Objectif

Mettre en œuvre l'analyse en composantes principales avec TANAGRA.

L'analyse en composantes principales est une technique de visualisation très populaire en analyse de données. Dans ce tutoriel, nous montrons comment la mettre en œuvre avec TANAGRA.

Fichier de données

Nous utilisons le fichier AUTOS_ACP.XLS tiré de l'ouvrage de SAPORTA[1] (Tableau 17.1, page 428). L'intérêt de ce fichier est que nous pouvons comparer directement nos résultats avec ceux du livre (pages 177 à 181). Nous nous contentons de montrer l'enchaînement des opérations et la lecture des tableaux de résultats dans ce tutoriel. Pour ce qui de l'interprétation, le mieux est de se référer à l'ouvrage.

Le tableau de données est le suivant :

Modele	CYL	PUISS	LONG	LARG	POIDS	V-MAX	FINITION	PRIX	R-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
Renault 16 TL	1565	55	424	163	1010	140	B	32000	18.36
Renault 30	2664	128	452	173	1320	180	TB	47700	10.31
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

La première colonne correspond à l'identifiant des observations, les variables actives sont en vert, les variables illustratives en bleu. Par rapport au fichier original, nous avons ajouté la variable R-POID.PUIS (rapport poids-puissance) qui indique la « vivacité » (sportivité) d'un modèle : plus faible est sa valeur, plus sportif est le véhicule.

Analyse en Composantes Principales avec TANAGRA

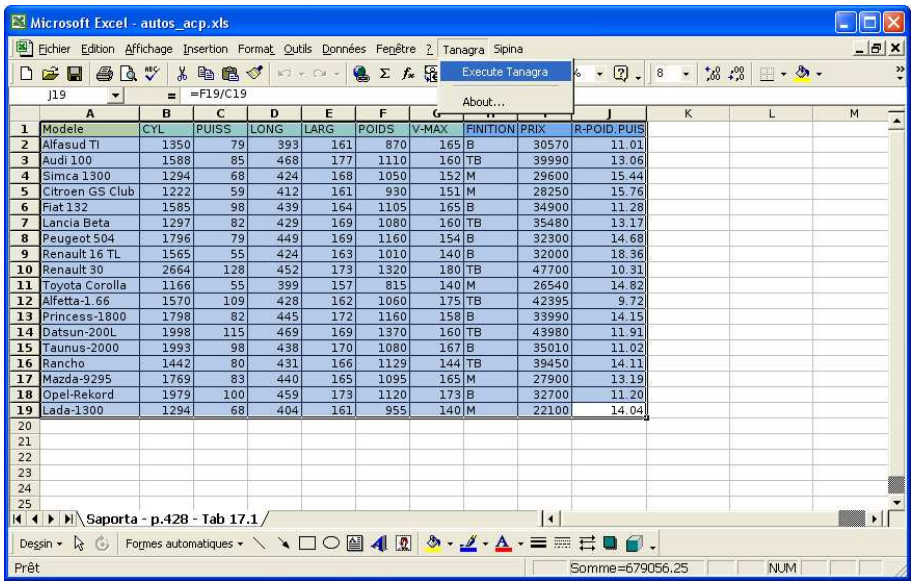
Créer un diagramme

Depuis la version 1.4.11, il est possible de démarrer TANAGRA à partir du tableur EXCEL[2]. C'est la procédure que nous choisissons ici : le diagramme est automatiquement créé, et les données importées.

[1] G. SAPORTA, « Probabilités, Analyse de données et Statistique », TECHNIP, 2006. C'est l'édition la plus récente du fameux ouvrage qui, depuis plusieurs dizaines d'années, fait référence en France en matière de traitement exploratoire de données.

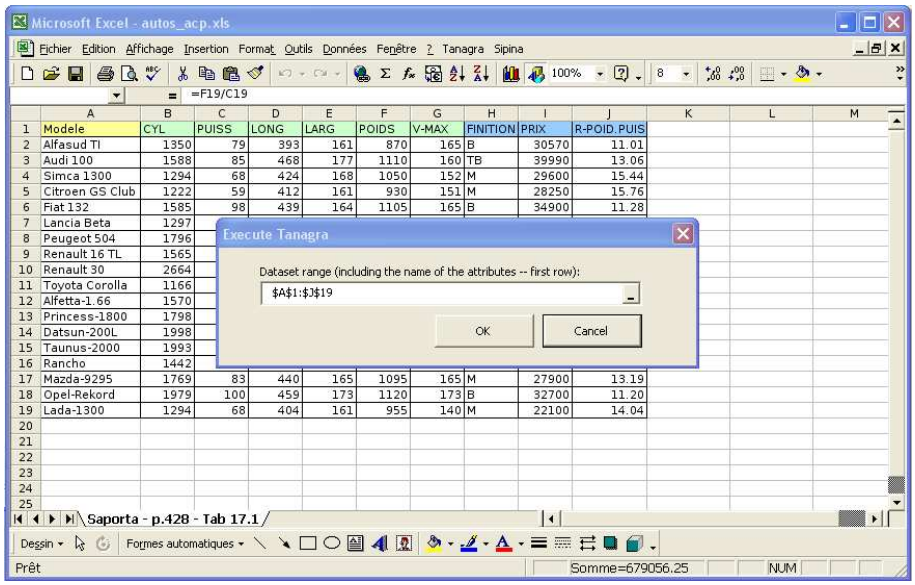
[2] Il faut bien entendu avoir référencé la macro-complémentaire (Add-In) TANAGRA dans EXCEL, voir le didacticiel adéquat sur le site web. La démarche est également valable avec le tableur CALC de OPEN OFFICE.

Pour ce faire, nous sélectionnons la plage de données dans EXCEL, puis nous cliquons sur le menu TANAGRA/ EXECUTE TANAGRA.



