1 Objectif

ACP (analyse en composantes principales) sous Excel avec la librairie XNUMBERS.

Tout le monde l'a bien compris, le tableur est pour moi avant tout un outil pédagogique pour l'enseignement de la statistique et du data mining. Les étudiants ne peuvent pas entrer des commandes ou cliquer frénétiquement au petit bonheur la chance. Ils doivent regarder de près les formules pour pouvoir les comprendre et les reproduire. Il n'y a pas mieux pour les amener à décortiquer les différentes étapes du calcul quelle que soit la méthode étudiée.

Nous avions analysé la Régression Linéaire Multiple sous Excel récemment (Régression Linéaire sous Excel, Mars 2018). Dans ce tutoriel, nous explorons la mécanique d'une autre méthode phare de la data science (voir Top Data Science and Machine Learning Used in 2017) : l'analyse en composantes principales (ACP). J'en profiterai pour présenter XNUMBERS, une librairie particulièrement performante pour le calcul scientifique sous Excel. Elle nous sera utile en particulier pour la factorisation des matrices à l'aide de la décomposition en valeurs singulières.

2 La librairie XNUMBERS

XNUMBERS est une libraire pour le calcul à très haute précision pour Excel. Il comprend un grand nombre de fonctions mathématiques et de méthodes numériques. Le projet a été développé à l'origine par la Foxes Team sous la houlette de Leonardo Volpi. Il a été abandonné en 2008 (version 5.6). Depuis, la librairie a été reprise (version 6.0 et suivantes) par un astronome dont le frère, John Beyers, a assuré le portage sur les versions les plus récentes d'Excel. Elle est accessible librement et est plutôt bien documentée, un fichier d'aide au format CHM l'accompagne. J'ai intégré la macro complémentaire **XN.xlam** accessible en ligne (http://www.thetropicalevents.com/Xnumbers60.htm) dans **Excel 2016 - 64 bits**, le tout a parfaitement fonctionné[1].

3 Données

Le fichier « **autos-acp-excel.xlsx** » (Feuille « **data** ») décrit $n = 18$ véhicules à l'aide de $p = 6$ variables (cylindrée, puissance, longueur, largeur, poids et vitesse maximale).

[1] Cf. https://www.excel-pratique.com/fr/fonctions-complementaires/installation-macro-complementaire.php

Modele	CYL	PUISS	LONG	LARG	POIDS	V.MAX
Alfasud TI	1350	79	393	161	870	165
Audi 100	1588	85	468	177	1110	160
Simca 1300	1294	68	424	168	1050	152
Citroen GS Club	1222	59	412	161	930	151
Fiat 132	1585	98	439	164	1105	165
Lancia Beta	1297	82	429	169	1080	160
Peugeot 504	1796	79	449	169	1160	154
Renault 16 TL	1565	55	424	163	1010	140
Renault 30	2664	128	452	173	1320	180
Toyota Corolla	1166	55	399	157	815	140
Alfetta 1.66	1570	109	428	162	1060	175
Princess 1800	1798	82	445	172	1160	158
Datsun 200L	1998	115	469	169	1370	160
Taunus 2000	1993	98	438	170	1080	167
Rancho	1442	80	431	166	1129	144
Mazda 9295	1769	83	440	165	1095	165
Opel Rekord	1979	100	459	173	1120	173
Lada 1300	1294	68	404	161	955	140

Figure 1 - Tableau de données - Feuille "data"

Il sert de données d'illustrations dans mon support de cours consacré à l'ACP (RAK, 2013) qui sera notre principale référencee. Nous pourrons ainsi vérifier nos calculs à chaque stade.

4 ACP sous Excel

L'ACP normée peut être traitée de deux manières : par la diagonalisation de la matrice des corrélations, ou par la décomposition en valeurs singulières de la matrice des données centrées et réduites. Nous optons pour cette seconde solution.

4.1 Préparation des données

La première étape passe par le centrage et réduction des variables la matrice des données X, les valeurs z_{ij} de la matrice Z sont calculées comme suit :

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}$$

Où $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ est la moyenne de la variable X_j, $\sigma_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}$ son écart-type.

Nous copions le tableau de données dans une nouvelle feuille « acp-svd » :

- A la ligne 21, nous calculons les moyennes. Pour CYL, nous insérons en B21 la formule =CNUM(xMean(B2:B19)). xMean() est une fonction de XNUMBERS qui effectue les calculs à haute précision. La fonction renvoie une chaîne de caractères, on la convertit avec CNUM().
- Pour l'écart-type, nous insérons en B22 la formule =CNUM(xStDevP(B2:B19)). xStDevP() calcule l'écart-type comme ci-dessus.
- Nous complétons les lignes 21 et 22 par copier-coller pour l'ensemble des variables.
- Il nous reste à produire la matrice Z. Pour le premier individu (Alfasud TI) et la première variable (CYL), nous appliquons en J2 la transformation =(B2-B$21)/B$22. Les références semi-absolues (voir les positions des $) permettent de compléter le tableau par copier-coller d'une traite.

J2 : =(B2-B$21)/B$22

	A	B	C	D	E	F	G
1	Modele	CYL	PUISS	LONG	LARG	POIDS	V.MAX
2	Alfasud TI	1350	79	393	161	870	165
3	Audi 100	1588	85	468	177	1110	160
4	Simca 1300	1294	68	424	168	1050	152
5	Citroen GS Club	1222	59	412	161	930	151
6	Fiat 132	1585	98	439	164	1105	165
7	Lancia Beta	1297	82	429	169	1080	160
8	Peugeot 504	1796	79	449	169	1160	154
9	Renault 16 TL	1565	55	424	163	1010	140
10	Renault 30	2664	128	452	173	1320	180
11	Toyota Corolla	1166	55	399	157	815	140
12	Alfetta 1.66	1570	109	428	162	1060	175
13	Princess 1800	1798	82	445	172	1160	158
14	Datsun 200L	1998	115	469	169	1370	160
15	Taunus 2000	1993	98	438	170	1080	167
16	Rancho	1442	80	431	166	1129	144
17	Mazda 9295	1769	83	440	165	1095	165
18	Opel Rekord	1979	100	459	173	1120	173
19	Lada 1300	1294	68	404	161	955	140
20							
21	Moyenne	1631.67	84.61	433.50	166.67	1078.83	158.28
22	Ecart-type	363.39	19.80	21.48	5.16	133.10	11.80

