1 Objectif

La page Excel'Ense de MODULAD (http://www.modulad.fr/). Utilisation de l'add-in d'analyse exploratoire des données de Jacques Vaillé (Revue Modulad, n°43, 2011).

La revue MODULAD est consacrée aux statistiques et traitement de données. Elle est éditée depuis 1988. Elle met l'accent sur le bon usage des méthodes et des outils. En janvier 2005, la version papier est abandonnée au profit d'une diffusion sur le web. Tout un chacun peut accéder librement aux articles et aux ressources associées sans avoir à s'enregistrer[1].

Une des forces de la revue est d'avoir su trouver un équilibre entre les fondements théoriques, la pratique, et les outils. Il ouvre la porte aux non-spécialistes, attachés aux aspects applicatifs, mais désireux de s'appuyer sur des références théoriques solides. La politique éditoriale de la revue évoque, bien avant l'heure, ce que l'on a coutume d'appeler aujourd'hui « data science » ou encore « big data analytics » : « Avec l'émergence du calcul intensif et du stockage massif, sur des machines omniprésentes et souvent connectées entre elles, les attitudes aussi bien que les objectifs du traitement de l'information ont beaucoup évolué : moins d'hypothèses formelles (qui sont souvent des hypothèses de commodité), davantage d'approches graphiques, des modélisations plus souples, et le souci de caractériser les structures révélées par les données elles-mêmes…. La Revue MODULAD veut assurer la diffusion des solutions qui tirent avantage des ressources nouvelles appliquées à des problèmes nouveaux. Soit que ces problèmes mettent en jeu des données volumineuses ou complexes, parfois peu structurées comme les textes, les sons ou les images. Soit qu'ils fassent appel à des méthodes historiques, restées d'usage courant, mais souvent revisitées dans des perspectives innovantes. ».

La page Excel'Ense (http://www.modulad.fr/excel.htm) montre l'attachement de la revue au caractère pratique de l'analyse de données. Il s'agit d'un « cahier d'information et d'échanges, ciblé sur l'exploitation des données et l'enseignement de la statistique avec un tableur, le prototype étant Excel ». Concrètement, des classeurs types et macros-complémentaires (add-ins) adaptés à différents problèmes statistiques sont mis à disposition des lecteurs (ex. calcul des coefficients d'autocorrélation, construction de plans d'expériences factoriels complets, tests de Wilcoxon, tirage d'un échantillon stratifié, etc.). Nous pouvons les charger et appliquer les techniques proposées sur nos propres données via Excel.

[1] Les archives sont disponibles : http://www.modulad.fr/archives.htm ; les numéros antérieurs à 2005 ont été scannés.

Dans ce tutoriel, nous étudions l'add-in « Explore.xla » de Jacques Vaillé (2011). L'auteur met à notre disposition plusieurs outils et méthodes statistiques pour l'analyse exploratoire des données. Certains d'entre eux, très simples, sont pourtant particulièrement utiles. Une macro par exemple permet de réaliser un graphique nuage de points étiquetés à l'aide des labels des observations, chose impossible à faire avec les fonctions standards d'Excel. Une documentation accompagne la librairie. Nous nous en tiendrons aux techniques d'analyse factorielle dans notre présentation.

L'add-in fonctionne aussi bien sous Excel 2003 qu'avec Excel 2007. Je ne l'ai pas testé, mais on peut penser qu'il sera tout aussi opérationnel sous Excel 2010.

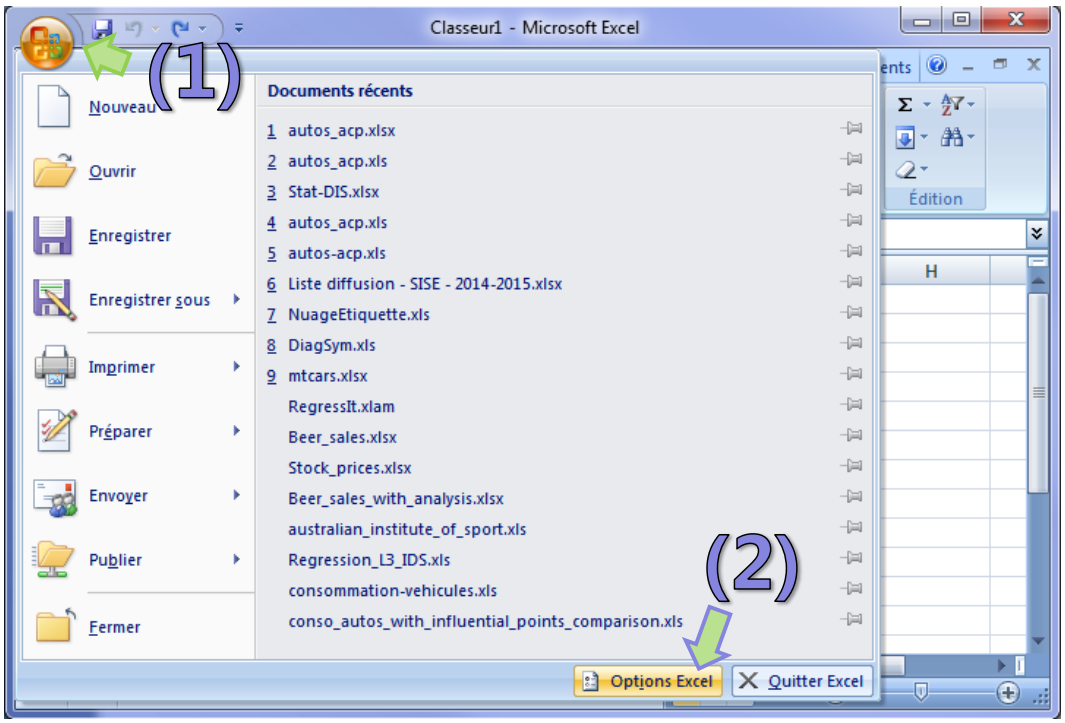
2 Chargement et installation de l'add-in « Explore.xla »

2.1 Chargement

Nous désarchivons le fichier « Vaillé-StatistiqueEtDonnes.zip » en provenance de la page Excel'Ense. Nous obtenons 2 fichiers : « **Explore.xla** » est la macro-complémentaire qu'il faudra intégrer dans Excel ; « **StatistiqueEtDonnes.pdf** » correspond à la documentation[2].

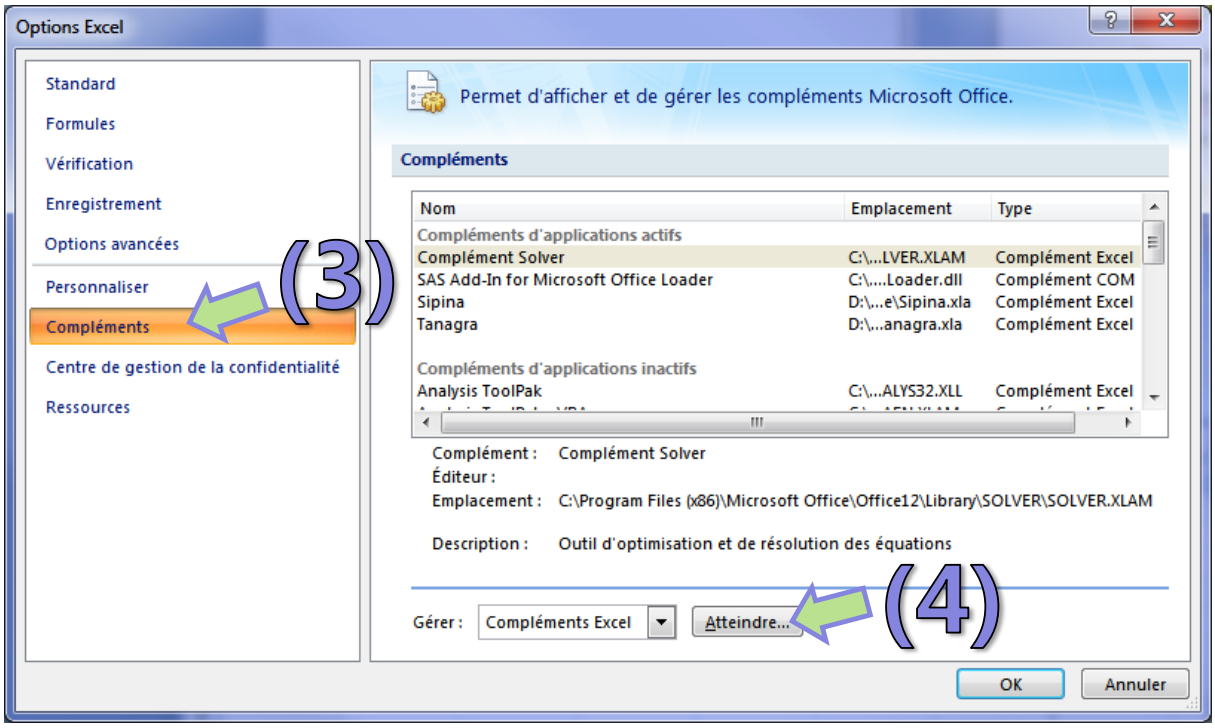
2.2 Intégration dans Excel

Nous utilisons la version 2007 d'Excel dans ce tutoriel. Après avoir démarré le logiciel, nous actionnons le menu « Options Excel ».

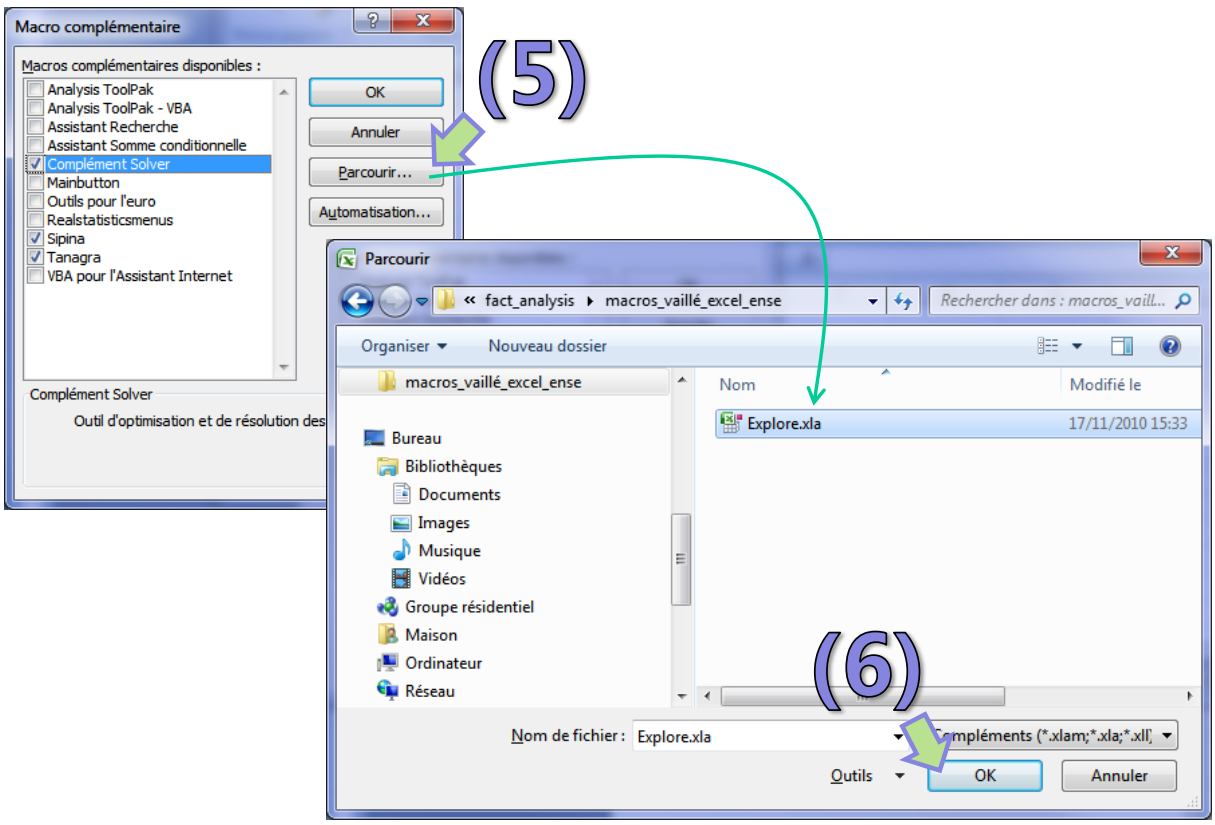


[2] Article publié dans le n°43 de la Revue MODULAD (2011), http://www.modulad.fr/index.htm

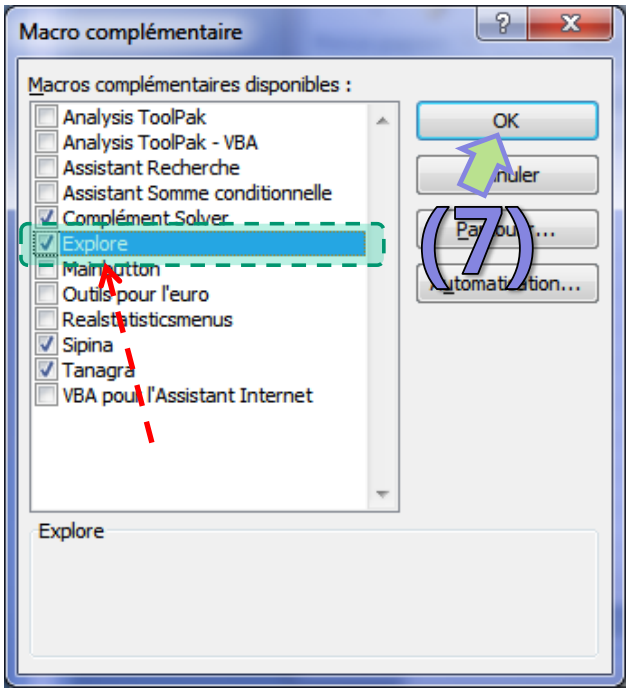
Puis, dans la boîte de paramétrage qui apparaît, nous sélectionnons l'option « **Compléments** » et nous cliquons sur le bouton « **Atteindre** ».