	A	B	C	D	E	F	G	H	I	J
1	Modele	CYL	PUISS	LONG	LARG	POIDS	V-MAX	FINITION	PRIX	R-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	Renault 16 TL	1565	55	424	163	1010	140	B	32000	18.36
10	Renault 30	2664	128	452	173	1320	180	TB	47700	10.31
11	Toyota Corolla	1166	55	399	157	815	140	M	26540	14.82
12	Alfetta-1.66	1570	109	428	162	1060	175	TB	42395	9.72
13	Princess-1800	1798	82	445	172	1160	158	B	33990	14.15
14	Datsun-200L	1998	115	469	169	1370	160	TB	43980	11.91
15	Taunus-2000	1993	98	438	170	1080	167	B	35010	11.02
16	Rancho	1442	80	431	166	1129	144	TB	39450	14.11
17	Mazda-9295	1769	83	440	165	1095	165	M	27900	13.19
18	Opel-Rekord	1979	100	459	173	1120	173	B	32700	11.20
19	Lada-1300	1294	68	404	161	955	140	M	22100	14.04

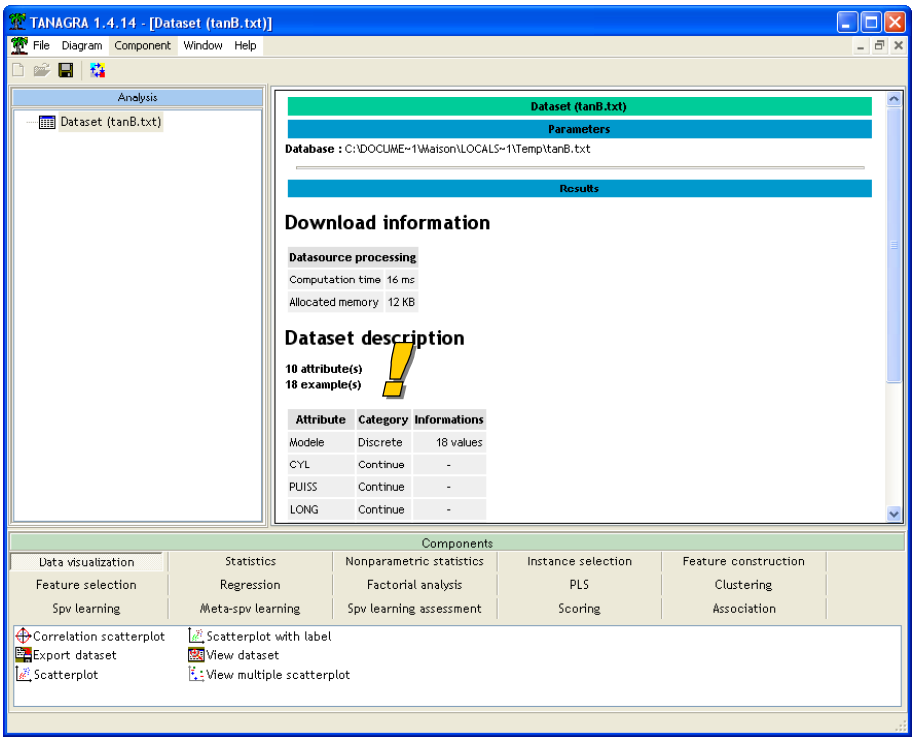
Une boîte de dialogue vient confirmer la sélection en affichant les références de la plage de cellules. Nous validons en cliquant sur OK.



	A	B	C	D	E	F	G	H	I	J
1	Modele	CYL	PUISS	LONG	LARG	POIDS	V-MAX	FINITION	PRIX	R-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								
8	Peugeot 504	1796								
9	Renault 16 TL	1565								
10	Renault 30	2664								
11	Toyota Corolla	1166								
12	Alfetta-1.66	1570								
13	Princess-1800	1798								
14	Datsun-200L	1998								
15	Taunus-2000	1993								
16	Rancho	1442								
17	Mazda-9295	1769	83	440	165	1095	165	M	27900	13.19
18	Opel-Rekord	1979	100	459	173	1120	173	B	32700	11.20
19	Lada-1300	1294	68	404	161	955	140	M	22100	14.04

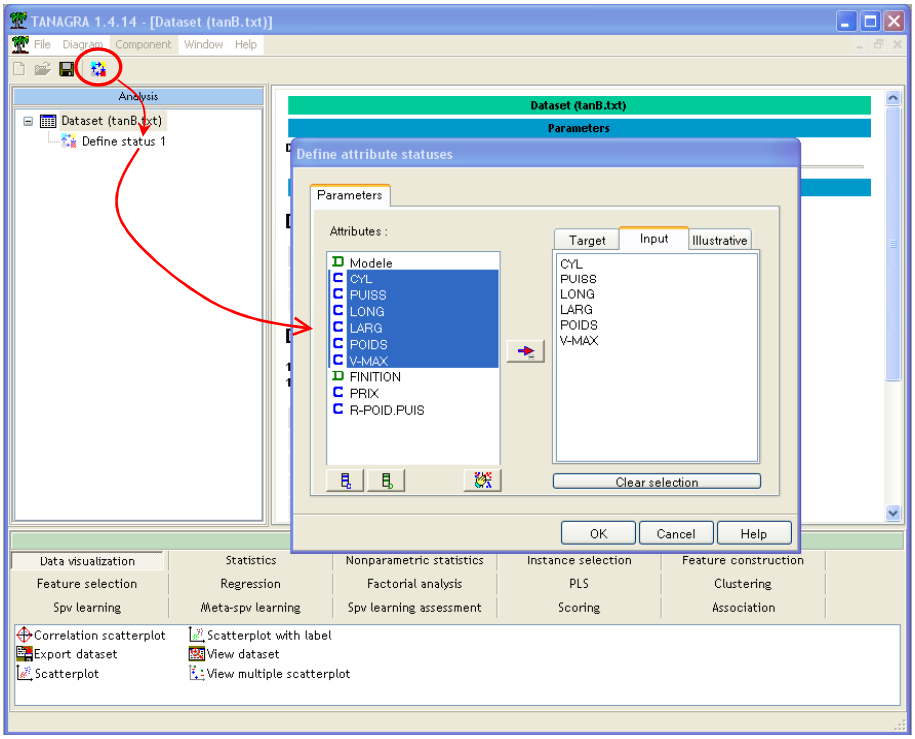
TANAGRA est alors démarré, nous vérifions que l'ensemble de données comprend bien 18 observations et 10 variables[3].

[3] Le type identifiant n'est pas reconnu par TANAGRA, la première colonne est donc codée comme une variable discrète, avec autant de modalités que d'observations. Cela ne pose pas de problème dans la pratique... sauf si nous traitons plus de 255 observations. Mais l'identifiant n'est vraiment utile que si l'on veut produire les plans factoriels en étiquetant les points. Si les observations sont nombreuses, ce graphique ne sera pas très lisible de toute manière, les étiquettes se superposant les uns sur les autres.

