Contingency table with a double partition on rows and columns. Visualization and comparison of the partial and global structures

Mónica Bécue[1], Jerôme Pagès[2], and Campo-Elías Pardo[3]

[1] EIO. Universitat Politècnica de Catalunya
08028 Barcelona - Spain
(e-mail: monica.becue@upc.edu)

[2] Agrocampus Rennes
65 rue de Saint-Brieuc, CS 84215
F-35042 Rennes cedex, France
(e-mail: jerome.pages@agrocampus-rennes.fr)

[3] Departamento de Estadística. Universidad Nacional de Colombia
Bogotá, Colombia
(e-mail: cepardot@unal.edu.co)

Abstract. Internal correspondence analysis (ICA) deals with frequency tables having a double partition structure on the columns and rows, offering their representation on principal axes which reflects the inner structure of the subtables as defined by the two partitions. We enrich this global representation by the superimposed representation of the rows (respectively, the columns) as described separately by every group of columns (respectively, by every group of rows). The new aids to interpretation that we propose, give information about the common and specific structures in the subtables.

Keywords: Correspondence analysis, Internal correspondence analysis, Multiple factor analysis, Common dispersion directions, Multicanonical analysis.

1 Introduction

Some applications lead to build up contingency tables having a double partition on the columns and rows. This characteristic induces objectives such as giving an account of the global structure of the table that takes into account its specificity as well as comparing all the partial row structures (respectively, partial column structures) induced on the rows by each group of columns (respectively, each group of rows). Concerning the first objective, internal correspondence analysis (ICA) [Cazes *et al.*, 1988] offers an interesting approach. However, ICA does not provide any result relative to the partial structures. In this work, in the framework of ICA, we propose several tools in order to compare them, favoring their simultaneous visualization.

§2 introduces the notation. After reminding the basic principles of ICA (§3), we propose a methodology to compare the global and partial points of wiew in §4. The §5 shows the interest of these tools as applied to an example. Finally, we conclude with some remarks.

2 Notation

We consider a table **F** of proportions (the overall sum amounts to 1) divided into $L \times J$ sub-tables (Fig. 1). The I rows are partitioned in L groups with, respectively, $I_1, I_2, \ldots, I_L$ rows. The K columns are structured in J groups with, respectively, $K_1, K_2, \ldots, K_J$ columns. The subtable (l, j) has I_l rows and K_j columns.

