

Analyse de données

TAI



Promo 2019

CHAZAL Baptiste – SIMON Andrieux

DORIAN Antoine – MOUSSEAUX Thomas

Sommaire

Introduction.....	3
Quel type d'analyse ?.....	5
Analyse de la matrice des corrélations.....	6
Choix des axes.....	7
Matrice des saturations.....	8
Cercles de corrélation.....	10
Choix des individus intéressants.....	12
Synthèse.....	13

Introduction

Dans le cadre du TAI d'Analyse de Données, nous en venons à réaliser une analyse complète de données statistiques. Pour cela, nous avons choisi des chiffres provenant d'Ipsos, un important institut de sondages français. Nous utilisons le logiciel XLStat afin de nous simplifier la génération des matrices.

Nous réaliserons donc ici une analyse des orientations politiques en fonctions des catégories socioprofessionnelles durant les dernières élections municipales.

Nos données, présentées sur la page suivante, sont uniquement des pourcentages.

		Extrême Gauche	Front de gauche	EELV et alliés	PS/PRG	LR, UDI et Modem	Debout la France	Front National	Autres Listes
Sexe	Femme	2	4	7	26	27	4	23	3
	Homme	1	4	6	21	27	4	31	3
Age	18-24 ans	1	4	9	21	21	5	35	3
	25-34 ans	3	3	10	27	21	3	28	2
	35-59 ans	2	4	6	24	22	4	32	2
	60 ans et +	1	4	6	24	35	4	20	3
	Agriculteur, Artisan, Chef d'entreprise	0	3	3	17	34	1	35	3
CSP	Cadre supérieur	1	3	10	30	29	5	17	2
	Profession intermédiaire	2	5	9	28	24	3	25	2
	Employé	3	3	6	20	21	4	36	2
	Ouvrier	4	3	4	20	17	4	43	2
	Retraité	1	4	6	25	33	4	21	3
Niveau de diplôme	Inférieur au BAC	3	4	4	20	23	4	36	3
	Baccalauréat	1	3	7	24	26	3	32	1
	Bac+ 2	1	4	7	26	28	4	23	3
	Au moins Bac +3	1	4	10	27	33	5	14	3

Quel type d'analyse ?

Tout d'abord, rappelons les conditions nécessaires à l'utilisation de tel ou tel type d'analyse de données statistiques. Dans le cas de données qualitatives, on utilisera une ACM (Analyse des Correspondances Multiples) ; Dans le cas où nous avons 2 variables quelconques, on utilisera une AFC (Analyse Factorielle des Correspondances) ; Enfin, dans le cas de variables quantitatives de même importance, on utilisera plutôt une ACP (Analyse en Composantes Principales). Dans notre cas, nous avons bien plus de deux variables, et celles-ci sont de même importance, nous optons donc pour une ACP.

A présent, nous devons nous demander s'il est nécessaire d'utiliser des données centrées réduites. Cela est nécessaire lorsque les unités sont différentes, lorsque les ordres de grandeurs sont différents, ou lorsque les écarts-type sont trop différents. Ici, notre tableau contient uniquement des pourcentages, il n'est donc pas nécessaire d'utiliser des données centrées réduites.

Notons les statistiques descriptives de cette analyse :

Variable	Minimum	Maximum	Moyenne	Ecart-type
Extrême Gauche	0,000	4,000	1,688	1,078
Front de gauche	3,000	5,000	3,688	0,602
EELV et alliés	3,000	10,000	6,875	2,217
PS/PRG	17,000	30,000	23,750	3,587
LR, UDI et Modem	17,000	35,000	26,313	5,425
Debout la France	1,000	5,000	3,813	0,981
Front National	14,000	43,000	28,188	8,085
Autres Listes	1,000	3,000	2,500	0,632

Analyse de la matrice des corrélations

Avant d'analyser la matrice des corrélations, il faut tout d'abord la construire. Rappelons sa formule :

$$R = \frac{1}{n} * t_X^S * x^S$$

On obtient la matrice des corrélations suivante :

