2 Analyse Factorielle des Correspondances

N.B. Pour éviter les dysfonctionnements de Statistica 10 lors de l'affichage des graphiques :
- Allez au menu Outils - Options
- Allez sur l'onglet Documents > Graphs > Affichage
- Désactivez la boîte à cocher : "Permettre un rendu avancé des graphiques".
Ces options sont normalement enregistrées avec votre profil, lorsque vous quittez le logiciel.

2.1 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 développée 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'AFC 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.

2.2 Exemple

2.2.1 Enoncé

Réf. Résultats publiés dans les quotidiens "Le Monde" datés des 24 avril et 8 mai 2012.

Les données qui suivent sont constituées par les résultats du premier tour des élections présidentielles de 2012. Pour chacune des 23 régions françaises (22 régions métropolitaines + 1 "région" Outremer), on donne les effectifs de suffrages pour chacun des 10 candidats (en colonnes). L'objectif est d'analyser la structure des votes ainsi que les liaisons entre candidats et régions. On a également prévu une colonne "bulletins blancs et nuls" afin d'étudier l'évolution du vote blanc entre les deux tours.

Données : résultats du premier tour des présidentielles 2012.

Region	Hollande	Sarkozy	Le Pen	Bayrou	Melenchon	Joly	Dupond-Aignan	Poutou	Arthaud	Cheminade	Blancs_Nuls
Alsace	191282	326313	219251	116115	72376	27168	18681	10817	6387	2977	19825
Aquitaine	593891	465683	296210	204805	231951	45067	31199	30281	9114	4371	36540
Auvergne	252117	192229	139868	81816	104730	15348	15276	10333	5509	2010	19031
Bourgogne	262816	250202	191148	81986	97185	17077	19101	11192	5937	2318	19160
Bretagne	628421	508072	262102	224902	217923	58396	35587	26693	12593	5085	36169
Centre	385182	403455	280094	137170	151969	26314	30608	17609	9827	3752	29194
Champ-Ard	178914	206171	172783	62093	62184	10150	14993	8611	4990	1752	12621
Corse	39029	50493	39210	8045	15843	3762	1728	1870	502	337	2919
Fr-Comte	172644	177701	141969	58233	73946	14379	13221	8459	4615	1797	14590

Ile-de-Fr	1695345	1549965	655835	492062	632181	144338	92469	44554	20129	13411	86766
Lang-Rouss	408662	384094	363879	105431	204169	35465	22228	17072	7027	3488	28276
Limousin	172150	94373	69377	34568	58007	7462	7722	5385	2655	991	10225
Lorraine	332003	330550	308405	118883	133564	22749	25860	18175	9435	3621	24321
Midi-Pyr	552933	397600	281088	165585	229283	44763	28324	21319	8295	4173	34859
Nord-PdeC	633438	524339	517136	156865	270990	34293	35280	26089	16791	5010	41527
Basse-Nor	242706	249473	150801	93171	88064	17278	19336	12031	6142	2312	16017
Haute-Nor	286908	266940	207519	82589	126333	16898	19520	13500	7107	2561	19599
Pays-Loire	609220	614201	308798	250002	217500	52826	44722	28640	14338	5093	47799
Picardie	282725	266735	266067	77127	110642	14084	21017	13424	8799	2546	19575
Poit-Char	329070	270310	173591	100801	112765	22036	22168	15347	6974	2531	23333
PACA	600744	846041	650320	182175	306341	59011	41441	23103	9494	5888	44836
Rhone-Alp	869820	968762	628311	340035	382676	97781	67700	35386	18289	8974	62954
Outremer	439365	256841	74016	55023	51173	19859	10757	8445	6476	3117	47673

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

Y a-t-il des régions dont les résultats sont proches des résultats nationaux ? Y a-t-il des régions "à part" (dont les résultats s'écartent notablement des résultats nationaux) ?

Y a-t-il des candidats dont les résultats se ressemblent : ils n'obtiennent pas nécessairement les mêmes scores, mais les régions où ils obtiennent de bons scores sont les mêmes ? Y a-t-il des candidats dont les résultats s'opposent ?

Y a-t-il des candidats pour lesquels la répartition des votes est la même dans toutes les régions ? Y a-t-il des candidats pour lesquels le vote est concentré dans certaines régions ?

Comment les régions "à part" et les candidats à "vote inégalement réparti" s'associent-ils ?

2.2.2 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

2.2.2.1 Tableau des fréquences

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

	Hollande	Sarkozy	Le Pen	Bayrou	Melenchon	Joly	Dupond-Aignan	Poutou	Arthaud	Cheminade	Blancs_Nuls	Total
Alsace	0,53	0,90	0,61	0,32	0,20	0,08	0,05	0,03	0,02	0,01	0,05	2,79
Aquitaine	1,64	1,29	0,82	0,57	0,64	0,12	0,09	0,08	0,03	0,01	0,10	5,39
Auvergne	0,70	0,53	0,39	0,23	0,29	0,04	0,04	0,03	0,02	0,01	0,05	2,32
Bourgogne	0,73	0,69	0,53	0,23	0,27	0,05	0,05	0,03	0,02	0,01	0,05	2,65
Bretagne	1,74	1,40	0,72	0,62	0,60	0,16	0,10	0,07	0,03	0,01	0,10	5,57
Centre	1,06	1,12	0,77	0,38	0,42	0,07	0,08	0,05	0,03	0,01	0,08	4,08
Champ-Ard	0,49	0,57	0,48	0,17	0,17	0,03	0,04	0,02	0,01	0,00	0,03	2,03
Corse	0,11	0,14	0,11	0,02	0,04	0,01	0,00	0,01	0,00	0,00	0,01	0,45
Fr-Comte	0,48	0,49	0,39	0,16	0,20	0,04	0,04	0,02	0,01	0,00	0,04	1,88
Ile-de-Fr	4,69	4,28	1,81	1,36	1,75	0,40	0,26	0,12	0,06	0,04	0,24	15,00
Lang-Rouss	1,13	1,06	1,01	0,29	0,56	0,10	0,06	0,05	0,02	0,01	0,08	4,37
Limousin	0,48	0,26	0,19	0,10	0,16	0,02	0,02	0,01	0,01	0,00	0,03	1,28
Lorraine	0,92	0,91	0,85	0,33	0,37	0,06	0,07	0,05	0,03	0,01	0,07	3,67
Midi-Pyr	1,53	1,10	0,78	0,46	0,63	0,12	0,08	0,06	0,02	0,01	0,10	4,89
Nord-PdeC	1,75	1,45	1,43	0,43	0,75	0,09	0,10	0,07	0,05	0,01	0,11	6,25
Basse-Nor	0,67	0,69	0,42	0,26	0,24	0,05	0,05	0,03	0,02	0,01	0,04	2,48
Haute-Nor	0,79	0,74	0,57	0,23	0,35	0,05	0,05	0,04	0,02	0,01	0,05	2,90

Pays-Loire	1,68	1,70	0,85	0,69	0,60	0,15	0,12	0,08	0,04	0,01	0,13	6,06
Picardie	0,78	0,74	0,74	0,21	0,31	0,04	0,06	0,04	0,02	0,01	0,05	2,99
Poit-Char	0,91	0,75	0,48	0,28	0,31	0,06	0,06	0,04	0,02	0,01	0,06	2,98
PACA	1,66	2,34	1,80	0,50	0,85	0,16	0,11	0,06	0,03	0,02	0,12	7,65
Rhone-Alp	2,40	2,68	1,74	0,94	1,06	0,27	0,19	0,10	0,05	0,02	0,17	9,62
Outremer	1,21	0,71	0,20	0,15	0,14	0,05	0,03	0,02	0,02	0,01	0,13	2,69
Total	28,08	26,54	17,68	8,93	10,92	2,23	1,77	1,13	0,56	0,24	1,93	100,00

2.2.2.2 Tableau des fréquences lignes

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

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

Les coordonnées du profil ligne moyen (dans le tableau des fréquences) sont calculées par :

$$f_{.j} = \frac{n_{.j}}{n_{..}} = \frac{Effectif\ \ de\ \ la\ \ colonne\ \ j}{Effectif\ \ total}$$

	Hollande	Sarkozy	Le Pen	Bayrou	Melenchon	Joly	Dupond-Aignan	Poutou	Arthaud	Cheminade	Blancs_Nuls	Total
Alsace	18,92	32,27	21,68	11,48	7,16	2,69	1,85	1,07	0,63	0,29	1,96	100,00
Aquitaine	30,47	23,89	15,20	10,51	11,90	2,31	1,60	1,55	0,47	0,22	1,87	100,00
Auvergne	30,08	22,93	16,69	9,76	12,49	1,83	1,82	1,23	0,66	0,24	2,27	100,00
Bourgogne	27,43	26,11	19,95	8,56	10,14	1,78	1,99	1,17	0,62	0,24	2,00	100,00
Bretagne	31,17	25,20	13,00	11,16	10,81	2,90	1,77	1,32	0,62	0,25	1,79	100,00
Centre	26,11	27,35	18,99	9,30	10,30	1,78	2,07	1,19	0,67	0,25	1,98	100,00
Champ-Ard	24,33	28,04	23,50	8,45	8,46	1,38	2,04	1,17	0,68	0,24	1,72	100,00
Corse	23,84	30,84	23,95	4,91	9,68	2,30	1,06	1,14	0,31	0,21	1,78	100,00
Fr-Comte	25,33	26,07	20,83	8,54	10,85	2,11	1,94	1,24	0,68	0,26	2,14	100,00
Ile-de-Fr	31,24	28,56	12,08	9,07	11,65	2,66	1,70	0,82	0,37	0,25	1,60	100,00
Lang-Rouss	25,87	24,31	23,03	6,67	12,92	2,24	1,41	1,08	0,44	0,22	1,79	100,00
Limousin	37,19	20,39	14,99	7,47	12,53	1,61	1,67	1,16	0,57	0,21	2,21	100,00
Lorraine	25,01	24,90	23,23	8,95	10,06	1,71	1,95	1,37	0,71	0,27	1,83	100,00
Midi-Pyr	31,27	22,49	15,90	9,36	12,97	2,53	1,60	1,21	0,47	0,24	1,97	100,00
Nord-PdeC	28,01	23,18	22,86	6,94	11,98	1,52	1,56	1,15	0,74	0,22	1,84	100,00
Basse-Nor	27,05	27,80	16,81	10,38	9,81	1,93	2,15	1,34	0,68	0,26	1,78	100,00
Haute-Nor	27,34	25,44	19,77	7,87	12,04	1,61	1,86	1,29	0,68	0,24	1,87	100,00
Pays-Loire	27,78	28,01	14,08	11,40	9,92	2,41	2,04	1,31	0,65	0,23	2,18	100,00
Picardie	26,11	24,64	24,57	7,12	10,22	1,30	1,94	1,24	0,81	0,24	1,81	100,00
Poit-Char	30,50	25,05	16,09	9,34	10,45	2,04	2,05	1,42	0,65	0,23	2,16	100,00
PACA	21,69	30,55	23,48	6,58	11,06	2,13	1,50	0,83	0,34	0,21	1,62	100,00
Rhone-Alp	24,99	27,83	18,05	9,77	10,99	2,81	1,95	1,02	0,53	0,26	1,81	100,00
Outremer	45,17	26,40	7,61	5,66	5,26	2,04	1,11	0,87	0,67	0,32	4,90	100,00

2.2.2.3 Tableau des fréquences colonnes

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

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

Les coordonnées du profil colonne moyen (dans le tableau des fréquences) sont calculées par :

$$f_{i.} = \frac{n_{i.}}{n_{..}} = \frac{Effectif\ \ de\ \ la\ \ ligne\ \ i}{Effectif\ \ total}$$

	Hollande	Sarkozy	Le Pen	Bayrou	Melenchon	Joly	Dupond-Aignan	Poutou	Arthaud	Cheminade	Blancs_Nuls
Alsace	1,88	3,40	3,43	3,60	1,83	3,37	2,92	2,65	3,17	3,38	2,84
Aquitaine	5,85	4,85	4,63	6,34	5,87	5,59	4,88	7,42	4,52	4,96	5,24
Auvergne	2,48	2,00	2,19	2,53	2,65	1,90	2,39	2,53	2,74	2,28	2,73
Bourgogne	2,59	2,61	2,99	2,54	2,46	2,12	2,99	2,74	2,95	2,63	2,75
Bretagne	6,19	5,29	4,10	6,96	5,51	7,24	5,57	6,54	6,25	5,77	5,18
Centre	3,79	4,20	4,38	4,25	3,85	3,26	4,79	4,31	4,88	4,26	4,18
Champ-Ard	1,76	2,15	2,70	1,92	1,57	1,26	2,35	2,11	2,48	1,99	1,81
Corse	0,38	0,53	0,61	0,25	0,40	0,47	0,27	0,46	0,25	0,38	0,42

