3 Analyse Factorielle des Correspondances

3.1 Introduction

L'analyse factorielle des correspondances (AFC), ou analyse des correspondances simples, est une méthode exploratoire d'analyse des tableaux de contingence. Elle a été développé essentiellement par J.-P. Benzecri durant la période 1970-1990.

Soient deux variables nominales X et Y, comportant respectivement p et q modalités. On a observé les valeurs de ces variables sur une population et on dispose d'un tableau de contingence à p lignes et q colonnes donnant les effectifs conjoints c'est-à-dire les effectifs observés pour chaque combinaison d'une modalité i de X et d'une modalité j de Y.
Les valeurs de ce tableau seront notées n_{ij}, l'effectif total sera noté N.

L'ACP vise à analyser ce tableau en apportant des réponses à des questions telles que :

- Y a-t-il des lignes du tableau (modalités de X) qui se "ressemblent", c'est-à-dire telles que les distributions des modalités de Y soient analogues ?
- Y a-t-il des lignes du tableau (modalités de X) qui s'opposent, c'est-à-dire telles que les distributions des modalités de Y soient très différentes ?
- Mêmes questions pour les colonnes du tableau.
- Y a-t-il des associations modalité de X - modalité de Y qui s'attirent (effectif conjoint particulièrement élevé) ou qui se repoussent (effectif conjoint particulièrement faible) ?

La méthode se fixe également comme but de construire des représentations graphiques mettant en évidence ces propriétés des données.

3.2 Exemple

3.2.1 Enoncé

Réf. Examen de Statistiques de mai 2004, Module MULT, Maîtrise de Psychologie, Université René Descartes. Site Web : http://piaget.psycho.univ-paris5.fr/Statistiques/

Les données qui suivent sont constituées par les résultats du premier tour des élections régionales de 2004 pour la région Ile de France. Pour chacun des huit départements de l'Ile de France (en lignes), on a les effectifs de suffrages pour chacune des huit listes candidates ainsi que les effectifs d'abstentions (en colonnes). L'objectif est d'analyser la structure des votes ainsi que les liaisons entre listes et départements. Voici les codes de désignation des départements et des listes :

Départements	Code
Paris (75)	PARI
Seine et Marne (77)	SMAR
Yvelines (78)	YVEL
Essonne (91)	ESSO
Hauts de Seine (92)	HTSS
Seine Saint-Denis (93)	STDE
Val de Marne (94)	VDMA
Val d'Oise (95)	VDOI

Listes	Tête de liste	Code
PS-Verts-MRG-MRC	Huchon	HUCH
UMP	Copé	COPE
UDF	Santini	SANT
FN	Le Pen	LEPE
PC-AGR-AC	Buffet	BUFF
LO-LCR	Laguiller	LAGU
GE-Les Bleus	Pelegrin	PELE
MNR	Bay	BAY
Abstentions		ABST

Données : résultats du premier tour des régionales 2004 en Ile de France

	HUCHON	COPE	SANTINI	LEPEN	BUFFET	LAGU	PELEG	BAY	ABSTEN	TOTAL
PARI	258495	184419	114222	57183	39052	22479	13277	5006	434078	1128211
SMAR	128715	114003	48782	71897	25732	19738	11980	7085	301478	729410
YVEL	150141	140634	96746	61676	23292	15998	13939	6486	329626	838538
ESSO	144581	95451	59967	54309	26732	17545	12108	5346	270414	686453
HTSS	143444	136677	122610	47279	32987	16438	11322	4690	314964	830411
STDE	107327	61507	40081	54412	49535	19619	8393	5176	287618	633668
VDMA	126569	93049	60234	47074	41897	17308	10969	4557	286913	688570
VDOI	111176	82524	47903	55165	24693	17018	9876	4825	262458	615638
TOTAL	1170448	908264	590545	448995	263920	146143	91864	43171	2487549	6150899

Y a-t-il des départements qui se ressemblent, c'est-à-dire dans lesquels les résultats (en pourcentages) des différentes listes sont voisins ? Y a-t-il au contraire des départements qui s'opposent (résultats très différents) ?

Y a-t-il des départements dont les résultats sont proches de ceux de la région tout entière ? Y a-t-il des départements "à part" (dont les résultats s'écartent notablement de ceux de la région) ?

Y a-t-il des listes qui se ressemblent : elles n'obtiennent pas nécessairement les mêmes scores, mais les départements où elles obtiennent de bons scores sont les mêmes ? Y a-t-il des listes qui s'opposent ?

Y a-t-il des listes dont l'audience est la même dans tous les départements ? Y a-t-il des listes pour lesquelles le vote est concentré dans certains départements ?

Comment les départements "à part" et les listes à "vote concentré" s'associent-ils ?

3.2.2 Etude descriptive du tableau de contingence

On fixe les notations suivantes :

n_{ij} : effectif de la cellule (i,j),

$n_{i\bullet}$: effectif total de la ligne i,

$n_{\bullet j}$: effectif total de la colonne j

$n_{\bullet\bullet}$: effectif total

3.2.2.1 Tableau des fréquences

Les fréquences sont calculées par : $f_{ij} = \frac{n_{ij}}{n_{\bullet\bullet}} = \frac{Effectif\ \ de\ \ la\ \ cellule\ \ (i,j)}{Effectif\ \ total}$

	HUCHON	COPE	SANTINI	LEPEN	BUFFET	LAGU	PELEG	BAY	ABSTEN	TOTAL
PARI	4,20%	3,00%	1,86%	0,93%	0,63%	0,37%	0,22%	0,08%	7,06%	18,34%
SMAR	2,09%	1,85%	0,79%	1,17%	0,42%	0,32%	0,19%	0,12%	4,90%	11,86%
YVEL	2,44%	2,29%	1,57%	1,00%	0,38%	0,26%	0,23%	0,11%	5,36%	13,63%
ESSO	2,35%	1,55%	0,97%	0,88%	0,43%	0,29%	0,20%	0,09%	4,40%	11,16%
HTSS	2,33%	2,22%	1,99%	0,77%	0,54%	0,27%	0,18%	0,08%	5,12%	13,50%
STDE	1,74%	1,00%	0,65%	0,88%	0,81%	0,32%	0,14%	0,08%	4,68%	10,30%
VDMA	2,06%	1,51%	0,10%	0,77%	0,68%	0,28%	0,18%	0,07%	4,66%	11,19%
VDOI	1,81%	1,34%	0,78%	0,90%	0,40%	0,28%	0,16%	0,08%	4,27%	10,01%
TOTAL	19,03%	14,77%	9,60%	7,30%	4,29%	2,38%	1,49%	0,70%	40,44%	100,00%

3.2.2.2 Tableau des fréquences lignes

Les fréquences lignes (ou coordonnées des profils lignes) sont calculées par :

$$fl_{ij} = \frac{n_{ij}}{n_{i\bullet}} = \frac{f_{ij}}{f_{i\bullet}} = \frac{Effectif\ \ de\ \ la\ \ cellule\ \ (i,j)}{Effectif\ \ de\ \ la\ \ ligne\ \ i}$$

Les coordonnées du profil ligne moyen sont calculées par : $f_{\bullet j} = \frac{n_{\bullet j}}{n_{\bullet\bullet}} = \frac{Effectif\ \ de\ \ la\ \ colonne\ \ j}{Effectif\ \ total}$

	HUCHON	COPE	SANTINI	LEPEN	BUFFET	LAGU	PELEG	BAY	ABSTEN	TOTAL
PARI	22,91%	16,35%	10,12%	5,07%	3,46%	1,99%	1,18%	0,44%	38,47%	100,00%
SMAR	17,65%	15,63%	6,69%	9,86%	3,53%	2,71%	1,64%	0,97%	41,33%	100,00%
YVEL	17,91%	16,77%	11,54%	7,36%	2,78%	1,91%	1,66%	0,77%	39,31%	100,00%
ESSO	21,06%	13,90%	8,74%	7,91%	3,89%	2,56%	1,76%	0,78%	39,39%	100,00%
HTSS	17,27%	16,46%	14,76%	5,69%	3,97%	1,98%	1,36%	0,56%	37,93%	100,00%
STDE	16,94%	9,71%	6,33%	8,59%	7,82%	3,10%	1,32%	0,82%	45,39%	100,00%
VDMA	18,38%	13,51%	8,75%	6,84%	6,08%	2,51%	1,59%	0,66%	41,67%	100,00%
VDOI	18,06%	13,40%	7,78%	8,96%	4,01%	2,76%	1,60%	0,78%	42,63%	100,00%
TOTAL	19,03%	14,77%	9,60%	7,30%	4,29%	2,38%	1,49%	0,70%	40,44%	100,00%

3.2.2.3 Tableau des fréquences colonnes

Les fréquences colonnes (ou coordonnées des profils colonnes) sont calculées par :

$$fc_{ij} = \frac{n_{ij}}{n_{\bullet j}} = \frac{f_{ij}}{f_{\bullet j}} = \frac{Effectif\ \ de\ \ la\ \ cellule\ \ (i,j)}{Effectif\ \ de\ \ la\ \ colonne\ \ j}$$

Les coordonnées du profil colonne moyen sont calculées par : $f_{i\bullet} = \frac{n_{i\bullet}}{n_{\bullet\bullet}} = \frac{Effectif\ \ de\ \ la\ \ ligne\ \ i}{Effectif\ \ total}$

	HUCHON	COPE	SANTINI	LEPEN	BUFFET	LAGU	PELEG	BAY	ABSTEN	TOTAL
PARI	22,09%	20,30%	19,34%	12,74%	14,80%	15,38%	14,45%	11,60%	17,45%	18,34%
SMAR	11,00%	12,55%	8,26%	16,01%	9,75%	13,51%	13,04%	16,41%	12,12%	11,86%
YVEL	12,83%	15,48%	16,38%	13,74%	8,83%	10,95%	15,17%	15,02%	13,25%	13,63%
ESSO	12,35%	10,51%	10,15%	12,10%	10,13%	12,01%	13,18%	12,38%	10,87%	11,16%
HTSS	12,26%	15,05%	20,76%	10,53%	12,50%	11,25%	12,32%	10,86%	12,66%	13,50%
STDE	9,17%	6,77%	6,79%	12,12%	18,77%	13,42%	9,14%	11,99%	11,56%	10,30%
VDMA	10,81%	10,24%	10,20%	10,48%	15,87%	11,84%	11,94%	10,56%	11,53%	11,19%
VDOI	9,50%	9,09%	8,11%	12,29%	9,36%	11,64%	10,75%	11,18%	10,55%	10,01%
TOTAL	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%

3.2.2.4 Distances entre profils. Métrique du Φ^2

Chaque ligne du tableau des fréquences lignes peut être vue comme la liste des coordonnées d'un point dans un espace à q dimensions. On obtient ainsi le nuage des individus-lignes. On définit de même le nuage des individus-colonnes à partir du tableau des fréquences colonnes.

Comme en ACP, on s'intéresse alors aux directions de "plus grande dispersion" de chacun de ces nuages de points. Mais, pour mesurer la "distance" entre deux individus, on utilise la *métrique du* Φ^2 au lieu de la distance habituelle (dite *métrique euclidienne*). La distance du Φ^2 entre la ligne i et la ligne i' est ainsi définie par :

$$d^2_{\Phi^2}(L_i, L_{i'}) = \sum_j \frac{(fl_{ij} - fl_{i'j})^2}{f_{\bullet j}}$$

Pourquoi utiliser cette métrique plutôt que la métrique euclidienne ? Deux raisons fortes peuvent être avancées :

- Avec la métrique du Φ^2, la distance entre deux lignes ne dépend pas des poids respectifs des colonnes. Ainsi, sur notre exemple, les différentes listes obtiennent des scores très différents et l'usage de la métrique euclidienne aurait donné trop de poids aux listes qui ont obtenu des scores élevés (ABST, HUCH, COPE).

- La métrique du Φ^2 possède la propriété d'*équivalence distributionnelle* : si on regroupe deux modalités lignes, les distances entre les profils-colonne, ou entre les autres profils-lignes restent inchangées.

