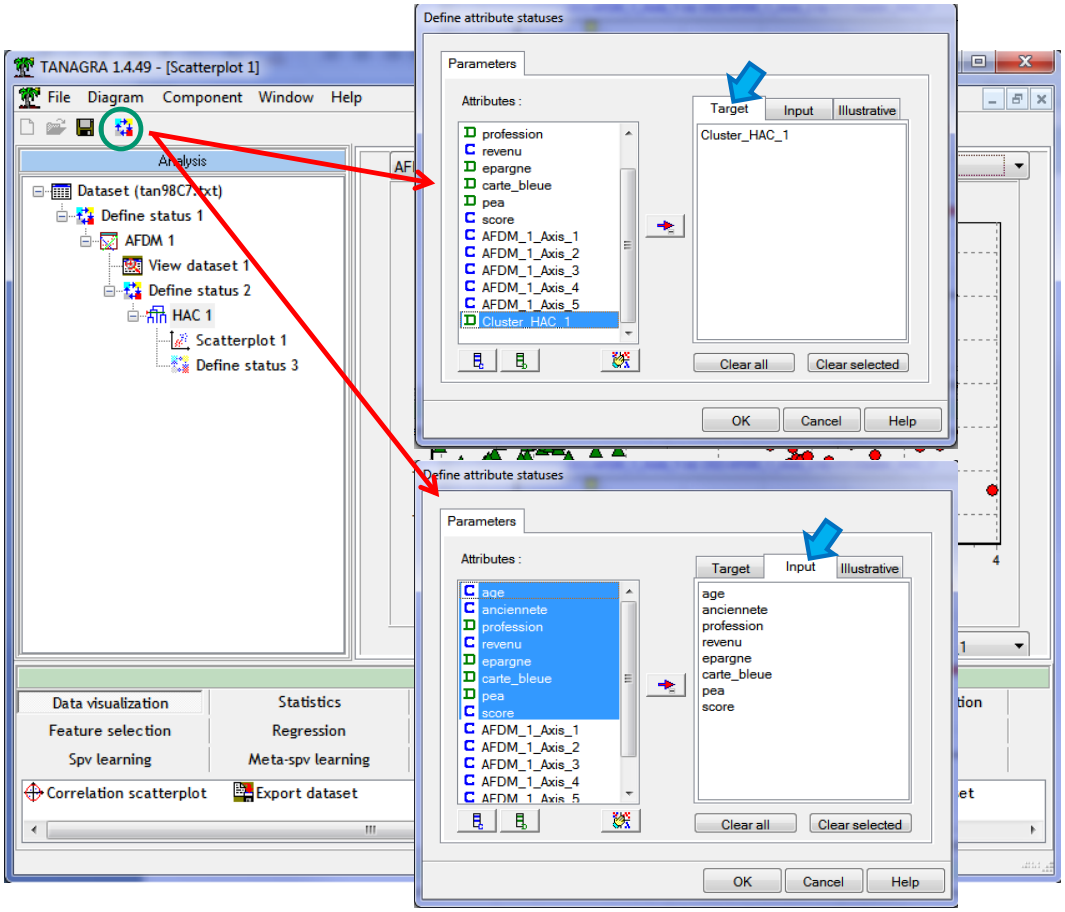
Figure 7 - Positionnement des groupes dans le premier plan factoriel

Nous observons nettement les 3 groupes d'observations mises en évidence par l'algorithme de classification. Le premier facteur permet d'isoler le premier cluster (C_HAC_1), le second permet de distinguer le second (C_HAC_2) du troisième (C_HAC_3). Ils sont parfaitement séparés – il n'y a pas d'empiètement entre les classes – dans le premier plan factoriel. Ce qui est tout à fait normal dans la mesure où nous n'avions utilisé que ces deux premiers facteurs pour la catégorisation. Enfin, si le 1er groupe est bien distinct, il y a quand même des chances que les 2nd et 3ème groupes présentent des traits relativement similaires.

