REVUE DE STATISTIQUE APPLIQUÉE

D. CHESSEL
M. HANAFI

Analyses de la co-inertie de K nuages de points

Revue de statistique appliquée, tome 44, n° 2 (1996), p. 35-60
<http://www.numdam.org/item?id=RSA_1996__44_2_35_0>

Rev. Statistique Appliquée, 1996, **XLIV** (2), 35-60

ANALYSES DE LA CO-INERTIE DE *K* NUAGES DE POINTS

D. Chessel et M. Hanafi

URA CNRS 1974, Bât 401C
Université Lyon I
69622 Villeurbanne Cedex

RÉSUMÉ

Soient K études statistiques $(\mathbf{X}_k, \mathbf{Q}_k, \mathbf{D})(1 \leqslant k \leqslant K)$ portant sur les mêmes n individus et K groupes de variables, comptant respectivement $p_1, p_2, \ldots, p_K$ descripteurs. Soit p la somme des p_k . Cet ensemble de données définit un nuage de K points dans l'ensemble des opérateurs D-symétriques de $\mathbb{R}^n$ utilisé par la méthode ACT-STATIS (Lavit *et Coll.* 1994). Ce même ensemble définit un nuage de p points de $\mathbb{R}^n$ et n nuages de K points de $\mathbb{R}^p$ qui sont étudiés par l'Analyse Factorielle MULTiple (Escofier et Pagès 1994). Cet ensemble fournit enfin K nuages de n points dans K espaces euclidiens $\mathbb{R}^{p_k}$ distincts. L'article définit l'analyse de co-inertie de ces K nuages (Analyse de CO-inertie Multiple) en utilisant une généralisation de l'analyse de co-inertie (Dolédec et Chessel 1994) à l'analyse de co-inertie multiple voisine de la généralisation de l'analyse canonique à l'analyse canonique généralisée (Carrol 1968). Les méthodes ACT-STATIS, AFMULT et ACOM sont disponibles dans le logiciel ADE-4 en diffusion libre sur Internet.

Mots-clés : *analyse des données, schéma de dualité, K tableaux, ACT-STATIS, analyse factorielle multiple, AFMULT, co-inertie, analyse de co-inertie multiple, ACOM, analyse canonique, AC, analyse canonique généralisée, ACG.*

SUMMARY

Let be statistical K triplets $(\mathbf{X}_k, \mathbf{Q}_k, \mathbf{D})(1 \leqslant k \leqslant K)$ with K sets of variables measured in the same sampling units. n is the number of samples; $p_1, p_2, \ldots, p_K$ represent the number of variables, p being the total of the p_k. This data set defines a cloud of K points in the Euclidean space of the $(n \times n)$ **D**-symmetric matrices, used by the ACT-STATIS method (Lavit *et al.* 1994). The same data set provides a cloud of p points in $\mathbb{R}^n$ and n clouds of K points in $\mathbb{R}^p$ wich are analyzed via Multiple Factor Analysis (Escofier et Pagès 1994). Finally, this data set results in K clouds of n points in K different Euclidean spaces $\mathbb{R}^{p_k}$. This paper increases the scope of co-inertia analysis (Dolédec et Chessel 1994) that enables the analysis of two tables to the co-inertia analysis of K tables (Multiple CO-inertia Analysis). Such an extension is similar to the generalization of Canonical Correlation Analysis by Carrol (1968). ACT-STATIS, MFA and MCOA are available in ADE-4 software in freeware distribution on Internet.

Keywords : *Data analysis, duality diagram, K-tables, ACT-STATIS, multiple factor analysis (MFA), co-inertia, multiple co-inertia analysis (MCOA), canonical correlation analysis (CCA), generalized canonical analysis (GCA).*

1. Introduction

1.1. Objectifs

Dans cet article nous proposons d'étendre l'analyse de co-inertie de deux tableaux à l'analyse simultanée de K tableaux. L'analyse de co-inertie de deux tableaux est le nom générique qui recouvre (Chessel et Mercier 1993) l'analyse inter-batterie de Tucker 1958, l'analyse canonique sur variables qualitatives de Cazes 1980 et l'analyse des correspondances de tableaux de profils écologiques (Mercier *et Coll.* 1992). Sa stabilité numérique (Kazi-Aoual *et Coll.* 1995), sa facilité d'emploi en terme de double analyse d'inertie à coordonnées covariantes (Dolédec et Chessel 1994, Prodon et Lebreton 1994), son universalité en termes de conditions numériques et de type d'analyses de départ, ses propriétés de point de départ de la régression PLS (Tenehaus et Coll. 1995) en font une bonne alternative à l'analyse canonique des corrélations (Hotelling 1936), souvent impossible ou numériquement instable ou difficilement interprétable. C'est Tucker (1958) qui a ouvert la brèche en substituant à la maximisation d'un coefficient de corrélation entre variables canoniques celle de la covariance entre combinaison de variables. Optimiser une covariance impose de ne pas se désintéresser des variances des combinaisons donc des inerties projetées.

1.2. Notations utilisées

On considère K études statistiques $(\mathbf{X}_k, \mathbf{Q}_k, \mathbf{D})(1 \leqslant k \leqslant K)$ portant sur les mêmes n individus et K groupes de variables, comptant respectivement $p_1, p_2, \ldots, p_K$ descripteurs. Les notations sont celles du schéma de dualité (Escoufier 1977, 1987). $\mathbf{X}_k$ est un tableau $n \times p_k$ dont les lignes sont des observations sur n individus et les colonnes sont les mesures sur p_k variables. $\mathbf{Q}_k$ est une matrice $p_k \times p_k$ définie positive utilisée pour mesurer les distances entre individus dans l'espace vectoriel $\mathbb{R}^{p_k}$. $\mathbf{D}$ est une matrice diagonale $n \times n$ à éléments positifs de trace unité dite matrice des poids associés aux individus et utilisée pour mesurer les distances entre variables dans l'espace vectoriel $\mathbb{R}^n$.

Le rang de $\mathbf{X}_k$ est noté ρ_k, les valeurs propres de l'analyse du triplet k sont par ordre décroissant λ_k^j, ses composantes principales $\mathbf{D}$-normées sont $\mathbf{f}_k^j$ et ses axes principaux $\mathbf{Q}_k$-normés sont $\mathbf{e}_k^j$. On notera π_k un poids positif attribué *a priori* à l'analyse de $\mathbf{X}_k$. On désigne par ρ le minimum des ρ_k.

On note $\mathbf{X} = [\mathbf{X}_1 | \mathbf{X}_2 | \ldots \mathbf{X}_K]$ le tableau de dimension $n \times p$ avec $p = p_1 + \ldots p_K$ obtenu par juxtaposition des K tableaux $\mathbf{X}_k$. $\mathbf{X}$ définit un triplet statistique $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$, avec $\mathbf{Q}$ la matrice bloc diagonale constituée des matrices $\mathbf{Q}_k$. Le tout forme un K-tableaux et définit K analyses séparées. Celles-ci sont d'un type quelconque, en particulier des analyses en composantes principales (ACP) centrées, doublement centrées ou normées, des analyses des correspondances multiples (ACM) qui utilisent les indicatrices des classes et les schémas de Tenehaus et Young 1985 ou des analyses des correspondances floues (ACF, Chevenet *et Coll.* 1994).

1.3. Principes géométriques

Analyser un triplet $(\mathbf{X}_k, \mathbf{Q}_k, \mathbf{D})$ c'est géométriquement trouver les axes d'inertie d'un nuage de n points de $\mathbb{R}^{p_k}$ (axes principaux) ou trouver les axes

d'inertie d'un nuage de p_k points de $\mathbb{R}^n$ (composantes principales). Les solutions implique la diagonalisation des opérateurs d'inertie $\mathbf{W}_K\mathbf{D} = \mathbf{X}_k\mathbf{Q}_k\mathbf{X}_k^t\mathbf{D}$ et $\mathbf{V}_k\mathbf{Q}_k = \mathbf{X}_k^t\mathbf{D}\mathbf{X}_k\mathbf{Q}_k$ (figure 1).

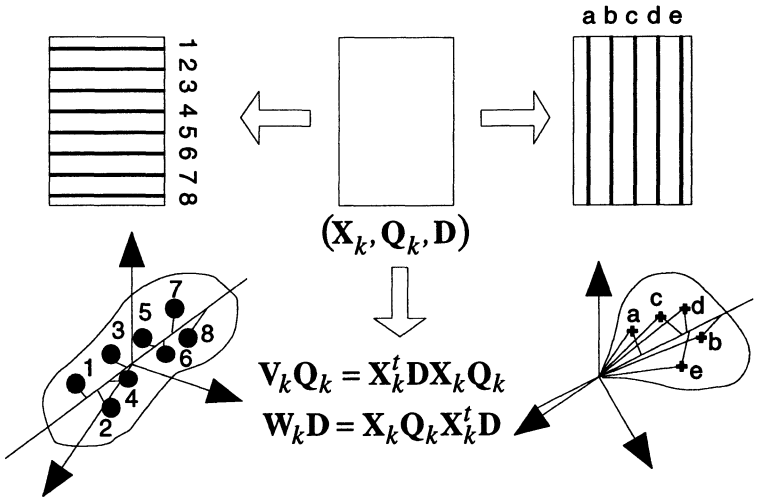

FIGURE 1 :
Deux nuages de points et deux opérateurs d'inertie sont associés à un triplet statistique

Généraliser la co-inertie à K tableaux permet d'identifier l'espace libre laissé entre ce que les revues récentes de Bove et Di Ciaccio (1994) et de Rizzi et Vichi (1995) donnent comme les principales méthodes d'ordination simultanée d'essence euclidienne respectivement ACT-STATIS et l'analyse factorielle multiple (AFMULT).

STATIS (Lavit 1988, Lavit *et coll.* 1994) s'appuie sur le passage du triplet $(\mathbf{X}_k, \mathbf{Q}_k, \mathbf{D})$ à son opérateur d'inertie $\mathbf{W}_k\mathbf{D}$ élément de l'espace euclidien des opérateurs $\mathbf{D}$-symétriques de $\mathbb{R}^n$ muni du produit scalaire d'Hilbert-Schmidt (Escoufier 1973). L'analyse du nuage d'opérateurs définit un système d'axes d'inertie (interstructure) dont le premier a des propriétés très particulière comme tout premier axe d'une ACP non centrée. Le passage dans l'espace des opérateurs a pour prix un retour dans les espaces de départ sur lequel nous reviendrons.

L'AFMULT (Escofier et Pagès 1984, 1986, 1994) reste dans l'espace commun des nuages initiaux donc dans $\mathbb{R}^n$ (figure 2). Le retour dans l'espace des variables se fait par le biais de l'association lignes-colonnes par le tableau de la figure 3 qui permet à Casin et Turlot 1986 d'identifier l'analyse canonique généralisée (ACG) comme un cas particulier de l'analyse discriminante, l'analyse discriminante étant connue comme un cas particulier de l'analyse canonique classique (Caillez et Pagès 1976) et l'AFMULT est alors un cas particulier de l'ACP inter-classes. Il vient à l'esprit que la notion de cas particulier en termes de théorie peut paraître curieux au praticien.