Par exemple, la distance entre la ligne PARI et la ligne SMAR est donnée par :

$$d^2_{\Phi^2}(PARI,SMAR) = \frac{(0,2291-0,1765)^2}{0,1903} + \ldots + \frac{(0,3847-0,4133)^2}{0,4044} = 0,0682$$

La distance entre PARI et le profil-ligne moyen est donnée par :

$$d^2_{\Phi^2}(PARI,Moyenne) = \frac{(0,2291-0,1903)^2}{0,1903} + \ldots + \frac{(0,3847-0,4044)^2}{0,4044} = 0,0215$$

Avec les transpositions nécessaires, ce qui vient d'être dit pour les lignes s'applique également aux colonnes. Par exemple, la distance entre la colonne BUFFET et la colonne SANTINI est :

$$d^2_{\Phi^2}(BUFFET,SANTINI) = \frac{(0,1480-0,1934)^2}{0,1834} + \ldots + \frac{(0,0936-0,0811)^2}{0,1001} = 0,2753$$

Notons qu'en revanche, il n'existe pas d'outil mesurant une "distance" entre une ligne et une colonne.

3.2.2.5 Taux de liaison et Phi-2

Les taux de liaison sont définis par : $t_{ij} = \frac{f_{ij} - f_{i\bullet}.f_{\bullet j}}{f_{i\bullet}.f_{\bullet j}}$

	HUCHON	COPE	SANTINI	LEPEN	BUFFET	LAGU	PELEG	BAY	ABSTEN
PARI	0,204	0,107	0,054	-0,306	-0,193	-0,161	-0,212	-0,368	-0,049
SMAR	-0,073	0,058	-0,303	0,350	-0,178	0,139	0,100	0,384	0,022
YVEL	-0,059	0,136	0,202	0,008	-0,353	-0,197	0,113	0,102	-0,028
ESSO	0,107	-0,058	-0,090	0,084	-0,092	0,076	0,181	0,110	-0,026
HTSS	-0,092	0,115	0,538	-0,220	-0,074	-0,167	-0,087	-0,195	-0,062
STDE	-0,110	-0,343	-0,341	0,176	0,822	0,303	-0,113	0,164	0,122
VDMA	-0,034	-0,085	-0,089	-0,063	0,418	0,058	0,067	-0,057	0,030
VDOI	-0,051	-0,092	-0,190	0,228	-0,065	0,163	0,074	0,117	0,054

Signification pratique du taux de liaison : le score de la liste Huchon à Paris est 20% plus élevé que le score théorique que l'on observerait si les votes étaient indépendants des départements. Au contraire, celui de la liste Le Pen est 30% moins élevé que le score théorique.
Par construction, les valeurs prises par le taux de liaison sont :

- des nombres positifs quelconques (un score observé peut être 200% ou 300% supérieur au score théorique)
- des nombres négatifs compris entre -1 et 0 (le "déficit" le plus extrême d'un score observé est d'être 100% moins élevé que le score théorique).

Notez que le coefficient $f_{i\bullet}.f_{\bullet j}$ représente le "poids théorique" de chaque cellule dans le tableau. La somme de ces coefficients vaut 1.
La moyenne de la série des taux de liaison pondérée par les coefficients $f_{i\bullet}.f_{\bullet j}$ est nulle. La variance de cette série (avec les mêmes pondérations) est le coefficient Φ^2 :

$$\Phi^2 = \sum_{i,j} f_{i\bullet}.f_{\bullet j}.t^2_{ij} = \sum_{i,j} \frac{(f_{ij} - f_{i\bullet}.f_{\bullet j})^2}{f_{i\bullet}.f_{\bullet j}} = \frac{X^2}{n_{\bullet\bullet}}$$

Ici, on obtient : $\Phi^2 = 0,02379$.
La méthode d'analyse factorielle des correspondances peut être vue comme une décomposition pertinente du Φ^2 selon plusieurs axes factoriels.

3.2.3 L'analyse factorielle des correspondances proprement dite

L'application de la méthode a deux effets :

- d'une part, on construit des images des nuages d'"individus-lignes" et d'"individus-colonnes" de départ, de façon que les distances entre images soient des distances euclidiennes et non plus des distances calculées selon la métrique du Φ^2;
- d'autre part, on recherche les directions de plus grande dispersion dans ces nuages de points images.

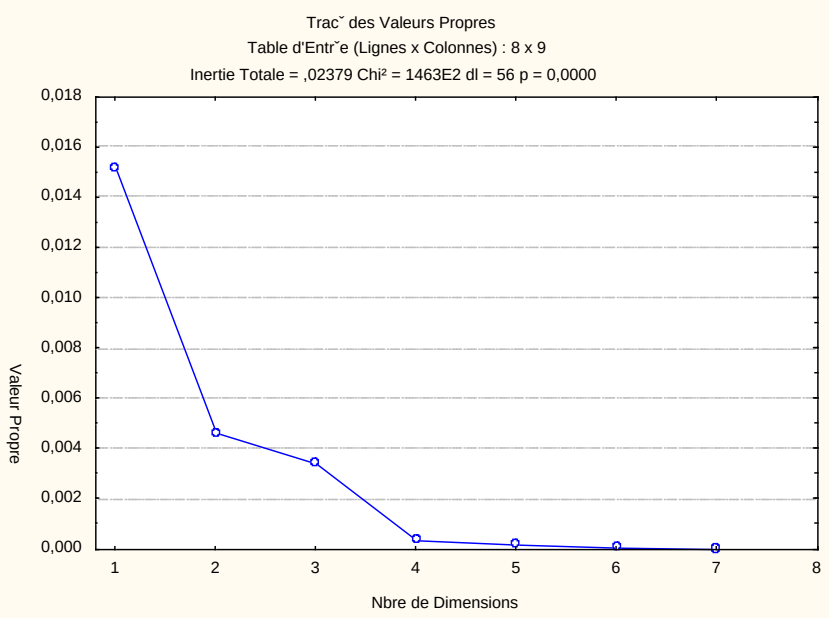
La matrice (tableau de valeurs) dont on recherche les valeurs propres et vecteurs propres est un objet mathématique "compliqué", qui ne possède pas de signification intuitive immédiate. De fait, on part de la matrice dont le terme à l'intersection de la ligne i et de la colonne j vaut : $\frac{f_{ij}}{\sqrt{f_{i\bullet}.f_{\bullet j}}}$ et on calcule des produits scalaires entre lignes (ou entre colonnes) de cette matrice.

3.2.3.1 Valeurs propres

Le nombre de valeurs propres produites par la recherche des facteurs principaux est égal au minimum du nombre de lignes et du nombre de colonnes du tableau de contingence. Cependant, la première valeur propre est systématiquement égale à 1, et n'est pas mentionnée dans les résultats. Les autres valeurs propres sont des nombres positifs inférieurs à 1 et leur somme est égale à Φ^2.

Valeurs Propres et Inertie de toutes les Dimensions (idf.sta)
Inertie Totale = ,02379 Chi^2 = 146308 dl = 56 p = 0,0000

Dimension	ValSing.	ValProp.	%age Inertie	%age Cumulé	Chi^2
1	0,122976	0,015123	63,58	63,58	93020
2	0,068237	0,004656	19,58	83,15	28640
3	0,058363	0,003406	14,32	97,47	20951
4	0,018685	0,000349	1,47	98,94	2147
5	0,012321	0,000152	0,64	99,58	934
6	0,008701	0,000076	0,32	99,90	466
7	0,004936	0,000024	0,10	100,00	150



Le choix du nombre d'axes factoriels à conserver se fait comme dans le cas de l'ACP. Ici, on observe une brusque décroissance des valeurs propres entre la 3è et la 4è valeur propre. On retient donc les 3 premiers axes factoriels.

3.2.3.2 Résultats relatifs aux individus-lignes

Coordonnées Ligne et Contributions à l'Inertie (idf.sta)
Standardisation : Profils ligne et colonne

	Ligne N°	Coord. Dim.1	Coord. Dim.2	Coord. Dim.3	Masse	Qualité	Inertie Relative	Inertie Dim.1	Cos^2 Dim.1	Inertie Dim.2	Cos^2 Dim.2	Inertie Dim.3	Cos^2 Dim.3
PARI	1	-0,1050	0,0027	0,1016	0,1834	0,9924	0,1659	0,1337	0,5122	0,0003	0,0003	0,5561	0,4799
SMAR	2	0,0821	-0,1181	-0,0332	0,1186	0,9678	0,1122	0,0528	0,2993	0,3550	0,6194	0,0385	0,0491
YVEL	3	-0,0960	-0,0397	-0,0555	0,1363	0,9811	0,0810	0,0830	0,6517	0,0461	0,1113	0,1234	0,2181
ESSO	4	0,0183	-0,0393	0,0355	0,1116	0,5895	0,0249	0,0025	0,0627	0,0369	0,2898	0,0413	0,2369
HTSS	5	-0,1586	0,0824	-0,0752	0,1350	0,9966	0,2140	0,2245	0,6668	0,1967	0,1799	0,2239	0,1498
STDE	6	0,2478	0,0954	-0,0017	0,1030	0,9979	0,3061	0,4184	0,8691	0,2014	0,1288	0,0001	0,0000
VDMA	7	0,0706	0,0667	0,0115	0,1119	0,9225	0,0488	0,0369	0,4803	0,1071	0,4293	0,0044	0,0128
VDOI	8	0,0854	-0,0513	-0,0206	0,1001	0,9264	0,0470	0,0482	0,6529	0,0565	0,2356	0,0124	0,0379

Le tableau ci-dessus rassemble tous les résultats relatifs aux individus-lignes.

La colonne "Masse" rappelle les fréquences marginales des lignes c'est-à-dire le profil colonne moyen. Contrairement à l'ACP normée, dans laquelle chaque individu était affecté du même poids, les départements ont ici un "poids" dépendant de l'effectif total d'électeurs inscrits dans le département.

La colonne "Qualité" indique les qualités de représentation des individus ligne par les trois premiers axes principaux. Ces qualités sont calculées par des formules du type (Li désigne ici la ligne N°i, Fj, le facteur principal N°j) :

$$QLT(L_i, F_1; F_2; F_3) = \frac{(Coord\,de\,L_i\,selon\,F_1)^2 + (Coord\,de\,L_i\,selon\,F_2)^2 + (Coord\,de\,L_i\,selon\,F_3)^2}{\sum_l (Coord\,de\,L_i\,selon\,F_l)^2}$$

Par exemple :

$$QLT(PARIS, F_1; F_2; F_3) = \frac{(-0{,}1050)^2 + (0{,}0027)^2 + (0{,}1016)^2}{(-0{,}1050)^2 + (0{,}0027)^2 + (0{,}1016)^2 + (0{,}0107)^2 + (0{,}0068)^2 + (-0{,}0017)^2 + (-0{,}0007)^2}$$

La colonne "Inertie relative" est calculée de la manière suivante :

- L'inertie d'une combinaison individu-ligne individu-colonne correspondant à une cellule du tableau de contingence est le carré du taux de liaison, multiplié par la pondération (fréquence-ligne x fréquence colonne) correspondante.
- L'inertie absolue d'un individu-ligne est la somme des inerties des cellules de la ligne
- L'inertie relative d'un individu ligne est obtenue en divisant l'inertie absolue de l'individu par la somme de toutes les inerties, c'est-à-dire par Φ^2.

Pour chacun des trois axes factoriels, le tableau nous donne également les coordonnées ou *scores factoriels* de l'individu-ligne selon cet axe. Ces coordonnées ont les propriétés suivantes :

- Selon chaque axe, la moyenne des coordonnées des individus-lignes pondérées par les masses, est nulle.
- Selon chaque axe, la moyenne des carrés des coordonnées des individus-lignes pondérées par les masses, est égale à la valeur propre correspondante.
- Les coordonnées selon deux axes différents, pondérées par les masses, forment deux séries statistiques indépendantes (covariance nulle)

Ainsi :

$$(-0{,}1050 \times 0{,}1834) + (0{,}0821 \times 0{,}1186) + \ldots + (0{,}0854 \times 0{,}1001) = 0$$

$$(-0{,}1050)^2 \times 0{,}1834 + (0{,}0821)^2 \times 0{,}1186 + \ldots + (0{,}0854)^2 \times 0{,}1001 = 0{,}015123$$
$$(-0{,}1050) \times (0{,}0027) \times 0{,}1834 + (0{,}0821) \times (-0{,}1181) \times 0{,}1186 + \ldots + (0{,}0854) \times (-0{,}0513) \times 0{,}1001 = 0$$

Le tableau donne également la contribution de chaque individu à la formation de l'axe, ou inertie selon cet axe. Cette valeur est définie par :

$$Ctr(L_i, F_k) = \frac{(Masse\ \ L_i) \times (Coord\ \ L_i\ \ selon\ \ F_k)^2}{Valeur\ \ propre\ \ relative\ \ à\ \ F_k}$$

Par exemple, pour Paris et l'axe factoriel N°1 :

$$Ctr(PARIS, F_1) = \frac{0{,}1834 \times (-0{,}1050)^2}{0{,}0151} = 0{,}1337$$

Ces valeurs sont des contributions relatives (la somme de la colonne vaut 1). On peut donc utiliser des colonnes pour rechercher quels sont les individus-lignes qui ont eu une influence supérieure à la moyenne dans la formation de l'axe factoriel considéré.