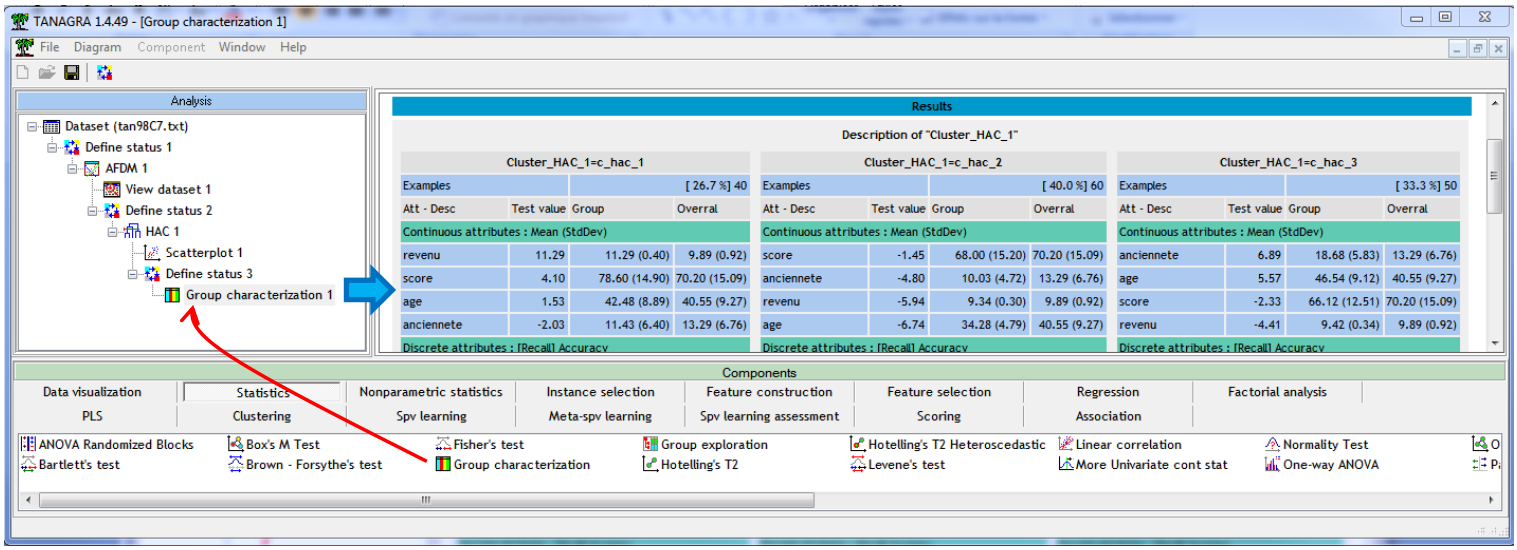
3.5 Caractérisation des classes – Variables actives et illustratives

Il s'agit justement de comprendre les caractéristiques sous jacentes aux classes, en utilisant d'une part les variables ayant participé à la construction de la typologie, d'autre part la variable supplémentaire SCORE indiquant l'appréciation du client par le conseiller clientèle.

Nous utilisons pour la 3ème fois le composant DEFINE STATUS. Nous plaçons en TARGET la variable CLUSTER_HAC_1. Elle associe chaque individu à la classe qui lui a été affectée. Nous mettons en INPUT toutes les variables de l'étude, y compris SCORE que nous utiliserons pour illustrer les groupes.



Nous insérons le composant GROUP CHARACTERIZATION (onglet STATISTICS). Le tableau est scindé en deux : pour les variables quantitatives (CONTINUOUS), les moyennes conditionnellement aux groupes sont comparées aux moyennes globales, calculées sur la totalité de l'échantillon ; pour les variables qualitatives (DISCRETE), les fréquences conditionnelles sont opposées aux fréquences globales.



Faisons un zoom sur le tableau fourni par le composant.

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 - Caractérisation des classes - Comparaison des moyennes et fréquences conditionnelles et globales

Détaillons le premier cluster pour bien préciser les idées :

