

Objectif

Mettre en œuvre l'analyse factorielle des correspondances (AFC) avec TANAGRA.

L'analyse factorielle des correspondances (AFC) est une technique de visualisation très populaire en analyse de données. Elle est adaptée à l'étude des tableaux de contingence. Un des objectifs est de produire une représentation, dans un repère unique, des catégories en lignes et en colonnes du tableau afin de mettre en évidence leurs positions respectives, et les éventuelles attractions-répulsions entre les caractéristiques.

A vrai dire, TANAGRA a été conçu pour traiter exclusivement des tableaux individus – variables. L'appréhension d'un tableau de contingence pose problème. L'informaticien, que je suis, renâcle face à une adaptation qui ne peut être que du bricolage. Mais l'enseignant de data mining, que je suis également, comprend que cette méthode, très largement répandue en analyse de données, ne peut pas être ignorée.

Au final, un compromis a été trouvé. Un tableau de contingence, dans TANAGRA, sera représenté par un fichier de données où une des variables, catégorielle, correspond aux lignes du tableau, les colonnes étant associées à une série de variables continues. L'intersection de la ligne et de la colonne contient l'effectif. Le type « variable catégorielle » (discrète) étant limité à 255 modalités dans TANAGRA, il ne sera donc pas possible de traiter un tableau de contingence avec plus de 255 modalités. On peut raisonnablement considérer néanmoins que cette limitation n'est pas excessivement pénalisante. Il n'y a pas de limitations en ce qui concerne le nombre de colonnes.

Compte tenu de ce micmac plus ou moins heureux, il n'a pas été possible de mettre en place un dispositif gérant les lignes ou les colonnes illustratives. Mais comme nous le verrons dans ce tutoriel, leur positionnement sur les axes factoriels peut être très facilement calculé à l'aide d'un tableur. L'association tableur-TANAGRA est d'ailleurs particulièrement avantageuse dans ce type d'analyse. De nombreux calculs peuvent être délégués au tableur. Il en est ainsi du calcul des profils lignes et colonnes.

Enfin, si votre fichier se présente sous la forme standard d'un fichier individus-variables, et que vous vous voulez étudier le croisement entre deux variables catégorielles, le plus simple est de passer par l'outil « Tableaux croisés dynamiques » d'un tableur¹ pour former le tableau de contingence. Puis, à partir du tableur, lancer TANAGRA en veillant à sélectionner la plage correspondant au tableau de contingence, en incluant les étiquettes.

Pour illustrer ce didacticiel, nous utilisons un exemple tiré de l'ouvrage de Lebart, Morineau et Piron, « Statistique Exploratoire Multidimensionnelle », Dunod, 2000. Ses auteurs font preuve d'une pédagogie remarquable. Ils nous permettent de suivre pas à pas les formules, pourtant ardues, de l'AFC. Cet ouvrage est également l'un des rares où l'on peut trouver, aussi clairement, une formulation symétrique de la matrice à diagonaliser (page 102), autrement plus facile à manipuler avec les bibliothèques de calcul usuelles. La partie relative à l'analyse factorielle des correspondances correspond à la section 1.3 (pages 67 à 107).

Fichier de données

Nous utilisons le fichier MEDIA_PROF_AFC.XLS tiré de notre ouvrage de référence (Tableau 1.3-10, page 104). L'intérêt de ce fichier est que nous pouvons comparer directement nos résultats avec

¹ « Tableaux croisés dynamiques » sous EXCEL. Sous OPEN OFFICE CALC, il faut utiliser le « Pilote de Données ».

ceux du livre (pages 104 à 107). Nous nous contentons de montrer l'enchaînement des opérations et la lecture des tableaux de résultats dans ce tutoriel. Pour ce qui des commentaires et de l'interprétation, le mieux est de se référer à l'ouvrage.