Fr-Comte	1,70	1,85	2,22	1,80	1,87	1,78	2,07	2,07	2,29	2,04	2,09
Ile-de-Fr	16,69	16,14	10,25	15,24	16,00	17,90	14,47	10,91	9,99	15,22	12,43
Lang-Rouss	4,02	4,00	5,69	3,26	5,17	4,40	3,48	4,18	3,49	3,96	4,05
Limousin	1,69	0,98	1,08	1,07	1,47	0,93	1,21	1,32	1,32	1,12	1,47
Lorraine	3,27	3,44	4,82	3,68	3,38	2,82	4,05	4,45	4,68	4,11	3,49
Midi-Pyr	5,44	4,14	4,39	5,13	5,80	5,55	4,43	5,22	4,12	4,74	5,00
Nord-PdeC	6,24	5,46	8,08	4,86	6,86	4,25	5,52	6,39	8,34	5,69	5,95
Basse-Nor	2,39	2,60	2,36	2,89	2,23	2,14	3,03	2,95	3,05	2,62	2,30
Haute-Nor	2,82	2,78	3,24	2,56	3,20	2,10	3,06	3,31	3,53	2,91	2,81
Pays-Loire	6,00	6,40	4,83	7,74	5,50	6,55	7,00	7,01	7,12	5,78	6,85
Picardie	2,78	2,78	4,16	2,39	2,80	1,75	3,29	3,29	4,37	2,89	2,81
Poit-Char	3,24	2,82	2,71	3,12	2,85	2,73	3,47	3,76	3,46	2,87	3,34
PACA	5,91	8,81	10,16	5,64	7,75	7,32	6,49	5,66	4,71	6,68	6,43
Rhone-Alp	8,56	10,09	9,82	10,53	9,68	12,12	10,60	8,67	9,08	10,18	9,02
Outremer	4,32	2,68	1,16	1,70	1,29	2,46	1,68	2,07	3,22	3,54	6,83
Total	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00

2.2.2.4 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 :

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

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 différents candidats obtiennent des scores très différents et l'usage de la métrique euclidienne aurait donné trop de poids aux candidats qui ont obtenu des scores élevés (Hollande, Sarkozy, Le Pen).
- 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 Alsace et la ligne Aquitaine est donnée par :

$$d^2_{\Phi^2}(Alsace, Aquitaine) = \frac{(0{,}1892 - 0{,}3047)^2}{0{,}2808} + \ldots + \frac{(0{,}0196 - 0{,}0187)^2}{0{,}0193} = 0{,}1232$$

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

$$d^2_{\Phi^2}(Alsace, Moyenne) = \frac{(0{,}1892 - 0{,}2808)^2}{0{,}2808} + \ldots + \frac{(0{,}0196 - 0{,}0193)^2}{0{,}0193} = 0{,}0729$$

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 Hollande et la colonne Sarkozy est :

$$d^2_{\Phi^2}(\text{Hollande}, \text{Sarkozy}) = \frac{(0{,}0188 - 0{,}0340)^2}{0{,}0279} + \ldots + \frac{(0{,}0432 - 0{,}0268)^2}{0{,}0269} = 0{,}0474$$

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

2.2.2.5 Taux de liaison et Phi-2

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

	Hollande	Sarkozy	Le Pen	Bayrou	Melenchon	Joly	Dupond-Aignan	Poutou	Arthaud	Cheminade	Blancs_Nuls
Alsace	-0,33	0,22	0,23	0,29	-0,34	0,21	0,05	-0,05	0,13	0,21	0,02
Aquitaine	0,09	-0,10	-0,14	0,18	0,09	0,04	-0,09	0,38	-0,16	-0,08	-0,03
Auvergne	0,07	-0,14	-0,06	0,09	0,14	-0,18	0,03	0,09	0,18	-0,02	0,18
Bourgogne	-0,02	-0,02	0,13	-0,04	-0,07	-0,20	0,13	0,03	0,11	-0,01	0,04
Bretagne	0,11	-0,05	-0,26	0,25	-0,01	0,30	0,00	0,17	0,12	0,04	-0,07
Centre	-0,07	0,03	0,07	0,04	-0,06	-0,20	0,17	0,06	0,20	0,04	0,03
Champ-Ard	-0,13	0,06	0,33	-0,05	-0,23	-0,38	0,15	0,04	0,22	-0,02	-0,11
Corse	-0,15	0,16	0,35	-0,45	-0,11	0,03	-0,40	0,01	-0,45	-0,15	-0,08
Fr-Comte	-0,10	-0,02	0,18	-0,04	-0,01	-0,05	0,10	0,10	0,22	0,08	0,11
Ile-de-Fr	0,11	0,08	-0,32	0,02	0,07	0,19	-0,04	-0,27	-0,33	0,01	-0,17
Lang-Rouss	-0,08	-0,08	0,30	-0,25	0,18	0,01	-0,20	-0,04	-0,20	-0,09	-0,07
Limousin	0,32	-0,23	-0,15	-0,16	0,15	-0,28	-0,06	0,03	0,03	-0,12	0,15
Lorraine	-0,11	-0,06	0,31	0,00	-0,08	-0,23	0,10	0,21	0,28	0,12	-0,05
Midi-Pyr	0,11	-0,15	-0,10	0,05	0,19	0,14	-0,09	0,07	-0,16	-0,03	0,02
Nord-PdeC	0,00	-0,13	0,29	-0,22	0,10	-0,32	-0,12	0,02	0,33	-0,09	-0,05
Basse-Nor	-0,04	0,05	-0,05	0,16	-0,10	-0,14	0,22	0,19	0,23	0,06	-0,07
Haute-Nor	-0,03	-0,04	0,12	-0,12	0,10	-0,28	0,05	0,14	0,22	0,00	-0,03
Pays-Loire	-0,01	0,06	-0,20	0,28	-0,09	0,08	0,15	0,16	0,17	-0,05	0,13
Picardie	-0,07	-0,07	0,39	-0,20	-0,06	-0,42	0,10	0,10	0,46	-0,03	-0,06
Poit-Char	0,09	-0,06	-0,09	0,05	-0,04	-0,08	0,16	0,26	0,16	-0,04	0,12
PACA	-0,23	0,15	0,33	-0,26	0,01	-0,04	-0,15	-0,26	-0,38	-0,13	-0,16
Rhone-Alp	-0,11	0,05	0,02	0,09	0,01	0,26	0,10	-0,10	-0,06	0,06	-0,06
Outremer	0,61	0,00	-0,57	-0,37	-0,52	-0,08	-0,37	-0,23	0,20	0,32	1,54

Signification pratique du taux de liaison : le score de Hollande en Alsace est 33% moins élevé que le score théorique que l'on observerait si les votes étaient indépendants des régions. Au contraire, celui de Sarkozy est 22% plus 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).

Ici, le taux de liaison maximum (1,54) est observé entre Outremer et Blancs_Nuls : les votes blancs et nuls y sont deux fois et demie plus nombreux qu'en moyenne nationale. La plus grande valeur suivante est 0,61, observée pour Hollande et Outremer : dans cette "région", Hollande obtient un score supéreur de 61% à sa moyenne nationale.

Le taux de liaison minimum est de -0,57 ; il est observé entre Le Pen et Outremer.

Notez que le coefficient $f_{i.} f_{.j}$ 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 $f_{i.} f_{.j}$ est nulle. La variance de cette série (avec les mêmes pondérations) est le coefficient Φ^2 :

$$\Phi^2 = \sum_{i,j} f_{i.} f_{.j} t_{ij}^2 = \sum_{i,j} \frac{(f_{ij} - f_{i.} f_{.j})^2}{f_{i.} f_{.j}} = \frac{\chi^2}{n_{..}}$$

Ici, on obtient : $\Phi^2 = 0{,}02928$.
La méthode d'analyse factorielle des correspondances peut être vue comme une décomposition pertinente du Φ^2 selon plusieurs axes factoriels.

2.2.3 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.

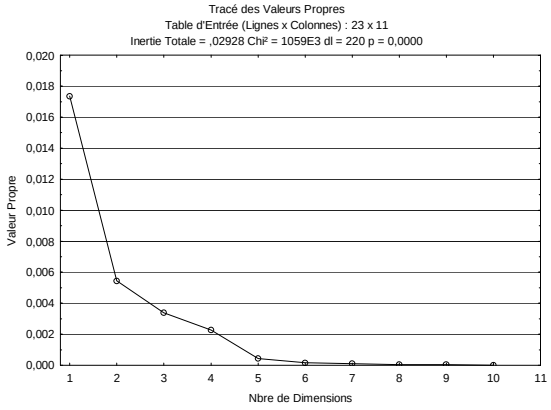
La matrice (tableau de valeurs) dont on recherche les valeurs propres et vecteurs propres est un objet mathématique "compliqué", qui ne possède pas de signification intuitive immédiate. De fait, on part de la matrice dont le terme à l'intersection de la ligne i et de la colonne j vaut : $\frac{f_{ij}}{\sqrt{f_{i\cdot}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

2.2.3.1 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 (Regions dans Presidentielles-2012.stw)
Table d'Entrée (Lignes x Colonnes) : 23 x 11
Inertie Totale = ,02928 Chi² = 1059E3 dl = 220 p = 0,0000

	ValSing.	ValProp.	%age Inertie	%age Cumulé	Chi²
1	0,13173	0,01735	59,26	59,26	627865
2	0,07382	0,00545	18,61	77,87	197176
3	0,05822	0,00339	11,58	89,45	122638
4	0,04775	0,00228	7,79	97,24	82506
5	0,02111	0,00045	1,52	98,76	16121
6	0,01250	0,00016	0,53	99,29	5649
7	0,01094	0,00012	0,41	99,70	4327
8	0,00684	0,00005	0,16	99,86	1692
9	0,00596	0,00004	0,12	99,98	1286
10	0,00229	0,00001	0,02	100,00	190



Le choix du nombre d'axes factoriels à conserver se fait comme dans le cas de l'ACP. Ici, 100% de l'inertie se répartit sur 10 axes ; on pourrait donc choisir de n'étudier que les axes représentant plus de 10% de l'inertie, c'est-à-dire les 3 premiers axes. Cependant, on observe une brusque décroissance des valeurs propres entre la 4è et la 5è valeur propre, alors que rien de tel n'apparaît entre la 3è et la 4è. On retient donc les 4 premiers axes factoriels.