- Le revenu moyen dans la population (en prenant en compte la totalité de l'échantillon) est de 9.89, avec un écart type de 0.92. Dans ce cluster, il devient 11.29 (avec un écart type de 0.40). De fait, les personnes du premier groupe présentent un revenu significativement plus élevé comme l'atteste la valeur test (TEST VALUE, dans le cas de la moyenne elle correspond à peu près au t de Student de comparaison de moyennes[8]) de 11.29.
- Le score moyen est de 70.20 dans la population, dans ce groupe il est de 78.60. Ces personnes sont mieux notées par les conseillers clientèle. Pas de surprise : plus on a de l'argent, plus on intéresse nos banquiers. Le contraire eut été très étonnant.
- Passons aux modalités des variables qualitatives. La proportion des cardes dans l'échantillon initial est de 26.7%. Dans ce groupe, il est de 100% (Accuracy). Toutes les personnes du groupe sont des cadres ! De plus, on constate que tous les cadres sont intégrés dans ce groupe (Recall = [100%]). La surreprésentation est très significative avec une valeur test de 12.21 (la valeur test correspond approximativement à la Z value de la comparaison de proportion pour les variables qualitatives).
- 24.7% des clients ont une épargne élevée [P(épargne = élevée) = 24.7%], dans ce groupe il passe à 35% [P(épargne = élevée / groupe = 1) = 35%]. Et 37.8% des personnes à épargne élevée se retrouvent dans ce groupe [P(appartenir au groupe = 1 / épargne = élevée) = 37.8%]. La surreprésentation est faiblement significative (valeur test = 1.76).

En étudiant le tableau de caractérisation des groupes, nous aboutissons à l'analyse suivante.

Groupe	Caractéristiques
Groupe 1	Ce groupe correspond aux personnes à revenu élevé [REVENU] qui intéressent la banque [SCORE]. Ce sont des clients assez récents [ANCIENNETE] qui sont exclusivement des cadres [PROFESSION = CAD], avec une épargne un peu plus élevée que la moyenne [EPARGNE = ELEVEE est plus fréquente]. Bref, on serait tenté de les voir comme des clients à fort potentiel envers lesquels il serait indiqué de promouvoir des produits financiers. On note par exemple que la pénétration du PEA n'y est pas décisive pour l'instant (42.5% dans ce groupe contre 38.7% globalement).
Groupe 2	Il s'agit de clients récents jeunes qui n'intéressent pas vraiment le banquier (valeur test de SCORE = -1.45). Les employés (profession = EMP) et les agriculteurs (profession = AGR, ils sont tous regroupés dans cette classe [100%]) y sont surreprésentés. Ils n'ont pas beaucoup d'épargne. Bref, il n'y a pas grand-chose à en tirer… pour l'instant.

[8] Voir http://tutoriels-data-mining.blogspot.fr/2008/04/interprter-la-valeur-test.html, les idées et formules sous-jacentes à la notion de valeur test y sont développées. Voir aussi l'article de référence d'Alain Morineau, http://www.deenov.com/analyse-de-donnees/article-valeur-test.aspx. *Très approximativement*, une valeur test supérieure à +2 ou inférieure à -2 indique un décalage significatif au risque 5%.

Groupe 3	Ce sont les clients historiques (âge élevé, ancienneté élevée) qui n'intéresse plus du tout le banquier (valeur test SCORE = -2.33). Ils ont des revenus faibles, mais se situe dans la moyenne concernant l'épargne (les proportions conditionnelles sont proches des proportions globales). Ce sont les braves personnes que la banque souvent « oublie », traversant les années sans jamais être contactés par leur conseiller clientèle[9].

Remarque : interprétation des groupes à partir des facteurs issus de l'AFDM. Les statistiques conditionnelles permettent de comprendre les caractéristiques des classes. Cette stratégie présente l'avantage de simplicité. Une autre piste de lecture serait d'interpréter les facteurs de l'AFDM en se référant au positionnement des groupes dans le repère factoriel. La démarche devient multivariée. Nous constatons par exemple que le premier cluster se distingue sur le 1er facteur (Figure 7, groupe rouge à droite). Nous notons que ce facteur est caractérisé par le revenu (corrélation positive de 0.960678, contribution de 42.8%) (Figure 1), la profession cadre (contribution PROFESSION CAD = 28.97%) et l'épargne élevée (contribution EPARGNE ELEVEE = 6.29%, par opposition l'EPARGNE FAIBLE = 5.29%) (Figure 2). Ces résultats sont parfaitement cohérents avec l'analyse univariée basée sur les comparaisons des moyennes et des proportions.