	I	J	K	L	M	N	O
1	Modele	CYL	PUISS	LONG	LARG	POIDS	V.MAX
2	Alfasud TI	-0.7751	-0.2834	-1.8851	-1.0973	-1.5690	0.5698
3	Audi 100	-0.1202	0.0196	1.6058	2.0010	0.2342	0.1460
4	Simca 1300	-0.9292	-0.8389	-0.4422	0.2582	-0.2166	-0.5321
5	Citroen GS Club	-1.1273	-1.2933	-1.0007	-1.0973	-1.1182	-0.6168
6	Fiat 132	-0.1284	0.6761	0.2560	-0.5164	0.1966	0.5698
7	Lancia Beta	-0.9209	-0.1319	-0.2095	0.4518	0.0088	0.1460
8	Peugeot 504	0.4522	-0.2834	0.7215	0.4518	0.6098	-0.3626
9	Renault 16 TL	-0.1835	-1.4953	-0.4422	-0.7100	-0.5172	-1.5492
10	Renault 30	2.8408	2.1911	0.8611	1.2264	1.8119	1.8411
11	Toyota Corolla	-1.2814	-1.4953	-1.6058	-1.8719	-1.9822	-1.5492
12	Alfetta 1.66	-0.1697	1.2316	-0.2560	-0.9037	-0.1415	1.4173
13	Princess 1800	0.4577	-0.1319	0.5353	1.0328	0.6098	-0.0235
14	Datsun 200L	1.0081	1.5346	1.6524	0.4518	2.1876	0.1460
15	Taunus 2000	0.9943	0.6761	0.2095	0.6455	0.0088	0.7393
16	Rancho	-0.5219	-0.2329	-0.1164	-0.1291	0.3769	-1.2102
17	Mazda 9295	0.3779	-0.0814	0.3025	-0.3227	0.1215	0.5698
18	Opel Rekord	0.9558	0.7771	1.1869	1.2264	0.3093	1.2478
19	Lada 1300	-0.9292	-0.8389	-1.3731	-1.0973	-0.9304	-1.5492

Figure 2 - Tableau des données centrées et réduites Z (J2:O19) - Feuille "acp-svd"

4.2 Principe de la décomposition en valeurs singulières

La décomposition en valeurs singulières (SVD, singular-value decomposition) est une méthode de factorisation très populaire en *data mining*. La matrice Z de dimension (n, p) est décomposée en 3 sous-matrices (RAK, 2013 ; page 20) :

$$Z = UDV^T$$

Sous la configuration usuelle où (n > p) : U est de dimension (n, p), elle positionne les individus dans le nouvel espace de représentation ; V est de dimension (p, p) et permet de situer le rôle des variables ; D est une matrice diagonale (p, p) et sert à évaluer la qualité de la représentation.

4.3 Qualité de la représentation

Voyons ce qu'il en de la matrice D. Nous utilisons xSVDD(). Comme il s'agit d'une fonction matricielle, nous devons valider la saisie avec la combinaison de touches CTRL + SHIFT + ENTREE.

En Q3:V8, nous insérons {=CNUM(xSVDD(J2:O19))}. Les accolades { } sont automatiquement ajoutées par Excel pour signifier que nous avons bien validé une fonction matricielle destinée à compléter automatiquement une plage de cellules.

Q3 : {=CNUM(xSVDD(J2:O19))}

	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	Modele	CYL	PUISS	LONG	LARG	POIDS	V.MAX		Qualité de la repésentation des composantes					
2	Alfasud TI	-0.7751	-0.2834	-1.8851	-1.0973	-1.5690	0.5698		F1	F2	F3	F4	F5	F6
3	Audi 100	-0.1202	0.0196	1.6058	2.0010	0.2342	0.1460		8.9205	0	0	0	0	0
4	Simca 1300	-0.9292	-0.8389	-0.4422	0.2582	-0.2166	-0.5321		0	3.9254	0	0	0	0
5	Citroen GS Clu	-1.1273	-1.2933	-1.0007	-1.0973	-1.1182	-0.6168		0	0	2.5914	0	0	0
6	Fiat 132	-0.1284	0.6761	0.2560	-0.5164	0.1966	0.5698		0	0	0	1.9623	0	0
7	Lancia Beta	-0.9209	-0.1319	-0.2095	0.4518	0.0088	0.1460		0	0	0	0	1.2924	0
8	Peugeot 504	0.4522	-0.2834	0.7215	0.4518	0.6098	-0.3626		0	0	0	0	0	0.8827
9	Renault 16 TL	-0.1835	-1.4953	-0.4422	-0.7100	-0.5172	-1.5492							
10	Renault 30	2.8408	2.1911	0.8611	1.2264	1.8119	1.8411	d_k	8.92	3.93	2.59	1.96	1.29	0.88
11	Toyota Corolla	-1.2814	-1.4953	-1.6058	-1.8719	-1.9822	-1.5492							
12	Alfetta 1.66	-0.1697	1.2316	-0.2560	-0.9037	-0.1415	1.4173	n	18					
13	Princess 1800	0.4577	-0.1319	0.5353	1.0328	0.6098	-0.0235							
14	Datsun 200L	1.0081	1.5346	1.6524	0.4518	2.1876	0.1460	lambda_k	**4.4209**	**0.8561**	0.3731	0.2139	0.0928	0.0433
15	Taunus 2000	0.9943	0.6761	0.2095	0.6455	0.0088	0.7393							
16	Rancho	-0.5219	-0.2329	-0.1164	-0.1291	0.3769	-1.2102	%inertie	**73.68%**	**14.27%**	6.22%	3.57%	1.55%	0.72%
17	Mazda 9295	0.3779	-0.0814	0.3025	-0.3227	0.1215	0.5698							
18	Opel Rekord	0.9558	0.7771	1.1869	1.2264	0.3093	1.2478							
19	Lada 1300	-0.9292	-0.8389	-1.3731	-1.0973	-0.9304	-1.5492							

Figure 3 - Calcul de la matrice D - Feuille "acp-svd"

En Q10:V10, nous récupérons les valeurs d_k sur la diagonale (k = 1, …, 6). Nous en déduisons λ_k qui correspond au pouvoir explicatif du facteur, avec (RAK, 2013 ; page 20) :

$$\lambda_k = \frac{d_k^2}{n}$$

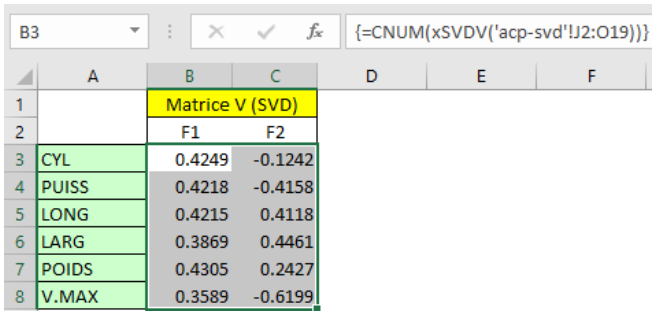
Ainsi, pour le premier facteur nous avons $\lambda_1 = 4.4209$, etc.

En Q16:V16, nous exprimons la qualité de représentation en proportion d'inertie expliquée (en Q16, nous avons =Q14/SOMME($Q$14:$V$14) ; la ligne est complétée par copier-coller). Nous observons

que les 2 premiers axes permettent de restituer **(73.68 + 14.27) = 87.95%** de l'information disponible. Nous nous en tiendrons à ces deux premiers facteurs (**K = 2**) dans le reste de l'étude.