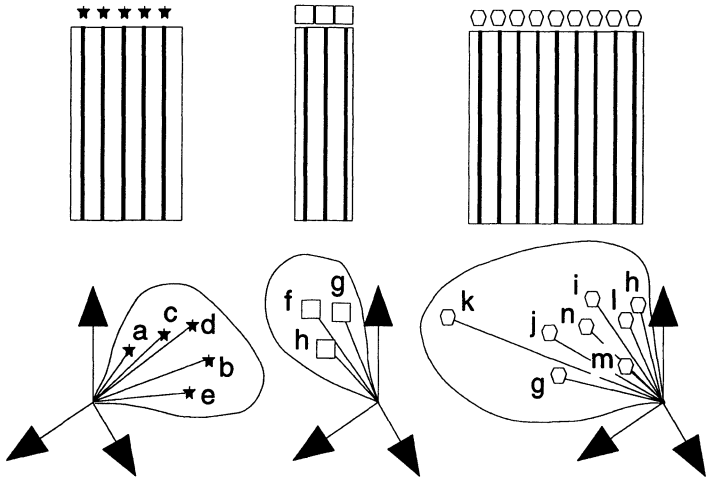

FIGURE 2 :

K nuages de p_k points de $\mathbb{R}^n$ sont réunis pour former la base de départ de l'AFMULT.

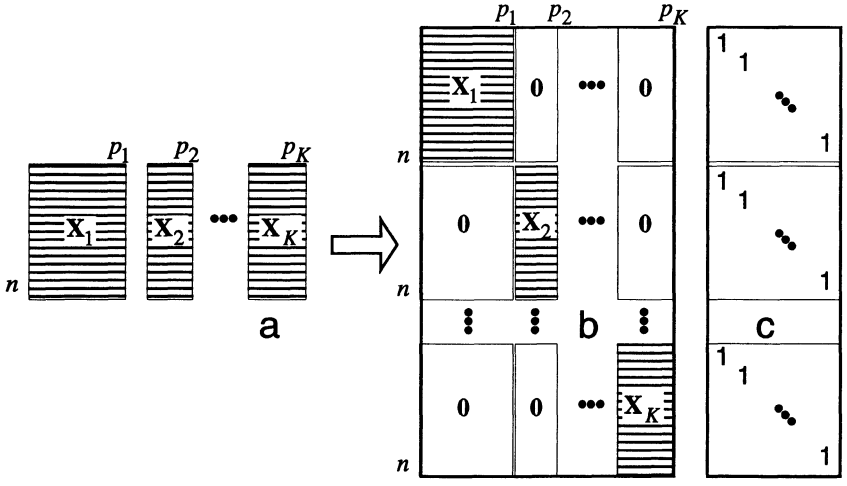

FIGURE 3 :

n groupes de K points de $\mathbb{R}^p$ sont utilisés dans l'AFMULT. a) Tableaux de départ. b) Tableau groupé dont les lignes donnent nK points de $\mathbb{R}^p$. c) Tableau des indicatrices des groupes de points utilisés implicitement par l'AFMULT pour faire l'analyse inter-classe. Cette analyse inter-classes définie par Escofier et Pagès 1984 est figurée dans la présentation de Casin et Turlot 1986.

Il reste alors un objet non envisagé, celui des K nuages dans K espaces séparés (figure 4). Dans une première partie, nous donnerons le principe géométrique de la co-inertie de K nuages de points appariés en utilisant les variables auxiliaires de l'Analyse Canonique Généralisée. Dans une seconde partie, sera présentée l'existence de la solution du problème d'optimisation associé et l'algorithme de calcul. Une troisième partie comparera sur un exemple les trois méthodes STATIS, l'AFMULT et ce que nous proposons d'appeler l'analyse de co-inertie multiple (ACOM).

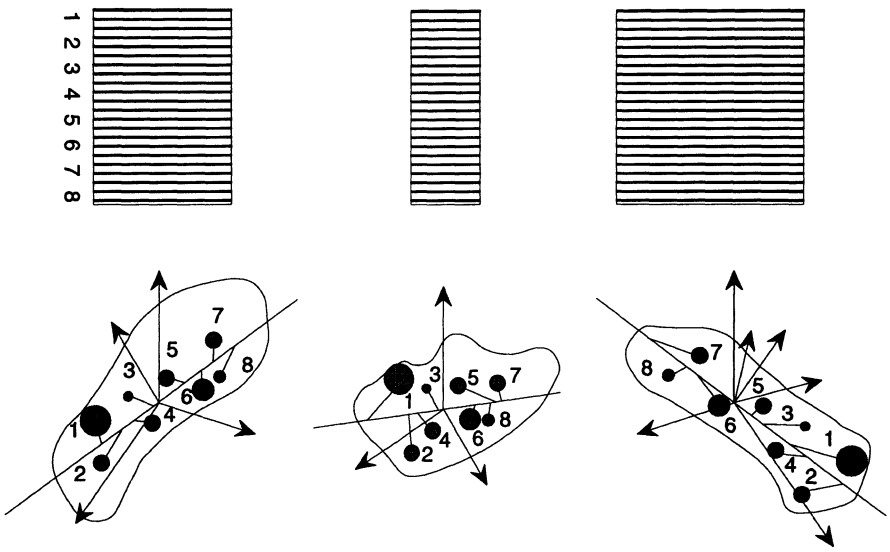

FIGURE 4 :
K nuages de n points des espaces $\mathbb{R}^{p_k}$ sont la base de départ de l'ACOM.

2. Variables auxiliaires et analyse canonique généralisée

Initialement l'Analyse Canonique de la Corrélation (ACC) s'occupe du cas $K = 2$ et est décrite par Hotteling (1936). Plusieurs généralisation ($K > 2$) sont présentées par Kettenring (1971) parmi lesquelles on retrouve celle de Carroll (1968). Celle-ci consiste à trouver d'abord au pas 1 des solutions $\mathbf{u}_k^1$ dans chaque espace $\mathbb{R}^{p_k}$, et une variable dite auxiliaire $\mathbf{v}^1$, $\mathbf{D}$-normée dans l'espace vectoriel $\mathbb{R}^n$, qui maximise la quantité :

$$f(\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_K, \mathbf{v}) = \sum_{k=1}^{K} \pi_k r^2(\mathbf{X}_k \mathbf{Q}_k \mathbf{u}_k, \mathbf{v})$$

où $r(.,.)$ désigne la corrélation linéaire entre les variables entre parenthèses. Au pas 2, on cherche des solutions $\mathbf{u}_k^2$ normés dans chaque espace $\mathbb{R}^{p_k}$, et une variable dite

auxiliaire $\mathbf{v}^2$, dans $\mathbb{R}^n$ qui maximise la même quantité sous contrainte d'orthogonalité. Au pas s, la même quantité est maximisée sous les contraintes :

$$\left(\mathbf{v}^j|\mathbf{v}^s\right)_{\mathbf{D}} = 0 \quad (1 \leqslant j < s)$$

Kiers *et Coll.* (1994, p. 332) en se référant à Tenenhaus et Young (1985 p. 100) rappelle que la maximisation successive énoncé ci-dessus est également la solution de la maximisation simultanée du critère :

$$\sum_{j=1}^{s}(\mathbf{u}_1^j, \mathbf{u}_2^j, \ldots, \mathbf{u}_K^j, \mathbf{v}_j) = \sum_{j=1}^{s}\sum_{k=1}^{K} \pi_k r^2(\mathbf{X}_k\mathbf{Q}_k\mathbf{u}_k^j, \mathbf{v}_j)$$

On trouve dans Saporta (1975) et Tenenhaus (1984) un exposé théorique de la méthode de Carroll, ainsi que des applications aux variables qualitatives et aux mélanges de type de variables. L'apport fondamental de l'ACG est l'utilisation des variables auxiliaires. Sabatier (1993), après Kettering (1971) rappelle qu'on peut effectivement s'en passer en voyant dans l'ACG au pas 1 la solution de la recherche du maximum de (Sabatier 1993, CR1, p. 109) :

$$\sum_{k=1}^{K}\sum_{l=1}^{K}\left(\mathbf{X}_k\mathbf{Q}_k\mathbf{u}_k^1|\mathbf{X}_l\mathbf{Q}_l\mathbf{u}_l^1\right)_{\mathbf{D}} = \left|\left|\sum_{k=1}^{K}\mathbf{X}_k\mathbf{Q}_k\mathbf{u}_k^1\right|\right|_{\mathbf{D}}^2$$

ceci dans le cas d'une pondération unitaire des tableaux avec la contrainte (Sabatier 1993, CO3, p. 109) :

$$\sum_{k=1}^{K}\left|\left|\mathbf{X}_k\mathbf{Q}_k\mathbf{u}_k^1\right|\right|_{\mathbf{D}}^2 = 1$$

Les généralisations de Kettering (1971) éliminent également cette notion de variables auxiliaires en maximisant, entre autres :

$$\sum_{k=1}^{K}\sum_{l=1}^{K} r\left(\mathbf{X}_k\mathbf{Q}_k\mathbf{u}_k^1, \mathbf{X}_l\mathbf{Q}_l\mathbf{u}_l^1\right) \qquad \text{[SUMCOR]}$$

$$\lambda_1\left(\left[\mathbf{X}_k\mathbf{Q}_k\mathbf{u}_k^1|\mathbf{X}_l\mathbf{Q}_l\mathbf{u}_l^1\right]_{\substack{1\leqslant k\leqslant K\\ 1\leqslant l\leqslant K}}\right) \qquad \text{[MAXVAR]}$$

où $\lambda_1(\mathbf{A})$ désigne la plus grande valeur propre de $\mathbf{A}$. Mais l'apport des variables auxiliaires dans l'interprétation est très important. L'ACG est la seule généralisation de l'analyse canonique qui en fait usage. L'analyse PRINQUAL de Tenenhaus (1977) étend cette notion en cherchant à maximiser, pour m fixé :

$$\sum_{k=1}^{K}\pi_k\sum_{j=1}^{m} r^2(\mathbf{X}_k\mathbf{Q}_k\mathbf{u}_k^1, \mathbf{v}_j) = \sum_{k=1}^{K}\pi_k\mathbb{R}^2\left(\mathbf{X}_k\mathbf{Q}_k\mathbf{u}_k^1, \mathbf{v}_1, \ldots, \mathbf{v}_m\right)$$

L'analyse discriminante de tableaux (ADT) de Casin (1995) reprend le concept en introduisant la contrainte :

$$(\mathbf{X}_k\mathbf{Q}_k\mathbf{u}_k^r|\mathbf{X}_k\mathbf{Q}_k\mathbf{u}_k^s)_{\mathbf{D}} = 0 \quad (1 \leqslant r < s)$$

Les variables auxiliaires sont encore utilisées pour étendre l'ACG à d'autres critères de corrélation par Lazraq *et Coll.* (1992).

L'ACG a pour autre propriété très favorable d'être de mise en œuvre aisée. En effet, elle traite de plusieurs tableaux et admet une résolution algorithmique simple. C'est une méthode aux valeurs propres demandant une diagonalisation en dimension n (diagonalisation d'une somme de projecteurs) ou en dimension p (Saporta 1975, Casin et Turlot 1986).

Pour l'utilisateur de l'analyse des données, par contre, l'ACG porte toutes les limitations associées aux méthodes qui utilisent les projecteurs ou implicitement les métriques de Mahalanobis $(\mathbf{X}_k^t\mathbf{D}\mathbf{X}_k)^-$, comme la régression multiple (alternatives dans la régression PLS ou la régression sur composantes), comme l'analyse discriminante (alternative dans l'analyse inter-classe), comme dans l'analyse canonique et les analyses sur variables instrumentales (alternative dans l'analyse de co-inertie). Elle ne s'impose que dans le cas d'un grand nombre d'individus par rapport au nombre maximal de variables dans un groupe ou dans le cas des variables orthogonales par tableaux (comme dans le cas des indicatrices des modalités qui redonne l'analyse des correspondances multiples).