3.6 Déploiement – Classement d'un individu supplémentaire

Nous souhaitions rattacher un individu supplémentaire ω décrit comme suit à une des catégories :

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

Etape 1 : Calcul des coordonnées factorielles. Nous devons utiliser les « Factor Scores » fournis par l'AFDM pour calculer les coordonnées factorielles de l'individu (Figure 3). Attention, (1) nous devons utiliser les paramètres de centrage et de réduction ; (2) pour les variables qualitatives, l'indicatrice concernée prend la valeur 1, les autres 0.

Voyons le détail des opérations pour le premier facteur :

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

Nous obtenons les coordonnées (F_1 : -0.453, F_2 : 1.618). Ce qui placerait ce nouvel individu plutôt dans le 3ème cluster si l'on se réfère au nuage des observations dans le premier plan factoriel (Figure 9).

[9] J'affabule, j'affabule…. L'idée est de construire une « histoire » à partir des résultats de l'analyse exploratoire des données, rien de plus.

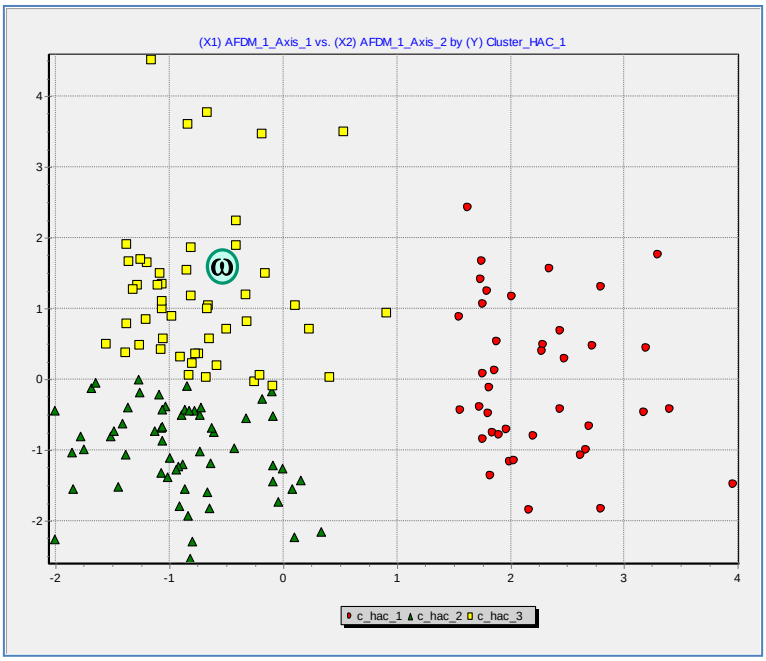


Figure 9 - Positionnement de l'individu supplémentaire ω dans le premier plan factoriel

Etape 2 : Calcul de la distance aux centres de classes G_k. Nous exploitons le tableau des coordonnées des barycentres conditionnels pour ce faire (Figure 6). Détaillons le calcul de la distance euclidienne pour le premier groupe :

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

Nous obtenons ainsi 3 valeurs : $d^2(G_1) = 9.934$; $d^2(G_2) = 6.985$; $d^2(G_3) = 0.257$. Manifestement, le barycentre du 3ème groupe est le plus proche (Figure 10). L'individu supplémentaire peut y être associé. Ce qui est logique au vu de ses caractéristiques : son âge et son ancienneté sont largement plus élevés que la moyenne, il fait partie des « vieux » clients de la banque.

