3 Analyse Factorielle des Correspondances

3.1 Introduction

L'analyse factorielle des correspondances (AFC), ou analyse des correspondances simples, est une méthode exploratoire d'analyse des tableaux de contingence. Elle a été 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.

3.2 Exemple

3.2.1 Enoncé

Réf. Résultats publiés dans "Le Monde" au lendemain du 22 avril 2007.

Les données qui suivent sont constituées par les résultats du premier tour des élections présidentielles de 2007. Pour chacune des 23 régions françaises (22 régions métropolitaines + 1 "région" Outremer), on donne les effectifs de suffrages pour chacun des 12 candidats (en colonnes). L'objectif est d'analyser la structure des votes ainsi que les liaisons entre candidats et régions.

Données : résultats du premier tour des présidentielles 2007.

	Sarkozy	Bayrou	Royal	Le Pen	Besanc.	Villiers	Voynet	Laguiller	Bove	Buffet	Nihous	Schivardi
Alsace	362391	214259	171282	135730	33310	22492	20382	13821	13758	6100	5142	2522
Aquitaine	532127	417546	557300	168664	78230	34028	28285	22046	27941	41791	35300	7572
Auvergne	238152	169395	225477	78704	41522	18730	12090	12936	13532	21920	12474	4207
Bourgogne	297544	175213	241094	119041	42246	24971	13690	14440	12296	18154	12079	3608
Bretagne	557507	451988	564100	143926	94205	41212	39026	25662	28484	31860	21207	5169
Centre	460425	278175	345352	168912	65347	45720	22655	22279	17395	30003	20567	5696
Cham-Ard	246680	122642	160280	114527	33424	20455	10727	13560	7414	12465	9016	2270
Corse	56819	18979	33493	23432	5941	1908	2119	1346	1659	5163	2260	450
Fr-Comte	212358	114148	165243	94212	30672	16361	12879	10880	10365	9204	7814	2446
Ile de Fr	1931429	1143081	1593033	430553	181247	89498	89885	52965	57453	110967	19890	12386
Lang-Rous	470017	234739	395509	214468	62597	28166	20787	17175	27412	38590	21356	11436
Limousin	123870	82445	142237	38525	24040	10789	6144	6566	6629	15695	8029	2284
Lorraine	403919	250195	315596	196696	70940	29183	21562	25574	16094	19229	10378	4211
Midi-Pyr	458093	341651	542038	154777	69177	30850	25267	18623	33687	34076	25280	8778
Nord-PdC	639390	340679	573071	335855	127881	40702	31388	52695	24591	74027	43595	6348
Basse-Nor	280914	184256	209308	88569	43997	25722	14389	14652	13180	11211	21449	3090
Haute-Nor	309924	184615	257664	126795	58312	25692	15707	20048	12563	26376	13717	3604
Pays-Loire	636934	457560	552280	158844	93685	107895	38952	28481	25811	26737	26674	6481

Picardie	331053	161236	251862	168699	57769	26731	14219	23982	11619	22334	20439	4189
Poit.-Cha	304493	194126	322212	88138	45638	38735	16333	13949	13934	15901	21571	3882
PACA	1010234	419161	579036	377831	88331	54851	38339	26467	36963	61968	26538	9935
Rhone-Alp	1121615	689984	807220	360646	123776	75959	63032	38114	51441	57654	32624	10969
Outremer	337711	103933	398110	36714	22159	5139	12383	10234	14904	14062	2698	1772

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 lesquelles le vote est concentré dans certaines régions ?

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

3.2.2 Etude descriptive du tableau de contingence

On fixe les notations suivantes :

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

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

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

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

3.2.2.1 Tableau des fréquences

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

	Sarkozy	Bayrou	Royal	Le Pen	Besanc.	Villiers	Voynet	Laguiller	Bove	Buffet	Nihous	Schivardi	Total
Alsace	1,00	0,59	0,47	0,37	0,09	0,06	0,06	0,04	0,04	0,02	0,01	0,01	2,75
Aquitaine	1,46	1,15	1,53	0,46	0,21	0,09	0,08	0,06	0,08	0,11	0,10	0,02	5,36
Auvergne	0,65	0,47	0,62	0,22	0,11	0,05	0,03	0,04	0,04	0,06	0,03	0,01	2,33
Bourgogne	0,82	0,48	0,66	0,33	0,12	0,07	0,04	0,04	0,03	0,05	0,03	0,01	2,68
Bretagne	1,53	1,24	1,55	0,40	0,26	0,11	0,11	0,07	0,08	0,09	0,06	0,01	5,51
Centre	1,27	0,76	0,95	0,46	0,18	0,13	0,06	0,06	0,05	0,08	0,06	0,02	4,07
Champ Ard.	0,68	0,34	0,44	0,31	0,09	0,06	0,03	0,04	0,02	0,03	0,02	0,01	2,07
Corse	0,16	0,05	0,09	0,06	0,02	0,01	0,01	0,00	0,00	0,01	0,01	0,00	0,42
Fr-Comte	0,58	0,31	0,45	0,26	0,08	0,04	0,04	0,03	0,03	0,03	0,02	0,01	1,89
Ile-de-Fr.	5,31	3,14	4,38	1,18	0,50	0,25	0,25	0,15	0,16	0,30	0,05	0,03	15,70
Lang-Rous	1,29	0,64	1,09	0,59	0,17	0,08	0,06	0,05	0,08	0,11	0,06	0,03	4,24
Limousin	0,34	0,23	0,39	0,11	0,07	0,03	0,02	0,02	0,02	0,04	0,02	0,01	1,28
Lorraine	1,11	0,69	0,87	0,54	0,19	0,08	0,06	0,07	0,04	0,05	0,03	0,01	3,75
Midi-Pyr	1,26	0,94	1,49	0,43	0,19	0,08	0,07	0,05	0,09	0,09	0,07	0,02	4,79
Nord-PdeCa	1,76	0,94	1,57	0,92	0,35	0,11	0,09	0,14	0,07	0,20	0,12	0,02	6,29
Basse-Nor	0,77	0,51	0,58	0,24	0,12	0,07	0,04	0,04	0,04	0,03	0,06	0,01	2,50
Haute-Nor	0,85	0,51	0,71	0,35	0,16	0,07	0,04	0,06	0,03	0,07	0,04	0,01	2,90
Pays Loire	1,75	1,26	1,52	0,44	0,26	0,30	0,11	0,08	0,07	0,07	0,07	0,02	5,94
Picardie	0,91	0,44	0,69	0,46	0,16	0,07	0,04	0,07	0,03	0,06	0,06	0,01	3,01
Poitou-Char	0,84	0,53	0,89	0,24	0,13	0,11	0,04	0,04	0,04	0,04	0,06	0,01	2,96
PACA	2,78	1,15	1,59	1,04	0,24	0,15	0,11	0,07	0,10	0,17	0,07	0,03	7,50
Rhone-Alp.	3,08	1,90	2,22	0,99	0,34	0,21	0,17	0,10	0,14	0,16	0,09	0,03	9,43
Outremer	0,93	0,29	1,09	0,10	0,06	0,01	0,03	0,03	0,04	0,04	0,01	0,00	2,64
Total	31,11	18,55	25,83	10,51	4,11	2,24	1,57	1,34	1,32	1,94	1,15	0,34	100,00

3.2.2.2 Tableau des fréquences lignes

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

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

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

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

	Sarkozy	Bayrou	Royal	Le Pen	Besancenot	Villiers	Voynet	Laguiller	Bove	Buffet	Nihous	Schivardi	Total
Alsace	36,20	21,40	17,11	13,56	3,33	2,25	2,04	1,38	1,37	0,61	0,51	0,25	100,00
Aquitaine	27,28	21,40	28,57	8,65	4,01	1,74	1,45	1,13	1,43	2,14	1,81	0,39	100,00
Auvergne	28,05	19,95	26,55	9,27	4,89	2,21	1,42	1,52	1,59	2,58	1,47	0,50	100,00
Bourgogne	30,54	17,98	24,74	12,22	4,34	2,56	1,41	1,48	1,26	1,86	1,24	0,37	100,00
Bretagne	27,81	22,55	28,14	7,18	4,70	2,06	1,95	1,28	1,42	1,59	1,06	0,26	100,00
Centre	31,06	18,76	23,29	11,39	4,41	3,08	1,53	1,50	1,17	2,02	1,39	0,38	100,00
Champ-Ard.	32,74	16,28	21,27	15,20	4,44	2,71	1,42	1,80	0,98	1,65	1,20	0,30	100,00
Corse	37,00	12,36	21,81	15,26	3,87	1,24	1,38	0,88	1,08	3,36	1,47	0,29	100,00
Fr-Comte	30,93	16,63	24,07	13,72	4,47	2,38	1,88	1,58	1,51	1,34	1,14	0,36	100,00
Ile-de-France	33,81	20,01	27,89	7,54	3,17	1,57	1,57	0,93	1,01	1,94	0,35	0,22	100,00
Lang-Rous.	30,48	15,22	25,64	13,91	4,06	1,83	1,35	1,11	1,78	2,50	1,38	0,74	100,00
Limousin	26,51	17,64	30,44	8,24	5,14	2,31	1,31	1,41	1,42	3,36	1,72	0,49	100,00
Lorraine	29,62	18,35	23,14	14,43	5,20	2,14	1,58	1,88	1,18	1,41	0,76	0,31	100,00
Midi-Pyr	26,29	19,61	31,11	8,88	3,97	1,77	1,45	1,07	1,93	1,96	1,45	0,50	100,00
Nord-PdC	27,92	14,88	25,02	14,66	5,58	1,78	1,37	2,30	1,07	3,23	1,90	0,28	100,00
Basse-Norm	30,84	20,23	22,98	9,72	4,83	2,82	1,58	1,61	1,45	1,23	2,36	0,34	100,00
Haute-Norm	29,38	17,50	24,42	12,02	5,53	2,44	1,49	1,90	1,19	2,50	1,30	0,34	100,00
Pays Loire	29,48	21,18	25,56	7,35	4,34	4,99	1,80	1,32	1,19	1,24	1,23	0,30	100,00
Picardie	30,26	14,74	23,02	15,42	5,28	2,44	1,30	2,19	1,06	2,04	1,87	0,38	100,00
Poitou-Char	28,22	17,99	29,86	8,17	4,23	3,59	1,51	1,29	1,29	1,47	2,00	0,36	100,00
PACA	37,01	15,36	21,21	13,84	3,24	2,01	1,40	0,97	1,35	2,27	0,97	0,36	100,00
Rhone-Alpes	32,67	20,10	23,51	10,51	3,61	2,21	1,84	1,11	1,50	1,68	0,95	0,32	100,00
Outremer	35,18	10,83	41,48	3,83	2,31	0,54	1,29	1,07	1,55	1,47	0,28	0,18	100,00