Enfin, ce tableau nous donne les cosinus-carrés ou qualités de représentation des individus-lignes par chaque axe factoriel. Ces valeurs sont définies par :

$$QLT(L_i, F_k) = \frac{(Coord\ de\ L_i\ selon\ F_k)^2}{\sum_l (Coord\ de\ L_i\ selon\ F_l)^2}$$

Par exemple :

$$QLT(PARIS, F_1) = \frac{(-0{,}1050)^2}{(-0{,}1050)^2 + (0{,}0027)^2 + (0{,}1016)^2 + (0{,}0107)^2 + (0{,}0068)^2 + (-0{,}0017)^2 + (-0{,}0007)^2} = 0{,}5122$$

L'interprétation géométrique de ces valeurs est analogue à celle développée pour l'ACP : c'est le carré du cosinus de l'angle du vecteur représentant "PARIS" dans l'espace à 7 dimensions de sa projection sur le premier axe factoriel.

3.2.3.3 Résultats relatifs aux individus-colonnes

Dans une AFC, les individus-lignes et les individus-colonnes jouent des rôles symétriques. Les résultats relatifs aux individus-colonnes s'interprètent donc de la même façon que les résultats relatifs aux individus-lignes.

Coordonnées Colonne et Contributions à l'Inertie (idf.sta)
Standardisation : Profils ligne et colonne

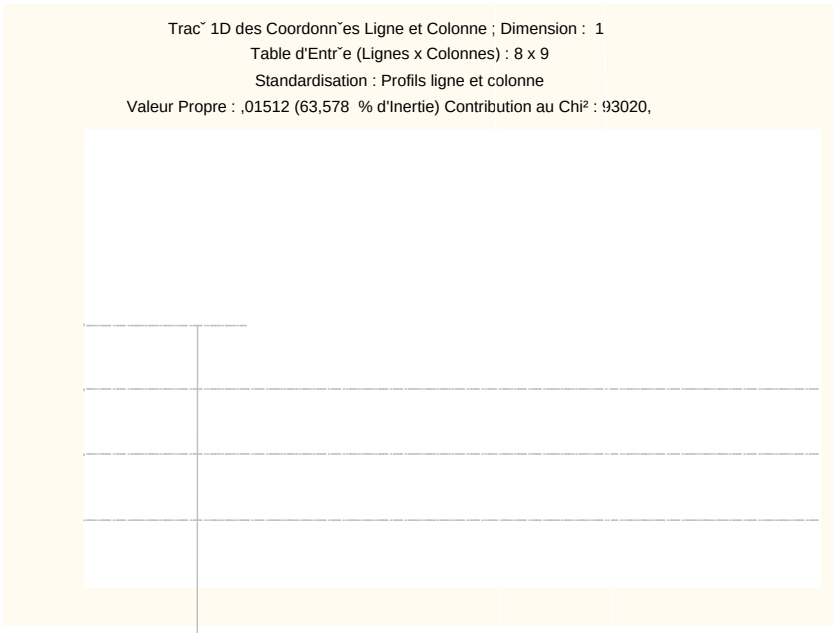
	Colon-ne N°	Coord. Dim.1	Coord. Dim.2	Coord. Dim.3	Masse	Qualité	Inertie Relative	Inertie Dim.1	Cos2 Dim.1	Inertie Dim.2	Cos2 Dim.2	Inertie Dim.3	Cos2 Dim.3
HUCHON	1	-0,0421	-0,0165	0,1024	0,1903	0,9795	0,1024	0,0223	0,1388	0,0112	0,0214	0,5857	0,8194
COPE	2	-0,1305	-0,0513	-0,0089	0,1477	0,9432	0,1299	0,1663	0,8139	0,0833	0,1256	0,0034	0,0038
SANTINI	3	-0,2388	0,0955	-0,0822	0,0960	0,9928	0,2964	0,3622	0,7768	0,1880	0,1241	0,1903	0,0919
LEPEN	4	0,1628	-0,1146	-0,0883	0,0730	0,9912	0,1469	0,1279	0,5539	0,2059	0,2745	0,1670	0,1628
BUFFET	5	0,2581	0,2259	0,0117	0,0429	0,9912	0,2144	0,1890	0,5605	0,4704	0,4295	0,0017	0,0012
LAGU	6	0,1655	-0,0084	-0,0066	0,0238	0,9401	0,0292	0,0431	0,9362	0,0004	0,0024	0,0003	0,0015
PELEG	7	0,0332	-0,0714	-0,0625	0,0149	0,5543	0,0114	0,0011	0,0604	0,0163	0,2796	0,0171	0,2144
BAY	8	0,1514	-0,1198	-0,1211	0,0070	0,9497	0,0161	0,0106	0,4190	0,0216	0,2626	0,0302	0,2681
ABSTEN	9	0,0538	0,0058	-0,0059	0,4044	0,9470	0,0533	0,0775	0,9251	0,0029	0,0106	0,0042	0,0113

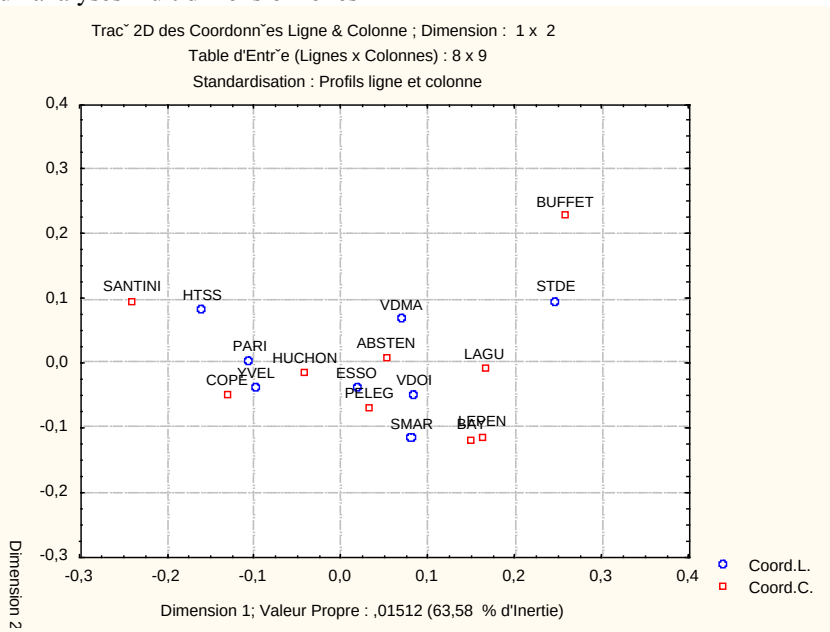
3.2.3.4 Résultats graphiques

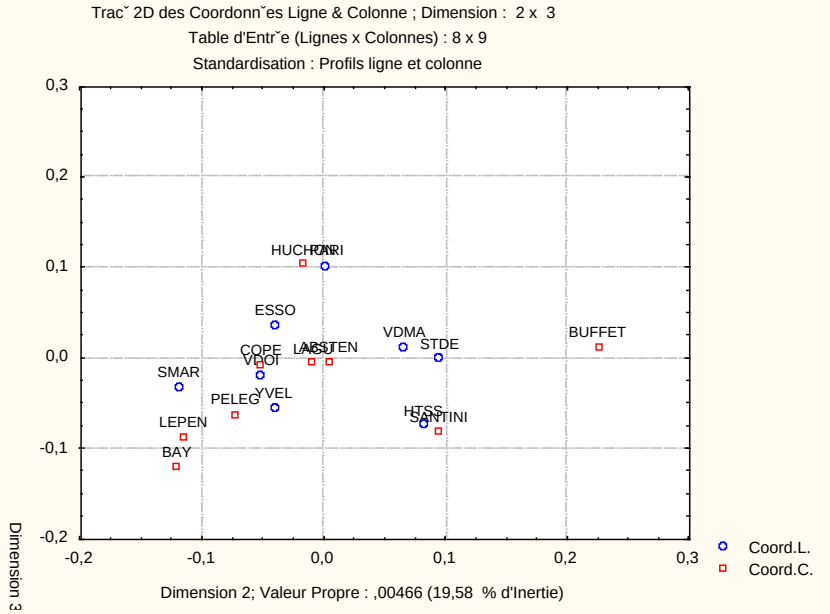
Les transformations et les pondérations introduites rendent tout à fait comparables les valeurs obtenues pour les individus lignes et les individus colonnes. Contrairement à l'ACP, les graphiques factoriels pourront être construits en faisant figurer sur un même graphique les individus lignes et les individus colonnes.

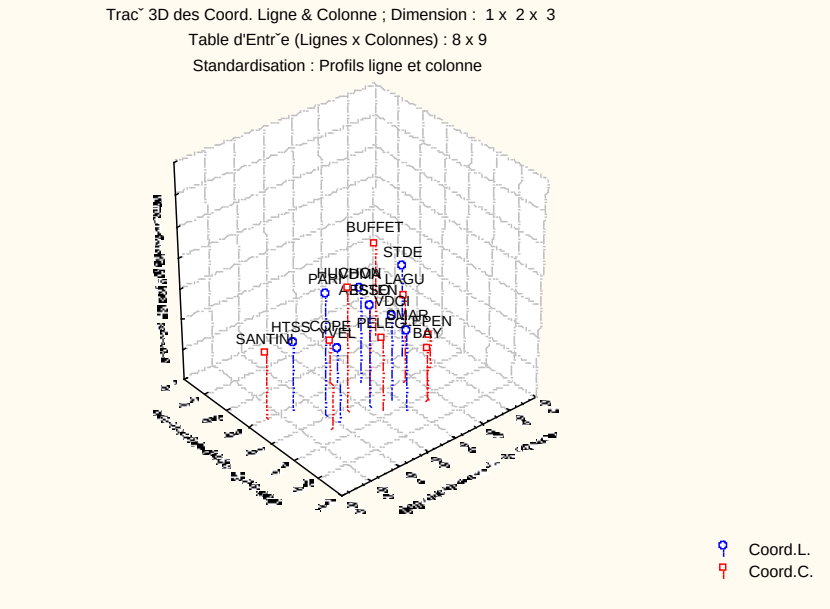
On peut réaliser et essayer d'interpréter des graphiques :

- en dimension 1 : on place les individus le long d'un axe factoriel,
- en dimension 2 : on place les individus dans un plan défini à partir de deux axes factoriels,
- éventuellement, en dimension 3 :on place les individus dans une représentation en perspective d'un espace à 3 dimensions.




Trac˘ 2D des Coordonn˘es Ligne & Colonne ; Dimension : 1 x 2
Table d'Entr˘e (Lignes x Colonnes) : 8 x 9
Standardisation : Profils ligne et colonne
BUFFET
SANTINI
HTSS
STDE
VDMA
PARI
ABSTEN
HUCHON
LAGU
YVEL
ESSO
COPE
VDOI
PELEG
SMAR
LEPEN
BAY
Dimension 2
Dimension 1; Valeur Propre : ,01512 (63,58 % d'Inertie)
Coord.L.
Coord.C.


Trac˘ 2D des Coordonn˘es Ligne & Colonne ; Dimension : 2 x 3
Table d'Entr˘e (Lignes x Colonnes) : 8 x 9
Standardisation : Profils ligne et colonne
ESSO
VDMA
BUFFET
STDE
SMAR
PELEG
YVEL
HTSS
SANTINI
LEPEN
BAY
Dimension 3
Dimension 2; Valeur Propre : ,00466 (19,58 % d'Inertie)
Coord.L.
Coord.C.