Variables	Extrême Gauche	Front de gauche	EELV et alliés	PS/PRG	LR, UDI et Modem	Debout la France	Front National	Autres Listes
Extrême Gauche	1	-0,160	-0,129	-0,108	-0,791	0,130	0,489	-0,342
Front de gauche	-0,160	1	0,169	0,270	0,154	0,233	-0,343	0,438
EELV et alliés	-0,129	0,169	1	0,800	-0,013	0,448	-0,608	-0,190
PS/PRG	-0,108	0,270	0,800	1	0,186	0,365	-0,803	-0,206
LR, UDI et Modem	-0,791	0,154	-0,013	0,186	1	-0,139	-0,688	0,476
Debout la France	0,130	0,233	0,448	0,365	-0,139	1	-0,331	0,161
Front National	0,489	-0,343	-0,608	-0,803	-0,688	-0,331	1	-0,254
Autres Listes	-0,342	0,438	-0,190	-0,206	0,476	0,161	-0,254	1

On note ainsi les variables présentant une forte corrélation entre elles (> 0.6, en vert), ou une forte anti-corrélation (< -0.6, en jaune). On obtient ainsi ceci :

Forte corrélation	Forte anti-corrélation
PS/PRG - EELV et alliés	Extrême gauche - LR, UDI et Modem
	EELV et alliés - Front national
	PS/PRG - Front national
	LR, UDI et Modem - Front national

Choix des axes

Afin de choisir les axes, nous devons tout d'abord calculer les valeurs propres de notre tableau statistique. Pour cela, on utilisera la formule suivante :

$$\det(R - \lambda I) = 0$$

Rappelons que R représente notre matrice des corrélations.

Ainsi, nous obtenons les valeurs propres suivantes :

	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8
Valeur propre	3,217	2,120	1,269	0,666	0,391	0,269	0,063	0,004
Qualité (%)	40,216	26,504	15,862	8,331	4,889	3,364	0,784	0,051
Qualité cumulée	40,216	66,720	82,582	90,913	95,802	99,165	99,949	100,000

Rappelons que la qualité d'une valeur propre λ_i vaut :

$$\frac{\lambda_i}{\sum \lambda} * 100$$

Un qge d'au moins 70 étant nécessaire, on choisira trois axes. Notre qge vaut donc 82,582%.

Matrice des saturations

Pour la suite de l'analyse, notamment la représentation graphique via des cercles de corrélation, il est nécessaire de construire la matrice des saturations. Rappelons sa formule :

$$S = \frac{1}{n} * t_{X^s} * F * \Lambda^{-\frac{1}{2}}$$

Λ est la matrice des valeurs propres, c'est une matrice diagonale équivalent à R la matrice des corrélations. F, quant à elle, représente la matrice des composantes principales, qui s'obtient comme ceci :

$$F = A^s * U$$

A représente notre matrice de base (le tableau initial), et U la matrice des vecteurs propres.

Notre matrice des vecteurs propres U est la suivante :

	Axe 1	Axe 2	Axe 3	Axe 4	Axe 5	Axe 6	Axe 7	Axe 8
Extrême Gauche	0,325	-0,411	0,280	0,075	0,689	0,154	-0,359	0,124
Front de gauche	-0,272	0,065	0,557	0,703	-0,170	-0,246	-0,165	0,020
EELV et alliés	-0,368	-0,428	-0,120	-0,007	-0,399	0,565	-0,409	0,147
PS/PRG	-0,426	-0,373	-0,172	0,200	0,292	-0,033	0,622	0,370
LR, UDI et Modem	-0,361	0,457	-0,240	-0,119	0,273	-0,208	-0,449	0,519
Debout la France	-0,225	-0,326	0,477	-0,630	-0,141	-0,437	-0,022	0,085
Front National	0,536	0,017	0,116	0,044	-0,367	0,029	0,129	0,738
Autres Listes	-0,188	0,436	0,519	-0,217	0,154	0,600	0,266	0,065

On obtient donc notre matrice des composantes principales :

