REVUE DE STATISTIQUE APPLIQUÉE

P. CAZES
D. CHESSEL
S. DOLEDEC

L'analyse des correspondances internes d'un tableau partitionné : son usage en hydrobiologie

Revue de statistique appliquée, tome 36, nº 1 (1988), p. 39-54
<http://www.numdam.org/item?id=RSA_1988__36_1_39_0>

© Société française de statistique, 1988, tous droits réservés.

L'accès aux archives de la revue « Revue de statistique appliquée » (http://www.sfds.asso.fr/publicat/rsa.htm) implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/conditions). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM
Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
http://www.numdam.org/

Rev. Statistique Appliquée, 1988, **XXXVI** (1), 39-54

L'ANALYSE DES CORRESPONDANCES INTERNES D'UN TABLEAU PARTITIONNÉ : SON USAGE EN HYDROBIOLOGIE

P. CAZES (1), D. CHESSEL (2) et S. DOLEDEC (2)

(1) Université Paris Dauphine, Place du Mal de Lattre de Tassigny, 75775, Paris Cedex 16.
(2) Ecologie des eaux douces, UA 367, Université Lyon 1, 69622 Villeurbanne Cedex.

RÉSUMÉ

Un tableau d'analyse des correspondances à I lignes et J colonnes est partitionné en L blocs lignes et K blocs colonnes. On ne veut tenir compte dans sa description que des structures internes des sous-tableaux définis par les deux partitions. La note introduit dans le schéma de dualité de l'analyse des correspondances deux projecteurs et précise les aides à l'interprétation qui en découlent. Les illustrations concernent l'analyse d'un tableau faunistique décomposé par groupes taxonomiques et par saison d'échantillonnage. Lorsque les structures observées du point de vue choisi sont faibles, par rapport aux structures globales définies par l'analyse des correspondances, l'analyse des correspondances internes donne des résultats radicalement différents.

Mots clés : Schéma de dualité, Projecteurs, Analyse des Correspondances, Hydrobiologie.

ABSTRACT

A table with I lines and J columns, to be processed by a dual scaling, was divided into L line-units and K column-units. To describe the table, only the inner structure of subtables such defined is taken into account. This paper introduces two projection operators in the duality diagram of Correspondence Analysis and provides interpretative assistance thereof. The analysis of a faunal table, divided by taxonomic groups and sampling seasons, is illustrated. If the structures observed from a choosen point of view are weak, with regard to the overall structure defined by Correspondence Analysis, the Internal Correspondence Analysis gives radically different results.

Keywords : Duality diagramm, Projection operators, Correspondence Analysis, Hydrobiology.

I. Introduction

Il est habituel de décrire les structures d'un tableau relevés-taxons par l'analyse factorielle des correspondances (AFC), que l'on y cherche des associations d'espèces et des groupes de relevés (ROUX et ROUX, 1967) ou que l'on désire y voir une ordination réciproque sur un gradient écologique (HILL 1973, 1974). Introduite en hydrobiologie par LEVEQUE et GABORIT (1972), l'AFC y est devenu un outil de base comme dans les autres secteurs de l'écologie. Dans l'étude des hydrosystèmes cependant deux questions particulières sont soulevées par la description des tableaux faunistiques.

La première touche aux difficultés des déterminations. Les organismes capturés dans les relevés ne peuvent souvent être identifiés à l'espèce que par des spécialistes de divers groupes faunistiques. Une partie des animaux restent identifiés au genre, ou à la famille, voire simplement à l'ordre. Le coût des tris et des identifications est tel qu'on s'interroge en permanence sur **la valeur typologique des groupes de taxons :** il serait souhaitable dans l'analyse des tableaux de ne tenir compte que des correspondances intervenant à l'intérieur des différents groupes d'espèces.

