1 Objectif

Analyse factorielle des correspondances (AFC) – Comparaison de logiciels.

L'analyse des correspondances est une méthode statistique de réduction de dimension. Elle propose une vision synthétique de l'information intéressante d'un tableau de contingence. Son pouvoir de séduction repose en grande partie sur les représentations graphiques qu'elle propose. Elles nous permettent de situer facilement (beaucoup le pensent en tous cas) les similarités (dissimilarités) et les attractions (répulsions) entre les modalités. L'AFC est bien une technique factorielle. Les facteurs – les variables latentes – qui en sont issus sont des combinaisons linéaires des points modalités (lignes ou colonnes) exprimés par des profils (lignes ou colonnes).

Il existe un très grand nombre de supports la décrivant sur le web[1]. On peut voir l'AFC comme une méthode d'ajustement des nuages de profils lignes et colonnes, permettant leur représentation simultanée, c.-à-d. dans le même repère. Nous pouvons aussi la voir sous l'angle de la décomposition orthogonale du χ^2 d'écart à l'indépendance. Ce prisme est intéressant parce qu'il indique clairement qu'avant de nous intéresser à la structure des écarts, il est très important de considérer leur importance et se poser la question : y a –t-il de l'information pertinente dans le tableau à analyser ? Si le χ^2 (ou le ϕ^2) est trop faible, l'étude subséquente est illusoire[2].

Dans ce tutoriel, nous décrivons la mise en œuvre de l'AFC dans plusieurs logiciels : la version **1.4.48** de **Tanagra** qui intègre des nouveautés destinées à améliorer la lecture des résultats ; le logiciel **R** avec les packages « **ca** » et « **ade4** » ; le logiciel **OpenStat** ; et le logiciel **SAS** qui servira de référence. Nous constaterons – comme toujours – que tous ces logiciels produisent exactement les mêmes résultats numériques (heureusement !). Les différences se situent essentiellement au niveau de la mise en valeur des sorties.

Ce document vient compléter deux précédents tutoriels dédiés à l'analyse factorielle des correspondances : « AFC – Association médias et professions » qui présente en détail la mise en œuvre de la méthode dans Tanagra (1.4.15), mettre en parallèle les copies d'écran permettra d'apprécier les améliorations apportées dans cette version 1.4.48 ; « Analyse factorielle des correspondances avec R » qui s'appuie sur le package « FactoMineR », raison pour laquelle nous avons choisi de décrire d'autres packages cette fois-ci.

2 Données

Nous utilisons les données décrites dans le tutoriel de Bendixen (1996)[3]. Nous pourrons confronter les sorties des différents outils avec ceux de l'auteur. Le lecteur pourra s'y référer également pour ce qui est de l'interprétation des résultats. Sujet sur lequel nous ne nous attarderons guère. Notre objectif est avant tout de situer le comportement des différents logiciels.

[1] Ex. P. Besse, « Exploration Statistique Multidimensionnelle », Chapitre « Analyse Factorielle des Correspondances » ; A. Bouchier, « AFC Simple », in R & Statistique, 2010.

[2] P. Cibois, « Les pièges de l'analyse des correspondances », Histoire & Mesure, 12 (3/4), pp. 99-320, 1997.

[3] M. Bendixen, « A practical guide to the use of the correspondence analysis in marketing research », Marketing Research On-Line, 1 (1), pp. 16-38, 1996 ; http://marketing-bulletin.massey.ac.nz/V14/MB_V14_T2_Bendixen.pdf

Le tableau de contingence - issu d'une enquête auprès de **100 ménages** - croise 8 types d'aliments pour petit déjeuner avec 14 mots clés destinés à les qualifier. Nous avons donc **L = 14 lignes** et **C = 8 colonnes**. Les réponses multiples étant autorisées, l'effectif total est de **n = 1760**. L'étude cherche à mettre en évidence les relations (attractions ou répulsions) les plus marquantes entre les mots clés et les aliments.

		Foods							
		Cereals	Muesli	Porridge	BaconEggs	ToastTea	FreshFruit	StewedFruit	Yoghurt
Statement	Healthy	14	38	25	18	8	31	28	34
	Nutritious	14	28	25	25	7	32	26	31
	GoodSummer	42	22	11	13	7	37	16	35
	GoodWinter	10	10	32	26	6	11	19	8
	Expensive	6	33	5	27	3	9	18	10
	QuickEasy	54	33	8	2	15	26	8	20
	Tasty	24	21	16	34	11	33	26	26
	Economical	24	3	20	3	16	7	3	7
	ForATreat	5	3	3	31	4	4	16	17
	ForWeekdays	47	24	15	9	13	11	6	10
	ForWeekends	12	5	8	56	16	10	23	18
	Tasteless	8	6	2	2	0	0	2	1
	TooLongToPrepare	0	0	9	35	1	0	10	0
	FamilyFavourite	14	4	10	31	5	7	2	5

Figure 1 – Aliments et qualificatifs (Bendixen, 1996)

3 Analyse des correspondances avec Tanagra (14.4.48)

3.1 Analyse factorielle des correspondances avec SAS

Nous commençons par le logiciel SAS puisque ses sorties serviront de jalons. Nous avons décrit dans un précédent document l'importation des données[4]. Nous soumettons les instructions suivantes pour lancer une AFC.

```
proc corresp data = mesdata.foods dimens=2;
var Cereals -- Yoghurt;
id Statement;
run;
```

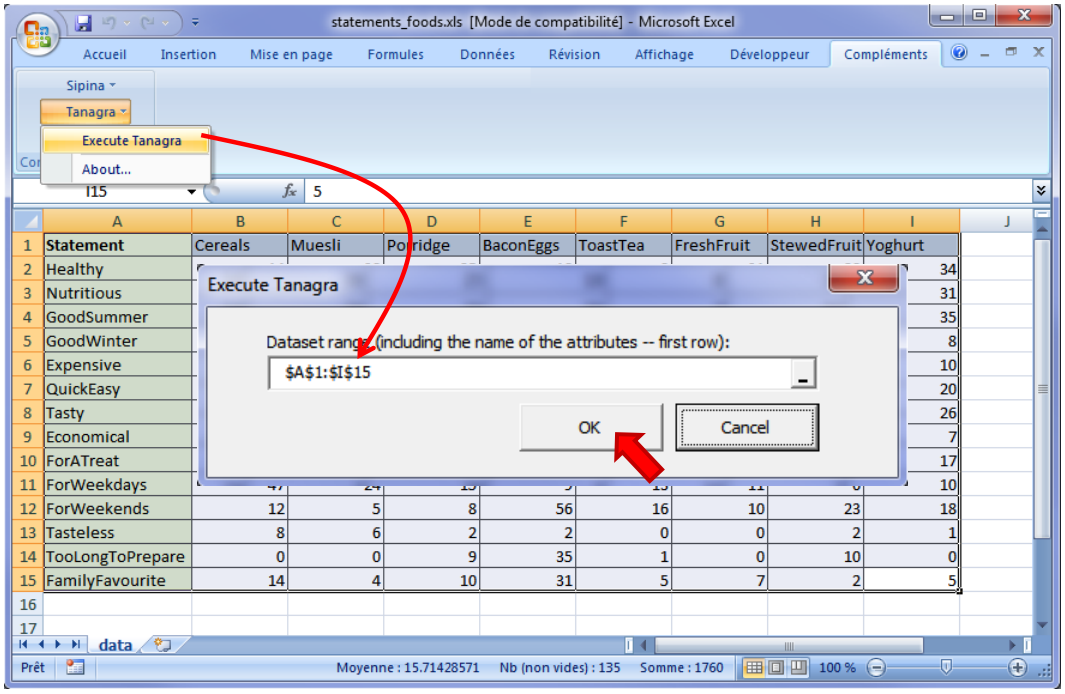
Nous décrirons les résultats dans ce qui suit, en les mettant en parallèle avec ceux de Tanagra.

3.2 Importation des données dans Tanagra

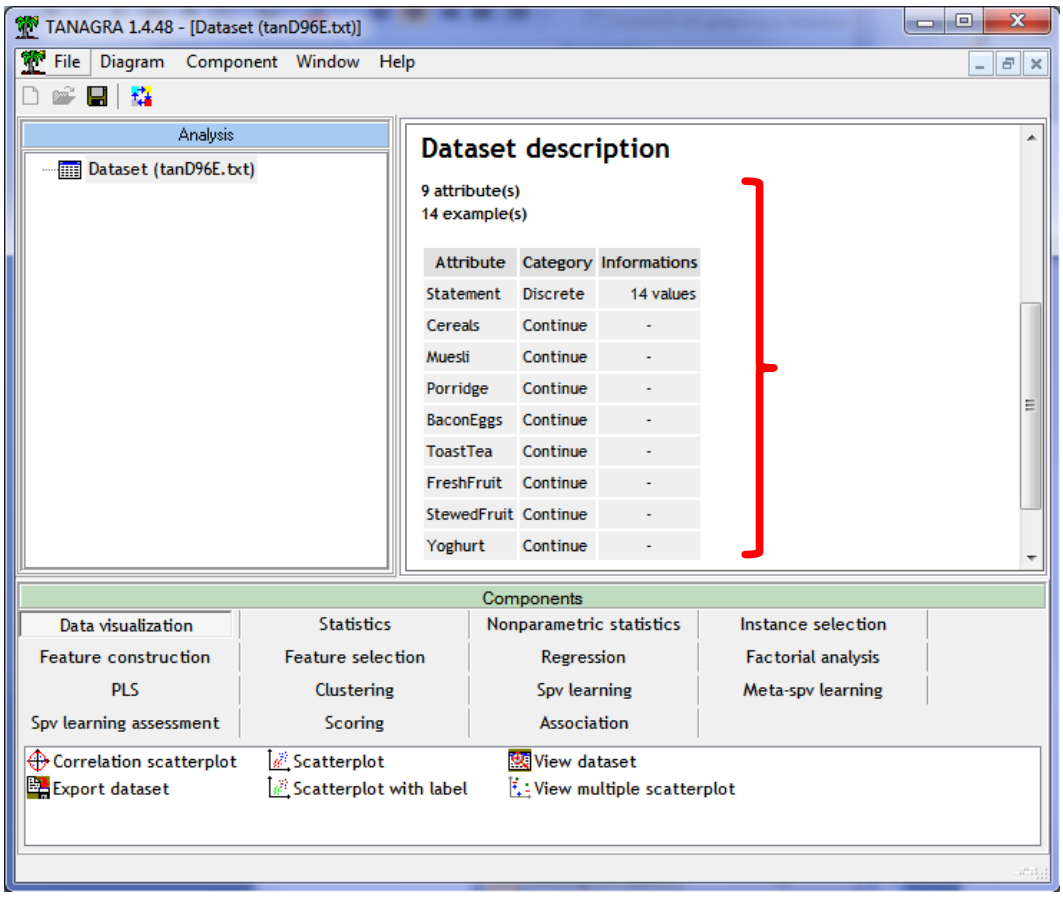
Le plus simple pour importer les données est d'ouvrir le fichier « statements_foods.xls » dans le tableur Excel et de l'envoyer à Tanagra via la macro complémentaire **Tanagra.xla**[5]. Une boîte de dialogue permet de vérifier les coordonnées des cellules (**$A$1:$I$15**), nous validons en cliquant sur le bouton OK.

[4] http://tutoriels-data-mining.blogspot.fr/2012/04/la-proc-logistic-de-sas-93.html, pages 2 et 3.

[5] Voir http://tutoriels-data-mining.blogspot.fr/2010/08/ladd-in-tanagra-pour-excel-2007-et-2010.html pour Excel 2007 et 2010 ; http://tutoriels-data-mining.blogspot.fr/2008/03/importation-fichier-xls-excel-macro.html pour Excel 2003 et version antérieures ; http://tutoriels-data-mining.blogspot.fr/2011/07/tanagra-addon-pour-openoffice-33.html pour Open Office et Libre Office.

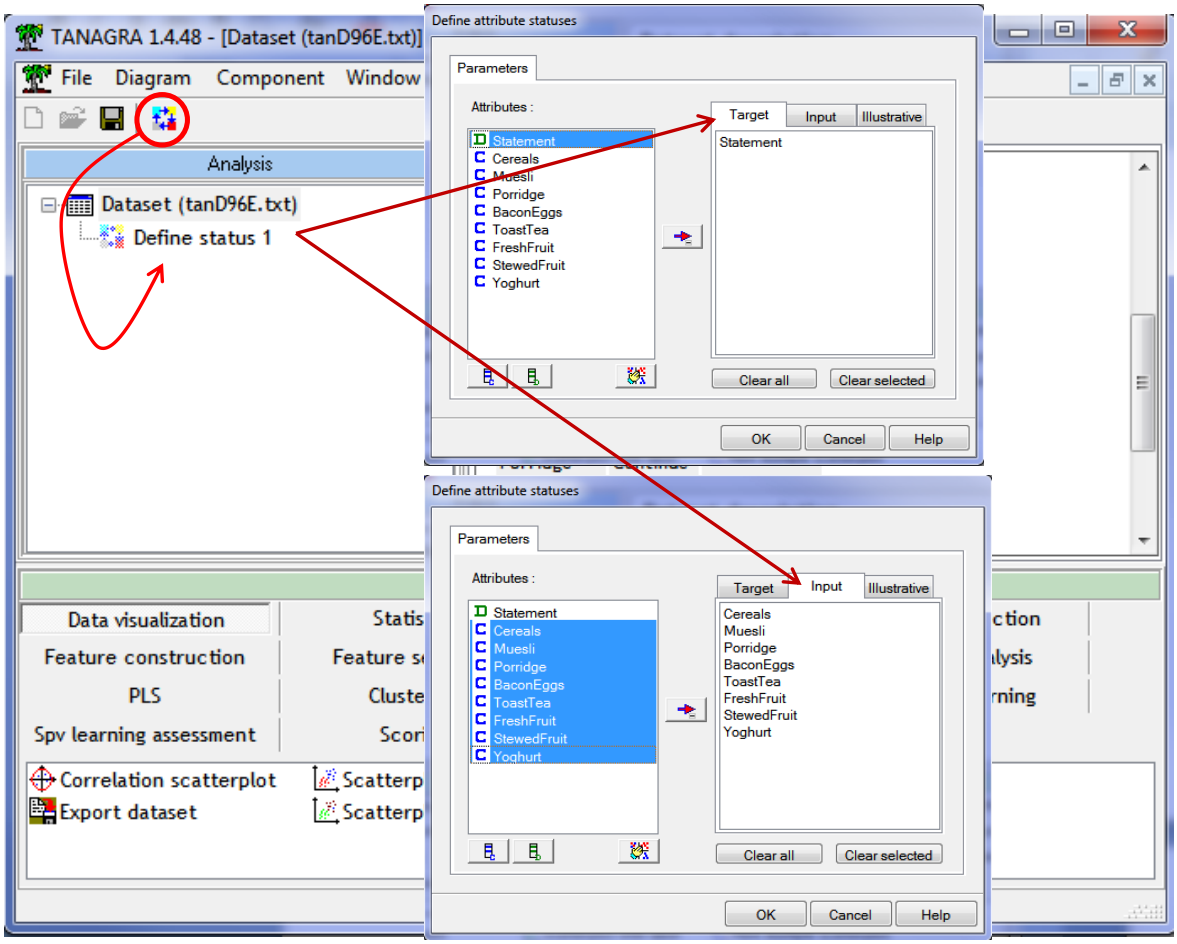


Tanagra est automatiquement démarré, 8 colonnes (9 en comptant l'étiquette des lignes) et 14 lignes (15 avec l'étiquette des colonnes) ont été chargées.