3.2.2.3 Tableau des fréquences colonnes

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

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

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

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

	Sarkozy	Bayrou	Royal	Le Pen	Besanc.	Villiers	Voynet	Laguiller	Bove	Buffet	Nihous	Schivardi
Alsace	3,20	3,17	1,82	3,55	2,23	2,76	3,57	2,84	2,87	0,86	1,22	2,05
Aquitaine	4,70	6,19	5,93	4,41	5,23	4,17	4,96	4,53	5,83	5,92	8,40	6,14
Auvergne	2,10	2,51	2,40	2,06	2,78	2,30	2,12	2,66	2,82	3,11	2,97	3,41
Bourgogne	2,63	2,60	2,56	3,11	2,83	3,06	2,40	2,97	2,57	2,57	2,88	2,93
Bretagne	4,92	6,70	6,00	3,76	6,30	5,05	6,84	5,27	5,95	4,52	5,05	4,19
Centre	4,07	4,12	3,67	4,42	4,37	5,60	3,97	4,58	3,63	4,25	4,90	4,62
Champ-Ard	2,18	1,82	1,70	2,99	2,24	2,51	1,88	2,79	1,55	1,77	2,15	1,84
Corse	0,50	0,28	0,36	0,61	0,40	0,23	0,37	0,28	0,35	0,73	0,54	0,36
Fr-Comte	1,88	1,69	1,76	2,46	2,05	2,01	2,26	2,24	2,16	1,30	1,86	1,98
Ile de France	17,06	16,93	16,94	11,26	12,13	10,97	15,76	10,89	11,99	15,73	4,73	10,05
Lang. Rous.	4,15	3,48	4,21	5,61	4,19	3,45	3,65	3,53	5,72	5,47	5,08	9,27
Limousin	1,09	1,22	1,51	1,01	1,61	1,32	1,08	1,35	1,38	2,22	1,91	1,85
Lorraine	3,57	3,71	3,36	5,14	4,75	3,58	3,78	5,26	3,36	2,73	2,47	3,42
Midi-Pyr	4,05	5,06	5,76	4,05	4,63	3,78	4,43	3,83	7,03	4,83	6,02	7,12
Nord-PdC	5,65	5,05	6,09	8,78	8,56	4,99	5,50	10,83	5,13	10,49	10,38	5,15
Basse-Norm	2,48	2,73	2,23	2,32	2,94	3,15	2,52	3,01	2,75	1,59	5,11	2,51
Haute-Norm	2,74	2,74	2,74	3,32	3,90	3,15	2,75	4,12	2,62	3,74	3,27	2,92
Pays Loire	5,62	6,78	5,87	4,15	6,27	13,23	6,83	5,85	5,39	3,79	6,35	5,26
Picardie	2,92	2,39	2,68	4,41	3,87	3,28	2,49	4,93	2,43	3,17	4,87	3,40
Poitou-Char	2,69	2,88	3,43	2,30	3,05	4,75	2,86	2,87	2,91	2,25	5,13	3,15
PACA	8,92	6,21	6,16	9,88	5,91	6,72	6,72	5,44	7,71	8,78	6,32	8,06
Rhone-Alpes	9,91	10,22	8,58	9,43	8,28	9,31	11,05	7,83	10,74	8,17	7,77	8,90

Outremer	2,98	1,54	4,23	0,96	1,48	0,63	2,17	2,10	3,11	1,99	0,64	1,44
Total	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00

3.2.2.4 Distances entre profils. Métrique du Φ^2

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

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

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

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

- Avec la métrique du Φ^2, la distance entre deux lignes ne dépend pas des poids respectifs des colonnes. Ainsi, sur notre exemple, les 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 (Sarkozy, Royal, Bayrou).
- 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 :

$$d^2_{\Phi^2}(Alsace, Aquitaine) = \frac{(0{,}3619 - 0{,}2728)^2}{0{,}3111} + \ldots + \frac{(0{,}0025 - 0{,}0038)^2}{0{,}0033} = 0{,}1315$$

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

$$d^2_{\Phi^2}(Alsace, Moyenne) = \frac{(0{,}3619 - 0{,}3111)^2}{0{,}3111} + \ldots + \frac{(0{,}0025 - 0{,}0033)^2}{0{,}0033} = 0{,}0668$$

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 Sarkozy et la colonne Bayrou est :

$$d^2_{\Phi^2}(Sarkozy, Bayrou) = \frac{(0{,}0320 - 0{,}0317)^2}{0{,}0275} + \ldots + \frac{(0{,}0298 - 0{,}0154)^2}{0{,}0264} = 0{,}0379$$

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

3.2.2.5 Taux de liaison et Phi-2

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

	Sarkozy	Bayrou	Royal	Le Pen	Besanc.	Villiers	Voynet	Laguiller	Bove	Buffet	Nihous	Schivardi
Alsace	0,16	0,15	-0,34	0,29	-0,19	0,00	0,30	0,03	0,04	-0,69	-0,56	-0,26
Aquitaine	-0,12	0,15	0,11	-0,18	-0,02	-0,22	-0,07	-0,15	0,09	0,11	0,57	0,15
Auvergne	-0,10	0,08	0,03	-0,12	0,19	-0,02	-0,09	0,14	0,21	0,33	0,27	0,46
Bourgogne	-0,02	-0,03	-0,04	0,16	0,06	0,14	-0,10	0,11	-0,04	-0,04	0,07	0,09
Bretagne	-0,11	0,22	0,09	-0,32	0,14	-0,08	0,24	-0,04	0,08	-0,18	-0,08	-0,24
Centre	0,00	0,01	-0,10	0,08	0,07	0,38	-0,02	0,12	-0,11	0,04	0,20	0,13
Champ. Ard	0,05	-0,12	-0,18	0,45	0,08	0,21	-0,09	0,35	-0,25	-0,15	0,04	-0,11
Corse	0,19	-0,33	-0,16	0,45	-0,06	-0,45	-0,12	-0,34	-0,18	0,73	0,27	-0,14
Fr-Comte	-0,01	-0,10	-0,07	0,31	0,09	0,06	0,20	0,19	0,15	-0,31	-0,01	0,05

Ile-de-France	0,09	0,08	0,08	-0,28	-0,23	-0,30	0,00	-0,31	-0,24	0,00	-0,70	-0,36
Lang-Rous	-0,02	-0,18	-0,01	0,32	-0,01	-0,19	-0,14	-0,17	0,35	0,29	0,20	1,19
Limousin	-0,15	-0,05	0,18	-0,22	0,25	0,03	-0,16	0,05	0,08	0,73	0,49	0,44
Lorraine	-0,05	-0,01	-0,10	0,37	0,27	-0,05	0,01	0,40	-0,10	-0,27	-0,34	-0,09
Midi-Pyr	-0,15	0,06	0,20	-0,15	-0,03	-0,21	-0,07	-0,20	0,47	0,01	0,26	0,49
Nord-PdC	-0,10	-0,20	-0,03	0,40	0,36	-0,21	-0,13	0,72	-0,18	0,67	0,65	-0,18
Basse-Norma	-0,01	0,09	-0,11	-0,07	0,18	0,26	0,01	0,20	0,10	-0,36	1,04	0,00
Haute-Norma	-0,06	-0,06	-0,05	0,14	0,35	0,09	-0,05	0,42	-0,10	0,29	0,13	0,01
Pays Loire	-0,05	0,14	-0,01	-0,30	0,06	1,23	0,15	-0,01	-0,09	-0,36	0,07	-0,11
Picardie	-0,03	-0,21	-0,11	0,47	0,29	0,09	-0,17	0,64	-0,19	0,05	0,62	0,13
Poitou-Char	-0,09	-0,03	0,16	-0,22	0,03	0,60	-0,03	-0,03	-0,02	-0,24	0,73	0,06
PACA	0,19	-0,17	-0,18	0,32	-0,21	-0,10	-0,10	-0,27	0,03	0,17	-0,16	0,07
Rhone-Alpes	0,05	0,08	-0,09	0,00	-0,12	-0,01	0,17	-0,17	0,14	-0,13	-0,18	-0,06
Outremer	0,13	-0,42	0,61	-0,64	-0,44	-0,76	-0,18	-0,20	0,18	-0,24	-0,76	-0,46

Signification pratique du taux de liaison : le score de Sarkozy en Alsace est 16% plus élevé que le score théorique que l'on observerait si les votes étaient indépendants des régions. Au contraire, celui de Royal est 34% moins élevé que le score théorique.
Par construction, les valeurs prises par le taux de liaison sont :

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

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

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

Ici, on obtient : $\Phi^2 = 0{,}03341$.
La méthode d'analyse factorielle des correspondances peut être vue comme une décomposition pertinente du Φ^2 selon plusieurs axes factoriels.