L'ACG est basée sur la notion de variables auxiliaires et de substitution des projecteurs $\mathbf{P}_k = \mathbf{X}_k(\mathbf{X}_k^t\mathbf{D}\mathbf{X}_k)^-\mathbf{X}_k^t\mathbf{D}$ aux tableaux (Van de Geer 1984) : nous voulons garder la première propriété mais remplacer la seconde par l'usage des opérateurs d'inertie $\mathbf{W}_k\mathbf{D} = \mathbf{X}_k\mathbf{Q}_k\mathbf{X}_k^t\mathbf{D}$. Cela définit l'analyse de co-inertie multiple.

3. L'analyse de co-inertie multiple

3.1. Définition

On appellera Analyse de CO-inertie multiple (ACOM) d'un multi-tableau au pas 1, la recherche de k vecteurs $\mathbf{u}_k^1$ normés dans chaque espace $\mathbb{R}^{p_k}$, et une variable dite auxiliaire $\mathbf{v}^1$, $\mathbf{D}$-normée dans l'espace vectoriel $\mathbb{R}^n$, qui maximise la quantité :

$$g(\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_K, \mathbf{v}) = \sum_{k=1}^{K} \pi_k \left(\mathbf{X}_k\mathbf{Q}_k\mathbf{u}_k|\mathbf{v}\right)_{\mathbf{D}}^2$$

Au pas 2, on cherche des solutions $\mathbf{u}_k^2$ normé dans chaque espace $\mathbb{R}^{p_k}$, et une variable dite auxiliaire $\mathbf{v}^2$ normée dans $\mathbb{R}^n$ qui maximise la même quantité sous contrainte d'orthogonalité.

Au pas s, la même quantité est maximisée sous les contraintes

$$\left(\mathbf{v}^j|\mathbf{v}^s\right)_{\mathbf{D}} = 0 \ \ (1 \leqslant j < s) \quad \text{et} \quad \left(\mathbf{u}_k^j|\mathbf{u}_k^s\right)_{\mathbf{Q}_k} = 0 \ \ (1 \leqslant j < s,\ 1 \leqslant k \leqslant K).$$

3.2. Solution d'ordre 1

Pour $\mathbf{v}$ un vecteur fixé, $\mathbf{D}$-normé dans $\mathbb{R}^n$, l'application de l'inégalité de Cauchy-Schwartz montre que la quantité $(\mathbf{X}_k\mathbf{Q}_k\mathbf{u}_k|\mathbf{V})^2_{\mathbf{D}}$ est majorée par $||\mathbf{X}_k^t\mathbf{D}\mathbf{v}||^2_{\mathbf{Q}_k}$. De plus ce majorant est atteint pour :

$$\mathbf{u}_k = \frac{\mathbf{X}_k^t\mathbf{D}\mathbf{v}}{||\mathbf{X}_k^t\mathbf{D}\mathbf{v}||_{\mathbf{Q}_k}} \tag{1}$$

Il s'ensuit que le vecteur $\mathbf{v}^1$, $\mathbf{D}$-normé dans l'espace vectoriel $\mathbb{R}^n$, qui maximise la fonction g, maximise également la quantité :

$$\sum_{k=1}^{K} \pi_k ||\mathbf{X}_k^t\mathbf{D}\mathbf{v}||^2_{\mathbf{Q}_k} = \sum_{k=1}^{K} \pi_k \left(\mathbf{X}_k\mathbf{Q}_K\mathbf{X}_k^t\mathbf{D}\mathbf{v}|\mathbf{v}\right)_{\mathbf{D}} = \sum_{k=1}^{K} \pi_k (\mathbf{W}_k\mathbf{D}\mathbf{v}|\mathbf{v})_{\mathbf{D}}$$

$$= \mathbf{v}^t\mathbf{D}\left(\sum_{k=1}^{K} \pi_k \mathbf{W}_k\mathbf{D}\right)\mathbf{v}$$

Donc $\mathbf{v}^1$ est la première composante $\mathbf{D}$-normé de l'ACP du tableau pondéré $\tilde{\mathbf{X}}$, où le bloc $\mathbf{X}_k$ est pondéré par $\sqrt{\pi_k}$. Il en découle que les axes $\mathbf{u}_k^1$, $\mathbf{Q}_k$- normés dans $\mathbb{R}^{p_k}$, sont les vecteurs de $\mathbb{R}^{p_k}$ normalisés par bloc du premier axe principal du tableau $\tilde{\mathbf{X}}$. La première variable auxiliaire de l'ACOM est donc identique à la première coordonnée normalisée des lignes de l'AFMULT et les premiers axes de co-inertie ont pour composantes la première coordonnée normalisée par blocs des colonnes de l'AFMULT.

3.3. Solution d'ordre 2

Proposition : les solutions d'ordre 2 existent et sont obtenues par les étapes :

1 – considérer les projecteurs $\mathbf{Q}_k$-orthogonaux notés $\mathbf{P}_k^1$ sur les sous-espaces vectoriels de $\mathbf{R}^{p_k}$ engendrés par le vecteur $\mathbf{u}_k^1$;

2 – définir le tableau $\mathbf{Z}$ de la manière suivante :

$$\mathbf{Z} = [\mathbf{Z}_1|\mathbf{Z}_2|\ldots|\mathbf{Z}_K] \quad \text{avec} \quad \mathbf{Z}_k = \mathbf{X}_k - \mathbf{X}_k\mathbf{P}_k^{1^t};$$

3 – calculer les solutions de rang 1 de co-inertie multiple du tableau $\mathbf{Z}$.

Démonstration : on définit les deux problèmes suivant :

Problème 1 : Trouver $\mathbf{a}_k^1 (1 \leqslant k \leqslant K)$ $\mathbf{Q}_k$-normés et $\mathbf{w}^1\mathbf{D}$-normé, solution de rang 1 de co-inertie multiple du tableau $\mathbf{Z} = [\mathbf{Z}_1|\mathbf{Z}_2|\ldots|\mathbf{Z}_K]$.

Problème 2 : Trouver a_k^2 $(1 \leqslant k \leqslant K)\mathbf{Q}_k$-normés et $\mathbf{w}^2$ $\mathbf{D}$-normé qui maximisent $\sum_{k=1}^{K}(\mathbf{X}_k\mathbf{Q}_k\mathbf{u}_k|\mathbf{v})_\mathbf{D}^2$ sous les contraintes $(\mathbf{v}|\mathbf{v}_1)_\mathbf{D} = 0$ et $\left(\mathbf{u}_k^1|\mathbf{u}_k\right)_{\mathbf{Q}_k} = 0$. $(1 \leqslant k \leqslant K)$.

Les solutions du problème 1 et du problème 2 existent toujours car les fonctions à maximiser sont continues et les contraintes définissent des ensembles compacts. Nous allons montrer que les solutions du problème 1 et du problème 2 sont identiques. Sachant que la solution du problème 1 est celle du paragraphe précédent, la solution du problème 2 sera alors acquise.

1) Commençons par montrer que $\mathbf{w}^1$ est $\mathbf{D}$-orthogonal à $\mathbf{v}^1$. En effet, $\mathbf{v}^1$ est vecteur propre de la somme des opérateurs d'inertie (§3.2) :

$$\left(\sum_{k=1}^{K}\pi_k\mathbf{W}_k\mathbf{D}\right)\mathbf{v}^1 = \lambda\mathbf{v}^1$$

De même compte tenu de l'égalité :

$$\mathbf{Z}_k\mathbf{Q}_k\mathbf{Z}_k^t = \mathbf{W}_k - \mathbf{X}_k\mathbf{Q}_k\mathbf{P}_k^1\mathbf{X}_k^t,$$

résultat obtenu en utilisant la $\mathbf{Q}_k$-symétrie de $\mathbf{P}_k^1$ on a :

$$\left(\sum_{k=1}^{K}\pi_k\mathbf{W}_k\right)\mathbf{D}\mathbf{w}^1 - \left(\sum_{k=1}^{K}\pi_k\mathbf{X}_k\mathbf{Q}_k\mathbf{P}_k^1\mathbf{X}_k^t\right)\mathbf{D}\mathbf{w}^1 = \mu\mathbf{w}^1$$

D'où, toujours grâce à la $\mathbf{Q}_k$-symétrie des opérateurs en jeu :

$$\left(\left(\sum_{k=1}^{K}\pi_k\mathbf{W}_k\right)\mathbf{D}\mathbf{v}^1|\mathbf{w}^1\right)_\mathbf{D} - \left(\left(\sum_{k=1}^{k}\pi_K\mathbf{X}_k\mathbf{Q}_k\mathbf{P}_k^1\mathbf{X}_k^t\right)\mathbf{D}\mathbf{v}^1|\mathbf{w}^1\right)_\mathbf{D} = \mu\left(\mathbf{v}^1|\mathbf{w}^1\right)_\mathbf{D} \tag{2}$$

D'autre part :

$$\left(\left(\sum_{k=1}^{K}\pi_k\mathbf{W}_k\right)\mathbf{D}\mathbf{v}^1|\mathbf{w}^1\right)_\mathbf{D} = \lambda\left(\mathbf{v}^1|\mathbf{w}^1\right)_\mathbf{D}$$

De plus, d'après (1) et puisque $\mathbf{P}_k^1\mathbf{u}_k^1 = \mathbf{u}_k^1$ par définition, on a :

$$\mathbf{P}_k^1\mathbf{X}_k^t\mathbf{D}\mathbf{v}^1 = ||\mathbf{X}_k^t\mathbf{D}\mathbf{v}^1||_{\mathbf{Q}_k}\mathbf{u}_K^1 = \mathbf{X}_k^t\mathbf{D}\mathbf{v}^1$$

D'où :

$$\mathbf{X}_k\mathbf{Q}_k\mathbf{P}_k^1\mathbf{X}_k^t\mathbf{D}\mathbf{v}^1 = \mathbf{W}_k\mathbf{D}\mathbf{v}^1$$

Donc :

$$\left(\left(\sum_{k=1}^{K}\pi_k\mathbf{X}_k\mathbf{Q}_k\mathbf{P}_k^1\mathbf{X}_k^t\right)\mathbf{D}\mathbf{v}^1|\mathbf{w}^1\right)_{\mathbf{D}} = \left(\sum_{k=1}^{K}\pi_k\mathbf{X}_k\mathbf{Q}_k\mathbf{P}_k^1\mathbf{X}_k^t\mathbf{D}\mathbf{v}^1|\mathbf{w}^1\right)_{\mathbf{D}}$$

$$= \sum_{k=1}^{K}\pi_k\left(\mathbf{X}_k\mathbf{Q}_k\mathbf{P}_k^1\mathbf{X}_k^t\mathbf{D}\mathbf{v}^1|\mathbf{w}^1\right)_{\mathbf{D}} = \sum_{k=1}^{K}\pi_k\left(\mathbf{W}_k\mathbf{D}\mathbf{v}^1|\mathbf{w}^1\right)_{\mathbf{D}} \qquad (3)$$

De (2) et (3), l'on déduit que $\mu\left(\mathbf{v}^1|\mathbf{w}^1\right)_{\mathbf{D}} = 0$, d'où en excluant le cas dégénéré $\mu = 0$:

$$\left(\mathbf{v}^1|\mathbf{w}^1\right)_{\mathbf{D}} = 0.$$

2) Montrons alors que, pour $1 \leqslant k \leqslant K$, $\mathbf{a}_k^1$ est $\mathbf{Q}_k$-orthogonal à $\mathbf{u}_k^1$.