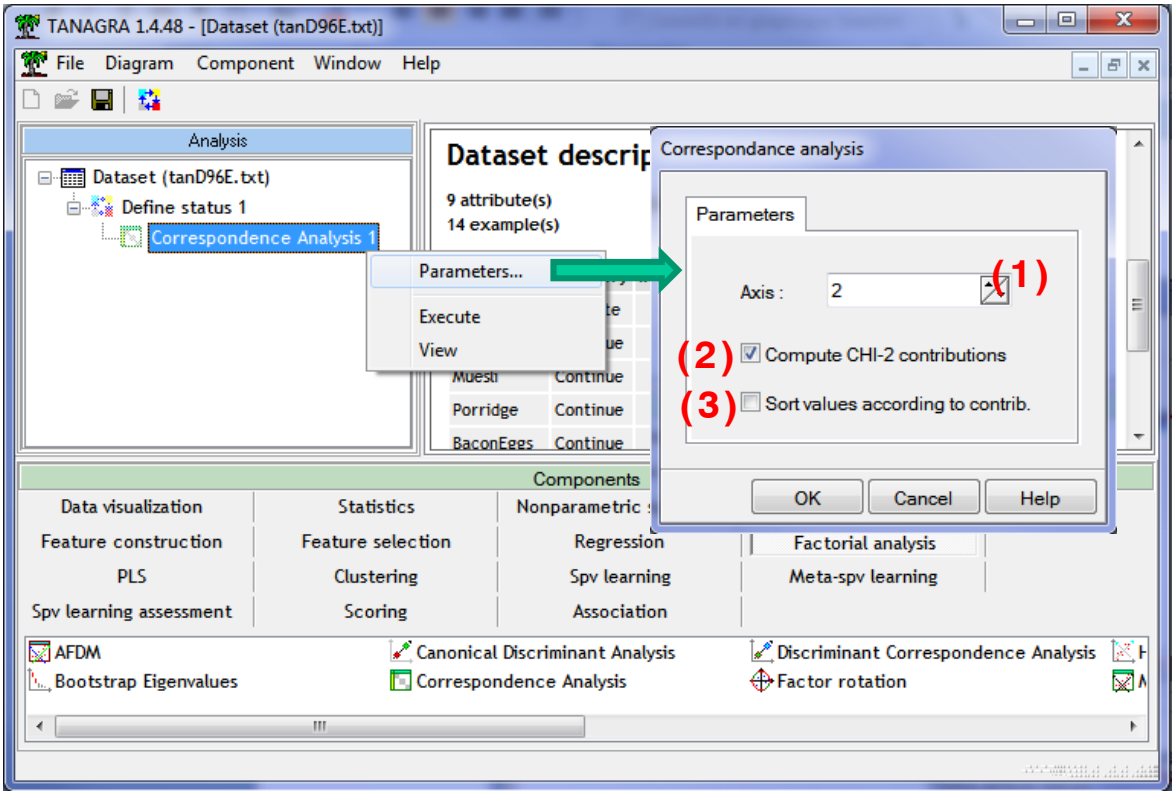


3.3 Paramétrage de l'analyse des correspondances

Nous devons définir le rôle des variables à l'aide du composant DEFINE STATUS. Nous plaçons STATEMENT (étiquette des lignes) en TARGET, les autres variables (les colonnes) en INPUT.



Ensuite, nous plaçons le composant CORRESPONDANCE ANALYSIS (onglet FACTORIAL ANALYSIS) dans le diagramme. Nous actionnons au menu contextuel PARAMETERS pour en définir les paramètres.



Nous restreignons le calcul à 2 axes factoriels (1) (nous expliquerons pourquoi plus loin) ; nous calculons les contributions au KHI-2 des cellules du tableau de contingence (2) ; nous ne souhaitons pas trier les tableaux pour l'instant (3).

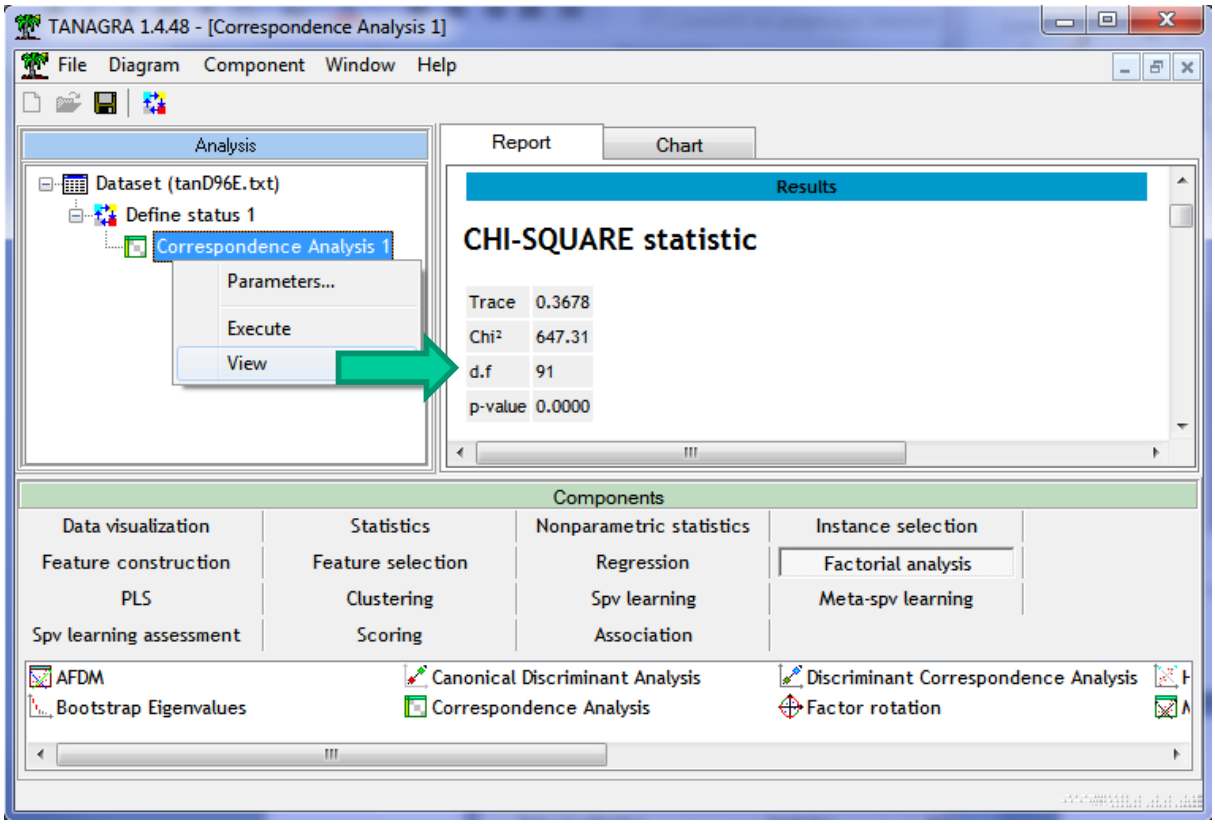
Il ne nous reste plus qu'à valider ces choix et à cliquer sur le menu VIEW pour accéder aux résultats.

3.4 Lecture des résultats

Les sorties sont subdivisées en plusieurs zones, nous allons les énumérer tour à tour dans ce qui suit.

3.4.1 KHI-2 (global) de l'écart à l'indépendance

Le premier tableau indique la statistique du test du χ^2 d'écart à l'indépendance[6]. Ce résultat est fondamental. En effet, si la liaison globale est trop faible, l'étude des relations entre les modalités ne sert à rien. Il faut s'assurer qu'il existe une information exploitable dans le tableau.



En l'occurrence, nous avons $\chi^2_{global} = 647.31$, avec un degré de liberté égal à 91 [= (14 - 1) x (8 – 1)], la liaison est très significative (p-value < 0.0001).

De plus, Tanagra fournit la valeur du coefficient ϕ^2 (**Trace**), avec $\phi^2 = \chi^2/n = 647.31 / 1760 = 0.3678$. L'AFC va décomposer cette quantité[7] - qui symbolise l'information disponible dans le tableau de contingence - sur les différents axes factoriels.

[6] Le test du χ^2 n'est pas strictement applicable ici. En effet chaque individu a pu choisir plusieurs couples de valeurs (statement x food). De fait, les observations ne sont pas indépendantes. Il faut dès lors voir le χ^2 plutôt comme un indicateur de la quantité d'information exploitable dans le tableau.

[7] A partir de $\phi > 0.2$, on peut penser que le tableau recèle des informations intéressantes (Bendixen, page 7).

3.4.2 Tableau des valeurs propres – Choix du nombre d'axes

Tableau des valeurs propres. Tanagra affiche ensuite le tableau des valeurs propres (λ_k). Elles expriment la part d'inertie expliquée par les axes. Ainsi, puisque la décomposition est orthogonale, elles s'additionnent et la somme (0.193095 + 0.077731 + … + 0.002363) est égale à $\phi^2 = 0.3678$.

Nous pouvons ré-écrire les valeurs en pourcentage d'inertie expliquée par les axes (ex. % axe 1 = 0.193095 / 0.3678 = 52.50 % ; % axe 2 = 0.07731 = 21.13 %).

Eigen values

Matrix trace = 0.3678
SQRT(Matrix trace) = 0.6065

Axis	Eigen value	% explained	Histogram	% cumulated
1	0.193095	52.50%		52.50%
2	0.077731	21.13%		73.64%
3	0.043854	11.92%		85.56%
4	0.032804	8.92%		94.48%
5	0.012257	3.33%		97.81%
6	0.005687	1.55%		99.36%
7	0.002363	0.64%		100.00%
Tot.	0.367791	-	-	-

(Tanagra)

The CORRESP Procedure

Inertia and Chi-Square Decomposition

Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent	11 22 33 44 55 ----+----+----+----+----+---
0.43943	0.19309	339.846	52.50	52.50	**********
0.27880	0.07773	136.806	21.13	73.64	**********
0.20941	0.04385	77.183	11.92	85.56	*****
0.18112	0.03280	57.735	8.92	94.48	****
0.11071	0.01226	21.572	3.33	97.81	**
0.07541	0.00569	10.010	1.55	99.36	*
0.04861	0.00236	4.159	0.64	100.00	
Total	0.36779	647.312	100.00		

Degrees of Freedom = 91

(SAS)

Notons une particularité de SAS. Elle propose une décomposition de l'information selon le χ^2. Il s'agit tout simplement de la valeur propre multipliée par l'effectif *n* (ex. χ^2_1 = 0.19309 x 1760 = 339.846 ; etc.). La somme des χ^2 individuels donne bien le χ^2_{global} = 647.312.

Choix du nombre d'axes – Règle de Kaiser. Vient alors une question récurrente : combien d'axes devons-nous retenir dans l'analyse ? Le nombre maximum d'axes factoriels que nous pouvons produire est K_{max} = MIN(L -1, C-1) = MIN (13, 7) = 7. Dès lors, une règle très simple consiste à choisir les axes pour lesquels la part d'inertie expliquée est supérieure à (1 / K_{max}) = 14.3 %, soit les axes 1 (52.50 %) et 2 (21.13 %).

Cette idée revient à appliquer la règle de Kaiser. Nous retenons les axes portés par une valeur propre supérieure à leur moyenne c.-à-d. dont la valeur propre est supérieure à (0.3678 / 7) = 0.0525 ; nous avons bien les axes 1 (λ_1=0.19309) et 2 (λ_2= 0.07773).

Choix du nombre d'axes – Scree plot. La règle du coude peut être également mise à contribution.

Il s'agit de repérer le « coude » dans la décroissance des valeurs propres. Dans notre cas, il survient au second axe, ce qui corrobore la conclusion de la règle de Kaiser.

Choix du nombre d'axes – Test de Malinvaud. Dans son ouvrage, Saporta (2006, page 209) décrit une règle de sélection d'axes moins empirique. Il s'agit d'une procédure permettant de tester séquentiellement l'intérêt des axes postérieurs aux K premiers de l'analyse c.-à-d. est-ce K axes sont suffisant (H0) ou il en faudrait plus (H1) ?

La statistique de test s'écrit

$$Q_K = n \times \left(\lambda_{K+1} + \cdots + \lambda_{K_{max}}\right)$$

Elle suit une loi du χ^2 à (L – K – 1) x (C – K – 1) degrés de liberté.

Les fondements théoriques sont plus solides. Certes. Malheureusement, comme tout test basé sur la statistique du χ^2, il a tendance à être exagérément significatif dès que l'effectif *n* augmente. Nous le constaterons sur notre exemple.