3.2.3 L'analyse factorielle des correspondances proprement dite

L'application de la méthode a deux effets :

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

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

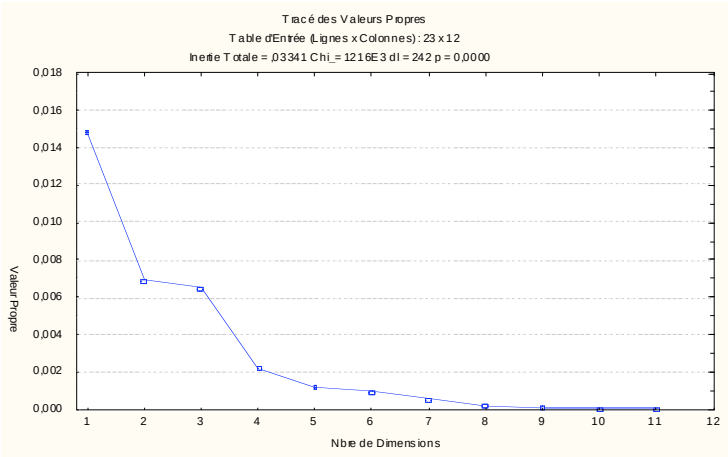
3.2.3.1 Valeurs propres

Le nombre de valeurs propres produites par la recherche des facteurs principaux est égal au minimum du nombre de lignes et du nombre de colonnes du tableau de contingence. Cependant, la première valeur

propre est systématiquement égale à 1, et n'est pas mentionnée dans les résultats. Les autres valeurs propres sont des nombres positifs inférieurs à 1 et leur somme est égale à Φ^2.

Valeurs Propres et Inertie de toutes les Dimensions (Par-Region dans Presidentielles-2007-v2.stw)
Inertie Totale = ,03341 Chi² = 1216E3 dl = 242 p = 0,0000

	ValSing.	ValProp.	%age Inertie	%age Cumulé	Chi²
1	0,12148	0,01476	44,17	44,17	537082
2	0,08322	0,00693	20,73	64,90	252072
3	0,08075	0,00652	19,52	84,42	237324
4	0,04709	0,00222	6,64	91,05	80692
5	0,03449	0,00119	3,56	94,61	43300
6	0,03083	0,00095	2,85	97,46	34602
7	0,02274	0,00052	1,55	99,01	18827
8	0,01333	0,00018	0,53	99,54	6465
9	0,00939	0,00009	0,26	99,80	3209
10	0,00667	0,00004	0,13	99,94	1619
11	0,00463	0,00002	0,06	100,00	781



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

3.2.3.2 Résultats relatifs aux individus-lignes

Coordonnées Ligne et Contributions à l'Inertie (Par-Region dans Presidentielles-2007-v2.stw)
Table d'Entrée (Lignes x Colonnes) : 23 x 12
Standardisation : Profils ligne et colonne

NomLigne	igne iméi	Coord. Dim.1	Coord. Dim.2	Coord. Dim.3	Masse	Qualité	Inertie Relative	Inertie Dim.1	Cosinus² Dim.1	Inertie Dim.2	Cosinus² Dim.2	Inertie Dim.3	Cosinus² Dim.3
Alsace	1	-0,1219	0,1067	-0,1876	0,0275	0,9192	0,0550	0,0277	0,2222	0,0452	0,1704	0,1485	0,5266
Aquitaine	2	0,0821	-0,0806	0,0276	0,0536	0,6685	0,0336	0,0245	0,3220	0,0503	0,3101	0,0063	0,0364
Auvergne	3	0,0251	-0,0807	0,0272	0,0233	0,6607	0,0083	0,0010	0,0526	0,0220	0,5461	0,0026	0,0620
Bourgogn	4	-0,0607	-0,0139	-0,0006	0,0268	0,8474	0,0037	0,0067	0,8052	0,0007	0,0421	0,0000	0,0001
Bretagne	5	0,1259	-0,0753	-0,0354	0,0551	0,8223	0,0456	0,0591	0,5724	0,0450	0,2046	0,0106	0,0453
Centre	6	-0,0593	-0,0350	-0,0397	0,0407	0,8317	0,0093	0,0097	0,4634	0,0072	0,1608	0,0099	0,2076
Champag	7	-0,1799	0,0291	-0,0207	0,0207	0,9107	0,0229	0,0454	0,8761	0,0025	0,0229	0,0014	0,0116
Corse	8	-0,1891	0,1485	0,0941	0,0042	0,8574	0,0098	0,0102	0,4598	0,0134	0,2838	0,0057	0,1139
Franche-(	9	-0,1048	0,0093	0,0012	0,0189	0,6643	0,0094	0,0140	0,6590	0,0002	0,0052	0,0000	0,0001
Ile-de-Fra	10	0,1263	0,0854	-0,0231	0,1570	0,9244	0,1209	0,1697	0,6201	0,1654	0,2836	0,0128	0,0207
Languedc	11	-0,1029	0,0235	0,0902	0,0424	0,7125	0,0343	0,0304	0,3914	0,0034	0,0205	0,0529	0,3006
Limousin	12	0,0673	-0,1103	0,1199	0,0128	0,8325	0,0143	0,0039	0,1214	0,0226	0,3259	0,0283	0,3853
Lorraine	13	-0,1264	-0,0054	-0,0137	0,0375	0,6174	0,0294	0,0406	0,6090	0,0002	0,0011	0,0011	0,0072
Midi-Pyre	14	0,1022	-0,0708	0,0738	0,0479	0,7606	0,0394	0,0339	0,3799	0,0347	0,1827	0,0399	0,1980
Nord-Pas	15	-0,1688	-0,0608	0,1367	0,0629	0,9286	0,1032	0,1215	0,5201	0,0336	0,0675	0,1803	0,3410
Basse-Nc	16	-0,0283	-0,0936	-0,0604	0,0250	0,5544	0,0179	0,0014	0,0336	0,0317	0,3676	0,0140	0,1531
Haute-No	17	-0,0824	-0,0522	0,0329	0,0290	0,7599	0,0121	0,0133	0,4867	0,0114	0,1957	0,0048	0,0775
Pays-de-l	18	0,0760	-0,1216	-0,1343	0,0594	0,7547	0,0909	0,0233	0,1131	0,1267	0,2890	0,1642	0,3526
Picardie	19	-0,1995	-0,0360	0,0648	0,0301	0,9326	0,0437	0,0811	0,8195	0,0056	0,0267	0,0193	0,0864
Poitou-Ch	20	0,0786	-0,1151	0,0190	0,0296	0,6629	0,0265	0,0124	0,2068	0,0567	0,4440	0,0016	0,0121
Provence	21	-0,1308	0,1378	-0,0105	0,0750	0,9131	0,0890	0,0870	0,4315	0,2056	0,4788	0,0013	0,0028
Rhone-Al	22	-0,0028	0,0338	-0,0659	0,0943	0,8336	0,0186	0,0001	0,0012	0,0156	0,1731	0,0629	0,6593
Outremer	23	0,3201	0,1623	0,2393	0,0264	0,9058	0,1622	0,1831	0,4989	0,1003	0,1282	0,2316	0,2787

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 le département.

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

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

Par exemple :

$$QLT(Alsace, F_1; F_2; F_3) = \frac{(-0{,}1219)^2 + (0{,}1067)^2 + (-0{,}1876)^2}{(-0{,}1219)^2 + (0{,}1067)^2 + (-0{,}1876)^2 + (-0{,}0339)^2 + \ldots + (0{,}0047)^2}$$

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 :

$$(-0{,}1219 \times 0{,}0275) + (0{,}0821 \times 0{,}0536) + \ldots + (0{,}3201 \times 0{,}0264) = 0$$

$$(-0{,}1219)^2 \times 0{,}0275 + (0{,}0821)^2 \times 0{,}0536 + \ldots + (0{,}3201)^2 \times 0{,}0264 = 0{,}01476$$

$$(-0{,}1219) \times (0{,}1067) \times 0{,}0275 + (0{,}0821) \times (-0{,}0806) \times 0{,}0536 + \ldots + (0{,}3201) \times (0{,}1623) \times 0{,}0264 = 0$$

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 :

$$Ctr(Alsace, F_1) = \frac{0{,}0275 \times (-0{,}1219)^2}{0{,}01476} = 0{,}0277$$

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 :

$$QLT(Alsace, F_1) = \frac{(-0{,}1219)^2}{(-0{,}1219)^2 + (0{,}1067)^2 + (-0{,}1876)^2 + \ldots + (0{,}0047)^2} = 0{,}2222$$

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 11 dimensions et sa projection sur le premier axe factoriel.