2.2.3.2 Résultats relatifs aux individus-lignes

NomLign Nom	Coordonnées Ligne et Contributions à l'Inertie (Regions dans Presidentielles-2012.stw) Table d'Entrée (Lignes x Colonnes) : 23 x 11 Standardisation : Profils ligne et colonne															
	Ligne Numéro	Coord. Dim.1	Coord. Dim.2	Coord. Dim.3	Coord. Dim.4	Masse	Qualité	Inertie Relative	Inertie Dim.1	Cosinus² Dim. 1	Inertie Dim.2	Cosinus² Dim. 2	Inertie Dim.3	Cosinus² Dim. 3	Inertie Dim.4	Cosinus² Dim. 4
Alsace	1	-0,161	0,170	0,107	0,070	0,028	0,979	0,070	0,042	0,356	0,149	0,399	0,095	0,158	0,059	0,067
Aquitaine	2	0,081	-0,002	-0,071	0,032	0,054	0,913	0,025	0,020	0,477	0,000	0,000	0,079	0,361	0,024	0,074
Auvergne	3	0,046	-0,047	-0,067	0,036	0,023	0,876	0,009	0,003	0,185	0,009	0,193	0,030	0,385	0,013	0,114
Bourgogi	4	-0,053	-0,028	0,020	0,028	0,026	0,914	0,005	0,004	0,540	0,004	0,149	0,003	0,075	0,009	0,150
Bretagne	5	0,135	0,046	-0,044	0,036	0,056	0,963	0,047	0,059	0,745	0,022	0,087	0,032	0,080	0,031	0,051
Centre	6	-0,048	0,016	0,019	0,032	0,041	0,817	0,007	0,005	0,472	0,002	0,055	0,004	0,075	0,018	0,215
Champa	7	-0,159	-0,009	0,070	0,050	0,020	0,934	0,024	0,030	0,723	0,000	0,002	0,029	0,139	0,022	0,070
Corse	8	-0,183	-0,035	0,108	-0,093	0,005	0,931	0,009	0,009	0,566	0,001	0,020	0,016	0,198	0,017	0,147
Franche-	9	-0,089	-0,014	0,006	0,024	0,019	0,930	0,006	0,009	0,842	0,001	0,021	0,000	0,004	0,005	0,063
Ile-de-Fr	10	0,137	0,047	-0,006	-0,069	0,150	0,983	0,134	0,162	0,715	0,061	0,084	0,001	0,001	0,315	0,183
Langued	11	-0,134	-0,082	-0,041	-0,050	0,044	0,956	0,045	0,045	0,596	0,054	0,222	0,021	0,055	0,048	0,083
Limousin	12	0,146	-0,169	-0,067	-0,003	0,013	0,987	0,024	0,016	0,387	0,067	0,518	0,017	0,082	0,000	0,000
Lorraine	13	-0,139	-0,033	-0,000	0,063	0,037	0,989	0,031	0,041	0,782	0,008	0,045	0,000	0,000	0,065	0,162
Midi-Pyre	14	0,075	-0,046	-0,088	-0,004	0,049	0,940	0,027	0,016	0,340	0,019	0,130	0,111	0,469	0,000	0,001
Nord-Pa	15	-0,110	-0,121	-0,030	0,001	0,063	0,988	0,060	0,044	0,433	0,168	0,523	0,017	0,032	0,000	0,000
Basse-N	16	0,007	0,054	0,013	0,050	0,025	0,803	0,006	0,000	0,007	0,013	0,412	0,001	0,026	0,027	0,359
Haute-N	17	-0,057	-0,051	-0,023	-0,001	0,029	0,828	0,008	0,006	0,423	0,014	0,335	0,005	0,069	0,000	0,000
Pays-de-	18	0,078	0,088	0,011	0,054	0,061	0,974	0,036	0,021	0,347	0,087	0,451	0,002	0,007	0,078	0,169
Picardie	19	-0,164	-0,097	0,017	0,038	0,030	0,971	0,040	0,046	0,688	0,052	0,240	0,003	0,007	0,018	0,036
Poitou-C	20	0,060	-0,023	-0,007	0,043	0,030	0,914	0,007	0,006	0,549	0,003	0,081	0,000	0,007	0,024	0,277
Provence	21	-0,195	0,017	0,058	-0,081	0,077	0,995	0,127	0,168	0,785	0,004	0,006	0,075	0,068	0,221	0,136
Rhone-A	22	-0,035	0,068	-0,003	-0,004	0,096	0,868	0,022	0,007	0,183	0,082	0,682	0,000	0,002	0,001	0,002
Outreme	23	0,395	-0,192	0,240	0,018	0,027	0,996	0,232	0,242	0,620	0,182	0,146	0,457	0,229	0,004	0,001

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 régions ont ici un "poids" dépendant de l'effectif total d'électeurs inscrits dans la région.

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

$$QLT(L_i, F_1; F_2; F_3; F_4) = \frac{(Coord\,L_i\,selon F_1)^2 + (Coord\,L_i\,selon F_2)^2 + (Coord\,L_i\,selon F_3)^2 + (Coord\,L_i\,selon F_4)^2}{\sum_l (Coord\ de\ L_i\,selon F_l)^2}$$

Par exemple :

$$QLT(Alsace, F_1; F_2; F_3; F_4) = \frac{(-0{,}161)^2 + (0{,}170)^2 + (-0{,}107)^2 + (0{,}070)^2}{(-0{,}161)^2 + (0{,}170)^2 + (-0{,}107)^2 + (0{,}070)^2 + ... + (-0{,}0002)^2}$$

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 :

$$(-0{,}161 \times 0{,}028) + (0{,}081 \times 0{,}054) + \ldots + (0{,}395 \times 0{,}027) = 0$$

$$(-0{,}161)^2 \times 0{,}028 + (0{,}081)^2 \times 0{,}054 + \ldots + (0{,}395)^2 \times 0{,}027 = 0{,}01735$$

$$(-0{,}161) \times (0{,}170) \times 0{,}028 + (0{,}081) \times (-0{,}002) \times 0{,}054 + \ldots + (0{,}395) \times (-0{,}192) \times 0{,}027 = 0$$

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 l'Alsace et l'axe factoriel N°1 :

$$Ctr(Alsace, F_1) = \frac{0{,}028 \times (-0{,}161)^2}{0{,}01735} = 0{,}042$$

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 :

$$QLT(Alsace, F_1) = \frac{(-0{,}161)^2}{(-0{,}161)^2 + (0{,}170)^2 + (-0{,}107)^2 + (0{,}070)^2 + ... + (-0{,}0002)^2} = 0{,}356$$

L'interprétation géométrique de ces valeurs est analogue à celle développée pour l'ACP : c'est le carré du cosinus de l'angle entre le vecteur représentant l'Alsace dans l'espace à 10 dimensions et sa projection sur le premier axe factoriel.

2.2.3.3 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.

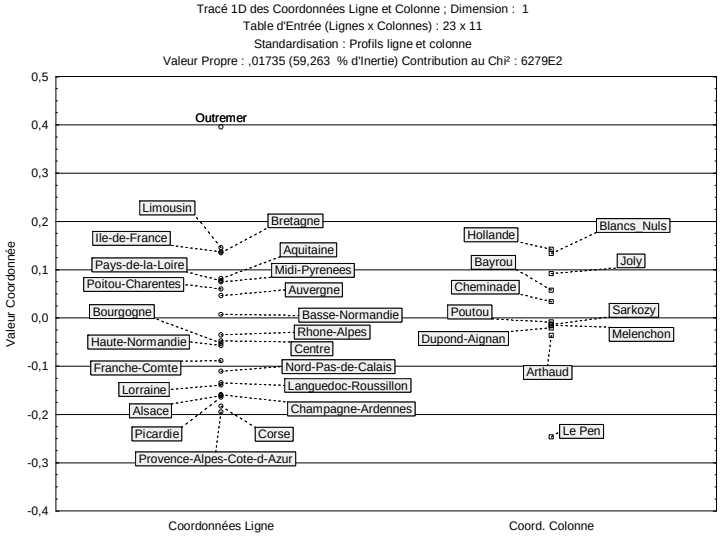
Nom Col Nom	Coordonnées Colonne et Contributions à l'Inertie (Regions dans Presidentielles-2012.stw) Table d'Entrée (Lignes x Colonnes) : 23 x 11 Standardisation : Profils ligne et colonne															
	Colonne Nu	Coord. Dim.1	Coord. Dim.2	Coord. Dim.3	Coord. Dim.4	Masse	Qualité	Inertie Relative	Inertie Dim.1	Cosinus² Dim.1	Inertie Dim.2	Cosinus² Dim.2	Inertie Dim.3	Cosinus² Dim.3	Inertie Dim.4	Cosinus² Dim.4
Hollande	1	0,142	-0,069	0,003	-0,000	0,281	0,996	0,241	0,327	0,805	0,248	0,191	0,001	0,000	0,000	0,000
Sarkozy	2	-0,014	0,067	0,058	-0,032	0,265	0,987	0,084	0,003	0,022	0,222	0,489	0,264	0,363	0,123	0,113
Le Pen	3	-0,247	-0,054	0,004	0,016	0,177	0,999	0,387	0,619	0,949	0,095	0,046	0,001	0,000	0,019	0,004
Bayrou	4	0,057	0,130	-0,059	0,090	0,089	0,996	0,097	0,017	0,103	0,276	0,527	0,091	0,109	0,320	0,256
Melenchc	5	-0,016	-0,016	-0,120	-0,061	0,109	0,983	0,071	0,002	0,013	0,005	0,013	0,466	0,761	0,180	0,197
Joly	6	0,092	0,140	-0,026	-0,060	0,022	0,741	0,033	0,011	0,194	0,081	0,450	0,004	0,016	0,035	0,081
Dupond-/	7	-0,020	0,066	-0,021	0,074	0,018	0,601	0,011	0,000	0,024	0,014	0,243	0,002	0,025	0,042	0,310
Poutou	8	-0,009	-0,024	-0,079	0,152	0,011	0,839	0,014	0,000	0,002	0,001	0,016	0,021	0,174	0,115	0,646
Arthaud	9	-0,036	-0,078	0,022	0,207	0,006	0,824	0,012	0,000	0,021	0,006	0,100	0,001	0,008	0,105	0,695
Chemina	10	0,034	0,014	0,048	0,033	0,002	0,571	0,001	0,000	0,137	0,000	0,023	0,002	0,282	0,001	0,129
Blancs_N	11	0,134	-0,122	0,161	0,084	0,019	0,878	0,049	0,020	0,239	0,053	0,199	0,147	0,345	0,060	0,095

2.2.3.4 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.

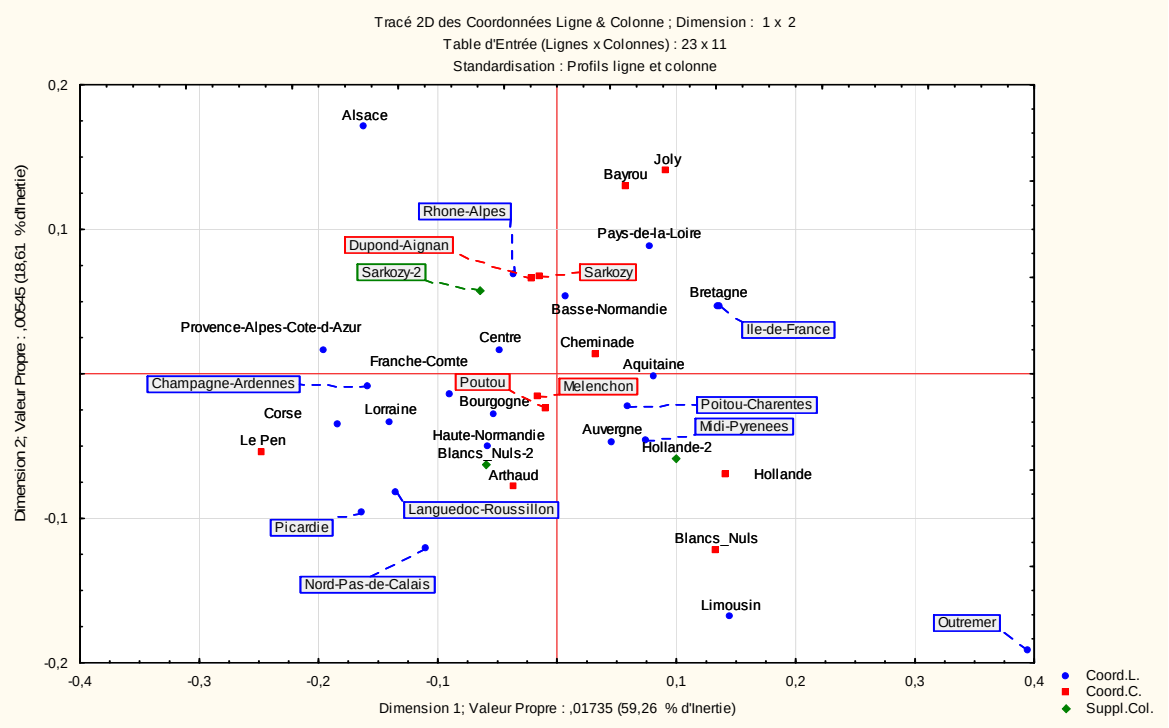
On peut réaliser et essayer d'interpréter des graphiques :

- 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.