Trac˘ 3D des Coord. Ligne & Colonne ; Dimension : 1 x 2 x 3
Table d'Entr˘e (Lignes x Colonnes) : 8 x 9
Standardisation : Profils ligne et colonne
BUFFET
STDE
LAGU
VDOI
SMAR
LEPEN
BAY
SANTINI
Coord.L.
Coord.C.

3.2.3.5 Interprétation géométrique

Les distances entre deux individus-lignes, ou entre un indivivu-ligne et l'origine des axes, peuvent être facilement interprétées. En effet : la distance euclidienne entre deux points-lignes, représentés par leurs coordonnées factorielles est égale à la distance du Φ^2 entre les profils-lignes initiaux.

Par exemple, nous avions vu que :

$$d^2_{\Phi^2}(PARI, SMAR) = \frac{(0{,}2291 - 0{,}1765)^2}{0{,}1903} + \ldots + \frac{(0{,}3847 - 0{,}4133)^2}{0{,}4044} = 0{,}0682$$

Or, le tableau (complet) des scores factoriels des lignes est :

	Facteur 1	Facteur 2	Facteur 3	Facteur 4	Facteur 5	Facteur 6	Facteur 7
PARI	-0,1050	0,0027	0,1016	0,0107	-0,0068	-0,0017	-0,0007
SMAR	0,0821	-0,1181	-0,0332	0,0231	0,0077	-0,0115	-0,0004
YVEL	-0,0960	-0,0397	-0,0555	0,0029	-0,0015	0,0148	-0,0062
ESSO	0,0183	-0,0393	0,0355	-0,0442	0,0149	-0,0026	-0,0016
HTSS	-0,1586	0,0824	-0,0752	-0,0042	-0,0011	-0,0104	0,0019
STDE	0,2478	0,0954	-0,0017	0,0006	-0,0096	-0,0026	-0,0070
VDMA	0,0706	0,0667	0,0115	0,0151	0,0208	0,0100	0,0066
VDOI	0,0854	-0,0513	-0,0206	-0,0134	-0,0231	0,0050	0,0092

On vérifie que :

$$d^2_{eucl}(PARI', SMAR') = (-0{,}1050 - 0{,}0821)^2 + \ldots + (-0{,}0007 + 0{,}0004)^2 = 0{,}0682$$

De même, on avait établi que :

$$d^2_{\Phi^2}(PARI, Moyenne) = \frac{(0{,}2291 - 0{,}1903)^2}{0{,}1903} + \ldots + \frac{(0{,}3847 - 0{,}4044)^2}{0{,}4044} = 0{,}0215$$

Et l'on a :

$$d^2_{eucl}(PARI', O) = (-0{,}1050)^2 + \ldots + (-0{,}0007)^2 = 0{,}0215$$

La même propriété s'applique aux colonnes. Le tableau complet des scores factoriels des colonnes est donné par :

	Facteur 1	Facteur 2	Facteur 3	Facteur 4	Facteur 5	Facteur 6	Facteur 7
HUCHON	-0,0421	-0,0165	0,1024	-0,0157	0,0024	-0,0023	-0,0020
COPE	-0,1305	-0,0513	-0,0089	0,0325	0,0108	-0,0038	0,0013
SANTINI	-0,2388	0,0955	-0,0822	-0,0225	-0,0035	-0,0032	-0,0009
LEPEN	0,1628	-0,1146	-0,0883	-0,0174	0,0017	-0,0101	-0,0034
BUFFET	0,2581	0,2259	0,0117	0,0178	0,0259	-0,0069	-0,0027
LAGU	0,1655	-0,0084	-0,0066	-0,0212	-0,0020	-0,0297	0,0204
PELEG	0,0332	-0,0714	-0,0625	-0,0499	0,0601	0,0423	0,0148
BAY	0,1514	-0,1198	-0,1211	-0,0160	0,0350	-0,0014	-0,0356
ABSTEN	0,0538	0,0058	-0,0059	0,0055	-0,0100	0,0060	0,0004

On avait établi que :

$$d^2_{\Phi^2}(BUFFET, SANTINI) = \frac{(0{,}1480 - 0{,}1934)^2}{0{,}1834} + \ldots + \frac{(0{,}0936 - 0{,}0811)^2}{0{,}1001} = 0{,}2753$$

On retrouve ici :

$$d^2_{eucl}(BUFFET', SANTINI') = (0{,}2581 + 0{,}2388)^2 + \ldots + (-0{,}0027 + 0{,}0009)^2 = 0{,}2753$$

La proximité entre un point-ligne L et un point-colonne C ne possède pas d'interprétation géométrique immédiate. En revanche, l'angle de sommet O dont les côtés passent par L et C a la propriété suivante :

- si l'angle (OL, OC) est aigu, la modalité-ligne L et la modalité colonne C s'attirent (taux de liaison positif)

- si l'angle (OL, OC) est obtus, la modalité-ligne L et la modalité colonne C se repoussent (taux de liaison négatif)
- si l'angle (OL, OC) est droit, la modalité-ligne L et la modalité colonne C n'interagissent pas (taux de liaison voisin de 0).

3.2.3.6 Reconstitution des données

Il est possible de reconstituer les données à partir des scores factoriels des lignes et des colonnes. En effet, on peut montrer la relation suivante entre les taux de liaison t_{ij}, les scores factoriels des lignes, les scores factoriels des colonnes et les valeurs propres :

$$t_{ij} = \sum_{Axes\ \ factoriels} \frac{(Score\ \ fact.\ ligne\ \ i\ \ selon\ \ axe\ \ \alpha)(Score\ \ fact.\ colonne\ \ j\ \ selon\ \ axe\ \ \alpha)}{\sqrt{Valeur\ \ propre\ \ associée\ \ à\ \ l'axe\ \ \alpha}}$$

Par exemple, le taux de liaison entre PARI et la liste HUCHON peut être retrouvé à l'aide du calcul suivant :

$$t_{11} = \frac{(-0{,}1050)(-0{,}0421)}{\sqrt{0{,}015123}} + \frac{(0{,}0027)(-0{,}0165)}{\sqrt{0{,}004656}} + \ldots + \frac{(-0{,}0007)(-0{,}0020)}{\sqrt{0{,}000024}} = 0{,}20406$$

Connaissant les profils moyens des lignes et des colonnes, et l'effectif total N, l'ensemble des données peut ainsi être retrouvé.

3.2.4 Interprétation des résultats de l'AFC

Au niveau global, on pourra noter que les inerties relatives les plus fortes sont observées sur la Seine St-Denis, les Hauts de Seine et Paris, pour les départements, et sur Santini, Buffet et Le Pen pour les listes. Ce sont donc essentiellement ces modalités lignes et modalités colonnes qui vont apparaître dans l'étude qui suit. En revanche, des modalités telles que l'abstention, proches du profil moyen, n'apparaîtront pas.

L'interprétation pourra être faite axe par axe, en étudiant d'abord séparément lignes et colonnes.
Pour chaque axe, on pourra dresser un tableau des individus qui ont apporté une contribution supérieure à la moyenne à la formation de cet axe.

3.2.4.1 Interprétation des axes

Pour le premier axe :

- Points lignes :

-	+
HTSS (22%)	STDE (42%)
PARI (13%)	

- Points colonnes :

-	+
SANTINI (36%)	BUFFET (19%)
COPE (17%)	LE PEN (13%)

Le premier axe oppose Paris et les Hauts de Seine à la Seine St Denis. Si on considère le positionnement des autres départements, cet axe oppose Paris et la banlieue Ouest (socialement assez favorisée) à la banlieue du nord et de l'est (socialement moins favorisée).

Pour les modalités colonnes, cet axe oppose deux listes proches de la majorité gouvernementale à deux listes de "forte opposition", voire de vote protestataire.

La synthèse entre l'analyse des lignes et des colonnes associe le vote protestataire à la Seine St Denis, tandis que le vote pour la majorité gouvernementale est mieux représenté dans l'ouest de la région.

Pour le deuxième axe :

- Points lignes :

-	+
SMAR (35%)	STDE (20%)
	HTSS (19%)

- Points colonnes :

-	+
LEPEN (20%)	BUFFET (47%)
	SANTINI (19%)

Cet axe oppose la Seine et Marne (grande banlieue, urbanisation plus diffuse et zones rurales) aux Hauts de Seine et à la Seine St Denis, très urbanisées. Le positionnement de l'ensemble des départements montre même une opposition entre les départements de la "grande couronne" et ceux de la "petite couronne".

Pour les modalités colonnes, cet axe oppose la liste Le Pen aux listes Santini et Buffet.

Cet axe oppose donc les zones moins urbanisés, où la liste Le Pen obtient ses meilleurs scores, aux zones plus urbanisés où le vote en dehors des listes "classiques" (UMP, PS) est surtout représenté par les listes Santini et Buffet.

Pour le troisième axe :

- Points lignes :

-	+
HTSS (22%)	PARI (55%)
YVEL (12%)	

- Points colonnes :

-	+
SANTINI (19%)	HUCHON (58%)
LE PEN (17%)	

Paris représente plus de la moitié de l'inertie de cet axe, qui est donc essentiellement représentatif des spécificités du vote à Paris "intra-muros". On note cependant que les trois départements figurant dans le tableau ci-dessus sont aussi les plus peuplés.
De même, la liste Huchon représente plus de la moitié de cet axe.

Le troisième axe associe donc la liste Huchon au vote à Paris. C'est effectivement dans ce département que cette liste obtient les meilleurs scores. L'indépendance entre les axes nous amène à nous demander s'il existe une spécificité du vote parisien. Mais, les résultats n'indiqueraient-ils pas plutôt que les spécificités rencontrées précédemment ne se retrouvent pas à Paris ?

3.2.4.2 Interprétation du premier plan factoriel

On peut aussi faire une interprétation globale du premier plan factoriel (axes 1 et 2) en distinguant les 4 quadrants :

1. Le vote protestataire de gauche, représenté par la Seine St Denis et la liste Buffet (et, dans une moindre mesure, la liste Laguiller) ;
2. Le cas spécifique du couple (Santini, Hauts de Seine)

3. Le vote pour les partis de gouvernement (faisant partie de la majorité, ou dans l'opposition). Ce vote est particulièrement représenté à Paris et dans les Yvelines
4. Le vote protestataire de droite (Le Pen, Bay), plutôt représenté dans la grande banlieue, et particulièrement en Seine et Marne.

3.2.4.3 Remarques :

1. Les grands partis "classiques" (Huchon, Copé) interviennent finalement assez peu dans l'analyse. C'est assez normal : d'une part, ils sont bien représentés dans tous les départements, d'autre part, les modalités-colonnes correspondantes ont une masse importante, et ils ont fortement contribué à la formation du profil-colonne moyen. Il n'est donc pas étonnant que les points colonnes qui les représentent soient proches de l'origine des axes. La même remarque s'applique aussi à l'abstention.
2. Les qualités de représentation sont bonnes (sauf pour l'Essonne). Cependant, il faut être prudent pour la liste Huchon et le département "Paris" : il faut attendre le 3è axe pour obtenir une qualité de représentation satisfaisante.
3. On notera sur les graphiques la proximité des deux listes d'extrême droite (Le Pen et Bay), qui n'est pas apparue dans les tableaux chiffrés en raison de la faiblesse numérique du vote "Bay".

3.2.5 Quelques principes d'interprétation supplémentaires

3.2.5.1 Forme générale du nuage

L'inertie totale (le Phi-2) est un indicateur de la dispersion totale du nuage. La comparaison des inerties de chacun des axes (c'est-à-dire des valeurs propres associées aux axes) renseigne sur la forme du nuage de points. Si les premières valeurs propres sont proches les unes des autres, la dispersion est relativement homogène : il n'y a pas vraiment de direction privilégiée et le nuage de points est approximativement sphérique. Si au contraire, les valeurs propres sont nettement différentes, cela traduit un nuage de points fortement allongé selon une (ou plusieurs) direction.