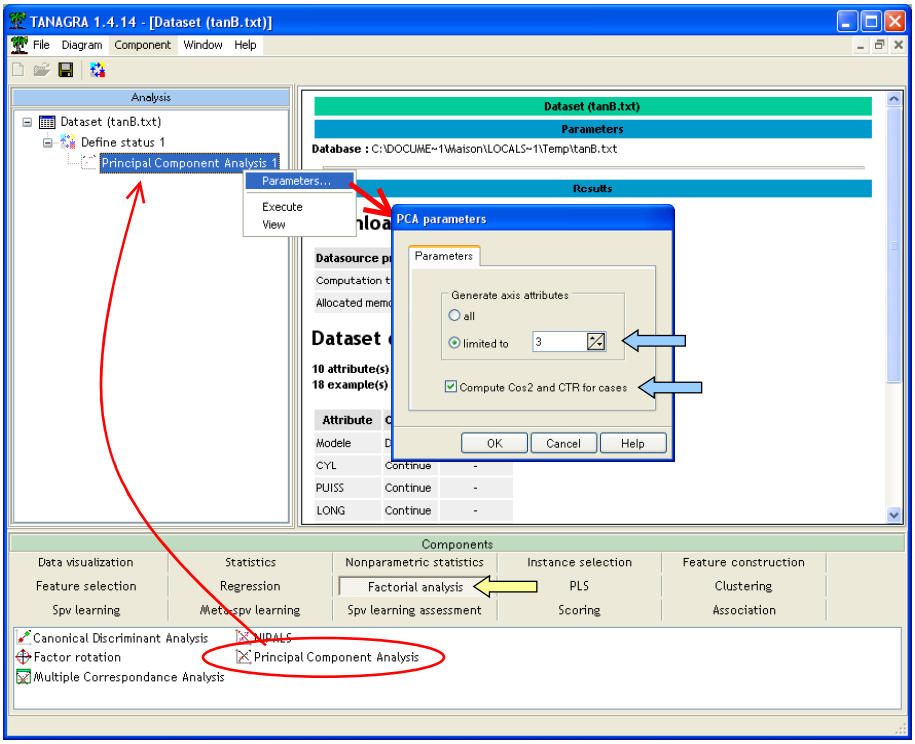


ACP

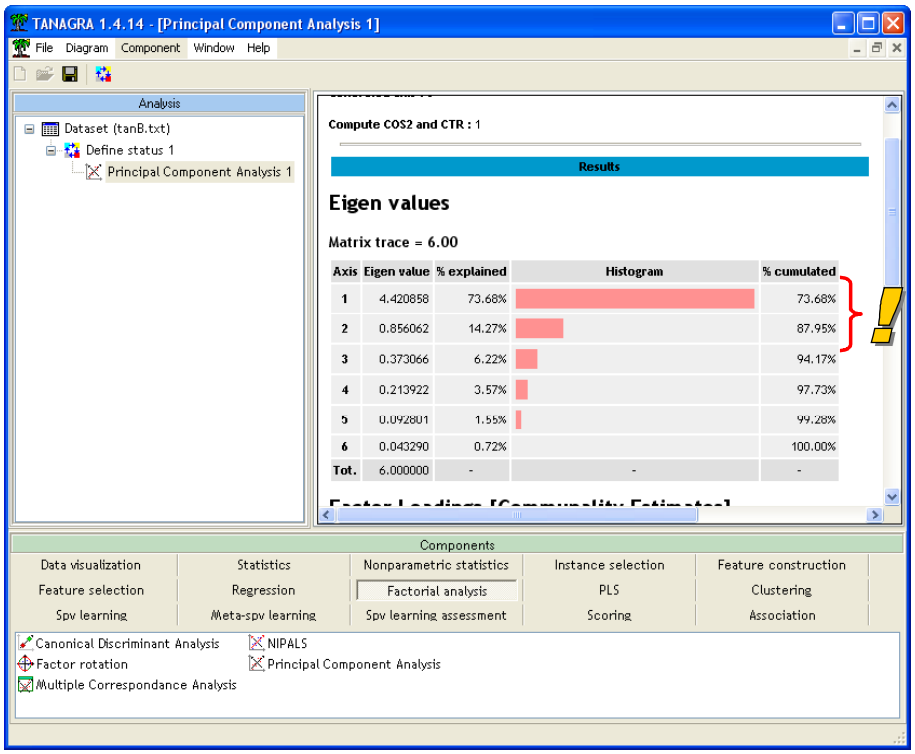
Pour initier une analyse, nous devons tout d'abord définir le rôle des variables. C'est le rôle du composant DEFINE STATUS accessible dans la barre d'outil. Nous mettons en INPUT les variables actives. Nous verrons plus tard comment utiliser les variables illustratives.



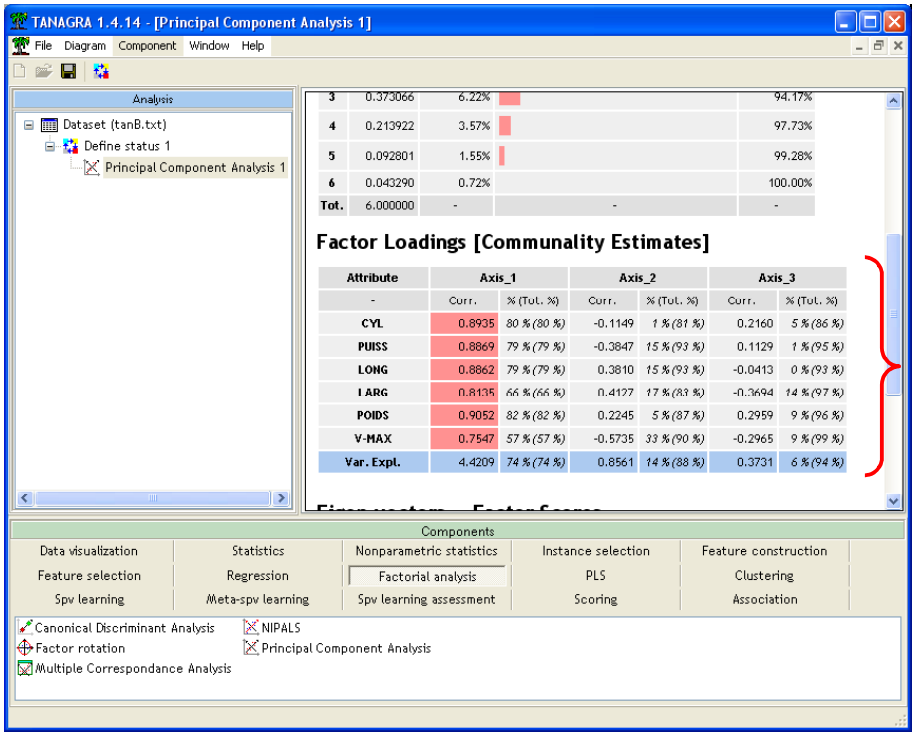
Puis nous plaçons le composant PRINCIPAL COMPONENT ANALYSIS (onglet FACTORIAL ANALYSIS) dans le diagramme de traitements. Nous cliquons sur le menu PARAMETERS : nous spécifions alors le nombre d'axes à produire (**3**) ; nous activons l'option qui permet de calculer les COS2 et les contributions pour chaque individu.