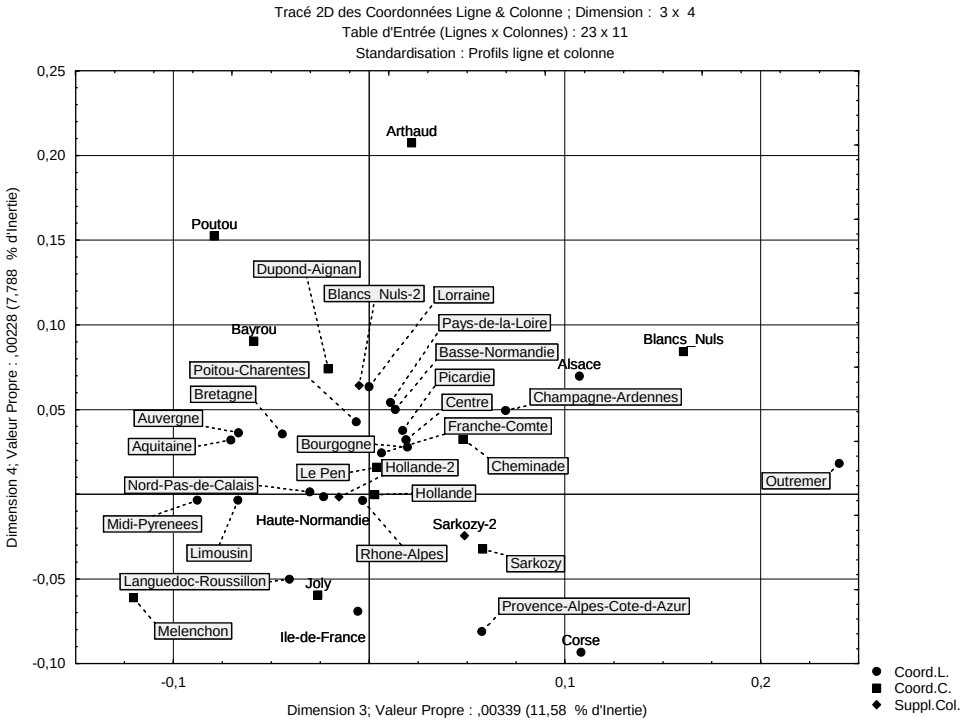


Pour les graphiques à deux dimensions, trois individus colonnes supplémentaires ont été introduits : les scores des deux candidats Hollande et Sarkozy au second tour des élections, ainsi que le nombre de bulletins blancs et nuls au second tour. Ces individus n'interviennent pas dans le calcul des axes factoriels de l'analyse. En revanche, leurs coordonnées subissent les mêmes transformations que celles des individus colonnes actifs et leurs positions sur les graphiques peuvent donner des indications sur les reports de voix au second tour.

Graphique selon les axes 1 et 2



Graphique selon les axes 3 et 4



2.2.3.5 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 :

$$d^2_{\Phi^2}(Alsace, Aquitaine) = \frac{(0{,}1892 - 0{,}3047)^2}{0{,}2808} + \ldots + \frac{(0{,}0196 - 0{,}0187)^2}{0{,}0193} = 0{,}1232$$

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

	Fact. 1	Fact. 2	Fact. 3	Fact. 4	Fact. 5	Fact. 6	Fact. 7	Fact. 8	Fact. 9	Fact. 10
Alsace	-0,1610	0,1704	0,1074	0,0696	-0,0312	-0,0149	-0,0024	-0,0174	-0,0004	0,0025
Aquitaine	0,0811	-0,0021	-0,0706	0,0320	-0,0122	-0,0040	-0,0320	0,0009	0,0028	0,0010
Auvergne	0,0462	-0,0472	-0,0666	0,0362	-0,0008	0,0327	0,0021	-0,0184	0,0038	0,0007
Bourgogne	-0,0526	-0,0277	0,0197	0,0278	0,0168	-0,0012	0,0043	0,0029	0,0112	-0,0018
Bretagne	0,1353	0,0462	-0,0442	0,0355	-0,0150	-0,0239	0,0030	0,0031	-0,0098	-0,0009
Centre	-0,0476	0,0163	0,0190	0,0321	0,0248	0,0159	0,0017	-0,0006	0,0022	0,0018
Champ.Ard	-0,1590	-0,0086	0,0697	0,0495	0,0393	-0,0237	-0,0035	-0,0028	0,0129	-0,0033
Corse	-0,1831	-0,0347	0,1082	-0,0933	-0,0236	-0,0292	-0,0409	0,0271	-0,0163	-0,0004
Fr-Comte	-0,0890	-0,0139	0,0064	0,0244	-0,0118	0,0171	0,0098	0,0100	-0,0040	0,0035
Ile de Fr	0,1369	0,0469	-0,0057	-0,0692	0,0197	-0,0057	0,0031	-0,0025	0,0007	0,0010
Lang.Rous	-0,1342	-0,0819	-0,0408	-0,0501	-0,0363	0,0011	-0,0003	0,0037	0,0001	0,0001
Limousin	0,1457	-0,1686	-0,0670	-0,0035	0,0208	-0,0059	0,0050	-0,0037	0,0140	-0,0046
Lorraine	-0,1394	-0,0334	0,0000	0,0634	0,0034	-0,0135	0,0009	0,0015	0,0069	0,0051
Midi-Pyr	0,0747	-0,0462	-0,0877	-0,0036	-0,0300	0,0016	0,0020	-0,0035	0,0078	-0,0001
Nord-PdC	-0,1102	-0,1212	-0,0302	0,0014	0,0073	-0,0059	0,0060	-0,0101	-0,0108	-0,0014
Basse-Nor	0,0069	0,0537	0,0134	0,0501	0,0362	0,0022	-0,0054	0,0049	-0,0017	0,0024
Haute-Nor	-0,0574	-0,0511	-0,0232	-0,0015	0,0254	0,0220	-0,0040	0,0058	-0,0114	0,0053
Pays Loire	0,0775	0,0884	0,0108	0,0541	0,0049	0,0185	-0,0053	-0,0014	-0,0053	-0,0046
Picardie	-0,1640	-0,0969	0,0171	0,0375	0,0277	-0,0145	0,0112	0,0041	-0,0022	-0,0018
Poit. Char	0,0601	-0,0230	-0,0066	0,0427	0,0136	0,0038	-0,0019	0,0187	0,0028	-0,0023
PACA	-0,1950	0,0165	0,0575	-0,0811	-0,0055	0,0059	-0,0126	-0,0002	0,0007	-0,0018
Rhone-Alp	-0,0353	0,0682	-0,0034	-0,0037	-0,0229	0,0049	0,0177	0,0055	0,0021	-0,0002
Outremer	0,3955	-0,1919	0,2401	0,0181	-0,0308	0,0045	0,0000	-0,0004	-0,0003	0,0011

On vérifie que :

$$d^2_{eucl}(Alsace', Aquitaine') = (-0{,}1610 - 0{,}0811)^2 + \ldots + (0{,}0025 + 0{,}0010)^2 = 0{,}1232$$

De même, on avait établi que :

$$d^2_{\Phi^2}(Alsace, Moyenne) = \frac{(0{,}1892 - 0{,}2808)^2}{0{,}2808} + \ldots + \frac{(0{,}0196 - 0{,}0193)^2}{0{,}0193} = 0{,}0729$$

Et l'on a :

$$d^2_{eucl}(Alsace', O) = (-0{,}1610)^2 + \ldots + (-0{,}0025)^2 = 0{,}0729$$

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

	Fact. 1	Fact. 2	Fact. 3	Fact. 4	Fact. 5	Fact. 6	Fact. 7	Fact. 8	Fact. 9	Fact. 10
Hollande	0,1422	-0,0693	0,0029	-0,0004	0,0049	-0,0080	0,0003	-0,0003	0,0012	-0,0002
Sarkozy	-0,0144	0,0675	0,0581	-0,0324	0,0100	0,0023	-0,0036	0,0002	-0,0018	0,0000
Le Pen	-0,2465	-0,0542	0,0039	0,0158	-0,0068	-0,0063	0,0009	-0,0010	0,0019	-0,0002
Bayrou	0,0575	0,1298	-0,0589	0,0905	-0,0052	-0,0023	-0,0032	-0,0091	0,0034	-0,0003
Melenchon	-0,0156	-0,0157	-0,1203	-0,0612	-0,0003	0,0175	-0,0001	-0,0013	-0,0023	0,0003
Joly	0,0922	0,1405	-0,0261	-0,0596	-0,0991	-0,0255	0,0246	0,0165	-0,0047	-0,0011
Dupond-Ai.	-0,0204	0,0655	-0,0210	0,0740	0,0607	0,0226	0,0421	0,0279	0,0179	-0,0013
Poutou	-0,0088	-0,0237	-0,0792	0,1525	-0,0017	0,0015	-0,0627	0,0405	-0,0151	0,0025
Arthaud	-0,0364	-0,0785	0,0217	0,2074	0,0498	0,0071	0,0646	-0,0086	-0,0642	-0,0010
Cheminade	0,0335	0,0136	0,0480	0,0325	-0,0142	-0,0087	0,0331	-0,0038	0,0066	0,0456
Blancs_Nuls	0,1337	-0,1219	0,1606	0,0842	-0,0731	0,0609	-0,0002	-0,0030	0,0035	-0,0005

On avait établi que :

$$d^2_{\Phi^2}(\text{Hollande},\text{Sarkozy}) = \frac{(0{,}0188-0{,}0340)^2}{0{,}0279} + \ldots + \frac{(0{,}0432-0{,}0268)^2}{0{,}0269} = 0{,}0474$$

On retrouve ici :

$$d^2_{eucl}(\text{Hollande}',\text{Sarkozy}') = (0{,}1422+0{,}0144)^2 + \ldots + (-0{,}0002-0{,}0000)^2 = 0{,}0474$$

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).

2.2.3.6 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 Alsace et le candidat Hollande peut être retrouvé à l'aide du calcul suivant :

$$t_{11} = \frac{(-0{,}1610)(0{,}1422)}{\sqrt{0{,}01735}} + \frac{(0{,}1704)(-0{,}0693)}{\sqrt{0{,}00545}} + \ldots + \frac{(0{,}0025)(-0{,}0002)}{\sqrt{0{,}00001}} = -0{,}3263$$

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

Remarque. On obtient ainsi une décomposition "additive" du taux de liaison. L'étude des différents termes de cette somme peut nous indiquer quels sont les axes sur lesquels apparaît le plus clairement la liaison entre la modalité ligne et la modalité colonne. Par exemple, pour la liaison entre Alsace et Hollande, les différents termes de la somme sont :

-0,1738 - 0,1600 + 0,0053 - 0,0006 - 0,0072 + 0,0095 - 0,0001 + 0,0009 - 0,0001 - 0,0002 = - 0,3263

Autrement dit, l'essentiel de la liaison entre ces deux modalités apparaît sur le plan (F_1, F_2).

2.2.4 Interprétation des résultats de l'AFC

Au niveau global, on pourra noter que les inerties relatives les plus fortes sont observées sur l'Outremer, l'Ile de France, Provence-Alpes Côte d'Azur, l'Alsace et le Nord Pas-de-Calais pour les régions, et sur Le Pen, Hollande, Bayrou et Sarkozy pour les candidats. Ce sont donc essentiellement ces modalités lignes et modalités colonnes qui vont apparaître dans l'étude qui suit. On pourra noter que ces modalités correspondent soit à des modalités de poids important (Ile de France, Nord Pas-de-Calais, Hollande, Sarkozy) soit à des modalités éloignées du profil moyen (Outremer, Alsace).

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.

2.2.4.1 Interprétation des axes

Pour le premier axe :

- Points lignes :

-	+
Provence - Alpes Côte d'Azur (17%) Picardie (5%) Languedoc-Roussillon (5%) Nord - Pas de Calais (4%) Alsace (4%) Lorraine (4%)	Outremer (24%) Ile de France (16%) Bretagne (6%)

- Points colonnes :

-	+
Le Pen (62%)	Hollande (33%)

Le premier axe oppose les régions du Nord et de l'Est, et les régions du Sud-Est (PACA, Languedoc-Roussillon) d'une part, à des régions telles que l'Outremer, l'Ile de France et la Bretagne.

Pour les modalités colonnes, cet axe est essentiellement unipolaire (la modalité Le Pen représente plus de la moitié de son inertie) et oppose les modalités Le Pen et Hollande, dont l'inertie est également importante (33%). Les autres modalités interviennent peu.

La synthèse entre l'analyse des lignes et des colonnes montre que cet axe oppose les régions où le vote pour la candidate Le Pen est supérieur à la moyenne nationale à celles où ce vote est inférieur à la moyenne (particulièrement l'Outremer, notamment). On constate que ces dernières sont également des régions de fort vote "Hollande". Il ne faudrait pas pour autant en conclure que les deux candidats recrutent leurs voix dans le même électorat. C'est vraisemblablement plutôt l'ensemble du corps électoral qui possède une sensibilité plus "à gauche" dans certaines régions. Il faut également remarquer que le candidat Sarkozy intervient très peu dans la formation de cet axe.