4.4 Analyse des variables

Pour analyser les variables, nous avons besoin de la sous-matrice V (v_{jk}) issue de la décomposition. Nous créons une nouvelle feuille « variables ». Nous listons les variables, puis nous insérons en B3:C8 la fonction {=CNUM(xSVDV('acp-svd'!J2:O19))}. Remarques : (1) V possède bien p = 6 lignes, mais elle est restreinte à K = 2 colonnes dans notre analyse ; (2) la formule prend en entrée la matrice Z des données centrées réduites (J2:019) située dans la feuille 'acp-svd'.



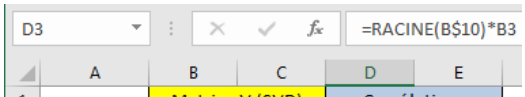
	A	B	C	D	E	F
1		Matrice V (SVD)				
2		F1	F2			
3	CYL	0.4249	-0.1242			
4	PUISS	0.4218	-0.4158			
5	LONG	0.4215	0.4118			
6	LARG	0.3869	0.4461			
7	POIDS	0.4305	0.2427			
8	V.MAX	0.3589	-0.6199			

Figure 4 - Calcul de la matrice V - Feuille "variables"

Corrélations variables – facteurs. Nous obtenons la corrélation (r_{jk}) des variables (X_j) avec les facteurs (F_k) via

$$r_{jk} = \sqrt{\lambda_k} \times v_{jk}$$

Dans notre feuille de calcul, après y avoir reporté les valeurs de λ_k (copier – collage spécial / valeurs) en B10:C10, nous appliquons =RACINE(B$10)*B3 en D3. Puis nous étendons par copier-coller.



	A	B	C	D	E
1		Matrice V (SVD)		Corrélations	
2		F1	F2	F1	F2
3	CYL	0.4249	-0.1242	0.8935	-0.1149
4	PUISS	0.4218	-0.4158	0.8869	-0.3847
5	LONG	0.4215	0.4118	0.8862	0.3810
6	LARG	0.3869	0.4461	0.8135	0.4127
7	POIDS	0.4305	0.2427	0.9052	0.2245
8	V.MAX	0.3589	-0.6199	0.7547	-0.5735
9					
10	lambda_k	4.4209	0.8561		

Figure 5 - Corrélations variables-facteurs - Feuille "variables"

Qualité de représentation des variables (COS²). La qualité de la représentation d'une variable (COS²) sur un facteur correspond au carré de la corrélation. Pour chaque variable X_j, la somme des COS² sur l'ensemble des p = 6 facteurs potentiels est égale à 1 ($\sum_{k=1}^{p} COS_{jk}^2 = 1$).

$$COS_{jk}^2 = r_{jk}^2$$

Nous passons au format « pourcentage » le contenu des cellules. On note par exemple que PUIS est parfaitement représentée puisque (78.7% + 14.8%) = 93.5% de l'information qu'elle véhicule (c'est le cas de le dire) est disponible dans le premier plan factoriel.

F3 : =D3^2

	A	B	C	D	E	F	G
1		Matrice V (SVD)		Corrélations		COS²	
2		F1	F2	F1	F2	F1	F2
3	CYL	0.4249	-0.1242	0.8935	-0.1149	79.8%	1.3%
4	PUISS	0.4218	-0.4158	0.8869	-0.3847	78.7%	14.8%
5	LONG	0.4215	0.4118	0.8862	0.3810	78.5%	14.5%
6	LARG	0.3869	0.4461	0.8135	0.4127	66.2%	17.0%
7	POIDS	0.4305	0.2427	0.9052	0.2245	81.9%	5.0%
8	V.MAX	0.3589	-0.6199	0.7547	-0.5735	57.0%	32.9%
9							
10	lambda_k	4.4209	0.8561				

Figure 6 - Qualité de représentation des variables (COS²) - Feuille "variables"

Contribution des variables aux axes. La contribution des variables est aussi dérivée de la corrélation, mais elle est normalisée par l'importance de l'axe :

$$CTR_{jk} = \frac{r_{jk}^2}{\lambda_k}$$

Pour chaque axe, la somme des contributions des variables est égale à 1 ($\sum_{j=1}^{p} CTR_{jk} = 1$).

H3 : =D3^2/B$10

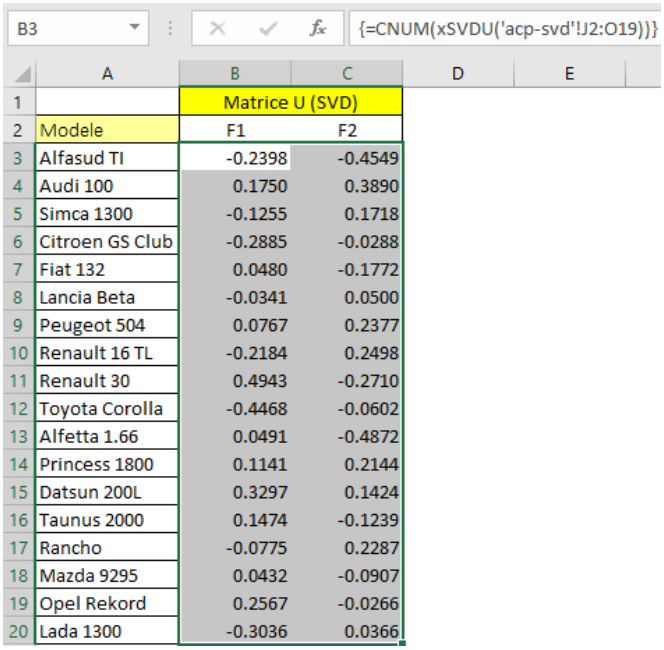
	A	B	C	D	E	F	G	H	I
1		Matrice V (SVD)		Corrélations		COS²		CTR	
2		F1	F2	F1	F2	F1	F2	F1	F2
3	CYL	0.4249	-0.1242	0.8935	-0.1149	79.8%	1.3%	18.1%	1.5%
4	PUISS	0.4218	-0.4158	0.8869	-0.3847	78.7%	14.8%	17.8%	17.3%
5	LONG	0.4215	0.4118	0.8862	0.3810	78.5%	14.5%	17.8%	17.0%
6	LARG	0.3869	0.4461	0.8135	0.4127	66.2%	17.0%	15.0%	19.9%
7	POIDS	0.4305	0.2427	0.9052	0.2245	81.9%	5.0%	18.5%	5.9%
8	V.MAX	0.3589	-0.6199	0.7547	-0.5735	57.0%	32.9%	12.9%	38.4%
9									
10	lambda_k	4.4209	0.8561						

Figure 7 - Contribution (CTR) des variables aux axes - Feuille "variables"

Toutes les variables pèsent peu ou prou de la même manière pour le premier facteur, à l'exception de V.MAX qui, elle, est déterminante pour le second facteur (CTR = 38.4%).

4.5 Analyse des individus

Nous utilisons la matrice U de la décomposition pour obtenir les coordonnées des individus. Nous créons une nouvelle feuille « individus » et nous y reportons les labels des véhicules. Dans notre cas, elle est de dimension (n, K) avec K = 2 puisque nous nous en tenons au premier plan factoriel.