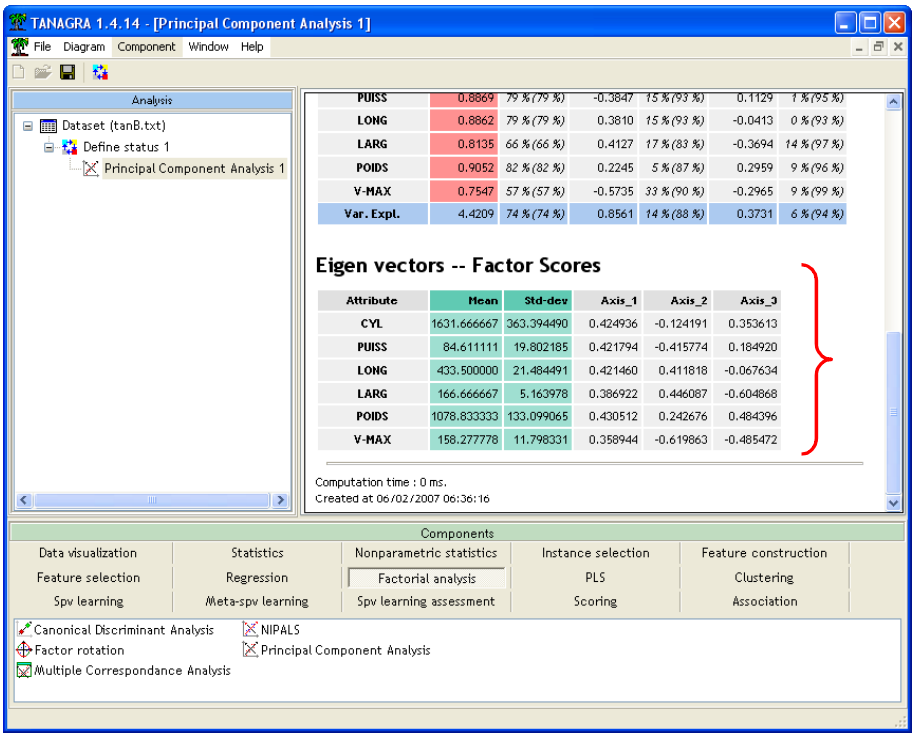
Valeurs propres. Nous activons le menu contextuel VIEW pour accéder aux résultats. La première partie des résultats correspond au tableau des valeurs propres. Nous constatons que les deux premiers axes restituent 87.95% de l'information disponible (p.178).



Corrélations variables / axes. La seconde partie des résultats indique la corrélation et le COS^2 -- en % et % cumulé -- des variables avec les axes factoriels (p.178).

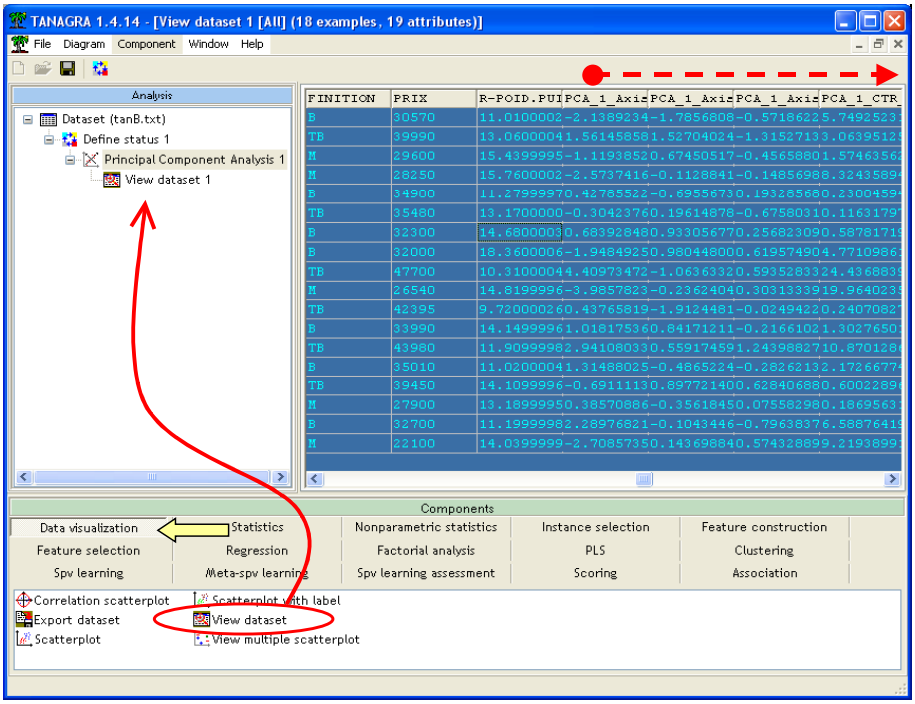


Coordonnées factorielles : projection, COS² et contribution. Un troisième tableau propose les paramètres des équations de projections pour chaque axe. Il est bien entendu nécessaire de centrer et réduire les données avec les moyennes et écarts-types affichés avant d'appliquer les coefficients.