Nous avons retranscrit la procédure de sélection dans le tableau suivant :

K	Factor	Eigen value	CHI-2	ddl	p-value
0	1	0.193095	647.31	91	0.0000
1	2	0.077731	307.46	72	0.0000
2	3	0.043854	170.66	55	0.0000
3	4	0.032804	93.48	40	0.0000
4	5	0.012257	35.74	27	0.1211
5	6	0.005687	14.17	16	0.5862
6	7	0.002363	4.16	7	0.7613

Le test initial, K = 0, permet de tester l'existence d'informations exploitables dans le tableau de contingence que nous analysons. Il correspond au test de significativité globale, en effet $Q_0 = \chi^2_{global}$.

Pour K = 3 (c.-à-d. 3 axes sont suffisants ou bien en faudrait-il plus ?), nous avons

$$Q_3 = 1760 \times (0.032804 + 0.012257 + 0.05687 + 0.002363) = 93.48$$

Avec (L – K – 1) x (C – K – 1) = (14 – 3 – 1) x (8 – 3 – 1) = 40 degrés de liberté, nous avons un p-value < 0.0001. Une modélisation en K = 3 axes n'est apparemment pas suffisante. Il faut sélectionner au moins un axe supplémentaire.

Pour K = 4 (c.-à-d. 4 axes sont suffisants ou bien en faudrait-il plus ?), nous avons

$$Q_4 = 1760 \times (0.012257 + 0.05687 + 0.002363) = 35.74$$

Avec (14 – 4 – 1) x (8 – 4 – 1) = 27 degrés de liberté, nous avons un p-value = 0.1211. On peut s'en tenir à un ajustement en K = 4 axes.

Manifestement, cette solution (K = 4 axes) est inappropriée. Elle ne correspond absolument pas au résultat suggéré par la règle de Kaiser et par le « scree test ». Comme le souligne notre référence, mieux vaut réserver cette procédure aux tableaux de contingence à effectifs relativement modérés.

3.4.3 Représentation des lignes

La représentation des lignes couvre plusieurs informations (Figure 2) : les statistiques sur les points lignes (poids, distance à l'origine, l'inertie) [A] ; les coordonnées factorielles [B] ; les contributions (CTR) aux axes (en %) [C] ; et la qualité de représentation (COS²) par axe et cumulée [D].

Rows analysis

Row Characterization				Coord.		Contributions (%)		COS²	
Values	Weight	Sq. Dist.	Inertia	coord 1	coord 2	ctr 1	ctr 2	cos² 1	cos² 2
Healthy	0.11136	0.16313	0.01817	0.08662	-0.34591	0.43	17.14	0.05 (0.05)	0.73 (0.78)
Nutritious	0.10682	0.10686	0.01141	-0.00861	-0.26886	0.00	9.93	0.00 (0.00)	0.68 (0.68)
GoodSummer	0.10398	0.22035	0.02291	0.34879	-0.13275	6.55	2.36	0.55 (0.55)	0.08 (0.63)
GoodWinter	0.06932	0.33945	0.02353	-0.27741	0.12713	2.76	1.44	0.23 (0.23)	0.05 (0.27)
Expensive	0.06307	0.40466	0.02552	-0.20067	-0.36125	1.32	10.59	0.10 (0.10)	0.32 (0.42)
QuickEasy	0.09432	0.46617	0.04397	0.64365	0.08476	20.24	0.87	0.89 (0.89)	0.02 (0.90)
Tasty	0.10852	0.03870	0.00420	-0.03963	-0.11273	0.09	1.77	0.04 (0.04)	0.33 (0.37)
Economical	0.04716	0.80423	0.03793	0.41692	0.65687	4.25	26.18	0.22 (0.22)	0.54 (0.75)
ForATreat	0.04716	0.54240	0.02558	-0.61946	-0.06492	9.37	0.26	0.71 (0.71)	0.01 (0.72)
ForWeekdays	0.07670	0.42038	0.03225	0.50675	0.33799	10.20	11.27	0.61 (0.61)	0.27 (0.88)
ForWeekends	0.08409	0.43297	0.03641	-0.56005	0.17123	13.66	3.17	0.72 (0.72)	0.07 (0.79)
Tasteless	0.01193	0.78910	0.00942	0.45096	0.15778	1.26	0.38	0.26 (0.26)	0.03 (0.29)
TooLongToPrepare	0.03125	1.82554	0.05705	-1.27778	0.33401	26.42	4.49	0.89 (0.89)	0.06 (0.96)
FamilyFavourite	0.04432	0.43902	0.01946	-0.38784	0.42183	3.45	10.15	0.34 (0.34)	0.41 (0.75)

(A) (B) (C) (D)

Figure 2 - Représentation des lignes - Tanagra

SAS fournit bien les mêmes informations, mais en les disséminant dans plusieurs tableaux (Figure 3).

Summary Statistics for the Row Points

	Quality	Mass	Inertia
Healthy	0.7795	0.1114	0.0494
Nutritious	0.6772	0.1068	0.0310
GoodSummer	0.6321	0.1040	0.0623
GoodWinter	0.2743	0.0693	0.0640
Expensive	0.4220	0.0631	0.0694
QuickEasy	0.9041	0.0943	0.1195
Tasty	0.3689	0.1085	0.0114
Economical	0.7527	0.0472	0.1031
ForATreat	0.7152	0.0472	0.0695
ForWeekdays	0.8826	0.0767	0.0877
ForWeekends	0.7921	0.0841	0.0990
Tasteless	0.2893	0.0119	0.0256
TooLongToPrepare	0.9555	0.0313	0.1551
FamilyFavourite	0.7480	0.0443	0.0529

(A)

Row Coordinates

	Dim1	Dim2
Healthy	0.0866	-0.3459
Nutritious	-0.0086	-0.2689
GoodSummer	0.3488	-0.1328
GoodWinter	-0.2774	0.1271
Expensive	-0.2007	-0.3612
QuickEasy	0.6437	0.0848
Tasty	-0.0396	-0.1127
Economical	0.4169	0.6569
ForATreat	-0.6195	-0.0649
ForWeekdays	0.5068	0.3380
ForWeekends	-0.5600	0.1712
Tasteless	0.4510	0.1578
TooLongToPrepare	-1.2778	0.3340
FamilyFavourite	-0.3878	0.4218

(B)

Partial Contributions to Inertia for the Row Points

	Dim1	Dim2
Healthy	0.0043	0.1714
Nutritious	0.0000	0.0993
GoodSummer	0.0655	0.0236
GoodWinter	0.0276	0.0144
Expensive	0.0132	0.1059
QuickEasy	0.2024	0.0087
Tasty	0.0009	0.0177
Economical	0.0425	0.2618
ForATreat	0.0937	0.0026
ForWeekdays	0.1020	0.1127
ForWeekends	0.1366	0.0317
Tasteless	0.0126	0.0038
TooLongToPrepare	0.2642	0.0449
FamilyFavourite	0.0345	0.1015

(C)

Squared Cosines for the Row Points

	Dim1	Dim2
Healthy	0.0460	0.7335
Nutritious	0.0007	0.6765
GoodSummer	0.5521	0.0800
GoodWinter	0.2267	0.0476
Expensive	0.0995	0.3225
QuickEasy	0.8887	0.0154
Tasty	0.0406	0.3284
Economical	0.2161	0.5365
ForATreat	0.7075	0.0078
ForWeekdays	0.6109	0.2717
ForWeekends	0.7244	0.0677
Tasteless	0.2577	0.0315
TooLongToPrepare	0.8944	0.0611
FamilyFavourite	0.3426	0.4053

(D)

Figure 3 - Représentation des lignes - SAS 9.3

Les modalités à inertie élevée vont souvent conditionner les résultats sur les premiers axes. Ce n'est pas un problème en soi. Il faut en avoir conscience simplement pour une lecture distanciée du

tableau. En l'occurrence, « TooLongToPrepare », « QuickEasy » et « Economical » vont quasiment à eux seuls définir les deux premiers facteurs si on se réfère aux contributions.

Les COS² indiquent la qualité de représentation individuelle et cumulée des modalités sur les K premiers facteurs. Dans notre exemple, seules les modalités « GoodWinter » et « TasteLess » sont mal résumées par les 2 premiers axes (moins que les autres modalités tout du moins).

Pour chaque facteur, Tanagra met en évidence les coordonnées des modalités répondant aux conditions suivantes : (CTR > 1/L)[8] et (COS² > $1/K_{max}$) c.-à-d. la modalité contribue plus que la moyenne (non pondérée), et l'information qu'elle véhicule est concentrée sur le facteur (plus que la moyenne). L'idée est d'attirer l'œil de l'utilisateur sur les éléments les marquants du tableau qui est, reconnaissons le, assez rébarbatif à lire.

Dans notre exemple, nous constatons que le premier facteur est défini par l'opposition entre (ForAtreat, ForWeekEnds, TooLongToPrepare) et (QuickEasy, ForWeekDays)[9]. Sur le second, nous avons une opposition entre (Healthy, Nutitrious, Expensive) et (Economical, FamilyFavourite, ForWeekADays). Attention, la lecture n'est pas aussi évidente qu'on pourrait le croire…

3.4.4 Représentation des colonnes

La représentation des colonnes obéit à la même logique.

Columns analysis

Row Characterization				Coord.		Contributions (%)		COS²	
Values	Weight	Sq. Dist.	Inertia	coord 1	coord 2	ctr 1	ctr 2	cos² 1	cos² 2
Cereals	0.15568	0.48968	0.07623	0.56272	0.35791	25.53	25.66	0.65 (0.65)	0.26 (0.91)
Muesli	0.13068	0.33926	0.04433	0.31310	-0.31869	6.63	17.08	0.29 (0.29)	0.30 (0.59)
Porridge	0.10739	0.35450	0.03807	-0.05363	0.21310	0.16	6.27	0.01 (0.01)	0.13 (0.14)
BaconEggs	0.17727	0.66963	0.11871	-0.78344	0.14814	56.35	5.00	0.92 (0.92)	0.03 (0.95)
ToastTea	0.06364	0.37744	0.02402	0.17534	0.44702	1.01	16.36	0.08 (0.08)	0.53 (0.61)
FreshFruit	0.12386	0.19655	0.02434	0.24619	-0.24493	3.89	9.56	0.31 (0.31)	0.31 (0.61)
StewedFruit	0.11534	0.19120	0.02205	-0.31540	-0.24281	5.94	8.75	0.52 (0.52)	0.31 (0.83)
Yoghurt	0.12614	0.15878	0.02003	0.08599	-0.26416	0.48	11.32	0.05 (0.05)	0.44 (0.49)

(A) (B) (C) (D)

Figure 4 - Représentation des colonnes - Tanagra

D'emblée, on sait que « BaconEggs » et, dans une moindre mesure, « Cereals » vont beaucoup peser dans l'étude. En effet, BaconEggs compte pour 32.3% (0.11881 / 0.3678) de l'inertie totale, « Cereals » pour 20.7% (0.07623 / 0.3678). Et effectivement, ces modalités déterminent en grande partie les deux premiers facteurs.

[8] Nous avons préféré cette règle à (CTR > poids de la modalité) (Saporta, 2006 ; page 207) pour éviter de mettre en évidence des modalités sous-représentées qui, de toute manière, contribuent faiblement au χ^2_{global}. Par exemple, avec cette seconde condition, nous aurions du mettre « tasteless » en évidence sur le premier axe (CTR = 1.26%, poids = 1.19%). Or, cette modalité pèse finalement très peu dans l'analyse. Elle ne correspond ni à une attraction ni à une répulsion réellement marquée avec l'une des modalités colonnes.

[9] Voilà la vérité cruelle de notre existence : se faire plaisir, prendre son temps, ça n'est pas possible en semaine.

La règle de mise en surbrillance des coordonnées devient (CTR > 1 / C) et (COS^2 > $1/K_{max}$) pour la représentation des colonnes.

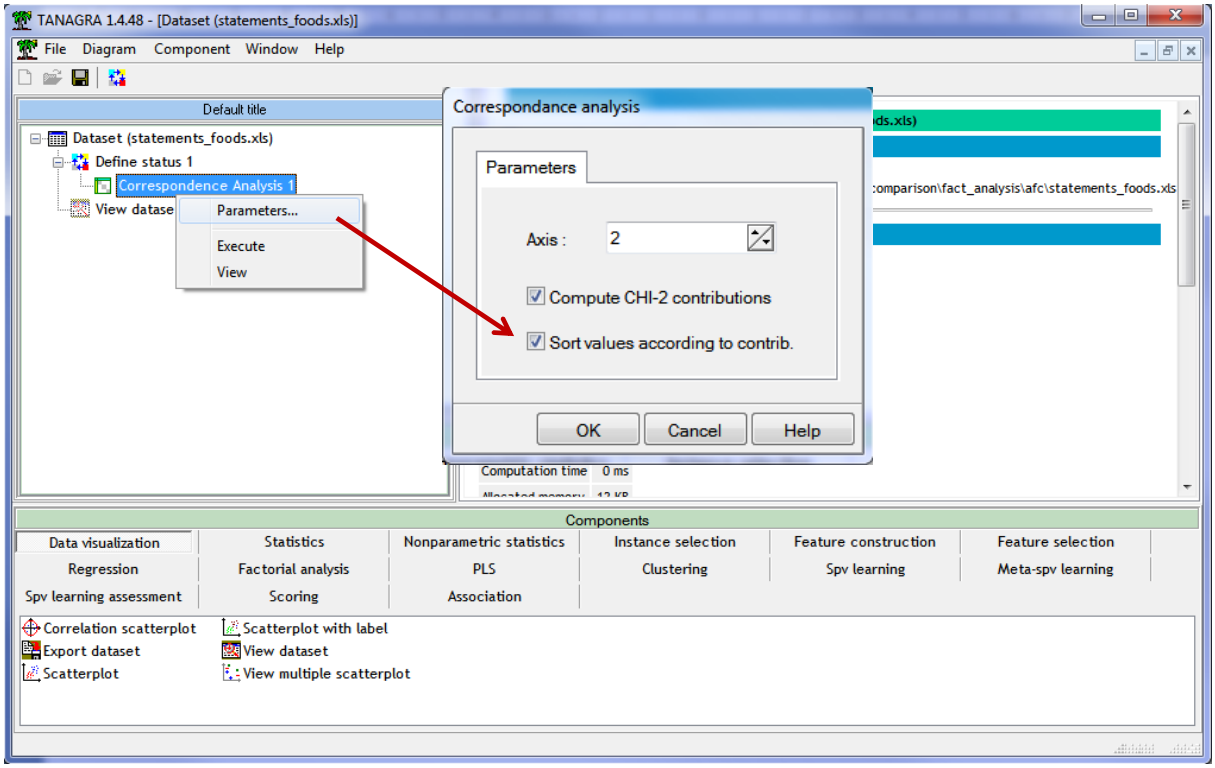
Concernant SAS, les résultats sont toujours répartis sur plusieurs tableaux. Ils sont cohérents avec ceux de Tanagra bien sûr. Il ne m'a pas semblé nécessaire de remettre la copie d'écran ici.