3.2.5.2 Valeurs propres proches de 1

Les valeurs propres sont toutes inférieures à 1. Mais, une valeur propre proche de 1 indique une dichotomie des données, c'est-à-dire un tableau de contingence qui, après reclassement des modalités, aurait l'allure suivante :

	0
0	

De même, l'existence de deux valeurs propres proches de 1 indique une partition des observations en 3 groupes. Si toutes les valeurs propres sont proches de 1, cela indique une correspondance entre chaque modalité ligne et une modalité colonne "associée". Avec une réorganisation convenable des modalités, les effectifs importants se trouvent alors le long de la diagonale.

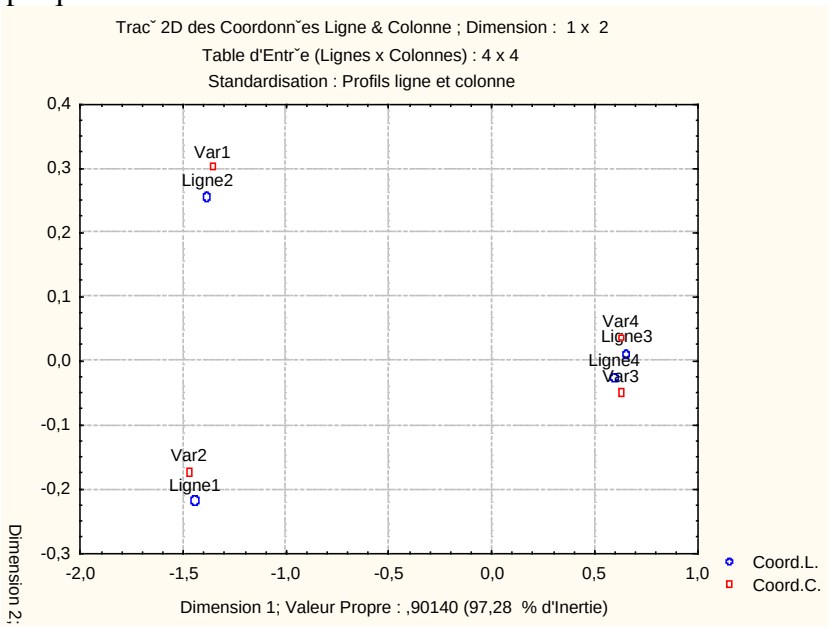
Exemple : Soit le tableau de contingence suivant :

	Var1	Var2	Var3	Var4
Ligne 1	20	45	2	0
Ligne 2	25	32	0	3
Ligne 3	1	0	78	112
Ligne 4	2	1	45	44

Les valeurs propres sont alors :

Nombre	Valeurs Propres et Inertie de toutes les Dimensions (dicho.sta)				
de Dims.	Inertie Totale = ,92657 Chi2 = 379,89 dl = 9 p = 0,0000				
	ValSing.	ValProp.	%age Inertie	%age Cumulé	Chi2
1	0,949423	0,901404	97,28374	97,2837	369,5757
2	0,132451	0,017543	1,89336	99,1771	7,1928
3	0,087320	0,007625	0,82290	100,0000	3,1261

La représentation graphique a l'allure suivante :



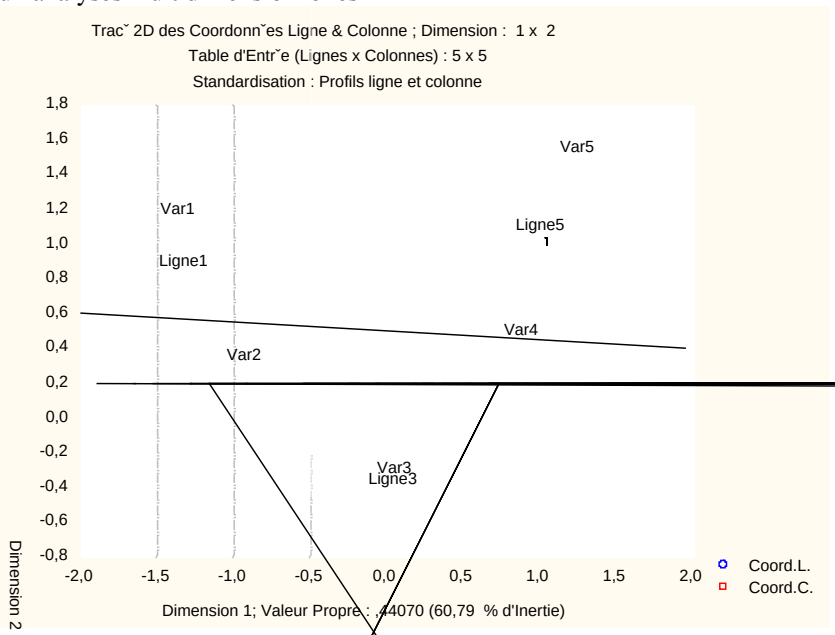
3.2.5.3 L'effet Guttman.

Un nuage de points de forme parabolique indique une redondance entre les deux variables étudiées : la connaissance de la ligne i donne pratiquement celle de la colonne j. Dans un tel cas, pratiquement toute l'information est contenue dans le premier facteur. Cette configuration se rencontre notamment lorsque les deux variables sont ordinales, et classent les sujets de la même façon. Dans ce cas, le premier axe oppose les valeurs extrêmes et classe les valeurs, tandis que le deuxième axe oppose les intermédiaires aux extrêmes.

Exemple :

	Var1	Var2	Var3	Var4	Var5
Ligne 1	10	30	7	0	0
Ligne 2	3	100	70	4	0
Ligne 3	2	32	200	35	1
Ligne 4	1	6	80	100	2
Ligne 5	0	3	5	25	5

Ce tableau conduit au nuage de points suivant :


Trac˜ 2D des Coordonn˜es Ligne & Colonne ; Dimension : 1 x 2
Table d'Entr˜e (Lignes x Colonnes) : 5 x 5
Standardisation : Profils ligne et colonne
1,8
1,6
1,4
1,2
1,0
0,8
0,6
0,4
0,2
0,0
-0,2
-0,4
-0,6
-0,8
Var5
Var1
Ligne5
Ligne1
Var4
Var2
Var3
Ligne3
-2,0
-1,5
-1,0
-0,5
0,0
0,5
1,0
1,5
2,0
Coord.L.
Coord.C.
Dimension 2
Dimension 1; Valeur Propre : ,44070 (60,79 % d'Inertie)

3.3 Analyse factorielle des correspondances avec Statistica

3.3.1 Présentation des données étudiées

Source : Lebart L., Morineau A., Piron M. Statistique Exploratoire Multidimensionnelle.

L'exemple concerne l'analyse d'un tableau de contingence qui croise 8 professions et catégories socioprofessionnelles (PCS) et 6 types de médias pour un échantillon de 12 388 "contacts média" relatifs à 4433 personnes interrogées. L'individu statistique sera pour nous le "contact média" et non la personne interrogée dans l'enquête. Les données sont extraites de l'Enquête Budget-temps Multimédia 1991-1992 du CESP.

Afin d'interpréter plus efficacement les représentations obtenues, on projettera en éléments supplémentaires certaines autres caractéristiques de la population enquêtée telles que le sexe, l'âge, le niveau d'instruction.

Tables de contingence croisant les types de contacts-média (colonnes) avec professions, sexe, âge, niveau d'éducation (lignes).

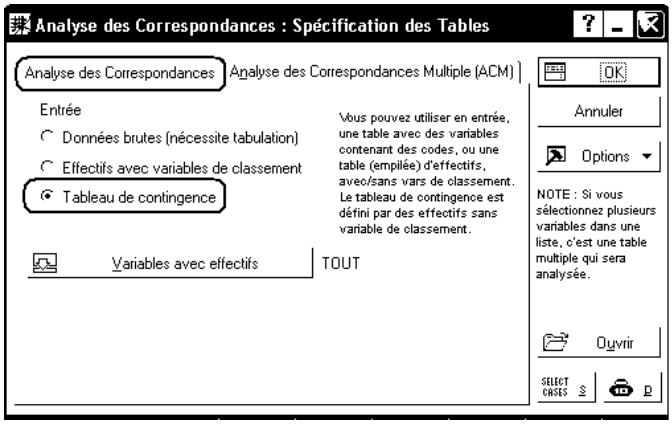
	Radio	Tél.	Quot.N.	Quot R.	P.Mag.	P.TV
Professions						
Agriculteur	96	118	2	71	50	17
Petit patron	122	136	11	76	49	41
Prof. Cad. S.	193	184	74	63	103	79
Prof. interm	360	365	63	145	141	184
Employé	511	593	57	217	172	306
Ouvrier qual	385	457	42	174	104	220
Ouvrier n-q	156	185	8	69	42	85
Inactif	1474	1931	181	852	642	782
Sexe						
Homme	1630	1900	285	854	621	776
Femme	1667	2069	152	815	683	938
Age						
15-24 ans	660	713	69	216	234	360
25-34 ans	640	719	84	230	212	380
35-49 ans	888	1000	130	429	345	466
50-64 ans	617	774	84	391	262	263
65 ans ou +	491	761	70	402	251	245
Education						
Primaire	908	1307	73	642	360	435
Secondaire	869	1008	107	408	336	494
Techn. prof.	901	1035	80	140	311	504
Supérieur	619	612	177	209	298	281

Nous disposons des tables de contingence suivantes (cf. tableau). Pour le premier bloc K de 8 lignes (lignes actives) on trouve, à l'intersection de la ligne i et de la colonne j le nombre kij d'individus appartenant à la catégorie i et ayant eu la veille (un jour de semaine) au moins un contact avec le type de média j. Les blocs suivants (lignes supplémentaires) s'interprètent de façon analogue. Une personne interrogée pouvant avoir des contacts avec plusieurs médias, les valeurs en ligne représentent des "nombres de contacts".

On cherche à décrire les éventuelles affinités entre les groupes socioprofessionnels et les différents types de médias

3.3.2 Traitement des données avec Statistica

Ouvrez le classeur Contacts-Medias-2006.stw et observez les données saisies.
Pour effectuer l'AFC, nous utilisons le menu Statistiques - Techniques exploratoires multivariées - Analyse des correspondances.



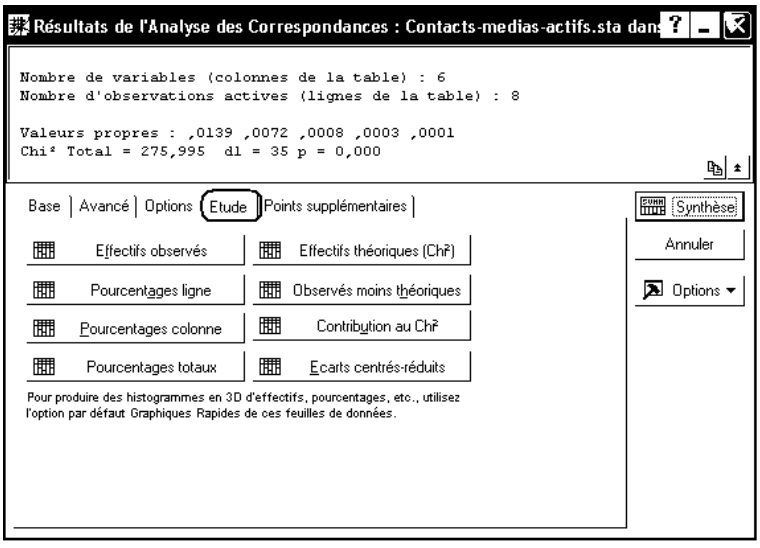
La fenêtre de dialogue permet d'indiquer la manière dont se présentent nos données. La situation la plus classique est celle d'un tableau de contingence : les modalités lignes sont indiquées comme noms d'observations (elles auraient pu être indiquées dans une variable spécifique), les modalités colonnes sont les variables du tableau, et la feuille de données contient les effectifs n_{ij}.

On indique également les variables qui participeront à l'analyse. Notez que les zéros sont obligatoires, car une cellule laissée vide est interprétée comme une valeur manquante, et c'est alors l'ensemble de la ligne qui est éliminé de l'analyse.

N.B. Ne fermez pas l'analyse en cours pendant la suite des manipulations. Ainsi, vous n'aurez pas à indiquer de nouveau les options ci-dessus, vos résultats seront cohérents entre eux et se rassembleront dans un même classeur.