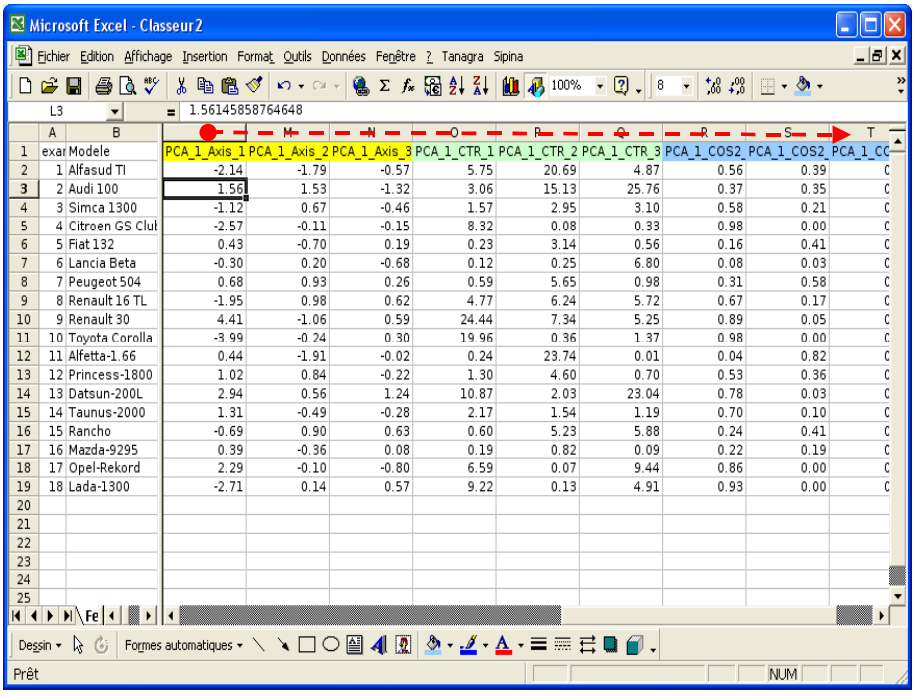


Mais il est possible d'accéder directement aux données calculées, c.-à-d. les projections dans le nouvel espace. En effet, le composant ACP rajoute automatiquement une série de variables à l'ensemble de données. Il s'agit, pour chaque individu et pour chaque axe demandé, des projections sur les axes, des contributions et des COS².

Pour visualiser le tableau de données associé, nous plaçons dans le diagramme le composant VIEW DATASET (onglet DATA VISUALIZATION). Nous cliquons sur le menu VIEW. Dans la dernière partie de la grille apparaissent les nouvelles colonnes.

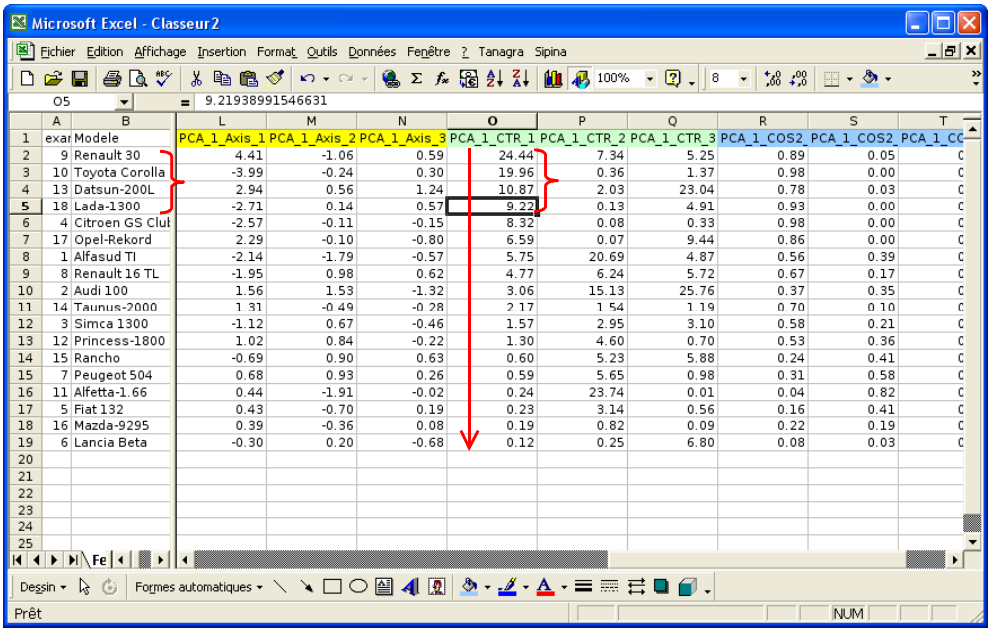


TANAGRA adopte le format scientifique. Cela peut gêner la lecture. Une manière simple de s'en sortir est de copier (menu COMPONENT / COPY RESULTS) et de coller les données de la grille dans le tableur de votre choix. En adaptant la précision à notre convenance, nous obtenons le tableau p.180.



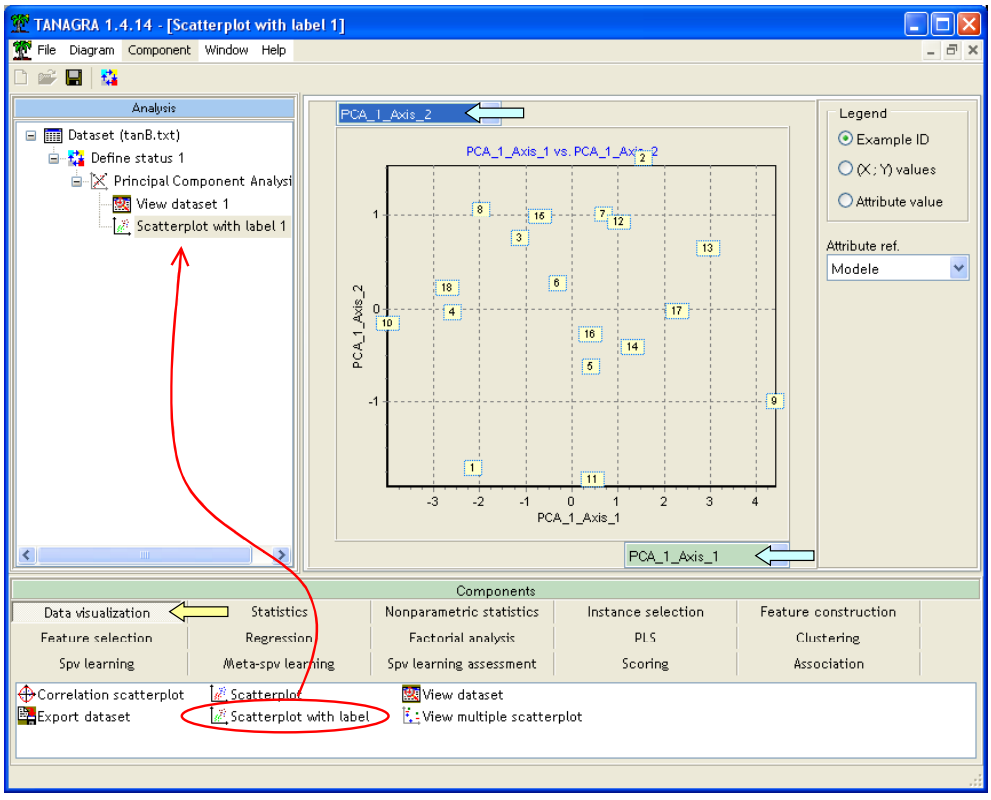
L'intérêt du tableur est manifeste ici. En effet, nous disposons de multiples possibilités de tri qui permettent de mettre en évidence les informations pertinentes. Par exemple, si nous trions de manière décroissante sur les contributions des individus au premier axe, nous constatons que le

facteur est en grande partie déterminé par l'opposition RENAULT 30 TS + DATSUN-200 L (voitures « imposantes ») et TOYOTA COROLLA + LADA-1300 (« petites » voitures).