Le second problème a trait aux rythmes de reproduction très diversifiés en œuvre dans les populations naturelles formant les communautés interspécifiques étudiées. Les expérimentateurs échantillonnent à plusieurs reprises dans l'année ces communautés pour assurer la capture des individus de chaque espèce indépendamment des rythmes de ponte, d'émergence, de déplacement,... On voudrait alors mesurer **la valeur typologique d'une date d'échantillonnage,** question symétrique du problème précédent. A quel moment et pour quel groupe obtient-on une capacité optimale à dissocier les stations sur leur contenu faunistique ? La typologie est-elle stable ? Posée à plusieurs reprises dans la littérature écologique, par exemple par DAKKI (1985, 1986), la question de la prise en compte de ces structures expérimentales dans l'analyse du tableau n'a pas reçu de solutions pratiques. L'objectif est globalement de comparer les structures internes des sous-tableaux définis par un groupe faunistique et une date d'échantillonnage, en quelque sorte d'effectuer simultanément les AFC des sous-tableaux et de mesurer quelle part chacun d'entre eux est capable de restituer de la structure définie en commun. La note propose une solution à ce problème. Nous présenterons d'abord un jeu de données de taille limitée pour illustrer la méthode, puis une approche théorique exprimée dans le cadre du schéma de dualité (CAILLIEZ et PAGES, 1976) et une procédure de calcul limitant le travail de programmation à la construction d'un tableau sur lequel on effectue l'AFC; nous donnons ensuite les propriétés de la méthode qui est une AFC sous contraintes et les perspectives ouvertes par cette pratique.

L'exemple est présenté dans la figure 1 : il concerne 35 prélèvements de macro-invertébrés benthiques réalisés au filet de type Surber (surface échantillonnée de 0,1 m^2). On a conservé parmi les organismes capturés 43 espèces réparties en 4 groupes taxonomiques. Les 35 prélèvements se répartissent sur 6 stations notées A, B, C, D, E, F et 6 dates d'échantillonnage numérotées $1, 2, \ldots, 6$; le relevé de la station F à la date 1 est manquant. L'ensemble de l'expérience dont sont extraites ces données est présenté dans la thèse de DOLEDEC (1986). La figure 1 contient le tableau 43-35 divisé en 24 sous-tableaux par la répartition des espèces en 4 groupes et celle des relevés en 6 campagnes d'échantillonnage. La figure 2 résume les codifications des objets biologiques manipulés.

Soulignons qu'on désire éliminer les structures faunistiques associées au temps, l'objet étant d'interpréter une **typologie stationnelle** et qu'on souhaite mesurer la capacité des groupes à servir d'**indicateurs biologiques.**