3.3.2.1 Statistiques descriptives

Les principaux résultats de statistiques descriptives pourront être obtenus à partir de l'onglet "Etude".
On peut ainsi obtenir les fréquences, les fréquences lignes, les fréquences colonnes et les profils moyens.



Par exemple, les fréquences et les profils ligne et colonne moyens sont :

	Pourcentages Totaux (Contacts-medias-actifs.sta dans Classeur1) Table d'Entrée (Lignes x Colonnes) : 8 x 6 Inertie Totale = ,02228 Chi² = 276,00 dl = 35 p = 0,0000						
	Radio	Tél.	Quot.N.	Quot R.	P.Mag.	P.TV	Total
Agriculteur	0,77	0,95	0,02	0,57	0,40	0,14	2,86
Petit patron	0,98	1,10	0,09	0,61	0,40	0,33	3,51
Prof. Cad. S.	1,56	1,49	0,60	0,51	0,83	0,64	5,62
Prof. interm	2,91	2,95	0,51	1,17	1,14	1,49	10,15
Employé	4,12	4,79	0,46	1,75	1,39	2,47	14,98
Ouvrier qual	3,11	3,69	0,34	1,40	0,84	1,78	11,16
Ouvrier n-q	1,26	1,49	0,06	0,56	0,34	0,69	4,40
Inactif	11,90	15,59	1,46	6,88	5,18	6,31	47,32
Total	26,61	32,04	3,54	13,46	10,52	13,84	100,00

3.3.2.2 Choix des valeurs propres

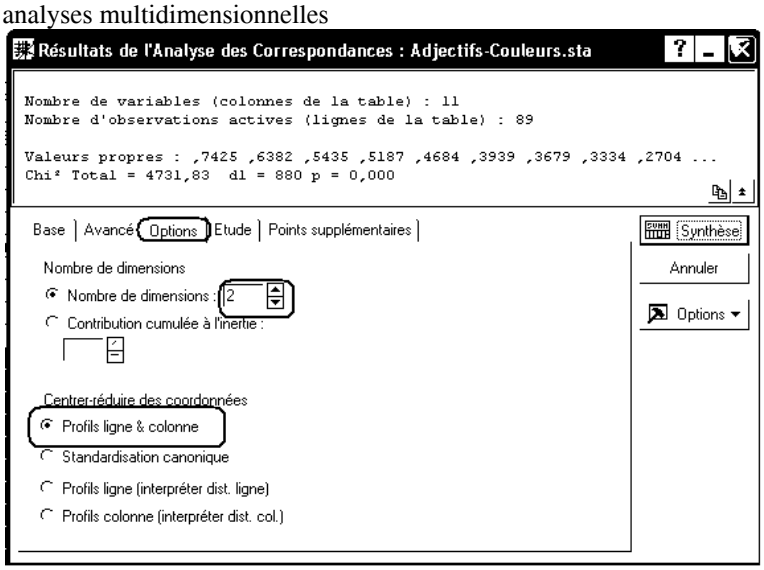
C'est ensuite l'onglet "Avancé" qui nous permettra d'afficher les valeurs propres, et donc de choisir le nombre d'axes à garder :

Nombre de Dims.	Valeurs Propres et Inertie de toutes les Dimensions Table d'Entrée (Lignes x Colonnes) : 8 x 6 Inertie Totale = ,02228 Chi² = 276,00 dl = 35 p = 0,0000				
	ValSing.	ValProp.	%age Inertie	%age Cumulé	Chi²
1	0,117717	0,013857	62,19818	62,1982	171,6641
2	0,084916	0,007211	32,36503	94,5632	89,3260
3	0,028718	0,000825	3,70179	98,2650	10,2168
4	0,017431	0,000304	1,36383	99,6288	3,7641
5	0,009094	0,000083	0,37117	100,0000	1,0244

On voit ici que seules les deux premières valeurs propres représentent plus de 20% d'inertie. Nous pourrons donc limiter l'étude au premier plan factoriel.

3.3.2.3 Résultats relatifs aux individus-lignes et aux individus-colonnes.

Pour les résultats qui suivent, on indique le nombre d'axes factoriels à conserver sous l'onglet "Base" ou sous l'onglet "Options". Ce dernier permet également de choisir plusieurs types d'échelles pour représenter lignes et colonnes. Le type de représentation vu en cours, qui fait jouer des rôles symétriques aux lignes et aux colonnes, correspond à la première option.

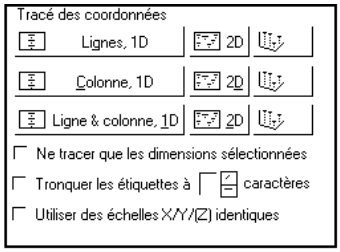


On retourne ensuite sous l'onglet "Avancé" pour afficher les coordonnées des individus-lignes et des individus-colonnes. On notera que Statistica produit deux tableaux de résultats, et on passera de l'un à l'autre à l'aide des onglets du classeur.

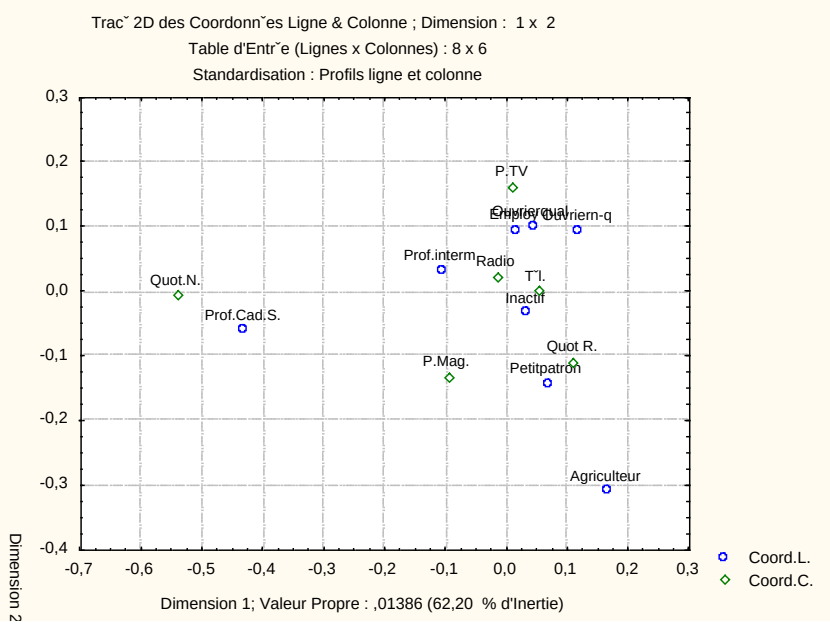
	Coordonnées Ligne et Contributions à l'Inertie Table d'Entrée (Lignes x Colonnes) : 8 x 6 Standardisation : Profils ligne et colonne									
NomLigne	Ligne Numéro	Coord. Dim.1	Coord. Dim.2	Masse	Qualité	Inertie Relative	Inertie Dim.1	Cosinus² Dim.1	Inertie Dim.2	Cosinus² Dim.2
Agriculteur	1	0,1661	-0,3096	0,0286	0,9549	0,1658	0,0569	0,2135	0,3799	0,7414
Petit patron	2	0,0684	-0,1432	0,0351	0,8281	0,0479	0,0118	0,1538	0,0998	0,6742
Prof. Cad. S.	3	-0,4300	-0,0609	0,0562	0,9978	0,4766	0,7496	0,9782	0,0289	0,0196
Prof. interm	4	-0,1066	0,0326	0,1015	0,8772	0,0646	0,0833	0,8022	0,0150	0,0750
Employé	5	0,0157	0,0955	0,1498	0,9542	0,0660	0,0027	0,0252	0,1894	0,9289
Ouvrier qual	6	0,0437	0,1014	0,1116	0,8820	0,0692	0,0154	0,1383	0,1590	0,7437
Ouvrier n-q	7	0,1178	0,0949	0,0440	0,9161	0,0493	0,0441	0,5557	0,0549	0,3604
Inactif	8	0,0326	-0,0334	0,4732	0,7632	0,0606	0,0363	0,3722	0,0732	0,3910

	Coordonnées Colonne et Contributions à l'Inertie Table d'Entrée (Lignes x Colonnes) : 8 x 6 Standardisation : Profils ligne et colonne									
Nom Col.	Colonne Numéro	Coord. Dim.1	Coord. Dim.2	Masse	Qualité	Inertie Relative	Inertie Dim.1	Cosinus² Dim.1	Inertie Dim.2	Cosinus² Dim.2
Radio	1	-0,0149	0,0221	0,2661	0,2454	0,0346	0,0043	0,0770	0,0180	0,1685
Tél.	2	0,0533	0,0021	0,3204	0,8521	0,0480	0,0656	0,8508	0,0002	0,0013
Quot.N.	3	-0,5407	-0,0062	0,0354	0,9931	0,4672	0,7459	0,9930	0,0002	0,0001
Quot R.	4	0,1088	-0,1096	0,1346	0,9806	0,1470	0,1150	0,4866	0,2244	0,4940
P.Mag.	5	-0,0948	-0,1325	0,1052	0,9354	0,1340	0,0682	0,3168	0,2561	0,6186
P.TV	6	0,0098	0,1616	0,1384	0,9622	0,1692	0,0009	0,0035	0,5011	0,9587

On utilise ensuite les boutons du bloc "Tracé des coordonnées" pour obtenir des représentations graphiques des résultats de l'AFC.



Les graphiques "par axe" pourront être obtenus à l'aide du bouton "Ligne & colonne, 1D". Le graphique dans un plan, superposant les résultats des lignes et des colonnes, pourra être obtenu à l'aide du bouton "2D" de la même ligne. En revanche, il n'est pas évident d'éliminer certaines étiquettes pour améliorer la lisibilité du graphique. La seule méthode paraît être de faire un clic droit sur une étiquette, de sélectionner l'item de menu "Propriétés..." puis d'éditer manuellement le tableau des étiquettes qui s'affiche.



3.3.2.4 Individus ligne et individus colonne supplémentaires

L'insertion d'individus-ligne ou d'individus-colonne supplémentaires peut poursuivre deux buts :

- d'une part, il peut être utile de positionner sur le graphique les groupes définis par une autre variable, telle que le sexe ou l'âge ou le niveau d'étude ;
- d'autre part, on peut remarquer que les modalités "Quotidiens nationaux" et "Prof. Cad. S." jouent un rôle prépondérant dans la formation du premier axe factoriel. On peut donc souhaiter réaliser l'AFC en ignorant ces modalités, puis en les réintroduisant comme éléments supplémentaires.