3.4.5 Trier les représentations lignes / colonnes selon les contributions

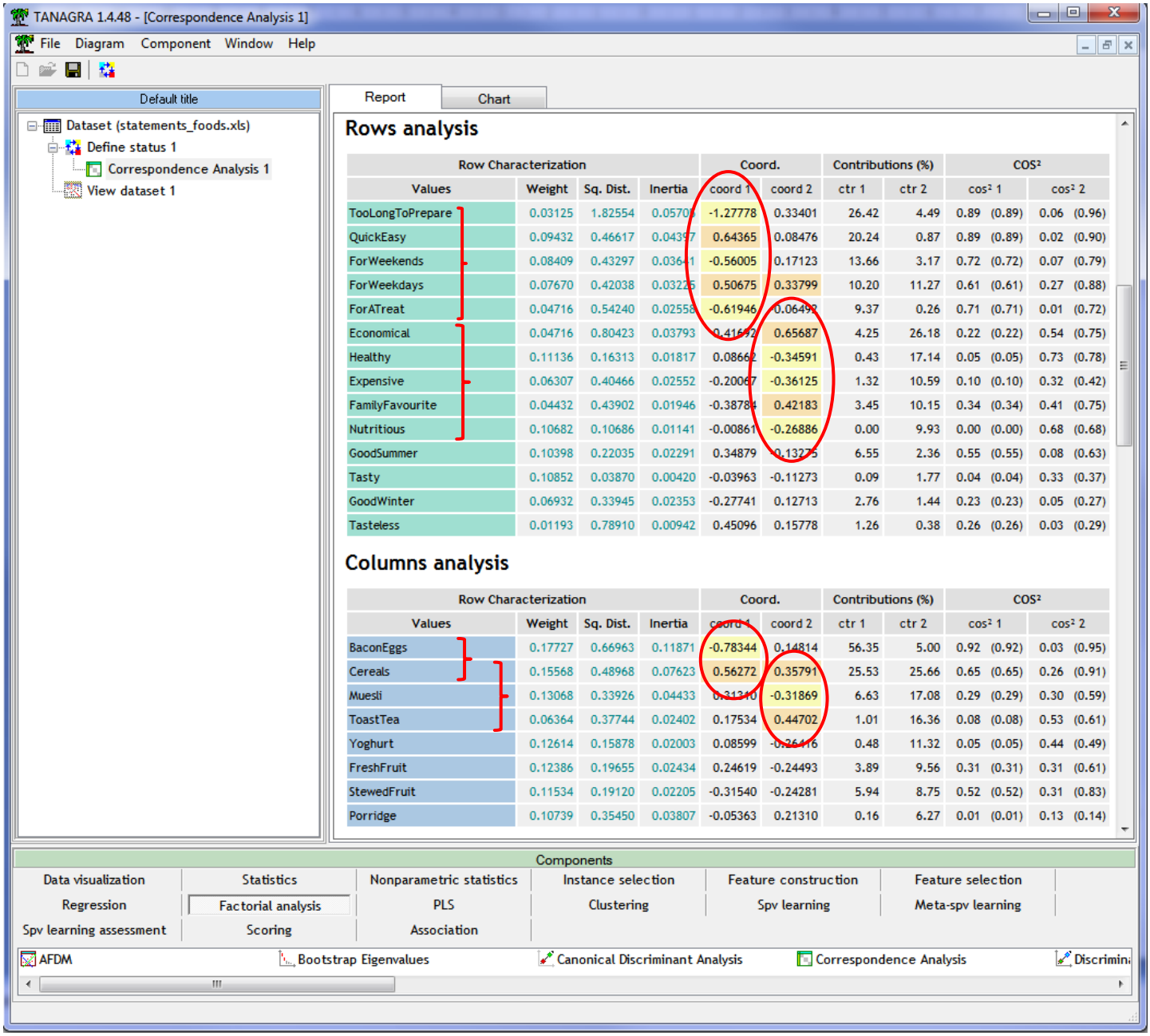
Lorsque le nombre de modalités lignes et colonnes est élevé, déchiffrer les grands tableaux de résultats devient particulièrement ardu. Surtout lorsque l'on cherche à identifier les attractions (répulsions) entre modalités. D'où le succès de la représentation graphique simultanée d'ailleurs. Néanmoins, cette dernière peut se révéler trompeuse parfois. On conseille généralement de revenir sur les éléments chiffrés (CTR et COS^2) pour identifier les vraies interactions.

Mettre en surbrillance certaines cases est déjà un premier pas vers une meilleure appréhension des résultats. Une seconde piste consiste à trier les représentations selon les contributions aux axes des modalités. Mais comme l'information est décomposée sur plusieurs facteurs, on ne peut pas se contenter d'effectuer un tri sur le premier uniquement. Tanagra s'appuie sur la stratégie suivante : pour le premier axe, il identifie les variables qui contribuent plus que la moyenne non pondérée (1/L pour les lignes, 1/C pour les colonnes), il trie les modalités selon leurs contributions ; parmi les modalités restantes, il identifie celles qui pèsent sur le second axe, il les trie de nouveau, etc. Ainsi, nous avons une représentation en cascade (en diagonale) qui permet d'identifier très rapidement la nature des informations véhiculées par les facteurs.

Nous revenons sur le composant « Correspondence Analysis », nous actionnons le menu PARAMETERS et nous cochons l'option « Sort values according to contrib. ».



Nous validons et nous actionnons le menu VIEW.



Nous distinguons mieux maintenant les « blocs » de modalités associées aux axes. Bien évidemment, le tri n'est pas parfait. Certaines modalités peuvent contribuer fortement sur deux (ou plusieurs) facteurs, la mise en surbrillance permet de les identifier rapidement.

3.4.6 Représentation simultanée

La popularité des analyses factorielles repose pour beaucoup sur les « cartes » graphiques. On a l'impression de tout comprendre en un coup d'œil. Il est toujours très stimulant de pouvoir associer visuellement des points. Certains outils proposent même la visualisation 3D. Alors là, on passe un temps fou à tourner le graphique dans tous les sens, à voir le nuage par en dessous, sur le côté, de travers, etc., en perdant complètement le fil de l'analyse au passage.

Tanagra propose la représentation pseudo-barycentrique. Nous activons l'onglet CHART de la fenêtre d'affichage des résultats. Nous sélectionnons « Axis 1 » en abscisse, « Axis 2 » en ordonnée.

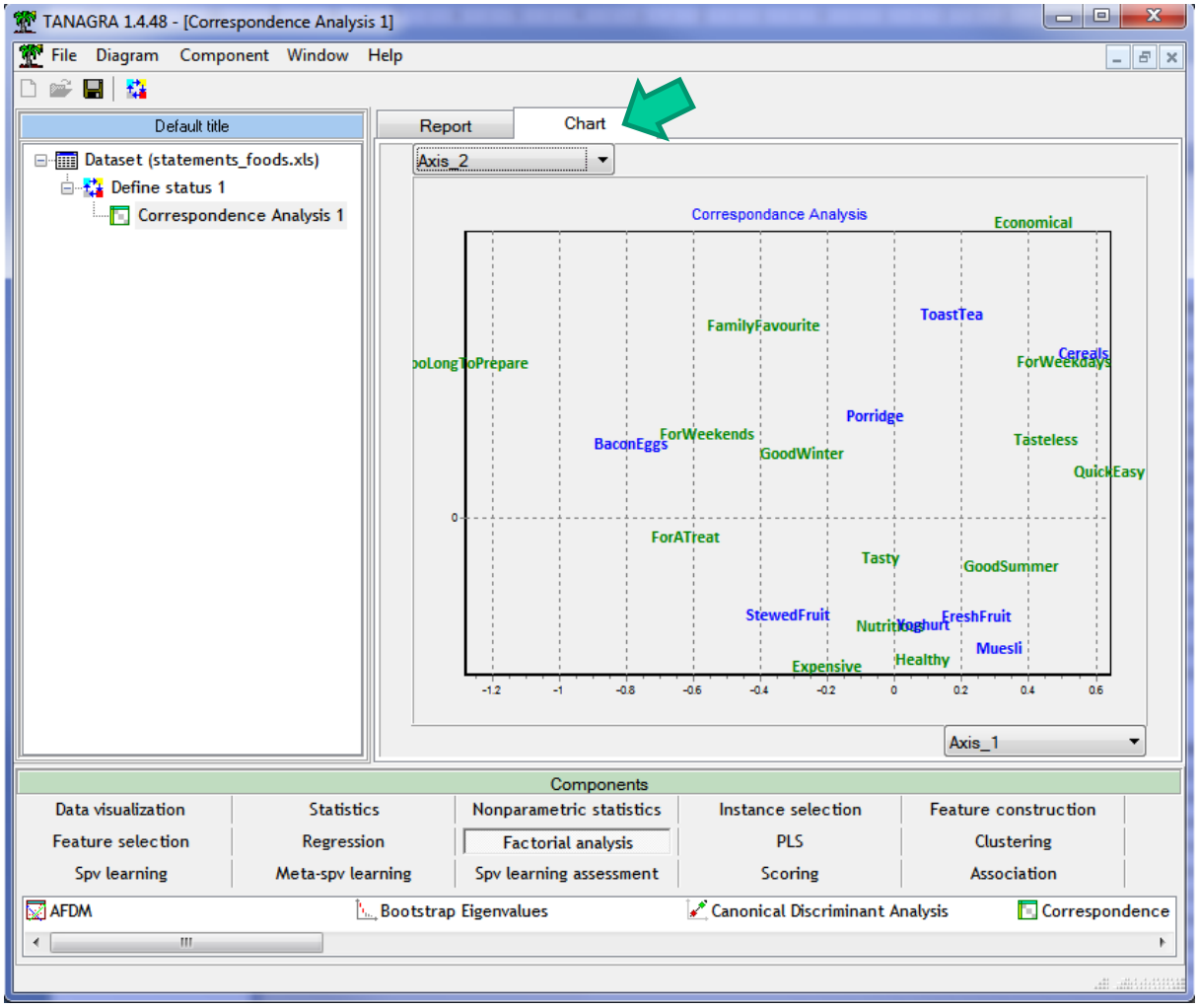


Figure 5 - Représentation simultanée - Tanagra

Tous les auteurs s'accordent à dire qu'il faut se méfier des proximités fallacieuses dans ce graphique, surtout lorsque l'on se rapproche de l'origine. Par exemple, le rapprochement entre « nutritious » et « healthy » d'une part, et « yoghurt » d'autre part, ne correspond à rien lorsqu'on analyse finement le tableau de contingence (voir section 3.4.7).

En revanche, lorsque nous sommes sur les extrémités, pour les points à forts CTR et COS², les attractions et répulsions entre les modalités se traduisent mieux en proximités ou éloignement géographiques. Tout en sachant qu'il faut toujours se cantonner à une vision relative des points (une modalité ligne est positionnée globalement par rapport à l'ensemble des modalités colonnes) en raison de la relation quasi-barycentrique (Saporta, 2006 ; page 208).

Prenons un exemple simple pour préciser cette idée. Nous reprenons ici les coordonnées des points colonnes sur le premier axe factoriel.

Coord.Colonnes.Axe.1							
0.563	0.313	-0.054	-0.783	0.175	0.246	-0.315	0.086
Cereals	Muesli	Porridge	BaconEggs	ToastTea	FreshFruit	StewedFruit	Yoghurt

Voyons maintenant comment positionner « TooLongToPrepare » par rapport à ces points. Nous avons besoin pour cela de son profil ligne.

Profil ligne - TooLongToPrepare								
	Cereals	Muesli	Porridge	BaconEggs	ToastTea	FreshFruit	StewedFruit	Yoghurt
TooLongToPrepare	0.00	0.00	0.16	0.64	0.02	0.00	0.18	0.00

Pas besoin d'être prix Nobel pour comprendre que la modalité « BaconEggs » va peser lourdement dans la balance, ***mais elle ne sera pas la seule*** ! Le calcul de la coordonnée de « TooLongToPrepare » nécessite la valeur propre du premier axe λ_1 = 0.1931. Ainsi, nous obtenons (*en excluant les coefficients nuls*) :

$$\frac{1}{\sqrt{0.1931}}[0.16 \times (-0.054) + 0.64 \times (-0.783) + 0.02 \times 0.175 + 0.18 \times (-0.315)] = -1.278$$

C'est bien la coordonnée que nous lisons dans la représentation des lignes (Figure 2). Les modalités (Pooridge, BaconEggs, StewedFruit) l'ont tiré vers les valeurs négatives, l'emportant sur l'influence positive de (ToastTea).