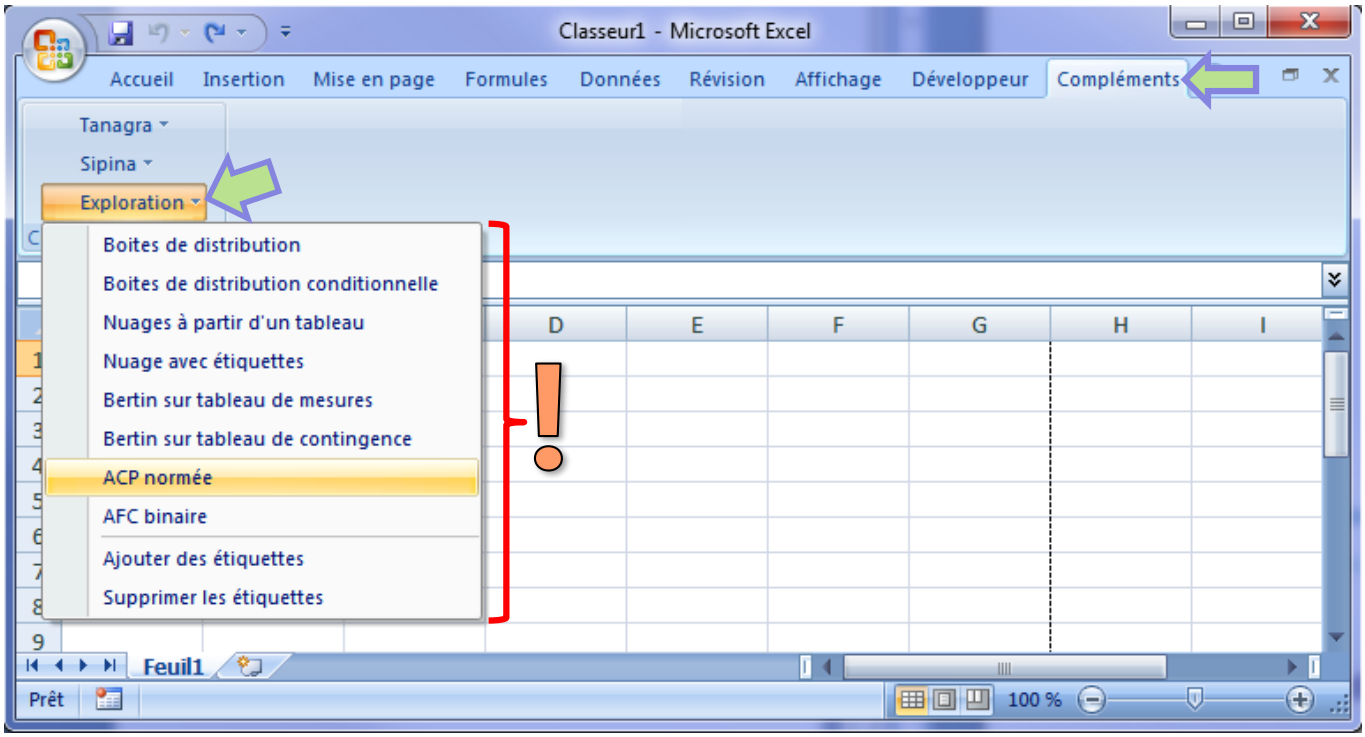
« Explore » n'est pas visible dans la liste des macros disponibles. Nous l'installons en allant piocher (bouton « **Parcourir** ») le fichier « **Explore.xla** » sur notre disque dur.



L'add-in **EXPLORE** apparaît dans la liste maintenant. Nous validons en cliquant sur « OK ».



Le menu « **Exploration** » dédié à l'analyse des données est maintenant visible dans l'onglet « Compléments » d'Excel 2007.



Dans ce qui suit, nous présentons deux méthodes phares de l'analyse factorielle proposées par Explore : l'analyse en composantes principales (ACP) sur un tableau individus-variables, incluant des variables et des individus supplémentaires ; et l'analyse factorielle des correspondances (AFC) sur un tableau de contingence.

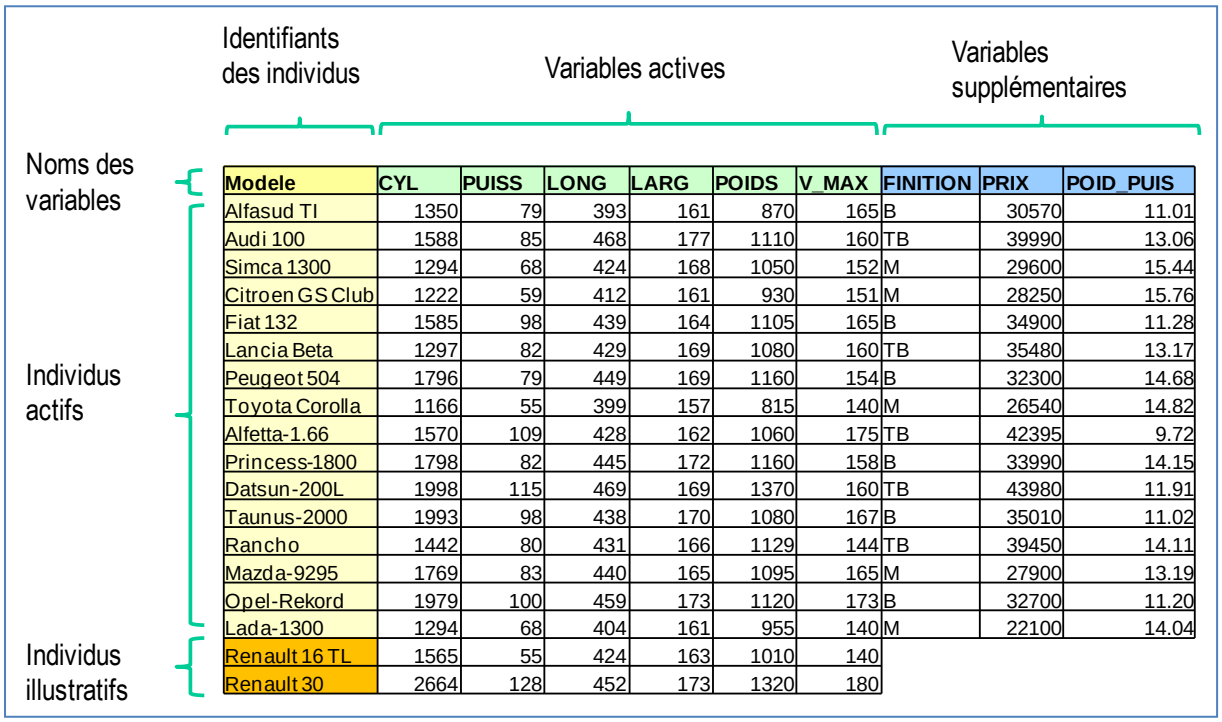
3 Analyse de données avec « Explore »

Nous nous contentons de présenter les sorties et les principaux outils d'Explore dans ce document. Pour le lecteur désireux d'approfondir les aspects méthodologiques, le mieux est de se référer aux très nombreux supports pédagogiques accessibles sur le web[3].

3.1 Analyse en composantes principales

3.1.1 Données

Nous utilisons « **autos_acp.xlsx** »[4] pour découvrir l'analyse en composantes principales (ACP) d'Explore : n = 16 véhicules sont caractérisés par p = 6 variables (cylindrée, puissance, longueur, largeur, poids, vitesse max) ; nous disposons de 3 variables supplémentaires (finition, qualitative ; prix et rapport poids-puissance, quantitatives) et 2 individus illustratifs (Renault 16 TL et Renault 30). La disposition des données doit obéir à des règles particulières.



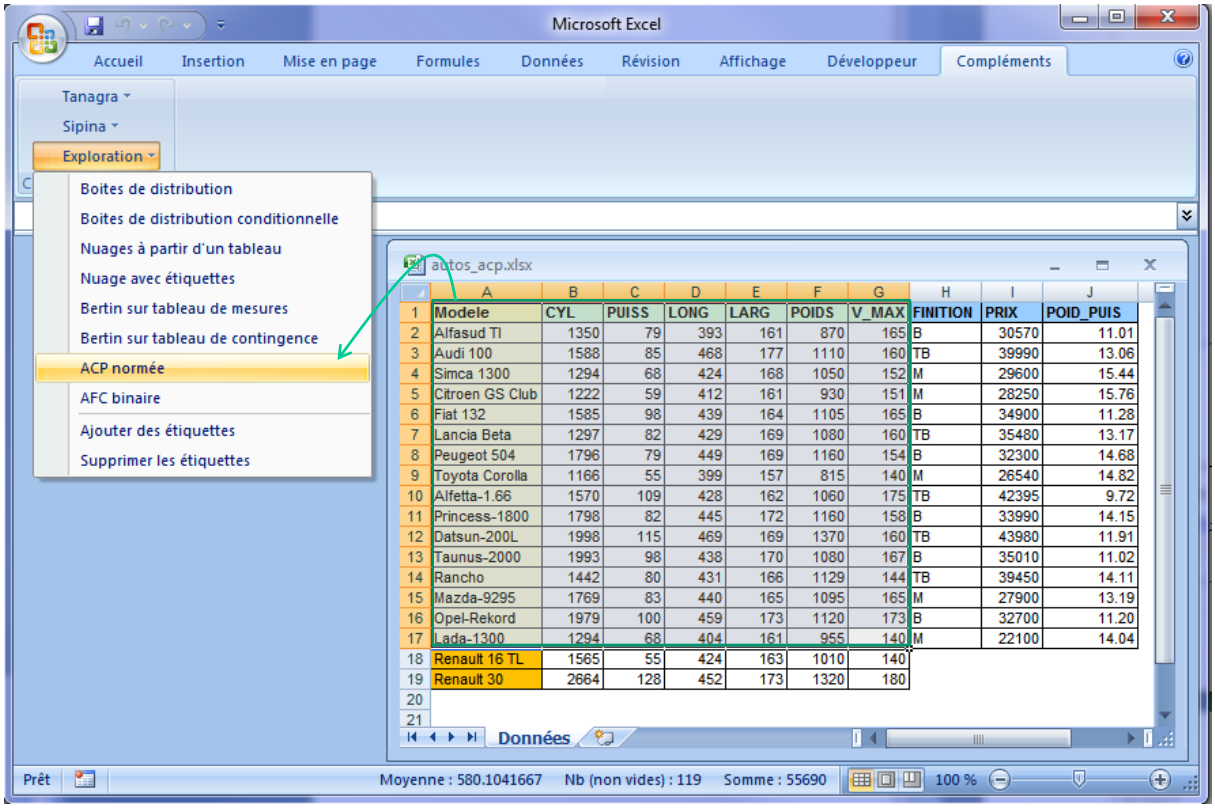
Modele	CYL	PUISS	LONG	LARG	POIDS	V_MAX	FINITION	PRIX	POID_PUIS
Alfasud TI	1350	79	393	161	870	165	B	30570	11.01
Audi 100	1588	85	468	177	1110	160	TB	39990	13.06
Simca 1300	1294	68	424	168	1050	152	M	29600	15.44
Citroen GS Club	1222	59	412	161	930	151	M	28250	15.76
Fiat 132	1585	98	439	164	1105	165	B	34900	11.28
Lancia Beta	1297	82	429	169	1080	160	TB	35480	13.17
Peugeot 504	1796	79	449	169	1160	154	B	32300	14.68
Toyota Corolla	1166	55	399	157	815	140	M	26540	14.82
Alfetta-1.66	1570	109	428	162	1060	175	TB	42395	9.72
Princess-1800	1798	82	445	172	1160	158	B	33990	14.15
Datsun-200L	1998	115	469	169	1370	160	TB	43980	11.91
Taunus-2000	1993	98	438	170	1080	167	B	35010	11.02
Rancho	1442	80	431	166	1129	144	TB	39450	14.11
Mazda-9295	1769	83	440	165	1095	165	M	27900	13.19
Opel-Rekord	1979	100	459	173	1120	173	B	32700	11.20
Lada-1300	1294	68	404	161	955	140	M	22100	14.04
Renault 16 TL	1565	55	424	163	1010	140			
Renault 30	2664	128	452	173	1320	180			

3.1.2 ACP avec Explore

Pour lancer l'analyse, nous devons sélectionner la partie active des données c.-à-d. la plage allant de la cellule A1 à G17 dans notre feuille Excel. Puis nous actionnons l'item « **ACP Normée** » du menu « **Exploration** ».

[3] Entres autres : http://tutoriels-data-mining.blogspot.fr/2013/07/analyse-en-composantes-principales.html pour l'ACP ; http://tutoriels-data-mining.blogspot.fr/2013/07/analyse-factorielle-des-correspondances.html pour l'AFC.

[4] G. Saporta, " Probabilités, Analyse de données et Statistique ", Dunod, 2006 ; tableau 17.1, page 428.