3.2.3.3 Résultats relatifs aux individus-colonnes

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

Coordonnées Colonne et Contributions à l'Inertie (Par-Region dans Presidentielles-2007-v2.stw)
Table d'Entrée (Lignes x Colonnes) : 23 x 12
Standardisation : Profils ligne et colonne

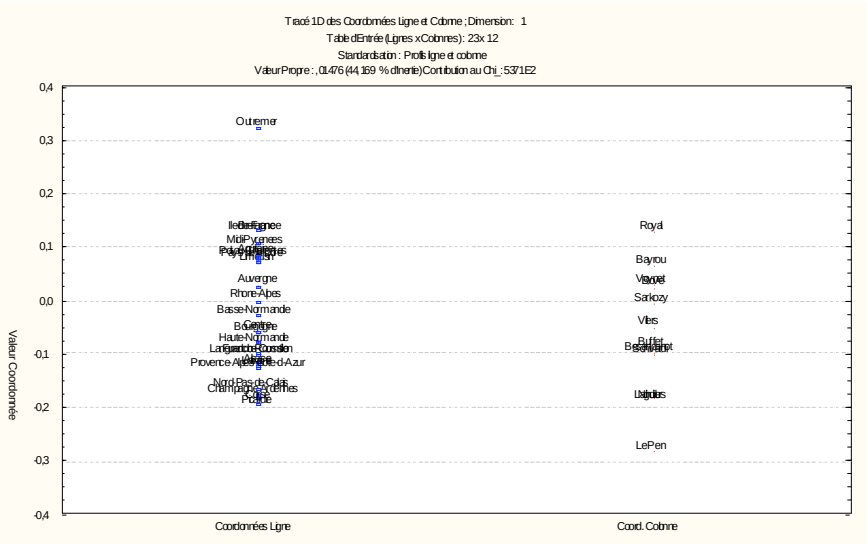
Nom Col.	Numéro	Coord. Dim.1	Coord. Dim.2	Coord. Dim.3	Masse	Qualité	Inertie Relative	Inertie Dim.1	Cosinus² Dim.1	Inertie Dim.2	Cosinus² Dim.2	Inertie Dim.3	Cosinus² Dim.3
Sarkozy	1	-0,0071	0,0911	-0,0274	0,3111	0,9194	0,0922	0,0011	0,0050	0,3731	0,8383	0,0359	0,0760
Bayrou	2	0,0610	-0,0500	-0,1032	0,1855	0,8085	0,1159	0,0468	0,1785	0,0670	0,1200	0,3028	0,5100
Royal	3	0,1244	-0,0146	0,0864	0,2583	0,9791	0,1829	0,2710	0,6547	0,0080	0,0091	0,2955	0,3154
Le Pen	4	-0,2839	0,0274	0,0297	0,1051	0,9824	0,2633	0,5739	0,9629	0,0114	0,0090	0,0142	0,0105
Besancer	5	-0,1026	-0,1497	0,0368	0,0411	0,8532	0,0494	0,0293	0,2619	0,1328	0,5576	0,0085	0,0337
Villiers	6	-0,0527	-0,2195	-0,2150	0,0224	0,6467	0,1008	0,0042	0,0185	0,1559	0,3205	0,1589	0,3077
Voynet	7	0,0269	-0,0095	-0,0991	0,0157	0,6394	0,0078	0,0008	0,0434	0,0002	0,0054	0,0236	0,5906
Laguiller	8	-0,1878	-0,1426	0,0812	0,0134	0,6823	0,0365	0,0319	0,3869	0,0392	0,2231	0,0135	0,0723
Bove	9	0,0222	-0,0298	0,0358	0,0132	0,0736	0,0143	0,0004	0,0136	0,0017	0,0245	0,0026	0,0355
Buffet	10	-0,0886	-0,0149	0,1921	0,0194	0,5316	0,0491	0,0103	0,0928	0,0006	0,0026	0,1097	0,4362
Nihous	11	-0,1880	-0,3514	0,1310	0,0115	0,7982	0,0762	0,0276	0,1602	0,2058	0,5602	0,0304	0,0778
Schivardi	12	-0,1060	-0,0930	0,0921	0,0034	0,2435	0,0118	0,0026	0,0964	0,0042	0,0743	0,0044	0,0728

3.2.3.4 Résultats graphiques

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

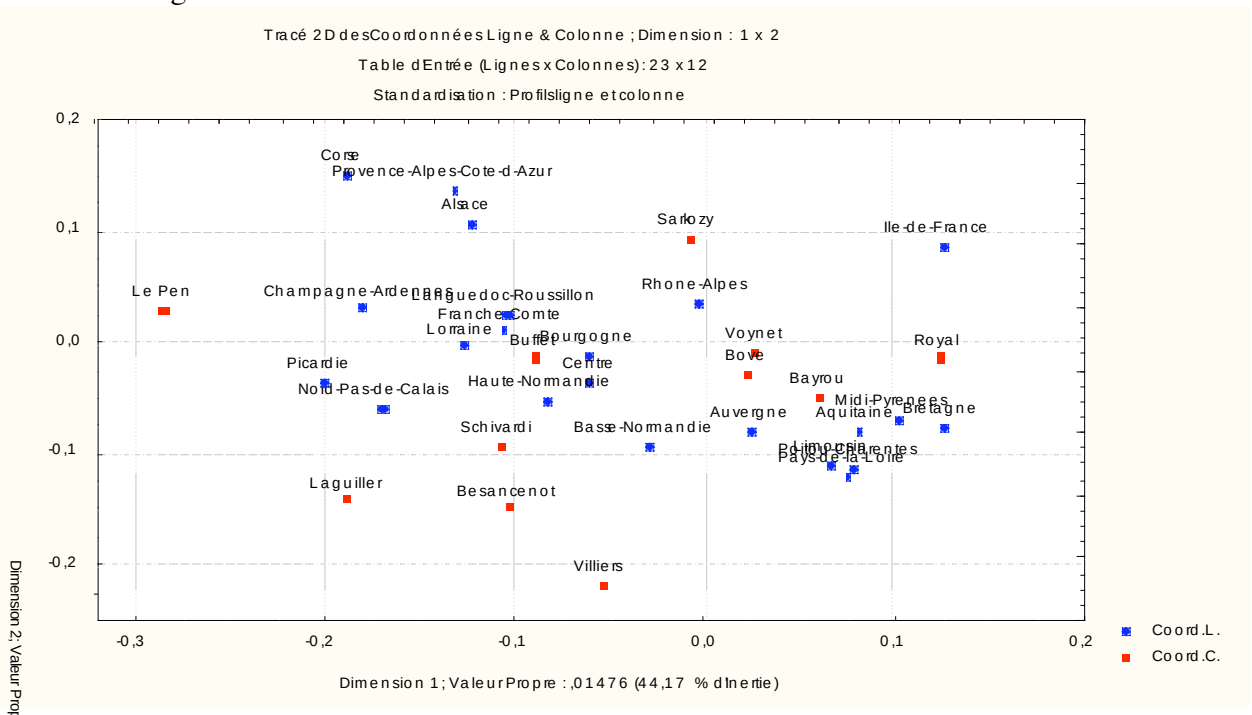
On peut réaliser et essayer d'interpréter des graphiques :

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

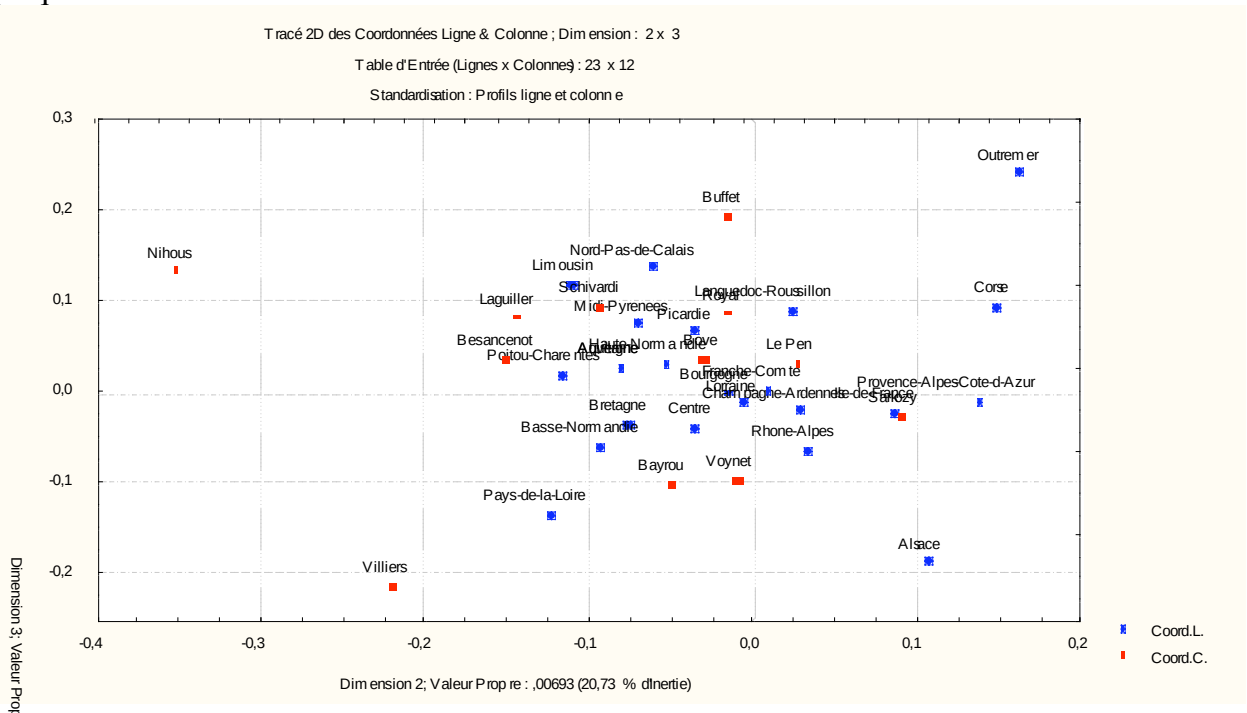


Graphique selon les axes 1 et 2

N.B. L'individu ligne "Outremer" et l'individu colonne "Nihous" sont en dehors du dessin.