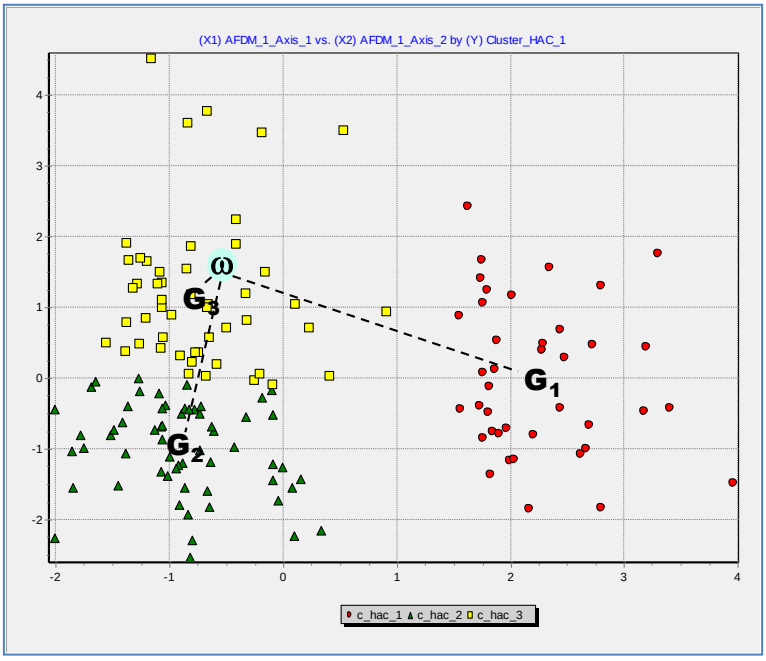


Figure 10 - Positionnement de l'individu supplémentaire ω par rapport aux centres de classes

4 Typologie sur données mixtes avec R

L'analyse factorielle des données mixtes est disponible dans plusieurs packages R. Il m'a semblé intéressant de les mettre en œuvre sur les mêmes données. Nous répétons la séquence AFDM [dudi.mix() avec le package ade4] + CAH [hclust() du package stats].

Voici le code source du programme :

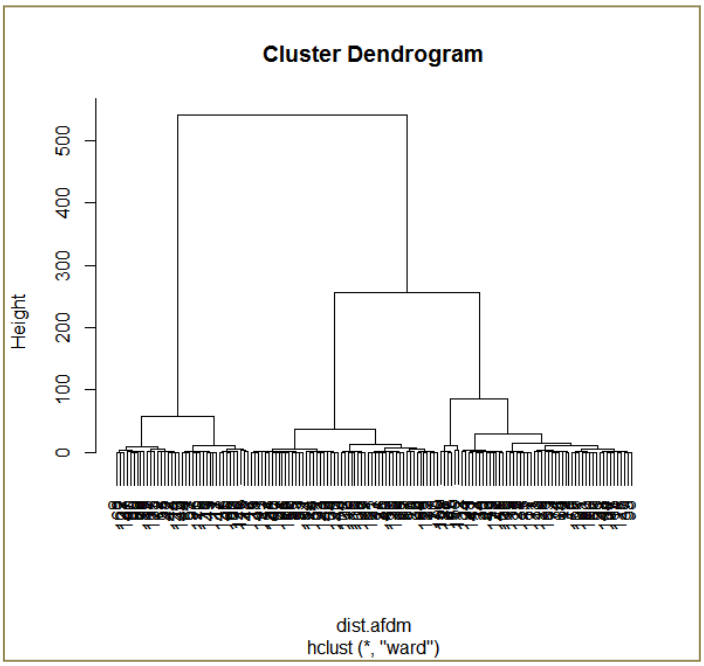
```
#chargement des données avec la librairie xlsx
library(xlsx)
bank <- read.xlsx(file="BankCustomer.xls",sheetIndex=1,header=T)

#stat. descriptives
summary(bank)

#variables actives
bank.active <- bank[,1:7]

#chargement du package ade4
library(ade4)

#AFDM, on retient les 2 premiers facteurs
bank.afdm <- dudi.mix(bank.active,scannf=F,nf=2)

#afficher les coordonnées factorielles des 5 premiers individus
print(head(bank.afdm$li,5))

#distance euclidenne entre les individus - 2 premiers facteurs
dist.afdm <- dist(bank.afdm$li[,1:2],method="euclidian")

#carré de la distance pour la méthode de Ward
#voir http://en.wikipedia.org/wiki/Ward's_method
dist.afdm <- dist.afdm^2

#CAH à partir de la matrice de distance
bank.tree <- hclust(dist.afdm,method="ward")
plot(bank.tree)

#découpage en 3 classes
bank.clusters <- cutree(bank.tree,k=3)

#distribution des classes
table(bank.clusters)

#projection des individus dans le plan
plot(bank.afdm$li[,1],bank.afdm$li[,2],col=c("red","yellow","green")[bank.clusters])

#positionnement des groupes par rapport au score
print(aggregate(x=bank$score,by=list(bank.clusters),FUN=mean))
```

Détaillons les résultats les plus importants du processus.