D'après la solution de rang 1 du tableau $\mathbf{Z}$ on $\mathbf{a}_k^1 = \dfrac{\mathbf{Z}_k^t\mathbf{D}\mathbf{w}^1}{||\mathbf{Z}_k^t\mathbf{D}\mathbf{w}^1||_{\mathbf{Q}_k}}$ et $\mathbf{w}^1$ est la première composante principale du tableau $\tilde{\mathbf{Z}} = \left[\tilde{\mathbf{Z}}_1|\tilde{\mathbf{Z}}_2|\ldots|\tilde{\mathbf{Z}}_K\right]$ avec $\tilde{\mathbf{Z}}_k = \sqrt{\pi_k}\mathbf{Z}_k$.

Le caractère idempotent des projecteurs implique :

$$\left(\mathbf{I}_k - \mathbf{P}_k^1\right)\mathbf{a}_k^1 = \frac{(\mathbf{I}_k - \mathbf{P}_k^1)\mathbf{Z}_k^t\mathbf{D}\mathbf{w}^1}{||\mathbf{Z}_k^t\mathbf{D}\mathbf{w}^1||_{\mathbf{Q}_k}} = \frac{(\mathbf{I}_k - \mathbf{P}_k^1)\mathbf{X}_k^t\mathbf{D}\mathbf{w}^1}{||\mathbf{Z}_k^t\mathbf{D}\mathbf{w}^1||_{\mathbf{Q}_k}} = \frac{\mathbf{Z}_k^t\mathbf{D}\mathbf{w}^1}{||\mathbf{Z}_k^t\mathbf{D}\mathbf{w}^1||_{\mathbf{Q}_k}} = \mathbf{a}_k^1$$

Le développement de $\mathbf{a}_k^1 = (\mathbf{I}_k - \mathbf{P}_k^1)\mathbf{a}_k^1$ donne alors :

$$\mathbf{a}_k^1 = \mathbf{a}_k^1 - \left(\mathbf{a}_k^1|\mathbf{u}_k^1\right)_{\mathbf{Q}_k}\ \mathbf{u}_k^1 \Longrightarrow \left(\mathbf{a}_k^1|\mathbf{u}_k^1\right)_{\mathbf{Q}_k} = 0$$

En conséqence, les solutions du problème 1 vérifient les contraintes du problème 2, donc par définition de la solution du problème 2 et compte tenu de ce que :

$$\mathbf{Z}_k\mathbf{Q}_k = \mathbf{X}_k(\mathbf{I}_k - \mathbf{P}_k^{1^t})\mathbf{Q}_k = \mathbf{X}_k\mathbf{Q}_k(\mathbf{I}_k - \mathbf{P}_k^1)$$

on a :

$$\sum_{k=1}^{K}\pi_k\left(\mathbf{Z}_k\mathbf{Q}_k\mathbf{a}_k^1|\mathbf{w}^1\right)_{\mathbf{D}}^2 = \sum_{k=1}^{K}\pi_k\left(\mathbf{X}_k\mathbf{Q}_k\mathbf{a}_k^1|\mathbf{w}^1\right)_D^2 \leqslant \sum_{k=1}^{K}\pi_k\left(\mathbf{X}_k\mathbf{Q}_k\mathbf{a}_k^2|\mathbf{w}^2\right)_{\mathbf{D}}^2$$

De même, par définition de la solution du problème 1, la solution du problème 2 vérifie :

$$\sum_{k=1}^{K} \left(\mathbf{Z}_k \mathbf{Q}_k \mathbf{a}_k^2 | \mathbf{w}^2\right)_{\mathbf{D}}^2 = \sum_{k=1}^{K} \pi_k \left(\mathbf{X}_k \mathbf{Q}_k (\mathbf{I} - \mathbf{P}_k^1) \mathbf{a}_k^2 | \mathbf{w}^2\right)_{\mathbf{D}}^2$$

$$= \sum_{k=1}^{K} \Pi_k \left(\mathbf{X}_k \mathbf{Q}_k \mathbf{a}_k^2 | \mathbf{w}^2\right)_{\mathbf{D}}^2 \leqslant \sum_{k=1}^{K} \pi_k (\mathbf{Z}_k \mathbf{Q}_k \mathbf{a}_k^1 | \mathbf{w}^1)_{\mathbf{D}}^2 = \sum_{k=1}^{K} \pi_k \left(\mathbf{X}_k \mathbf{Q}_k \mathbf{a}_k^1 | \mathbf{w}^1\right)_{\mathbf{D}}^2$$

Donc :

$$\sum_{k=1}^{K} \pi_k \left(\mathbf{X}_k \mathbf{Q}_k \mathbf{a}_k^2 | \mathbf{w}^2\right)_{\mathbf{D}}^2 = \sum_{k=1}^{K} \pi_k \left(\mathbf{Z}_k \mathbf{Q}_k \mathbf{a}_k^1 | \mathbf{w}^1\right)_{\mathbf{D}}^2$$

3.4. Solution d'ordre s

Au pas s, pour $s \geqslant 2$, la solution est alors obtenue en effectuant les trois étapes suivantes :

1 – considérer les projecteurs $\mathbf{Q}_k$-orthogonaux notés $\mathbf{P}_k^{s-1}$ sur les sous-espaces vectoriels de $\mathbb{R}^{p_k}$ engendrés par les systèmes orthonormés $\left\{\mathbf{u}_k^1, \ldots, \mathbf{u}_k^{s-1}\right\}$;

2 – remplacer le tableau $\mathbf{X}$ par le tableau $\mathbf{X}^s$ de la manière suivante :

$$\mathbf{Z} = \left[\mathbf{X}_1 - \mathbf{X}_1 \mathbf{P}_1^{s-1^t} | \mathbf{X}_2 - \mathbf{X}_2 P_2^{s-1^t} | \ldots | \mathbf{X}_k - \mathbf{X}_K \mathbf{P}_K^{s-1^t}\right] ;$$

3 – calculer les solutions de rang 1 de co-inertie multiple du tableau $\mathbf{Z}$.

3.5. Aides à l'interprétation

Globalement l'algorithme de l'ACOM fonctionne comme une AFMULT de rang 1 dans une boucle. Il fournit au rang s, pour chaque triplet de départ une base orthonormée d'axes de co-inertie sur lesquelles on projette les nuages de points initiaux et on obtient K analyses simples coordonnées. On projette de même les axes d'inertie des analyses simples et on peut comparer la valeur des systèmes d'inertie et des systèmes de co-inertie par les inerties projetées sur les deux systèmes. Les cartes factorielles dans chaque espace peuvent se superposer après normalisation des coordonnées et se superposer ensemble avec les variables auxiliaires de même rang pour exprimer la part qui revient à la corrélation coordonnée/variable auxiliaire dans la covariance correspondante. Il suffit de penser qu'optimiser un carré de covariance entre coordonnée et variable auxiliaire normée, c'est optimiser le produit variance de la coordonnée (inertie projetée) par carré de corrélation.

La question pratique essentielle est la même que pour l'AFMULT et touche la pondération des tableaux. Il convient d'éviter qu'un tableau de forte inertie (tant par les

normes des variables que par leur nombre) ne prenne une importance disproportionnée dans la définition des variables auxiliaires. Le module KTA du logiciel ADE-4 propose trois options classiques, respectivement la pondération uniforme, la pondération par l'inverse de l'inertie totale et la pondération par l'inverse de la première valeur propre de l'analyse séparée. La seconde option est proposée par défaut et a été utilisée dans l'illustration.

L'ACOM est programmée pour l'assemblage de triplets statistiques initiaux de nature arbitraire (variables quantitatives, qualitatives ou distributionnelles) avec des pondérations colonnes paramétrables. On s'en tiendra ici à une illustration comparée des trois méthodes STATIS, AFMULT (avec l'utilisation de la troisième pondération) et ACOM sur un exemple. L'essentiel de l'ACOM étant de coordonner K analyses élémentaires par l'usage des variables auxiliaires, la méthode ne pose aucune difficulté particulière d'interprétation pour les praticiens des analyses de données classiques.

4. Illustrations comparées

4.1. Données traitées

Les illustrations utilisent les données publiées par L.E. Friday (1987) qui posent clairement la question de la valeur typologique des groupes faunistiques en zoologie. Dans $n = 16$ étangs, on a mesuré l'abondance de $p = 91$ espèces réparties en $K = 10$ groupes faunistiques comportant p_k taxons, avec respectivement $p_1 = 11$ (Hémiptera), $p_2 = 7$ (Odonata), $p_3 = 13$ (Trichoptera), $p_4 = 4$ (Ephemeroptera), $p_5 = 13$ (Coleoptera), $p_6 = 22$ (Diptera), $p_7 = 4$ (Hydracarina), $p_8 = 3$ (Malacostraca), $p_9 = 8$ (Mollusca), et $p_{10} = 6$ (Oligochaeta). Pour respecter l'article cité les étangs portent les étiquettes A, B, ..., R, les lettres I et O n'étant pas utilisées. L'ACP centrée par variables (espèces) est utilisée sur 10 tableaux simultanément et on cherche à caractériser la capacité de chaque groupe de descripteurs à produire ou reproduire la typologie de stations éventuellement induite partiellement ou totalement par tout ou partie de l'ensemble des tableaux. Les données utilisées sont strictement celles de leur auteur (Friday 1987 p. 96 et 97) à l'exclusion près des groupes ne comportant qu'un seul ou deux taxons.

4.3. Utilisation de STATIS

STATIS, contrairement à l'AFMULT et l'ACOM, construit son compromis avant d'en faire l'analyse (Pagès 1995). C'est un point essentiel. En laissant faire l'interstructure (typologie d'opérateurs, figure 5) avant la description du compromis (moyenne d'opérateur) STATIS indique d'abord si il est légitime de faire une moyenne de structure (*a common pattern of correlation*, Génard *et Coll.* 1994). Ce faisant, la méthode propose une approche typologique (interstructure) avant la notion plus simple de moyenne (compromis), ce qui est une source de difficultés pour l'utilisateur non professionnel. Le tableau 1 donne pour chaque tableau, le nombre de variables, le carré de la norme d'Hilbert-Schmidt des opérateurs associés, le coefficient de corrélation vectorielle avec les autres (RV) et le poids du tableau dans la constitution du compromis.

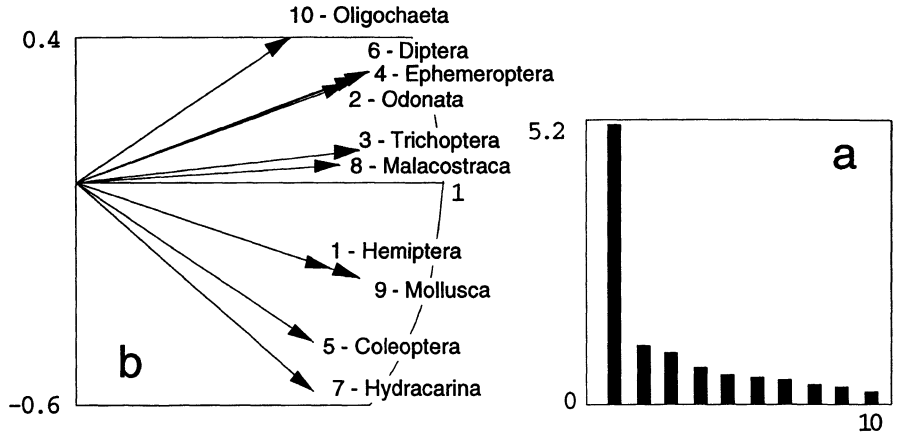

FIGURE 5 :

Interstructure de STATIS. a) Valeurs propres de la matrice des coefficients RV. b) Image euclidienne des 10 opérateurs d'inertie (vecteurs propres de la matrice des coefficients RV).