	Axe 1	Axe 2	Axe 3	Axe 4	Axe 5	Axe 6	Axe 7	Axe 8
Femme	-0,947	-0,014	0,674	0,182	0,657	0,282	0,252	-0,056
Homme	0,074	1,132	0,828	-0,130	-0,611	-0,051	0,023	-0,021
18-24 ans	0,010	-0,322	1,495	-0,644	-1,816	0,531	0,007	0,058
25-34 ans	0,495	-2,105	-1,233	0,279	0,487	1,370	-0,017	0,024
35-59 ans	0,737	-0,728	0,342	0,589	-0,259	-0,709	0,225	-0,088
60 ans et +	-1,597	1,483	0,151	-0,201	0,572	-0,438	-0,305	0,052
Agriculteur, Artisan, Chef d'entreprise	1,738	4,087	-1,796	0,226	-0,134	0,591	0,043	-0,005
Cadre supérieur	-2,273	-1,652	-1,441	-1,263	0,053	-0,233	0,296	0,039
Profession intermédiaire	-1,113	-1,142	0,236	2,593	-0,191	-0,025	-0,175	0,027
Employé	2,349	-0,882	-0,043	-0,730	0,118	-0,046	-0,401	-0,152
Ouvrier	3,757	-1,210	0,623	-0,521	0,613	-0,240	0,093	0,094
Retraité	-1,514	1,204	0,208	-0,092	0,505	-0,365	0,062	0,055
Inférieur au BAC	1,779	0,514	1,781	0,082	0,553	-0,095	-0,039	0,049
Baccalauréat	0,991	-0,665	-2,469	0,234	-0,707	-0,848	-0,092	0,028
Bac+ 2	-1,326	0,466	0,360	0,087	0,050	0,095	0,511	-0,075
Au moins Bac +3	-3,160	-0,166	0,284	-0,693	0,109	0,183	-0,481	-0,028

Et notre matrice des saturations :

	Axe 1	Axe 2	Axe 3	Axe 4	Axe 5	Axe 6	Axe 7	Axe 8
Extrême Gauche	0,583	-0,598	0,315	0,061	0,431	0,080	-0,090	0,008
Front de gauche	-0,488	0,095	0,628	0,574	-0,107	-0,128	-0,041	0,001
EELV et alliés	-0,659	-0,623	-0,135	-0,006	-0,249	0,293	-0,102	0,009
PS/PRG	-0,763	-0,543	-0,193	0,163	0,182	-0,017	0,156	0,024
LR, UDI et Modem	-0,648	0,666	-0,270	-0,097	0,171	-0,108	-0,112	0,033
Debout la France	-0,404	-0,474	0,537	-0,515	-0,088	-0,227	-0,006	0,005
Front National	0,962	0,024	0,131	0,036	-0,229	0,015	0,032	0,047
Autres Listes	-0,338	0,634	0,585	-0,177	0,097	0,311	0,067	0,004

Cercles de corrélations

A partir du tableau de saturation, on relève les variables illustrant le mieux les axes, c'est-à-dire les variables avec lesquelles l'axe présente une forte corrélation, ou une forte anti-corrélation :