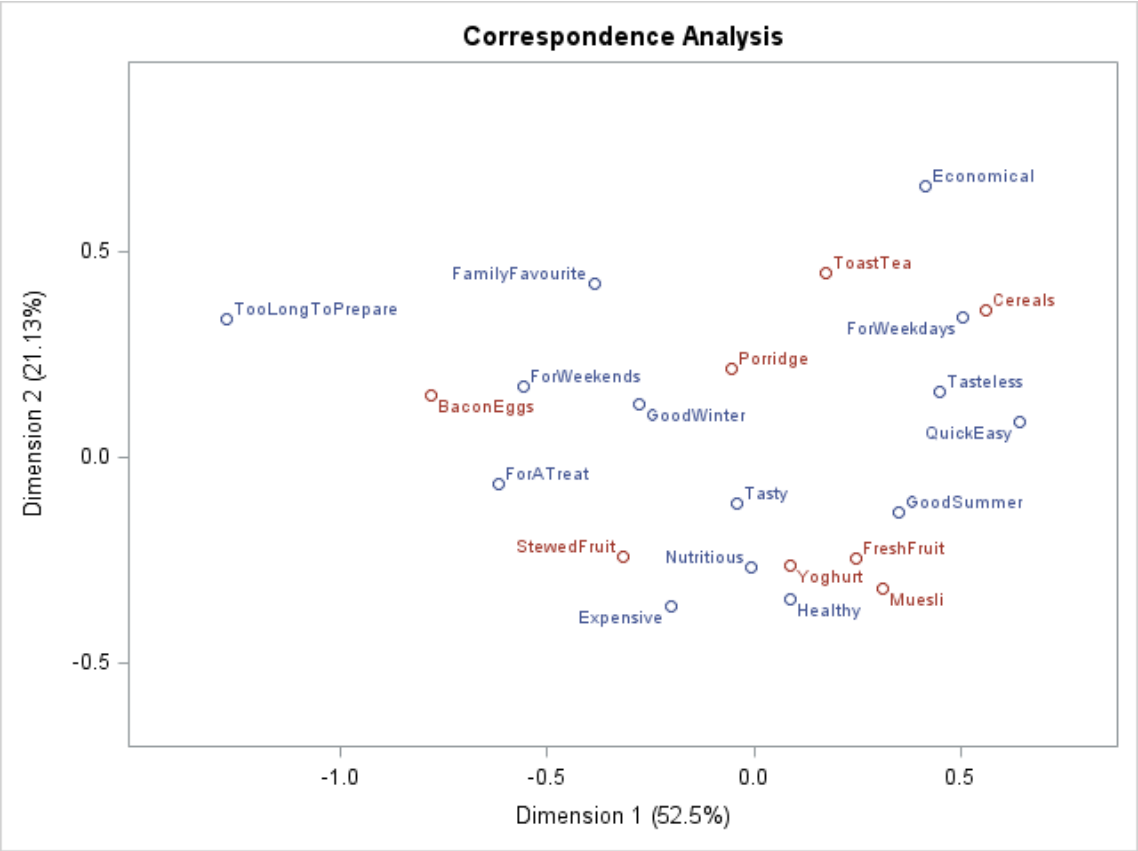


Figure 6 - Représentation simultanée - SAS

Nous retrouvons la même représentation simultanée sous SAS (Figure 6).

3.4.7 Contributions au KHI-2

Tanagra fournit en option le tableau des contributions au χ^2_{global}. Pour chaque combinaison des modalités lignes / colonnes, Tanagra propose : l'effectif observé, l'effectif théorique (sous indépendance), le résidu standardisé[10], la contribution absolue au χ^2 [(+) attraction ou (-) répulsion],

[10] Qui suit très approximativement une loi normale. On peut donc considérer qu'une valeur supérieure à 2 en valeur absolue indique une attraction ou répulsion significative pour un test à 5%. Le résidu standardisé me semble plus

et la contribution relative (en pourcentage). Il trie le tableau de manière décroissante pour que les informations les plus marquantes apparaissent au premier plan. Seules les contributions supérieures à la moyenne [contribution relative > 1 / (L x C)] sont affichées.

CHI-2 contributions

Id	Row	Column	Value	Expected	Std.Resid.	Contrib.		%
1	TooLongToPrepare	BaconEggs	35	9.8	8.09	(+)	65.39	10.10
2	ForWeekends	BaconEggs	56	26.2	5.81	(+)	33.77	5.22
3	ForWeekdays	Cereals	47	21.0	5.67	(+)	32.12	4.96
4	QuickEasy	Cereals	54	25.8	5.54	(+)	30.68	4.74
5	GoodWinter	Porridge	32	13.1	5.22	(+)	27.26	4.21
6	QuickEasy	BaconEggs	2	29.4	-5.06	(-)	25.56	3.95
7	Expensive	Muesli	33	14.5	4.86	(+)	23.58	3.64
8	Economical	ToastTea	16	5.3	4.66	(+)	21.75	3.36
9	FamilyFavourite	BaconEggs	31	13.8	4.62	(+)	21.33	3.29
10	ForATreat	BaconEggs	31	14.7	4.25	(+)	18.03	2.78
11	Economical	Porridge	20	8.9	3.71	(+)	13.79	2.13
12	GoodSummer	BaconEggs	13	32.4	-3.41	(-)	11.65	1.80
13	ForWeekends	Muesli	5	19.3	-3.26	(-)	10.63	1.64
14	Economical	Cereals	24	12.9	3.08	(+)	9.50	1.47
15	Economical	BaconEggs	3	14.7	-3.05	(-)	9.33	1.44
16	ForWeekdays	BaconEggs	9	23.9	-3.05	(-)	9.32	1.44
17	GoodSummer	FreshFruit	37	22.7	3.01	(+)	9.06	1.40
18	Healthy	Cereals	14	30.5	-2.99	(-)	8.94	1.38
19	TooLongToPrepare	Cereals	0	8.6	-2.93	(-)	8.56	1.32
20	Healthy	BaconEggs	18	34.7	-2.84	(-)	8.07	1.25
21	Nutritious	Cereals	14	29.3	-2.82	(-)	7.96	1.23
22	Expensive	Cereals	6	17.3	-2.71	(-)	7.36	1.14
23	TooLongToPrepare	Muesli	0	7.2	-2.68	(-)	7.19	1.11
24	TooLongToPrepare	Yoghurt	0	6.9	-2.63	(-)	6.94	1.07
25	Tasteless	Cereals	8	3.3	2.62	(+)	6.85	1.06
26	TooLongToPrepare	FreshFruit	0	6.8	-2.61	(-)	6.81	1.05
27	QuickEasy	StewedFruit	8	19.1	-2.55	(-)	6.49	1.00
28	GoodSummer	Cereals	42	28.5	2.53	(+)	6.41	0.99
29	GoodSummer	Yoghurt	35	23.1	2.48	(+)	6.15	0.95
30	Healthy	Muesli	38	25.6	2.45	(+)	5.99	0.93
31	QuickEasy	Muesli	33	21.7	2.43	(+)	5.89	0.91
32	ForWeekdays	StewedFruit	6	15.6	-2.43	(-)	5.88	0.91

Figure 7 - Contributions au χ^2 - Tanagra

Très rapidement, on se rend compte que l'information portée par le tableau de contingence est en réalité « écrasée » par l'attraction de « BaconEggs » avec « TooLongToPrepare » (10.10%).

intéressant que le résidu ajusté - qui s'approxime pourtant mieux avec la loi normale - dans le contexte de l'analyse factorielle. En effet, l'analyse des correspondances correspond aussi à la décomposition en valeurs singulières de la matrice des résidus standardisés (ex. http://www.unesco.org/webworld/idams/advguide/Chapt6_5_8.htm)

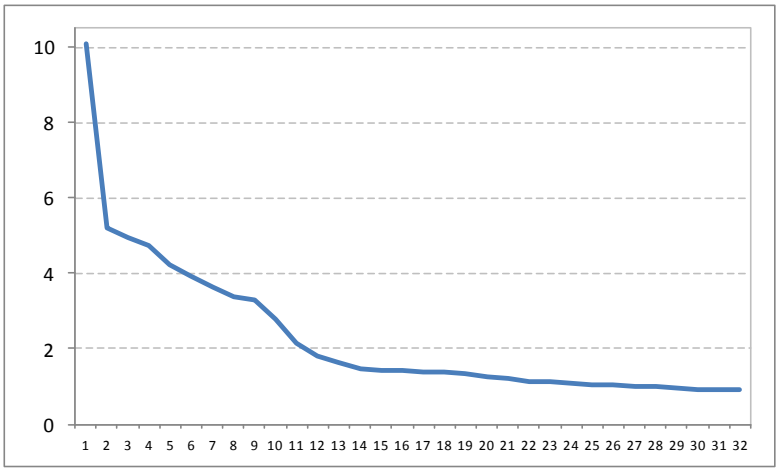


Figure 8 - Décroissance de la contribution au KHI-2

On s'en rend mieux compte en traçant la courbe de décroissance des contributions (Figure 8, graphique non fourni par Tanagra). *Avec le recul, on peut se demander - à juste titre - s'il n'est pas judicieux de mettre en colonne supplémentaire « BaconEggs » pour obtenir une vision plus juste des relations entre les modalités*[11].

L'idée que le yaourt (yoghourt) soit nourrissant (nutritious) et bon à la santé (healthy) semblait s'imposer dans la représentation simultanée (Figure 6). Elle correspond plus ou moins à ce que l'on croit savoir de l'alimentation d'ailleurs (merci la pub). Mais est-ce que réellement cette information est présente dans le tableau de contingence que nous analysons ? Nous explorons le tableau des contributions. Force est de constater que ces attractions n'apparaissent pas (tout du moins parmi les contributions supérieures à la moyenne). Avec l'outil filtre automatique d'Excel (après avoir copié les résultats), nous constatons que Yoghourt se distingue avant tout par sa répulsion avec TooLongToPrepare et, a contrario, son attraction avec GoodSummer. Et c'est tout !

Id	Row	Column	Value	Expected	Std.Resid	Contrib.		%
24	TooLongToPrepare	Yoghurt	0	6.9	-2.63	(-)	6.94	1.07
29	GoodSummer	Yoghurt	35	23.1	2.48	(+)	6.15	0.95

3.4.8 Projection des lignes supplémentaires

Calcul des coordonnées. L'analyse des correspondances nous permet de projeter des points supplémentaires dans le repère factoriel. Dans son tutoriel, Bendixen (page 11) décrit un tableau croisant la fréquence de consommation avec le type d'aliment.

	Cereals	Muesli	Porridge	BaconEggs	ToastTea	FreshFruit	StewedFruit	Yoghurt
I	24	3	4	8	18	2	9	11
II	58	15	8	13	16	10	10	29
III	6	10	12	46	8	14	15	8
IV	2	4	28	9	4	47	4	2
V	10	68	48	24	54	27	62	50

I : Daily ; II : Several times per week ; III : Several times per month ; IV : Every few months ; V : Never.

[11] Certaines modalités, masquées jusqu'à présent, joueront alors un rôle plus actif dans l'analyse.

Comment positionner ces points lignes (I, II, III, IV, V) par rapport aux modalités existantes ?

A partir de la relation quasi-barycentrique, il est possible de les positionner sur les 2 premiers facteurs à l'aide des coordonnées des modalités colonnes[12]. Mais l'idée n'est pas toujours très bien comprise, surtout que la plupart des outils proposent de le faire directement en interne, masquant le mécanisme de projection aux utilisateurs. Cette solution, forte intéressante au demeurant, ne permet pas un déploiement des fonctions scores dans des outils externes

Tanagra propose une approche différente. Il fournit les coefficients de fonctions score c.-à-d. les fonctions permettant de projeter les points supplémentaires dans le repère factoriel. Ces coefficients peuvent être exportés facilement dans d'autres outils (ex. dans le tableur Excel), permettant de calculer les coordonnées des points supplémentaires à partir de leurs profils.

Factor score coefficients for supplementary row

From column values (relative frequency)

Column value	Factor 1	Factor 2
Cereals	1.280579	1.283725
Muesli	0.712525	-1.143072
Porridge	-0.122041	0.764323
BaconEggs	-1.782883	0.531349
ToastTea	0.399026	1.603371
FreshFruit	0.560261	-0.878505
StewedFruit	-0.717755	-0.870916
Yoghurt	0.195685	-0.947476

Figure 9 - Fonctions scores pour les modalités lignes supplémentaires

Reprenons les fréquences de consommation, après leur transformation en profils lignes.

	Cereals	Muesli	Porridge	BaconEggs	ToastTea	FreshFruit	StewedFruit	Yoghurt
I	0.304	0.038	0.051	0.101	0.228	0.025	0.114	0.139
II	0.365	0.094	0.050	0.082	0.101	0.063	0.063	0.182
III	0.050	0.084	0.101	0.387	0.067	0.118	0.126	0.067
IV	0.020	0.040	0.280	0.090	0.040	0.470	0.040	0.020
V	0.029	0.198	0.140	0.070	0.157	0.079	0.181	0.146

Figure 10 - Profils lignes des modalités lignes supplémentaires

Nous pouvons dès lors calculer leurs coordonnées pour les deux premiers axes.

	Scores	
	Factor.1	Factor.2
I	0.280	0.551
II	0.448	0.321
III	-0.562	0.082
IV	0.114	-0.161
V	0.042	-0.157

Figure 11 - Coordonnées des modalités lignes supplémentaires

Détaillons le calcul pour « III : Several times per month » sur le 1er facteur.

[12] Voir également : Tutoriel Tanagra, « AFC – Association médias et professions » (pages 8 et 9).

C(III, Factor 1) = 1.280579 x 0.050 + 0.712525 x 0.084 + (-0.122041) x 0.101 + (-1.782883) x 0.387 + 0.399026 x 0.067 + 0.560261 x 0.118 + (-0.717755) x 0.126 + 0.195685 x 0.067 = **-0.562**

Ce qui le situerait plutôt du côté de (TooLongToPrepare, ForATreat, ForWeekends). Cela semble logique dans la mesure où « BaconEggs » est l'aliment le plus prisé dans ce cas (38.7 % dixit le tableau des profils lignes, Figure 10).

Qualité de la représentation. Pour apprécier la qualité de la représentation des points sur chaque facteur, nous calculons leur COS² c.-à-d. leur part d'inertie associée au facteur[13]. Pour ce faire, nous devons tout d'abord calculer leur distance au centre de gravité représenté par la marge colonne du tableau (ex. 274 réponses ont désigné l'aliment «Cereals », elles représentent 15.6% des réponses).

Marge colonne	Cereals	Muesli	Porridge	BaconEggs	ToastTea	FreshFruit	StewedFruit	Yoghurt	Total
Effectif	274	230	189	312	112	218	203	222	1760
Profil	0.156	0.131	0.107	0.177	0.064	0.124	0.115	0.126	-

Figure 12 - Marge colonne du tableau de contingence

Nous calculons la distance (du χ^2) de chaque point supplémentaire à cette marge.