Pour le deuxième axe :

- Points lignes :

-	+
Outremer (18%) Nord-Pas-de-Calais (17%) Limousin (7%) Languedoc-Roussillon (5%) Picardie (5%)	Alsace (15%) Pays de la Loire (9%) Rhone-Alpes (8%) Ile de France (6%)

- Points colonnes :

-	+
Hollande (25%) Le Pen (10%)	Bayrou (28%) Sarkozy (22%)

En ce qui concerne les modalités lignes, cet axe ne correspond pas à une répartition géographique. S'agissant des modalités colonnes, il oppose les votes pour Bayrou et Sarkozy aux votes pour Hollande et, dans une mesure moindre pour Le Pen.

La partie positive de l'axe correspond aux régions où les votes Sarkozy et Bayrou sont supérieurs à la moyenne nationale, tandis que la partie négative correspond à des régions où ce vote est inférieur à la moyenne, au profit de Hollande. Il semble par ailleurs qu'un faible vote pour Sarkozy ou Bayrou soit lié à un vote plus significatif pour Le Pen.

Pour le troisième axe :

- Points lignes :

-	+
Midi-Pyrénées (11%) Aquitaine (8%)	Outremer (46%) Alsace (10%) Provence-Alpes Côte d'Azur (7%)

- Points colonnes :

-	+
Mélenchon (47%) Bayrou (9%)	Sarkozy (26%) Blancs et nuls (15%)

Le troisième axe est fortement influencé par le vote pour Mélenchon, bien représenté dans des régions telles que Midi-Pyrénées et Aquitaine (partie négative de l'axe), faible dans la "région" Outremer, l'Alsace et Provence Alpes Côte d'Azur.(partie positive de l'axe). S'ajoute "en miroir" à cet effet celui du vote pour Sarkozy, bien représenté dans des régions telles que l'Alsace et Provence Alpes Côte d'Azur (partie positive de l'axe), faible dans les régions Midi-Pyrénées et Aquitaine. Enfin, le vote "Blanc et nuls" est fortement associé à la "région" Outremer sur la partie positive de l'axe.

Pour le quatrième axe :

- Points lignes :

-	+
Ile de France (32%) Provence-Alpes Côte d'Azur (22)% Languedoc-Roussillon (5%)	Pays de la Loire (8%) Lorraine (6%) Alsace (6%)

- Points colonnes :

-	+
Mélenchon (18%) Sarkozy (12%)	Bayrou (32%) Poutou (12%) Arthaud (10%)

Le quatrième axe combine plusieurs effets : d'une part, le vote pour Bayrou, élevé en Pays de la Loire et en Alsace, faible en Provence-Alpes Côte d'Azur et en Languedoc Roussillon, d'autre part le vote Mélenchon, fort en Languedoc Roussillon et les votes Poutou et Arthaud, faibles numériquement mais surreprésentés en Lorraine et en Pays de la Loire. On peut également noter la valeur élevée du cos^2 concernant ces modalités colonnes sur cet axe (respectivement 0,65 et 0,70) : très peu représentées sur les trois premiers axes, elles apparaissent sur le quatrième. La présence de la modalité colonne Sarkozy est plus étonnante, mais il s'agit sans doute d'une correction par rapport aux axes 2 et 3.

L'interprétation des points colonnes supplémentaires

Sur les graphiques, les points étiquetés Hollande-2, Sarkozy-2 et Blancs_Nuls-2 sont des modalités colonnes supplémentaires correspondant aux scores des deux candidats et aux effectifs de bulletins blancs et nuls observés au second tour de l'élection. Il est intéressant d'observer comment se positionnent ces modalités colonnes supplémentaires par rapport aux modalités colonnes actives correspondantes.

Sur le premier axe, les points Hollande-2, Sarkozy-2 et Blancs_Nuls-2 sont décalés vers la partie négative de l'axe par rapport aux points Hollande, Sarkozy et Blancs_Nuls. Nous avions vu précédemment que cet axe opposait principalement le vote Le Pen (partie négative) au vote Hollande (partie positive). Ce décalage est faible pour Hollande. En revanche, il est nettement plus important pour Sarkozy et pour Blancs_Nuls, ce qui semble confirmer :

- d'une part qu'une partie des électeurs qui ont voté Le Pen au premier tour a voté blanc au second tour
- d'autre part qu'une autre partie de ces électeurs a voté pour Sarkozy au second tour.

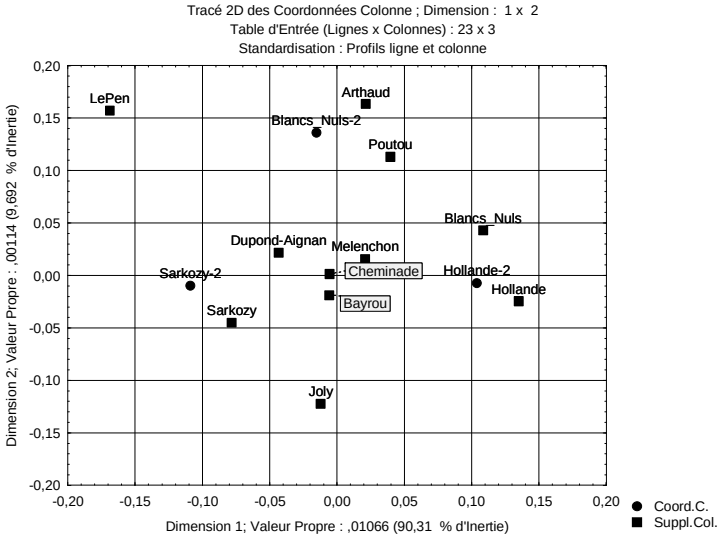
Cette observation pourrait être dans une certaine mesure confirmée par l'étude du deuxième axe : le point Sarkozy-2 y est décalé vers la partie négative par rapport au point Sarkozy, c'est-à-dire en direction du point Le Pen.

Sur le troisième axe, nous avions observé que la position de Blancs_Nuls était notamment due à l'importance de ce vote dans la "région" Outremer. Le point Blancs_Nuls-2 se positionne à proximité de l'origine de l'axe, ce qui semble montrer que cette spécificité du vote Outremer tend à disparaître au second tour. Par ailleurs, le point Hollande-2 a une coordonnée négative, alors que le point Hollande a une coordonnée positive, sans doute en raison des reports de voix des électeurs de Mélenchon.

2.2.4.2 Remarques :

1. Etant donné le poids des suffrages obtenus par les candidats Hollande et Sarkozy (respectivement 28% et 27% de l'ensemble), on aurait pu s'attendre à ce que ces modalités colonnes aient une grande influence dans la détermination du profil moyen et donc que les points représentant ces candidats soient très proches de l'origine. On remarque malgré tout que ces points restent bien distincts de l'origine.
2. Il est tout à fait remarquable que l'étude ne montre pas d'opposition entre les votes pour les deux candidats arrivés en tête. Mais, dans des régions telles que l'Outremer ou l'Ile de France, ces candidats obtiennent tous les deux des scores supérieurs ou égaux à leur moyenne nationale, alors qu'ils obtiennent simultanément des scores inférieurs ou égaux à leur moyenne nationale dans le Nord Pas de Calais. Et, l'Ile de France et le Nord Pas-de-Calais sont très importantes numériquement.
3. L'AFC présente souvent l'inconvénient de mettre en évidence des modalités de faible fréquence, mais présentant des taux de liaison élevés. Ici, cet effet est très limité, puisqu'il faut attendre le quatrième axe pour faire apparaître le rôle joué par les "petits" candidats.
4. On peut mener une étude analogue en prenant comme modalités actives les scores du second tour et comme modalités supplémentaires les scores des candidats du premier tour. Il n'y a alors plus

que deux axes factoriels, et la projection des modalités supplémentaires sur le plan (F1, F2) confirme dans une large mesure ce qui a été dit précédemment et montre même plus clairement les reports de voix des électeurs de Mélenchon, Poutou, Arthaud pour Hollande et ceux de Dupond-Aignan et Le Pen pour Sarkozy. Par contre, les qualités de représentation des points supplémentaires peuvent être très faibles (seulement 1% pour Bayrou, par exemple).



2.2.5 Quelques principes d'interprétation supplémentaires

2.2.5.1 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.

2.2.5.2 Situations où il vaut mieux éviter d'utiliser l'AFC

L'AFC peut être utilisée dans des situations variées, y compris sur des données qui ne constituent pas strictement un tableau de contingence. En revanche, comme l'indique Philippe Cibois dans son article "les pièges de l'AFC", il existe des situations où il vaut mieux s'abstenir d'utiliser cette méthode :

- L'AFC mettra toujours en évidence des attractions - répulsions entre modalités lignes et modalités colonne. Mais, lorsqu'on travaille sur un échantillon et que le khi-2 du tableau de contingence n'est pas significatif, l'effet mis en évidence n'est rien d'autre que le fruit du hasard.

- L'AFC n'a d'intérêt que si notre étude porte sur les liaisons existant entre lignes et colonnes. Au contraire, s'il s'agit de faire un classement multicritère sur un ensemble d'objets statistiques (par exemple, classer les pays selon leurs succès en termes de prix Nobel), la méthode ne fournit aucun résultat pertinent.

2.2.5.3 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.

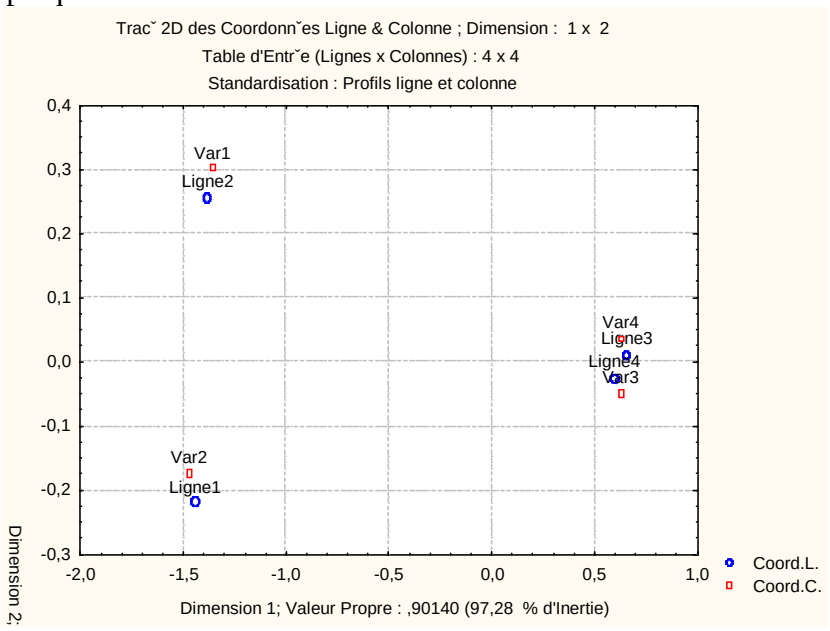
Exemple : Soit le tableau de contingence suivant :

	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 :



2.2.5.4 L'effet Guttman.

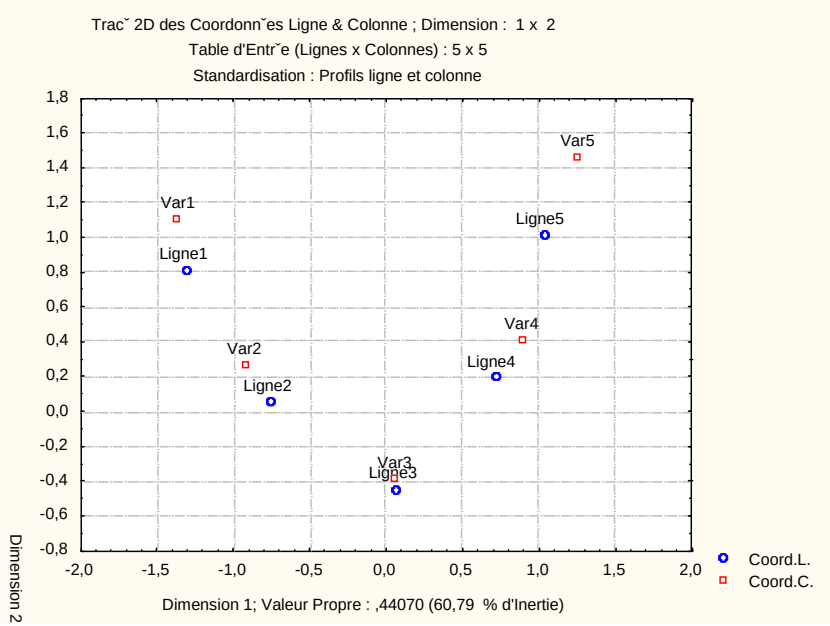
Un nuage de points de forme parabolique indique une redondance entre les deux variables étudiées : la connaissance de la ligne i donne pratiquement celle de la colonne j. Dans un tel cas, pratiquement toute l'information est contenue dans le premier facteur. Cette configuration se rencontre notamment 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 :



2.3 Analyse factorielle des correspondances avec Statistica

2.3.1 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 (Centre d'Etude des Supports de Publicité).