B3 | {=CNUM(xSVDU('acp-svd'!J2:O19))}

	A	B	C	D	E
1		Matrice U (SVD)			
2	Modele	F1	F2		
3	Alfasud TI	-0.2398	-0.4549		
4	Audi 100	0.1750	0.3890		
5	Simca 1300	-0.1255	0.1718		
6	Citroen GS Club	-0.2885	-0.0288		
7	Fiat 132	0.0480	-0.1772		
8	Lancia Beta	-0.0341	0.0500		
9	Peugeot 504	0.0767	0.2377		
10	Renault 16 TL	-0.2184	0.2498		
11	Renault 30	0.4943	-0.2710		
12	Toyota Corolla	-0.4468	-0.0602		
13	Alfetta 1.66	0.0491	-0.4872		
14	Princess 1800	0.1141	0.2144		
15	Datsun 200L	0.3297	0.1424		
16	Taunus 2000	0.1474	-0.1239		
17	Rancho	-0.0775	0.2287		
18	Mazda 9295	0.0432	-0.0907		
19	Opel Rekord	0.2567	-0.0266		
20	Lada 1300	-0.3036	0.0366		

Figure 8 - Calcul de la matrice U - Feuille "individus"

En (B3:C20), nous avons inséré {=CNUM(xSVDU('acp-svd'!J2:O19))}, toujours validée par la combinaison de touche CTRL + SHIFT + ENTREE. Les données centrées-réduites sont récupérées dans la feuille « acp-svd ».

Coordonnées factorielles des individus. Nous obtenons la coordonnée factorielle F_{ik} de l'individu n°i sur l'axe n°k par le produit des matrices D et U :

$$F_{ik} = d_k \times u_{ik}$$

Nous reportons donc les valeurs de d_k dans la nouvelle feuille, en B22:C22 (copier – collage spécial / valeurs). Puis nous créons les deux nouvelles colonnes.

D3 | =B$22*B3

	A	B	C	D	E
1		Matrice U (SVD)		Coordonnées	
2	Modele	F1	F2	F1	F2
3	Alfasud TI	-0.2398	-0.4549	-2.1389	-1.7857
4	Audi 100	0.1750	0.3890	1.5615	1.5270
5	Simca 1300	-0.1255	0.1718	-1.1194	0.6745
6	Citroen GS Club	-0.2885	-0.0288	-2.5737	-0.1129
7	Fiat 132	0.0480	-0.1772	0.4279	-0.6956
8	Lancia Beta	-0.0341	0.0500	-0.3042	0.1961
9	Peugeot 504	0.0767	0.2377	0.6839	0.9331
10	Renault 16 TL	-0.2184	0.2498	-1.9485	0.9804
11	Renault 30	0.4943	-0.2710	4.4097	-1.0636
12	Toyota Corolla	-0.4468	-0.0602	-3.9858	-0.2362
13	Alfetta 1.66	0.0491	-0.4872	0.4377	-1.9124
14	Princess 1800	0.1141	0.2144	1.0182	0.8417
15	Datsun 200L	0.3297	0.1424	2.9411	0.5592
16	Taunus 2000	0.1474	-0.1239	1.3149	-0.4865
17	Rancho	-0.0775	0.2287	-0.6911	0.8977
18	Mazda 9295	0.0432	-0.0907	0.3857	-0.3562
19	Opel Rekord	0.2567	-0.0266	2.2898	-0.1043
20	Lada 1300	-0.3036	0.0366	-2.7086	0.1437
21					
22	d_k	8.9205	3.9254		

Figure 9 - Coordonnées factorielles des individus - Feuille "individus"

Puisque nous sommes sous Excel, nous pouvons construire une graphique nuage de points qui permet de situer les positions relatives des individus.

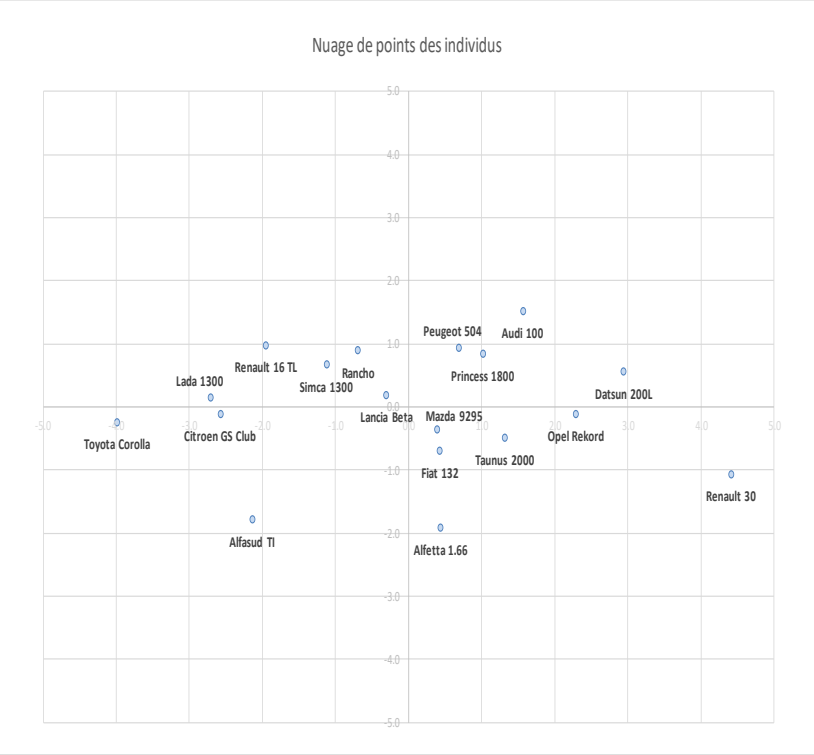


Figure 10 - Représentation des individus dans le premier plan factoriel

Qualité de représentation des individus (COS²). Pour calculer le COS² de l'individu n°i sur l'axe n°k, nous devons tout d'abord calculer le carré de la distance à l'origine o_i^2 de chaque individu. Il nous renseigne sur la participation de chaque observation dans l'inertie totale du nuage de points.

$$o_i^2 = \sum_{j=1}^{p} z_{ij}^2$$

Nous l'obtenons en effectuant la SOMME.CARRES() des coordonnées centrées réduites des individus, lesquelles sont disponibles dans la feuille « acp-svd ».