TABLEAU 1 :

Quelques paramètres numériques associés à STATIS. k-numéro du tableau. Nvar – Nombre de variables du groupe. HS-Carré de la norme de l'opérateur. RV-Matrice des coefficients RV (× 1000). Poids-Poids de l'opérateur d'inertie dans la constitution du compromis. On a une vision équilibrée du rôle des groupes dans la définition du compromis par les poids, le groupe 4 jouant un grand rôle par la norme.

k	Nvar	HS	RV										Poids
1	11	0.107	1000										0.305
2	7	0.232	442	1000									0.326
3	13	0.179	527	509	1000								0.341
4	4	1.507	439	571	543	1000							0.341
5	13	0.043	498	406	463	305	1000						0.285
6	22	0.296	428	640	614	624	451	1000					0.353
7	4	0.086	502	307	433	316	473	338	1000				0.284
8	3	0.643	347	432	510	596	310	494	418	1000			0.318
9	8	0.362	407	491	424	610	496	528	594	638	1000		0.339
10	6	0.536	389	410	442	407	254	514	284	335	242	1000	0.258

La diagonalisation du compromis (valeurs propres dans la figure 6a) donne un plan de $\mathbb{R}^n$ ($n = 16$) défini par les deux premiers vecteurs propres : sur ce plan se projettent les composantes principales de rang 1 (figure 6b) et de rang 2 (figure 6c) des analyses séparées. Le nuage de variables de chaque tableau accompagné de ses deux premières composantes est projeté sur le même plan (figure 6d).

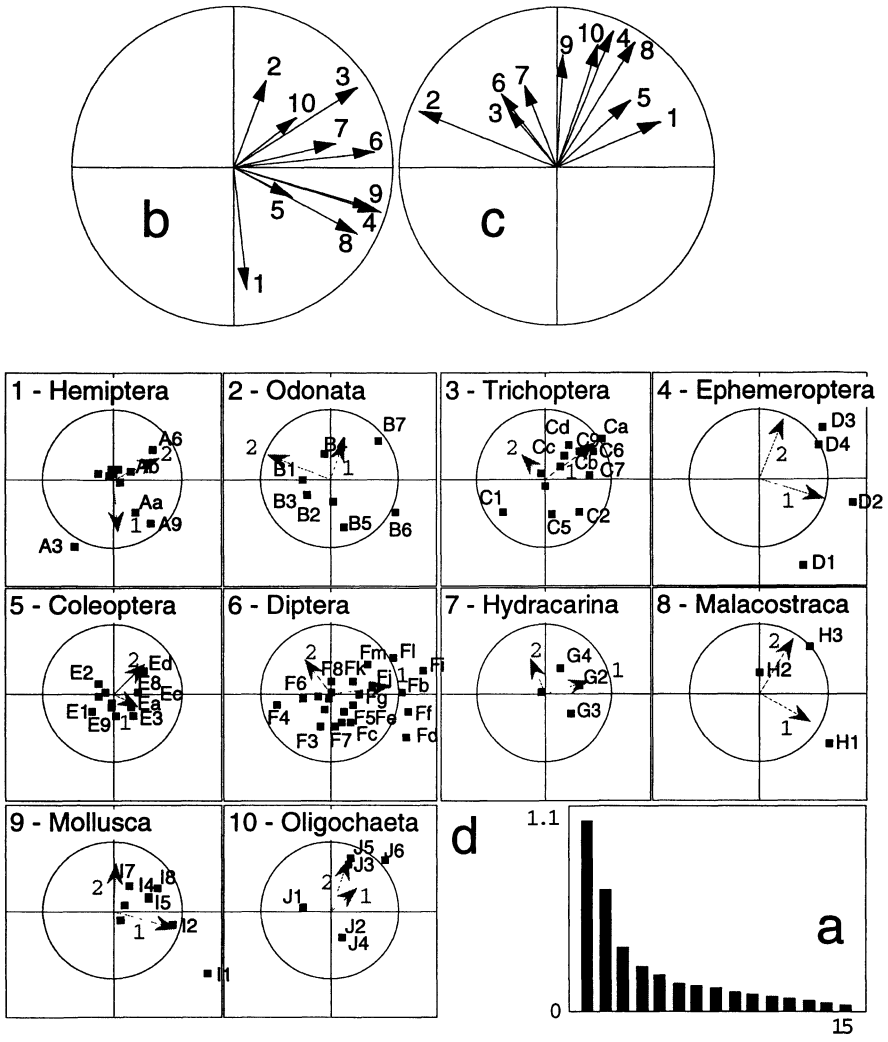

FIGURE 6 :

Analyse du compromis de STATIS. a) Valeur propre de l'opérateur compromis. b) Projection sur le plan des deux premiers vecteurs propres du compromis de 10 composantes principales de rang 1 des analyses séparées. c) Projection sur le même plan des 10 composantes principales de rang 2 des analyses séparées. d) Projection sur le même plan des nuages des variables de chaque tableau et des deux premières composantes principales de ces nuages. Les cercles de rayon unité donnent l'échelle mais ne sont utiles que pour la lecture de la projection des composantes principales qui sont des vecteurs de longueur unité. Les projections des variables respectent les variances initiales et peuvent être extérieures au cercle. Les groupes 4 et 8, bien que possédant que peu d'espèces définissent des plans 1-2 presque identique au plan du compromis.

Les composantes des vecteurs propres du compromis sont des codes numériques normés et non corrélés des stations. Ils ne correspondent à aucune approche géométrique. Leur pertinence globale ne fait pas de doute (figure 7a). La notion de trajectoires (représentation des stations par chacun des tableaux (figure 7b) n'est pas obtenue par projection ou représentation euclidienne. Sa seule propriété est que les centres de gravité des représentations du même point (exemple : figure 7c) redonne la carte globale.

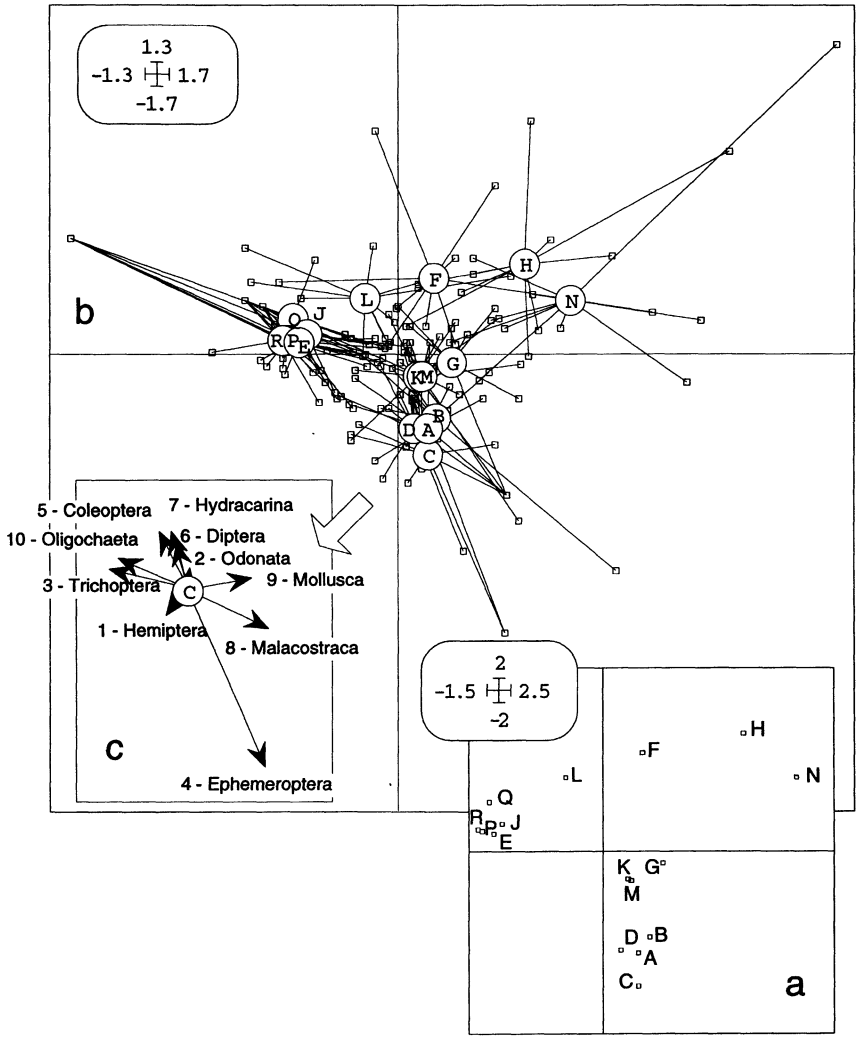

FIGURE 7 : *a) Position synthétique des stations vue par STATIS (vecteurs propres de l'analyse du compromis). b) Représentation multiple (une position par tableau) des stations par la technique des trajectoires de Lavit (1988). c) Détail de b pour la station C. La représentation simultanée est la partie faible de la méthode ACT-STATIS.*

L'absence de critère optimisé dans la représentation simultanée des lignes par tableau (ni en terme de proximité ni en terme d'éloignement) est le point faible de STATIS. La proposition de Place (1980) de positionner les individus de chaque étude par la prévision des vecteurs propres du compromis par régression multiple sur chaque tableau ne peut résoudre la question car cette prévision est parfaite pour les tableaux comptant plus de variables que d'individus, ce qui se rencontre ici.

4.3. Utilisation de l'AFMULT

L'AFMULT fait une analyse d'inertie (valeurs propres : figure 8a) dans $\mathbb{R}^n$. La base des composantes principales synthétiques permet la projection des composantes de chaque tableau de rang 1 (figure 8b) et de rang 2 (figure 8c), comme la projection de chaque nuage de variables accompagné de ses deux premières composantes (figure 8d). La cohérence avec STATIS est remarquable et le plan défini par le compromis de l'une ou l'analyse d'inertie de l'autre sont voisins.

