Subject

Using correspondence analysis with TANAGRA.

Correspondence analysis is a visualization technique. It enables to see the association between rows and columns in a large contingency table. It belongs to "factorial analysis" approach. The method computes some axes, which are latent variables that we interpret in order to understand the proximities between rows and/or columns.

Here a short description of the approach: http://www.statsoft.com/textbook/stathome.html, see Correspondence Analysis section.

TANAGRA is not really intended for contingency table. So we use an artifice. The rows are specified from a discrete attribute, and the columns correspond to several continuous attributes in our dataset. We cannot treat a contingency table with more than 255 rows.

This tutorial is suggested by the presentation of Lebart, Morineau and Piron, in their book, « Statistique Exploratoire Multidimensionnelle », Dunod, 2000. Unfortunately, I don't think there is an English translation of this very good teaching book, which is really popular in France. However, I hope this tutorial is understandable without the book.

If you read French, the description of the correspondence analysis is available at section 1.3 (p. 67 to 107).

Dataset

We use MEDIA_PROF_AFC.XLS from the book (Tableau 1.3-10, page 104). We observe the association between the occupation and the kind of media used.

Here is the dataset:

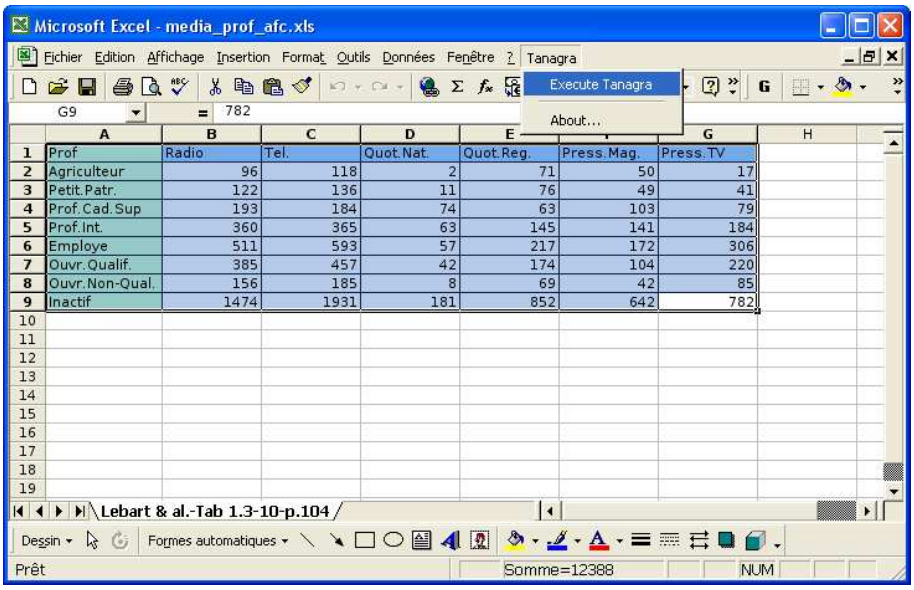
Prof	Radio	Tel.	Quot.Nat.	Quot.Reg.	Press.Mag.	Press.TV
Agriculteur	96	118	2	71	50	17
Petit.Patr.	122	136	11	76	49	41
Prof.Cad.Sup	193	184	74	63	103	79
Prof.Int.	360	365	63	145	141	184
Employe	511	593	57	217	172	306
Ouvr.Qualif.	385	457	42	174	104	220
Ouvr.Non-Qual.	156	185	8	69	42	85
Inactif	1474	1931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

In the first column of the dataset, we have the row identifier i.e. the occupation. In the first row, in blue, we have the column identifier i.e. the media. There are 9 rows and 6 columns in this contingency table.