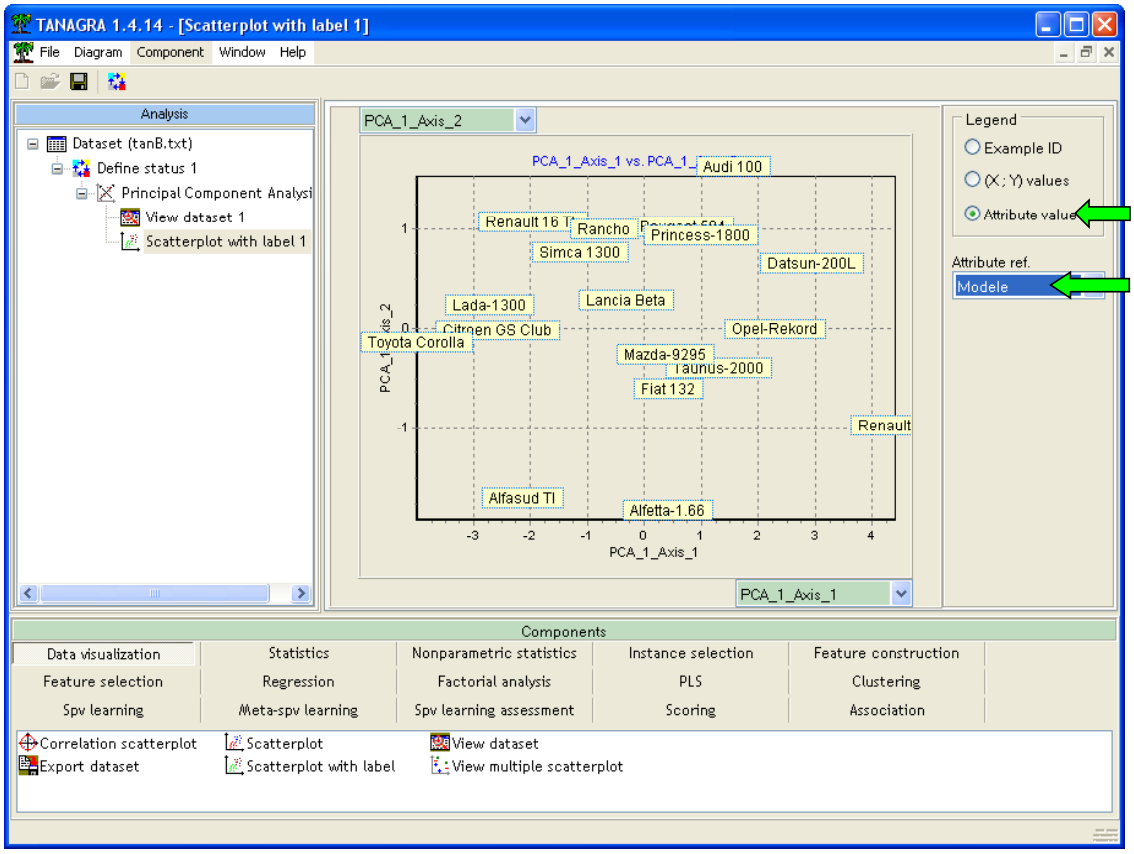


	A	B	L	M	N	O	P	Q	R	S	T
1	exar	Modele	PCA_1_Axis_1	PCA_1_Axis_2	PCA_1_Axis_3	PCA_1_CTR_1	PCA_1_CTR_2	PCA_1_CTR_3	PCA_1_COS2_	PCA_1_COS2_	PCA_1_CC
2	9	Renault 30	4.41	-1.06	0.59	24.44	7.34	5.25	0.89	0.05	0
3	10	Toyota Corolla	-3.99	-0.24	0.30	19.96	0.36	1.37	0.98	0.00	0
4	13	Datsun-200L	2.94	0.56	1.24	10.87	2.03	23.04	0.78	0.03	0
5	18	Lada-1300	-2.71	0.14	0.57	9.23	0.13	4.91	0.93	0.00	0
6	4	Citroen GS Clul	-2.57	-0.11	-0.15	8.32	0.08	0.33	0.98	0.00	0
7	17	Opel-Rekord	2.29	-0.10	-0.80	6.59	0.07	9.44	0.86	0.00	0
8	1	Alfasud TI	-2.14	-1.79	-0.57	5.75	20.69	4.87	0.56	0.39	0
9	8	Renault 16 TL	-1.95	0.98	0.62	4.77	6.24	5.72	0.67	0.17	0
10	2	Audi 100	1.56	1.53	-1.32	3.06	15.13	25.76	0.37	0.35	0
11	14	Taunus-2000	1.31	-0.49	-0.28	2.17	1.54	1.19	0.70	0.10	0
12	3	Simca 1300	-1.12	0.67	-0.46	1.57	2.95	3.10	0.58	0.21	0
13	12	Princess-1800	1.02	0.84	-0.22	1.30	4.60	0.70	0.53	0.36	0
14	15	Rancho	-0.69	0.90	0.63	0.60	5.23	5.88	0.24	0.41	0
15	7	Peugeot 504	0.68	0.93	0.26	0.59	5.65	0.98	0.31	0.58	0
16	11	Alfetta-1.66	0.44	-1.91	-0.02	0.24	23.74	0.01	0.04	0.82	0
17	5	Fiat 132	0.43	-0.70	0.19	0.23	3.14	0.56	0.16	0.41	0
18	16	Mazda-9295	0.39	-0.36	0.08	0.19	0.82	0.09	0.22	0.19	0
19	6	Lancia Beta	-0.30	0.20	-0.68	0.12	0.25	6.80	0.08	0.03	0

Plans factoriels. La popularité de l'ACP repose en grande partie sur les représentations graphiques qu'elle propose. Elles nous permettent d'apprécier visuellement les proximités entre les observations. Dans notre cas, nous projetons les observations dans le premier plan factoriel. Nous voulons associer les identifiants aux points. Nous utilisons pour cela le composant SCATTERPLOT WITH LABEL (onglet DATA VISUALIZATION) que nous plaçons en dessous de l'ACP. Nous le paramétrons de manière à avoir en abscisse le premier facteur, en ordonnée le second facteur. Notons qu'il est très aisé de passer d'un plan factoriel à un autre.