Graphique selon les axes 2 et 3



3.2.3.5 Interprétation géométrique

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

Par exemple, nous avions vu que :

$$d^2_{\Phi^2}(Alsace, Aquitaine) = \frac{(0{,}3619 - 0{,}2728)^2}{0{,}3111} + \ldots + \frac{(0{,}0025 - 0{,}0038)^2}{0{,}0033} = 0{,}1315$$

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	Fact. 11
Alsace	-0,1219	0,1067	-0,1876	-0,0339	0,0053	-0,0585	-0,0228	0,0057	0,0037	0,0143	0,0047
Aquitaine	0,0821	-0,0806	0,0276	-0,0641	-0,0400	0,0079	-0,0240	-0,0243	0,0021	0,0015	-0,0003
Auvergne	0,0251	-0,0807	0,0272	-0,0399	-0,0114	0,0286	0,0130	0,0281	-0,0180	0,0145	-0,0034
Bourgogne	-0,0607	-0,0139	-0,0006	0,0131	0,0025	-0,0112	0,0105	-0,0149	-0,0052	0,0022	-0,0056
Bretagne	0,1259	-0,0753	-0,0354	-0,0588	0,0245	-0,0184	-0,0034	0,0180	0,0038	-0,0130	0,0012
Centre	-0,0593	-0,0350	-0,0397	0,0262	-0,0007	0,0224	0,0035	-0,0023	-0,0059	0,0050	0,0034
Champ.Ard	-0,1799	0,0291	-0,0207	0,0379	0,0281	-0,0215	-0,0021	-0,0243	-0,0035	0,0018	-0,0013
Corse	-0,1891	0,1485	0,0941	0,0158	-0,0393	0,0815	-0,0221	0,0048	0,0168	-0,0430	-0,0030
Fr-Comte	-0,1048	0,0093	0,0012	0,0214	0,0065	-0,0675	0,0074	0,0076	0,0156	-0,0113	0,0070
Ile de Fr	0,1263	0,0854	-0,0231	-0,0237	0,0252	0,0248	0,0043	-0,0096	-0,0044	-0,0003	0,0019
Lang.Rous	-0,1029	0,0235	0,0902	-0,0011	-0,0725	-0,0161	0,0440	0,0049	-0,0129	-0,0024	0,0117
Limousin	0,0673	-0,1103	0,1199	-0,0064	-0,0097	0,0694	0,0272	0,0223	-0,0073	-0,0038	0,0000
Lorraine	-0,1264	-0,0054	-0,0137	-0,0266	0,0617	-0,0667	0,0279	-0,0095	-0,0103	-0,0070	-0,0075
Midi-Pyr	0,1022	-0,0708	0,0738	-0,0392	-0,0527	-0,0405	0,0213	-0,0058	0,0056	0,0058	-0,0084
Nord-PdC	-0,1688	-0,0608	0,1367	-0,0191	0,0462	0,0316	-0,0103	0,0009	0,0167	0,0055	0,0016
Basse-Nor	-0,0283	-0,0936	-0,0604	0,0141	-0,0312	-0,0051	-0,0926	0,0179	-0,0228	-0,0029	-0,0024
Haute-Nor	-0,0824	-0,0522	0,0329	-0,0046	0,0454	0,0203	0,0131	0,0230	-0,0113	-0,0025	-0,0041
Pays Loire	0,0760	-0,1216	-0,1343	0,1073	0,0087	0,0163	0,0257	0,0007	0,0052	0,0029	-0,0010
Picardie	-0,1995	-0,0360	0,0648	0,0353	0,0272	-0,0189	-0,0245	-0,0121	-0,0126	0,0018	0,0047
Poit. Char	0,0786	-0,1151	0,0190	0,0905	-0,0271	-0,0078	-0,0220	-0,0213	0,0044	-0,0104	0,0040
PACA	-0,1308	0,1378	-0,0105	0,0295	-0,0411	0,0282	-0,0025	0,0030	0,0026	-0,0044	-0,0073
Rhone-Alp	-0,0028	0,0338	-0,0659	-0,0203	-0,0190	-0,0061	-0,0030	0,0116	0,0117	0,0024	0,0016
Outremer	0,3201	0,1623	0,2393	0,1202	0,0258	-0,0574	-0,0250	0,0170	0,0012	0,0066	-0,0008

On vérifie que :

$$d^2_{eucl}(Alsace', Aquitaine') = (-0{,}1219 - 0{,}0821)^2 + \ldots + (0{,}0047 + 0{,}0003)^2 = 0{,}1315$$

De même, on avait établi que :

$$d^2_{\Phi^2}(Alsace, Moyenne) = \frac{(0{,}3619 - 0{,}3111)^2}{0{,}3111} + \ldots + \frac{(0{,}0025 - 0{,}0033)^2}{0{,}0033} = 0{,}0668$$

Et l'on a :

$$d^2_{eucl}(Alsace', O) = (-0{,}1219)^2 + \ldots + (-0{,}0047)^2 = 0{,}0668$$

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	Fact. 11
Sarkozy	-0,0071	0,0911	-0,0274	0,0219	-0,0015	0,0112	-0,0127	0,0045	-0,0033	-0,0001	-0,0001
Bayrou	0,0610	-0,0500	-0,1032	-0,0628	-0,0018	0,0026	0,0008	-0,0065	-0,0008	0,0020	-0,0006
Royal	0,1244	-0,0146	0,0864	0,0146	0,0040	-0,0136	0,0059	-0,0065	0,0015	-0,0007	0,0001
Le Pen	-0,2839	0,0274	0,0297	-0,0159	-0,0053	-0,0272	0,0153	-0,0141	0,0049	-0,0013	-0,0007
Besancenot	-0,1026	-0,1497	0,0368	-0,0185	0,0643	-0,0009	0,0037	0,0278	-0,0178	-0,0173	-0,0044
Villiers	-0,0527	-0,2195	-0,2150	0,2222	-0,0078	0,0323	0,0499	-0,0058	0,0070	0,0027	-0,0022
Voynet	0,0269	-0,0095	-0,0991	-0,0196	0,0185	-0,0342	-0,0012	0,0386	0,0431	-0,0149	0,0232
Laguiller	-0,1878	-0,1426	0,0812	0,0180	0,1538	-0,0308	-0,0433	0,0199	-0,0030	0,0411	0,0073
Bove	0,0222	-0,0298	0,0358	-0,0256	-0,1394	-0,0855	0,0229	0,0693	0,0152	0,0154	-0,0175
Buffet	-0,0886	-0,0149	0,1921	-0,0728	-0,0097	0,1774	0,0482	0,0134	0,0146	0,0062	0,0001
Nihous	-0,1880	-0,3514	0,1310	0,0318	-0,1420	0,0417	-0,1461	-0,0131	0,0039	-0,0050	0,0035
Schivardi	-0,1060	-0,0930	0,0921	-0,0164	-0,2420	-0,0376	0,1260	0,0237	-0,0955	0,0151	0,0458

On avait établi que :

$$d^2_{\Phi^2}(Sarkozy, Bayrou) = \frac{(0{,}0320 - 0{,}0317)^2}{0{,}0275} + \ldots + \frac{(0{,}0298 - 0{,}0154)^2}{0{,}0264} = 0{,}0379$$

On retrouve ici :

$$d^2_{eucl}(Sarkozy', Bayrou') = (-0{,}0071 - 0{,}0610)^2 + \ldots + (-0{,}0001 + 0{,}0006)^2 = 0{,}0379$$

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

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

3.2.3.6 Reconstitution des données

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

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

Par exemple, le taux de liaison entre Alsace et le candidat Sarkozy peut être retrouvé à l'aide du calcul suivant :

$$t_{11} = \frac{(-0{,}1219)(-0{,}0071)}{\sqrt{0{,}01476}} + \frac{(0{,}1067)(-0{,}0911)}{\sqrt{0{,}00693}} + \ldots + \frac{(0{,}0047)(-0{,}0001)}{\sqrt{0{,}00002}} = 0{,}16$$

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

3.2.4 Interprétation des résultats de l'AFC

Au niveau global, on pourra noter que les inerties relatives les plus fortes sont observées sur l'Outremer, l'Ile de France, le Nord Pas de Calais et les Pays de la Loire pour les régions, et sur Le Pen, Royal, Bayrou, Villiers 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, Royal, Sarkozy) soit à des modalités éloignées du profil moyen (Outremer, Pays de Loire, Villiers, Le Pen).

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

3.2.4.1 Interprétation des axes

Pour le premier axe :

- Points lignes :

-	+
Nord - Pas de Calais (12%) Provence - Alpes Côte d'Azur (9%) Picardie (8%) Champagne Ardennes (5%)	Outremer (18%) Ile de France (17%) Bretagne (6%)

- Points colonnes :

-	+
Le Pen (57%)	Royal (27%)

Le premier axe oppose les régions du Nord et de l'Est, et la région PACA 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 Royal.