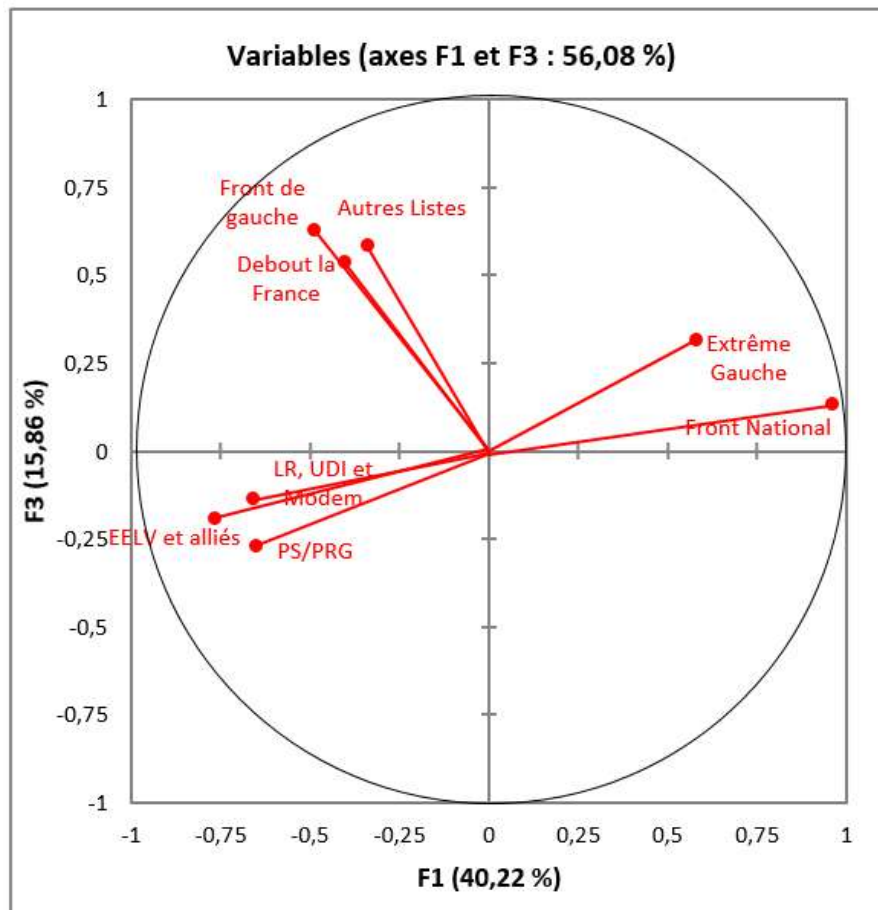
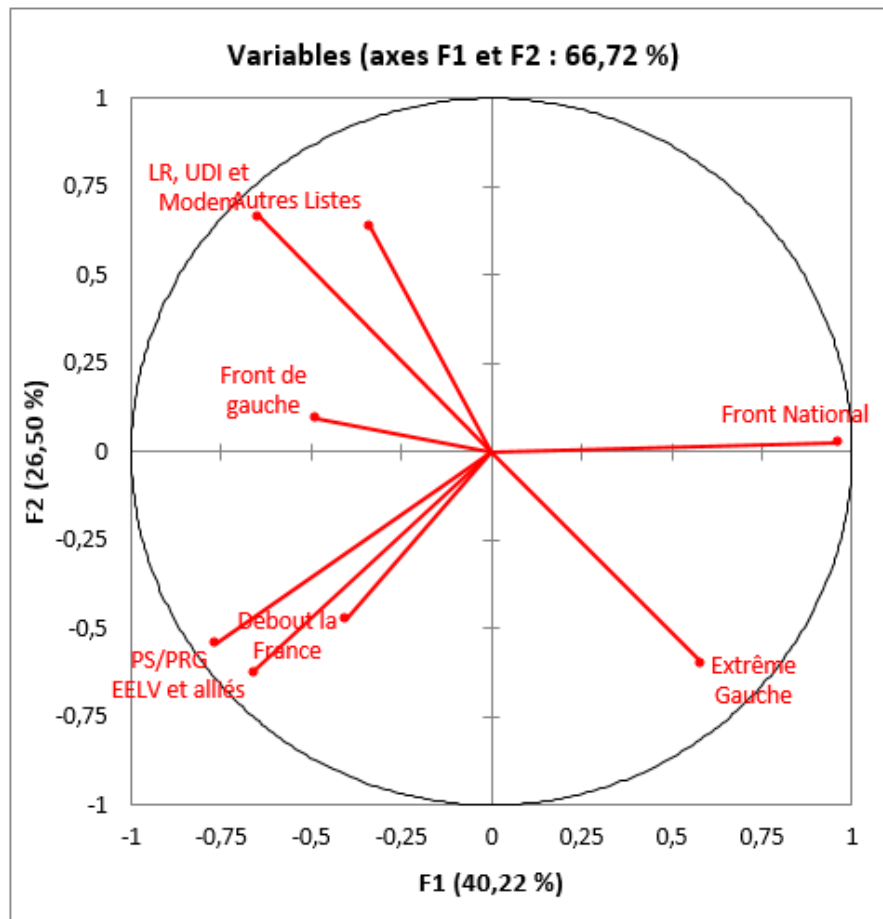
Axe 1 : Front national - EELV et alliés, PS/PRG

Axe 2 : LR, UDI & Modem, Autres listes - Extrême gauche

Axe 3 : Front de gauche

(Axe i : Corrélations – Anti-corrélations)

On place ensuite les variables sur ces axes en utilisant comme coordonnées les valeurs de la matrice des saturations.