Le tableau de données est le suivant :

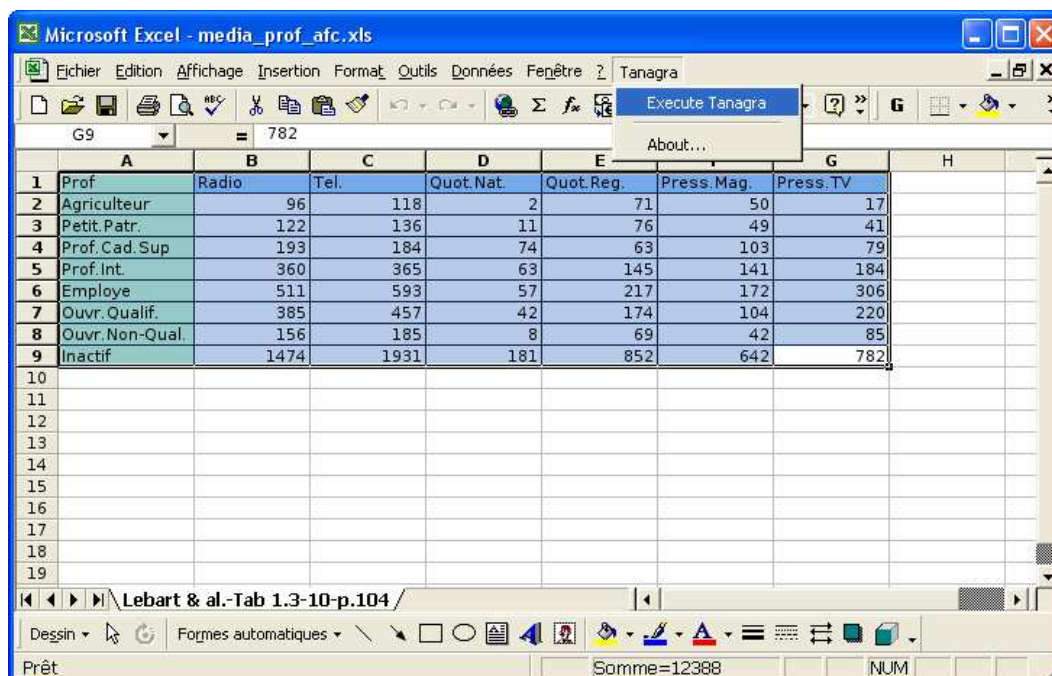
Prof	Radio	Tel.	Quot.Nat.	Quot.Reg.	Press.Mag.	Press.TV
Agriculteur	96	118	2	71	50	17
Petit.Patr.	122	136	11	76	49	41
Prof.Cad.Sup	193	184	74	63	103	79
Prof.Int.	360	365	63	145	141	184
Employe	511	593	57	217	172	306
Ouvr.Qualif.	385	457	42	174	104	220
Ouvr.Non-Qual.	156	185	8	69	42	85
Inactif	1474	1931	181	852	642	782

La première colonne, en vert, correspond à l'identifiant des lignes du tableau croisé. Les colonnes sont en bleu. A l'intersection d'une ligne et d'une colonne, nous lisons l'effectif associé à deux caractéristiques ex. 96 agriculteurs écoutent la radio. Notre tableau comporte 9 lignes et 6 colonnes.

Analyse Factorielle des Correspondances avec TANAGRA

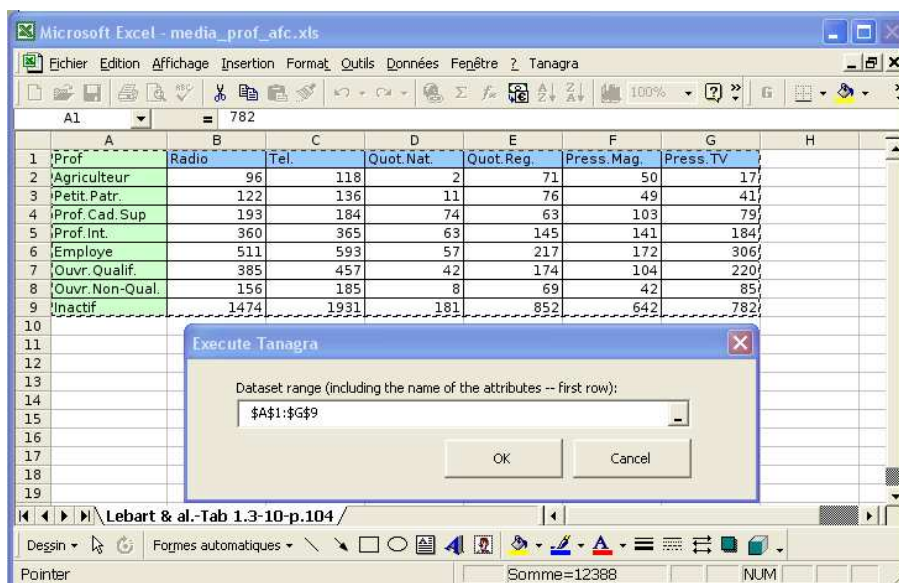
Créer un diagramme

Depuis la version 1.4.11, il est possible de démarrer TANAGRA à partir du tableur EXCEL². C'est la procédure que nous choisissons ici : le diagramme est automatiquement créé, et les données importées. Pour ce faire, nous sélectionnons la plage de données dans EXCEL, puis nous cliquons sur le menu TANAGRA/ EXECUTE TANAGRA.