	DATE 1					DATE 2						DATE 3						DATE 4						DATE 5						DATE 6					
	STATIONS																																		
	A	B	C	D	E	A	B	C	D	E	F	A	B	C	D	E	F	A	B	C	D	E	F	A	B	C	D	E	F	A	B	C	D	E	F
1	9	9	6	6	9	7	3	6	6	8	5	11	8	9	10	10	8	6	8	8	11	7	4	9	8	8	8	7	7	1	5	4	.	4	6
2	7	4	6	6	8	8	4	7	7	8	1	4	1	2	3	1	1	9	8	9	5	8	3	6	5	7	6	5	5	9	9	9	9	10	9
3	6	2	5	4	7	5	.	1	5	7	.	4	.	.	.	2	1	4	3	3	2	6	3	5	4	2	2	2	5	7	5	6	6	5	2
4	.	.	1	1	1	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	2	1	1	10	10	9	9	9	8	8	6	9	9	7	2
5	5	.	5	5	4	.	.	.	3	2	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	7	5	9	10	10	8
6	.	.	2	.	3	4	.	.	5	5	.	.	1	2	4	2	1	6	6	6	5	6	2	3	1	6	2	4	3	3	3	.	6	1	.
7	6	7	.	4	5	1	1	.	.	6	7	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	6	4	.	1	1
8	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	8	6	7	1	6	6	8	.	7	4	3	.
9	1	1	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.
10	.	3	.	.	4	.	3	.	.	5	1	4	.	.	.	.	.	.	.	1	.	.	.	2	2	1	.	.	.	2	6	5	3	5	7
11	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	6	.	.	4	3	.	.	.	.	.	.
12	3	9	7	3	6	1	6	5	5	8	1	1	.	.	.	.	.	1	.	1	.	.	.	7	7	5	.	4	4	7	10	7	6	4	8
13	3	7	5	4	5	.	2	1	5	5	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	5	5	5	4	4	5	5	2	3
14	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	5	3	3	.	3	3	4	4	.	.	3	.	.	.	.	.	.	.
15	4	5	3	.	6	7	5	4	.	6	.	3	.	.	2	.	.	.	4	4	5	3	.	7	5	5	.	3	4	6	5	7	4	6	.
16	8	9	8	5	8	9	8	6	9	9	5	6	.	1	4	5	4	4	7	5	8	4	2	7	8	6	7	5	6	5	5	6	6	7	7
17	9	9	5	9	8	9	9	6	10	9	8	7	1	5	7	5	4	5	9	6	9	5	5	7	8	5	7	5	5	6	8	5	8	4	6
18	9	7	4	6	5	9	7	4	4	6	7	4	.	4	4	4	.	4	6	8	6	1	.	6	8	5	5	.	2	6	5	6	7	4	5
19	5	3	.	.	4	4	1	.	.	3	.	1	.	.	.	.	1	.	3	2	1	1	.	2	3	1	1	2	.	3	3	4	2	4	1
20	8	4	2	9	6	8	5	4	9	7	7	8	3	4	7	4	6	5	4	6	8	4	3	6	4	6	7	5	4	.	.	2	4	.	.
21	7	7	5	4	6	8	5	4	8	7	.	4	1	1	4	1	1	2	.	3	6	.	.	3	3	3	6	1	1	7	7	7	7	7	5
22	7	.	.	6	3	8	2	1	5	6	1	1	.	1	3	.	.	.	.	3	7	.	.	3	.	.	6	.	2	5	1	7	7	.	2
23	1	.	.	4	2	1	1	.	.	1	.	.	.	.	.	.	.	.	1	3	.	.	.	.	.	2	2	.	.	.	.	2	.	.	.
24	2	2	.	.	1	1	5	1	1	.	.	3	.	.	5	.	.	.	.	.	1	.	.	1	3	.	.	.	.	.	2	.	.	5	.
25	5	.	.	4	.	5	.	.	3	1	.	1	.	1	1	.	.	.	.	1	5	.	.	1	.	1	3	1	1	2	.	.	4	.	.
26	3	1	.	4	.	5	.	.	.	1	.	3	.	.	3	.	.	.	.	1	3	.	.	.	.	.	.	.	.	2	.	2	4	1	.
27	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	1	1	.	.	3	.	.	1	.	.	1	.	.	.	1	.	.	1	.
28	8	3	6	7	8	10	1	3	8	8	2	8	4	5	7	8	6	2	2	5	6	5	2	5	1	4	5	4	4	7	4	8	8	5	6
29	9	5	8	7	8	8	.	2	6	6	1	8	.	3	5	5	4	4	.	2	7	3	1	6	4	5	7	5	5	4	4	6	6	5	3
30	8	1	5	3	4	9	1	1	5	.	.	6	.	2	5	3	.	2	.	2	5	.	.	5	3	4	6	3	.	7	1	3	7	4	.
31	.	2	1	.	.	3	1	.	.	2	.	3	1	2	2	.	.	.	.	3	7	.	.	5	.	4	7	1	1	.	.	.	2	.	.
32	6	3	4	4	6	6	.	5	4	.	.	6	.	.	4	2	.	2	2	4	5	1	.	5	3	3	4	1	.	5	4	6	6	2	2
33	4	7	1	6	5	2	8	1	.	4	.	2	1	3	3	3	.	.	5	.	4	.	.	1	4	.	3	.	.	.	.	.	.	.	.
34	4	5	2	3	4	1	1	1	2	5	1	.	.	.	.	.	.	.	.	.	.	.	.	1	4	1	.	.	1	.	3	3	.	.	1
35	.	6	.	2	.	1	1	.	.	1	.	.	.	1	.	.	.	1	1	3	5	1	.	.	3	.	5	1	.	.	.	.	.	.	.
36	2	.	3	4	4	3	1	4	.	3	2	2	.	.	2	4	.	.	.	.	7	.	.	1	.	.	5	.	.	.	.	.	.	.	.
37	.	2	.	.	8	3	.	.	5	7	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	3	.	.	1	.	.	1	.	1	.
38	1	.	.	.	3	.	.	.	4	.	4	2	.	.	2	.	.	.	.	.	6	.	.	1	1	1	6	.	.	.	.	.	.	.	3
39	2	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	1	1	3	4	3	5	.	1	2	4	1	3	1	.	.	.	.	.	.
40	.	.	.	.	4	.	.	.	1	2	.	3	2	.	.	.	.	3	.	.	2	1	1	.	.	.	.	1	.	4	.	.	2	.	.
41	.	.	.	1	4	.	.	.	.	4	.	.	.	.	.	3	.	.	.	.	5	.	.	.	.	.	4	.	.	.	.	.	.	.	.
42	1	.	1	2	5	6	.	.	.	5	.	2	.	1	2	1	2	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.
43	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	3	.	3	.	.	.	1	1	1	.	.