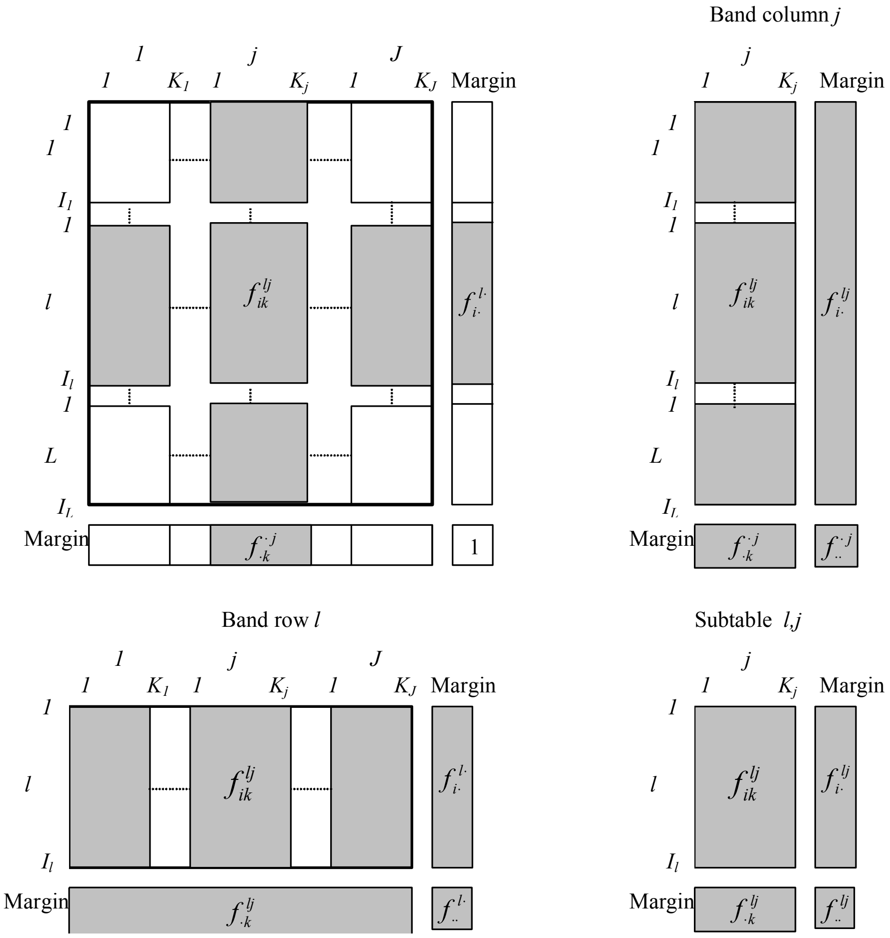

Fig. 1. The global table **F** of proportions, as partitioned into rows and columns groups

3 Internal correspondence analysis (ICA)

3.1 Correspondence analysis with respect to a model

The classical CA refers to the independence model, as given by the products of the margins. Other models can be considered. The CA of **F** with respect to any model **A** having the same margins than **F** [Escofier, 1984], is equivalent to PCA(**X**,**M**,**D**), **X** being the matrix with the general term $x^{lj}_{ik} = \frac{f^{lj}_{ik} - a^{lj}_{ik}}{f^{l\cdot}_{i\cdot} f^{\cdot j}_{\cdot k}}$, using the metric $\mathbf{M} = diag(f^{\cdot j}_{\cdot k})$ and the weights $\mathbf{D} = diag(f^{l\cdot}_{i\cdot})$ in the row space (respectively, the metric $\mathbf{D} = diag(f^{l\cdot}_{i\cdot})$ and the weights $\mathbf{M} = diag(f^{\cdot j}_{\cdot k})$ in the column space).

3.2 Within and double within-tables correspondence analysis

A particular case arises when analyzing the row-wise juxtaposition of J tables (respectively, the column-wise juxtaposition of L tables). The within-tables CA [Benzécri, 1983], [Escofier and Drouet, 1983], [Escofier and Pagès, 1998, p.229], takes as model the within-table independence in order to globally study the deviations of every columns subcloud to their own centroid. For example, if we only take into account the partition on the columns of **F** and consider this table as the row-wise juxtaposition of J tables, the general term of the within-table independence model is $a^{lj}_{ik} = \frac{f^{lj}_{i\cdot} f^{\cdot j}_{\cdot k}}{f^{\cdot j}_{\cdot\cdot}}$. This model has the same margins as table **F**.

In order to take into account both partitions on the rows and columns, [Cazes *et al.*, 1988] propose the internal correspondence analysis (ICA) that considers the model whose general term is given in (1):

$$a^{lj}_{ik} = \frac{f^{lj}_{\cdot k} f^{l\cdot}_{i\cdot}}{f^{l\cdot}_{\cdot\cdot}} + \frac{f^{lj}_{i\cdot} f^{\cdot j}_{\cdot k}}{f^{\cdot j}_{\cdot\cdot}} - \frac{f^{l\cdot}_{i\cdot} f^{\cdot j}_{\cdot k}}{f^{l\cdot}_{\cdot\cdot} f^{\cdot j}_{\cdot\cdot} / f^{lj}_{\cdot\cdot}} \qquad (1)$$

This model has the same margins as table **F**. ICA can be seen as a double within correspondence analysis. The matrix **X** analysed in the PCA(**X**,**M**,**D**) corresponding to ICA inherits the double partition structure of **F**.

4 Comparison of the partial and global structures

ICA offers a representation of the global structure, of the rows and columns, on principal planes in a CA-like way. This global representation can be enriched by looking for representing the rows (resp., the columns) as described separately by every group of columns (resp., every group of rows). For that goal, we adopt a MFA-like point of view [Escofier and Pagès, 1994] by looking for a simultaneous visualization of the partial structures (of rows or of columns) on the principal planes corresponding to the global analysis. We enrich this simultaneous representation with a series of aids to interpretation.