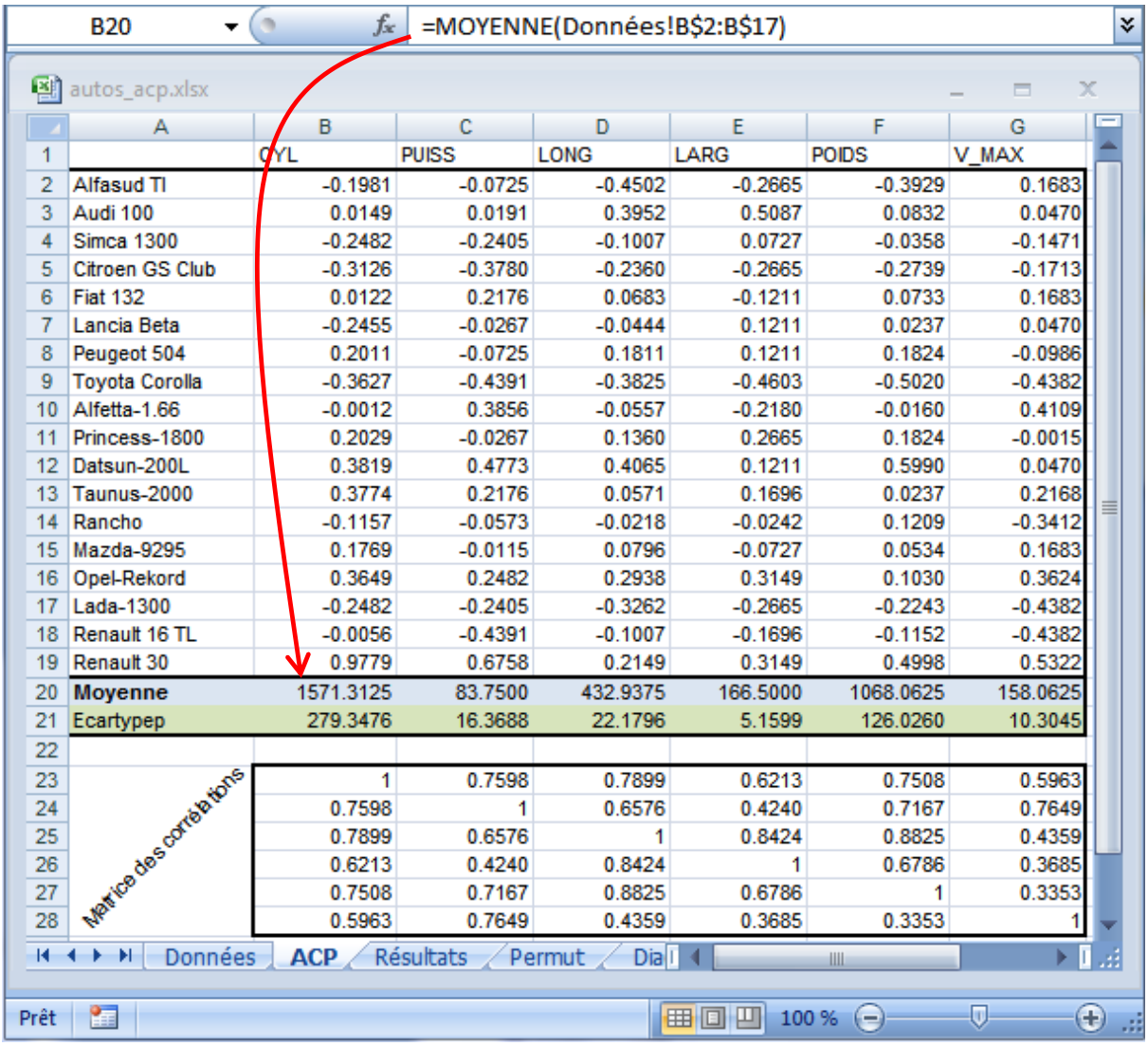
	A	B	C	D	E	F	G	H	I	J
1	Modele	CYL	PUISS	LONG	LARG	POIDS	V_MAX	FINITION	PRIX	POID_PUIS
2	Alfasud TI	1350	79	393	161	870	165	B	30570	11.01
3	Audi 100	1588	85	468	177	1110	160	TB	39990	13.06
4	Simca 1300	1294	68	424	168	1050	152	M	29600	15.44
5	Citroen GS Club	1222	59	412	161	930	151	M	28250	15.76
6	Fiat 132	1585	98	439	164	1105	165	B	34900	11.28
7	Lancia Beta	1297	82	429	169	1080	160	TB	35480	13.17
8	Peugeot 504	1796	79	449	169	1160	154	B	32300	14.68
9	Toyota Corolla	1166	55	399	157	815	140	M	26540	14.82
10	Alfetta-1.66	1570	109	428	162	1060	175	TB	42395	9.72
11	Princess-1800	1798	82	445	172	1160	158	B	33990	14.15
12	Datsun-200L	1998	115	469	169	1370	160	TB	43980	11.91
13	Taunus-2000	1993	98	438	170	1080	167	B	35010	11.02
14	Rancho	1442	80	431	166	1129	144	TB	39450	14.11
15	Mazda-9295	1769	83	440	165	1095	165	M	27900	13.19
16	Opel-Rekord	1979	100	459	173	1120	173	B	32700	11.20
17	Lada-1300	1294	68	404	161	955	140	M	22100	14.04
18	Renault 16 TL	1565	55	424	163	1010	140			
19	Renault 30	2664	128	452	173	1320	180			
20										
21										

L'outil se charge automatique de détecter les éventuels individus et variables supplémentaires. Il génère alors plusieurs feuilles dans notre classeur Excel.

3.1.2.1 La feuille ACP

La feuille ACP contient les données centrées et réduites, ainsi que la matrice des corrélations. Attention, tous les indicateurs statistiques (moyennes, écarts-type, corrélations) sont calculés à partir des individus actifs comme nous pouvons le voir dans la copie d'écran ci-dessous. La moyenne de CYL est calculée sur la plage « B2:B17 » de la feuille « Données », excluant les deux dernières lignes. Mais l'outil calcule également les valeurs centrées et réduites pour les individus illustratifs (les Renault).

Remarque 1 : J'ai rajouté après coup les étiquettes « moyennes », « écart-type » et « matrice des corrélations ». Dans le même ordre d'idée, j'ai aussi effectué une petite mise en forme pour rendre la lecture plus attrayante.

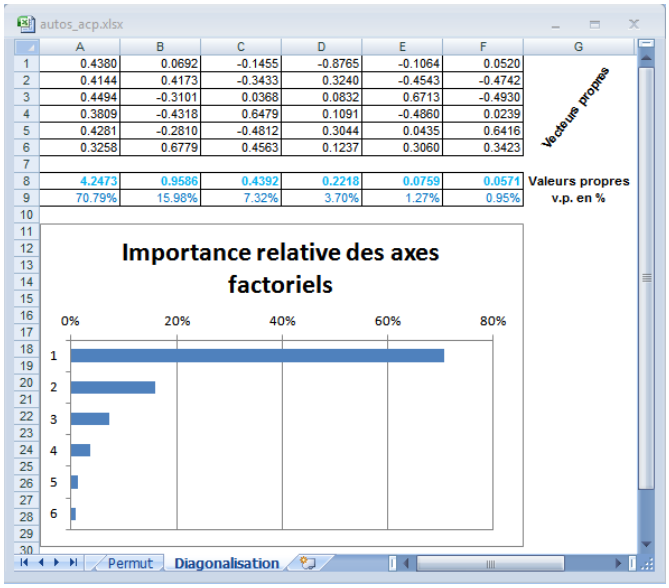


	A	B	C	D	E	F	G
1		CYL	PUISS	LONG	LARG	POIDS	V_MAX
2	Alfasud TI	-0.1981	-0.0725	-0.4502	-0.2665	-0.3929	0.1683
3	Audi 100	0.0149	0.0191	0.3952	0.5087	0.0832	0.0470
4	Simca 1300	-0.2482	-0.2405	-0.1007	0.0727	-0.0358	-0.1471
5	Citroen GS Club	-0.3126	-0.3780	-0.2360	-0.2665	-0.2739	-0.1713
6	Fiat 132	0.0122	0.2176	0.0683	-0.1211	0.0733	0.1683
7	Lancia Beta	-0.2455	-0.0267	-0.0444	0.1211	0.0237	0.0470
8	Peugeot 504	0.2011	-0.0725	0.1811	0.1211	0.1824	-0.0986
9	Toyota Corolla	-0.3627	-0.4391	-0.3825	-0.4603	-0.5020	-0.4382
10	Alfetta-1.66	-0.0012	0.3856	-0.0557	-0.2180	-0.0160	0.4109
11	Princess-1800	0.2029	-0.0267	0.1360	0.2665	0.1824	-0.0015
12	Datsun-200L	0.3819	0.4773	0.4065	0.1211	0.5990	0.0470
13	Taunus-2000	0.3774	0.2176	0.0571	0.1696	0.0237	0.2168
14	Rancho	-0.1157	-0.0573	-0.0218	-0.0242	0.1209	-0.3412
15	Mazda-9295	0.1769	-0.0115	0.0796	-0.0727	0.0534	0.1683
16	Opel-Rekord	0.3649	0.2482	0.2938	0.3149	0.1030	0.3624
17	Lada-1300	-0.2482	-0.2405	-0.3262	-0.2665	-0.2243	-0.4382
18	Renault 16 TL	-0.0056	-0.4391	-0.1007	-0.1696	-0.1152	-0.4382
19	Renault 30	0.9779	0.6758	0.2149	0.3149	0.4998	0.5322
20	**Moyenne**	1571.3125	83.7500	432.9375	166.5000	1068.0625	158.0625
21	Ecartypep	279.3476	16.3688	22.1796	5.1599	126.0260	10.3045
22							
23		1	0.7598	0.7899	0.6213	0.7508	0.5963
24		0.7598	1	0.6576	0.4240	0.7167	0.7649
25		0.7899	0.6576	1	0.8424	0.8825	0.4359
26		0.6213	0.4240	0.8424	1	0.6786	0.3685
27		0.7508	0.7167	0.8825	0.6786	1	0.3353
28		0.5963	0.7649	0.4359	0.3685	0.3353	1

Remarque 2 : Explore divise chaque valeur par $\sqrt{n}$ lors de la transformation, pour pouvoir obtenir directement la matrice des corrélations par simple produit matriciel derrière. Je vois bien l'artifice. En revanche, cette correction ne sera pas sans conséquences sur le calcul des coordonnées factorielles, qui seront également divisées par le même facteur par rapport aux sorties des autres logiciels tels que SPAD, R (avec les packages reconnus), Tanagra, … Est-ce vraiment un problème ? Non en vérité, ce sont les positions relatives des individus qui importent. Si les valeurs sont toutes déflatés de la même manière, cela n'altère en rien la perception fournie par l'analyse.

3.1.2.2 La feuille diagonalisation

Elle contient les valeurs et vecteurs propres, calculées à partir de la matrice des corrélations. Explore propose l'éboulis des valeurs propres sous forme de diagramme en bâtons.



Explore

Axis	Eigen value	Difference	Proportion (%)	Histogram	Cumulative (%)
1	4.247253	3.288634	70.79 %		70.79 %
2	0.958619	0.519386	15.98 %		86.76 %
3	0.439233	0.217401	7.32 %		94.09 %
4	0.221832	0.145904	3.70 %		97.78 %
5	0.075927	0.018790	1.27 %		99.05 %
6	0.057137	-	0.95 %		100.00 %
Tot.	6.000000	-	-	-	-

Tanagra

Les résultats sont complètement cohérents avec ceux de Tanagra (comme avec les autres outils que j'ai testé). Sachant que la procédure, notamment la diagonalisation, a été réalisée en VBA (Visual Basic pour Applications) dans Explore, on se rend compte qu'on peut faire des choses intéressantes en VBA. Bon, soyons lucides, l'affaire sera tout autre lorsqu'il s'agira de traiter des volumétries un peu plus importantes (section 3.1.3).

3.1.2.3 La feuille Résultats

La feuille « Résultats » est très dense. Elle est composée de plusieurs parties.

Informations sur les individus. Pour les 6 facteurs, puisque nous avions p = 6 variables actives, nous disposons des coordonnées factorielles (divisées par $\sqrt{n}$ par rapport aux sorties des autres logiciels), des contributions (en %) et des cos² (qualité de représentation).