FIGURE 1

Tableau d'AFC partitionné. 43 lignes (taxons) sont réparties en 4 groupes faunistiques. 35 colonnes (relevés) sont réparties en 6 groupes (dates d'échantillonnage). Les lettres A,...,F désignent les stations. On cherche à ne tenir compte que des structures internes de chaque sous-tableau.

II. Analyse des correspondances et projecteurs

II.1. Rappels sur l'analyse des correspondances

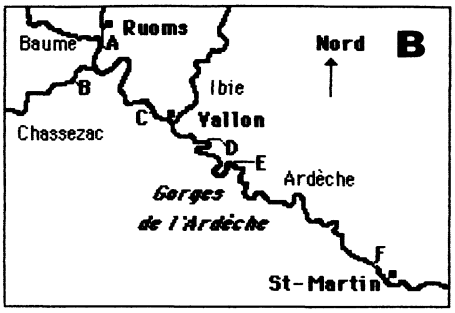
La revue bibliographique de BUYSE (1983), les ouvrages de NISHISATO (1980) et GREENACRE (1984) et la synthèse de référence de TENENHAUS et YOUNG (1985) montrent à la fois la multiplicité des approches possibles de l'AFC et la prise en compte de cette diversité dans le schéma de dualité. Nous utiliserons les notations les plus habituelles. T est un tableau de nombres positifs ou nuls, défini sur le produit de deux ensembles I et J, n_{ij} son terme générique, $n_{i.}$ et $n_{.j}$ les sommes marginales, n la somme de tous les éléments du tableau, p_{ij} la quantité n_{ij}/n, $p_{i.}$ et $p_{.j}$ les fréquences marginales, P le tableau des p_{ij}, D_I et D_J les matrices diagonales contenant respectivement les $p_{i.}$ et les $p_{.j}$.

Groupe 1 : Ephemeropteres
1 *Caenis luctuosa*
2 *Baetis fuscatus*
3 *Ecdyonurus spp.*
4 *Ephemerella ignita*
5 *Ephoron virgo*
6 *Baetis lutheri*
7 *Procloeon bifidum*
8 *Oligoneuriella rhenana*
9 *Centroptilum pennulatum*
A *Choroterpes picteti*
B *Rhithrogena spp.*

C

Groupe 2 . Plécoptères
1 *Leuctra fusca*
2 *Euleuctra geniculata*
3 *Nemoura spp*

Groupe 3 : Coléoptères
1 *Esolus parallelepipedus* (a)
2 *Esolus pygmaeus* (a)
3 *Esolus spp.* (l)
4 *Limnius opacus* (l)
5 *Limnius opacus* (a)
6 *Oulimnius troglodytes* (l)
7 *Oulimnius troglodytes* (a)
8 *Stenelmis canaliculata* (l)
9 *Stenelmis canaliculata* (a)
A *Dryops spp*
B *Limnius volkmari*
C *Elmis maugetii*
D *Hydroscapha granulum*



Groupe 4 : Trichoptères
1 *Hydropsyche exocellata*
2 *Cheumatopsyche lepida*
3 *Psychomyia pusilla*
4 *Orthotrichia spp.*
5 *Hydropsyche pellucidula*
6 *Setodes argentipunctellus*
7 *Polycentropus flavomaculatus*
8 *Athripsodes spp.*
9 *Ceraclea annulicornis*
A *Chimarra marginata*
B *Ceraclea dissimilis*
C *Rhyacophila dorsalis*
D *Hydropsyche siltalai*
E *Oecetis spp.*
F *Ecnomus spp*
G *Hydroptila spp.*