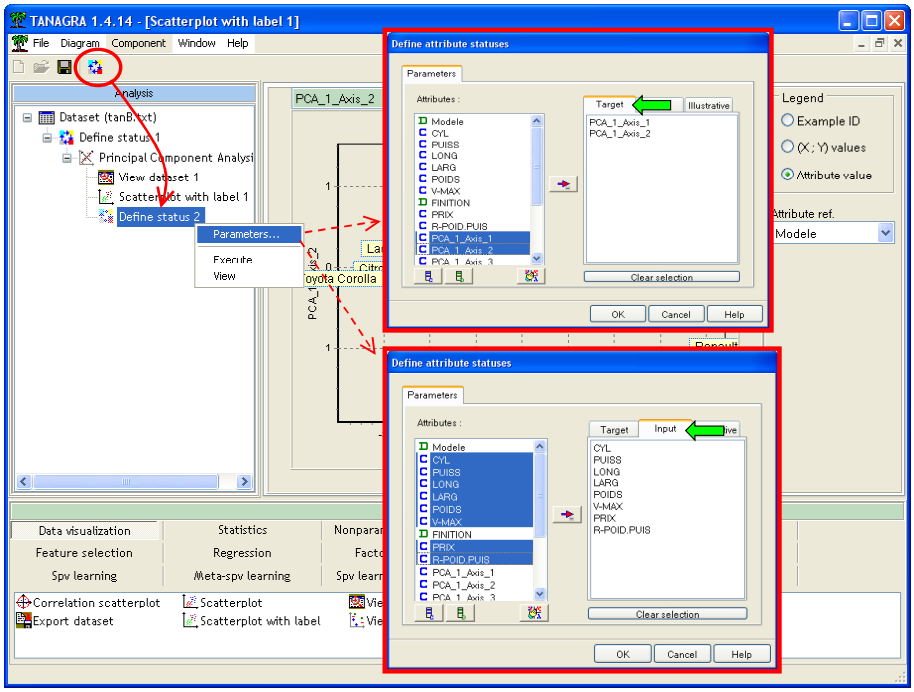
Au départ, les points sont identifiés par leur numéro. En activant l'option LEGEND / ATTRIBUTE VALUE, et en choisissant la référence MODELE, nous obtenons la carte des points étiquetés par leurs identifiants (p.181). Bien entendu, cette option est pratique tant que le nombre de points reste raisonnable. Au-delà d'un certain nombre d'observations, le graphique serait illisible.



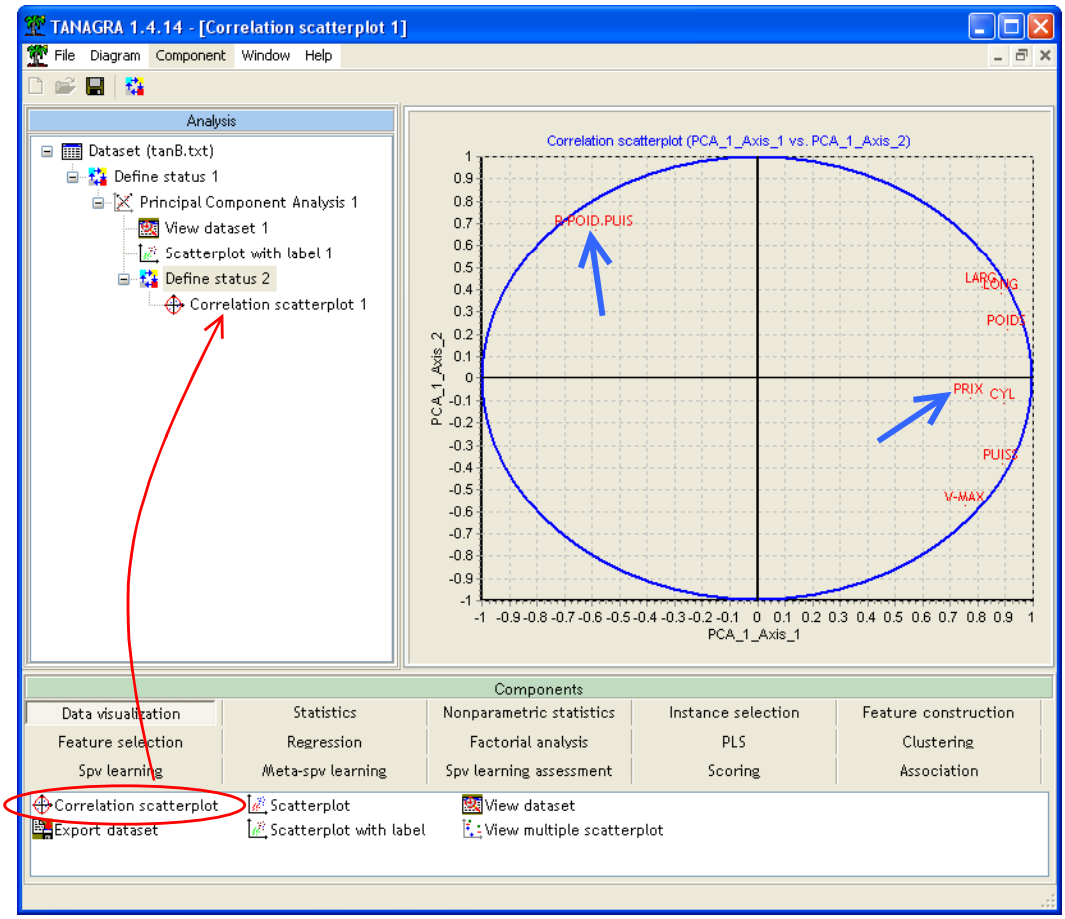
Il est possible de modifier la taille des étiquettes avec les raccourcis CTRL+Q et CTRL+W.

Cercle des corrélations et variables illustratives quantitatives. Le cercle de corrélations est un outil graphique qui permet de comprendre la nature des axes. Il sert à interpréter les axes. Il est généralement dévolu au positionnement des variables actives. Mais nous pouvons lui associer également les variables illustratives pour en préciser l'interprétation.