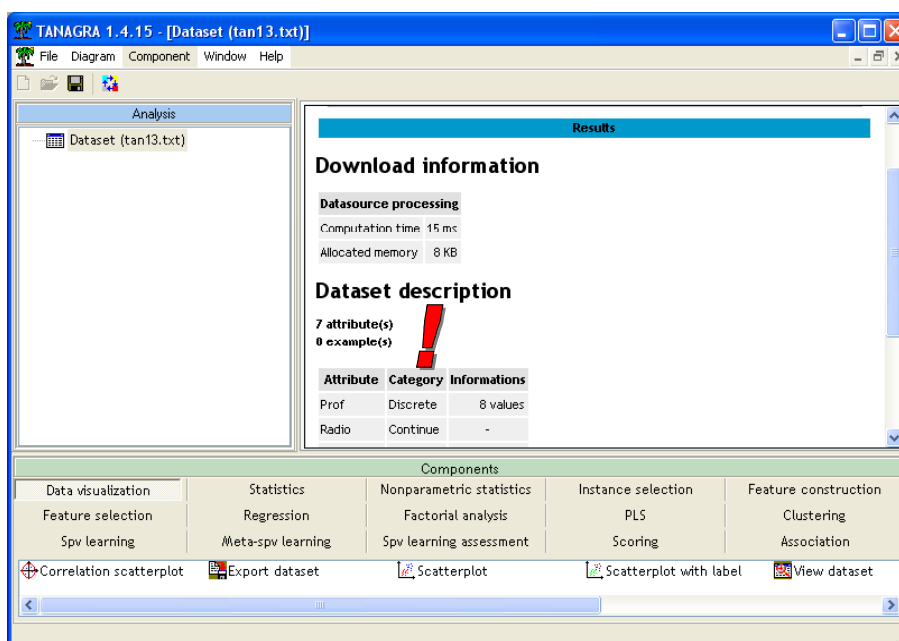


² Il faut bien entendu avoir référencé la macro-complémentaire (Add-In) TANAGRA dans EXCEL, voir le didacticiel adéquat sur le site web. La démarche est également valable avec le tableur CALC de OPEN OFFICE.

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



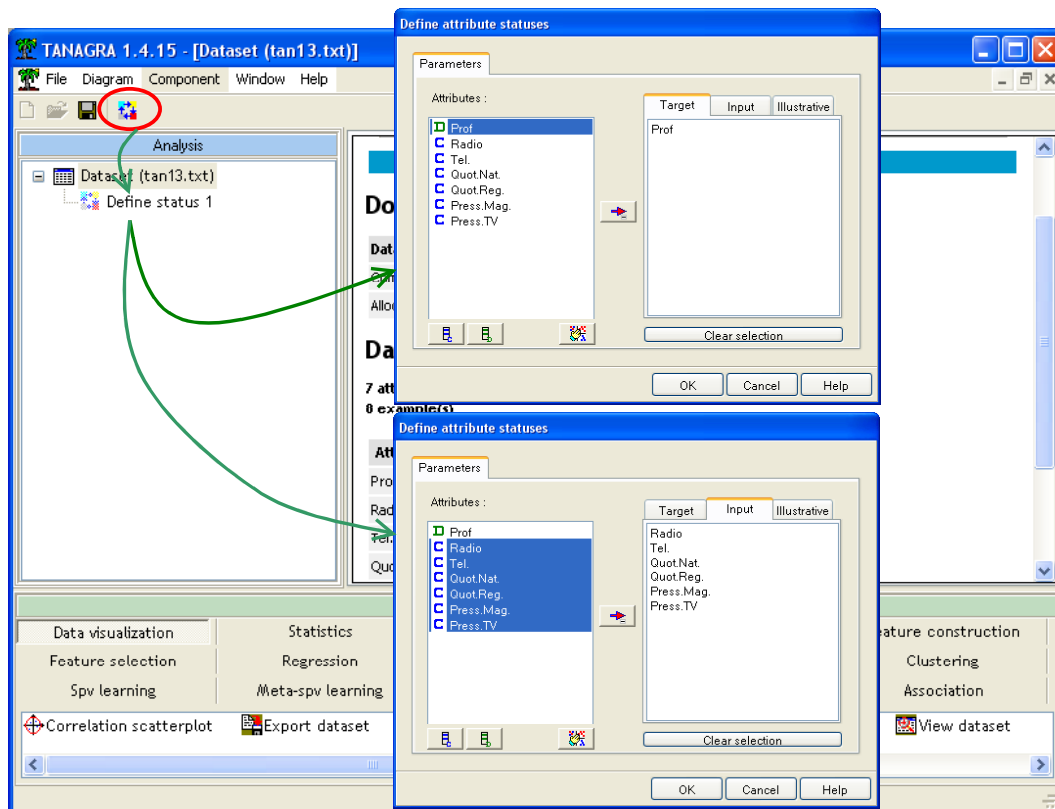
TANAGRA est alors démarré, nous vérifions que l'ensemble de données comprend bien 8 observations et 7 variables. Ce qui correspond à un tableau de contingence avec 8 lignes et 6 colonnes, une des variables, discrète, étant réservée pour identifier les lignes.