TABLEAU 2 :

Décomposition des valeurs propres dans l'AFMULT (Total). k-numéro du tableau. A – Valeur du lien pour le facteur 1 : inertie projetée sur la première composante ramenée en pourcentage de la première valeur propre (réduite à l'unité par la pondération adoptée) de l'analyse de chaque tableau. B – Valeur du lien pour le facteur 2 : inertie projetée sur la deuxième composante ramenée en pourcentage de la première valeur propre de l'analyse (réduite à l'unité par la pondération adoptée) de chaque tableau. L'interprétation de ce paramètre est surtout pertinente pour le premier facteur car il est réellement, dans ce cas et dans ce seul cas, un rapport à l'optimum possible.

k	1	2	3	4	5	6	7	8	9	10	Total
A	0.468	0.715	0.715	0.861	0.481	0.812	0.634	0.657	0.753	0.364	6.459
B	0.786	0.617	0.377	0.369	0.421	0.242	0.475	0.489	0.333	0.478	4.586

Chaque tableau définit son lien avec une composante synthétique $\mathbf{z}$ par :

$$\mathbb{L}(\mathbf{z},\, k) = \sum_{j \in P_k} \frac{p_j}{\alpha_k} \left(\mathbf{x}^j | \mathbf{z}\right)^2_{\mathbf{D}} = \frac{1}{\alpha_k} ||\mathbf{X}_k^t \mathbf{D} \mathbf{Z}||^2_{\mathbf{Q}_k} = \frac{1}{\alpha_k} \left(\mathbf{W}_k \mathbf{D} \mathbf{z} | \mathbf{z}\right)_{\mathbf{D}}$$

où $\mathbf{x}^j$ désigne une variable, $j \in P_k$ indique que la variable $\mathbf{x}^j$, dans la numérotation totale, appartient au tableau k, p_j est le poids de la variable $\mathbf{x}^j$, α_k est la première valeur propre de l'analyse séparée du tableau k (Escofier et Pagès 1984, p.43, Escofier et Pagès 1994, p.125). Le tableau 2 donne la décomposition des valeurs propres entre les valeurs du lien qui est utilisée pour représenter les tableaux (figure 9c), dans la logique de l'ACM (Escofier 1979), conformément aux propositions et justifications théoriques des auteurs (Escofier et Pagès 1985 p.9, 1994 p.138). Le module AFM du logiciel ADE-4 a utilisé le rapport très détaillé de 1985 et l'article de 1994 pour toutes les vérifications nécessaires.

L'AFMULT est la seule des trois méthodes à avoir une double canonicité dans la représentation des éléments par tableaux. Par le biais de l'analyse inter-classe, chaque

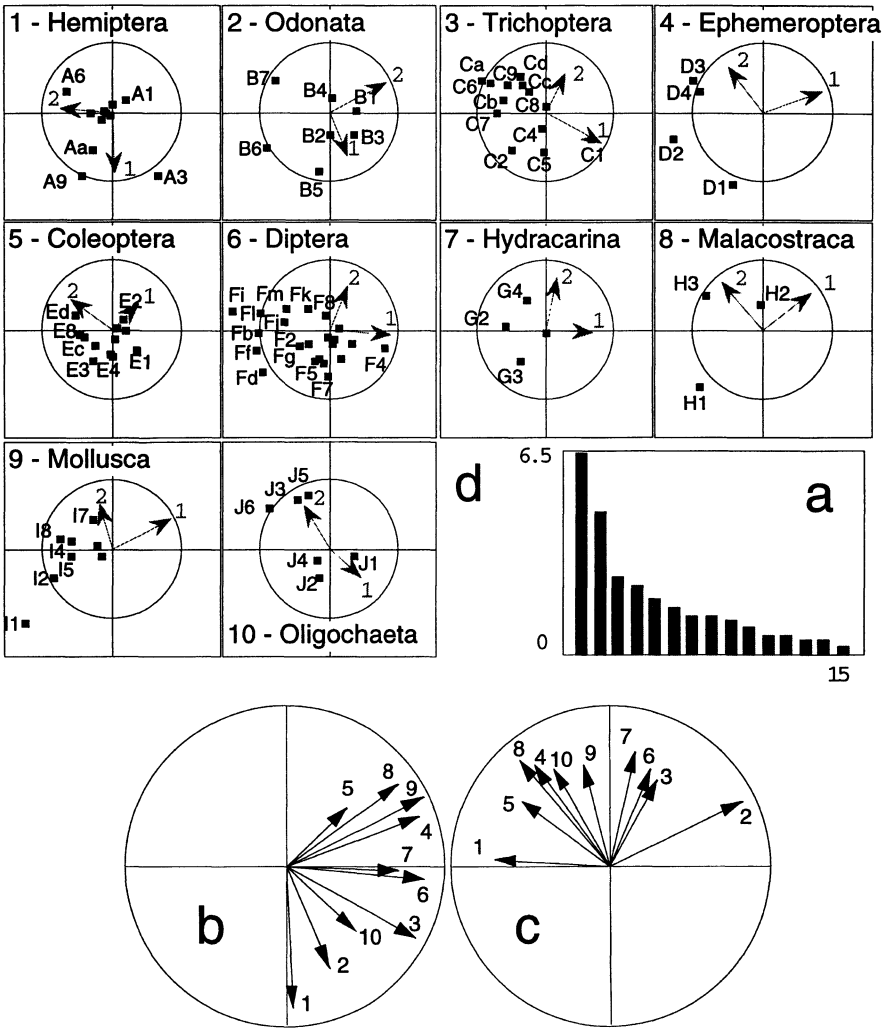

FIGURE 8 : *Analyse d'inertie pondérée de l'AFMULT. a) Valeurs propres de la somme pondérée des opérateurs d'inertie. L'équivalent dans STATIS (figure 6a) est plus net par propriété d'optimalité de STATIS (combinaison d'opérateurs optimisant la somme des carrés des valeurs propres). b) Projection sur le plan des deux premières composantes principales des 10 composantes principales de rang 1 des analyses séparées. c) Projection sur le même plan des 10 composantes principales de rang 2 des analyses séparées. d) Projection sur le même plan des nuages des variables de chaque tableau et des deux premières composantes principales de ces nuages. Les cercles de rayon unité donnent l'échelle mais ne sont utiles que pour la lecture de la projection des composantes principales qui sont des vecteurs de longueur unité. Les projections des variables respectent les variances initiales et peuvent être extérieures au cercle. Les groupes 4 et 8, montrent des plans 1-2 moins ajustés que dans le compromis de STATIS (figure 6d) alors que les tableaux avec plus de variables ont été mieux pris en compte. L'essentiel des deux approches est cependant cohérent.*

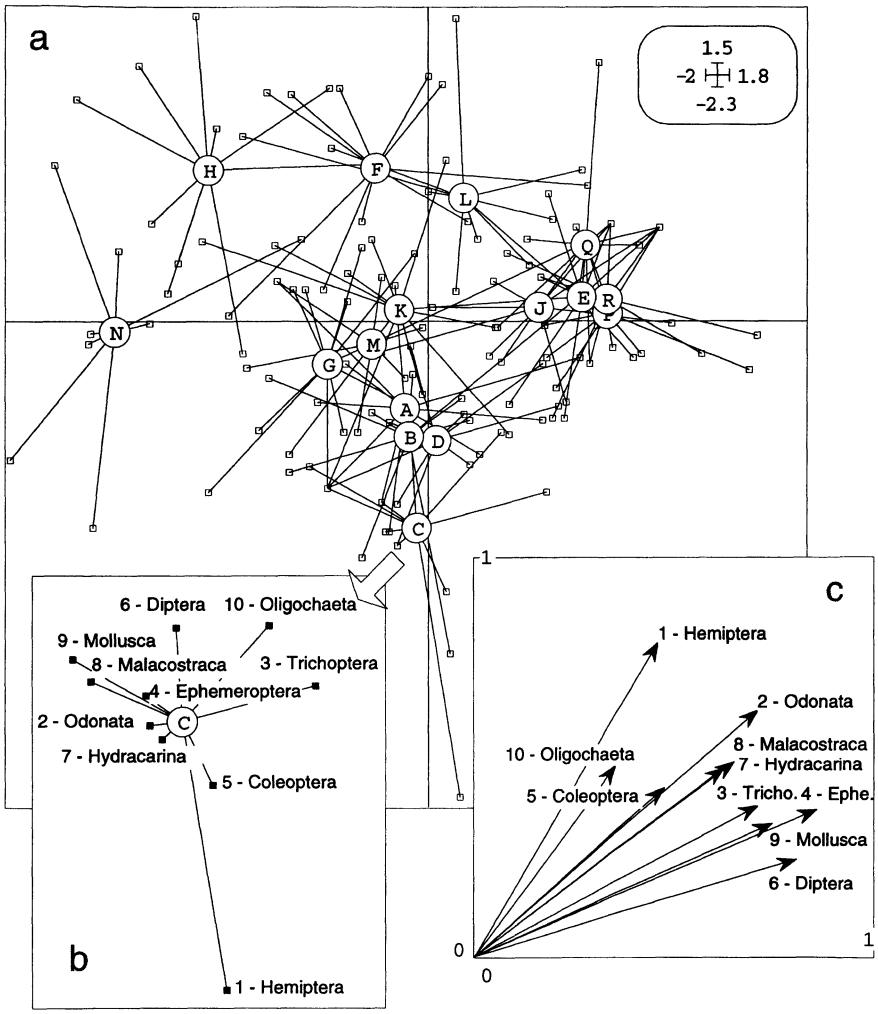

FIGURE 9 :

Analyse inter-classe duale de l'analyse d'inertie de la figure 8 assurée par l'AFMULT. a) Position synthétique des stations et représentation multiple (une position par tableau) exprimant l'optimisation de l'inertie inter-classe. b) Détail a pour la station C. La double canonicité des représentations des individus et des variables est la partie forte de l'AFMULT. c) Représentation des tableaux par les valeurs du lien avec chaque composante principale de synthèse. La pondération par l'inverse des racines carrées des premières valeurs propres joue ici un grand rôle en gommant les variations de structures entre tableaux.

station est représentée par chaque groupe de descripteur (figure 9) et le résultat est facilement interprétable.

4.4. Utilisation de l'ACOM

L'ACOM fait K analyses d'inertie coordonnées (ici K ACP) et définit dans chaque espace un système d'axes de co-inertie sur lequel on projette les nuages de points et leurs systèmes d'axe d'inertie (figure 10).

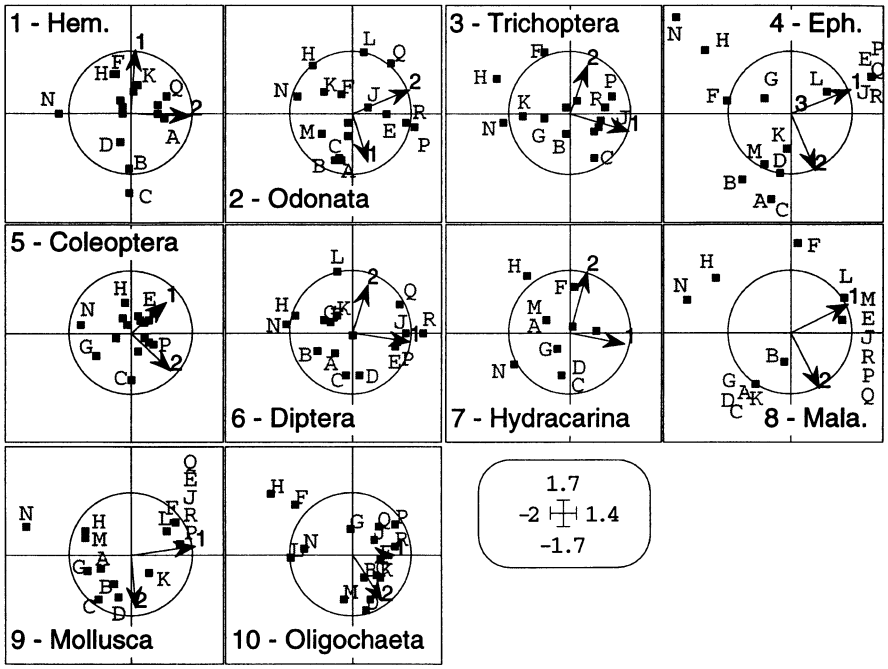

FIGURE 10 :

10 Analyses d'inertie coordonnées assurées par l'ACOM. Chaque fenêtre est un plan défini par deux axes de co-inertie dans chaque espace. Les vecteurs sont les projections des deux premiers axes principaux de chaque analyse et les carrés gris sont les projections des lignes de chaque tableau. Les cercles de rayon unité ne sont utiles que pour la lecture de la projection des axes principaux qui sont des vecteurs de longueur unité. Le groupe 10 apparaît encore une fois comme globalement non concerné par la typologie commune des stations. La variance des coordonnées reliée à l'inertie initiale de chaque tableau fait partie, contrairement à l'approche de l'AFMULT, de la valeur typologique des groupes de descripteurs.