La synthèse entre l'analyse des lignes et des colonnes montre que cet axe oppose les régions où le vote pour le candidat 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 "Royal". 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 peu dans la formation de cet axe.

Pour le deuxième axe :

- Points lignes :

-	+
Pays de la Loire (13%) Poitou-Charentes (6%) Aquitaine (5%) Bretagne (5%)	Provence Alpes Côte d'Azur (21%) Ile de France (17%) Outremer (10%) Alsace (5%)

- Points colonnes :

-	+
Nihous (20%) Villiers (15%) Besancenot (13%)	Sarkozy (37%)

Cet axe oppose les régions de l'Ouest à des régions telles que PACA, l'Ile de France, l'Outremer et l'Alsace.

Pour les modalités colonnes, cet axe oppose certains "petits" candidats au candidat Sarkozy, mais il est, dans une certaine mesure unipolaire (Sarkozy représente 37% de son inertie)..

Cet axe est essentiellement organisé autour du vote pour le candidat Sarkozy : la partie positive de l'axe correspond aux régions où le vote Sarkozy est supérieur à la moyenne nationale, tandis que la partie négative correspond à des régions où ce vote est inférieur à la moyenne. Il semble par ailleurs qu'un faible vote pour Sarkozy soit lié à un vote plus significatif pour certains petits candidats.

Pour le troisième axe :

- Points lignes :

-	+
Pays de la Loire (16%) Alsace (14%) Rhône Alpes (6%)	Outremer (23%) Nord Pas de Calais (18%) Languedoc Roussillon (5%)

- Points colonnes :

-	+
Bayrou (30%)	Royal (29%)
Villiers (16%)	

Le troisième axe oppose nettement le vote pour Bayrou, bien représenté dans des régions telles que les Pays de la Loire et l'Alsace (partie négative de l'axe) au vote pour Royal, particulièrement élevé dans la "région" Outremer (partie positive de l'axe).

3.2.4.2 Remarques :

1. Etant donné le poids des suffrages obtenus par le candidat Sarkozy (31% de l'ensemble), on aurait pu s'attendre à ce que cette modalité colonne ait une grande influence dans la détermination du profil moyen et donc que le point représentant le candidat soit très proche de l'origine. On remarque malgré tout que ce point reste bien distinct 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 à leur moyenne nationale, alors qu'ils obtiennent simultanément des scores inférieurs 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.

3.2.5 Quelques principes d'interprétation supplémentaires

3.2.5.1 Forme générale du nuage

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

3.2.5.2 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.

3.2.5.3 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.

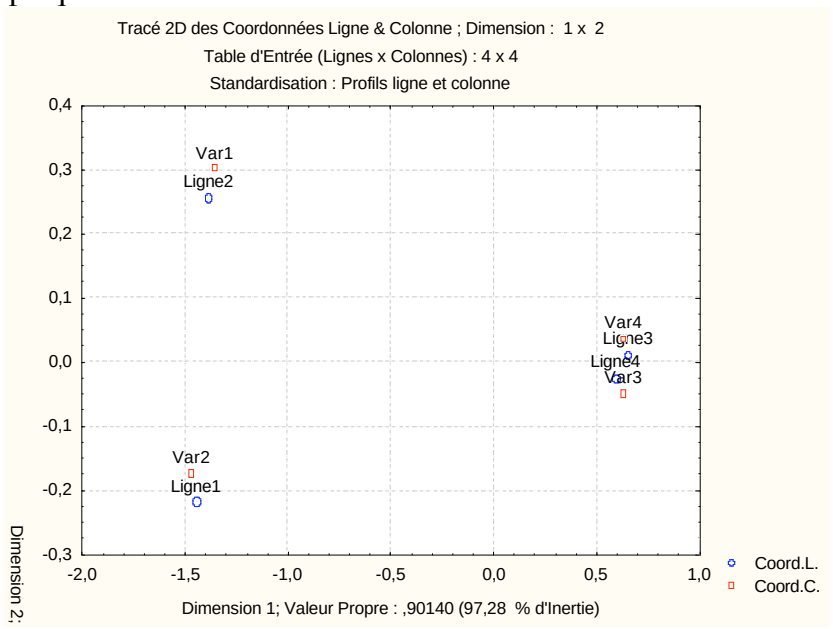
Exemple : Soit le tableau de contingence suivant :

	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 :



3.2.5.4 L'effet Guttman.

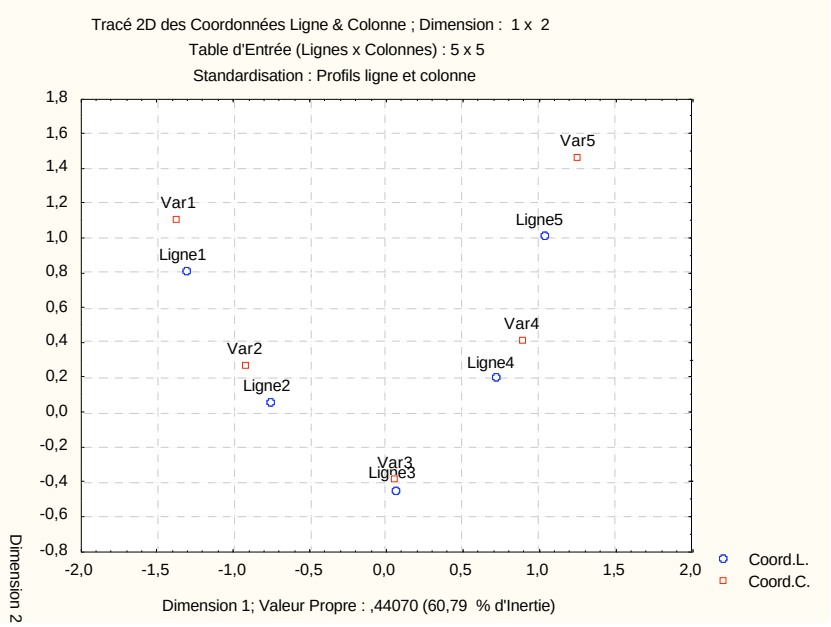
Un nuage de points de forme parabolique indique une redondance entre les deux variables étudiées : la connaissance de la ligne i donne pratiquement celle de la colonne j. Dans un tel cas, pratiquement toute l'information est contenue dans le premier facteur. Cette configuration se rencontre notamment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



3.3 Analyse factorielle des correspondances avec Statistica

3.3.1 Présentation des données étudiées

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

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

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

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

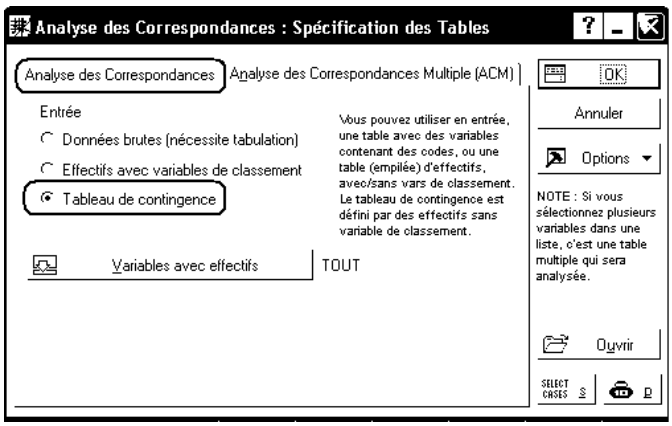
	Radio	Tél.	Quot.N.	Quot R.	P.Mag.	P.TV
Professions						
Agriculteur	96	118	2	71	50	17
Petit patron	122	136	11	76	49	41
Prof. Cad. S.	193	184	74	63	103	79
Prof. interm	360	365	63	145	141	184
Employé	511	593	57	217	172	306
Ouvrier qual	385	457	42	174	104	220
Ouvrier n-q	156	185	8	69	42	85
Inactif	1474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

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

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

3.3.2 Traitement des données avec Statistica

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



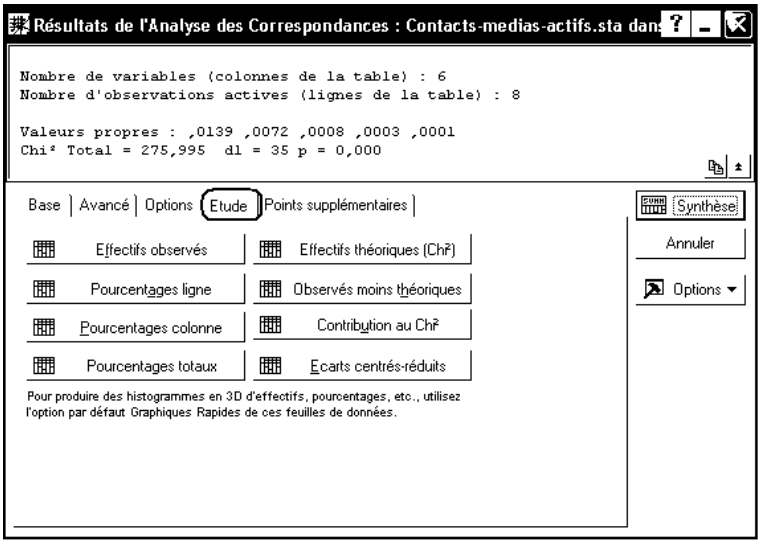
La fenêtre de dialogue permet d'indiquer la manière dont se présentent nos données. La situation la plus classique est celle d'un tableau de contingence : les modalités lignes sont indiquées comme noms d'observations (elles auraient pu être indiquées dans une variable spécifique), les modalités colonnes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