Correspondence Analysis with TANAGRA

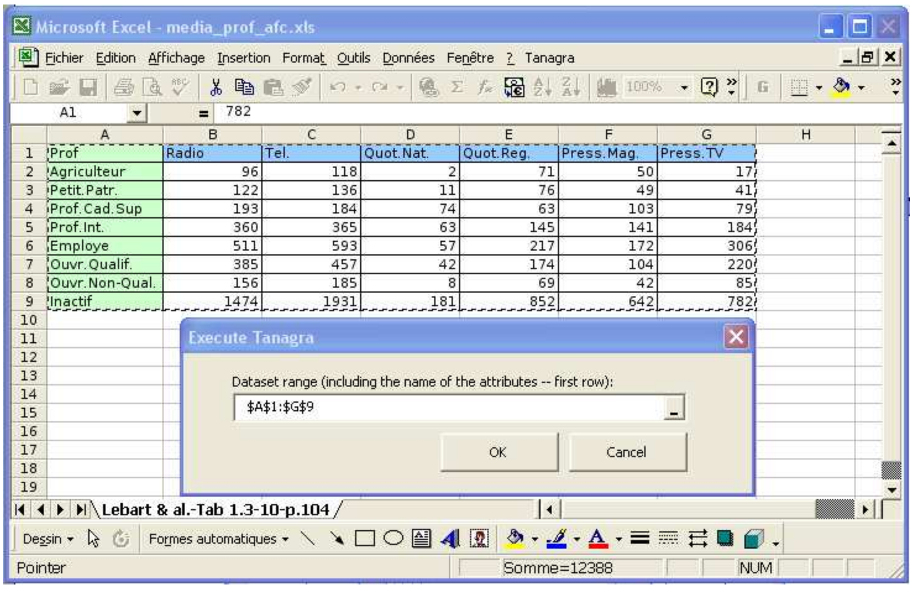
Diagram

We can start TANAGRA (version 1.4.11 and later) from EXCEL with the TANAGRA.XLA add-in. We use this option here. The diagram is automatically built. To do this, we select the cells range and click on the new menu TANAGRA / EXCECUTE TANAGRA in EXCEL.


Microsoft Excel - media_prof_afc.xls
Fichier Edition Affichage Insertion Format Outils Données Fenêtre ? Tanagra
Execute Tanagra
About...
G9
782
Lebart & al.-Tab 1.3-10-p.104
Dessin
Formes automatiques
Prêt
Somme=12388
NUM

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

A dialog box appears. We confirm the selection (OK button).

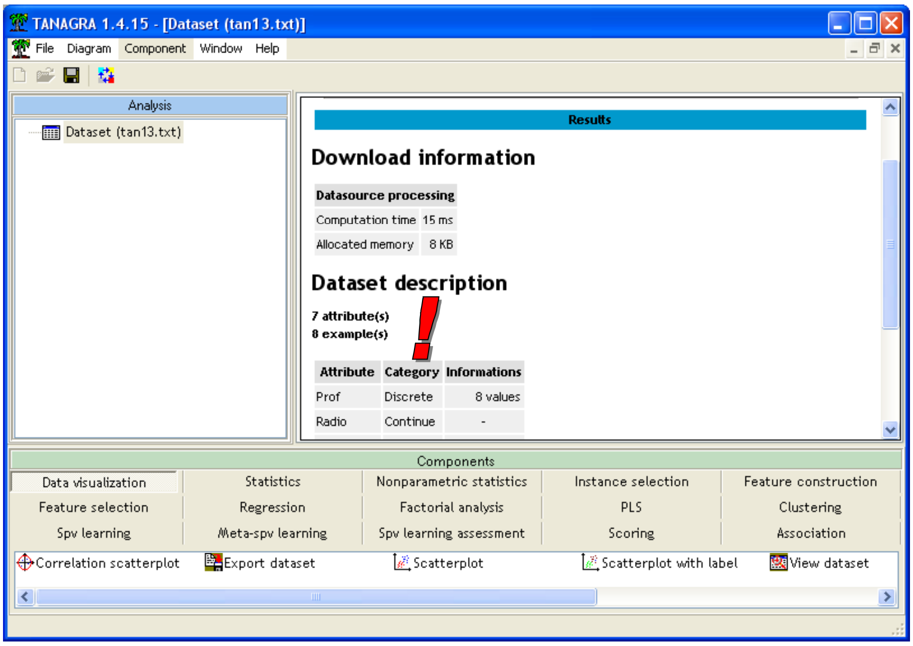

Microsoft Excel - media_prof_afc.xls
Fichier Edition Affichage Insertion Format Outils Données Fenêtre ? Tanagra
A1
782
Lebart & al.-Tab 1.3-10-p.104
Dessin
Formes automatiques
Pointer
Somme=12388
NUM
OK
Cancel