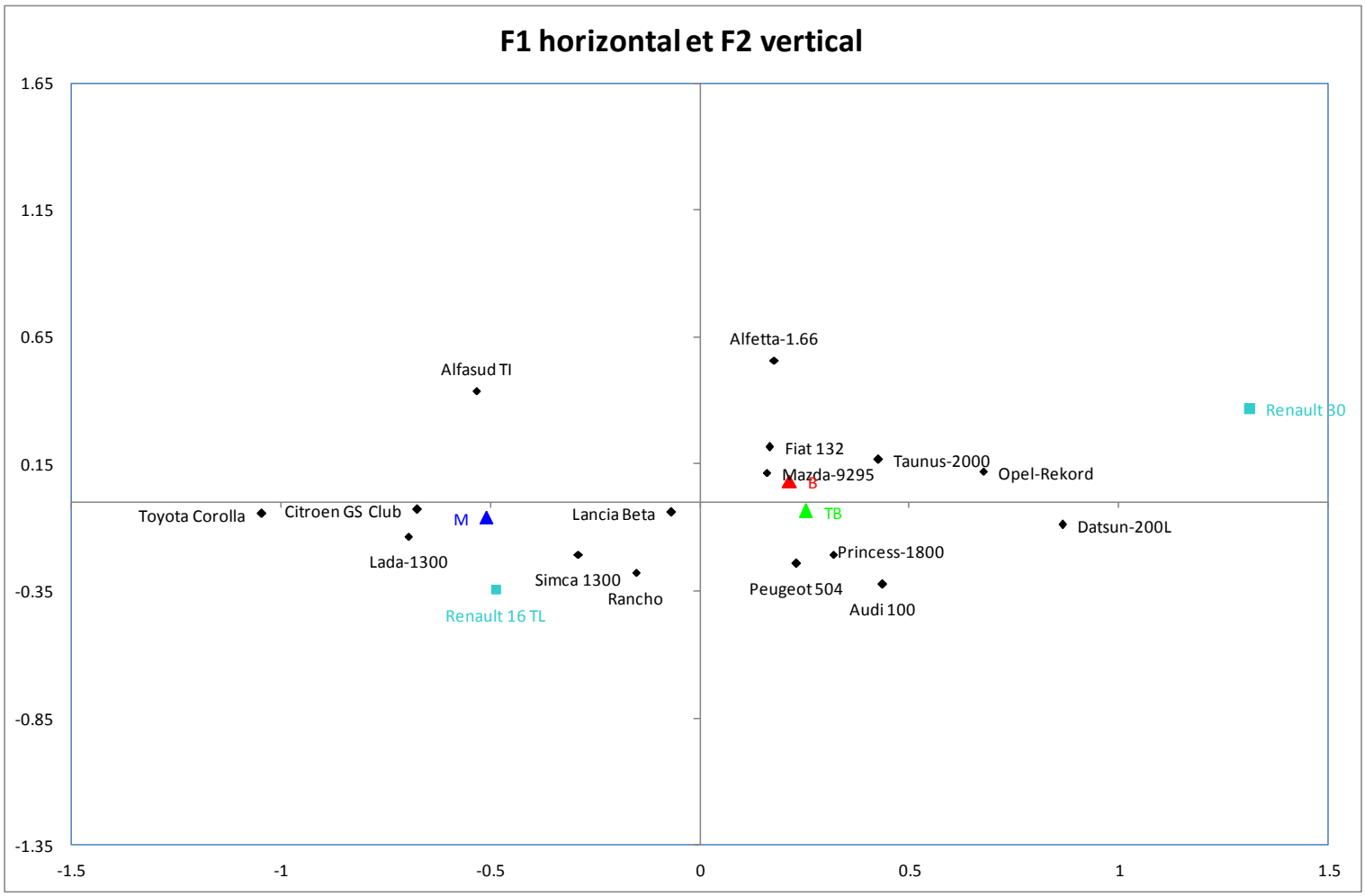
	Coordonnées factorielles des lignes						Contributions						Qualité de représentation					
	F1	F2	F3	F4	F5	F6	CTR1	CTR2	CTR3	CTR4	CTR5	CTR6	COR1	COR2	COR3	COR4	COR5	COR6
Alfasud TI	-0.53	0.44	0.13	-0.02	-0.08	0.05	6.7%	19.8%	3.9%	0.1%	9.3%	3.6%	0.57	0.38	0.03	0.00	0.01	0.00
Audi 100	0.44	-0.32	0.32	0.11	0.03	-0.12	4.5%	11.0%	22.9%	5.7%	0.9%	25.8%	0.45	0.25	0.24	0.03	0.00	0.03
Simca 1300	-0.29	-0.21	0.11	0.11	-0.01	0.08	2.0%	4.5%	2.9%	5.5%	0.3%	11.0%	0.53	0.27	0.08	0.08	0.00	0.04
Citroen GS Club	-0.67	-0.03	0.05	0.00	0.11	0.04	10.7%	0.1%	0.5%	0.0%	16.4%	2.6%	0.96	0.00	0.00	0.00	0.03	0.00
Fiat 132	0.17	0.22	-0.11	0.10	0.06	-0.03	0.7%	4.9%	2.8%	4.1%	4.6%	2.1%	0.28	0.47	0.12	0.09	0.03	0.01
Lancia Beta	-0.07	-0.04	0.13	0.23	-0.03	0.06	0.1%	0.2%	4.0%	23.7%	1.6%	5.5%	0.06	0.02	0.22	0.65	0.02	0.04
Peugeot 504	0.23	-0.24	-0.05	-0.13	0.05	0.04	1.3%	6.2%	0.6%	7.4%	3.6%	3.1%	0.39	0.43	0.02	0.12	0.02	0.01
Toyota Corolla	-1.05	-0.05	-0.07	-0.11	0.05	-0.11	25.7%	0.2%	1.0%	5.8%	3.2%	19.3%	0.97	0.00	0.00	0.01	0.00	0.01
Alfetta-1.66	0.18	0.56	-0.08	0.14	0.02	-0.03	0.7%	32.2%	1.5%	9.3%	0.5%	1.6%	0.09	0.84	0.02	0.06	0.00	0.00
Princess-1800	0.32	-0.21	0.07	-0.09	-0.04	0.08	2.4%	4.5%	1.1%	3.7%	2.1%	10.9%	0.61	0.26	0.03	0.05	0.01	0.04
Datsun-200L	0.87	-0.09	-0.39	0.06	0.00	0.00	17.6%	0.8%	35.1%	1.4%	0.0%	0.0%	0.82	0.01	0.17	0.00	0.00	0.00
Taunus-2000	0.43	0.17	0.07	-0.20	-0.12	-0.02	4.3%	2.9%	1.1%	18.6%	17.6%	0.6%	0.68	0.10	0.02	0.15	0.05	0.00
Rancho	-0.15	-0.28	-0.19	0.07	-0.06	-0.01	0.6%	8.2%	8.6%	2.4%	5.3%	0.1%	0.16	0.53	0.25	0.04	0.03	0.00
Mazda-9295	0.16	0.11	-0.01	-0.12	0.13	0.07	0.6%	1.3%	0.1%	6.8%	21.9%	7.5%	0.34	0.17	0.00	0.20	0.22	0.06
Opel-Rekord	0.68	0.12	0.19	-0.10	0.01	-0.05	10.8%	1.5%	8.4%	4.9%	0.1%	3.7%	0.88	0.03	0.07	0.02	0.00	0.00
Lada-1300	-0.70	-0.14	-0.16	-0.04	-0.10	-0.04	11.4%	1.9%	5.7%	0.7%	12.6%	2.6%	0.90	0.03	0.05	0.00	0.02	0.00
Renault 16 TL	-0.49	-0.34	-0.11	-0.25	0.08	0.03							0.54	0.27	0.03	0.15	0.01	0.00
Renault 30	1.31	0.37	-0.16	-0.37	-0.24	0.13							0.82	0.06	0.01	0.06	0.03	0.01

Les observations illustratives sont insérées dans le tableau. Visiblement, la Renault 16 TL ne se situe pas dans le même segment que la Renault 30, ils sont à l'opposé l'un de l'autre dans le premier plan factoriel (F1, F2).

Informations sur les variables. Ici aussi, nous avons la trilogie : coordonnées (corrélations des variables avec les axes), qualité de représentation (cos² = carré de la corrélation), et contribution (carré de la corrélation divisé par la valeur propre).

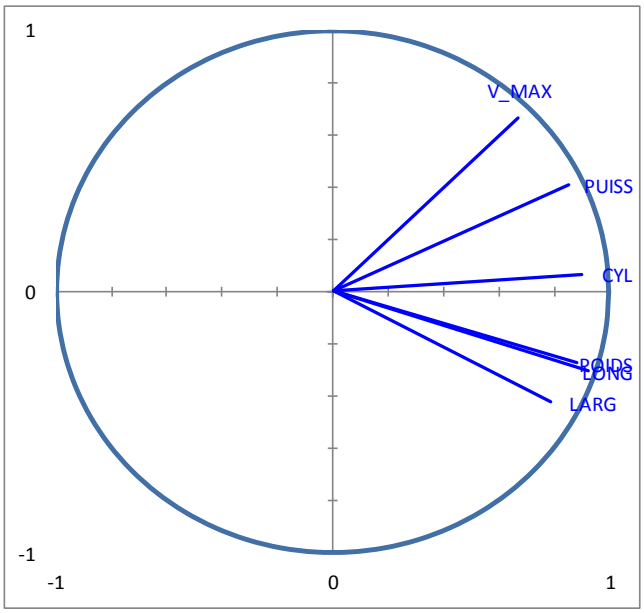
	Coordonnées factorielles des colonnes						Contributions						Qualité de représentation					
	G1	G2	G3	G4	G5	G6	CTR1	CTR2	CTR3	CTR4	CTR5	CTR6	COR1	COR2	COR3	COR4	COR5	COR6
CYL	0.90	0.07	-0.10	-0.41	-0.03	0.01	19%	0%	2%	77%	1%	0%	0.81	0.00	0.01	0.17	0.00	0.00
PUISS	0.85	0.41	-0.23	0.15	-0.13	-0.11	17%	17%	12%	10%	21%	22%	0.73	0.17	0.05	0.02	0.02	0.01
LONG	0.93	-0.30	0.02	0.04	0.18	-0.12	20%	10%	0%	1%	45%	24%	0.86	0.09	0.00	0.00	0.03	0.01
LARG	0.79	-0.42	0.43	0.05	-0.13	0.01	15%	19%	42%	1%	24%	0%	0.62	0.18	0.18	0.00	0.02	0.00
POIDS	0.88	-0.28	-0.32	0.14	0.01	0.15	18%	8%	23%	9%	0%	41%	0.78	0.08	0.10	0.02	0.00	0.02
V_MAX	0.67	0.66	0.30	0.06	0.08	0.08	11%	46%	21%	2%	9%	12%	0.45	0.44	0.09	0.00	0.01	0.01

Carte des individus. Nous pouvons choisir les axes à représenter à l'aide d'une liste déroulante. La carte des individus nécessite un petit nettoyage pour ne visualiser que les individus actifs et illustratifs. Une connaissance élémentaire de la manipulation d'Excel permet de le réaliser facilement.



Il est possible d'y faire figurer les informations concernant les variables supplémentaires : comme ici, les moyennes conditionnelles (M, B, TB de FINITION) lorsqu'elles sont qualitatives ; ou un dégradé de couleurs indexé sur les valeurs lorsqu'elles sont quantitatives.

Carte des variables. Il s'agit du cercle des corrélations.



Graphique de Bertin. Il propose une autre vision du tableau de données en réordonnant les individus et les variables en fonction de leur position sur un des axes factoriels, que nous pouvons choisir. Voici la représentation obtenue pour le 1er axe.

Choix de l'axe de ré-ordonnancement

Seuils sur valeurs centrées et réduites pour la colorisation

Ordonner suivant l'axe : 1 Bornes en écart-types pour les couleu 1 -0.9 0.9

	LONG	CYL	POIDS	PUISS	LARG	V_MAX	LONG	CYL	POIDS	PUISS	LARG	V_MAX
Datsun-200L	469	1998	1370	115	169	160	1.63	1.53	2.40	1.91	0.48	0.19
Opel-Rekord	459	1979	1120	100	173	173	1.18	1.46	0.41	0.99	1.26	1.45
Audi 100	468	1588	1110	85	177	160	1.58	0.06	0.33	0.08	2.03	0.19
Taunus-2000	438	1993	1080	98	170	167	0.23	1.51	0.09	0.87	0.68	0.87
Princess-1800	445	1798	1160	82	172	158	0.54	0.81	0.73	-0.11	1.07	-0.01
Peugeot 504	449	1796	1160	79	169	154	0.72	0.80	0.73	-0.29	0.48	-0.39
Alfetta-1.66	428	1570	1060	109	162	175	-0.22	0.00	-0.06	1.54	-0.87	1.64
Fiat 132	439	1585	1105	98	164	165	0.27	0.05	0.29	0.87	-0.48	0.67
Mazda-9295	440	1769	1095	83	165	165	0.32	0.71	0.21	-0.05	-0.29	0.67
Lancia Beta	429	1297	1080	82	169	160	-0.18	-0.98	0.09	-0.11	0.48	0.19
Rancho	431	1442	1129	80	166	144	-0.09	-0.46	0.48	-0.23	-0.10	-1.36
Simca 1300	424	1294	1050	68	168	152	-0.40	-0.99	-0.14	-0.96	0.29	-0.59
Alfasud TI	393	1350	870	79	161	165	-1.80	-0.79	-1.57	-0.29	-1.07	0.67
Citroen GS Club	412	1222	930	59	161	151	-0.94	-1.25	-1.10	-1.51	-1.07	-0.69
Lada-1300	404	1294	955	68	161	140	-1.30	-0.99	-0.90	-0.96	-1.07	-1.75
Toyota Corolla	399	1166	815	55	157	140	-1.53	-1.45	-2.01	-1.76	-1.84	-1.75

Tableau réorganisé suivant les positions sur l'axe F1

Tableau de données initiales (couleurs pour identifier les valeurs « significativement » faibles [bleu] ou élevées [rouge])

Tableau de données centrées et réduites

Comment lire ce tableau ?