F3 : =SOMME.CARRES('acp-svd'!J2:O2)

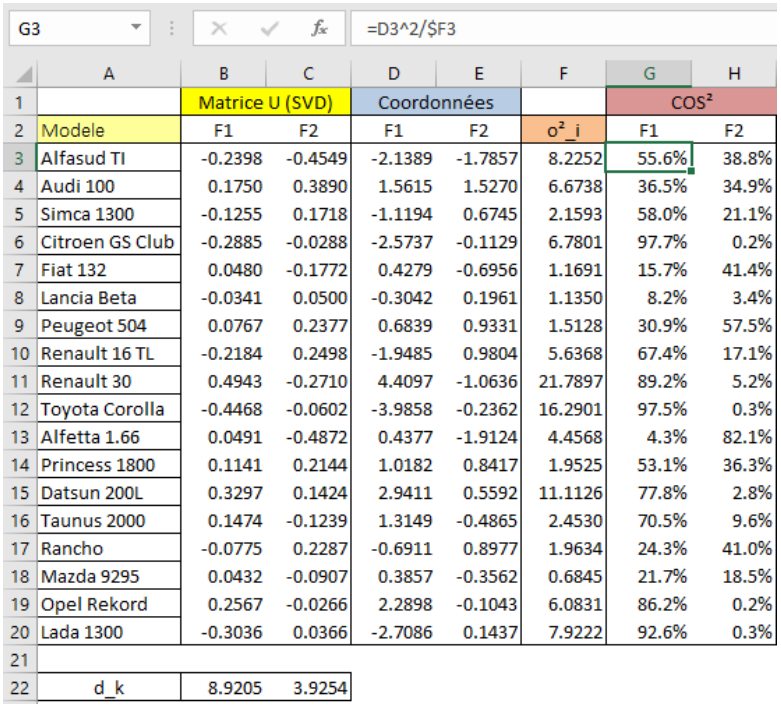
	A	B	C	D	E	F
1		Matrice U (SVD)		Coordonnées		
2	Modele	F1	F2	F1	F2	o²_i
3	Alfasud TI	-0.2398	-0.4549	-2.1389	-1.7857	8.2252
4	Audi 100	0.1750	0.3890	1.5615	1.5270	6.6738
5	Simca 1300	-0.1255	0.1718	-1.1194	0.6745	2.1593
6	Citroen GS Club	-0.2885	-0.0288	-2.5737	-0.1129	6.7801
7	Fiat 132	0.0480	-0.1772	0.4279	-0.6956	1.1691
8	Lancia Beta	-0.0341	0.0500	-0.3042	0.1961	1.1350
9	Peugeot 504	0.0767	0.2377	0.6839	0.9331	1.5128
10	Renault 16 TL	-0.2184	0.2498	-1.9485	0.9804	5.6368
11	Renault 30	0.4943	-0.2710	4.4097	-1.0636	21.7897
12	Toyota Corolla	-0.4468	-0.0602	-3.9858	-0.2362	16.2901
13	Alfetta 1.66	0.0491	-0.4872	0.4377	-1.9124	4.4568
14	Princess 1800	0.1141	0.2144	1.0182	0.8417	1.9525
15	Datsun 200L	0.3297	0.1424	2.9411	0.5592	11.1126
16	Taunus 2000	0.1474	-0.1239	1.3149	-0.4865	2.4530
17	Rancho	-0.0775	0.2287	-0.6911	0.8977	1.9634
18	Mazda 9295	0.0432	-0.0907	0.3857	-0.3562	0.6845
19	Opel Rekord	0.2567	-0.0266	2.2898	-0.1043	6.0831
20	Lada 1300	-0.3036	0.0366	-2.7086	0.1437	7.9222
21						
22	d_k	8.9205	3.9254			

Figure 11 - Carré des distances à l'origine des individus (o_i^2) - Feuille "individus"

Nous pouvons alors déduire la qualité de représentation des individus :

$$COS_{ik}^2 = \frac{F_{ik}^2}{o_i^2}$$

Que nous mettons en pourcentage. La somme des COS² pour un individu sur l'ensembles des facteurs est égale à 1.

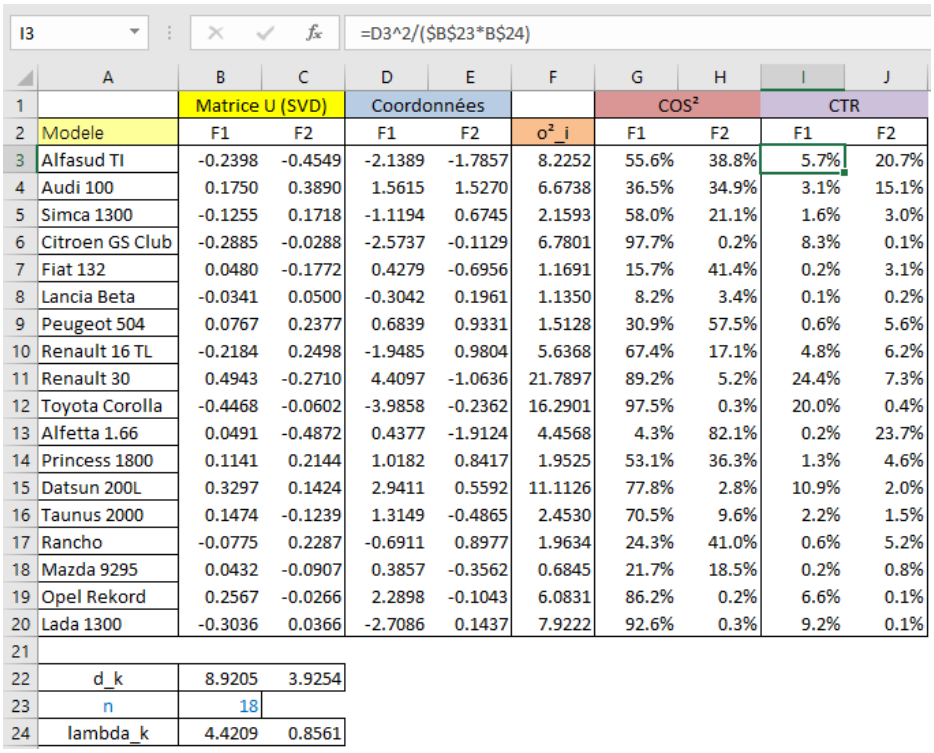


G3 : =D3^2/$F3

	A	B	C	D	E	F	G	H
1		Matrice U (SVD)		Coordonnées			COS²	
2	Modele	F1	F2	F1	F2	o²_i	F1	F2
3	Alfasud TI	-0.2398	-0.4549	-2.1389	-1.7857	8.2252	55.6%	38.8%
4	Audi 100	0.1750	0.3890	1.5615	1.5270	6.6738	36.5%	34.9%
5	Simca 1300	-0.1255	0.1718	-1.1194	0.6745	2.1593	58.0%	21.1%
6	Citroen GS Club	-0.2885	-0.0288	-2.5737	-0.1129	6.7801	97.7%	0.2%
7	Fiat 132	0.0480	-0.1772	0.4279	-0.6956	1.1691	15.7%	41.4%
8	Lancia Beta	-0.0341	0.0500	-0.3042	0.1961	1.1350	8.2%	3.4%
9	Peugeot 504	0.0767	0.2377	0.6839	0.9331	1.5128	30.9%	57.5%
10	Renault 16 TL	-0.2184	0.2498	-1.9485	0.9804	5.6368	67.4%	17.1%
11	Renault 30	0.4943	-0.2710	4.4097	-1.0636	21.7897	89.2%	5.2%
12	Toyota Corolla	-0.4468	-0.0602	-3.9858	-0.2362	16.2901	97.5%	0.3%
13	Alfetta 1.66	0.0491	-0.4872	0.4377	-1.9124	4.4568	4.3%	82.1%
14	Princess 1800	0.1141	0.2144	1.0182	0.8417	1.9525	53.1%	36.3%
15	Datsun 200L	0.3297	0.1424	2.9411	0.5592	11.1126	77.8%	2.8%
16	Taunus 2000	0.1474	-0.1239	1.3149	-0.4865	2.4530	70.5%	9.6%
17	Rancho	-0.0775	0.2287	-0.6911	0.8977	1.9634	24.3%	41.0%
18	Mazda 9295	0.0432	-0.0907	0.3857	-0.3562	0.6845	21.7%	18.5%
19	Opel Rekord	0.2567	-0.0266	2.2898	-0.1043	6.0831	86.2%	0.2%
20	Lada 1300	-0.3036	0.0366	-2.7086	0.1437	7.9222	92.6%	0.3%
21								
22	d_k	8.9205	3.9254					