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

Execute Tanagra

Dataset range (including the name of the attributes -- first row): $A$1:$G$9

TANAGRA is then started. We check that there are 8 observations and 7 variables in the dataset. It corresponds to a contingency table with 8 rows and 6 columns. The first discrete variable is in fact the row identifier.



Results

Download information

Datasource processing	
Computation time	15 ms
Allocated memory	8 KB

Dataset description

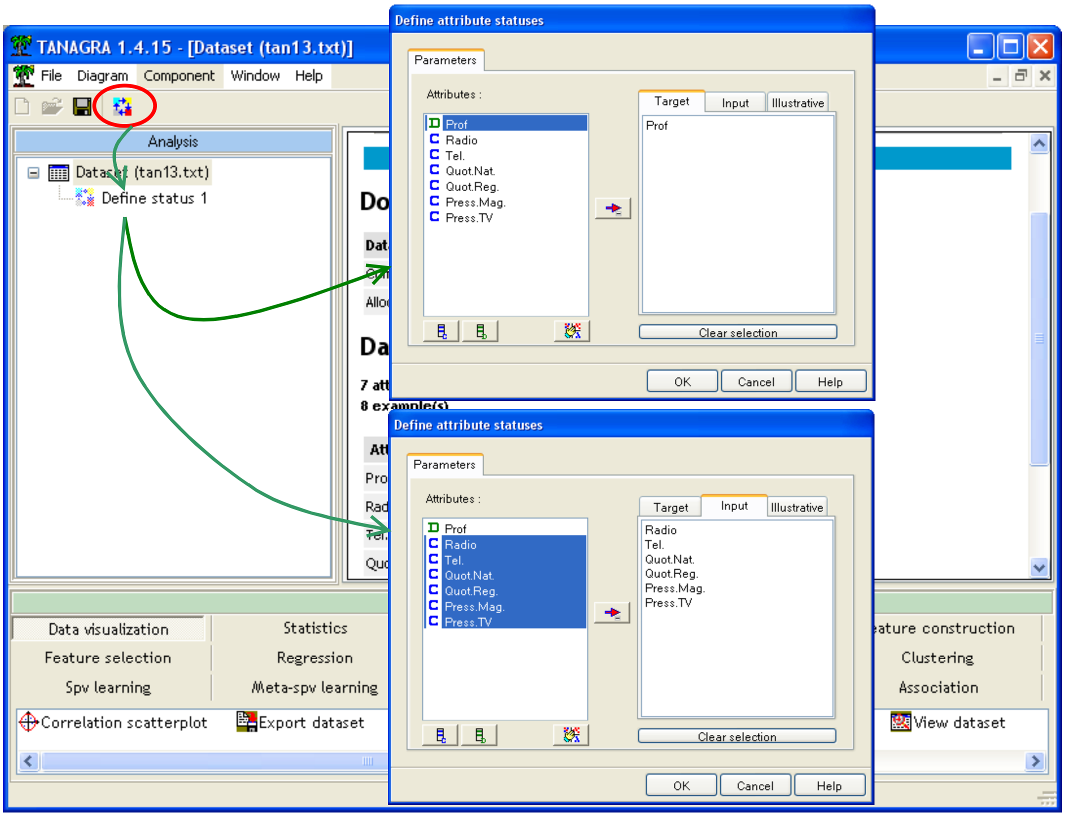
7 attribute(s)
8 example(s)

Attribute	Category	Informations
Prof	Discrete	8 values
Radio	Continue	-

Correspondence Analysis

In order to specify the analysis, we must to appropriately select the row and the column of the contingency table. We set PROF (discrete) as TARGET attribute; all the continuous variables are INPUT.