4.1 Superimposed representation of the partial and global rows on a common referential

To each column-band matrix j of $\mathbf{X}$ (as defined in ICA applied to $\mathbf{F}$), we associate the cloud N_I^j of the rows as described by only the columns of this matrix. As this cloud lies in the subspace $\mathbf{R}^{Kj}$ of $\mathbf{R}^K$, we assimilate it to the cloud of the rows of the matrix $\tilde{\mathbf{X}}_\mathbf{j}$, having the same dimension as $\mathbf{X}$ and derived from $\mathbf{X}_\mathbf{j}$ in the following way:

$$\tilde{\mathbf{X}}_j = \begin{array}{|c|c|c|c|}\hline \mathbf{0} & \mathbf{0} & \mathbf{X}_j & \mathbf{0} \\ \hline \end{array}$$

The coordinates of the partial rows, belonging to the cloud N_I^j, on the s-axis issued from the global analysis, are $\tilde{\mathbf{F}}_s^j = \tilde{\mathbf{X}}_j\mathbf{M}\mathbf{u}_s$. To every row (l, i), we associate the $N_{(l,i)}^J$ cloud of its J partial points. In order to obtain a superimposed representation in such a way that the global point corresponds to the centroid of the subcloud $N_{(l,i)}^J$, the coordinates $\tilde{F}_s^j(l, i)$ are amplified by J and then projected on the global representation.

4.2 Aids to interpretation of superimposed representation of the partial and global rows

Quality of representation of the partial clouds: the quality of representation of every cloud N_I^j on the s-axis is measured, in a classical way, through the ratio between the projected inertia and the total inertia.

Measure of the similarity between the partial clouds: the union of the whole of the $N_{(l,i)}^J$ clouds (i.e. the cloud of all the partial row-points noted N_I^J) contains $I \times J$ partial points. These $I \times J$ partial points can be divided into I subclouds, with J points $(l, i)^j$ in every subcloud, corresponding to the same row (l, i). So, the total inertia of N_I^J can be decomposed into within-inertia (inertia within the $N_{(l,i)}^J$ subclouds) and between-inertia (inertia between the $N_{(l,i)}^J$ subclouds). The ratio [between-inertia/total-inertia], calculated axis by axis, measures the proximity of the partial points corresponding to a same row and so, the global similarity between the J partial clouds as projected on this axis. If this ratio is close to 1, the homologous points $(l, i)^j$ are close to one another and the s-axis represents a structure common to the different groups of columns.

Selection of rows and of partial rows with a high contribution to the within-inertia: the within-inertia can be decomposed into the contributions of every row, in order to detect those whose behavior varies from the different points of view represented by the groups of columns. So, the more heterogeneous (respectively, more homogeneous) rows on every axis can be identified in order to interpret the global ratios.

4.3 Superimposed representation of the partial and global columns

The superimposed representation of the partial and global columns clouds and its interpretation aids are obtained in a symmetric way.

5 Application

To illustrate the superimposed representation and their interpretation aids, we utilize the example Ardèche [Cazes *et al.*, 1988], a faunal table crossing species (43 rows) and dates×sites (35 columns, corresponding to 35 dates×sites samplings). This data set is available in [Cazes *et al.*, 1988]. The 43 species are distributed in 4 taxonomic groups (*Ephemeroptera*, *Plecoptera*, *Coleoptera*, *Trichoptera*) which induce the partition on the rows (4 groups). 6 sites (A, B, C, D, E, F) are observed at 6 dates (*jul82*, *aug82*, *nov82*, *feb83*, *apr83*, *jul83*) chosen in different seasons, but the observation of the site F at date 1 is missing. We consider the partition on the columns induced by the different dates (6 groups).

Global representation through ICA

By recentering the subclouds corresponding to a same date, ICA solves the problem of eliminating the time-associated faunal structures and allows for interpreting the spatial typology and for assessing the ability of the taxonomic groups to be used as biological descriptors.

Figure 2 shows the dates×sites on the first principal plane issued from ICA. As [Cazes *et al.*, 1988] note, ICA puts to the fore the originality of the site B, mainly contrasting with A and D. Site D presents a very specific faunal composition in winter (*D-feb83* and *D-apr83*). Mainly F, but also E present outstanding differences between winter (at the left of the first axis: *F-feb83*, *E-feb83* , *F-apr83*, and *E-apr83*) and summer (at the right of the first axis: *F-aug82* and *F-jul83*). Finally, the rise of the water in November standardizes the faunal distribution and, therefore, the subcloud *Nov82*×sites is close to the centroid.

Concerning the species, the inertia on the first axis is mainly due to the great dispersion of the trichopterans: in this group, the species with sheath are attracted by the sites presenting sand or stones with vegetation, while the free trichopterans prefer hard substratum soil (see Fig. 3). *Coleoptera* dispersion strongly contributes to the inertia of the second axis, contrasting the species depending on their preference for strong current or not.

[Cazes *et al.*, 1988] conclude that there is a summer typology, mainly defined by *Coleoptera* and a winter typology, due to *Trichoptera* and corresponding to the originality of site D and the standardization of the fauna in November.