Figure 12 – COS² des individus dans le premier plan factoriel - Feuille "individus"

Contribution des individus aux axes (CTR). La contribution s'appuie toujours sur les coordonnées factorielles, mais la normalisation est différente : $CTR_{ik} = \frac{F_{ik}^2}{n \times \lambda_k}$



I3 : =D3^2/($B$23*B$24)

	A	B	C	D	E	F	G	H	I	J
1		Matrice U (SVD)		Coordonnées			COS²		CTR	
2	Modele	F1	F2	F1	F2	o²_i	F1	F2	F1	F2
3	Alfasud TI	-0.2398	-0.4549	-2.1389	-1.7857	8.2252	55.6%	38.8%	5.7%	20.7%
4	Audi 100	0.1750	0.3890	1.5615	1.5270	6.6738	36.5%	34.9%	3.1%	15.1%
5	Simca 1300	-0.1255	0.1718	-1.1194	0.6745	2.1593	58.0%	21.1%	1.6%	3.0%
6	Citroen GS Club	-0.2885	-0.0288	-2.5737	-0.1129	6.7801	97.7%	0.2%	8.3%	0.1%
7	Fiat 132	0.0480	-0.1772	0.4279	-0.6956	1.1691	15.7%	41.4%	0.2%	3.1%
8	Lancia Beta	-0.0341	0.0500	-0.3042	0.1961	1.1350	8.2%	3.4%	0.1%	0.2%
9	Peugeot 504	0.0767	0.2377	0.6839	0.9331	1.5128	30.9%	57.5%	0.6%	5.6%
10	Renault 16 TL	-0.2184	0.2498	-1.9485	0.9804	5.6368	67.4%	17.1%	4.8%	6.2%
11	Renault 30	0.4943	-0.2710	4.4097	-1.0636	21.7897	89.2%	5.2%	24.4%	7.3%
12	Toyota Corolla	-0.4468	-0.0602	-3.9858	-0.2362	16.2901	97.5%	0.3%	20.0%	0.4%
13	Alfetta 1.66	0.0491	-0.4872	0.4377	-1.9124	4.4568	4.3%	82.1%	0.2%	23.7%
14	Princess 1800	0.1141	0.2144	1.0182	0.8417	1.9525	53.1%	36.3%	1.3%	4.6%
15	Datsun 200L	0.3297	0.1424	2.9411	0.5592	11.1126	77.8%	2.8%	10.9%	2.0%
16	Taunus 2000	0.1474	-0.1239	1.3149	-0.4865	2.4530	70.5%	9.6%	2.2%	1.5%
17	Rancho	-0.0775	0.2287	-0.6911	0.8977	1.9634	24.3%	41.0%	0.6%	5.2%
18	Mazda 9295	0.0432	-0.0907	0.3857	-0.3562	0.6845	21.7%	18.5%	0.2%	0.8%
19	Opel Rekord	0.2567	-0.0266	2.2898	-0.1043	6.0831	86.2%	0.2%	6.6%	0.1%
20	Lada 1300	-0.3036	0.0366	-2.7086	0.1437	7.9222	92.6%	0.3%	9.2%	0.1%
21										
22	d_k	8.9205	3.9254							
23	n	18								
24	lambda_k	4.4209	0.8561							

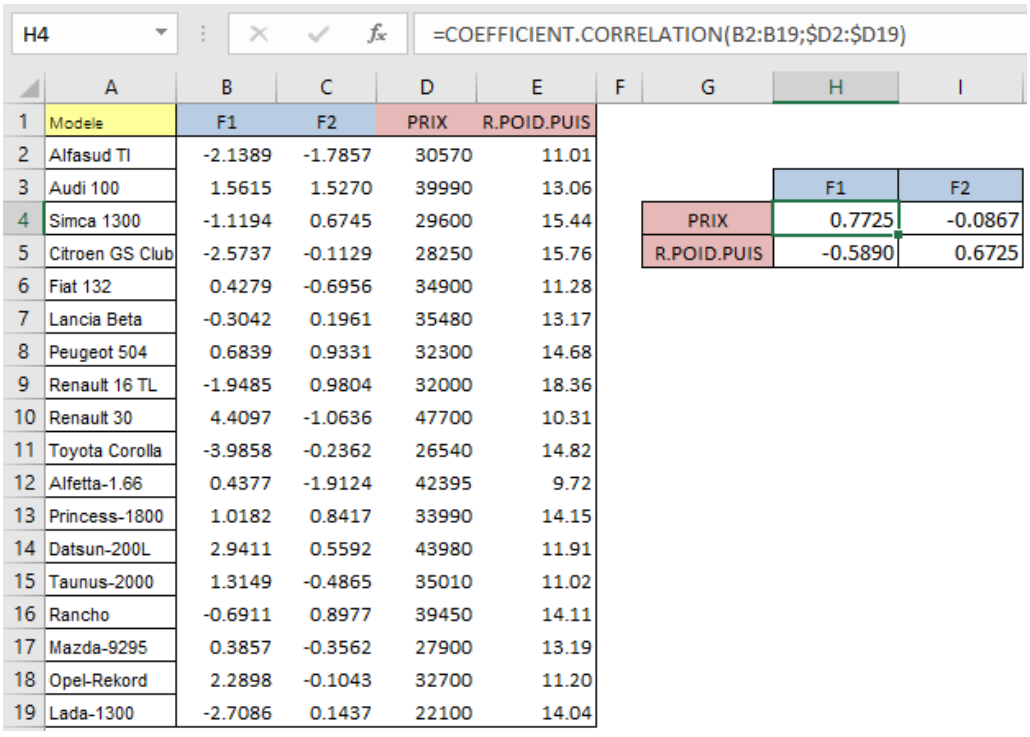
Figure 13 - CTR des individus aux facteurs - Feuille "individus"

Sans surprise, la Renault 30 et la Toyota Corolla, situés aux deux extrémités, sont déterminants pour le premier facteur. Le second, lui, repose surtout sur l'Alfetta 1.66 et l'Alfasud TI (ouh là là, c'étaient des bonnes voitures ça, elles avaient du caractère !).