- LONG est la variable la plus corrélée avec le premier facteur, puis CYL, puis POIDS, etc. (voir « Informations sur les variables » ou « Carte des variables »).
- De la même manière, Datsun 200L et Toyota Corolla sont aux deux extrémités des données sur le 1er axe (cf. « Informations sur les individus » ou « Carte des individus »).
- Ce double tri a pour objectif d'identifier les individus et les variables qui caractérisent le mieux le 1er facteur.
- Nous avons fixés les seuils (-0.9 ; +0.9). Les valeurs centrées réduites en-deçà de -0.09 seront considérées comme respectivement significativement faibles, coloriées en bleu dans le tableau des données initiales ; au-delà de 0.9, elles sont considérées élevées et coloriées en rouge.
- L'auteur qualifie les seuils de « bornes en écarts-type » parce que nous pouvons également les interpréter comme des écarts à la moyenne exprimés en écarts-type.

Dans notre exemple, l'opposition marquée entre (Datsun-200L, Opel Rekord, Audi 100) et (Alfasud TI, Citroën GS, Lada-1300, Toyota Corolla) repose sur quasiment l'ensemble des variables (un peu moins concernant VMAX, parce que justement c'est celle qui est la moins corrélée avec le 1er facteur). Nous avons un « effet taille » très pesant dans cette étude.

Sur le 2nd axe, nous obtenons le graphique de Bertin suivant :

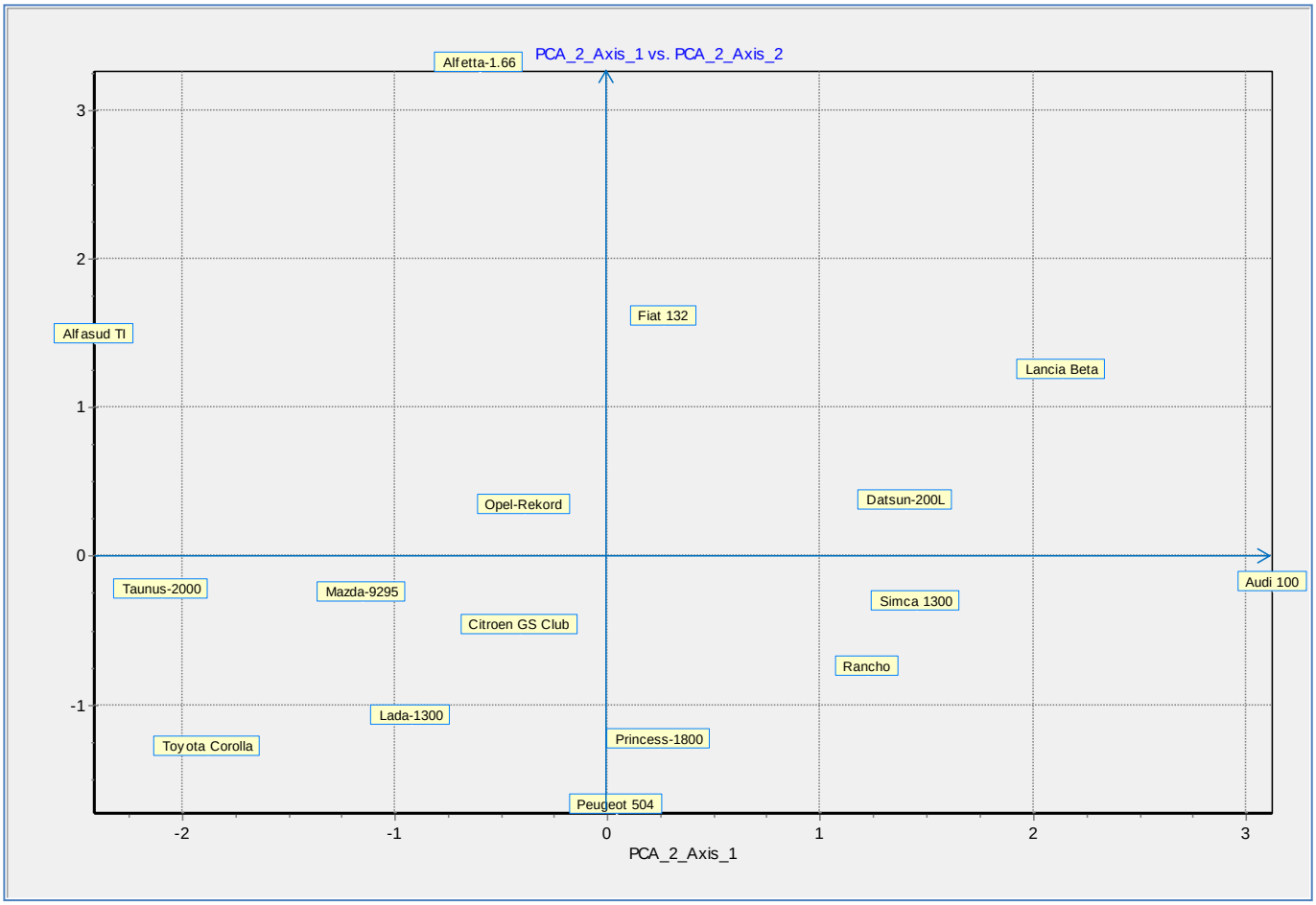
Ordonner suivant l'axe : 2 — Bornes en écart-types pour les couleu 2 — -0.9 — 0.9

	V_MAX	PUISS	CYL	POIDS	LONG	LARG	V_MA	PUISS	CYL	POIDS	LONG	LARG
Alfetta-1.66	**175**	**109**	1570	1060	428	162	1.64	1.54	0.00	-0.06	-0.22	-0.87
Alfasud TI	165	79	1350	*870*	*393*	*161*	0.67	-0.29	-0.79	-1.57	-1.80	-1.07
Fiat 132	165	98	1585	1105	439	164	0.67	0.87	0.05	0.29	0.27	-0.48
Taunus-2000	167	98	**1993**	1080	438	170	0.87	0.87	1.51	0.09	0.23	0.68
Opel-Rekord	**173**	**100**	**1979**	1120	**459**	**173**	1.45	0.99	1.46	0.41	1.18	1.26
Mazda-9295	165	83	1769	1095	440	165	0.67	-0.05	0.71	0.21	0.32	-0.29
Citroen GS Club	151	*59*	*1222*	*930*	*412*	*161*	-0.69	-1.51	-1.25	-1.10	-0.94	-1.07
Lancia Beta	160	82	*1297*	1080	429	169	0.19	-0.11	-0.98	0.09	-0.18	0.48
Toyota Corolla	*140*	*55*	*1166*	*815*	*399*	*157*	-1.75	-1.76	-1.45	-2.01	-1.53	-1.84
Datsun-200L	160	**115**	**1998**	**1370**	**469**	169	0.19	1.91	1.53	2.40	1.63	0.48
Lada-1300	*140*	*68*	*1294*	955	*404*	*161*	-1.75	-0.96	-0.99	-0.90	-1.30	-1.07
Princess-1800	158	82	1798	1160	445	**172**	-0.01	-0.11	0.81	0.73	0.54	1.07
Simca 1300	152	*68*	*1294*	1050	424	168	-0.59	-0.96	-0.99	-0.14	-0.40	0.29
Peugeot 504	154	79	1796	1160	449	169	-0.39	-0.29	0.80	0.73	0.72	0.48
Rancho	*144*	80	1442	1129	431	166	-1.36	-0.23	-0.46	0.48	-0.09	-0.10
Audi 100	160	85	1588	1110	**468**	**177**	0.19	0.08	0.06	0.33	1.58	2.03

Les voitures italiennes se démarquent (Alfetta, Alfasud, Fiat 132).

Remarque – ACP sur les corrélations partielles. Dans ce genre de configuration où une information « écrase » toutes les autres, il peut être intéressant de travailler sur les

corrélations partielles[5]. Ainsi, si l'on retranche l'impact de la cylindrée (CYL), nous obtiendrons la carte des individus suivante :



Sur le 1er facteur, nous distinguons l'opposition entre les automobiles plus ou moins volumineuses (largeur, longues, lourdes) *relativement à leur cylindrée* ; sur le 2nd, les automobiles plus ou moins performantes (vitesse max, puissance), toujours *relativement à leur cylindrée*. Ce type d'analyse amène une autre vision des structures sous-jacentes aux données.

Mais revenons à la macro complémentaire Explore.

Variables supplémentaires qualitatives. L'outil positionne les modalités des variables illustratives qualitatives en calculant les moyennes conditionnelles.

FINITION	F1	F2	F3	F4	F5	F6
B	0.21	0.08	0.05	-0.07	-0.02	0.01
TB	0.25	-0.04	-0.04	0.12	-0.01	-0.02
M	-0.51	-0.06	-0.02	-0.03	0.04	0.01

[5] Tutoriel Tanagra, « Travailler sur les corrélations partielles », 03/2008 ; « ACP sur corrélations partielles (suite) », 06/2012.

Nous observons surtout une opposition entre les véhicules de FINITION moyenne (**M**oyenne) et les véhicules luxueux (Finition = **B**ien ou **T**rès **B**ien) sur le 1er axe.

Variables supplémentaires quantitatives. Pour les variables supplémentaires quantitatives, le tableau des valeurs est accolé aux coordonnées factorielles. Il est aisé de calculer les corrélations avec les axes. Voyons ce qu'il en est pour PRIX.

	PRIX	F1	F2	F3	F4	F5	F6
Alfasud TI	30570	-0.53	0.44	0.13	-0.02	-0.08	0.05
Audi 100	39990	0.44	-0.32	0.32	0.11	0.03	-0.12
Simca 1300	29600	-0.29	-0.21	0.11	0.11	-0.01	0.08
Citroen GS Club	28250	-0.67	-0.03	0.05	0.00	0.11	0.04
Fiat 132	34900	0.17	0.22	-0.11	0.10	0.06	-0.03
Lancia Beta	35480	-0.07	-0.04	0.13	0.23	-0.03	0.06
Peugeot 504	32300	0.23	-0.24	-0.05	-0.13	0.05	0.04
Toyota Corolla	26540	-1.05	-0.05	-0.07	-0.11	0.05	-0.11
Alfetta-1.66	42395	0.18	0.56	-0.08	0.14	0.02	-0.03
Princess-1800	33990	0.32	-0.21	0.07	-0.09	-0.04	0.08
Datsun-200L	43980	0.87	-0.09	-0.39	0.06	0.00	0.00
Taunus-2000	35010	0.43	0.17	0.07	-0.20	-0.12	-0.02
Rancho	39450	-0.15	-0.28	-0.19	0.07	-0.06	-0.01
Mazda-9295	27900	0.16	0.11	-0.01	-0.12	0.13	0.07
Opel-Rekord	32700	0.68	0.12	0.19	-0.10	0.01	-0.05
Lada-1300	22100	-0.70	-0.14	-0.16	-0.04	-0.10	-0.04
Corrélation		**0.64**	**0.00**	**-0.34**	**0.49**	**-0.34**	**-0.29**

Les véhicules imposants (cylindrée, longueur, poids… ; axe 1) ont tendance à être plus coûteux que les autres. Ce n'est pas une surprise.

3.1.3 Gestion de la volumétrie

Un programme enchaînant des calculs aussi complexes développé en VBA (langage interprété) et accédant à des cellules Excel (faiblement typées) peut laisser sceptique quant à sa capacité à traiter de grands volumes de données. Pour en avoir le cœur net, j'ai lancé l'ACP sur un fichier de n = 5000 observations et p = 61 variables (**wavenoisy.xlsx**).

Le programme s'est interrompu au bout de 1 mn 45 sec (Processeur Intel Q9400, Windows 7 - 64 bits, Excel 2007 - 32 bits). Sans surprise, la diagonalisation de la matrice des corrélations est la partie la plus gourmande. Le calcul des coordonnées factorielles des individus et des variables ont été menées à leur terme. C'est au niveau de l'élaboration du graphique de Bertin qu'un problème est apparu. Un débordement de capacité est annoncé.

Ca ne me paraît pas irrémédiable compte tenu du potentiel d'Excel 2007 (1.048.576 lignes). Cette expérience dessine surtout le cadre dans lequel Explore s'exprime le mieux. Il s'agit avant tout d'un outil pédagogique, facile d'accès, simple à mettre en œuvre, permettant d'obtenir des résultats rapidement sur des jeux de données de taille modérée. Bref, il aurait tout à fait sa place dans des enseignements par exemple. Il n'est pas très adapté par contre

pour le traitement des données volumineuses. Il faut en être conscient simplement, et savoir choisir son outil en fonction du contexte dans lequel on se situe.

3.2 Analyse factorielle des correspondances (AFC)

3.2.1 Données

Le fichier « **job_edu_afc.xlsx** » croise le niveau d'études (éducation, K = 16 modalités, en ligne) atteint par de n = 48.842 individus avec le type d'emploi (occupation, L = 14 modalités, en colonne) qu'ils occupent[6]. Attention : (1) l'outil n'accepte pas les cellules vides, il faut spécifier explicitement la valeur 0 lorsqu'un effectif est nul ; (2) les marges ne doivent pas figurer dans les données. Voici le tableau de contingence[7] :

EDU_JOB	Adm_clerical	Armed_Forces	Craft_repair	Exec_managerial	Farming_fishing	Handlers_cleaners	Machine_op_inspct	Other_service	Priv_house_serv	Prof_specialty	Protective_serv	Sales
10th	59	0	239	42	71	108	152	280	8	164	12	120
11th	100	0	270	51	67	177	153	368	18	215	18	232
12th	52	1	92	18	29	55	61	129	8	71	11	70
1st_4th	6	0	28	6	33	26	36	55	14	22	1	8
5th_6th	8	0	71	6	52	59	95	98	20	43	1	17
7th_8th	20	0	172	28	106	66	129	149	17	124	11	40
9th	20	0	144	23	44	72	102	142	16	73	9	47
Assoc_acdm	281	0	167	240	25	34	51	110	3	279	50	209
Assoc_voc	269	1	375	234	85	43	95	160	5	328	67	163
Bachelors	765	1	332	2025	113	79	99	259	12	2486	147	1268
Doctorate	6	0	4	84	1	0	1	2	1	468	1	16
HS_grad	2047	5	2911	1192	573	943	1531	1936	91	1156	326	1580
Masters	105	2	34	779	14	5	12	35	1	1369	20	206
Preschool	3	0	6	1	17	5	12	22	2	11	0	2
Prof_school	12	1	9	69	7	0	1	7	0	691	1	23
Some_college	1858	4	1258	1288	253	400	492	1171	26	1481	308	1503

Ce tableau est suffisamment grand pour qu'une inspection manuelle des profils soit difficile. Le recours à l'AFC est justifié. Autre particularité importante, si les modalités du niveau d'éducation sont clairement ordonnées, cela est moins sûr concernant le type d'emploi, même si on peut imaginer qu'il faut une certaine qualification pour accéder à certains postes. L'AFC permettra – peut-être – de suggérer un ré-ordonnancement des lignes et de colonnes de manière à faire apparaître la nature de la relation entre ces deux variables.