Comparison of the partial structures

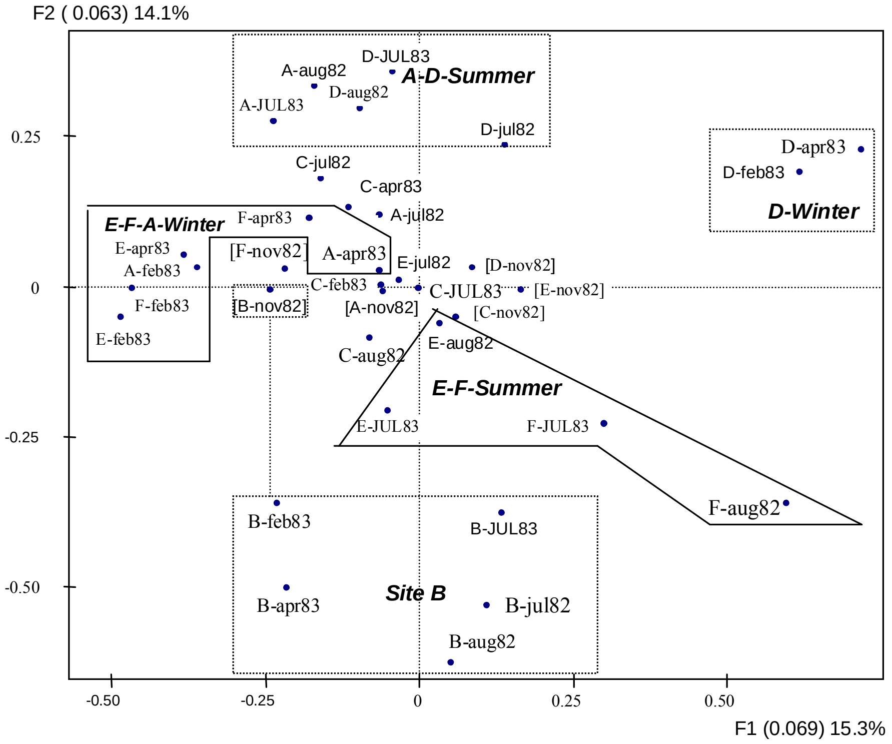

Fig. 2. The column-points on the first principal plane issued from ICA

However, the specific structure of the table, leads to other kinds of questions. For example, *D-feb83* and *D-apr83* lie in close positions from a global point of view (from the whole of the taxonomic groups), but are they also close from the point of view of every taxonomic group? In the same way, it is interesting to know, for example, if the species *Nemoura spp.* and *Eleuctra fusca*, very different from a global point of view (from the whole of the sites×dates) are alike at some date. The superimposed representations of the global and partial row-points (respectively of the global and partial column-points) will contribute to answer these questions.

5.1 Superimposed representation of the species

The global similarity between the six clouds of species, as induced by every date (partial row clouds) and as projected on the first and second axes, is measured by the ratio [between-inertia/total-inertia]. This ratio is equal to 31.9% and 38.8%, respectively. These relatively low values indicate that it exists a notable difference between the inter-species distances from one date to the other.