Choix des individus intéressants

A partir de la matrice des composantes principales, on détermine les QLT de nos classes pour chacun des trois axes :

	Axe 1	Axe 2	Axe 3	cum.
Femme	0,457	0,000	0,231	0,688
Homme	0,002	0,542	0,290	0,834
18-24 ans	0,000	0,016	0,353	0,369
25-34 ans	0,029	0,528	0,181	0,738
35-59 ans	0,251	0,245	0,054	0,550
60 ans et +	0,470	0,405	0,004	0,879
Agriculteur, Artisan, Chef d'entreprise	0,129	0,715	0,138	0,982
Cadre supérieur	0,441	0,233	0,177	0,851
Profession intermédiaire	0,132	0,139	0,006	0,277
Employé	0,785	0,111	0,000	0,896
Ouvrier	0,846	0,088	0,023	0,957
Retraité	0,547	0,346	0,010	0,904
Inférieur au BAC	0,457	0,038	0,458	0,953
Baccalauréat	0,112	0,050	0,692	0,854
Bac+ 2	0,736	0,091	0,054	0,881
Au moins Bac +3	0,920	0,003	0,007	0,930

On calcule pour chacune d'elles le QLT cumulé et on peut ainsi déterminer (au-dessus du seuil de 0.8) les individus intéressants pour faire une synthèse de l'analyse.

Synthèse

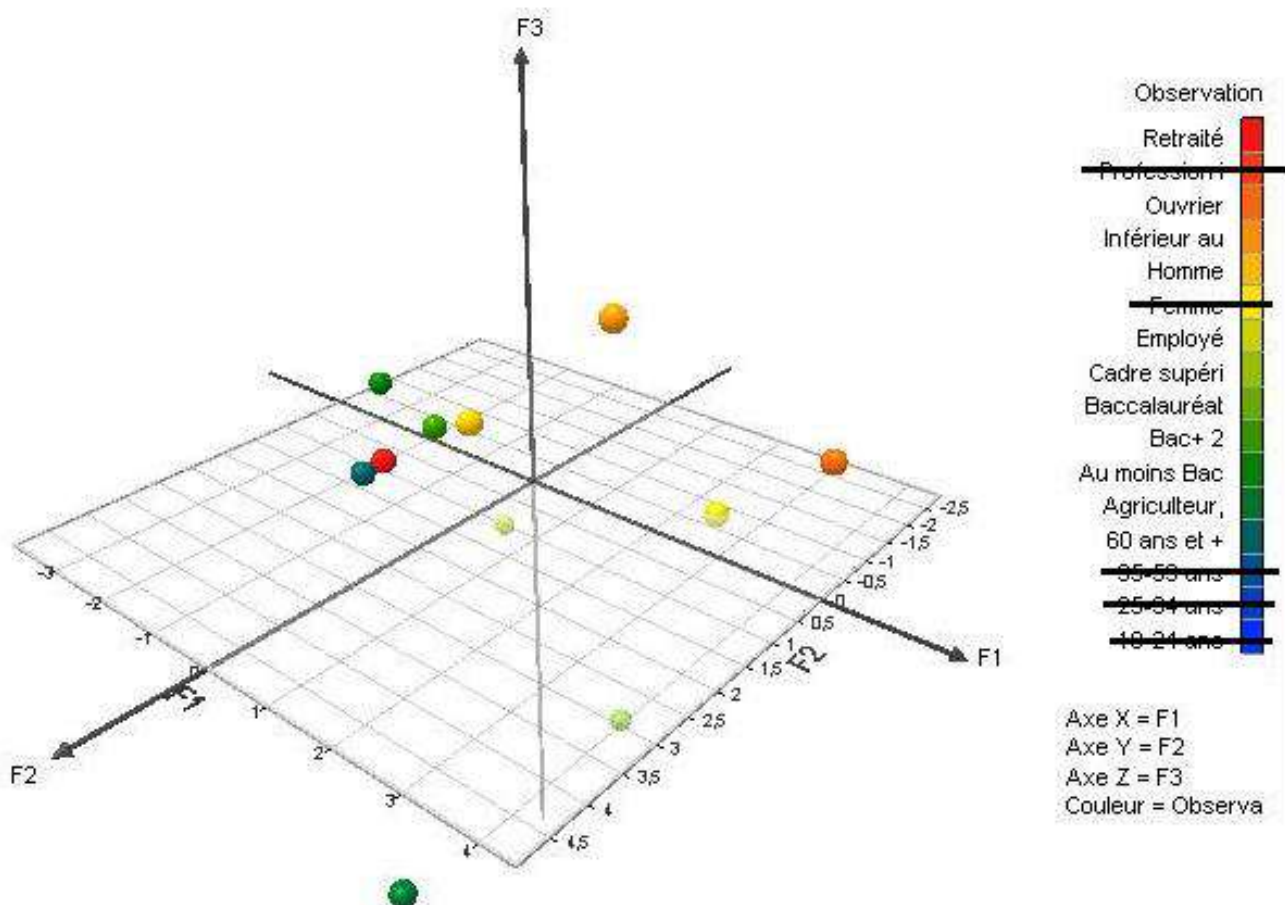
Rappelons ici la signification des axes :