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

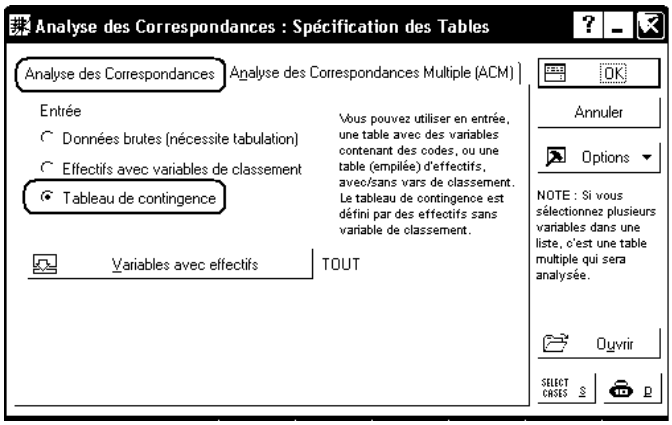
Ouvrier qual	385	457	42	174	104	220
Ouvrier n-q	156	185	8	69	42	85
Inactif	1474	1931	181	852	642	782
Sexe						
Homme	1630	1900	285	854	621	776
Femme	1667	2069	152	815	683	938
Age						
15-24 ans	660	713	69	216	234	360
25-34 ans	640	719	84	230	212	380
35-49 ans	888	1000	130	429	345	466
50-64 ans	617	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

2.3.2 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.



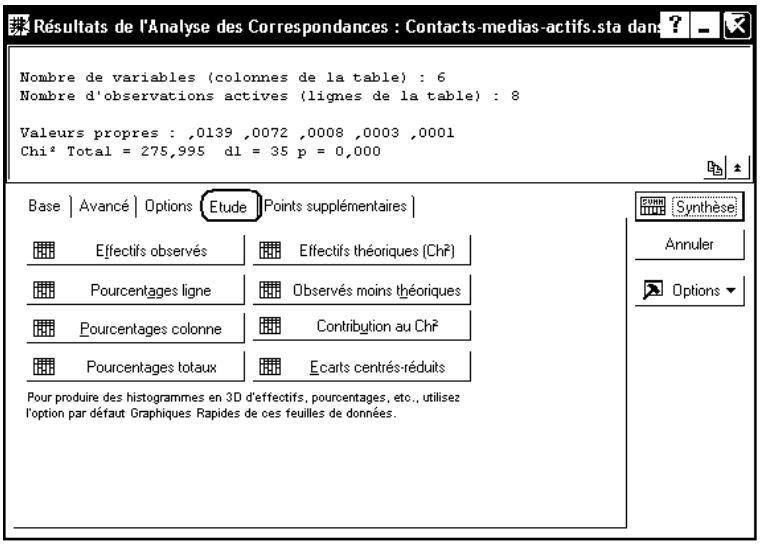
La fenêtre de dialogue permet d'indiquer la manière dont se présentent nos données. La situation la plus classique est celle d'un tableau de contingence : les modalités lignes sont indiquées comme noms d'observations (elles auraient pu être indiquées dans une variable spécifique), les modalités colonnes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

2.3.2.1 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

Statistica ne permet pas d'obtenir directement le tableau des taux de liaison, qui est pourtant un outil exploratoire intéressant. Mais on peut utiliser les tableaux "Observés moins théoriques" et "Effectifs théoriques". Le tableau "Observés moins théoriques" fournit le signe des taux de liaison et, on peut même recopier ces deux tableaux dans une feuille Excel et diviser chaque cellule du premier par la cellule correspondante du second pour obtenir le tableau des taux de liaison.

Observés moins théoriques (recopié depuis Statistica)							
	Radio	Tél.	Quot.N.	Quot R.	P.Mag.	P.TV	
Agriculteur	1,7848	4,5817	-10,5163	23,3637	12,7654	-31,9793	
Petit patron	6,2271	-3,3700	-4,3802	17,4639	3,2456	-19,1865	
Prof. Cad. S.	7,7633	-38,9919	49,3917	-30,6577	29,7930	-17,2984	

Prof. interm	25,1900	-38,0515	18,5211	-24,2837	8,6805	9,9435	
Employé	17,0355	-1,6451	-8,6222	-32,7540	-23,2186	49,2044	
Ouvrier qual	17,1881	14,2201	-6,8631	-11,9698	-41,3621	28,7869	
Ouvrier n-q	10,9512	10,3871	-11,2695	-4,3383	-15,3244	9,5940	
Inactif	-86,1400	52,8697	-26,2615	63,1758	25,4206	-29,0646	
Effectifs théoriques (recopié depuis Statistica)							
	Radio	Tél.	Quot.N.	Quot R.	P.Mag.	P.TV	Total
Agriculteur	94,215	113,418	12,5163	47,636	37,235	48,979	354,00
Petit patron	115,773	139,370	15,3802	58,536	45,754	60,186	435,00
Prof. Cad. S.	185,237	222,992	24,6083	93,658	73,207	96,298	696,00
Prof. interm	334,810	403,052	44,4789	169,284	132,320	174,057	1258,00
Employé	493,964	594,645	65,6222	249,754	195,219	256,796	1856,00
Ouvrier qual	367,812	442,780	48,8631	185,970	145,362	191,213	1382,00
Ouvrier n-q	145,049	174,613	19,2695	73,338	57,324	75,406	545,00
Inactif	1560,140	1878,130	207,2615	788,824	616,579	811,065	5862,00
Total	3297,000	3969,000	438,0000	1667,000	1303,000	1714,000	12388,00
Taux de liaison (calculé sous Excel - division terme à terme)							
	Radio	Tél.	Quot.N.	Quot R.	P.Mag.	P.TV	
Agriculteur	0,0189	0,0404	-0,8402	0,4905	0,3428	-0,6529	
Petit patron	0,0538	-0,0242	-0,2848	0,2983	0,0709	-0,3188	
Prof. Cad. S.	0,0419	-0,1749	2,0071	-0,3273	0,4070	-0,1796	
Prof. interm	0,0752	-0,0944	0,4164	-0,1434	0,0656	0,0571	
Employé	0,0345	-0,0028	-0,1314	-0,1311	-0,1189	0,1916	
Ouvrier qual	0,0467	0,0321	-0,1405	-0,0644	-0,2845	0,1505	
Ouvrier n-q	0,0755	0,0595	-0,5848	-0,0592	-0,2673	0,1272	
Inactif	-0,0552	0,0282	-0,1267	0,0801	0,0412	-0,0358	

2.3.2.2 Choix des valeurs propres

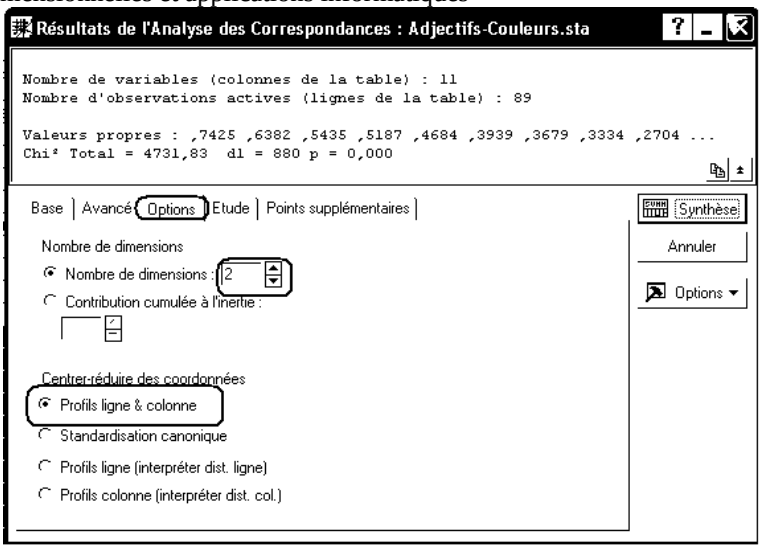
C'est ensuite l'onglet "Avancé" qui nous permettra d'afficher les valeurs propres, et donc de choisir le nombre d'axes à garder :

	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				
Nombre de Dims.	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.

2.3.2.3 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.

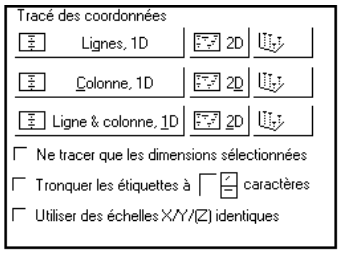


On retourne ensuite sous l'onglet "Avancé" pour afficher les coordonnées des individus-lignes et des individus-colonnes. On notera que Statistica produit deux tableaux de résultats, et on passera de l'un à l'autre à l'aide des onglets du classeur.

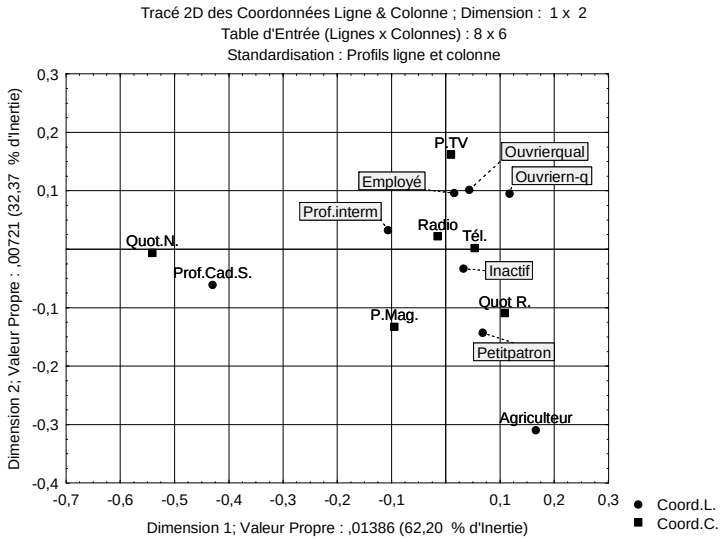
	Coordonnées Ligne et Contributions à l'Inertie Table d'Entrée (Lignes x Colonnes) : 8 x 6 Standardisation : Profils ligne et colonne									
NomLigne	Ligne Numéro	Coord. Dim.1	Coord. Dim.2	Masse	Qualité	Inertie Relative	Inertie Dim.1	Cosinus² Dim.1	Inertie Dim.2	Cosinus² Dim.2
Agriculteur	1	0,1661	-0,3096	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

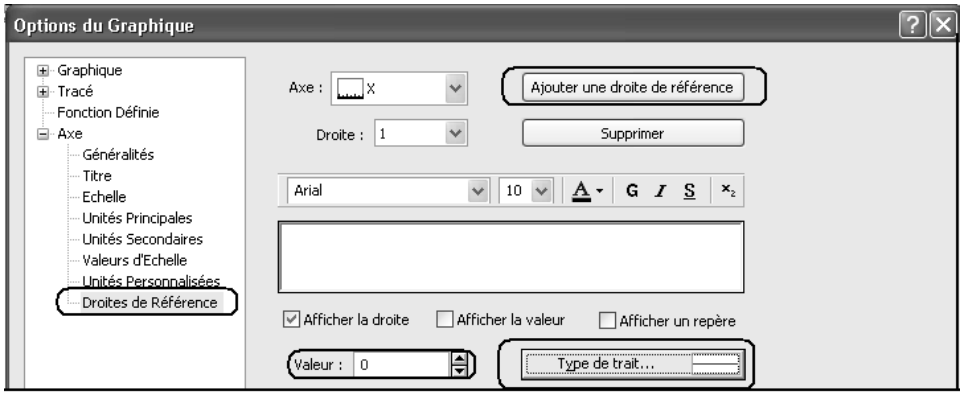
On utilise ensuite les boutons du bloc "Tracé des coordonnées" pour obtenir des représentations graphiques des résultats de l'AFC.



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. On peut, à l'aide de la souris, déplacer certaines étiquettes. 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.