On remarquera qu'on a jusqu'à présent représenté les projections des composantes principales des analyses séparées sur un même sous-espace compromis, alors qu'on représente ici les axes des analyses séparées sur des sous-espaces séparés. Ceci est plus précis dans chaque espace mais n'autorise pas la superposition faite dans les cas précédents. Nous n'avons pas non plus de répartition d'inertie globale mais une discussion tableau par tableau de la valeur en terme d'inertie des axes de co-inertie (tableau 3).

TABLEAU 3 :

Quelques paramètres numériques associés au facteur 1 de l'ACOM. k-numéro du tableau. Iner Max-Inertie projetée sur le premier axe d'inertie. Inertie projetée sur le premier axe de co-inertie. ProScal2-Carré de la covariance entre la coordonnée sur le premier axe de co-inertie et la première variable auxiliaire. Cos2-Carré de corrélation correspondant.

k	1	2	3	4	5	6	7	8	9	10
Iner Max	0.241	0.324	0.372	1.005	0.144	0.426	0.195	0.702	0.561	0.536
Inertie	0.165	0.272	0.343	0.959	0.098	0.400	0.177	0.636	0.552	0.377
ProScal2	0.107	0.231	0.247	0.879	0.068	0.344	0.125	0.499	0.449	0.165
Cos2	0.65	0.85	0.72	0.92	0.69	0.86	0.71	0.78	0.81	0.44

La liaison entre les typologies par tableau se fait en représentant simultanément les coordonnées normalisées par tableau et les variables auxiliaires de même rang, ce qui exprime la part corrélation dans la co-inertie (figure 11). La position donnée par la variable auxiliaire n'est pas la position moyenne mais la représentation en étoile est justifiée par l'expression de la corrélation sous la forme :

$$||\mathbf{z}^k - \mathbf{v}||^2_{\mathbf{D}} = 2 - r^2(\mathbf{z}^k, \mathbf{v})$$

La cohérence entre les trois typologies moyennes de stations souligne la vocation commune des trois méthodes qui cherchent, chacune à leur manière, la partie commune des analyses des triplets de départ.

Si STATIS est privée d'optimalité dans la représentation simultanée des individus, l'ACOM est privée d'optimalité dans la représentation simultanée des variables. Nous avons utilisé (figure 12a) la projection des nuages de variables sur le plan défini par les variables auxiliaires. On aurait pu utiliser les poids canoniques, c'est-à-dire les composantes des vecteurs $\mathbf{u}^j_k$. Le critère optimisé s'exprime par contre par la représentation des covariances entre vecteurs de coordonnées $\mathbf{X}_k\mathbf{Q}_k\mathbf{u}^j_k$ et variables auxiliaires de même rang $\mathbf{v}^j$ (figure 12 b).

5. Conclusions

Nous ferons pour conclure deux remarques. La première porte sur les relations entre les trois méthodes, la seconde sur le rôle des logiciels.

Toute comparaison entre les trois méthodes est marquée par les habitudes implicites que donne la pratique de l'analyse d'un seul tableau. La double analyse d'inertie (maximiser $||\mathbf{X}^t\mathbf{Dz}||^2_{\mathbf{Q}}$ et $||\mathbf{XQu}||^2_{\mathbf{D}}$), le théorème de reconstitution du tableau (minimiser $||\mathbf{X} - \sigma\mathbf{z}\mathbf{u}^t||^2_{HS}$) et de ses deux opérateurs, la relation de dualité (maximiser $(\mathbf{XQu}|\mathbf{z})_{\mathbf{D}}$ ou $(\mathbf{X}^t\mathbf{Dz}|\mathbf{u})_{\mathbf{Q}}$ conduisent à un résultat unique. Dans un K-tableaux, la généralisation de l'un des critères se fait aux dépends des autres. Il est tentant de compléter la partie canonique de chaque méthode par des éléments empruntés aux autres.

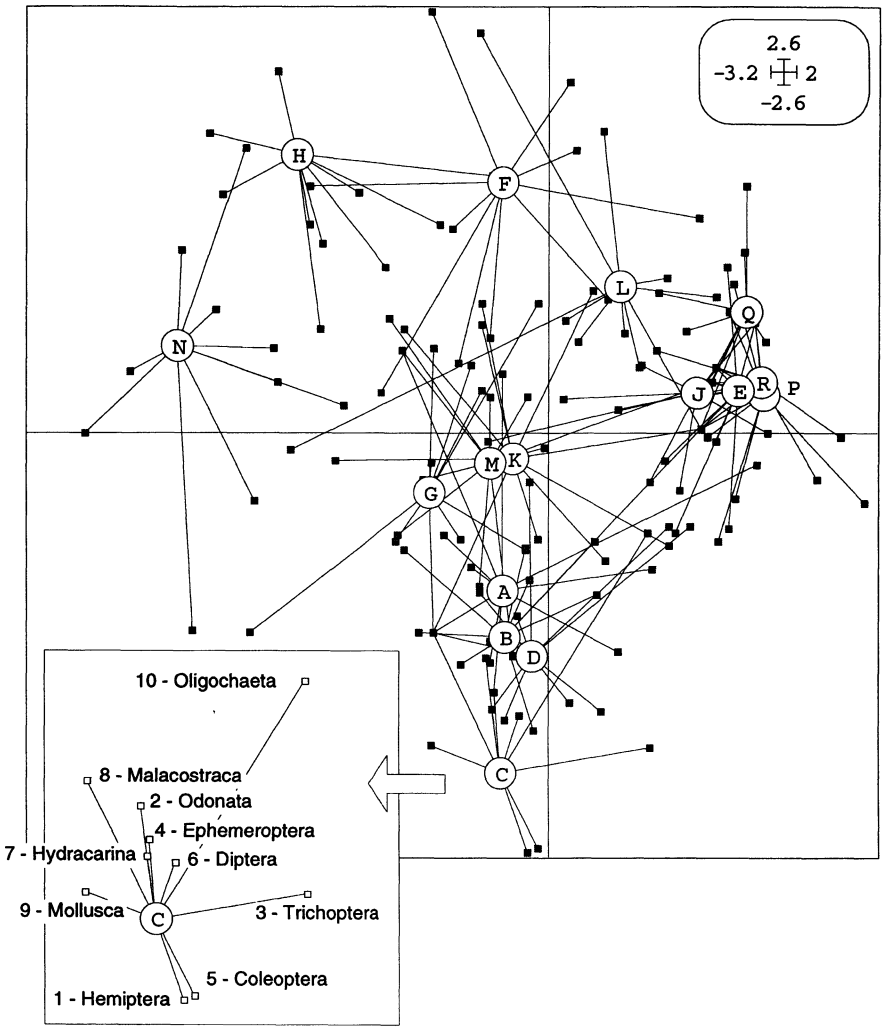

FIGURE 11 :

Représentation simultanée des stations comme expression de la corrélation entre coordonnées des projections sur un axe de co-inertie et la variable auxiliaire de même rang. Chaque station est représentée une fois pour chaque tableau (les coordonnées de variance unité définies par $\mathbf{X}_k\mathbf{Q}\mathbf{u}_k^j / ||\mathbf{X}_k\mathbf{Q}_k\mathbf{u}_k^j||$ *positionnent les petits carrés) et une nouvelle fois (cercles blancs) par les variables auxiliaires qui sont normées. Le résultat est très sensiblement celui de l'AFMULT (figure 9), les deux l'emportant sur celui de STATIS (Figure 7).*

Le compromis de STATIS conduit à la meilleur carte de synthèse des lignes des tableaux. La carte des colonnes de l'AFMULT est la meilleur approche simultanée

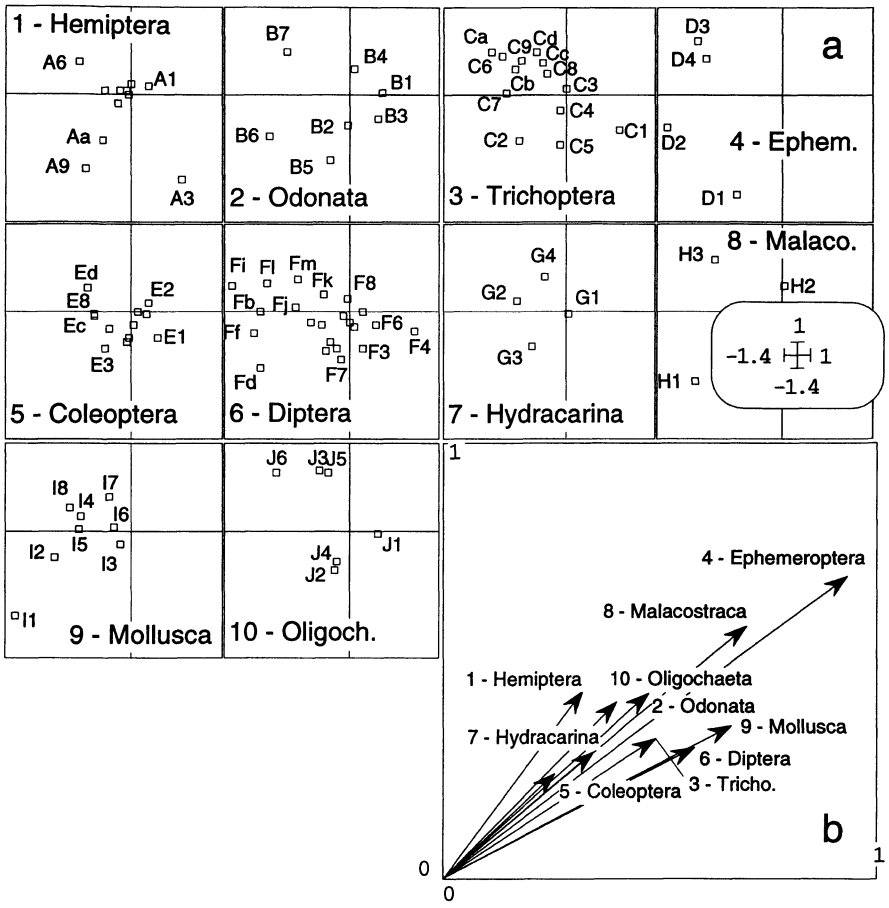

FIGURE 12 :

Utilisation des variables auxiliaires dans l'ACOM. a) Projection des nuages de variables par tableau sur le plan des variables auxiliaires de rang 1 et 2. La représentation n'a pas de caractère d'optimalité propre tout en étant très voisine de celle de l'AFMULT (Figure 8d). b) Représentation des tableaux par les valeurs du carré de la covariance entre chaque système de coordonnées (Figure 10) et variable auxiliaire de même rang. La covariance, liée à la variance (qui dépend de l'inertie initiale de chaque tableau) et à la corrélation, exprime la valeur typologique des groupes de descripteurs. Les groupes 4 et 8 l'emportent.

des variables. Les cartes séparés de l'ACOM donne la meilleure idée des variations de la capacité typologique des groupes. Mais une seule méthode ne peut faire cela simultanément. Aucune des trois n'assure la reconstitution simultanée des tableaux. STATIS ouvre d'autre part la possibilité de faire de la typologie de structures, ce qui n'est pas le cas des deux autres. STATIS comme l'ACOM s'étendent aux tableaux ayant les mêmes variables (et non les mêmes individus) ce qui n'a pas été

envisagé pour l'AFMULT. Cette dernière reste cependant la plus accessible pour une première approche des K-tableaux par un utilisateur non professionnel de l'analyse des données.