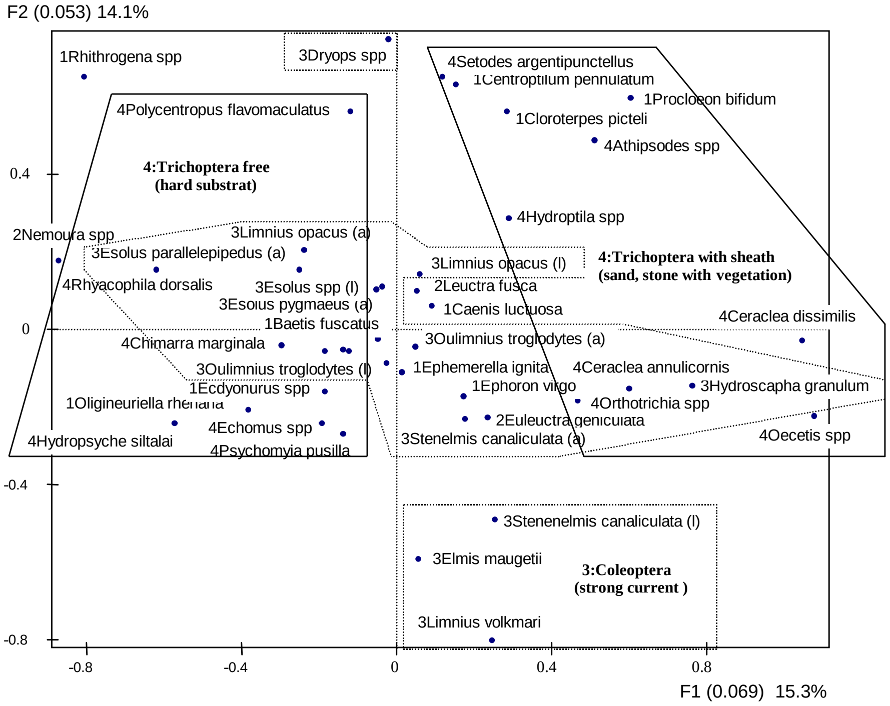

Fig. 3. The species on the first principal plane issued from ICA

In order to identify the species whose behavior varies more depending on the date, and to interpret these specific behaviors, we look for those which present the largest within-inertia on the first axes. Then, we search the partial point(s) responsible of the high dispersion of the concerned species, that are not in accordance to the other homologous partial points. For example, the species *Nemoura spp.* presents the second highest within-inertia on the first axis (equal to 10.4% of the total within-inertia on this axis, as summed up on all the rows). Furthermore, the partial point *Nemoura spp.-feb83* brings 75.9% of the within-inertia due to this species on the first axis. So, the position of this taxon (Fig. 4) suggests that it is a good indicator in February and in April, but not in summer: in fact, this species was almost never observed in summer (discarding one case). Moreover, discarding two cases, this taxon is the only plecopteran observed in February, which explains the more characteristic position of this partial point.

Regarding *Leuctra fusca* the most homogeneous of plecopteran (1.4% of the total within-inertia of the first axis), its partial points are globally close to the origin, except *feb83* (65.0% of within-inertia of this specie on first axis). In fact, this species was frequently observed, except in November and February. As not any other plecopteran was observed in November, *Leuctra fusca* is characteristic (by its absence) only in February. These examples

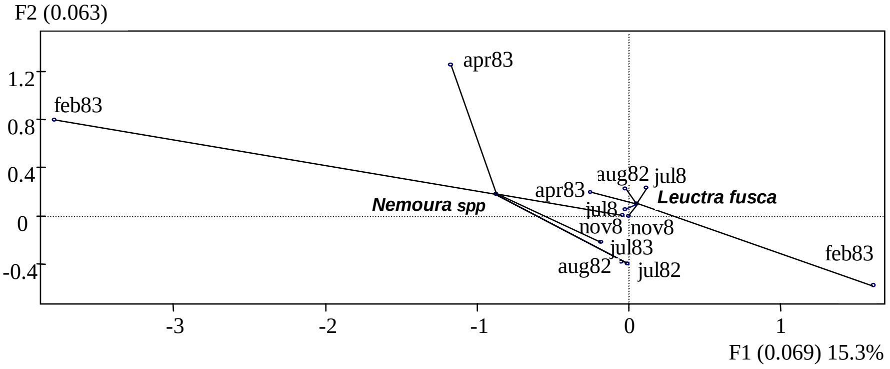

Fig. 4. Superimposed representation of two species belonging to *Plecoptera* on the first principal plane issued from ICA

5.2 Superimposed representation of the columns

The ratios [between-inertia/total-inertia] corresponding to the column clouds (the clouds of dates×sites as induced by each taxonomic group) as projected on the first and second axes are equal to 49.5% and 50.1%, respectively.

Figure 5 presents an excerpt of the superimposed representation, on the first principal plane, of the two dates×sites presenting the highest within-inertia on the first axis (see Table 1): *D-apr83* and *D-feb83* (i.e. the same site at different dates) and also of the dates×sites *C-jul82* and *D-jul82* to illustrate the case of different sites at the same date. Table 1 completes this representation with some information about dates×sites.

We can note that *D-apr83* and *D-feb83* are very similar as described globally (i.e. from all the taxonomic groups) but also as described by any taxonomic group: the partial points corresponding to the different taxonomic groups are close in every case. According to the graph, these two couple (date, site) are mainly characterized by *Trichoptera*. The two trichopterans having a high positive coordinate along axis 1 are *Caraclea dissimilis* and *Oecetis spp.* (Fig. 3). In fact, in February and April, these two taxa were observed quite only on the site *D*. From another point of view, the usual correspondence analysis applied only to the subtable (*Trichoptera*, *Feb83*) or the subtable (*Trichoptera*, *Apr83*) provides a first plane clearly showing the association between the site *D* and these two taxa. Concerning the sites *C* and *D* at July 82, we can see that, they have quite similar profiles in *Plecoptera* but different in *Coleoptera*.