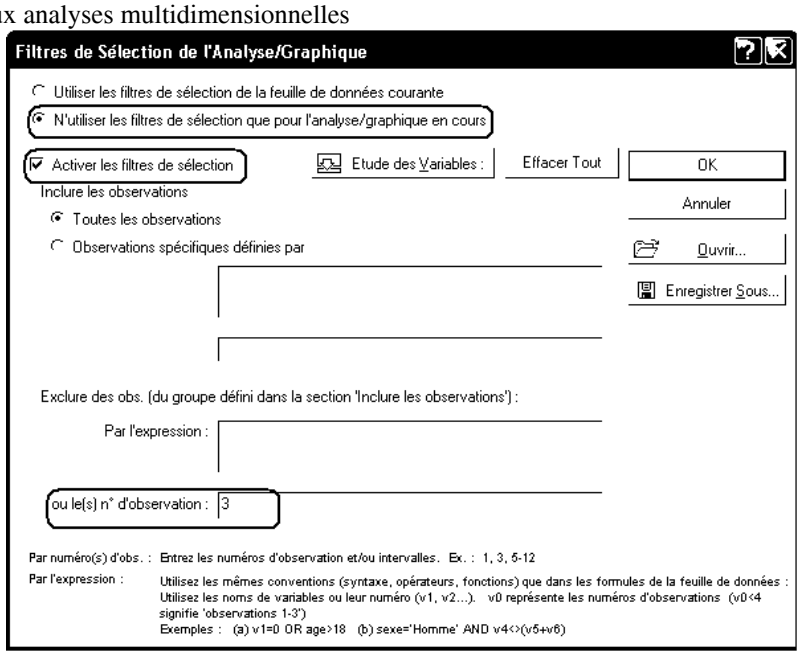
Positionner les groupes définis par l'âge, le sexe, le niveau d'étude

L'insertion d'éléments supplémentaires dans une AFC n'est pas très commode avec Statistica. Ici, on pourra procéder de la manière suivante :

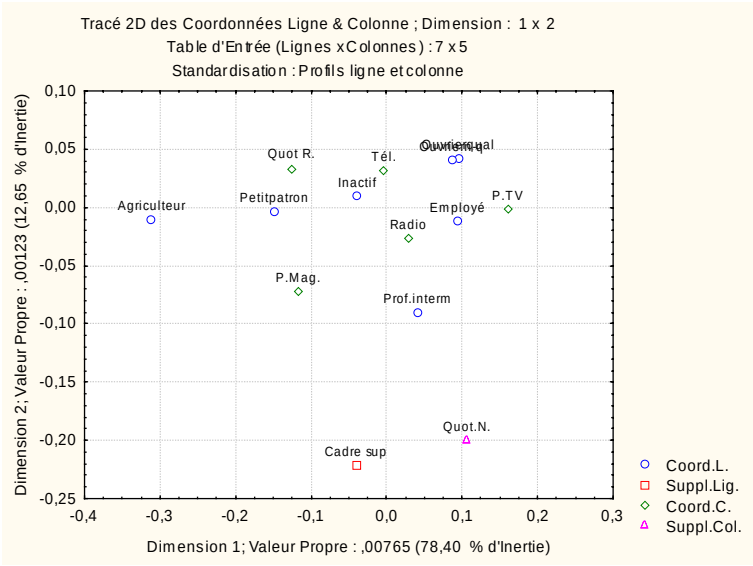
- Ouvrez le fichier de données Contacts-medias-supplementaires.sta, et copiez son contenu.
- Dans l'analyse en cours, activez l'onglet "Points supplémentaires", puis cliquez sur le bouton "Ajouter des points-ligne".
- Collez les données précédemment copiées à l'aide de la combinaison de touches Ctrl+V.

Refaites l'analyse, en réalisant notamment un graphique 2D avec l'ensemble des points lignes.
Le graphique qui suit ne représente que les points supplémentaires. Il a été obtenu en réalisant un graphique pour tous les points, puis en modifiant les options d'affichage de façon à faire disparaître les individus lignes et colonnes actifs :


Tracˇ 2D des Coordonnˇes Ligne & Colonne ; Dimension : 1 x 2
Table d'Entrˇe (Lignes x Colonnes) : 8 x 6
Standardisation : Profils ligne et colonne
Suppl.Lig.



On peut alors réintroduire ces points à l'aide de l'onglet "Points supplémentaires" vu précédemment. On notera cependant que l'effectif conjoint des deux modalités (74 contacts avec un "Quot. N." pour les "Prof. Cad. S.") n'intervient alors plus dans l'étude. L'analyse qui en résulte fournit des résultats assez différents des précédents, résumés dans le graphique ci-dessous :



3.4 Exercices et prolongements

3.4.1 Structures possibles pour les données d'entrée

On étudie la répartition de 296 prix Nobel selon le pays (4 pays : USA, Grande-Bretagne, République Fédérale Allemande, France) et la discipline (5 disciplines : Médecine, Physique, Chimie, Littérature, Sciences Economiques). Source : Rouanet, Le Roux, Bert (1987) d'après Le Monde

Sous forme de tableau de contingence, les données sont les suivantes :

PAYS	MEDE	PHYS	CHIM	LITT	SECO
USA	55	43	24	8	9
GB	19	20	21	6	2

RFA	11	14	24	7	0
FRAN	7	9	6	11	0

Dans le répertoire Nobel du serveur de TD, on trouve les fichiers Nobel-contingence.sta, Nobel-effectifs.sta, Nobel-protocole.sta ainsi que le fichier Excel Nobel.xls.

Observez le contenu de ces trois fichiers, et celui des trois onglets du classeur Excel. Il s'agit des mêmes données, mais structurées différemment.

Réalisez une AFC en utilisant successivement chacune des trois sources de données.
Interprétez les résultats de l'AFC, en répondant notamment aux questions suivantes :

- La répartition des prix Nobel par discipline est-elle la même pour les 4 pays ?
- Quels sont les pays les plus proches du point de vue du type de prix Nobel reçu ?
- Quels sont les pays les plus atypiques ?

3.4.2 Exercice à traiter à l'aide de Statistica

Le tableau de contingence suivant indique la répartition, en fonction des états-civils des conjoints, des 300513 mariages célébrés en France en 1983 :

	HCEL	HVEU	HDIV
FCEL	239767	1778	19807
FVEU	1954	1435	1597
FDIV	16837	2212	15126

Variable en ligne : Etat-civil de la femme
- FCEL : femme célibataire
- FVEU : femme veuve
- FDIV : femme divorcée

Variable en colonne : Etat-civil de l'homme
- HCEL : homme célibataire
- HVEU : homme veuve
- HDIV : homme divorcé

Source : INSEE, cité par Rouanet, Le Roux, Bert, 1987.

Les mariages se font-ils indépendamment de l'état-civil antérieur du conjoint ? Si non, quels états-civils "s'attirent", quels états-civils se repoussent ?

3.4.3 Exercice : associations Adjectifs-couleurs

Références : Extrait de [Fénelon, "Qu'est-ce que l'analyse des données ?", Lefonen] trouvé à l'adresse : http://www.escna.fr/fr/nte/cours/MKT/Ana_Don/adp6.htm .

L'exemple qui suit rassemble des résultats d'une expérience d'association couleur-adjectif.

Ouvrez la feuille de données adjectifs-couleurs.sta et observez les données saisies.

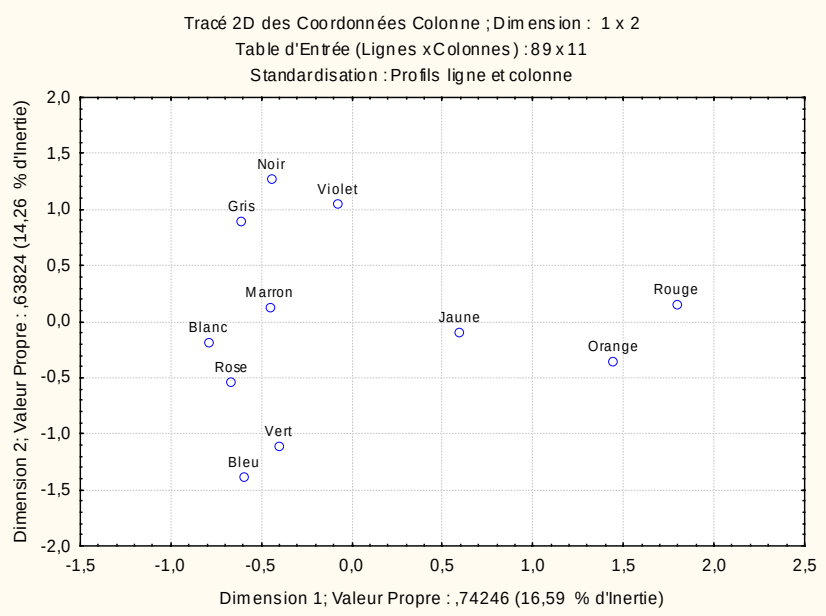
3.4.3.1 Choix des valeurs propres

Nombre de Dims.	Valeurs Propres et Inertie de toutes les Dimensions (Adjectifs-Couleurs.sta) Inertie Totale = 4,4767 Chi^2 = 4731,8 dl = 880 p = 0,0000				
	ValSing.	ValProp.	%age Inertie	%age Cumulé	Chi^2
1	0,8617	0,7425	16,5850	16,5850	784,7761
2	0,7989	0,6382	14,2569	30,8420	674,6150
3	0,7372	0,5435	12,1402	42,9822	574,4531
4	0,7202	0,5187	11,5876	54,5697	548,3037
5	0,6844	0,4684	10,4623	65,0320	495,0595
6	0,6276	0,3939	8,7991	73,8312	416,3597
7	0,6066	0,3679	8,2191	82,0503	388,9151
8	0,5774	0,3334	7,4481	89,4984	352,4327
9	0,5200	0,2704	6,0396	95,5380	285,7839
10	0,4469	0,1997	4,4620	100,0000	211,1346

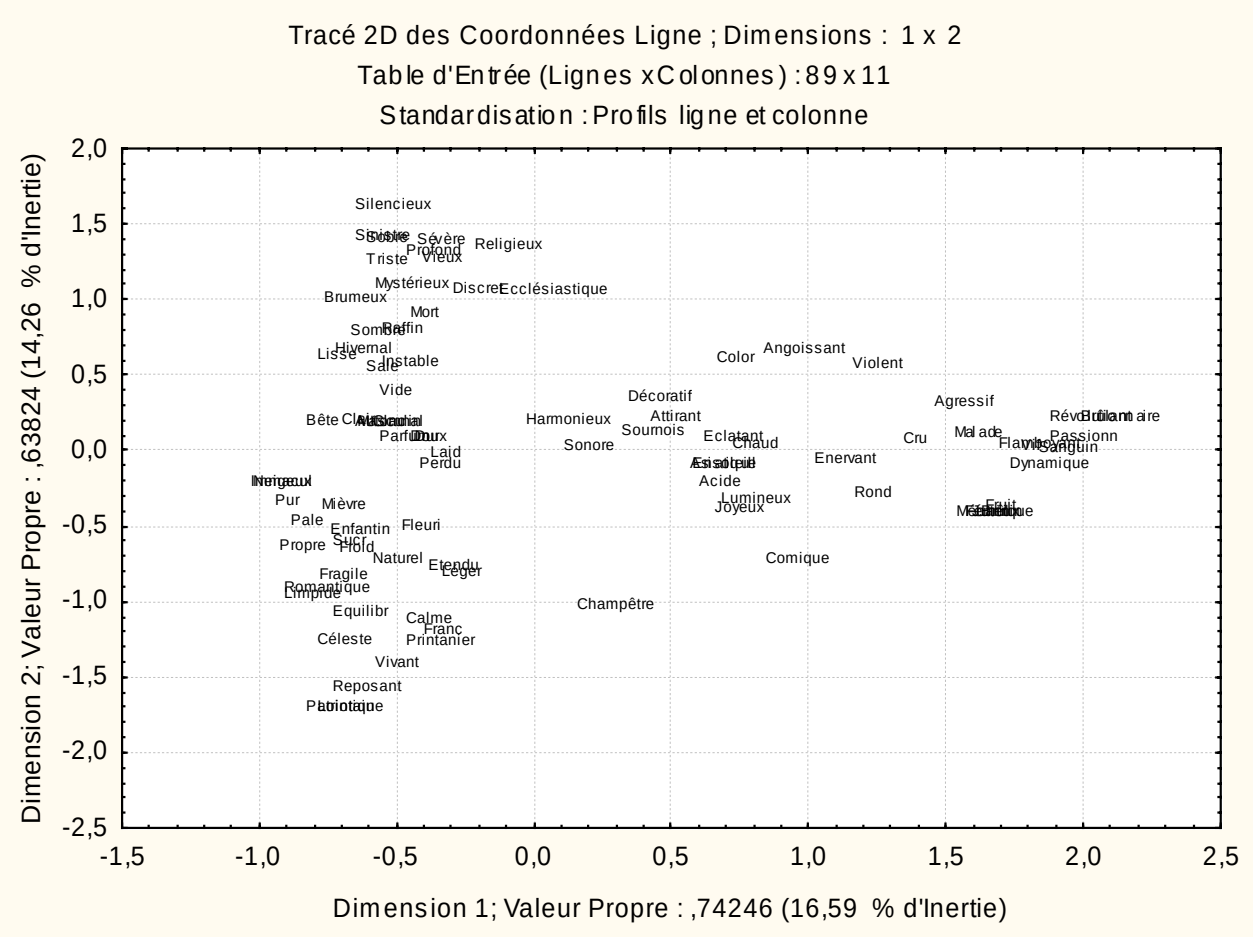
On voit ici que la décroissance des valeurs propres est très lente. Selon les règles énoncées précédemment, il faudrait conserver au moins 5 axes. Mais nous pouvons convenir de ne rechercher que les propriétés les plus caractéristiques des associations adjectifs - couleurs en n'étudiant que les deux premiers axes.