	Dist²
I	0.773
II	0.473
III	0.364
IV	1.616
V	0.407

Figure 13 - Carré de la distance au barycentre des lignes supplémentaires

Voyons le détail pour le point « III : Several times per month » (à partir des informations apparaissant dans Figure 10 et Figure 12) :

$$D^2(III, Marge) = \frac{(0.050 - 0.156)^2}{0.156} + \frac{(0.084 - 0.131)^2}{0.131} + \cdots + \frac{(0.050 - 0.126)^2}{0.126} = 0.364$$

Dès lors, le COS² du point « III » sur le premier facteur est déduite de sa coordonnée sur l'axe et la distance au barycentre :

$$COS^2(III, Facteur\ 1) = \frac{(-0.562)^2}{0.364} = 0.866$$

De la même manière, pour l'ensemble des points supplémentaire sur les deux facteurs, nous avons :

Cos²	Factor.1	Factor.2
I	0.101	0.393
II	0.425	0.218
III	0.866	0.019
IV	0.008	0.016
V	0.004	0.061

Figure 14 - Qualité de représentation des lignes supplémentaires

[13] Plus ou moins. En effet, la modalité n'ayant pas participé à la construction des facteurs, la somme totale des COS² (sur l'ensemble des axes) n'est pas strictement égale à 1.

L'information véhiculée par « III : Several times per month » est particulièrement bien traduite par le 1er facteur (cos^2 = 0.866). De fait, son rapprochement avec les modalités (TooLongToPrepare, ForATreat, ForWeekends) - qui sont dans son voisinage immédiat - est tout à fait crédible ; il en est de même pour son attraction avec « BaconEggs ».

3.4.9 Projection des colonnes supplémentaires

Le mécanisme de projection peut également s'appliquer aux colonnes supplémentaires. Tanagra fournit les coefficients de fonctions scores à cet égard.

Factor score coefficients for supplementary column

From row values (relative frequency)

Row value	Factor 1	Factor 2
Healthy	0.197122	-1.240706
Nutritious	-0.019587	-0.964348
GoodSummer	0.793731	-0.476147
GoodWinter	-0.631303	0.455973
Expensive	-0.456668	-1.295717
QuickEasy	1.464761	0.304003
Tasty	-0.090190	-0.404344
Economical	0.948796	2.356032
ForATreat	-1.409708	-0.232841
ForWeekdays	1.153220	1.212284
ForWeekends	-1.274503	0.614179
Tasteless	1.026245	0.565924
TooLongToPrepare	-2.907840	1.198004
FamilyFavourite	-0.882613	1.513020

Figure 15 - Fonctions scores pour les modalités colonnes supplémentaires - Tanagra

On pourra ainsi positionner un nouveau type d'aliment à partir de son profil colonne.

4 Analyse des correspondances avec R

La multiplicité des packages est à la fois un atout et une faiblesse de R. Un atout parce que nous disposons de plusieurs points de vue. Cela ne peut qu'enrichir l'analyse. Une faiblesse parce qu'obtenir des résultats apparemment dissemblables pour un même problème est toujours compliqué, surtout lorsqu'on a peu de recul par rapport aux techniques. Parfois, les sorties de certains packages ne correspondent pas du tout aux présentations que l'on retrouve dans les ouvrages de référence. Cela peut rapidement jeter la confusion dans l'esprit de l'utilisateur.

Dans cette section, nous utiliserons deux packages très prisées pour l'analyse des correspondances. Nous comparerons les résultats entre eux d'une part, avec SAS et Tanagra d'autre part. Nous constaterons que, moyennant quelques transformations intermédiaires dans certains cas, nous retrouvons exactement les mêmes indicateurs numériques. C'est ce qui importe finalement.

4.1 Test du KHI-2

Dans un premier temps, nous chargeons le fichier de données et nous réalisons le test d'indépendance du χ^2.

```
foods <- read.table(file="statements_foods.txt",header=T,sep="\t",row.names=1)
#chi-squared test
print(chisq.test(foods))
```

R affiche un message d'avertissement parce que les effectifs théoriques dans plusieurs cellules sont faibles, inférieures à 5. Il s'agit surtout des effectifs associés à la modalité « Tasteless » qui est sous-représentée, elle correspond à 1.2% des réponses.

```
> print(chisq.test(foods))

        Pearson's Chi-squared test

data:  foods
X-squared = 647.3121, df = 91, p-value < 2.2e-16

Message d'avis :
In chisq.test(foods) : Chi-squared approximation may be incorrect
```

4.2 Le package 'ca'

La package « ca[14] » est du à Michael Greenacre et Oleg Nenadic. Le premier est un acteur très dynamique de l'analyse factorielle avec plusieurs ouvrages à son actif.

4.2.1 Analyse des correspondances – 2 facteurs

Après avoir chargé la librairie, nous lançons l'analyse en demandant 2 facteurs.

```
#perform the correspondence analysis
library(ca)
foods.ca <- ca(foods,nd=2)
#eigen values and cumulative proportion of variance explained in percentage
print(cbind(foods.ca$sv^2,(100.0*cumsum(foods.ca$sv^2)/sum(foods.ca$sv^2))))
```

Nous obtenons les valeurs propres et l'évolution de la variance expliquée en pourcentage cumulé.

```
             [,1]      [,2]
[1,] 0.193094526  52.50116
[2,] 0.077730798  73.63567
[3,] 0.043854131  85.55932
[4,] 0.032804216  94.47858
[5,] 0.012256794  97.81112
[6,] 0.005687400  99.35749
[7,] 0.002363091 100.00000
```

4.2.2 Représentation des lignes

Nous exploitons l'objet '**foods.ca**' (de type « ca ») pour accéder aux différents résultats de l'analyse. Ainsi pour les indicateurs (poids, carré de la distance au centre de gravité, inertie, coordonnées) des modalités lignes :

```
#row analysis
attach(foods.ca)
row.ca <- round(cbind(rowmass,rowdist^2,rowinertia,rowcoord[,1]*sv[1],rowcoord[,2]*sv[2]),5)
```

[14] Voir aussi http://www.carme-n.org/

```
colnames(row.ca) <- c("weight","sq.dist","inertia","coord.1","coord.2")
rownames(row.ca) <- rownames
print(row.ca)
```

Nous obtenons :

	weight	sq.dist	inertia	coord.1	coord.2
Healthy	0.11136	0.16313	0.01817	0.08662	0.34591
Nutritious	0.10682	0.10686	0.01141	-0.00861	0.26886
GoodSummer	0.10398	0.22035	0.02291	0.34879	0.13275
GoodWinter	0.06932	0.33945	0.02353	-0.27741	-0.12713
Expensive	0.06307	0.40466	0.02552	-0.20067	0.36125
QuickEasy	0.09432	0.46617	0.04397	0.64365	-0.08476
Tasty	0.10852	0.03870	0.00420	-0.03963	0.11273
Economical	0.04716	0.80423	0.03793	0.41693	-0.65687
ForATreat	0.04716	0.54240	0.02558	-0.61946	0.06492
ForWeekdays	0.07670	0.42038	0.03225	0.50675	-0.33799
ForWeekends	0.08409	0.43297	0.03641	-0.56005	-0.17123
Tasteless	0.01193	0.78910	0.00942	0.45096	-0.15778
TooLongToPrepare	0.03125	1.82554	0.05705	-1.27778	-0.33401
FamilyFavourite	0.04432	0.43902	0.01946	-0.38784	-0.42183

Figure 16 - Représentation des lignes - Package 'ca'

Valeurs que nous pouvons rapprocher avec ceux de Tanagra (Figure 2).

4.2.3 Représentation des colonnes

```
#column analysis
col.ca <- round(cbind(colmass,coldist^2,colinertia,colcoord[,1]*sv[1],colcoord[,2]*sv[2]),5)
colnames(col.ca) <- c("weight","sq.dist","inertia","coord.1","coord.2")
rownames(col.ca) <- colnames
print(col.ca)
```

De même, pour la représentation des lignes, les résultats (Figure 17) sont cohérents avec ceux de Tanagra (Figure 4).

	weight	sq.dist	inertia	coord.1	coord.2
Cereals	0.15568	0.48968	0.07623	0.56272	-0.35791
Muesli	0.13068	0.33926	0.04433	0.31310	0.31869
Porridge	0.10739	0.35450	0.03807	-0.05363	-0.21310
BaconEggs	0.17727	0.66963	0.11871	-0.78344	-0.14814
ToastTea	0.06364	0.37744	0.02402	0.17534	-0.44702
FreshFruit	0.12386	0.19655	0.02434	0.24619	0.24493
StewedFruit	0.11534	0.19120	0.02205	-0.31540	0.24281
Yoghurt	0.12614	0.15878	0.02003	0.08599	0.26416

Figure 17 - Représentation des colonnes - Package 'ca'

4.2.4 Carte des modalités

La procédure plot() est surchargée pour prendre en compte directement la représentation simultanée à partir de l'objet « analyse des correspondances ».

```
#plot rows and columns
plot(foods.ca)
```

Nous retrouvons le positionnement relatif des modalités.

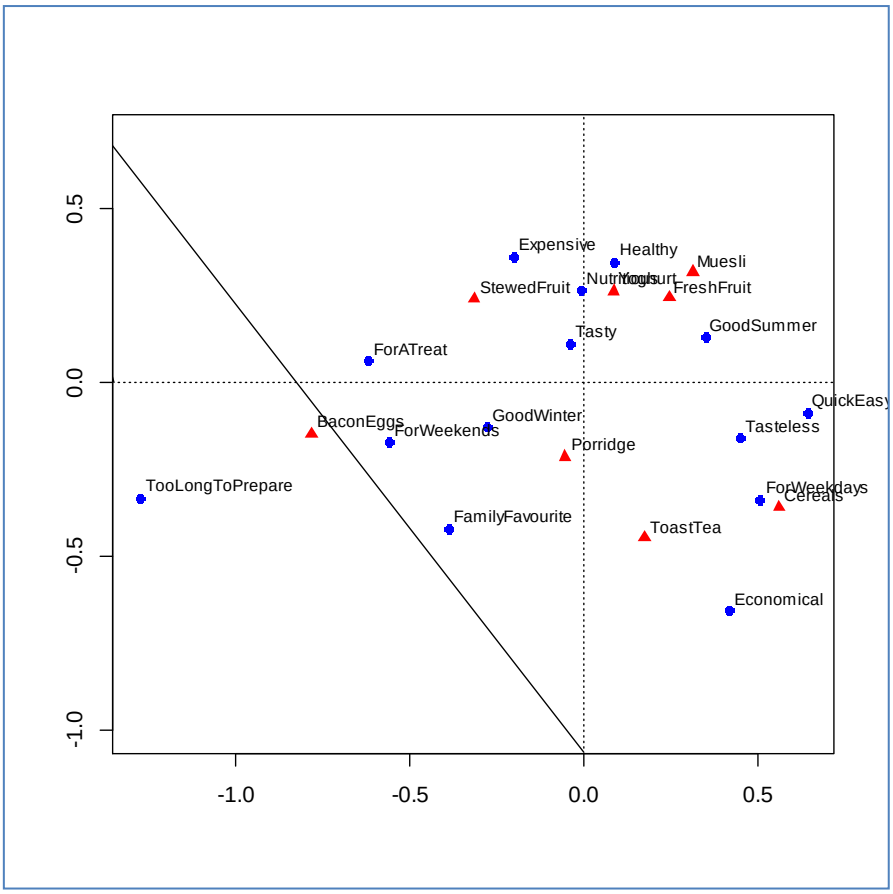


Figure 18 - Représentation simultanée - Package 'ca'

4.3 Le package 'ade4'

La librairie « ade4 » est issue du logiciel ADE-4, disponible sur le web depuis de nombreuses années. Son intégration dans R via le système des packages a été un virage très important qui lui assure une très large diffusion. Le sérieux du travail effectué en fait un outil de référence. Une très abondante documentation sous forme de tutoriels est accessible en ligne[15].

4.3.1 Analyse - Représentation des lignes et des colonnes

Nous lançons les calculs avec la procédure **dudi.coa()**[16]. L'objet qui en découle nous donne accès aux résultats de l'analyse.

```
library(ade4)
foods.coa <- dudi.coa(foods,scannf=F,nf=2)

#eigen values and cumulative proportion of variance explained in percentage
print(round(cbind(foods.coa$eig,100.0*cumsum(foods.coa$eig)/sum(foods.coa$eig)),4))
```

[15] http://pbil.univ-lyon1.fr/R/

[16] Cf. André Bouchier, « AFC Simple », accessible sur le site « R & Statistiques » (http://rstat.ouvaton.org/).

```
#row analysis - coordinates and contributions
print(cbind(foods.coa$li,inertia.dudi(foods.coa,row.inertia=T)$row.abs))

#column analysis - coordinates and contributions
print(cbind(foods.coa$co,inertia.dudi(foods.coa,col.inertia=T)$col.abs))
```

Nous retrouvons les résultats des outils précédents. Les contributions sont en « *pour 10 000* ».

```
> library(ade4)
> foods.coa <- dudi.coa(foods,scannf=F,nf=2)
>
> #eigen values and cumulative proportion of variance explained in percentage
> print(round(cbind(foods.coa$eig,100.0*cumsum(foods.coa$eig)/sum(foods.coa$eig)),4))
       [,1]     [,2]
[1,] 0.1931  52.5012
[2,] 0.0777  73.6357
[3,] 0.0439  85.5593
[4,] 0.0328  94.4786
[5,] 0.0123  97.8111
[6,] 0.0057  99.3575
[7,] 0.0024 100.0000
>
> #row analysis - coordinates and contributions
> print(cbind(foods.coa$li,inertia.dudi(foods.coa,row.inertia=T)$row.abs))
                      Axis1       Axis2 Axis1 Axis2
Healthy          -0.086620542 -0.34591187    43  1714
Nutritious        0.008607025 -0.26886266     0   993
GoodSummer       -0.348785262 -0.13275108   655   236
GoodWinter        0.277410291  0.12712635   276   144
Expensive         0.200671515 -0.36124906   132  1059
QuickEasy        -0.643652755  0.08475682  2024    87
Tasty             0.039631673 -0.11273214     9   177
Economical       -0.416925032  0.65686751   425  2618
ForATreat         0.619461485 -0.06491653   937    26
ForWeekdays      -0.506754024  0.33798769  1020  1127
ForWeekends       0.560048658  0.17123472  1366   317
Tasteless        -0.450957714  0.15778118   126    38
TooLongToPrepare  1.277778139  0.33400636  2642   449
FamilyFavourite   0.387842351  0.42183368   345  1015
>
> #column analysis - coordinates and contributions
> print(cbind(foods.coa$co,inertia.dudi(foods.coa,col.inertia=T)$col.abs))
                 Comp1       Comp2 Comp1 Comp2
Cereals     -0.56271870  0.3579056  2553  2566
Muesli      -0.31310163 -0.3186912   663  1708
Porridge     0.05362789  0.2130951    16   627
BaconEggs    0.78344380  0.1481415  5635   500
ToastTea    -0.17534202  0.4470236   101  1636
FreshFruit  -0.24619292 -0.2449293   389   956
StewedFruit  0.31539949 -0.2428134   594   875
Yoghurt     -0.08598906 -0.2641586    48  1132
```