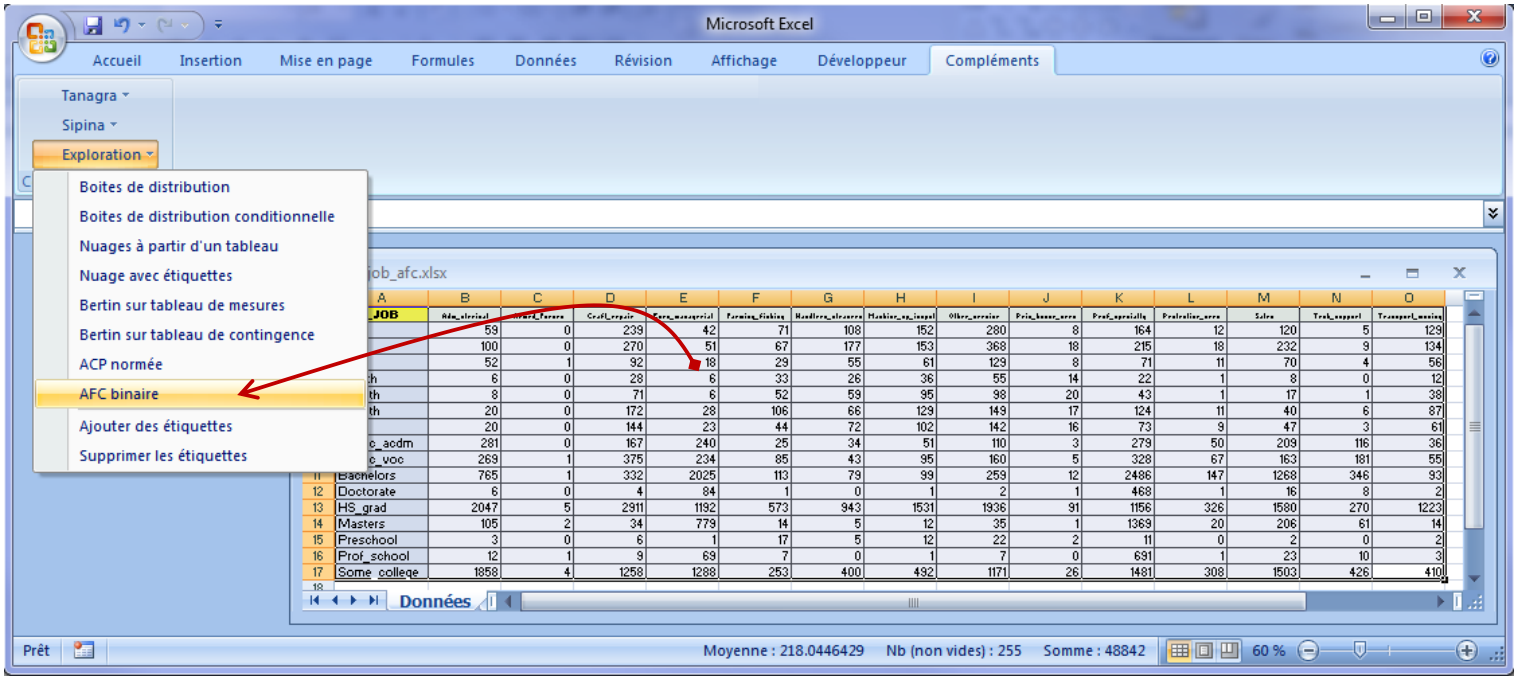
Enfin, nous n'exploitons pas cette possibilité ici mais, tout comme l'ACP, l'AFC sait traiter des lignes et des colonnes supplémentaires accolées au tableau de contingence.

3.2.2 AFC avec Explore

Nous sélectionnons la plage des données en incluant les en-têtes de lignes et de colonnes, puis nous actionnons le menu **Exploration / AFC Binaire**.

[6] Il s'agit du croisement des variables « education » et « occupation » de la base « Adult Data Set » du dépôt UCI : https://archive.ics.uci.edu/ml/datasets/Adult

[7] L'auteur conseille d'avoir plus de lignes que de colonnes dans le tableau de contingence, et de le transposer au préalable si cette condition n'est pas respectée.



Plusieurs feuilles de résultats sont générées.

3.2.3 La feuille AFC

La feuille AFC contient les données préparatoires.

	Adm_cle	Armed_I	Craft_re	Exec_ma	Farming	Handler	Machine	Other_s	Priv_hou	Prof_spe	Protecti	Sales	Tech_su	Transpor	**Marge**
10th	-0.0360	-0.0030	0.0224	-0.0451	0.0199	0.0289	0.0322	0.0535	0.0019	-0.0259	-0.0137	-0.0132	-0.0255	0.0343	1389
11th	-0.0339	-0.0034	0.0130	-0.0526	0.0071	0.0517	0.0175	0.0621	0.0136	-0.0293	-0.0138	0.0088	-0.0276	0.0226	1812
12th	-0.0122	0.0080	0.0049	-0.0319	0.0091	0.0233	0.0144	0.0349	0.0119	-0.0205	-0.0028	-0.0021	-0.0159	0.0196	657
1st_4th	-0.0190	-0.0012	-0.0024	-0.0202	0.0420	0.0217	0.0240	0.0273	0.0523	-0.0157	-0.0081	-0.0170	-0.0122	0.0001	247
5th_6th	-0.0299	-0.0018	0.0041	-0.0326	0.0419	0.0364	0.0512	0.0295	0.0498	-0.0237	-0.0131	-0.0241	-0.0164	0.0123	509
7th_8th	-0.0388	-0.0025	0.0217	-0.0377	0.					-0.0176	-0.0085	-0.0295	-0.0190	0.0273	955
9th	-0.0325	-0.0022	0.0230	-0.0332	0.					-0.0253	-0.0072	-0.0187	-0.0185	0.0184	756
Assoc_acdm	0.0324	-0.0032	-0.0107	0.0130	-0.					-0.0041	0.0142	0.0096	0.0451	-0.0212	1601
Assoc_voc	0.0095	0.0021	0.0330	-0.0064	0.					-0.0118	0.0179	-0.0206	0.0695	-0.0201	2061
Bachelors	-0.0234	-0.0042	-0.0960	0.1467	-0.				0.1190	-0.0052	0.0547	0.0318	-0.0676	8025	
Doctorate	-0.0341	-0.0019	-0.0369	0.0053	-0.0182	-0.0227	-0.0267	-0.0338	-0.0051	0.1553	-0.0143	-0.0282	-0.0103	-0.0225	594
HS_grad	0.0248	0.0003	0.0953	-0.0791	0.0189	0.0478	0.0803	0.0391	0.0065	-0.1467	0.0021	-0.0213	-0.0413	0.0758	15784
Masters	-0.0519	0.0059	-0.0741	0.1114	-0.0337	-0.0459	-0.0538	-0.0644	-0.0152	0.1802	-0.0207	-0.0244	-0.0090	-0.0456	2657
Preschool	-0.0096	-0.0007	-0.0062	-0.0131	0.0411	0.0036	0.0137	0.0213	0.0112	-0.0049	-0.0058	-0.0109	-0.0071	-0.0045	83
Prof_school	-0.0387	0.0067	-0.0422	-0.0155	-0.0165	-0.0269	-0.0319	-0.0380	-0.0092	0.1964	-0.0174	-0.0331	-0.0134	-0.0266	834
Some_college	0.0779	0.0016	-0.0127	-0.0083	-0.0196	-0.0129	-0.0316	0.0102	-0.0172	-0.0525	0.0272	0.0358	0.0262	-0.0226	10878
Marge	5611	15	6112	6086	1490	2072	3022	4923	242	8981	983	5504	1446	2355	48842

$$R = (r_{kl} = \frac{n_{kl} - e_{kl}}{\sqrt{n \times e_{kl}}})$$

0.1149	0.0003	0.1251	0.1246	0.0305	0.0424	0.0619	0.1008	0.0050	0.1839	0.0201	0.1127	0.0296	0.0482
0.0202	0.0000	0.0074	-0.0032	-0.0042	-0.0013	-0.0013	0.0009	-0.0049	-0.0283	0.0073	0.0077	0.0079	0.0008
0.0000	0.0002	-0.0003	0.0001	-0.0004	-0.0004	-0.0005	-0.0005	-0.0003	0.0016	0.0000	-0.0004	0.0000	-0.0003
0.0074	-0.0003	0.0300	-0.0331	0.0124	0.0184	0.0247	0.0227	0.0058	-0.0545	0.0027	-0.0055	-0.0057	0.0211
-0.0032	0.0001	-0.0331	0.0507	-0.0182	-0.0262	-0.0317	-0.0335	-0.0109	0.0534	-0.0005	0.0105	0.0123	-0.0260
-0.0042	-0.0004	0.0124	-0.0182	0.0146	0.0116	0.0162	0.0147	0.0092	-0.0229	-0.0013	-0.0064	-0.0055	0.0104
-0.0013	-0.0004	0.0184	-0.0262	0.0116	0.0179	0.0206	0.0216	0.0086	-0.0358	-0.0011	-0.0045	-0.0100	0.0165
-0.0013	-0.0005	0.0247	-0.0317	0.0162	0.020			0.0107	-0.0445	-0.0010	-0.0081	-0.0116	0.0210
0.0009	-0.0005	0.0227	-0.0335	0.0147	0.021			0.0097	-0.0472	-0.0001	-0.0043	-0.0103	0.0195
-0.0049	-0.0003	0.0058	-0.0109	0.0092	0.0086	0.0107	0.0097	0.0083	-0.0116	-0.0019	-0.0044	-0.0051	0.0060
-0.0283	0.0016	-0.0545	0.0534	-0.0229	-0.0358	-0.0445	-0.0472	-0.0116	0.1375	-0.0104	-0.0053	0.0048	-0.0378
0.0073	0.0000	0.0027	-0.0005	-0.0013	-0.0011	-0.0010	-0.0001	-0.0019	-0.0104	0.0030	0.0029	0.0043	-0.0004
0.0077	-0.0004	-0.0055	0.0105	-0.0064	-0.0045	-0.0081	-0.0043	-0.0044	-0.0053	0.0029	0.0102	0.0052	-0.0050
0.0079	0.0000	-0.0057	0.0123	-0.0055	-0.0100	-0.0116	-0.0103	-0.0051	0.0048	0.0043	0.0052	0.0135	-0.0101
0.0008	-0.0003	0.0211	-0.0260	0.0104	0.0165	0.0210	0.0195	0.0060	-0.0378	-0.0004	-0.0050	-0.0101	0.0183

$$C = R^T R$$

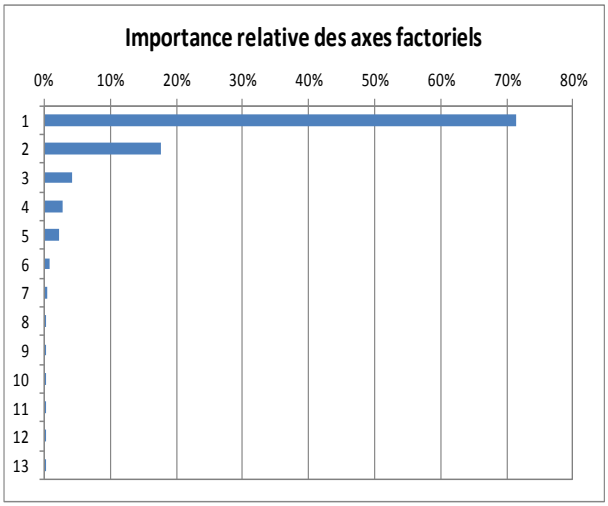
Ayant moi-même programmé l'AFC dans Tanagra[8], j'étais très curieux de voir comment procédait l'auteur d'Explore :

- Explore calcule tout d'abord la quantité (r_{kl}) qui est égale au résidu standardisé du tableau de contingence[9], divisé par la quantité $\sqrt{n}$ (n_{kl} sont les effectifs observés, e_{kl} correspondent aux effectifs théoriques sous l'hypothèse d'indépendance) ;
- Puis il calcule la matrice $C = R^T \times R$ de dimension (L x L) ;
- Qui est soumise à l'algorithme de diagonalisation.

On comprend pourquoi la condition ($L \leq K$) est souhaitée. La matrice à diagonaliser sera plus petite, et on sait de toute manière que le nombre de valeurs propres non nulles est au maximum égal à $H_{max} = (L - 1)$, soit $H_{max} = 13$ pour notre jeu de données.