4.6 Traitement des variables illustratives

Nous souhaitons renforcer l'interprétation des facteurs à l'aide de variables qui n'ont pas participé à l'étude (RAK, 2013 ; pages 32 et suivantes). Nous avons besoin des coordonnées des individus pour positionner ces variables dites « illustratives ».

Illustratives quantitatives. Nous créons une nouvelle feuille Excel « var.illus.quanti » dans lequel nous reportons les coordonnées factorielles des individus dans le plan (F1, F2) et les variables additionnelles PRIX et R.POIDS.PUIS (rapport poids-puissance). Nous calculons ensuite les coefficients de corrélation linéaire entre les facteurs, d'une part, et les variables, d'autre part.



H4 : =COEFFICIENT.CORRELATION(B2:B19;$D2:$D19)

	A	B	C	D	E
1	Modele	F1	F2	PRIX	R.POID.PUIS
2	Alfasud TI	-2.1389	-1.7857	30570	11.01
3	Audi 100	1.5615	1.5270	39990	13.06
4	Simca 1300	-1.1194	0.6745	29600	15.44
5	Citroen GS Club	-2.5737	-0.1129	28250	15.76
6	Fiat 132	0.4279	-0.6956	34900	11.28
7	Lancia Beta	-0.3042	0.1961	35480	13.17
8	Peugeot 504	0.6839	0.9331	32300	14.68
9	Renault 16 TL	-1.9485	0.9804	32000	18.36
10	Renault 30	4.4097	-1.0636	47700	10.31
11	Toyota Corolla	-3.9858	-0.2362	26540	14.82
12	Alfetta-1.66	0.4377	-1.9124	42395	9.72
13	Princess-1800	1.0182	0.8417	33990	14.15
14	Datsun-200L	2.9411	0.5592	43980	11.91
15	Taunus-2000	1.3149	-0.4865	35010	11.02
16	Rancho	-0.6911	0.8977	39450	14.11
17	Mazda-9295	0.3857	-0.3562	27900	13.19
18	Opel-Rekord	2.2898	-0.1043	32700	11.20
19	Lada-1300	-2.7086	0.1437	22100	14.04

	F1	F2
PRIX	0.7725	-0.0867
R.POID.PUIS	-0.5890	0.6725

Figure 14 - Positionnement des variables illustratives quantitatives - Feuille "var.illus.quanti"

Nous utilisons la fonction COEFFICIENT.CORRELATION() d'Excel. On peut lire par exemple que le PRIX est fortement lié au premier axe (corrélation = 0.7725). Ce dernier induit une différenciation des véhicules selon le prix.

Illustratives qualitatives. Les moyennes des facteurs conditionnellement aux modalités des variables illustratives font l'affaire dans le cas des variables qualitatives. Pour nos données, nous essayons de qualifier les facteurs à l'aide des finitions (FINITION) des véhicules (Moyen, Bonne, Très Bonne).

	A	B	C	D	E	F	G	H
1	Modele	F1	F2	FINITION		Étiquettes	Moyenne de F1	Moyenne de F2
2	Alfasud TI	-2.1389	-1.7857	2_B		1_M	-2.0004	0.0226
3	Audi 100	1.5615	1.5270	3_TB		2_B	0.2353	-0.0453
4	Simca 1300	-1.1194	0.6745	1_M		3_TB	1.3924	0.0340
5	Citroen GS Club	-2.5737	-0.1129	1_M				
6	Fiat 132	0.4279	-0.6956	2_B				
7	Lancia Beta	-0.3042	0.1961	3_TB				
8	Peugeot 504	0.6839	0.9331	2_B				
9	Renault 16 TL	-1.9485	0.9804	2_B				
10	Renault 30	4.4097	-1.0636	3_TB				
11	Toyota Corolla	-3.9858	-0.2362	1_M				
12	Alfetta-1.66	0.4377	-1.9124	3_TB				
13	Princess-1800	1.0182	0.8417	2_B				
14	Datsun-200L	2.9411	0.5592	3_TB				
15	Taunus-2000	1.3149	-0.4865	2_B				
16	Rancho	-0.6911	0.8977	3_TB				
17	Mazda-9295	0.3857	-0.3562	1_M				
18	Opel-Rekord	2.2898	-0.1043	2_B				
19	Lada-1300	-2.7086	0.1437	1_M				

Figure 15 - Moyenne des facteurs conditionnellement à FINITION - Feuille "var.illus.quali"

Nous avons utilisé un tableau croisé dynamique pour obtenir les moyennes conditionnelles. A l'évidence, la différenciation sur les finitions est une lecture possible du premier axe factoriel : de gauche à droite, les véhicules ont un niveau de finition croissant sur le premier axe, avec des écarts marqués (plus marqués en tous les cas que sur le second facteur).

4.7 Traitement des individus illustratifs

Nous avons besoin des coordonnées des variables pour positionner les individus illustratifs dans le plan factoriel (F1, F2), plus précisément des coefficients de la matrice V de la décomposition en valeurs singulières (Figure 4) (RAK, 2013 ; pages 36 et suivantes). Nous souhaitons situer 2 nouvelles Peugeot par rapport aux véhicules de notre fichier initial :

Modele	CYL	PUISS	LONG	LARG	POIDS	V.MAX
Peugeot 604	2664	136	472	177	1410	180
Peugeot 304 S	1288	74	414	157	915	160

Figure 16 - Caractéristiques des individus illustratifs - Feuille "ind.illustratifs"

Plusieurs étapes sont nécessaires. Il faut tout d'abord centrer et réduire les descriptions, mais en utilisant les moyennes et écarts-type calculés sur nos données initiales (Figure 2). Pour un nouvel individu i*, nous transformons ses coordonnées (x_{i^*j}) à l'aide de : $z_{i^*j} = \frac{x_{i^*j} - \bar{x}_j}{\sigma_j}$

Nous récupérons en (B5:G6) les moyennes et écarts-type calculés dans la feuille « acp-svd » (copier – collage spécial valeurs). Nous appliquons la formule de transformation en (B10:G11) :

B10 : =(B2-B$5)/B$6

	A	B	C	D	E	F	G
1	Modele	CYL	PUISS	LONG	LARG	POIDS	V.MAX
2	Peugeot 604	2664	136	472	177	1410	180
3	Peugeot 304 S	1288	74	414	157	915	160
4							
5	Moyenne	1631.67	84.61	433.50	166.67	1078.83	158.28
6	Ecart-type	363.39	19.80	21.48	5.16	133.10	11.80
7							
8							
9	Val. CR	zCYL	zPUISS	zLONG	zLARG	zPOIDS	zV.MAX
10	Peugeot 604	2.8408	2.5951	1.7920	2.0010	2.4881	1.8411
11	Peugeot 304 S	-0.9457	-0.5359	-0.9076	-1.8719	-1.2309	0.1460

Figure 17 - Coordonnées centrées et réduites des individus illustratifs - Feuille "ind.illustratifs"

Puis, après avoir copié en (J2:K7) les coefficients de la matrice V en provenance la feuille « variables » (Figure 4). Nous appliquons la formule :

$$F_{i^*k} = \sum_{j=1}^{p=6} z_{i^*j} \times v_{jk}$$

Pour la Peugeot 604, elle se traduit en J10 par {=CNUM(xMatMult($B10:$G10;J$2:J$7))}.