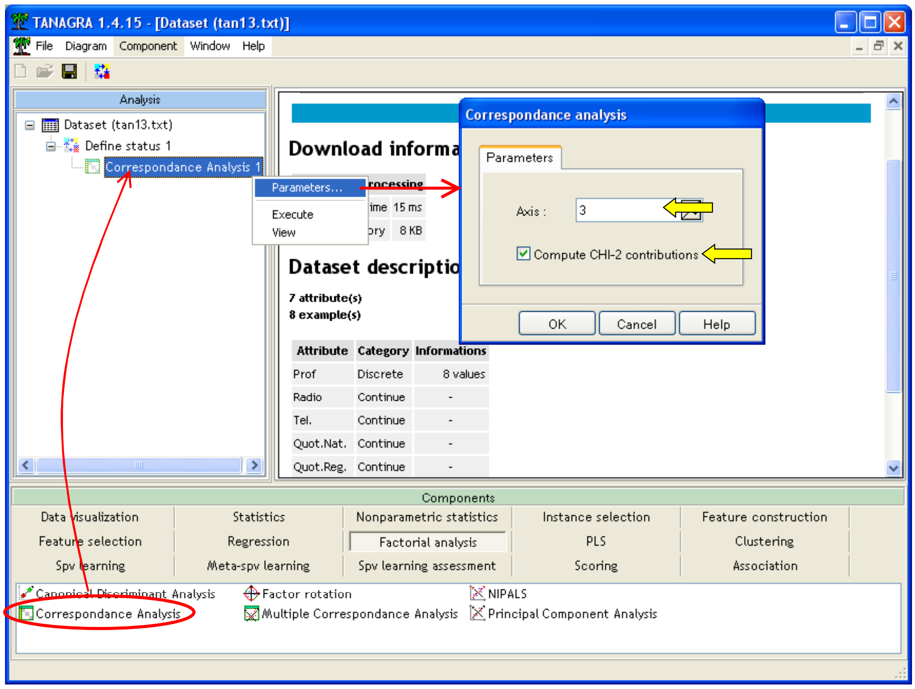
This selection has not really a signification in the analysis; it is not a prediction or a supervised learning scheme. It simply enables to define the rows and the columns of the cross-table.



Target: Prof

Input: Radio, Tel., Quot.Nat., Quot.Reg., Press.Mag., Press.TV

We add the CA (Correspondence Analysis – FACTORIAL ANALYSIS tab) component into the diagram. We click on the PARAMETERS menu. We set the number of axes (**3**), and we select the CHI-2 contributions computation. We see below the signification of this option.



Correspondance analysis

Parameters

Axis : 3

☑ Compute CHI-2 contributions

Dataset descriptio

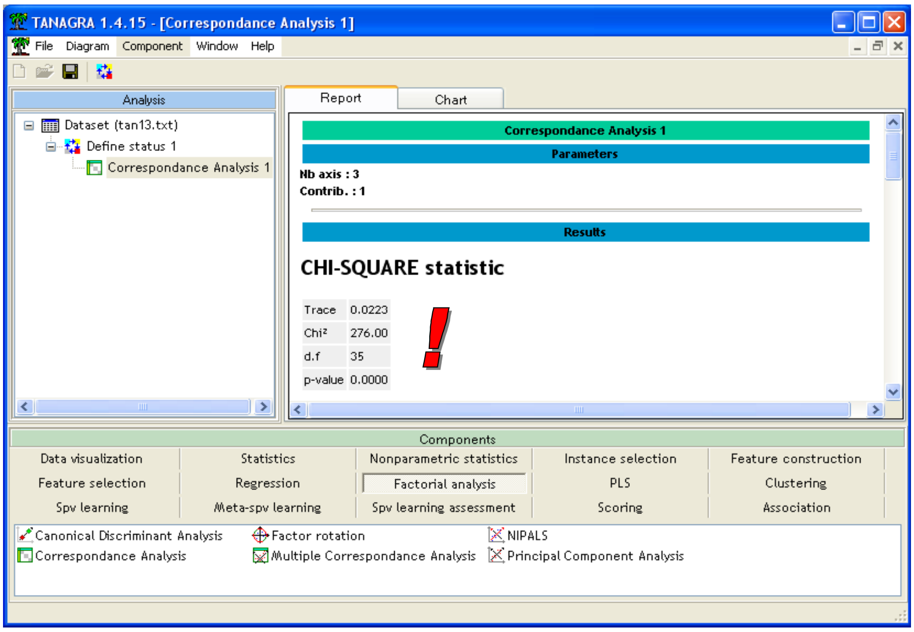
7 attribute(s)
8 example(s)