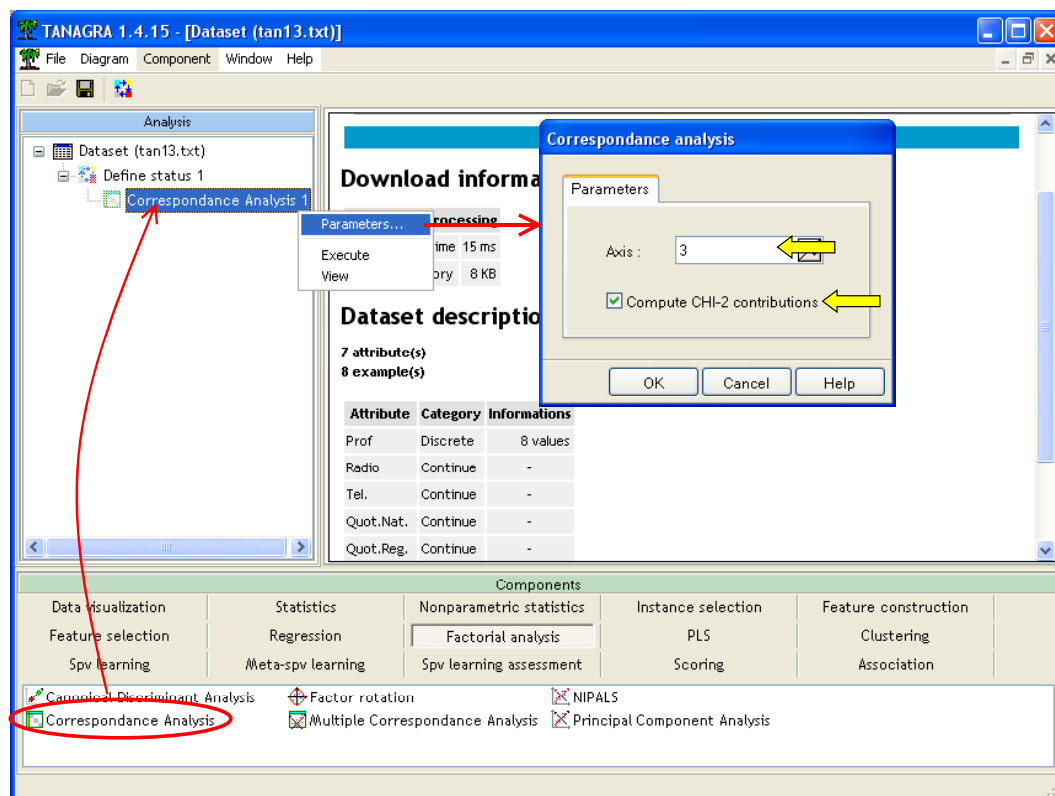
AFC

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 TARGET la variable discrète identifiant les lignes (PROF) ; en INPUT les variables continues, associées aux effectifs (RADIO ... PRESS.TV) dans les colonnes.

Attention, il ne s'agit pas de définir une analyse supervisée ici. C'est un artifice qui permet de spécifier les lignes et les colonnes du tableau de contingence.