Figure 19 - Résultats du package 'ade-4'

4.3.2 Carte des modalités – Représentation simultanée

« Ade4 » propose différents types de graphiques. Pour la représentation quasi-barycentrique, nous sélectionnons l'option « method = 1 » dans la commande **scatter.coa()**.

```
#plotting rows and columns
scatter.coa(foods.coa,method=1)
```

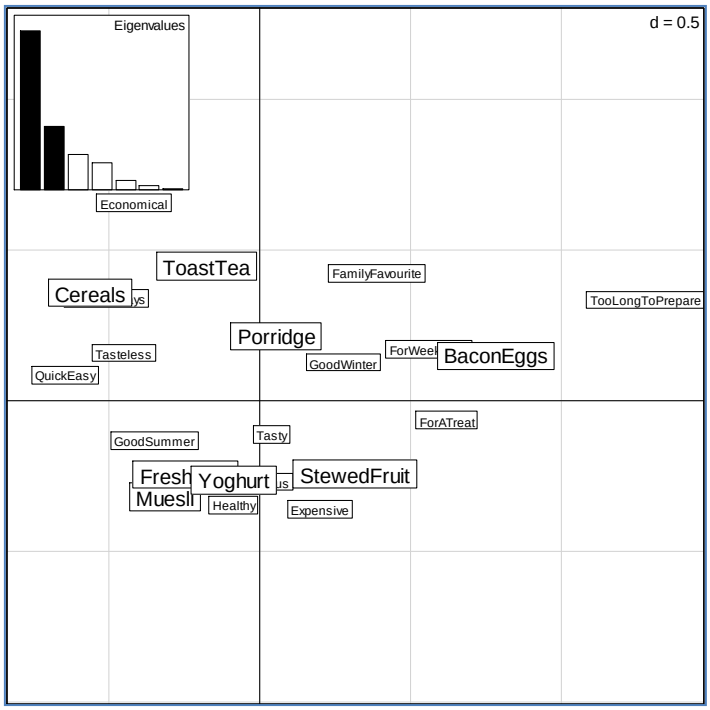


Figure 20 - Représentation simultanée - Package 'ade4'

4.3.3 Visualisation rapide du rôle des modalités

Tanagra trie les modalités pour que l'on identifie rapidement leur impact sur chaque facteur. Ade4 propose une fonctionnalité similaire. Mais, à la différence de Tanagra, l'ordonnancement est restreint à chaque axe et s'appuie sur les coordonnées.

```
#canonical graph
score.coa(foods.coa,xax=1,dotchart=T)
```

Nous obtenons une variante de la représentation simultanée.

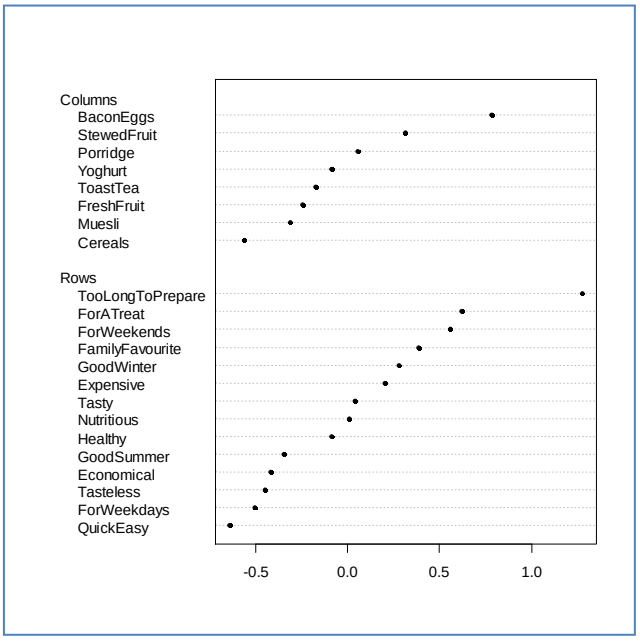


Figure 21 – Variante de la représentation simultanée - Package 'ade4'

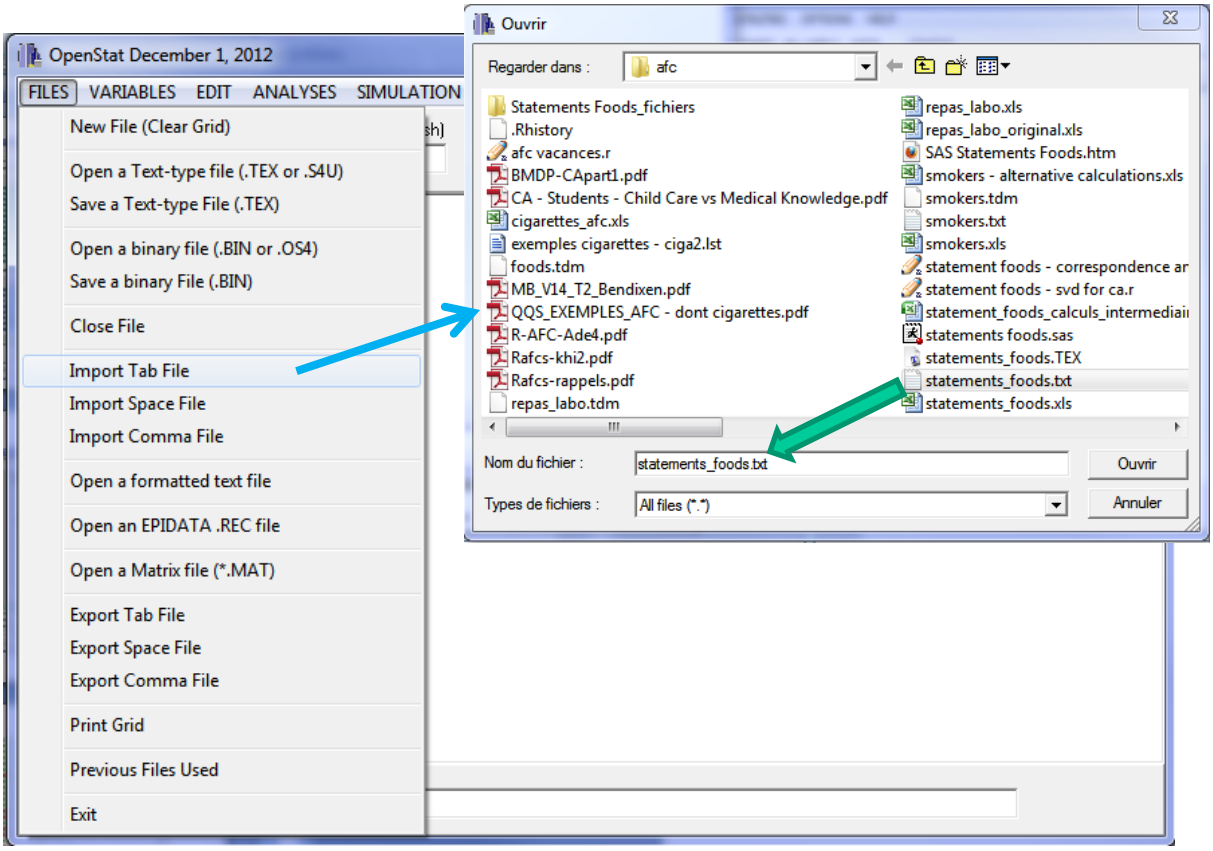
5 Analyse des correspondances avec OpenStat

OpenStat est un logiciel de statistique développé par William Miller, présent sur le web depuis de nombreuses années. Au–delà de l'outil, son auteur a mis en ligne une documentation considérable, rédigée ou sous forme de vidéos[17], accompagnée des jeux de données servant à illustrer les techniques. Il y a réellement matière à travailler pour les étudiants.

OpenStat est piloté par menu. Son utilisation est à la fois simple et assez particulière. Les résultats s'enchaînent dans des fenêtres qui s'affichent successivement. Elles sont relativement nombreuses. Nous n'en afficherons que certaines - celles qui sont en relation directe avec les sorties des autres outils - dans ce descriptif.

5.1 Importation des données

Après avoir démarré OpenStat (*version du 01 décembre 2012*), nous actionnons le menu FILES / IMPORT TAB FILE pour accéder au fichier « statements_foods.txt ».

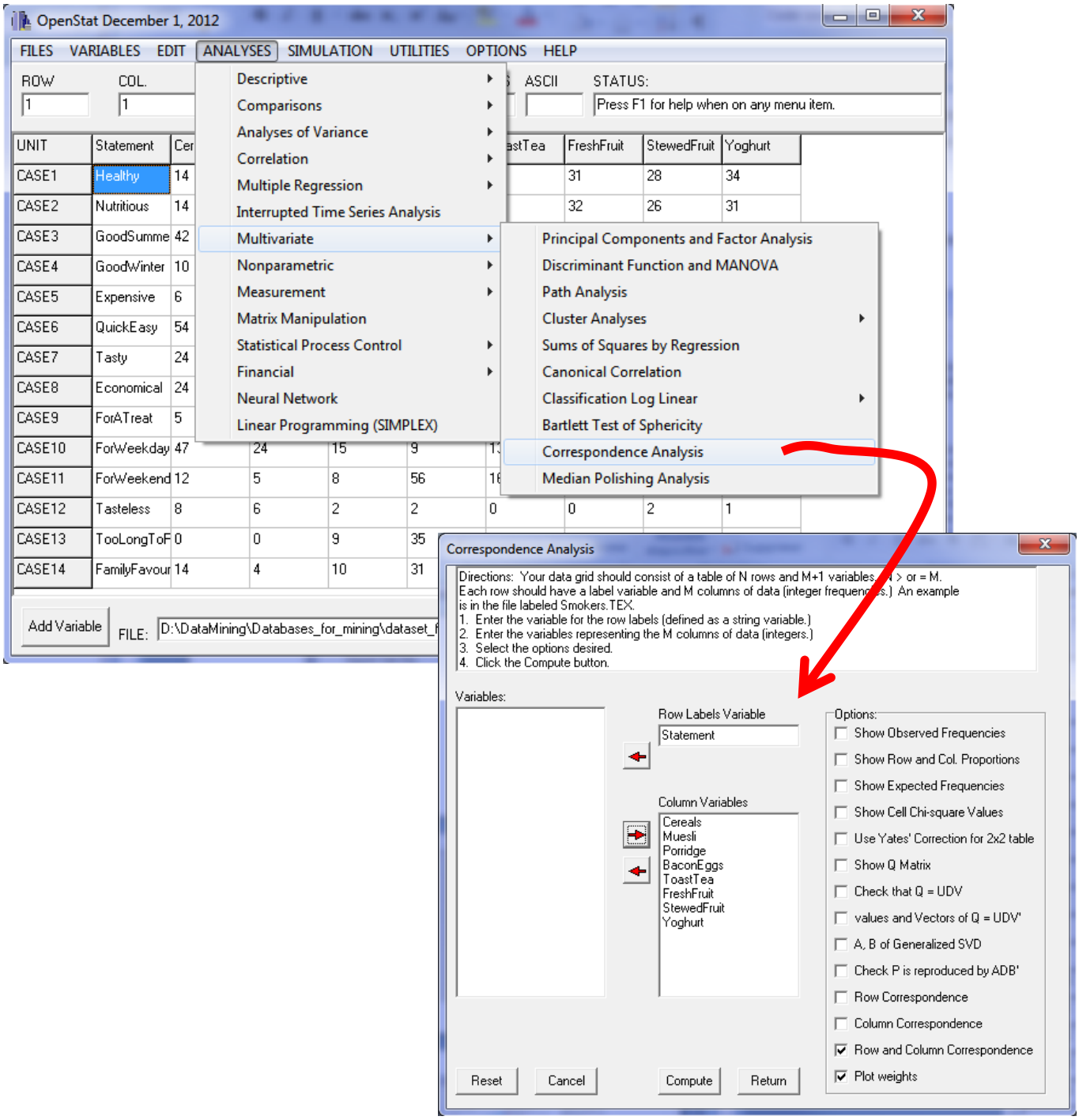


Les données sont chargées dans la grille de la fenêtre principale de l'application.

5.2 Paramétrage de la méthode

Nous actionnons le menu ANALYSES / MULTIVARIATE / CORRESPONDENCE ANALYSIS pour lancer l'analyse des correspondances. La colonne STATEMENT correspond à l'étiquette des lignes, les autres variables (CEREALS … YOGHOURT) aux colonnes du tableau de contingence.

[17] http://www.statprograms4u.com/ ; certains de ces documents seront publiés chez Springer incessamment (début 2013) : « OpenStat Reference Manual » et « Statistics and Measurement Concept with OpenStat ».



5.3 Lecture des résultats

Nous cliquons sur COMPUTE pour lancer les traitements. Plusieurs fenêtres vont alors s'enchaîner, décrivant tour à tour les différents aspects des résultats de l'analyse. Pour passer d'une fenêtre à l'autre, il faut cliquer sur le bouton RETURN en haut à droite.