On pourra également rendre le graphique plus clair en ajoutant des "droites de référence" correspondant à la position des axes. Pour cela :
- Faites un double-clic sur l'axe horizontal (par exemple) :
- Dans la fenêtre "Options du graphique" qui s'affiche alors, sélectionnez l'option Axe > Droites de référence.
- Cliquez sur le bouton "Ajouter une droite de référence". Indiquez comme valeur : 0 et un type de trait clairement visible :



2.3.2.4 Individus ligne et individus colonne supplémentaires

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

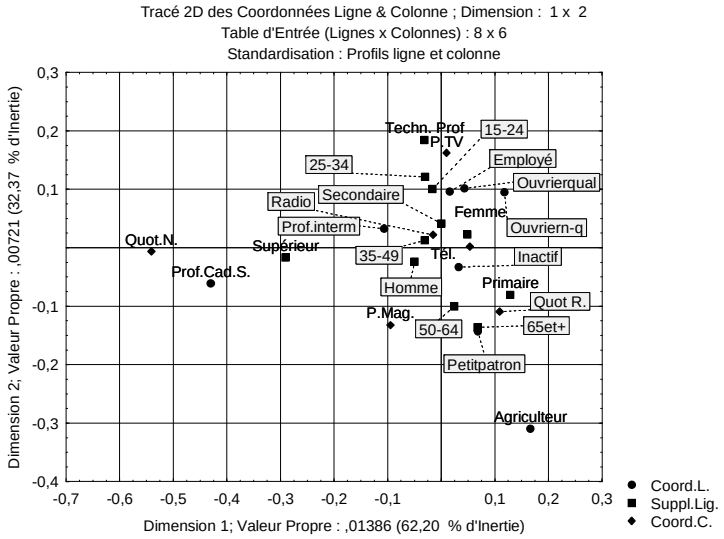
- d'une part, il peut être utile de positionner sur le graphique les groupes définis par une autre variable, telle que le sexe ou l'âge ou le niveau d'étude ;
- d'autre part, on peut remarquer que les modalités "Quotidiens nationaux" et "Prof. Cad. S." jouent un rôle prépondérant dans la formation du 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.

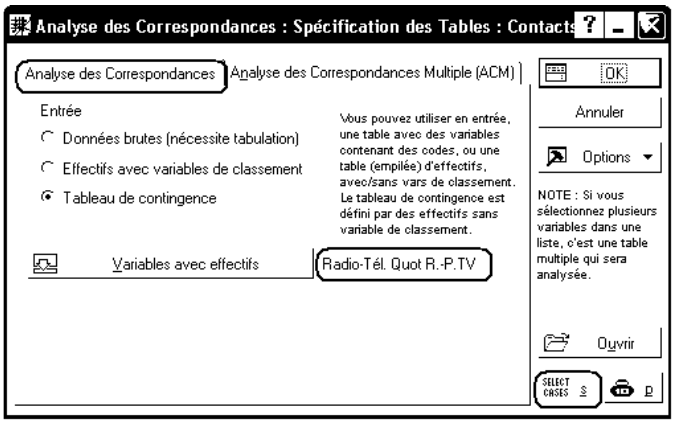
Refaites l'analyse, en réalisant notamment un graphique 2D avec l'ensemble des points lignes.

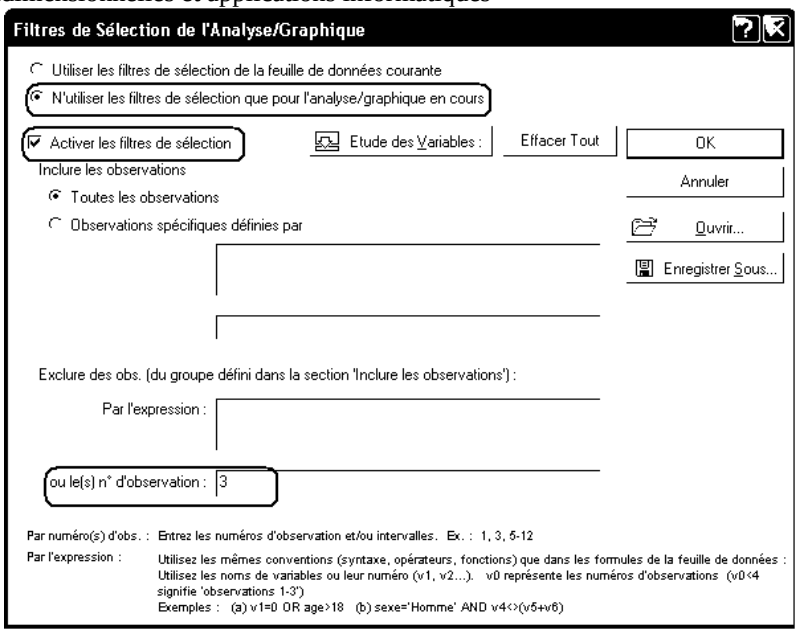


Rendre inactives certaines modalités des variables

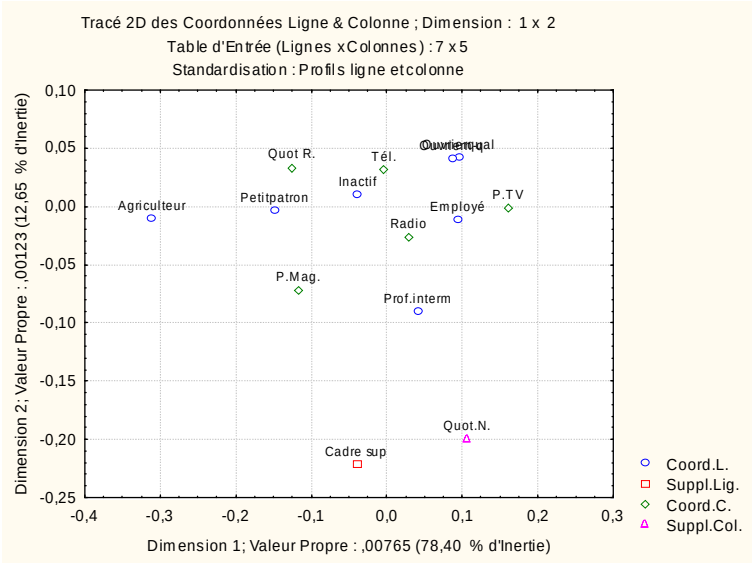
On veut réaliser l'étude en plaçant en éléments supplémentaires l'individu-ligne "Prof.Cad.S." et l'individu colonne "Quot.N.".

Pour rendre inactif un individu colonne, il suffit d'exclure la variable correspondante dans le premier dialogue affiché par l'ACP. Pour rendre inactif un individu ligne, on peut utiliser le bouton "Select cases" et définir un filtre excluant l'observation correspondante :





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 :



2.4 Exercices et prolongements

2.4.1 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 classeur Nobel.stw du serveur des salles de TD, on trouve les données saisies sous trois formes (protocole, effectifs, tableau de contingence). Ces données se trouvent aussi dans 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 ?

2.4.2 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 ?

2.4.3 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 le classeur Adjectifs-couleurs.stw et observez les données saisies.

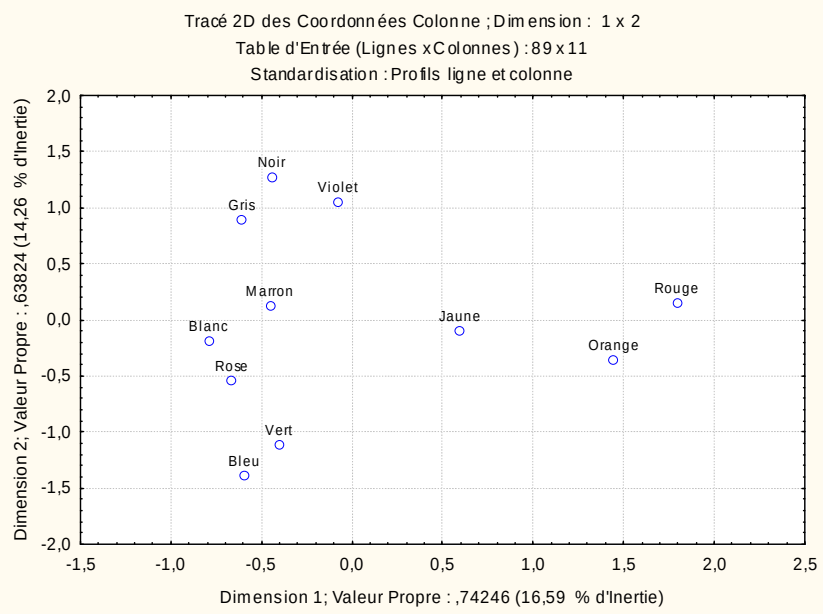
2.4.3.1 Choix des valeurs propres

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

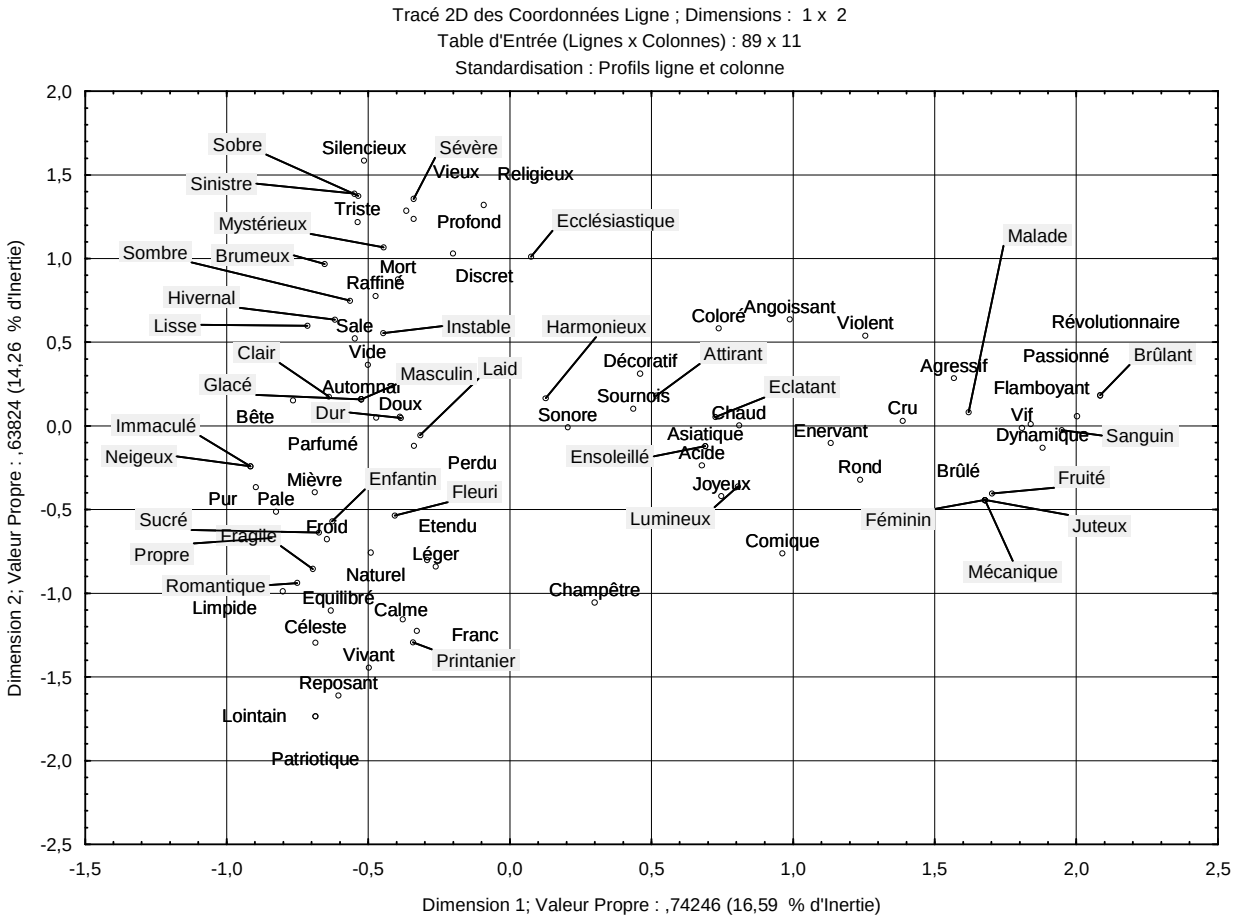
On voit ici que la décroissance des valeurs propres est très lente. Selon les règles énoncées précédemment, il faudrait conserver au moins 5 axes. Mais nous pouvons convenir de ne rechercher que les propriétés les plus caractéristiques des associations adjectifs - couleurs en n'étudiant que les deux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

2.4.3.2 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 :



2.4.3.3 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.