3.2.4 La feuille diagonalisation

La feuille diagonalisation contient les valeurs et vecteurs propres issus du traitement de la matrice C. Un « scree plot » sous forme de diagramme en bâtons est proposé. Les résultats sont cohérents avec ceux des différents logiciels que nous avons testés.



Explore

Matrix trace = 0.3801
SQRT(Matrix trace) = 0.6165

Axis	Eigen value	% explained	Histogram	% cumulated
1	0.271563	71.44%		71.44%
2	0.067033	17.64%		89.08%
3	0.015937	4.19%		93.27%
4	0.010504	2.76%		96.04%
5	0.008309	2.19%		98.22%
6	0.003349	0.88%		99.10%
7	0.001658	0.44%		99.54%
8	0.001114	0.29%		99.83%
9	0.000271	0.07%		99.90%
10	0.000174	0.05%		99.95%
11	0.000096	0.03%		99.97%
12	0.000065	0.02%		99.99%
13	0.000033	0.01%		100.00%
Tot.	0.380107	-	-	-

Tanagra

89.08% de l'information disponible est resituée par les deux premiers axes. Nous devrions obtenir une représentation assez fidèle de la réalité dans le premier plan factoriel.

3.2.4.1 La feuille Résultats

Elle contient les éléments et les outils qui nous permettent d'appréhender les résultats de l'AFC. La feuille est particulièrement touffue, d'autant plus que nous avons un nombre élevé

[8] Voir entres autres, Tutoriel Tanagra, « Analyse des correspondances – Comparaisons », décembre 2012.

[9] R. Rakotomalala, « Etudes des dépendances – Variables qualitatives », mars 2011 ; section 2.3.2, page 12.

de facteurs. Pour simplifier la lecture, nous n'avons conservé que les 3 premiers axes pour les tableaux des modalités lignes et colonnes.

Coordonnées factorielles des lignes. Ce tableau recense les coordonnées des modalités lignes, leurs contributions aux axes, et la qualité de leur représentation (cos²).

	oordonnées factorielles des ligne			Contributions			Qualité de représentation		
	F1	F2	F3	CTR1	CTR2	CTR3	COR1	COR2	COR3
10th	0.4950	0.3642	-0.0359	2.57%	5.63%	0.23%	0.5885	0.3186	0.0031
11th	0.4412	0.2846	-0.0638	2.66%	4.48%	0.95%	0.5324	0.2216	0.0111
12th	0.4679	0.2622	-0.0363	1.08%	1.38%	0.11%	0.6521	0.2048	0.0039
1st_4th	0.6838	0.6931	-0.4590	0.87%	3.62%	6.69%	0.3027	0.3110	0.1364
5th_6th	0.7315	0.6723	-0.3266	2.05%	7.03%	6.98%	0.4369	0.3691	0.0871
7th_8th	0.5302	0.5341	-0.1370	2.02%	8.32%	2.30%	0.4170	0.4233	0.0278
9th	0.5933	0.4508	-0.1296	2.01%	4.69%	1.63%	0.5961	0.3441	0.0284
Assoc_acdm	-0.1703	-0.3287	0.1014	0.35%	5.28%	2.11%	0.1699	0.6330	0.0602
Assoc_voc	0.0002	-0.1854	0.1747	0.00%	2.16%	8.08%	0.0000	0.1724	0.1530
Bachelors	-0.6352	-0.1503	-0.1533	24.41%	5.54%	24.23%	0.8935	0.0500	0.0520
Doctorate	-1.3490	0.7763	0.3498	8.15%	10.93%	9.34%	0.7133	0.2362	0.0479
HS_grad	0.4036	0.0061	0.0289	19.38%	0.02%	1.69%	0.9395	0.0002	0.0048
Masters	-1.0688	0.2244	-0.1038	22.88%	4.09%	3.68%	0.9302	0.0410	0.0088
Preschool	0.5796	0.7140	-0.3727	0.21%	1.29%	1.48%	0.1888	0.2865	0.0781
Prof_school	-1.3366	0.8758	0.4894	11.23%	19.54%	25.66%	0.6356	0.2729	0.0852
Some_college	0.0383	-0.2194	0.0589	0.12%	15.99%	4.85%	0.0230	0.7535	0.0544

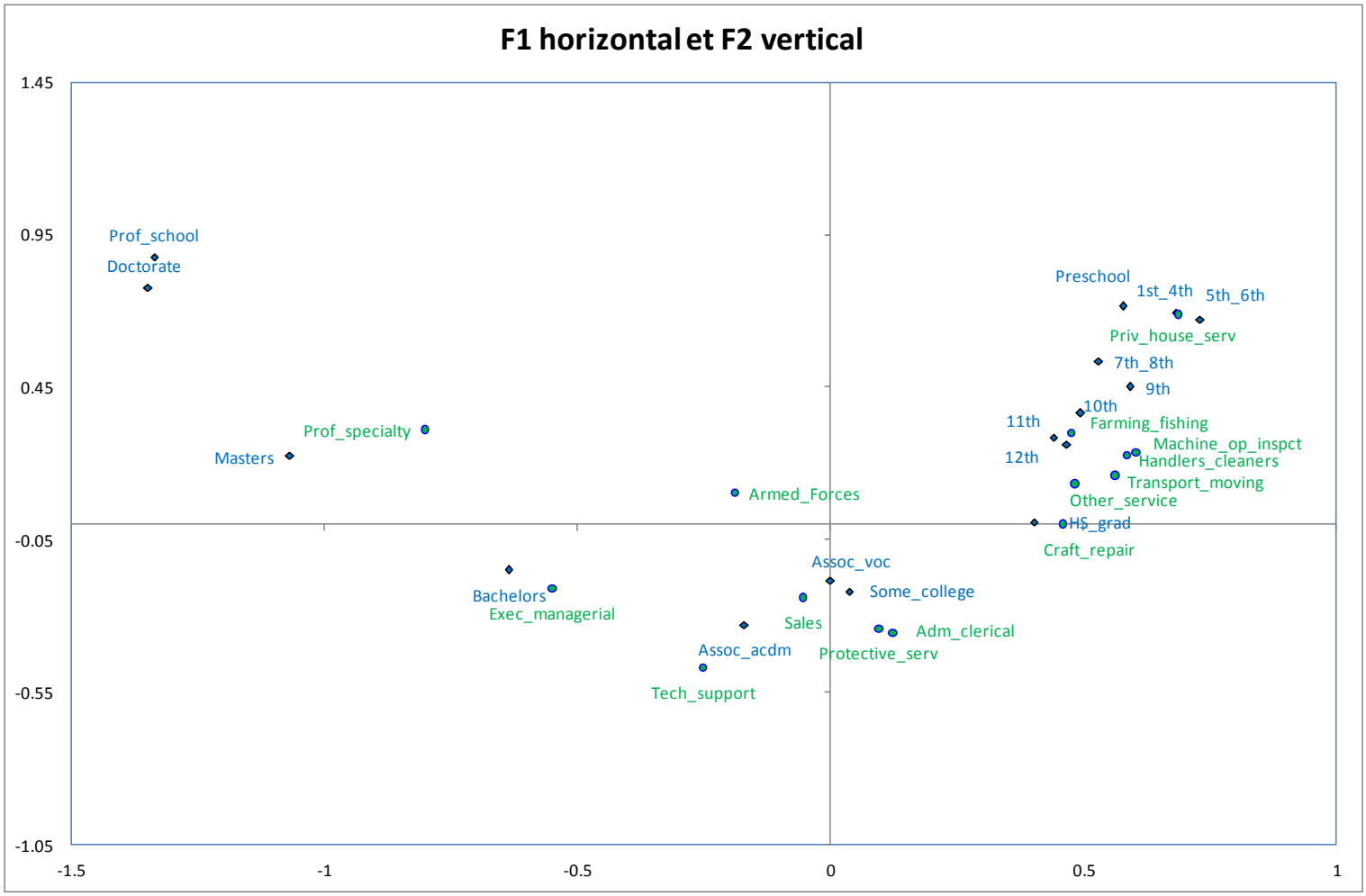
La somme des contributions (CTR) des modalités pour un facteur est égale à 100%.

L'addition des COS² d'une modalité indique la qualité de représentation d'une modalité sur les axes considérés. Par exemple, 91% (0.5885 + 0.3186 + 0.0031) de l'information véhiculée par (Education = 10th) est reproduite par les 3 premiers axes.

Coordonnées factorielles des colonnes.

	Coordonnées factorielles des colonnes			Contributions			Qualité de représentation		
	G1	G2	G3	CTR1	CTR2	CTR3	COR1	COR2	COR3
Adm_clerical	0.1228	-0.3571	0.1435	0.64%	21.85%	14.84%	0.0857	0.7245	0.1170
Armed_Forces	-0.1881	0.1040	0.3416	0.00%	0.00%	0.22%	0.0496	0.0152	0.1636
Craft_repair	0.4589	0.0014	0.1067	9.70%	0.00%	8.94%	0.8772	0.0000	0.0474
Exec_managerial	-0.5504	-0.2112	-0.2213	13.90%	8.29%	38.27%	0.7443	0.1096	0.1203
Farming_fishing	0.4769	0.3007	-0.1740	2.55%	4.12%	5.80%	0.4749	0.1888	0.0632
Handlers_cleaners	0.5866	0.2259	-0.0745	5.38%	3.23%	1.48%	0.8144	0.1207	0.0131
Machine_op_inspct	0.6038	0.2347	-0.0550	8.31%	5.08%	1.17%	0.8296	0.1253	0.0069
Other_service	0.4834	0.1330	-0.0285	8.67%	2.66%	0.51%	0.8308	0.0629	0.0029
Priv_house_serv	0.6874	0.6885	-0.5266	0.86%	3.50%	8.62%	0.2809	0.2818	0.1649
Prof_specialty	-0.8013	0.3107	0.0952	43.47%	26.47%	10.46%	0.8584	0.1290	0.0121
Protective_serv	0.0957	-0.3408	0.1218	0.07%	3.49%	1.87%	0.0612	0.7756	0.0992
Sales	-0.0542	-0.2399	-0.0839	0.12%	9.67%	4.97%	0.0325	0.6355	0.0777
Tech_support	-0.2515	-0.4706	0.1213	0.69%	9.78%	2.74%	0.1391	0.4868	0.0324
Transport_moving	0.5631	0.1605	0.0184	5.63%	1.85%	0.10%	0.8356	0.0679	0.0009

Carte factorielle. Nous disposons d'une représentation simultanée des modalités lignes et colonnes. Nous pouvons spécifier les facteurs en abscisse et on ordonnée. Si l'on s'en tient au premier plan factoriel, nous obtenons.



Ouh là, le bel « effet Guttman ». Nous savions qu'il y avait un ordonnancement sous-jacent aux modalités du niveau d'étude. Manifestement, nous retrouvons le même phénomène pour le type d'occupation. Ce résultat n'est pas gênant en soi. L'enjeu maintenant pour nous est d'identifier les associations entre les deux variables. On les devine plus ou moins dans la carte factorielle. Les graphiques de Bertin, où les modalités sont ordonnées selon leurs positions sur les axes factoriels, devraient confirmer / infirmer les impressions visuelles.

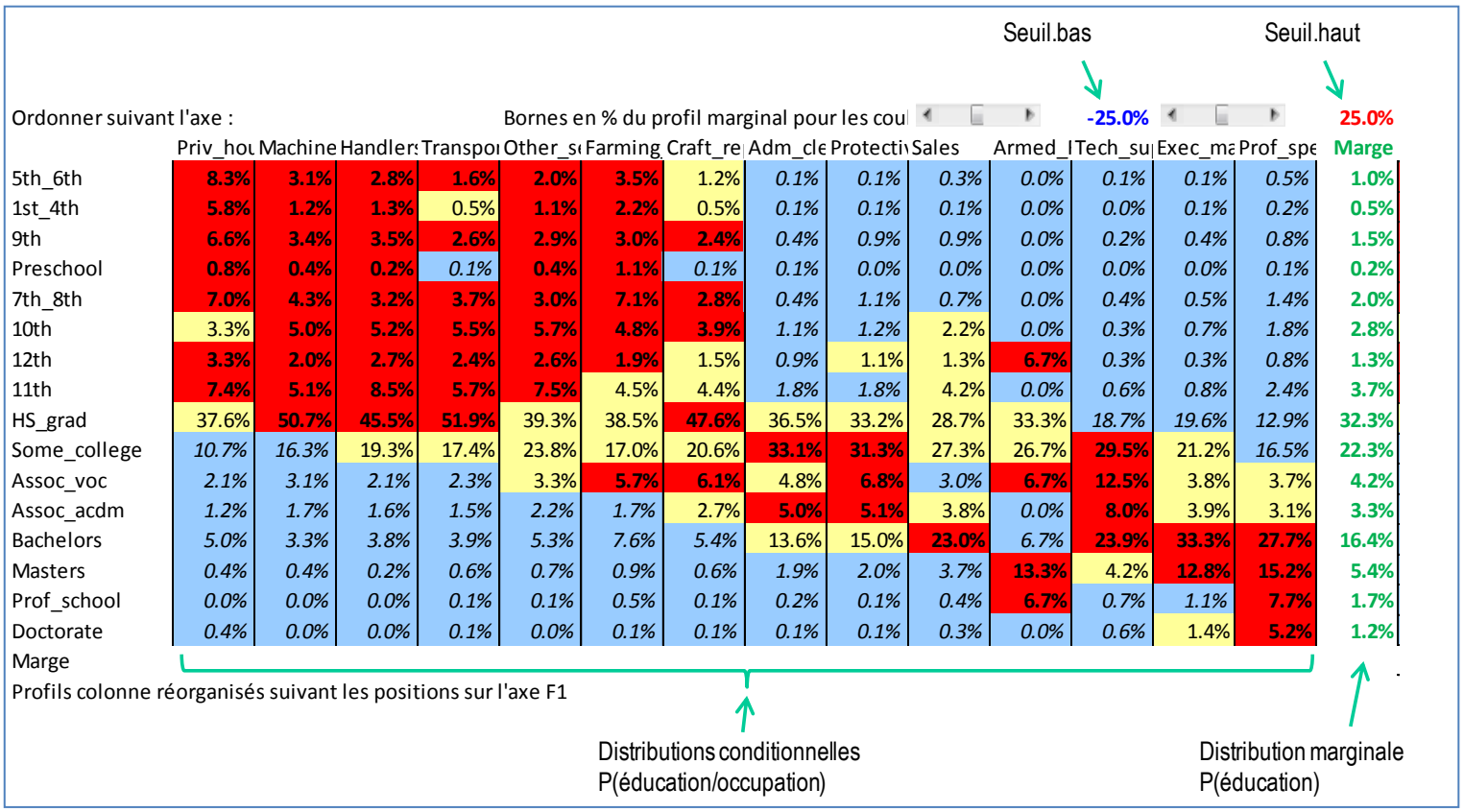
Graphique de Bertin – Profils colonnes. Nous observons les estimations des probabilités conditionnelles P(éducation / occupation) dans le tableau des profils colonnes. La somme des pourcentages en colonne fait 100%. La référence est la distribution marginale des modalités d'éducation P(éducation). On décide qu'une modalité est sous représentée dans une colonne (c.-à-d. conditionnellement à une des valeurs de occupation), et sera mise en fond bleu, lorsque