Attribute	Category	Informations
Prof	Discrete	8 values
Radio	Continue	-
Tel.	Continue	-
Quot.Nat.	Continue	-
Quot.Reg.	Continue	-

We click on the VIEW menu in order to obtain the results.

Results of Correspondence Analysis

CHI-2. The first result is the CHI-2 statistic. It enables to see if there is a global interaction between the rows and the columns of the cross-table (page 104).



Correspondance Analysis 1

Parameters

Nb axis : 3
Contrib. : 1

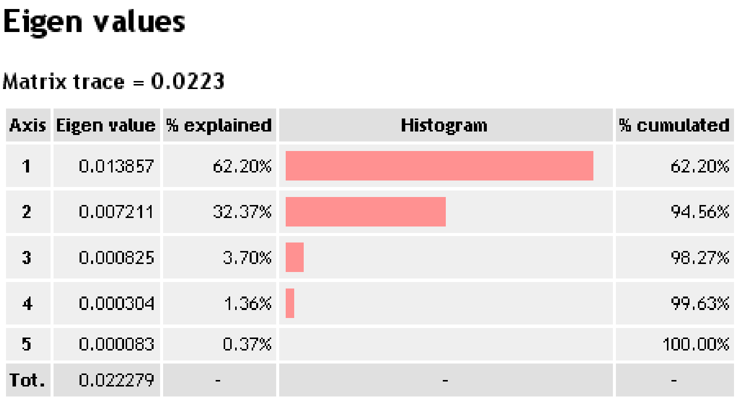
Results

CHI-SQUARE statistic

Trace	0.0223
Chi²	276.00
d.f	35
p-value	0.0000

The p-value, very small, seems to show that the rows and the columns of the cross-table are not independent of each other.

Eigen values. The next result is the Eigen values table. We observe the reported information on each axis. In this dataset, we observe that 94.56% of the whole information contained in the table is reported by the two first factors (Tableau 1.3-11, page 104).