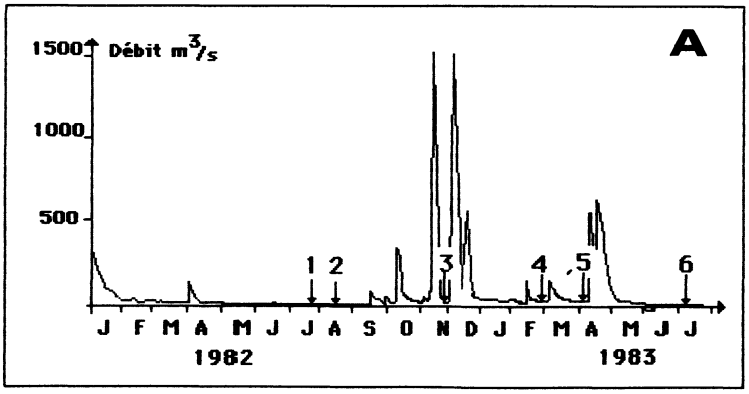


FIGURE 2
Codes des éléments du tableau traité. A : Dates d'échantillonnage situées sur la chronique du débit de la rivière. B : Stations d'échantillonnage. C : Code des taxons répartis par groupes faunistiques.

Pour éviter toute ambiguïté on peut appeler « AFC-PROGRAMME » la procédure.

1) Diagonaliser $D_J^{-1/2} P' D_I^{-1} P D_J^{-1/2}$.

2) Eliminer la plus grande valeur propre $\mu_0 = 1$ et conserver $\mu_1, \mu_2, \ldots, \mu_f$ dans la matrice diagonale S_f.

3) Eliminer le premier vecteur propre et conserver $u_1, u_2, \ldots, u_f$, vecteurs propres normés, en colonnes dans la matrice U_f.

4) Calculer $C_f = D_J^{-1/2} U_f S_f^{1/2}$ et $L_f = D_I^{-1} P C_f S_f^{-1/2}$ comptant respectivement J lignes et f colonnes (coordonnées factorielles des colonnes) et I lignes et f colonnes (coordonnées factorielles des lignes).

Ces calculs sont exécutés dans la plupart des programmes d'analyse des correspondances et ils donnent des coordonnées factorielles de variances égales aux valeurs propres. Ils ne préjugent pas de l'emploi qui en sera fait suivant le problème traité. En particulier on peut considérer :

1) L'ACP $(D_I^{-1} P D_J^{-1}, D_J, D_I)$ qui analyse le tableau des données « corrigées » $p_{ij}/(p_{i.}p_{.j})$ avec les pondérations D_I et D_J :

2) L'ACP $(PD_J^{-1}, D_J, D_I^{-1})$ qui analyse le nuage des distributions conditionnelles par colonnes avec la métrique du Khi2 D_I^{-1} et la pondération D_J ;

3) L'ACP $(D_I^{-1}P, D_J^{-1}, D_I)$ qui analyse le nuage des distributions conditionnelles par lignes avec la métrique du Khi2 D_J^{-1} et la pondération D_I ;

4) L'ACP (P, D_J^{-1}, D_I^{-1}) dont les facteurs D_J-normés et les cofacteurs D_I-normés sont des codages numériques des colonnes et des lignes normalisés pour les distributions marginales qui maximisent successivement la corrélation calculée sur la distribution bivariée P (corrélation canonique) ou respectivement la variance des moyennes conditionnelles par lignes (ou par colonnes) du tableau P.

L'exécution de l'une des quatre analyses donnent les éléments propres (axes, composantes, facteurs, cofacteurs) des trois autres à une transformation linéaire simple (mettant en jeu une matrice diagonale) près. Seul le schéma de dualité rend très simplement compte des relations entre les quatre analyses :

Ces remarques préliminaires permettent de prévoir que l'introduction de projecteurs dans un quelconque de ces modèles permet de conserver l'essentiel des propriétés de l'AFC en imposant des contraintes sur les facteurs. Rappelons encore que :

1) Les axes factoriels et les composantes principales de variance 1 issus de la première analyse, correspondent dans la terminologie de BENZECRI (1982) aux facteurs de variance 1 sur I et J respectivement, facteurs que nous noterons a^I, b^J, ou $\underline{a}$, $\underline{b}$ suivant la terminologie de CAILLIEZ et PAGES (1976) que nous adopterons ici. Ils correspondent également aux cofacteurs et aux facteurs de la quatrième analyse.