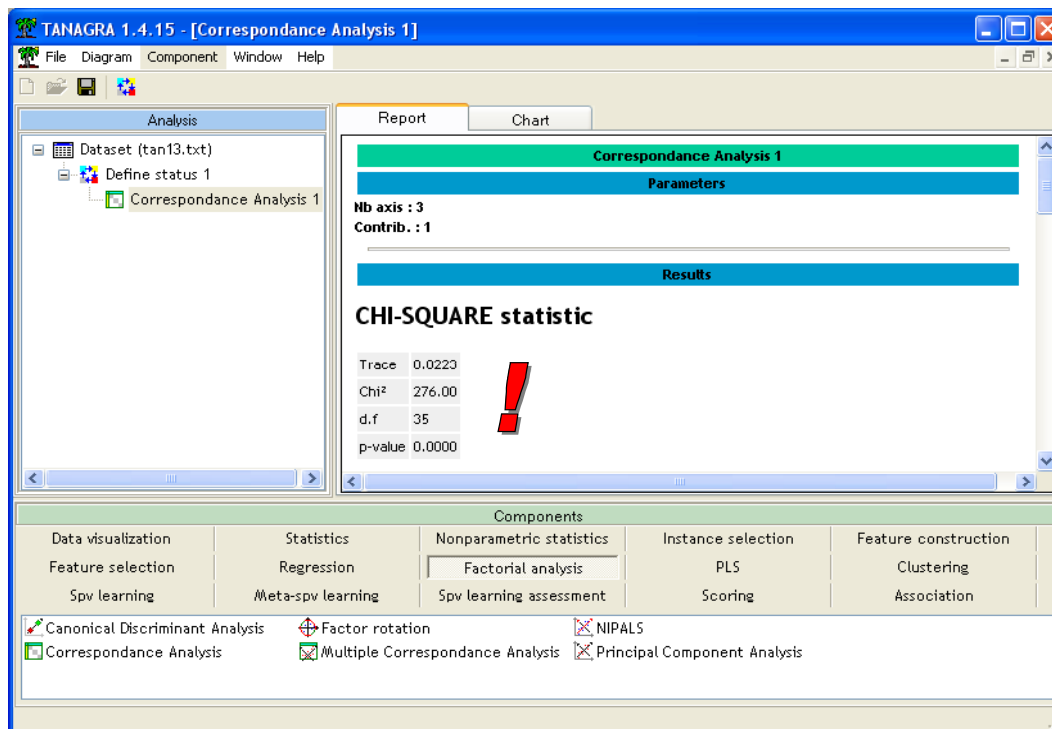
Puis nous plaçons le composant CORRESPONDANCE ANALYSIS (onglet FACTORIAL ANALYSIS) dans le diagramme de traitements. Nous cliquons sur le menu PARAMETERS. Nous spécifions alors le nombre d'axes à produire (3). Nous activons également l'option qui permet de calculer les contributions au CHI-2, nous détaillerons son rôle plus loin.



Nous cliquons sur le menu contextuel VIEW pour accéder aux résultats. Attention, il y a bien deux onglets dans la fenêtre d’affichage.

Résultats numériques de l’AFC

CHI-2 d’écart à l’indépendance. Comme nous travaillons sur un tableau de contingence, la première question est de savoir s’il existe un lien entre les lignes et les colonnes du tableau. L’indicateur usuel est le CHI-2 d’écart à l’indépendance, il est d’autant plus indiqué dans l’AFC puisque c’est la grandeur que nous décomposons par la suite (page 104).



La TRACE indique la somme des valeurs propres, multiplié à l’effectif total, il fournit au CHI-2 d’écart à l’indépendance bien connu. Pour tester l’hypothèse d’indépendance, nous utilisons la procédure usuelle, la p-value est affichée.

Valeurs propres. Plus bas dans la fenêtre, nous pouvons lire le tableau des valeurs propres. Nous observons la valeur propre calculée, le pourcentage d’inertie associé à chaque axe et le pourcentage cumulé qui permet de se donner une idée du nombre d’axes à retenir. Dans notre exemple, les deux premiers axes résument 94.56% de l’information disponible (Tableau 1.3-11, page 104).

Eigen values

Matrix trace = 0.0223

Axis	Eigen value	% explained	Histogram	% cumulated
1	0.013857	62.20%		62.20%
2	0.007211	32.37%		94.56%
3	0.000825	3.70%		98.27%
4	0.000304	1.36%		99.63%
5	0.000083	0.37%		100.00%
Tot.	0.022279	-	-	-

Coordonnées factorielles, contributions et COS². Dans la troisième partie des résultats, nous retrouvons les coordonnées factorielles de chaque modalité. Pour les lignes tout d'abord (Tableau 1.3 – 10, page 104).