Nous procédons en deux temps. Tout d'abord nous ajoutons le composant DEFINE STATUS en dessous de l'ACP. Nous définissons les deux premiers axes comme TARGET, les variables actives et les variables illustratives sont placées en INPUT.



Dans un deuxième temps, nous ajoutons le composant CORRELATION SCATTERPLOT dans le diagramme. Nous obtenons le cercle des corrélations.



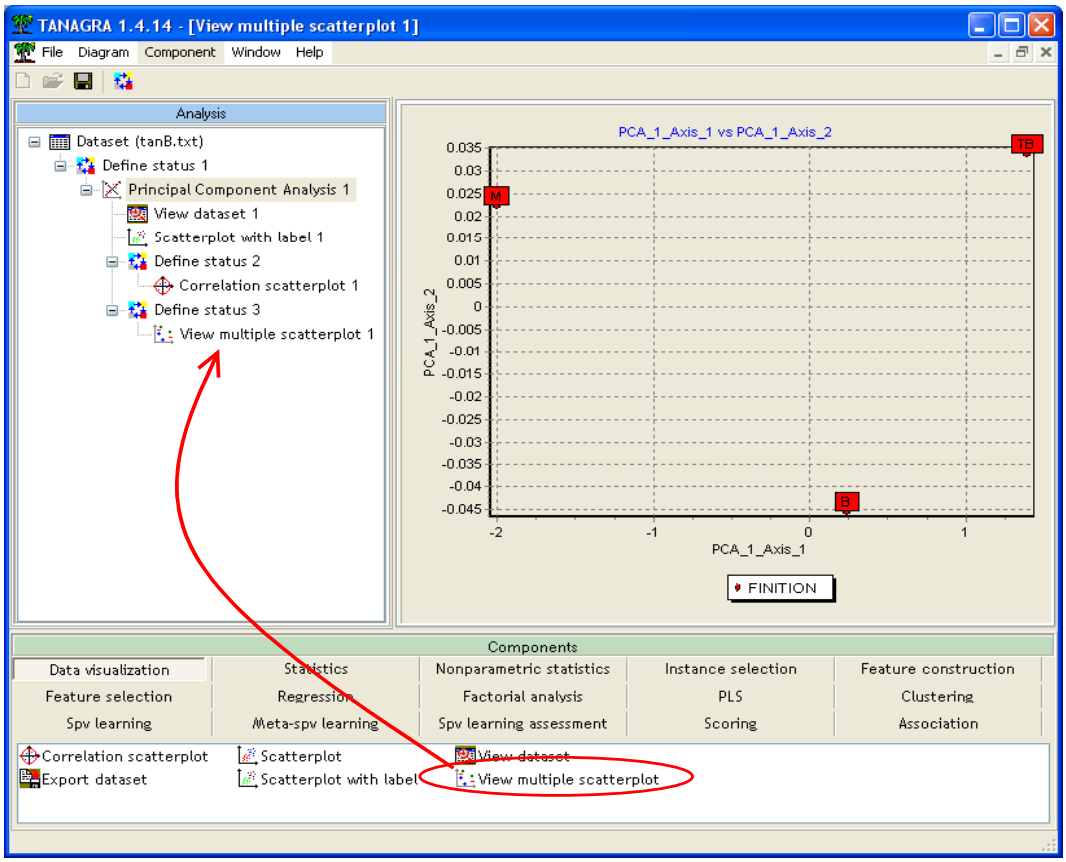
Le premier axe traduit bien un effet de taille, il oppose les voitures imposantes aux petites voitures selon les critères de largeur, longueur, poids, puissance, ... et de prix, puisque cette dernière est corrélée positivement avec le premier axe (p.179).

Le second axe était plus associé à la « sportivité » des véhicules. Les voitures italiennes ALFASUD et ALFETTA étant opposées à l'AUDI 100L (cf. les contributions sur le second axe). Cette interprétation est confortée par la forte corrélation négative avec le ratio RAPPORT POIDS-PUISSANCE. Rappelons qu'une valeur faible du ratio indique une voiture nerveuse.

Variables illustratives qualitatives. Autre possibilité intéressante, nous voulons positionner les véhicules selon leur qualité de finition, décrite par une variable à 3 modalités (Bien, Très Bien et Moyenne). La variable étant qualitative, il n'est possible d'utiliser le cercle de corrélations. Nous disposons d'un autre outil, le composant VIEW MULTIPLE SCATTERPLOT.

Tout comme le composant cercle de corrélations, nous agissons en deux temps. Tout d'abord nous plaçons le composant DEFINE STATUS dans le diagramme, nous plaçons les deux axes factoriels en TARGET. Puis, nous mettons en INPUT la ou les variables illustratives qualitatives.

Dans un second temps, nous insérons le composant VIEW MULTIPLE SCATTERPLOT. Il propose une représentation des modalités de chaque variable illustrative en calculant les moyennes conditionnelles associées aux modalités.



Nous observons que la qualité de finition s'étire sur le premier axe factoriel. Les petits véhicules sont de qualité moyenne, les voitures imposantes sont de meilleure qualité (p.181).

Individus illustratifs. Nous ne l'avons pas mis en œuvre dans ce tutoriel, mais il est tout à fait possible de travailler avec les données illustratives en analyse factorielle. Le plus simple serait alors de définir une variable indicatrice dans une nouvelle colonne, avec deux modalités (« actif » et « illustratif ») ; puis d'utiliser les composants INSTANCE SELECTION pour sélectionner les individus actifs (DISCRETE SELECT EXAMPLES ou RULE BASED SELECTION) lors de la construction des axes ; et enfin, de projeter la totalité des observations, ou uniquement les individus illustratifs, dans les différents plans factoriels en définissant le sous-ensemble adéquat avec le composant RECOVER EXAMPLES.

Conclusion

TANAGRA ne prétend pas fournir des outils de reporting et de déploiement à la hauteur des logiciels commerciaux. En se contentant de proposer des résultats standards, repris dans des ouvrages qui font référence, nous essayons de donner aux utilisateurs les principaux codes de lecture d'une analyse factorielle.

Pouvoir reprendre les résultats dans un tableur est certainement une des fonctionnalités les plus intéressantes du logiciel. En effet, il nous donne accès à des outils (tri, mise en forme, etc.) dans un environnement bien connu des praticiens du traitement des données. Par exemple, la possibilité de trier les différents tableaux selon les contributions et les COS2 s'avère réellement pratique lorsque l'on souhaite interpréter les axes.