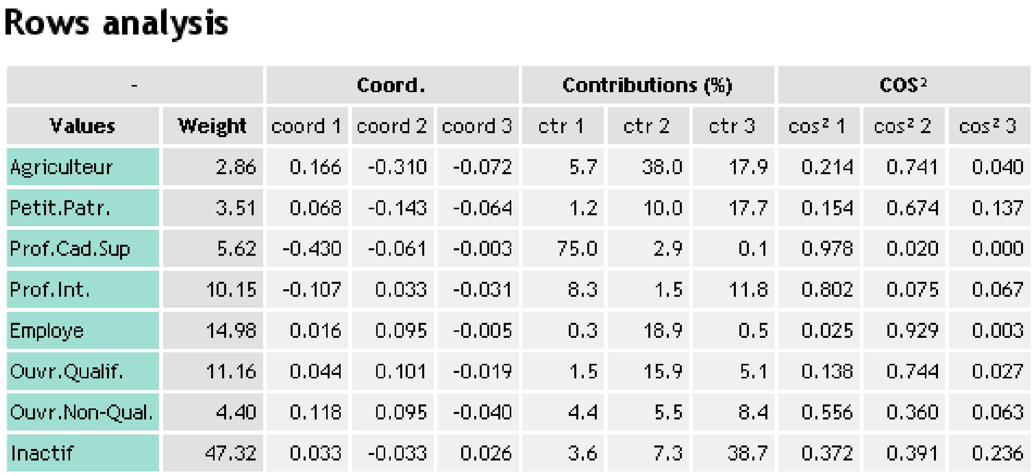


Eigen values

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	-	-	-

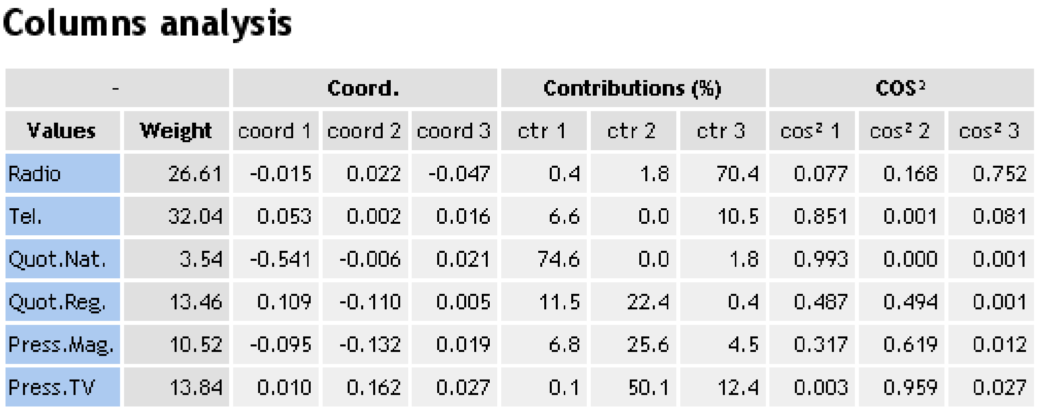
Coordinates, contributions and COS². The next results are the coordinates of the points (rows or columns) related to the factors. The contributions enable to interpret the factors i.e. which are the points that influence heavily the factor? The cosines enable to judge the quality of the representation of a point related to a factor (Tableau 1.3 – 10, page 104). We obtain the following table for the rows.



Rows analysis

-		Coord.			Contributions (%)			COS²		
Values	Weight	coord 1	coord 2	coord 3	ctr 1	ctr 2	ctr 3	cos² 1	cos² 2	cos² 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

And the following one for the columns.