Les logiques numériques différentes conduisent l'interprétation de manières séparées. Par contre la logique graphique permet assez aisément d'identifier l'accord sur la recherche des typologies communes. Ces méthodes voient leur efficacité sensiblement accrue par l'usage d'outils graphiques appropriés (Thioulouse *et Coll.* 1991). Le logiciel ADE-4 (Thioulouse *et Coll.* 1995), fusion des programmes ADE (Chessel et Dolédec , 1993), GraphMu et MacMul (Thioulouse 1989, 1990) réunit les fonctions graphiques multifenêtrées et les calculs des trois analyses K-tableaux discutées ici. Il doit rendre l'usage de ces méthodes accessibles à un public large.

Logiciel

On trouvera dans ADE-4, outre les utilitaires de manipulation de données, les analyses à un tableau, les ordinations sous contraintes spatiales, les analyses discriminantes linéaires, les régressions multiple et PLS, les ACPVI et les analyses de co-inertie à deux tableaux. Une interface HyperCard contient les jeux de données de la documentation et un millier de références bibliographiques en analyse des données. Cet ensemble pour micro-ordinateurs de gamme Apple est librement diffusé sur Internet sur l'URL :

ftp ://biom3.univ-lyon1.fr/pub/mac/ADE/ADE4

ou le serveur WWW :

http ://biomserv.univ-lyon1.fr/ADE-4.html

La diffusion des disquettes et de la documentation (700p.) est assurée par J.M. Olivier, URA 1974, Bât 401C, Université Lyon 1, 69622 Villeurbanne Cédex.

Remerciements

Nous remercions vivement Mr. P. Cazes pour ses précieux conseils.

Références

BOVE, G. et DI CIACCIO, A. (1994) A user-oriented overview of multiway methods and software. *Computational Statistics and Data Analysis*, 18, 15-37.

CAILLIEZ, F. et PAGÈS, J.P. (1976) *Introduction à l'analyse des données*. SMASH, 9 rue Duban, 75016 Paris. 1-616.

CARROL, J.D. (1968) A generalization of canonical correlation analysis to three or more sets of variables. *Proceeding of the 76th Convention of the American Psychological Association*, 3, 227-228.

CASIN, Ph. (1995) Une méthode de comparaison de tableaux : l'Analyse Discriminante de Tableaux. In : *XXVIIe Journées de Statistique*, Jouy-en-Josas, 15-19 mai 1995. Groupe HEC, 1 rue de la libération, 78351 Jouy-en-Josas cedex, France. 160-163.

CASIN, Ph., TURLOT, J.C. (1986) Une présentation de l'analyse canonique généralisée dans l'espace des individus. *Revue de Statistiques Appliquée*, 35, 65-75.

CAZES, P. (1980) L'analyse de certains tableaux rectangulaires décomposé en blocs : généralisation des propriétés rencontrées dans l'étude des correspondances multiples. I. Définitions et applications à l'analyse canonique des variables qualitatives. II Questionnaires : variantes des codages et nouveaux calculs de contributions. *Les Cahiers de l'Analyse des Données*, 5, 145-161 et 387-406.

CHESSEL, D. et MERCIER, P. (1993) Couplage de triplets statistiques et liaisons espèces-environement. In : *Biométrie et Environnement*. LEBRETON, J.D. et ASSELAIN, B. (Eds.) Masson, Paris. 15-44.

CHESSEL, D. et DOLÉDEC, S. (1993) *ADE Version 3.6 : HyperCard ©Stacks and Programme library for the Analysis of Environmental Data*. Manuel d'utilisation. 8 fascicules. URA CNRS 1451, Université Lyon 1, 69622 Villeurbanne cedex. 750 p.

CHEVENET, F., DOLÉDEC, S. et CHESSEL, D. (1994) A fuzzy coding approach for the analysis of long-term ecological data. *Freshwater Biology*, 31, 295-309.

DOLÉDEC, S. et CHESSEL, D. (1994) Co-inertia analysis : an alternative method for studying species-environment relationships. *Freshwater Biology*, 31, 277-294.

ESCOFIER , B. (1979) Une représentation des variables dans l'analyse des correspondances multiples. *Revue de Statistique Appliquée*, 27, 4, 37-47.

ESCOFIER, B. et PAGÈS J. (1984) L'analyse factorielle multiple : une méthode de comparaison de groupes de variables. In : *Data Analysis and Informatics III*. DIDAY, E. *et Coll.* (Eds.) Elsevier, North-Holland. 41-55.

ESCOFIER, B. et PAGÈS, J. (1985) Mise en œuvre de l'analyse factorielle multiple pour les tableaux numériques qualitatifs ou mixtes. *Rapport de recherche n°429*. INRIA, Domaine de Voluceau-Rocquencourt, BP 105, 78153 Le chesnay cedex, France. 1-54 + annexes.

ESCOFIER, B. et PAGÈS, J. (1986) Le traitement des variables qualitatives et des tableaux mixtes par analyse factorielle multiple. In : *Data Analysis and Informatics* IV. Diday, E. *et Coll.* (Eds.) Elsevier, North-Holland. 179-191.

ESCOFIER, B. et PAGÈS, J. (1994) Multiple factor analysis (AFMULT package). *Computational Statistics and Data Analysis*, 18, 121-140.

ESCOUFIER, Y. (1973) Le traitement des variables vectorielles. *Biometrics*, 29, 750-760.

ESCOUFIER, Y. (1977) Operators related to a data matrix. In : *Recent developments in Statistics*. BARRA, J.R. *et Coll.* (Eds.), North-Holland, 125-131.

ESCOUFIER, Y. (1987) The duality diagramm : a means of better practical applications. In : *Development in numerical ecology*. LEGENDRE, P. et LEGENDRE, L. (Eds.) NATO advanced Institute , Serie G, Springer Verlag, Berlin. 139-156.

FRIDAY, L.E. (1987) The diversity of macroinvertebrate and macrophyte communities in ponds. *Freshwater Biology*, 18, 87-104.

GÉNARD, M., SOUTY, M., HOLMES, S., REICH, M. et BREUILS, L. (1994) Correlation among quality parameters of peach fruit. *Journal of the science of food and agriculture*, 66, 241-245.

HOTELLING, H. (1936) Relations between two sets of variates. *Biometrika*, 28, 321-377.

KAZI-AOUAL, F., HITIER, S., SABATIER, R. et LEBRETON, J.D. (1995) Refined approximations to permutation tests for multivariate inference. *Computational Statistics and Data Analysis*, in press.

KETTERING, R.J. (1971) Canonical analysis of several sets of variables. *Biometrika*, 58, 433-451.

KIERS, H.A.L, CLÉROUX, R. et TEN BERGE, M.F. (1994) Generalized analysis based on optimizing matrix correlations and a relation with IDIOSCAL. *Computational Statistics and Data Analysis*, 18, 331-340.

LAVIT, Ch. (1988) Analyse conjointe de tableaux quantitatifs. Masson, Paris. 1-240.

LAVIT, Ch., ESCOUFIER, Y., SABATIER, R. et TRAISSAC, P. (1994) The ACT (Statis method). *Computational Statistics and Data Analysis*, 18, 97-119.

LAZRAQ, A., CLÉROUX, R. et KIERS, H.A.L. (1992) Mesures de liaison vectorielle et généralisation de l'analyse canonique. *Revue de Statistique Appliquée*, 39, 23-35.

MERCIER P., CHESSEL , D. et DOLÉDEC S. (1992) Complete correspondence analysis of an ecological profile data table : a central ordination method. *Acta Œcologica*, 13, 25-44.

PAGÈS , J. (1995) Eléments de comparaison de l'Analyse Factorielle Multiple et de la méthode STATIS. In : *XXVIIe Journées de Statistique*, Jouy-en-Josas, 15-19 mai 1995. Groupe HEC, 1 rue de la libération, 78351 Jouy-en-Josas cedex, France. 492-496.

PLACE, M.C. (1980) *Contribution algorithmique à la mise en œuvre de la méthode STATIS*. Thèse de troisième cycle, Université de Montellier II.

PRODON, R. et LEBRETON, J.D. (1994) Analyses multivariées des relations espèces-milieu : structure et interprétation écologique. *Vie Milieu*, 44, 69-91.

RIZZI, A. et VICHI, M. (1995) Representation, synthesis, variability and data preprocessing of a three-way data set. *Computational Statistics and Data Analysis*, 19, 203-222.

SABATIER, R. (1993) Critères et contraintes pour l'ordination simultanée de K tableaux. In : *Biométrie et Environnement*. LEBRETON, J.D. et ASSELAIN , B. (Eds.) Masson, Paris. 101-121.

SAPORTA, G. (1975) Liaisons entre plusieurs ensembles de variables et codage de données qualitatives. Thèse de 3°cycle, Université Pierre et Marie Curie, Paris VI. 1-102.

TENENHAUS, M. (1977) Analyse en composantes principales d'un ensemble de variables nominales ou numériques. *Revue de Statistique Appliquée*, 25, 39-56.

TENENHAUS, M. (1984) L'analyse canonique généralisée de variables numériques, nominales ou ordinales par des méthodes de codage optimal. In : *Data Analysis and Informatics*, III. DIDAY, E. et Coll. (Eds.) Elsevier Science Publishers B.V., North-Holland. 71-84.

TENENHAUS, M. et YOUNG, F.W. (1985) An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis ans other methods for quantifying categorical multivariate data. *Psychometrika*, 50, 1, 91-119.

TENENHAUS, M., GAUCHI, J.P. et MÉNARDO, C. (1995) Régression PLS et applications. *Revue de Statistique Appliquée*, 43, 7-63.

THIOULOUSE, J. (1989) Statistical analysis and graphical display of multivariate data on the MacIntosh. *Computer Applications in the BIOSciences* : 5, 4, 287-292.

THIOULOUSE, J. (1990) MacMul and GraphMu : two Macintosh programms for the display and analysis of multivariate data. *Computers and Geosciences* : 8, 1235-1240.

THIOULOUSE, J., DEVILLERS, J., CHESSEL , D. et AUDA, Y. (1991) Graphical techniques for multidimensional data analysis. In : *Applied Multivariate Analysis in SAR and Environmental Studies*. DEVILLERS, J. et KARCHER, W. (Eds.) Kluwer Academic Publishers. 153-205.

THIOULOUSE, J., DOLÉDEC, S., CHESSEL , D. et OLIVIER, J.M. (1995) ADE Software : multivariate analysis and graphical display of environmental data. In : *Software per l'ambiante*. GUARISO, G. et RIZZOLI, A. (Eds.) Pàtron Editore, Bologna. 57-62.

TUCKER, L.R. . (1958) An inter-battery method of factor analysis. *Psychometrika*, 23, 111- 136.

VAN DE GEER, J.P. (1984) Linear relations among K sets of variables. *Psychometrika*, 49, 79-94.