2) Dans les quatre analyses précédentes, ainsi que dans la procédure AFC-PROGRAMME, on peut, sans changer les résultats remplacer le tableau P par le tableau P^- de terme général :

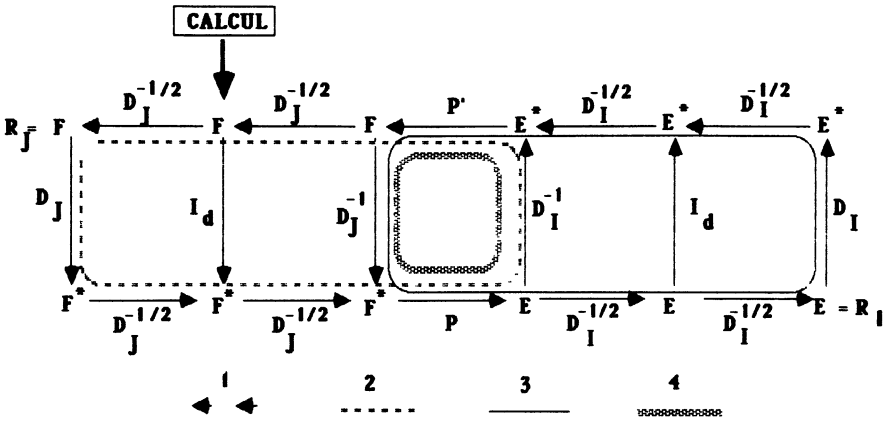


FIGURE 3

Schémas de dualité associés à une AFC. Suivant l'espace qu'on choisit référencé par la base canonique on reconnait : 1) L'ACP($D_I^{-1}PD_J^{-1}$, D_J, D_I). 2) L'ACP(PD_J^{-1}, D_J, D_I^{-1}), 3) L'ACP($D_I^{-1}P$, D_J^{-1}, D_I), 4) L'ACP(P, D_J^{-1}, D_I^{-1}). Les matrices qui y figurent correspondent soit à des normes, soit à des changements de base.

$$p_{ij}^{-} = p_{ij} - p_{i.}p_{.j}$$

puisque l'effet de cette transformation est de remplacer la valeur propre triviale 1 (valeur propre qu'on élimine, et qui est associée au facteur dont toutes les composantes valent 1) par la valeur propre nulle.

II.2. L'analyse des correspondances internes

Considérons maintenant une partition des lignes sous la forme d'une variable qualitative qui engendre un tableau disjonctif complet X comportant I lignes et L colonnes (paquets de lignes) pour une seule variable à L modalités. De même soit une partition des colonnes sous la forme d'une variable qualitative engendrant un tableau disjonctif complet Y comportant J lignes et K colonnes (paquets de colonnes) pour une seule variable à K modalités. X et Y′ sont représentés sur la figure 4 qui précise les notations utiles.

Soit P_X le D_I-projecteur sur le sous-espace vectoriel engendré par les colonnes du tableau X dans $E^* = R^I$. Dans la base canonique la matrice de ce projecteur est :

$$P_X = X(X'D_IX)^{-1}X'D_I = XD_L^{-1}X'D_I$$

où D_L est la matrice diagonale de terme général $p_{\ell+}$ somme des $p_{i.}$ pour les lignes i appartenant à la classe ℓ. Le terme général de la matrice P_X est défini, comme il est aisé de le voir par :

$$\begin{aligned}(P_X)_{ii'} &= p_{i'.}/p_{\ell+} && \text{si i et i' appartiennent à la même classe } \ell \\ &= 0 && \text{sinon}\end{aligned}$$

Soit P_Y le D_J-projecteur sur le sous-espace vectoriel engendré par les colonnes du tableau Y dans $F^* = R^J$. Dans la base canonique la matrice de ce projecteur est :