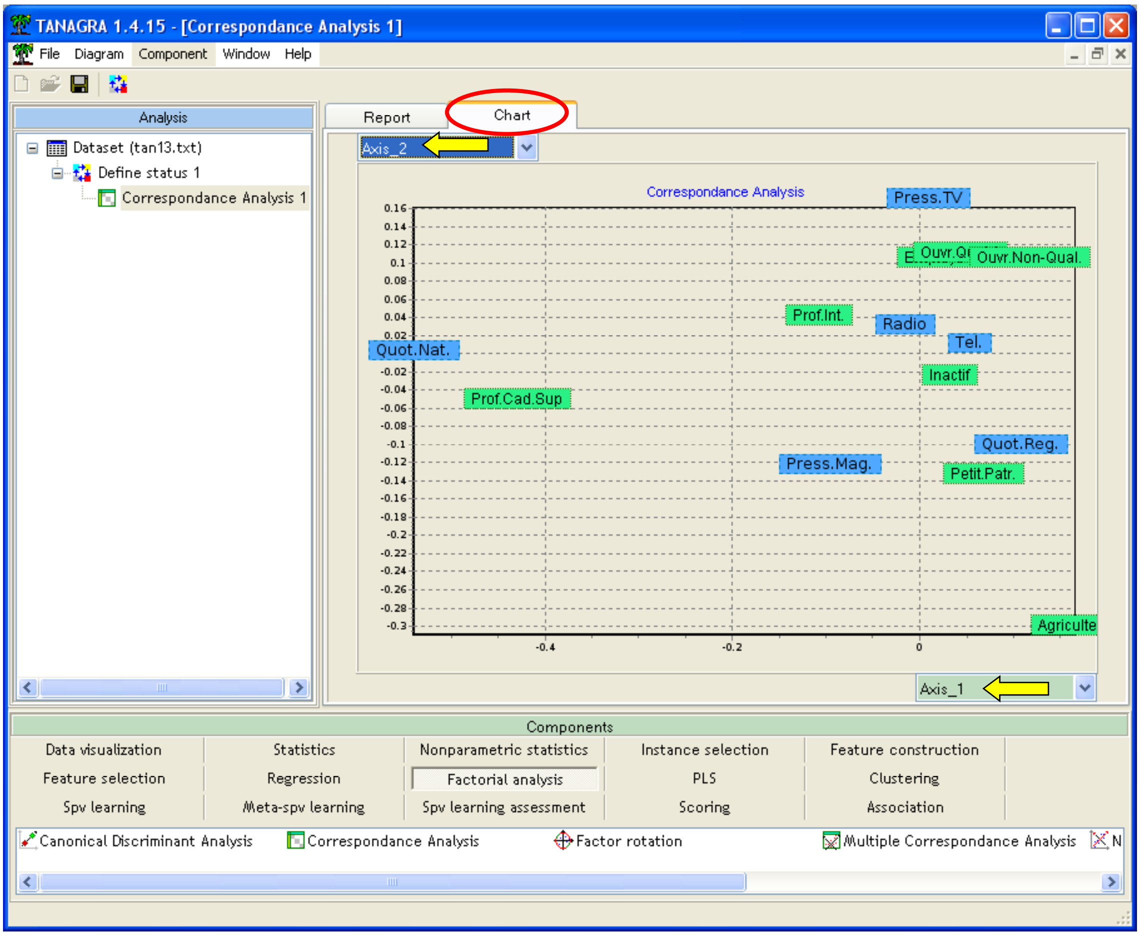
Columns analysis

-		Coord.			Contributions (%)			COS²		
Values	Weight	coord 1	coord 2	coord 3	ctr 1	ctr 2	ctr 3	cos² 1	cos² 2	cos² 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

Graphical representation

The popularity of the factorial analysis relies on the possibility to plot the points in a chart. We can then observe the proximity between the points. In the correspondence analysis, on the one part, we can plot the rows; and the other part, we can plot the columns. The originality is that we can also produce a result where the coordinates of the rows and the columns are compatible. Then we plot the rows and the columns in the same chart, we see if there is particular association between some rows and columns.

We click on the CHART tab. We can select the factors.

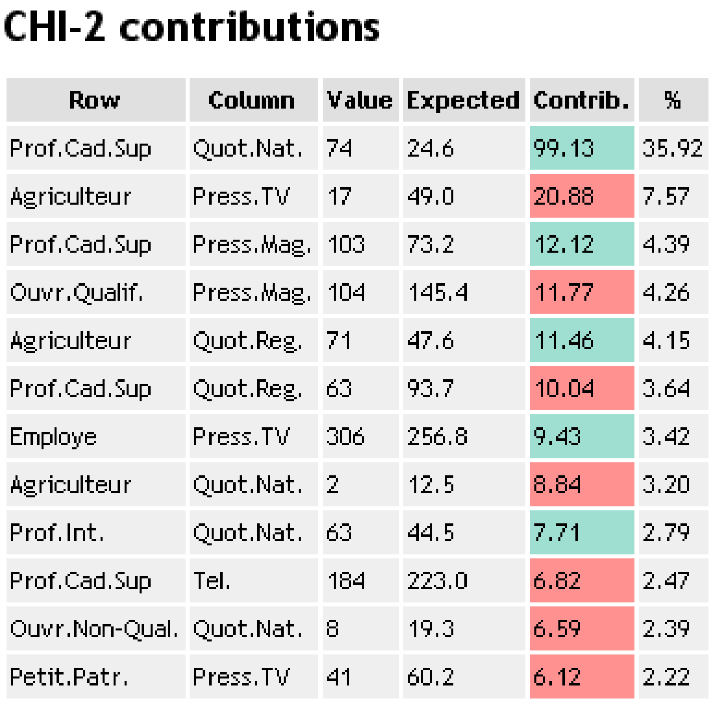


In this dataset, we see on the first two factors that the Executive Occupation and the Managers read mainly the national newspapers (Figure 1.3 – 23, page 106).