Rows analysis

-		Coord.			Contributions (%)			COS ²		
Values	Weight	coord 1	coord 2	coord 3	ctr 1	ctr 2	ctr 3	cos ² 1	cos ² 2	cos ² 3
Agriculteur	2.86	0.166	-0.310	-0.072	5.7	38.0	17.9	0.214	0.741	0.040
Petit.Patr.	3.51	0.068	-0.143	-0.064	1.2	10.0	17.7	0.154	0.674	0.137
Prof.Cad.Sup	5.62	-0.430	-0.061	-0.003	75.0	2.9	0.1	0.978	0.020	0.000
Prof.Int.	10.15	-0.107	0.033	-0.031	8.3	1.5	11.8	0.802	0.075	0.067
Employe	14.98	0.016	0.095	-0.005	0.3	18.9	0.5	0.025	0.929	0.003
Ouvr.Qualif.	11.16	0.044	0.101	-0.019	1.5	15.9	5.1	0.138	0.744	0.027
Ouvr.Non-Qual.	4.40	0.118	0.095	-0.040	4.4	5.5	8.4	0.556	0.360	0.063
Inactif	47.32	0.033	-0.033	0.026	3.6	7.3	38.7	0.372	0.391	0.236

Pour les colonnes par la suite.

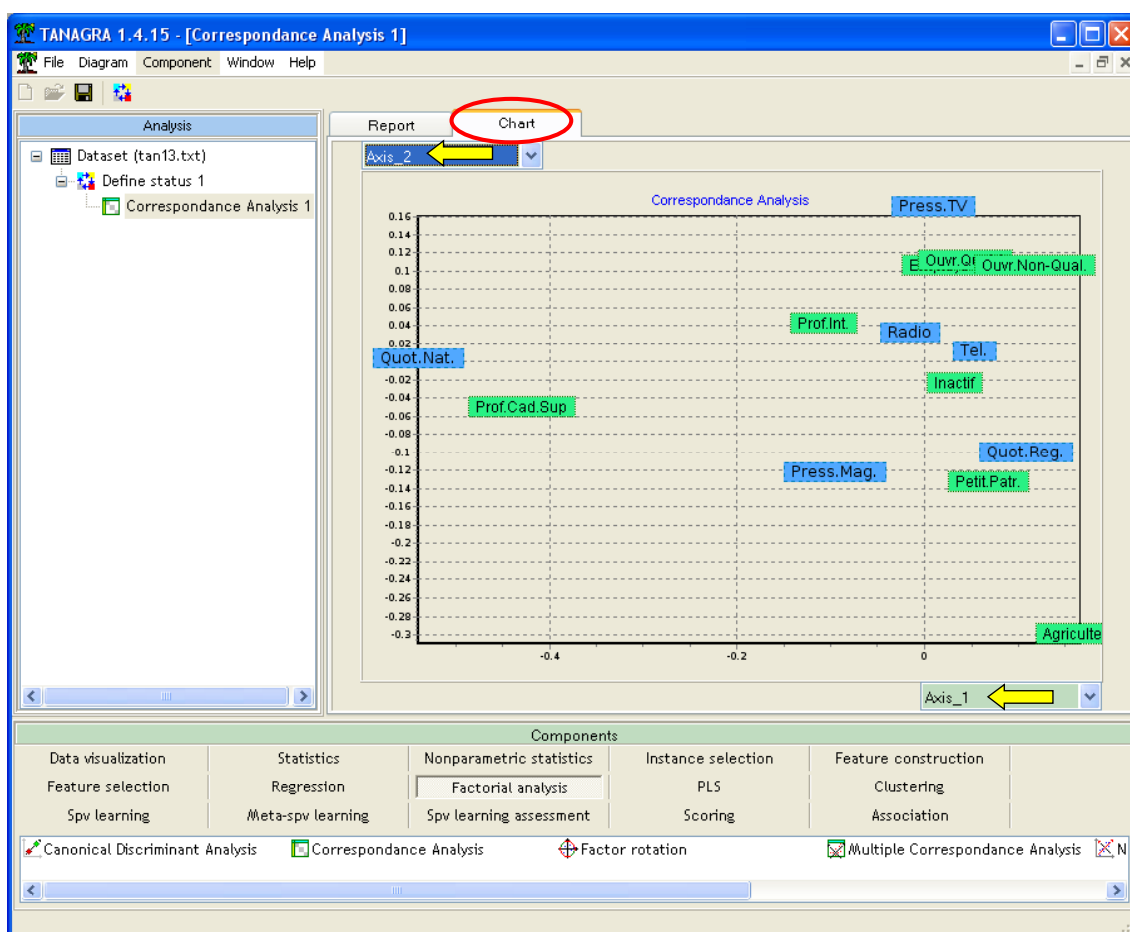
Columns analysis

-		Coord.			Contributions (%)			COS ²		
Values	Weight	coord 1	coord 2	coord 3	ctr 1	ctr 2	ctr 3	cos ² 1	cos ² 2	cos ² 3
Radio	26.61	-0.015	0.022	-0.047	0.4	1.8	70.4	0.077	0.168	0.752
Tel.	32.04	0.053	0.002	0.016	6.6	0.0	10.5	0.851	0.001	0.081
Quot.Nat.	3.54	-0.541	-0.006	0.021	74.6	0.0	1.8	0.993	0.000	0.001
Quot.Reg.	13.46	0.109	-0.110	0.005	11.5	22.4	0.4	0.487	0.494	0.001
Press.Mag.	10.52	-0.095	-0.132	0.019	6.8	25.6	4.5	0.317	0.619	0.012
Press.TV	13.84	0.010	0.162	0.027	0.1	50.1	12.4	0.003	0.959	0.027

Représentation graphique

Le pouvoir de séduction de l'analyse factorielle repose en grande partie sur les représentations graphiques qu'elle propose. Dans le cas de l'AFC, nous disposons d'une série de plans factoriels, que nous interprétons à l'aide des contributions. Elles permettent de situer les lignes (respectivement les colonnes) entre elles. Mais elles permettent également, grâce aux relations de transitions (page 85), de tracer simultanément les points « lignes » et les points « colonnes » dans le même repère. Nous pouvons ainsi évaluer en un coup d'œil les attractions et les répulsions qu'il peut y avoir entre certaines lignes et colonnes du tableau.

Pour accéder aux graphiques, nous sélectionnons l'onglet CHART dans la fenêtre de visualisation. Avec les boîtes listes situés en abscisse et en ordonnées, nous avons le choix du plan factoriel à étudier.