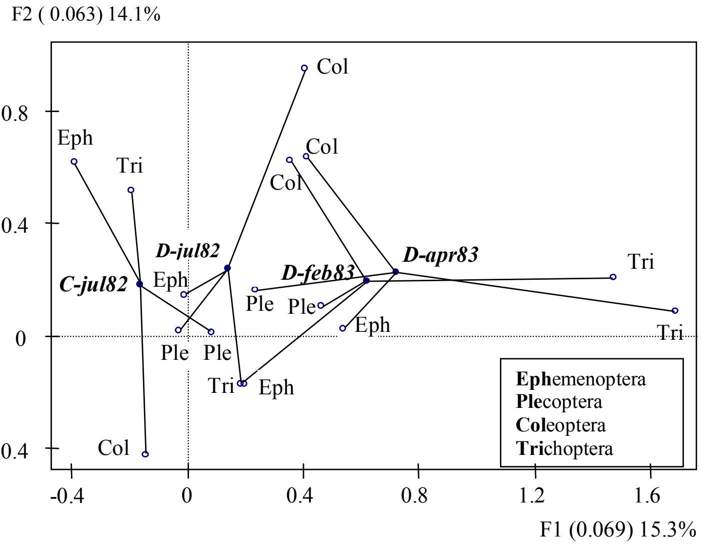

Fig. 5. Excerpt of the superimposed representation of the partial column points on the first principal plane issued from ICA

Table 1. Some interpretation aids of columns represented in Fig. 5

Column	Within Inertia x100000		Coordinates x1000		Contribution %		Weight %
	Axis-1	Axis-2	Axis-1	Axis-2	Axis-1	Axis-2	
D-apr83	309	55	720	227	28.6	30.8	3.8
D-feb83	268	88	619	190	23.7	24.2	4.2
C-jul82	19	117	-160	179	10.0	13.5	2.7
D-jul82	27	171	140	234	9.9	30.1	3.5

6 Conclusion

The comparison of the partial rows and columns structures enriches the results from ICA. This comparison induces a representation of the row-profiles (respectively, the column-profiles) not only from the global but also the partial points of view as induced by each group of columns (respectively, rows). In the case of the Ardèche example, the superimposed representation of the partial columns allows for better visualizing the taxonomic groups which are responsible of the differences observed among the sites according to the date.

Software note

The calculations are performed with ADE4 [Thioulouse *et al.*, 2004], in R environment [R Development Core Team, 2004].

References

[Benzécri, 1983]J.P. Benzécri. Analyse de l'inertie intraclasse par l'analyse d'un tableau de correspondances. *Les Cahiers de l'Analyse des Données*, 8(3):351–358, 1983.

[Cazes *et al.*, 1988]P. Cazes, D. Chessel, and S. Dolédec. L'analyse des correspondances internes d'un tableau partitionné. Son usage en hydrobiologie. *Revue de Statistique Appliquée*, 36(1):39–54, 1988. http://pbil.univ-lyon1.fr/R/articles/arti054.pdf.

[Escofier and Drouet, 1983]B. Escofier and D. Drouet. Analyse des différences entre plusieurs tableaux de fréquence. *Les Cahiers de l'Analyse des Données*, 8(4):491–499, 1983.

[Escofier and Pagès, 1994]B. Escofier and J. Pagès. Multiple factor analysis: afmult package. *Comput. Statist. Data Anal*, 18:121–140, 1994.

[Escofier and Pagès, 1998]B. Escofier and J. Pagès. *Analyses factorielles simples et multiples. Objectifs, méthodes et interprétation.* Dunod, Paris, 3 edition, 1998.

[Escofier, 1984]B. Escofier. Analyse factorielle en référence à un modèle. Application à l'analyse de tableaux d'échanges. *Revue de Statistique Appliquée*, 32(4):25–36, 1984.

[R Development Core Team, 2004]R Development Core Team. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, 2004. http://www.R-project.org.

[Thioulouse *et al.*, 2004]J. Thioulouse, A.B. Dufour, and D. Chessel. *ADE4: Analysis of Environmental Data : Exploratory and Euclidean method Multivariate data analysis and graphical display.* Lyon, France, November 2004. http://cran.univ-lyon1.fr/src/contrib/Descriptions/ade4.html.