2.4.4 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 ?

2.4.5 Exercice : représentations sociales de trois moyens d'échapper à l'impôt

Dans un article publié en 2003[1] , des chercheurs autrichiens se sont intéressés aux représentations sociales de trois moyens d'échapper à l'impôt : l'exploitation de niches fiscales, la fraude fiscale et l'évasion fiscale. L'utilisation de niches fiscales se réfère à la réduction du montant de l'impôt par l'utilisation de moyens légaux, par exemple des déductions ou des incitations fiscales alors que la fraude fiscale renvoie à des moyens illégaux tels que la sous-estimation des revenus ou la surestimation des charges. L'évasion fiscale réfère au recours à une délocalisation dans le seul but d'échapper à l'impôt.

D'un point de vue macro-économique, ces trois moyens ont les mêmes effets sur le budget de l'état. Certains économistes suggèrent donc d'analyser leurs effets conjointement. Toutefois, d'un point de vue psychologique, on suppose que les contribuables les perçoivent différemment.

Les chercheurs ont mené une enquête auprès de 242 sujets. Les sujets étaient issus de 4 groupes socio-économiques différents : des agents du fisc, des étudiants en gestion, des avocats d'affaires et des propriétaires de PME. On présentait à chaque participant un scénario relatif à l'un des trois moyens d'échapper à l'impôt. Douze groupes de sujets ont ainsi été constitués en croisant le statut du sujet et la nature du scénario. On demandait ensuite aux participants de produire des associations spontanées à partir

[1] Kirchler E., Maciejovsky B., Schneider F. : Everyday representations of tax avoidance, tax evasion, and tax flight : Do legal differences matter ?, Journal of Economic Psychology, No 24, pp 535-553.

du scénario. 507 associations différentes ont ainsi été produites, et ont pu ensuite être regroupées en catégories de synonymes. 34 catégories sémantiques ont ainsi été définies.

Les feuilles de données du classeur Tax-avoidance.stw du serveur de fichiers de TD donnent le nombre d'occurrences de chaque catégorie sémantique pour chacun des 12 groupes de sujets. Les deux feuilles fournies diffèrent par la manière dont les catégories sémantiques ont été saisies : dans une variable nominale pour l'une, comme noms d'observations pour l'autre. Notez également que les intitulés des catégories n'ont pas été traduits de l'anglais. L'équivalent en français est donné ci-dessous.

Codage des groupes :

NF-F	Utilisation de niches fiscales - Agents du fisc
FF-F	Fraude fiscale - Agents du fisc
EF-F	Evasion fiscale - Agents du fisc
NF-E	Utilisation de niches fiscales - Etudiants
FF-E	Fraude fiscale - Etudiants
EF-E	Evasion fiscale - Etudiants
NF-A	Utilisation de niches fiscales - Avocats
FF-A	Fraude fiscale - Avocats
EF-A	Evasion fiscale - Avocats
NF-P	Utilisation de niches fiscales - Patrons de PME
FF-P	Fraude fiscale - Patrons de PME
EF-P	Evasion fiscale - Patrons de PME

Codage des catégories sémantiques :

1	Fraude fiscale intentionnelle
2	Fraude fiscale involontaire
3	Opacité du système fiscal
4	Paradis fiscaux
5	Réduction d'impôts légale
6	Avantages économiques de l'évasion fiscale
7	Evasion à l'étranger
8	Injustice
9	Justice verticale
10	Justice horizontale
11	Intelligence
12	Contrôle et sanction
13	Justification individuelle
14	Avantage personnel
15	Illégal
16	Peccadille
17	Conséquences négatives de l'évasion fiscale
18	Conséquences économiques
19	Types d'impôt
20	Harmonisation du système fiscal
21	Bureaucratie
22	Critique du système fiscal
23	Non profitable
24	Refus des réductions d'impôts
25	Opportunité

26	Faire usage des autorisations fiscales
27	Désir de réduire la charge fiscale
28	Goût du risque
29	Echappatoire fiscale
30	Acceptation des réductions d'impôts
31	Argent sale
32	Réactance
33	Code fiscal
34	Service du fisc

1) Traitez ce tableau par une analyse factorielle des correspondances et répondez aux questions qui suivent.
2) Comment peut-on qualifier la décroissance des valeurs propres ? Selon les critères généralement utilisés, combien de valeurs propres semblerait-il pertinent de retenir ?
N.B. Indépendamment du résultat de cette question, le reste de l'étude est mené sur les deux premières valeurs propres.
3) Examiner les qualités de représentation des modalités colonnes dans le plan défini par les deux axes retenus dans les tableaux de résultats. Quel commentaire peut-on porter globalement sur ces qualités de représentation ?
4) Etude des individus colonnes
a) Pour le premier axe factoriel, quels sont les individus colonnes dont la contribution est supérieure à la moyenne ? Pour chacun d'eux, préciser le signe de la coordonnée correspondante.
b) Même question pour le deuxième axe.
c) Les 12 individus colonnes ont été obtenus en croisant le statut de la personne interrogée et le type de scénario qui lui a été proposé. De ces deux variables indépendantes, laquelle semble avoir eu l'effet le plus important du point de vue de l'inertie du nuage ?
5) Etude des individus lignes
a) Citer quelques individus lignes dont la contribution à la formation du premier axe est particulièrement importante. Pour chacun d'eux, préciser le signe de la coordonnée correspondante.
b) Même question pour le deuxième axe.
6) Pour chacun des deux axes étudiés, proposer un couple de termes antinomiques permettant de qualifier l'opposition entre les deux extrémités de l'axe.
7) Quelle autre méthode multidimensionnelle aurait-on pu utiliser si on avait souhaité mieux séparer le statut de la personne interrogée et le type de scénario qui lui a été proposé ?

2.4.6 Exercice : la représentation sociale des maghrébins

Ci-dessous figurent des extraits de l'article "LA DÉSIGNATION DES MAGHRÉBINS: EFFET DU CHOIX LEXICAL SUR LES ÉLÉMENTS ACTIVÉS DANS LA REPRÉSENTATION SOCIALE" publié en 2003 par Edith Salès-Wuillemin, Philippe Castel et Marie-Françoise Lacassagne (Laboratoire de Psychologie Clinique et Sociale - Université de Bourgogne (Dijon, France)).

Cet article concerne la représentation sociale des Maghrébins, il montre l'incidence du choix du mot inducteur dans une épreuve d'associations verbales. Six cent cinquante participants (répartis en treize groupes indépendants) ont réalisé cette tâche à partir du mot inducteur MAGHREBIN ou de l'un de ses équivalents. Les sujets devaient produire 5 verbes, 5 adjectifs et 5 substantifs (l'ordre a été randomisé). L'analyse des résultats fait ressortir que des termes différents, même s'ils sont synonymes, mobilisent des zones et des dimensions différentes (contrastées) de la représentation sociale.

Six cent cinquante sujets, tous volontaires, étudiants de première année en sciences humaines, de langue maternelle française (hommes et femmes, ayant entre 17 et 21 ans, âge médian 19 ans) ont participé à cette étude. Dans une première phase, la tâche des sujets consistait à donner, sans limite de temps, les

mots qu'ils estimaient être synonymes du vocable MAGHREBIN. Après élimination des termes les plus péjoratifs (pour des raisons déontologiques) et ceux désignant des nationalités précises (Turcs), nous en avons retenu 12: ARABE, BEUR, BOUGNOUL, CLANDESTIN, ETRANGER, GRIS, IMMIGRE, INTEGRISTE, ISLAMISTE, MUSULMAN, NORD-AFRICAIN, REBEU. Durant la deuxième phase, chacun des 13 mots (MAGHREBIN et ses 12 synonymes) a servi de mot inducteur dans une tâche d'associations verbales contraintes. La tâche demandée aux sujets consistait à écrire les premiers mots qui leur venaient à l'esprit à partir du mot stimulus présenté, et pour qu'il n'y ait pas d'ambiguïté, le mot inducteur était inscrit à la suite, en gras. Afin de couvrir différentes dimensions de la représentation, chaque sujet devait donner les cinq premiers substantifs, les cinq premiers adjectifs, et les cinq premiers verbes qui lui venaient à l'esprit dès qu'il entendait le mot inducteur.

Nous avons d'abord éliminé les réponses correspondant à des désignations de Maghrébins, c'est-à-dire les mots fournis dans la phase de préparation (immigré, islamiste, ...). Puis, pour accéder à la dimension sociale, nous avons opéré une sélection sur les mots restants. Concrètement, nous avons retenu ceux cités par plus de 10 % de l'effectif dans au moins une des conditions (à savoir six sujets sur 50) ce qui nous a permis de retenir 75 mots.

Nous avons pris en compte le nombre de sujets ayant produit chacun des mots induits. Ceci a été réalisé pour chacun des mots inducteurs présentés. Les données ainsi produites sont rassemblées dans la feuille de données du classeur Statistica Designation-Maghrebins.stw

1) Traitez ces données par une analyse factorielle des correspondances et répondez aux questions qui suivent.

2) Etude du tableau des taux de liaison
a) A l'aide d'Excel, former le tableau des taux de liaisons.
b) Quelle est, dans chacune des colonnes, la valeur minimale des taux de liaison ? Quelle est la signification de cette valeur ?
c) Quel est le taux de liaison le plus élevé ? Quelle autre interprétation peut-on donner de cette valeur ?
d) Pour chacun des mots inducteurs, quel est le mot réponse qui lui est le plus fortement associé ?

3) Au vu du tableau des valeurs propres, combien d'axes factoriels faudrait-il étudier ? Justifier.
N.B. L'étude ci-dessous ne portera que sur les deux premiers axes.

4) On a calculé le carré de la distance euclidienne des points lignes à l'origine des axes, dans l'espace factoriel de dimension 12. Les valeurs extrêmes obtenues sont :

$$d^2(\text{s'intégrer}, O) = 0{,}16 \quad ; \quad d^2(\text{cachés}, O) = 13{,}01$$

Comment pouvait-on prévoir (qualitativement) ces résultats en examinant le tableau de contingence ?

5) Etude des qualités de représentation
a) Parmi les mots inducteurs, quel est celui qui est le mieux représenté dans le premier plan factoriel ? Quel est celui qui est le plus mal représenté ? Justifier.
b) De même, quelles sont les 7 modalités lignes les mieux représentées dans le premier plan factoriel ?

6) Etude de la première variable factorielle
a) On considère le nuage des mots inducteurs. 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 inducteurs ?
b) On considère le nuage des mots réponses. Quels sont les individus dont la contribution relative à la formation de l'axe est supérieure à 2% ? Pour chacun d'eux, préciser le signe de la coordonnée correspondante.

7) Etude de la deuxième variable factorielle : mener une étude analogue pour la deuxième variable factorielle.

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

9) Réaliser un graphique donnant la position de tous les individus lignes et colonnes dans le premier plan factoriel, sans indiquer les étiquettes des points. Indiquer sur ce graphique une douzaine d'étiquettes d'individus lignes et 5 ou 6 étiquettes d'individus colonnes choisies de manière à illustrer les résultats de l'étude précédente.

2.4.7 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 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.

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 des réponses aux questions 1 à 7.

2.5 Travail à rendre : comment passez-vous vos vacances ?

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, Professions Intermédiaires, 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. Second.	Parents	Amis	Camping	Séjour organisé	Autres	*Total*
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*
Prof. inter.	65	43	21	294	79	57	18	6	*583*
Autres actifs	77	60	189	839	53	124	28	53	*1423*
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) Utilisez le tableau des "Observés moins théoriques" et le tableau des "Effectifs théoriques" pour obtenir, sous Excel, le tableau des taux de liaison.

b) Indiquez une modalité ligne et une modalité colonne qui s'attirent. Indiquez de même une modalité ligne et une modalité colonne qui se repoussent.

c) Le taux de liaison entre "Agriculteurs" et "Résidence secondaire" est de -0,9868. 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 0,1227, alors que pour l'individu "Résidence secondaire", ces valeurs sont respectivement 0,0718 et 0,2318. 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 classeur Excel contenant le calcul des taux de liaison.
- Un fichier Word contenant votre interprétation des résultats avec notamment les réponses aux questions 1 à 8 ci-dessus.