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

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

3.3.2.1 Statistiques descriptives

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



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

	Pourcentages Totaux (Contacts-medias-actifs.sta dans Classeur1) Table d'Entrée (Lignes x Colonnes) : 8 x 6 Inertie Totale = ,02228 Chi? = 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	

3.3.2.2 Choix des valeurs propres

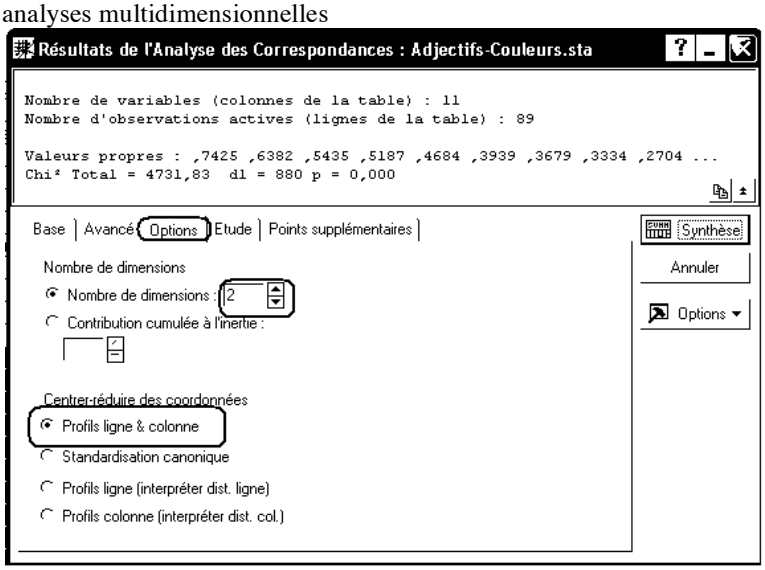
C'est ensuite l'onglet "Avancé" qui nous permettra d'afficher les valeurs propres, et donc de choisir le nombre d'axes à garder :

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

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

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

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

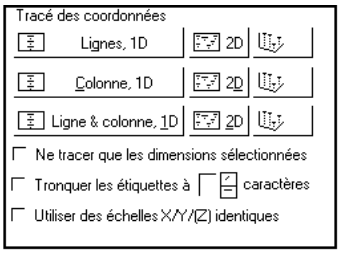


On retourne ensuite sous l'onglet "Avancé" pour afficher les coordonnées des individus-lignes et des individus-colonnes. On notera que Statistica produit deux tableaux de résultats, et on passera de l'un à l'autre à l'aide des onglets du classeur.

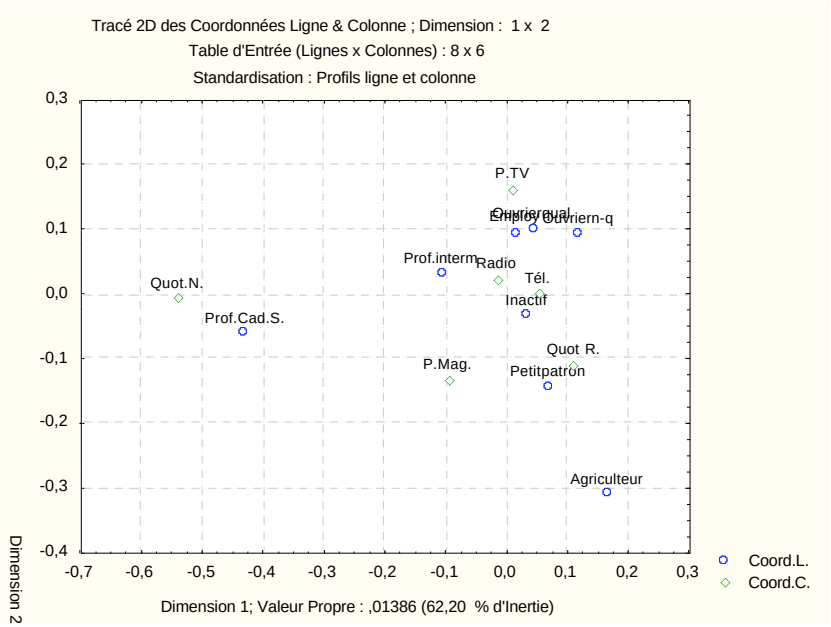
	Coordonnées Ligne et Contributions à l'Inertie Table d'Entrée (Lignes x Colonnes) : 8 x 6 Standardisation : Profils ligne et colonne									
NomLigne	Ligne Numéro	Coord. Dim.1	Coord. Dim.2	Masse	Qualité	Inertie Relative	Inertie Dim.1	Cosinus² Dim.1	Inertie Dim.2	Cosinus² Dim.2
Agriculteur	1	0,1661	-0,3096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

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

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



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



3.3.2.4 Individus ligne et individus colonne supplémentaires

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

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

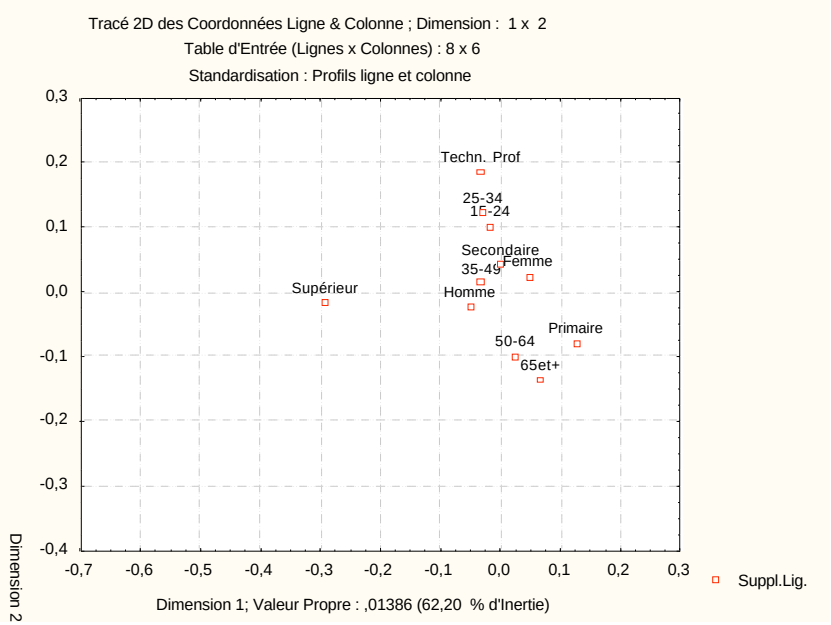
Positionner les groupes définis par l'âge, le sexe, le niveau d'étude

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

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

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

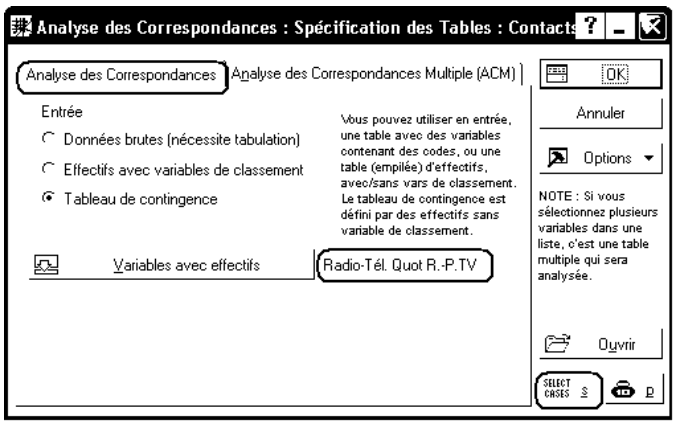
Le graphique qui suit ne représente que les points supplémentaires. Il a été obtenu en réalisant un graphique pour tous les points, puis en modifiant les options d'affichage de façon à faire disparaître les individus lignes et colonnes actifs :