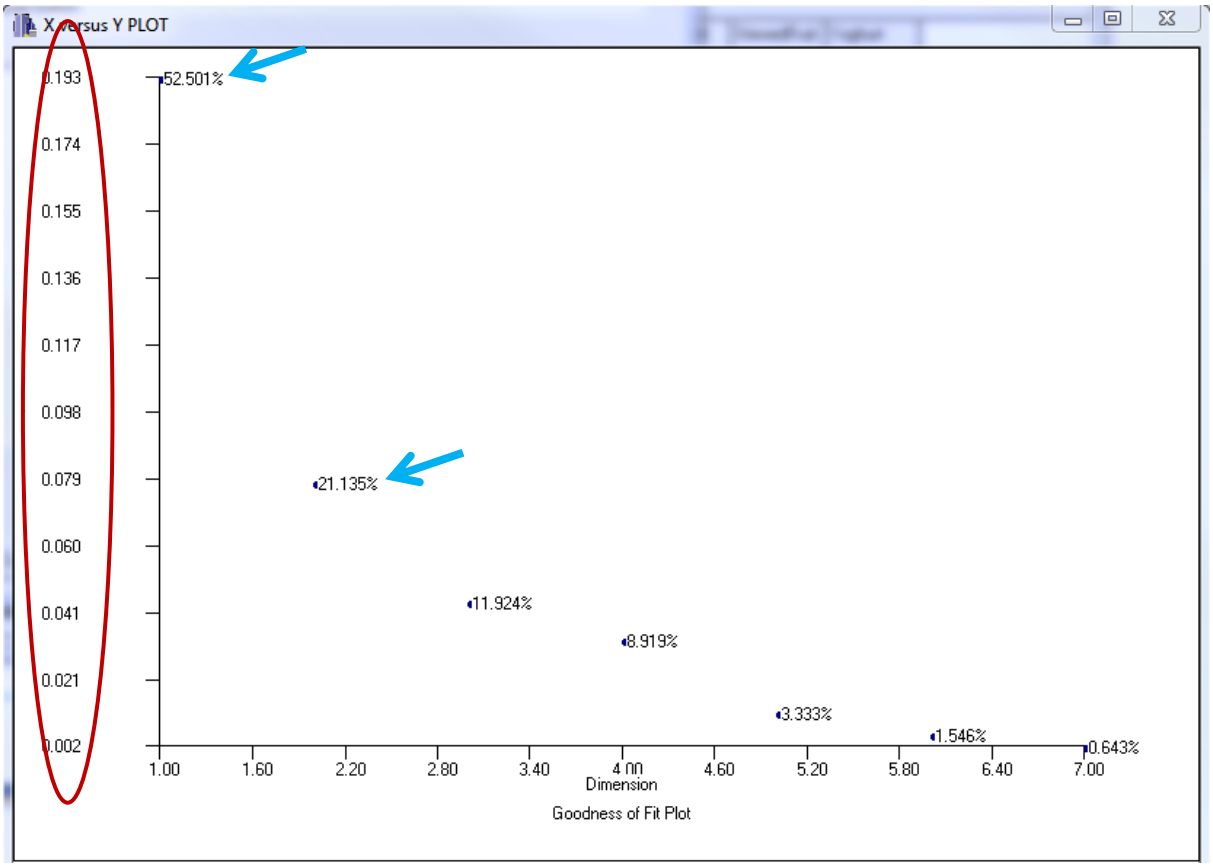
5.3.1 Test du χ^2

OpenStat met en œuvre plusieurs formulations du test d'indépendance. Nous retrouvons la statistique du $\chi^2 = 647.312$; le $\phi^2 = (0.6065)^2 = 0.3678$ correspond à l'inertie totale.

```
Chi-square =  647.312 with D.F. = 91. Prob. > value =    0.000

Liklihood Ratio =  652.128 with prob. > value = 0.0000

phi correlation = 0.6065

Pearson Correlation r = -0.0888

Mantel-Haenszel Test of Linear Association =   13.879 with probability > value = 0.0002

The coefficient of contingency =    0.519

Cramer's V =    0.229
```

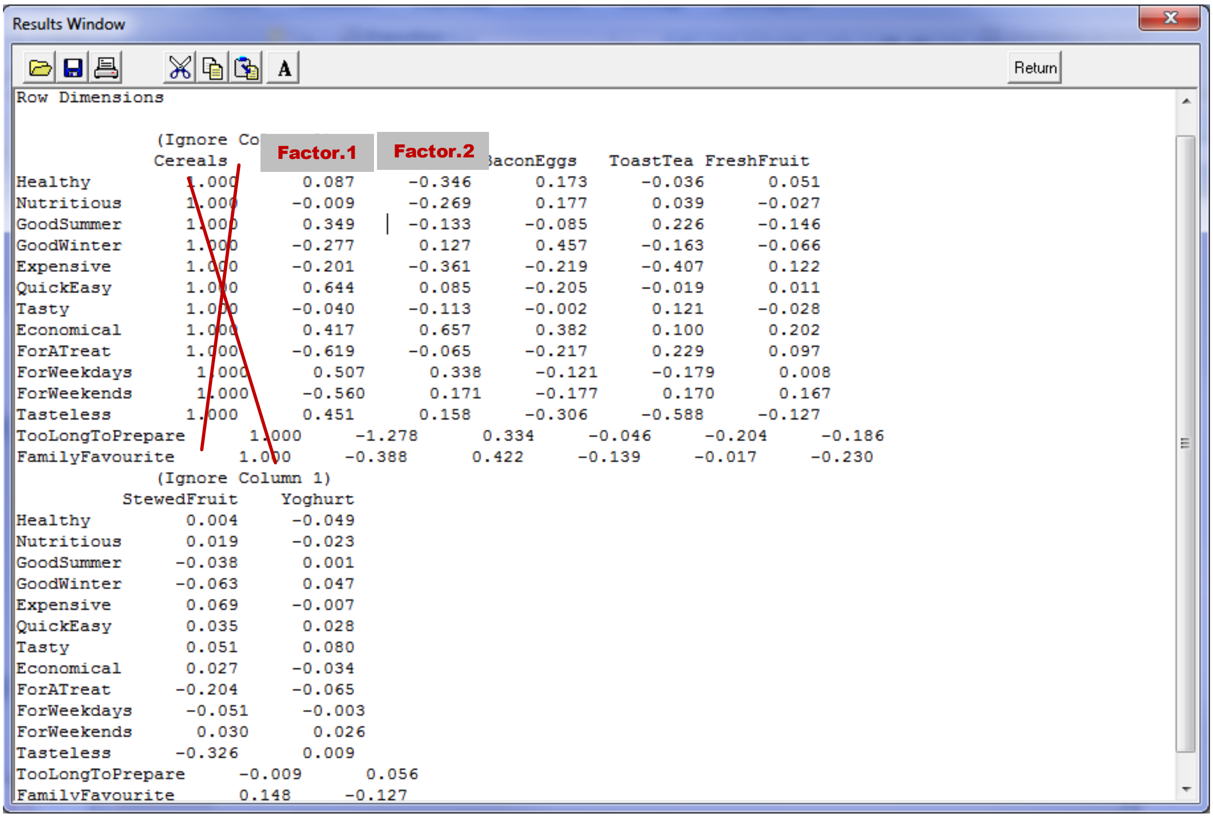
5.3.2 Décroissance du pourcentage d'inertie expliquée

OpenStat propose l'éboulis des valeurs propres (« Goodness of fit plot »). Les points sont étiquetés avec la proportion d'inertie expliquée.



5.3.3 Représentation des lignes

OpenStat inclut la solution triviale dans la représentation des lignes, en précisant cependant que nous devons l'ignorer. Il affiche les coordonnées des modalités lignes pour tous les facteurs. Nous retrouvons les résultats des autres logiciels. *En revanche, contrairement à ce qui est indiqué par le logiciel, les en-têtes de chaque colonne devraient être associés aux axes factoriels et non aux modalités colonnes*.

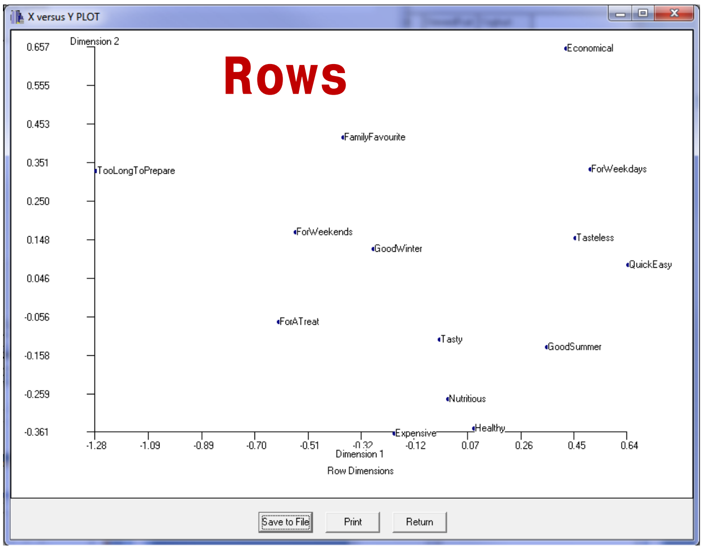


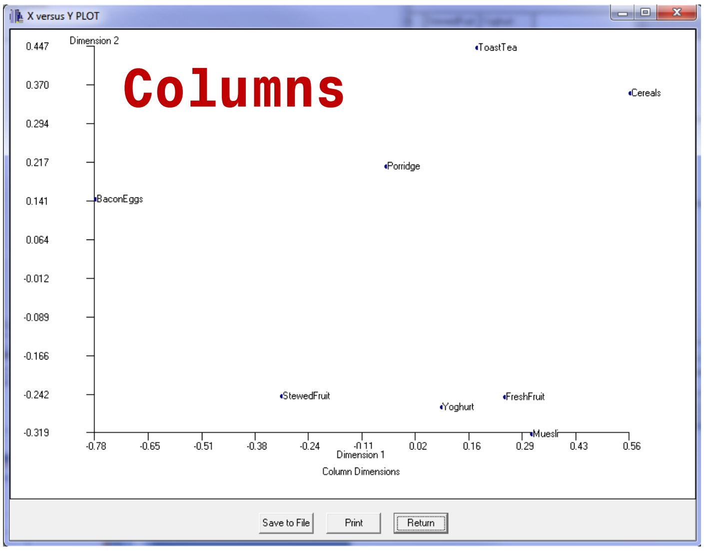
```
Row Dimensions

              (Ignore Co  Factor.1    Factor.2
              Cereals                           BaconEggs   ToastTea FreshFruit
Healthy          1.000      0.087     -0.346     0.173     -0.036      0.051
Nutritious       1.000     -0.009     -0.269     0.177      0.039     -0.027
GoodSummer       1.000      0.349     -0.133    -0.085      0.226     -0.146
GoodWinter       1.000     -0.277      0.127     0.457     -0.163     -0.066
Expensive        1.000     -0.201     -0.361    -0.219     -0.407      0.122
QuickEasy        1.000      0.644      0.085    -0.205     -0.019      0.011
Tasty            1.000     -0.040     -0.113    -0.002      0.121     -0.028
Economical       1.000      0.417      0.657     0.382      0.100      0.202
ForATreat        1.000     -0.619     -0.065    -0.217      0.229      0.097
ForWeekdays       1.000      0.507      0.338    -0.121     -0.179      0.008
ForWeekends       1.000     -0.560      0.171    -0.177      0.170      0.167
Tasteless        1.000      0.451      0.158    -0.306     -0.588     -0.127
TooLongToPrepare      1.000     -1.278      0.334    -0.046     -0.204     -0.186
FamilyFavourite      1.000     -0.388      0.422     -0.139     -0.017     -0.230
              (Ignore Column 1)
           StewedFruit    Yoghurt
Healthy          0.004     -0.049
Nutritious       0.019     -0.023
GoodSummer      -0.038      0.001
GoodWinter      -0.063      0.047
Expensive        0.069     -0.007
QuickEasy        0.035      0.028
Tasty            0.051      0.080
Economical       0.027     -0.034
ForATreat       -0.204     -0.065
ForWeekdays      -0.051     -0.003
ForWeekends       0.030      0.026
Tasteless       -0.326      0.009
TooLongToPrepare      -0.009      0.056
FamilyFavourite       0.148     -0.127
```

Nous avons le même type d'affichage pour les modalités colonnes.

5.3.4 Représentations graphiques

OpenStat préfère deux graphiques séparés pour positionner les modalités lignes et colonnes. Mais nous pouvons très facilement les mettre en parallèle puisque nous disposons des coordonnées adéquates pour une représentation simultanée.





6 Conclusion

L'analyse des correspondances est une technique très populaire. Elle permet de débroussailler rapidement les grands tableaux de contingence en mettant en avant les relations les plus saillantes.

In fine, nous le constatons encore une fois dans ce didacticiel, les logiciels fournissent les mêmes résultats numériques. Je ne le répéterai jamais assez, il n'y a pas de bons ou de mauvais logiciels. En revanche, de par le prisme adopté, ils éclairent avec plus ou moins d'acuité différentes facettes de l'analyse. Il nous appartient par la suite d'en apprécier pleinement la teneur. Cela n'est possible que si nous comprenons parfaitement les tenants et aboutissants de la technique sous-jacente.

Enfin, un dernier mot sur R. Avec le système des packages, il est possible de disposer de différents outils pour aborder un seul et même problème. Je pense que c'est un atout même si cela peut parfois ajouter à la confusion (après la vacuité du débat « qui est le meilleur logiciel », on ne va pas se lancer dans la discussion, tout au aussi stérile, « qui est le meilleur package » ?). Tous ces outils sont accessibles librement. Certains, je pense en particulier à ade4, proposent une documentation très abondante et de très grande qualité. C'est à ce niveau que se fait la différence. Leur site est une vraie mine d'or pour nos étudiants.

7 Références

L. Lebart, M. Piron, A. Morineau, « Statistique Exploratoire Multidimensionnelle », Dunod, 2006.
G. Saporta, « Probabilités, Analyse des Données et Statistique », Dunod, 2006.
M. Tenenhaus, « Statistique : Méthodes pour décrire, expliquer et prévoir », Dunod, 2006.