$$\frac{P(education/occupation) - P(education)}{P(education)} < seuil.bas$$

« Seuil.bas » est exprimé en pourcentage, il est modifiable interactivement par l'utilisateur.

A contrario, on considère qu'il y a surreprésentation, la cellule sera mise en fond rouge, si l'écart relatif est supérieur à « seuil.haut ».

Voyons ce qu'il en est de notre tableau où les modalités sont ordonnées selon le 1er facteur, avec (seuil.bas = -25%, seuil.haut = 25%) :



	Priv_hou	Machine	Handlers	Transpor	Other_s	Farming	Craft_re	Adm_cle	Protecti	Sales	Armed_F	Tech_su	Exec_ma	Prof_spe	Marge
5th_6th	8.3%	3.1%	2.8%	1.6%	2.0%	3.5%	1.2%	0.1%	0.1%	0.3%	0.0%	0.1%	0.1%	0.5%	1.0%
1st_4th	5.8%	1.2%	1.3%	0.5%	1.1%	2.2%	0.5%	0.1%	0.1%	0.1%	0.0%	0.0%	0.1%	0.2%	0.5%
9th	6.6%	3.4%	3.5%	2.6%	2.9%	3.0%	2.4%	0.4%	0.9%	0.9%	0.0%	0.2%	0.4%	0.8%	1.5%
Preschool	0.8%	0.4%	0.2%	0.1%	0.4%	1.1%	0.1%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.2%
7th_8th	7.0%	4.3%	3.2%	3.7%	3.0%	7.1%	2.8%	0.4%	1.1%	0.7%	0.0%	0.4%	0.5%	1.4%	2.0%
10th	3.3%	5.0%	5.2%	5.5%	5.7%	4.8%	3.9%	1.1%	1.2%	2.2%	0.0%	0.3%	0.7%	1.8%	2.8%
12th	3.3%	2.0%	2.7%	2.4%	2.6%	1.9%	1.5%	0.9%	1.1%	1.3%	6.7%	0.3%	0.3%	0.8%	1.3%
11th	7.4%	5.1%	8.5%	5.7%	7.5%	4.5%	4.4%	1.8%	1.8%	4.2%	0.0%	0.6%	0.8%	2.4%	3.7%
HS_grad	37.6%	50.7%	45.5%	51.9%	39.3%	38.5%	47.6%	36.5%	33.2%	28.7%	33.3%	18.7%	19.6%	12.9%	32.3%
Some_college	10.7%	16.3%	19.3%	17.4%	23.8%	17.0%	20.6%	33.1%	31.3%	27.3%	26.7%	29.5%	21.2%	16.5%	22.3%
Assoc_voc	2.1%	3.1%	2.1%	2.3%	3.3%	5.7%	6.1%	4.8%	6.8%	3.0%	6.7%	12.5%	3.8%	3.7%	4.2%
Assoc_acdm	1.2%	1.7%	1.6%	1.5%	2.2%	1.7%	2.7%	5.0%	5.1%	3.8%	0.0%	8.0%	3.9%	3.1%	3.3%
Bachelors	5.0%	3.3%	3.8%	3.9%	5.3%	7.6%	5.4%	13.6%	15.0%	23.0%	6.7%	23.9%	33.3%	27.7%	16.4%
Masters	0.4%	0.4%	0.2%	0.6%	0.7%	0.9%	0.6%	1.9%	2.0%	3.7%	13.3%	4.2%	12.8%	15.2%	5.4%
Prof_school	0.0%	0.0%	0.0%	0.1%	0.1%	0.5%	0.1%	0.2%	0.1%	0.4%	6.7%	0.7%	1.1%	7.7%	1.7%
Doctorate	0.4%	0.0%	0.0%	0.1%	0.0%	0.1%	0.1%	0.1%	0.1%	0.3%	0.0%	0.6%	1.4%	5.2%	1.2%
Marge															

Profils colonne réorganisés suivant les positions sur l'axe F1

Nous observons une surreprésentation des (bachelors, masters, prof school et doctorate) parmi la profession « prof speciality ». On le devinait plus ou moins en inspectant la carte factorielle, cette profession s'éloignait ostensiblement des différents niveaux d'éducation, et se situait du côté des 4 précités. L'attraction apparaît clairement dans le graphique de Bertin.

La proximité entre (occupation = « sales ») et (éducation = « assoc voc » ou « some college ») dans le plan factoriel en revanche s'avère trompeuse. Il n'en est rien à la lecture du graphique de Bertin, ce serait plutôt la modalité « bachelors » qui serait surreprésentée.

Remarque : Dans le même ordre d'idée, un graphique de Bertin basé sur les indices d'attraction / répulsion, ou sur les résidus standardisés aurait été intéressant également.

Graphique de Bertin – Profils lignes. Le tableau des profils est contigu au précédent dans la feuille **Résultats**. Les probabilités conditionnelles sont inversées [P(occupation / éducation)], la distribution marginale de référence est P(occupation). Les mêmes règles de mise en évidence des sous et sur représentation sont utilisées.

	Priv_house_serv	Machine_op_inspct	Handlers_cleaners	Transport_moving	Other_service	Farming_fishing	Craft_repair	Adm_clerical	Protective_serv	Sales	Armed_Forces	Tech_support	Exec_managerial	Prof_specialty
5th_6th	**3.9%**	**18.7%**	**11.6%**	**7.5%**	**19.3%**	**10.2%**	13.9%	*1.6%*	*0.2%*	*3.3%*	*0.0%*	*0.2%*	*1.2%*	*8.4%*
1st_4th	**5.7%**	**14.6%**	**10.5%**	4.9%	**22.3%**	**13.4%**	11.3%	*2.4%*	*0.4%*	*3.2%*	*0.0%*	*0.0%*	*2.4%*	*8.9%*
9th	**2.1%**	**13.5%**	**9.5%**	**8.1%**	**18.8%**	**5.8%**	**19.0%**	*2.6%*	*1.2%*	*6.2%*	*0.0%*	*0.4%*	*3.0%*	*9.7%*
Preschool	**2.4%**	**14.5%**	**6.0%**	*2.4%*	**26.5%**	**20.5%**	*7.2%*	*3.6%*	*0.0%*	*2.4%*	*0.0%*	*0.0%*	*1.2%*	*13.3%*
7th_8th	**1.8%**	**13.5%**	**6.9%**	**9.1%**	**15.6%**	**11.1%**	**18.0%**	*2.1%*	*1.2%*	*4.2%*	*0.0%*	*0.6%*	*2.9%*	*13.0%*
10th	0.6%	**10.9%**	**7.8%**	**9.3%**	**20.2%**	**5.1%**	**17.2%**	*4.2%*	*0.9%*	8.6%	*0.0%*	*0.4%*	*3.0%*	*11.8%*
12th	**1.2%**	**9.3%**	**8.4%**	**8.5%**	**19.6%**	**4.4%**	14.0%	*7.9%*	1.7%	10.7%	**0.2%**	*0.6%*	*2.7%*	*10.8%*
11th	**1.0%**	**8.4%**	**9.8%**	**7.4%**	**20.3%**	3.7%	14.9%	*5.5%*	*1.0%*	12.8%	*0.0%*	*0.5%*	*2.8%*	*11.9%*
HS_grad	0.6%	**9.7%**	**6.0%**	**7.7%**	12.3%	3.6%	**18.4%**	13.0%	2.1%	10.0%	0.0%	*1.7%*	*7.6%*	*7.3%*
Some_college	*0.2%*	*4.5%*	3.7%	3.8%	10.8%	2.3%	11.6%	**17.1%**	**2.8%**	13.8%	0.0%	**3.9%**	11.8%	*13.6%*
Assoc_voc	*0.2%*	*4.6%*	*2.1%*	*2.7%*	7.8%	**4.1%**	**18.2%**	13.1%	**3.3%**	*7.9%*	**0.0%**	**8.8%**	11.4%	15.9%
Assoc_acdm	*0.2%*	*3.2%*	*2.1%*	*2.2%*	*6.9%*	*1.6%*	10.4%	**17.6%**	**3.1%**	13.1%	*0.0%*	**7.2%**	15.0%	17.4%
Bachelors	*0.1%*	*1.2%*	*1.0%*	*1.2%*	*3.2%*	*1.4%*	*4.1%*	9.5%	1.8%	**15.8%**	*0.0%*	**4.3%**	**25.2%**	**31.0%**
Masters	*0.0%*	*0.5%*	*0.2%*	*0.5%*	*1.3%*	*0.5%*	*1.3%*	*4.0%*	*0.8%*	*7.8%*	**0.1%**	2.3%	**29.3%**	**51.5%**
Prof_school	*0.0%*	*0.1%*	*0.0%*	*0.4%*	*0.8%*	*0.8%*	*1.1%*	*1.4%*	*0.1%*	*2.8%*	**0.1%**	*1.2%*	*8.3%*	**82.9%**
Doctorate	*0.2%*	*0.2%*	*0.0%*	*0.3%*	*0.3%*	*0.2%*	*0.7%*	*1.0%*	*0.2%*	*2.7%*	*0.0%*	*1.3%*	14.1%	**78.8%**
	0.5%	**6.2%**	**4.2%**	**4.8%**	**10.1%**	**3.1%**	**12.5%**	**11.5%**	**2.0%**	**11.3%**	**0.0%**	**3.0%**	**12.5%**	**18.4%**

Profils ligne réorganisés suivant les positions sur l'axe F1

Distribution conditionnelle P(occupation/éducation)

Référence : distribution marginale P(occupation)

Nous lisons par exemple P(prof speciality / doctorate) = 78.8%, alors que les enseignants (occupation = prof speciality) ne représentent que 18.4% de la population : ceux qui effectuent un doctorat se destinent plus spécialement aux métiers de l'enseignement.

4 Conclusion

Je passe mon temps à défendre Excel auprès des irréductibles de R… et à défendre R auprès des fondus d'Excel. En réalité, ce type de débat est totalement stérile[10]. Il nous appartient de cerner ce que l'on peut attendre d'un outil. Excel, le tableur en général devrait-on dire, est très commode pour tout ce qui est manipulation et préparation des données. Il est moins compétitif pour ce qui est des calculs statistiques, si on s'en tient aux fonctions natives[11]. Il devient nettement plus intéressant lorsqu'on lui adjoint des add-ins (macros-complémentaires) spécifiques de statistique et de data mining. J'en ai moi-même parlé à plusieurs reprises[12]. Le tableur devient alors un outil particulièrement performant tant que l'on ne cherche pas à traiter de très grands volumes. Nous avons pu le vérifier encore une fois

[10] http://robjhyndman.com/hyndsight/rvsexcel/ … certes, certes. En attendant, Excel apparaît quasi-systématiquement dans toutes les offres d'emploi associées au mot-clé « statistique » sur le site de l'APEC.

[11] B. McCullogh, D. Heiser, « On the accuracy of statistical procedures in Microsoft Excel 2007 », Computational Statistics and Data Analysis 52, pp. 4570–4578 2008.

[12] « L'add-in Real Statistics pour Excel », juin 2014 ; « SQL Server Data Mining Add-ins », juillet 2014 ; etc.

dans ce tutoriel où nous avons mis en œuvre l'add-in dédié à « l'analyse exploratoire des données » de Jacques Vaillé, un outil pédagogique de très grande qualité – la présentation des données sous forme de graphique de Bertin est particulièrement astucieuse – accessible sur la page Excel'Ense du site Modulad.

5 Bibliographie

Jacques Vaillé, « La statistique au service des données : quelques macros Excel pour faire de l'analyse exploratoire des données », La revue MODULAD, n°43, 2011.

La revue MODULAD, la page Excel'Ense : http://www.modulad.fr/excel.htm