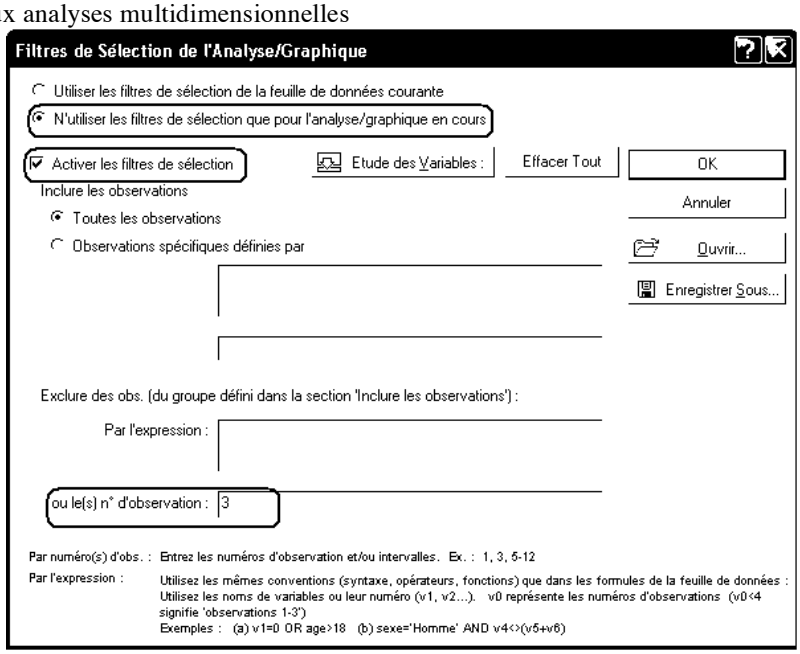


Rendre inactives certaines modalités des variables

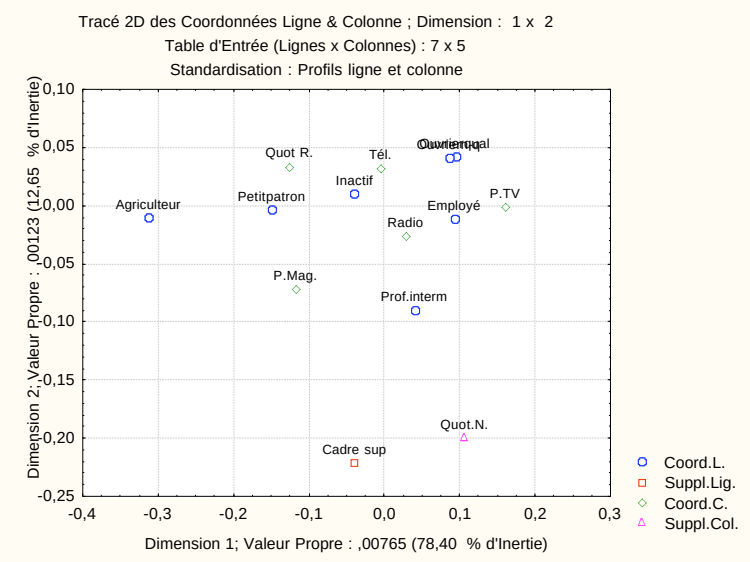
On veut réaliser l'étude en plaçant en éléments supplémentaires l'individu-ligne "Prof.Cad.S." et l'individu colonne "Quot.N.".

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 :



3.4 Exercices et prolongements

3.4.1 Structures possibles pour les données d'entrée

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

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

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

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

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

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

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

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

3.4.2 Exercice à traiter à l'aide de Statistica

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

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

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

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

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

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

3.4.3 Exercice : associations Adjectifs-couleurs

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

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

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

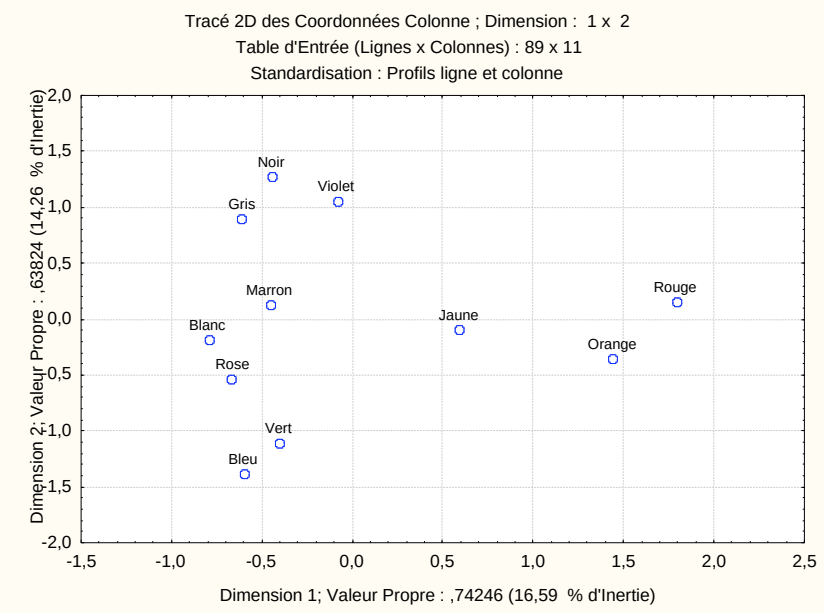
3.4.3.1 Choix des valeurs propres

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

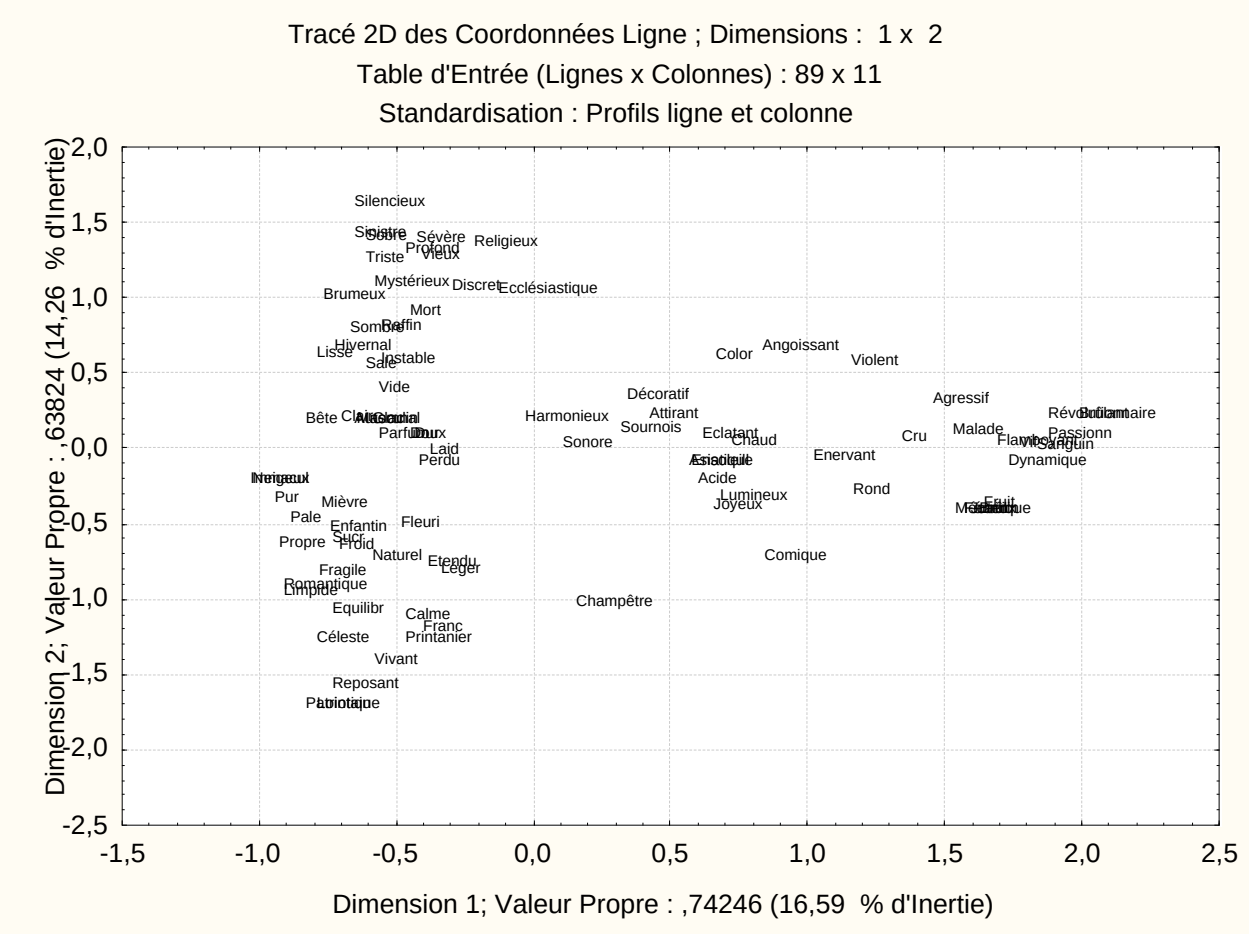
On voit ici que la décroissance des valeurs propres est très lente. Selon les règles énoncées précédemment, il faudrait conserver au moins 5 axes. Mais nous pouvons convenir de ne rechercher que les propriétés les plus caractéristiques des associations adjectifs - couleurs en n'étudiant que les deux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

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

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



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



3.4.3.3 Quelques éléments d'interprétation

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

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

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

3.4.4 Exercice : le cas Environnement

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

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

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

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

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

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

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

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

3.4.5 Exercice : étude des réponses à une question ouverte

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

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

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

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

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

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

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

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

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

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

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

5) Etude de la première variable factorielle

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

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

6) Mener une étude analogue pour la deuxième variable.

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

3.5 Travail à rendre

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

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

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

Les données recueillies sont les suivantes :

	Hôtel	Location	Résid. 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*
Autres actifs	142	103	210	1133	132	181	46	59	*2006*
Inactifs	741	332	327	1789	311	236	102	102	*3940*
Total	*4535*	*3377*	*2232*	*12780*	*1858*	*3856*	*1336*	*1105*	*31079*

Traitez ce tableau par une analyse factorielle des correspondances et répondez aux questions suivantes :
1) Comment ont été obtenues les premières valeurs respectives (0,63%, 18,47% et 4,30%) du tableau des fréquences, de celui des fréquences lignes et de celui des fréquences colonnes ?

2) a) Utilisez cependant 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 0,1276, alors que pour l'individu "Résidence secondaire", ces valeurs sont respectivement 0,0718 et 0,2236. Comment peut-on interpréter ces résultats ?

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

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

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

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

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

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

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

- Un classeur Statistica contenant les résultats numériques de l'AFC et les graphiques.
- Un 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.