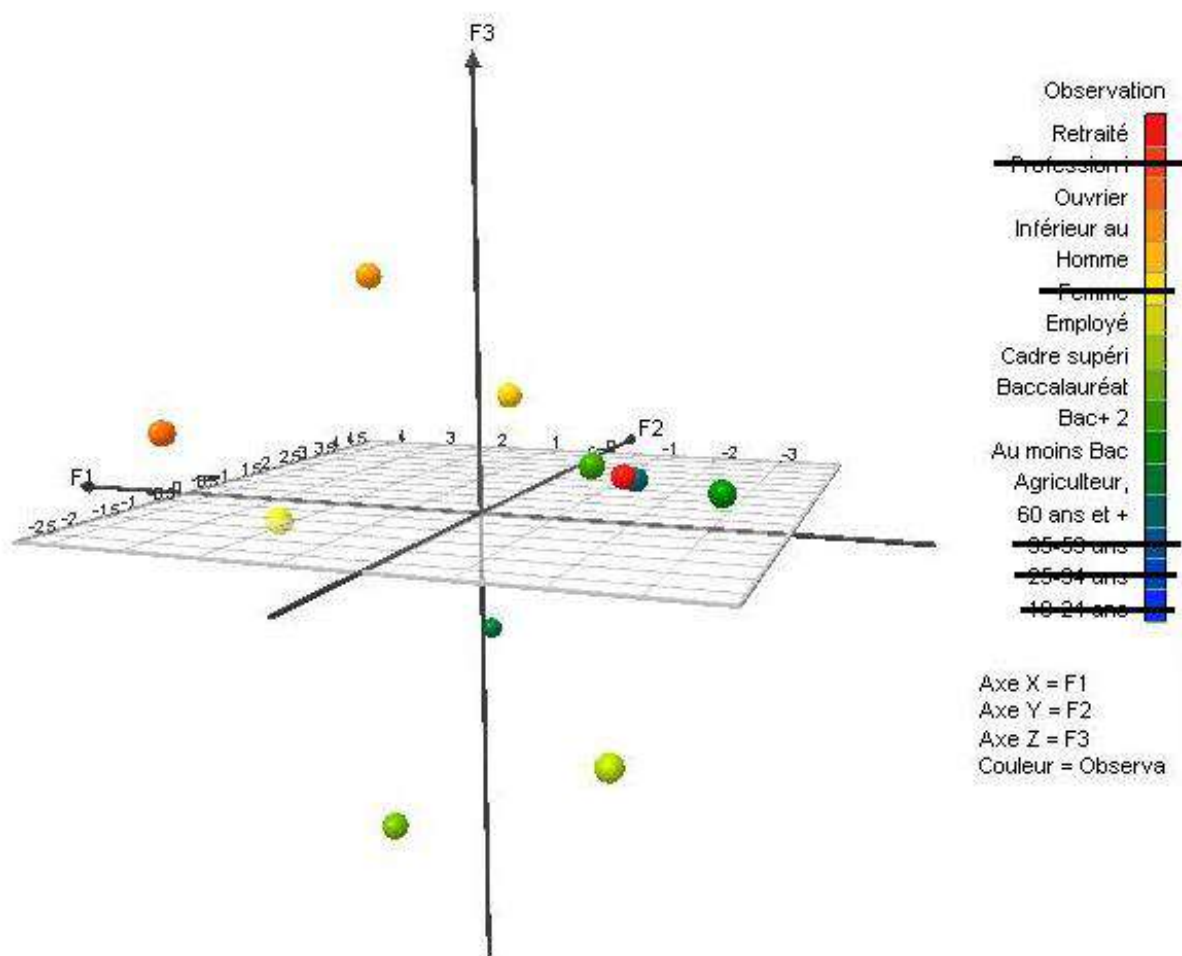
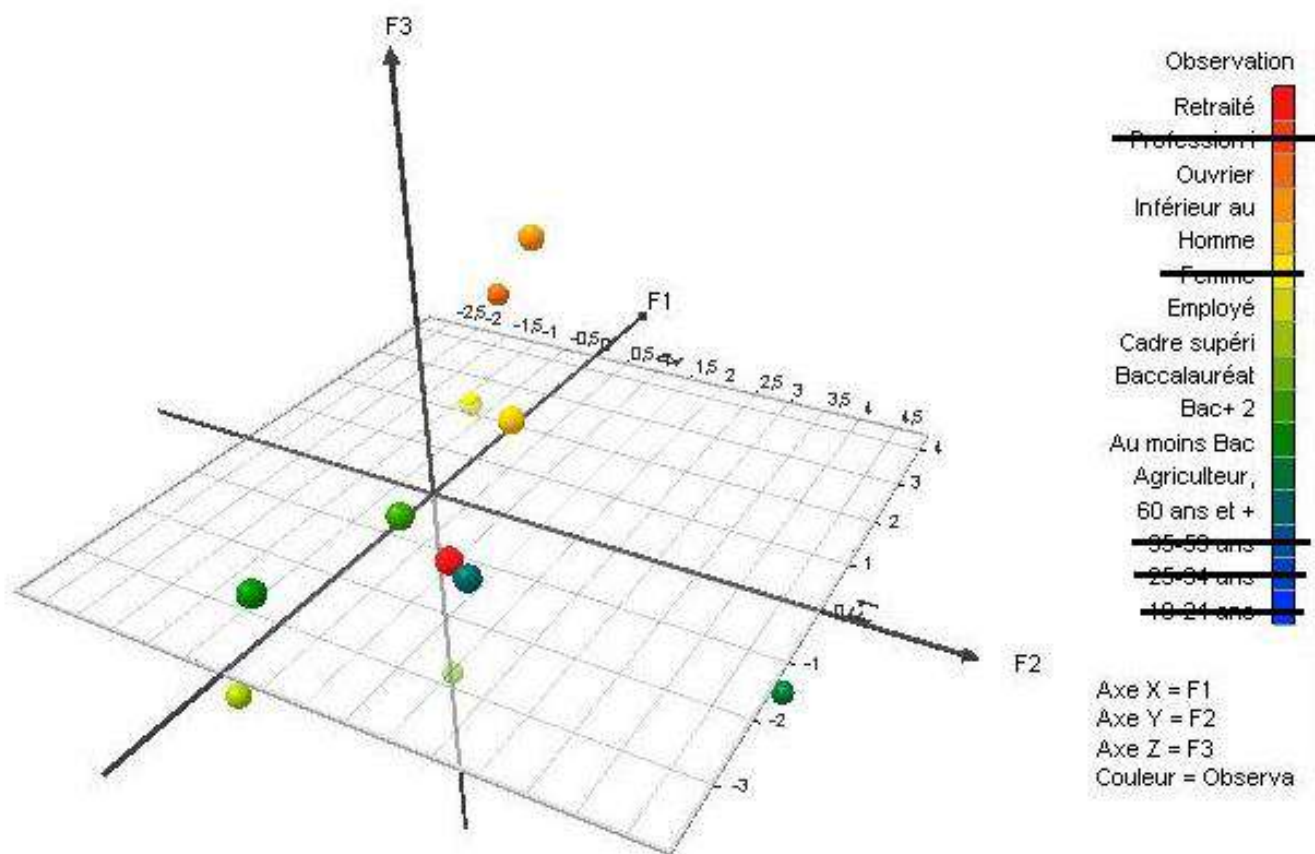
Axe 1 : Front national - EELV et alliés, PS/PRG

Axe 2 : LR, UDI & Modem, Autres listes - Extrême gauche

Axe 3 : Front de gauche

Puis, puisque nous avons trois axes, nous générons un graphique 3D à partir de XLStat. Les coordonnées que nous utilisons sont les valeurs de la matrice des composantes principales.





Ainsi, à partir de ce graphique 3D, on repère les tendances des différents groupes :

Les individus avec un niveau d'étude élevé, tout comme les personnes âgées, ont tendance à éviter les extrêmes, que ce soit extrême gauche ou extrême droite. Les individus de classe sociale plus élevée, quant à elles, ont plutôt tendance à voter à droite. On note une proximité dans les catégories socioprofessionnelles entre les personnes votant extrême gauche et front national (ouvrier et niveau d'étude inférieur au bac). Quant aux agriculteurs, on note une très forte tendance à éviter l'extrême gauche, et à voter à droite, que ce soit extrême ou non.

XLSTAT[>]