J10 : {=CNUM(xMatMult($B10:$G10;J$2:J$7))}

	A	B	C	D	E	F	G	H	I	J	K
1	Modele	CYL	PUISS	LONG	LARG	POIDS	V.MAX			F1	F2
2	Peugeot 604	2664	136	472	177	1410	180		CYL	0.4249	-0.1242
3	Peugeot 304 S	1288	74	414	157	915	160		PUISS	0.4218	-0.4158
4									LONG	0.4215	0.4118
5	Moyenne	1631.67	84.61	433.50	166.67	1078.83	158.28		LARG	0.3869	0.4461
6	Ecart-type	363.39	19.80	21.48	5.16	133.10	11.80		POIDS	0.4305	0.2427
7									V.MAX	0.3589	-0.6199
8											
9	Val. CR	zCYL	zPUISS	zLONG	zLARG	zPOIDS	zV.MAX		Modele	F1	F2
10	Peugeot 604	2.8408	2.5951	1.7920	2.0010	2.4881	1.8411		Peugeot 604	5.5633	-0.3386
11	Peugeot 304 S	-0.9457	-0.5359	-0.9076	-1.8719	-1.2309	0.1460		Peugeot 304 S	-2.2122	-1.2578

Figure 18 - Coordonnées factorielles des individus illustratifs - Feuille "ind.illustratifs"

xMatMult() de la librairie XNUMBERS est l'équivalent de PRODUITMAT() d'Excel.

Nous pouvons distinguer les Peugeot dans le nuage de points des individus du premier plan factoriel.

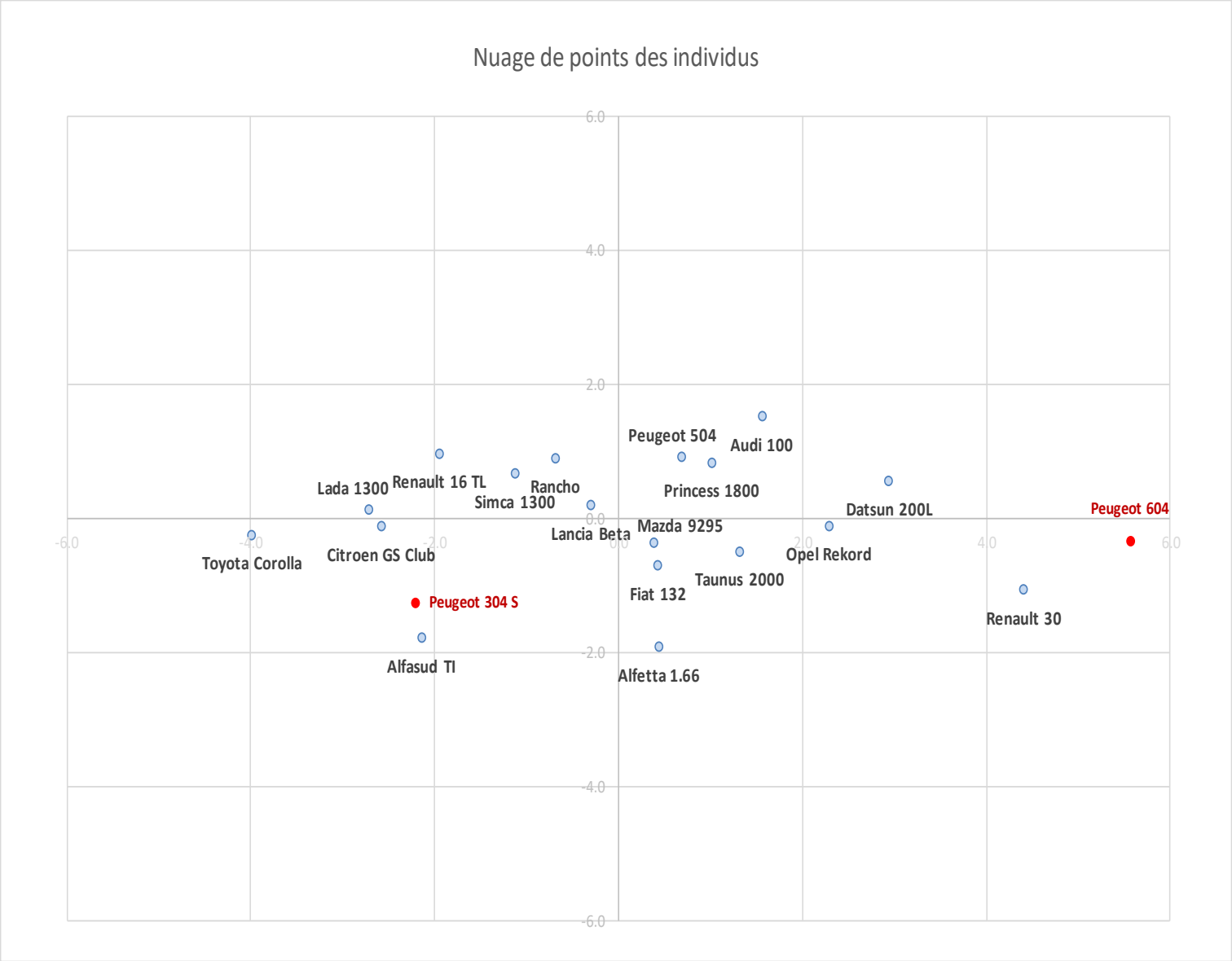


Figure 19 - Position des individus illustratifs dans le plan factoriel - Feuille "individus - Plan factoriel (2)"

La Peugeot 604 est plutôt un véhicule statutaire, proche de la Renault 30 ; la Peugeot 304 S (le S est très important) est une compacte sportive, similaire à l'Alfasud TI (c'était vraiment une bonne voiture). C'est ce que nous dit le graphique factoriel en tous les cas.

5 Conclusion

Le but premier de ce tutoriel était pédagogique : décortiquer pas à pas la mécanique de l'ACP en reproduisant les formules sous le tableur Excel. Les calculs ne sont pas sorciers finalement lorsqu'on les regarde de plus près. Nous avons pu réaliser une étude complète relativement simplement.

C'était aussi pour moi l'occasion de mettre en avant la librairie XNUMBERS que j'avais découverte en lisant un ouvrage sur le calcul scientifique sous Excel (de Levie, 2008). Elle est particulièrement puissante et précise. Nous avons ainsi retrouvé avec un niveau de qualité largement suffisant les principaux résultats présentés dans mon support dédié à l'ACP (RAK, 2013), où j'avais utilisé les logiciels phares de la statistique sur les mêmes données.

6 Références

John Beyers, « Xnumbers – Version 6.0 » (http://www.thetropicalevents.com/Xnumbers60.htm).

(RAK, 2013) Ricco Rakotomalala, « Analyse en composantes principales – Diapos », Juillet 2013.

Robert de Levie, « Advanced Excel for scientific data analysis », Oxford University Press, 2008.

Wikipedia, « XNUMBERS ».