Nous observons ici le premier plan factoriel (Figure 1.3 – 23, page 106). La lecture des quotidiens nationaux par les cadres supérieurs est l'information qui prédomine dans ce tableau, elle a tendance à écraser les autres enseignements que l'on pourrait en tirer.

Notons qu'il est possible de copier le graphique dans un traitement de texte. Il est également possible de modifier la taille de la police des étiquettes des points et de zoomer sur certaines parties du graphique.

Tableau des contributions au CHI-2

Lors du paramétrage de la méthode, nous avons activé une option, le calcul des contributions au CHI-2. Revenons sur l'onglet des résultats numériques. Il y a un dernier tableau que nous n'avons pas encore commenté.

Le tableau des contributions recense la contribution au CHI-2 de chaque case du tableau de contingence, en confrontant la valeur observée et la valeur espérée sous l'hypothèse d'indépendance. Il s'agit d'une autre manière de détecter les informations importantes. L'intérêt ici est que nous disposons des résultats pour chaque couple ligne x colonne du tableau, et ils sont surtout triés par ordre décroissant d'importance. Nous pouvons dès lors visualiser les cases les plus informatives du tableau, en termes d'attraction (en vert) ou de répulsion (en rouge).

Pour ne pas alourdir inutilement l'affichage, seules les contributions supérieures à la moyenne (CHI-2 divisé par le nombre de cases du tableau) sont affichées.

CHI-2 contributions

Row	Column	Value	Expected	Contrib.	%
Prof.Cad.Sup	Quot.Nat.	74	24.6	99.13	35.92
Agriculteur	Press.TV	17	49.0	20.88	7.57
Prof.Cad.Sup	Press.Mag.	103	73.2	12.12	4.39
Ouvr.Qualif.	Press.Mag.	104	145.4	11.77	4.26
Agriculteur	Quot.Reg.	71	47.6	11.46	4.15
Prof.Cad.Sup	Quot.Reg.	63	93.7	10.04	3.64
Employe	Press.TV	306	256.8	9.43	3.42
Agriculteur	Quot.Nat.	2	12.5	8.84	3.20
Prof.Int.	Quot.Nat.	63	44.5	7.71	2.79
Prof.Cad.Sup	Tel.	184	223.0	6.82	2.47
Ouvr.Non-Qual.	Quot.Nat.	8	19.3	6.59	2.39
Petit.Patr.	Press.TV	41	60.2	6.12	2.22

Nous retrouvons bien la forte attraction entre la presse nationale et les cadres supérieurs, préoccupés par le destin de la nation. Elle capte 35% de l'information qu'apporte le tableau de contingence.