Let us note that it is possible to copy the chart. It is also possible to modify the font size of the labels.

Contributions to CHI-2

There is another way to extract the main information in a cross-table. We compute the part of CHI-2 associated to each combination of row and column. We can observe which combination is really different to the independence situation between rows and columns and heavily contributes to the computation of the CHI-2 statistic. Then, we sort the results according to the contribution.

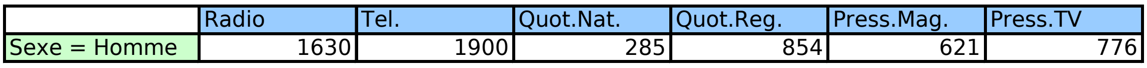


CHI-2 contributions

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

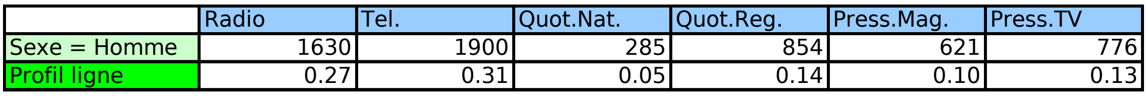
Supplementary points

We can compute the coordinates of a new point that is not included in the previous analysis. For instance, we want to evaluate the men's behavior according to the media used. We have the following values:



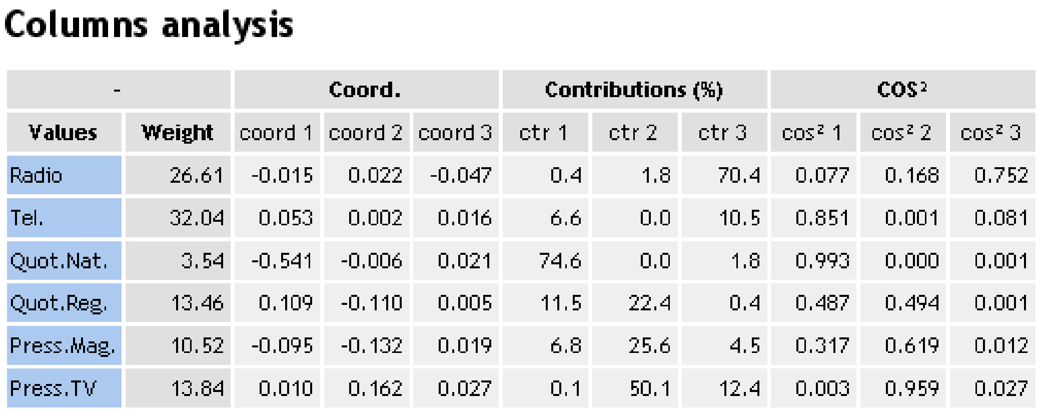
	Radio	Tel.	Quot.Nat.	Quot.Reg.	Press.Mag.	Press.TV
Sexe = Homme	1630	1900	285	854	621	776

We compute their relative frequencies:



	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

We use the columns coordinates in order to compute the coordinate of this new point.



Columns analysis

-		Coord.			Contributions (%)			COS²		
Values	Weight	coord 1	coord 2	coord 3	ctr 1	ctr 2	ctr 3	cos² 1	cos² 2	cos² 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

The formula is the following:

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

0.0139 is the first Eigen value. With this principle, we obtain the following coordinate on the first two factors: (-0.05; -0.02). We note that the men have not a particular behavior according to the media used.

Let us analyze now people that have high educational level (Tableau 1.3 – 10, page 104).

	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

Now, their coordinate on the first two factors is (-0.29; -0.02). There are close to people with high occupation level.

Conclusion

TANAGRA proposes simple outputs of the factorial analysis.

One of the main advantages of the software is that we can copy the results in a spreadsheet and perform various additional analyses.