3.4.3.2 Résultats relatifs aux individus-lignes et aux individus-colonnes.

Ici, pour interpréter les colonnes (couleurs), on pourra s'appuyer sur le graphique 2D limité aux seuls individus-colonnes :



La représentation des adjectifs pose plus de problèmes, étant donné leur nombre. Par exemple, on pourra afficher les étiquettes en caractères de taille 6, et supprimer les symboles des points (style assez classique pour ce type de schéma). On obtient ainsi le schéma suivant :



3.4.3.3 Quelques éléments d'interprétation

Le commentaire qui suit a été trouvé sur le site Web cité plus haut. Mais, il ne s'agit évidemment pas d'une interprétation complète des résultats obtenus.

C'est la structure triangulaire des données qui doit être soulignée. Apparemment, les couleurs rouge et orange s'opposent à toutes les autres (en tant que couleurs, elles sont donc fortement distinctives). Sur la gauche du mapping, on note une opposition entre des "non couleurs" ("noir", "gris") en haut et des couleurs pastel en bas.

On trouve à l'occasion sur le graphique des proximités couleur/adjectif justifiées par des associations fortes (par exemple, "asiatique" et "jaune" ou "marron" et "glacé"). Mais ce n'est pas toujours le cas. Les analyses factorielles travaillant sur des projections, une proximité dans l'espace d'origine se traduit forcément par une proximité sur les graphes factoriels, mais l'inverse n'est pas vrai (surtout vers le centre des graphiques). Il serait par exemple erroné de soutenir que "marron" et "perdu" sont fortement liés à cause de leur proximité sur le mapping. Un coup d'oeil à la matrice des données (appelée aussi, sous cette forme "tableau de contingence") montre qu'ils ne sont jamais associés l'un à l'autre.

3.4.4 Exercice : le cas Environnement

Les données suivantes ont été recueillies pour étudier la relation entre la catégorie socio-professionnelle (CSP) et la principale source d'information sur les problèmes d'environnement.
Sept CSP sont étudiées : agriculteur (AGRI), cadre supérieur (CSUP), cadre moyen (CMOY), employé (EMPL), ouvrier (OUVR), retraité (RETR), chômeur (CHOM).

Les 1283 personnes interrogées devaient indiquer leur principale source d'information sur les problèmes d'environnement, parmi les six sources suivantes : télévision (TEL), journaux (JOU), radio (RAD), livres (LIV), associations (ASS) et mairie (MAI).

CSP	TEL	JOU	RAD	LIV	ASS	MAI	Total
AGRI	26	18	9	5	4	6	68
CSUP	19	49	4	16	5	3	96
CMOY	44	87	4	39	14	3	191
EMPL	83	87	13	24	5	1	213
OUVR	181	107	16	31	7	7	349
RETR	167	95	29	15	7	7	320
CHOM	27	9	4	2	2	2	46
Total	547	452	79	132	44	29	1283

Saisissez ces données dans une feuille de données Statistica, sous une forme permettant d'effectuer ensuite une AFC.

Analysez ces données à l'aide d'une AFC sous Statistica, puis rédigez, dans un document Word, une interprétation des résultats obtenus, en répondant notamment aux questions suivantes :

1) On a décidé de ne retenir que les deux premiers axes principaux. Justifiez ce choix.

2) Etude de la première variable factorielle.
a) On considère d'abord le nuage des CSP. Quels sont les individus dont la contribution est supérieure à la moyenne ? Pour chacun d'eux, précisez le signe de la coordonnée correspondante.
b) Mêmes questions pour le nuage des sources d'information.
c) Indiquer ce que suggère principalement cette analyse de la première variable factorielle.

3)Etude de la seconde variable factorielle.
a) Du point de vue du nuage des CSP, un individu unique a une contribution prédominante. Lequel ?
b) Commenter de même les contributions des sources d'information.
c) Quelle interprétation de la seconde variable factorielle cette analyse
suggère-t-elle ? Pourquoi faut-il se montrer très prudent avant d'accepter cette interprétation ?

3.4.5 Exercice : étude des réponses à une question ouverte

Source : Lebart, L., Salem, A. (1988), Analyse des données textuelles, Paris, Dunod, repris par Corroyer D., Université Paris V.
Voir aussi le fichier W:\PSY4\PSRS83B\Mots\Mots-Corroyer.stw sur le serveur de TD.

On a posé deux questions à un échantillon de plusieurs centaines de personnes :
- "Quelles sont les raisons qui, selon vous, peuvent faire hésiter une femme ou un couple à avoir un enfant ?"
- "Quel est votre niveau d'études ?"

Pour la deuxième question, les réponses possibles étaient : sans diplôme (SANS), certificat d'études primaires (CEP), brevet d'études du premier cycle (BEPC), baccalauréat ou équivalent (BAC), université, grandes écoles ou équivalent (UNIV).
Pour la première question, les réponses ont été analysées. On a retenu 15 des mots utilisés : Peur, Santé, Avenir, Argent, Emploi, Guerre, Chômage, Travail, Egoïsme, Finances, Logement, Difficile, Economique, Financières, Conjoncture. Chaque personne peut avoir utilisé plusieurs de ces mots. Le tableau suivant indique, pour chacun des 15 mots retenus, le nombre d'occurrences d'utilisation en fonction du niveau d'étude.

MOTS	SANS	CEP	BEPC	BAC	UNIV	*TOTAL*
PEUR	25	45	38	38	13	*159*
SANTE	18	27	20	19	9	*93*
AVENIR	53	90	78	75	22	*318*
ARGENT	51	64	32	29	17	*193*
EMPLOI	12	35	19	6	7	*79*
GUERRE	4	7	7	6	2	*26*
CHOMAGE	71	111	50	40	11	*283*
TRAVAIL	35	61	29	14	12	*151*
EGOISME	21	37	14	26	9	*107*
FINANCES	10	7	7	3	1	*28*
LOGEMENT	8	22	7	10	5	*52*
DIFFICILE	7	11	4	3	2	*27*
ECONOMIQUE	7	13	12	11	11	*54*
FINANCIERES	21	32	42	47	30	*172*
CONJONCTURE	1	7	5	5	4	*22*
TOTAL	*344*	*569*	*364*	*332*	*155*	*1764*

Traiter ce tableau par une analyse factorielle des correspondances et répondez aux questions suivantes :

1) Caractériser qualitativement le profil du mot "Economique" par rapport au profil moyen.

2) Compte tenu des informations fournies, est-il légitime de ne s'intéresser qu'aux deux premiers axes factoriels ? Justifier.

3) Dans le tableau des résultats relatifs aux lignes, la colonne "masse" indique la valeur 0,0306 pour l'individu "Economique". Comment peut-on retrouver cette valeur ?

4) a) Les mots "Guerre" et "Peur" sont très proches l'un de l'autre sur le graphique, alors que "Economique" et "Finances" sont très éloignés. Expliquer pourquoi, en vous appuyant sur les tableaux des fréquences lignes et colonnes et sur le tableau des scores factoriels étendu à l'ensemble des facteurs.

b) Les mots "Santé" et "Egoïsme" sont tous deux proches de l'origine des axes.

Comment peut-on expliquer cette proximité pour chacun des deux mots ?

5) Etude de la première variable factorielle

a) On considère le nuage des mots. Quels sont les individus dont la contribution est supérieure à la moyenne ? Pour chacun d'eux, préciser le signe de la coordonnée correspondante. Comment peut-on interpréter cet axe en termes d'opposition entre mots ?

b) Même question pour le nuage des niveaux d'étude.

6) Mener une étude analogue pour la deuxième variable.

7) Faire une synthèse des deux études précédentes en décrivant les résultats obtenus dans le premier plan factoriel.

3.5 Travail à rendre

Source des données : Léopold Simar, Angélique Baclin :
http://www.stat.ucl.ac.be/cours/stat2411/index.html

On a mené auprès de 31079 sujets une enquête relative à leurs habitudes concernant la façon dont ils passent leurs vacances. On s'intéresse ici aux réponses obtenues aux deux questions suivantes :

- "Quelle est votre catégorie socio-professionnelle ?" (réponses possibles : Agriculteurs, Petits patrons, Cadres supérieurs, Cadres moyens, Employés, Ouvriers, Autres actifs, Inactifs).
- "Quel mode de villégiature avez-vous choisi lors de vos dernières vacances ?" (réponses possibles : A l'hôtel, En location, Dans une résidence secondaire, Chez des parents, Chez des amis, En camping/caravaning, En séjour organisé ou village vacances, Autre).

Les données recueillies sont les suivantes :

	Hôtel	Location	Résid.	Parents	Amis	Camping	Séjour	Autres	*Total*

			Second.				organisé		
Agriculteurs	195	62	1	499	44	141	49	65	*1056*
Petits patrons	700	354	229	959	185	292	119	140	*2978*
Cadres sup.	961	471	633	1580	305	360	162	148	*4620*
Cadres moy.	572	537	279	1689	206	748	155	112	*4298*
Employés	441	404	166	1079	178	434	178	92	*2972*
Ouvriers	783	1114	387	4052	497	1464	525	387	*9209*
Autres actifs	142	103	210	1133	132	181	46	59	*2006*
Inactifs	741	332	327	1789	311	236	102	102	*3940*
Total	*4535*	*3377*	*2232*	*12780*	*1858*	*3856*	*1336*	*1105*	*31079*

Traitez ce tableau par une analyse factorielle des correspondances et répondez aux questions suivantes :

1) Comment ont été obtenues les premières valeurs respectives (0,63%, 18,47% et 4,30%) du tableau des fréquences, de celui des fréquences lignes et de celui des fréquences colonnes ?

2) a) Statistica ne fournit pas directement le tableau des taux de liaison. Utilisez cependant le tableau des "Observés moins théoriques" pour indiquer une modalité ligne et une modalité colonne qui "s'attirent". Indiquez de même une modalité ligne et une modalité colonne qui "se repoussent".

b) Le taux de liaison entre "Agriculteurs" et "Résidence secondaire" est de -0,9868. Vérifiez cette valeur. Comment pourrait-on exprimer d'une autre façon ce résultat ?

3) Compte tenu des informations fournies, est-il légitime de ne s'intéresser qu'aux deux premiers axes factoriels ? Justifiez.

4) Dans le tableau des résultats relatifs aux lignes, la colonne "masse" indique la valeur 0,2963 pour l'individu-ligne "Ouvriers". Comment peut-on retrouver cette valeur ?

5) a) Sur le graphique, le point "Agriculteurs" apparaît assez proche de l'origine des axes. Peut-on en conclure que cet individu-ligne a un profil proche du profil-ligne moyen ?

b) Pour l'individu-colonne "Parents", le tableau des résultats indique une masse de 0,4112 et une inertie relative de 0,1276, alors que pour l'individu "Résidence secondaire", ces valeurs sont respectivement 0,0718 et 0,2236. Comment peut-on interpréter ces résultats ?

6) Etude de la première variable factorielle

a) On considère le nuage des catégories socio-professionnelles. Quels sont les individus dont la contribution est supérieure à la moyenne ? Pour chacun d'eux, préciser le signe de la coordonnée correspondante. Comment peut-on interpréter cet axe en termes d'opposition entre catégories socio-professionnelles.

b) Même question pour le nuage des modes d'hébergement.

7) Etude de la deuxième variable factorielle

a) L'un des individus-lignes a eu une influence importante dans la formation de cette variable. Lequel ?

b) Comment peut-on interpréter le deuxième axe factoriel en termes d'opposition entre modes d'hébergement.

c) L'individu-ligne "Autres actifs" semble occuper une position particulière sur le graphique : il est placé dans le bas du graphique, à l'écart des autres individus lignes et aucun individu colonne n'apparaît dans cette partie du graphique. De quelle façon le tableau des taux de liaison permet-il d'expliquer la position de ce point ?

8) Faites une synthèse des deux études précédentes en décrivant les résultats obtenus dans le premier plan factoriel.

Travail à rendre par mail à votre enseignant (Francois.Carpentier@univ-brest.fr) :

- Un classeur Statistica contenant les résultats numériques de l'AFC et les graphiques.
- Un fichier Word contenant votre interprétation des résultats avec notamment les réponses aux questions 1 à 8 ci-dessus.