Coordonnées factorielles des individus. Nous observons les coordonnées factorielles des 5 premières observations. Les valeurs sont cohérentes par rapport à Tanagra (Figure 4), en revanche le signe est inversé pour le second facteur. Ce n'est pas un problème en soi. Le plus important est que les proximités entre les individus soient respectées.

```
> 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
```

Dendrogramme de la CAH. La CAH s'appuie sur la matrice de distance euclidienne des individus pris deux à deux. Nous utilisons la méthode Ward, également implémentée dans Tanagra. Le découpage en 3 classes est manifestement pertinent.

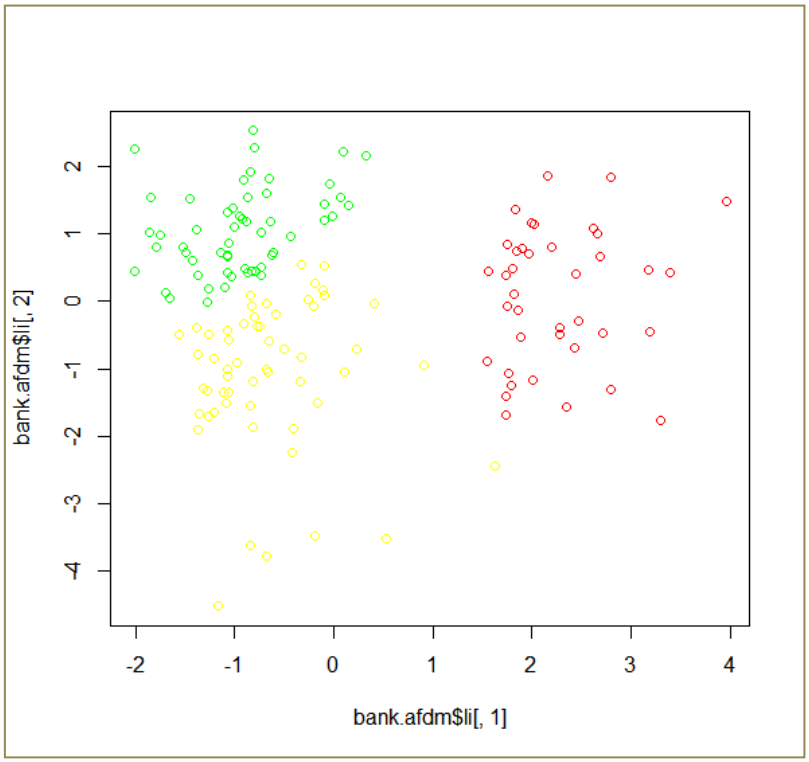


Découpage en 3 groupes, distribution. Nous réalisons une partition en 3 groupes (**cutree**), nous calculons les effectifs par classe avec **table**().

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

Notons que les valeurs diffèrent de celles de Tanagra parce que ce dernier, à la différence de cutree(), effectue une réallocation aux centres conditionnels les plus proches après la CAH afin de raffermir la cohésion des groupes ; avant réallocation, les partitions sont identiques (Figure 5).

Projection des groupes dans le plan factoriel. Du fait que les signes des valeurs soient inversés sur le second axe, les localisations des 2nd et 3ème classes sont interverties. Il n'en reste pas moins que les positions relatives et les formes générales des nuages sont bien identiques.



Moyennes de SCORE conditionnellement aux groupes. Nos utilisons la fonction **aggregate()** pour calculer les moyennes conditionnelles de la variable illustrative. La différenciation du SCORE selon le groupe d'appartenance est très perceptible encore une fois. Les valeurs sont très légèrement différentes de celle de Tanagra (Figure 8) parce que R travaille sur les clusters issus directement du dendrogramme, sans réallocation.

```
> 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

Avoir à traiter une base de données avec des variables hétérogène est une situation fréquemment rencontrée dans les études réelles. Dans ce tutoriel, nous présentons une approche basée sur la combinaison de l'analyse factorielle des données mixtes (AFDM) et la classification ascendante hiérarchique (CAH). Ce n'est certainement pas la solution miracle, d'autres pistes sont envisageables, le plus important est d'obtenir des résultats exploitables. De plus, la démarche est opérationnelle. Il est possible de classer automatiquement les individus supplémentaires.