Nous observons également que les agriculteurs sont fâchés avec la presse TV, etc.

Coordonnées des lignes/colonnes supplémentaires

Un autre intérêt des techniques factorielles est la possibilité de placer de nouvelles observations dans le repère qui a été défini par l'analyse. Dans notre cas, nous essayons de voir la situation des hommes dans le premier plan factoriel ; puis la situation des personnes ayant suivi des études supérieures. Les coordonnées sont les suivantes :

	Radio	Tel.	Quot.Nat.	Quot.Reg.	Press.Mag.	Press.TV
Sexe = Homme	1630	1900	285	854	621	776

TANAGRA ne peut pas calculer directement les coordonnées de cette nouvelle observation. En revanche, nous avons en main tous les éléments pour réaliser les calculs.

Calculons le profil associé à cette observation :

	Radio	Tel.	Quot.Nat.	Quot.Reg.	Press.Mag.	Press.TV
Sexe = Homme	1630	1900	285	854	621	776
Profil ligne	0.27	0.31	0.05	0.14	0.10	0.13

Nous utilisons le tableau des coordonnées factorielles des colonnes (cf. plus haut, pour rappel nous le recopions ici)

Columns analysis

-		Coord.			Contributions (%)			COS ²		
Values	Weight	coord 1	coord 2	coord 3	ctr 1	ctr 2	ctr 3	cos ² 1	cos ² 2	cos ² 3
Radio	26.61	-0.015	0.022	-0.047	0.4	1.8	70.4	0.077	0.168	0.752
Tel.	32.04	0.053	0.002	0.016	6.6	0.0	10.5	0.851	0.001	0.081
Quot.Nat.	3.54	-0.541	-0.006	0.021	74.6	0.0	1.8	0.993	0.000	0.001
Quot.Reg.	13.46	0.109	-0.110	0.005	11.5	22.4	0.4	0.487	0.494	0.001
Press.Mag.	10.52	-0.095	-0.132	0.019	6.8	25.6	4.5	0.317	0.619	0.012
Press.TV	13.84	0.010	0.162	0.027	0.1	50.1	12.4	0.003	0.959	0.027

Sa coordonnée sur le premier axe factoriel est alors :

$$H_1 = \frac{1}{\sqrt{0.0139}} [0.27 \times -0.015 + 0.31 \times 0.053 + 0.05 \times -0.541 + 0.14 \times 0.109 + 0.1 \times -0.095 + 0.13 \times 0.01]$$

$$= -0.05$$

La valeur 0.0139 est la première valeur propre, relative au premier axe. Nous obtenons ainsi le point (-0.05 ; -0.02) dans le premier plan factoriel (Tableau 1.3 – 12, page 105). La coordonnée est assez proche de l'origine, les hommes ne se comportent pas de manière particulière par rapport à l'accès aux médias.

Prenons un second exemple, nous nous intéressons maintenant aux personnes ayant suivi des études supérieures (Tableau 1.3 – 10, page 104). Voici son profil :

	Radio	Tel.	Quot.Nat.	Quot.Reg.	Press.Mag.	Press.TV
Etud.Sup	619	612	177	209	298	281
Profil ligne	0.28	0.28	0.08	0.10	0.14	0.13

Toujours avec la même démarche, sa coordonnée dans le premier plan factoriel est (-0.29 ; -0.02). On constate une certaine proximité avec les cadres supérieurs (-0.43 ; -0.06) concernant leur comportement face à l'accès aux médias : on leur a appris à s'intéresser aux grandes destinées de la nation.

Conclusion

TANAGRA ne prétend pas fournir des outils de reporting et de déploiement à la hauteur des logiciels commerciaux. Le module graphique reste assez fruste. 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. Entre autres, la possibilité de projeter les individus supplémentaires, en effectuant des calculs très simples sous un tableur, permet d'étendre la portée de l'analyse.