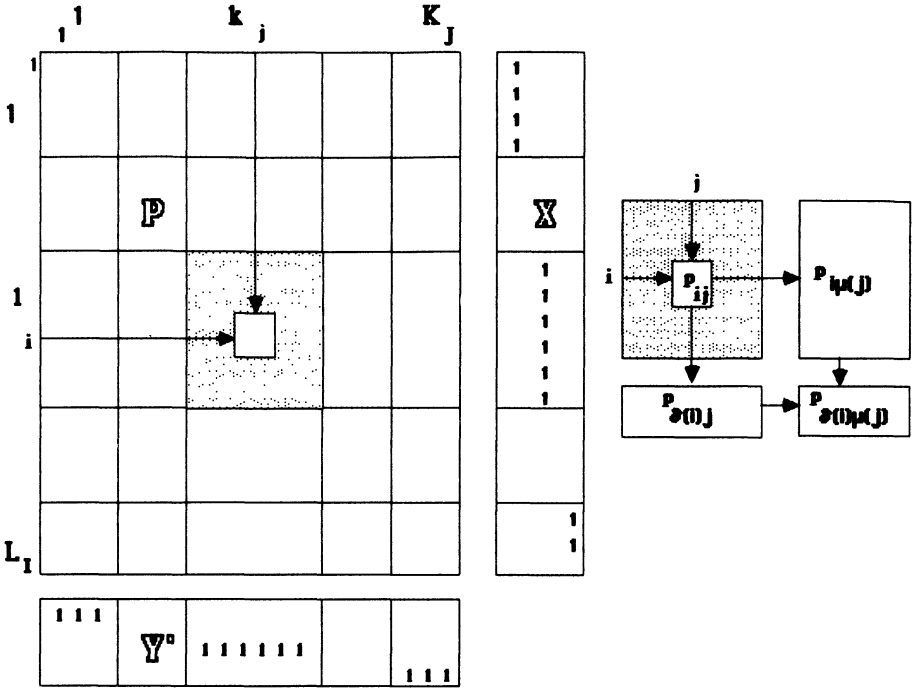


FIGURE 4
Notations utilisées dans le calcul de P°.

$$P_Y = Y(Y'D_JY)^{-1}Y'D_J = YD_K^{-1}Y'D_J$$

où D_K est la matrice diagonale de terme général p_{+k} somme des $p_{.j}$ pour les colonnes j appartenant à la classe k. On a de même, pour le terme général de P_Y :

$$\begin{aligned}(P_Y)_{jj'} &= p_{.j'}/p_{+k} && \text{si j et j' appartiennent à la même classe k}\\ &= 0 && \text{sinon}\end{aligned}$$

Soient Q_X et Q_Y les projecteurs définis par

$$P_X + Q_X = Id_I \text{ et } P_Y + Q_Y = Id_J$$

où Id_I et Id_J sont respectivement les applications identités de R^I et R^J.

Posons : $$P^\circ = Q_X'PQ_Y$$

soit $$P^\circ = (Id_I - P_X')\,P\,(Id_J - P_Y) = P - PP_Y - P_X'P + P_X'PP_Y$$

ce qui permet d'observer que P° est un tableau à I lignes et J colonnes de terme général $P^\circ(i,j)$ qui vaut, compte tenu de la définition de P_X et P_Y :

$$p_{ij} - p_{\delta(i)j}(p_{i.}/p_{\delta(i)+}) - p_{i\mu(j)}(p_{.j}/p_{+\mu(j)}) + p_{\delta(i)\mu(j)}(p_{.j}/p_{+\mu(j)})(p_{i.}/p_{\delta(i)+})$$

avec les notations définies dans la figure 4 et

$$\begin{aligned}p_{\delta(i)+} &= p_{\ell+} && \text{si la ligne i appartient à la classe } \ell\\ p_{+\mu(j)} &= p_{+k} && \text{si la colonne j appartient à la classe k}\end{aligned}$$

qui définissent la « version table de contingence » du double centrage d'un tableau d'analyse de variance à deux facteurs orthogonaux.

On appelera Analyse des Correspondances Internes (ACI) de P par rapport à X et Y

$$\text{l'ACP}\,(Q_X D_I^{-1} P D_J^{-1} Q_Y', D_J, D_I)$$

dont il convient de définir la procédure de calcul et les propriétés.

Cette ACP correspond à la première ACP définie au § II.1., mais où $D_I^{-1} P D_J^{-1}$ est remplacé par $Q_X D_I^{-1} P D_J^{-1} Q_Y'$. Elle peut donc être considérée comme une double analyse en composantes principales sur variables instrumentales (BONIFAS *et al.*, 1984), ou comme une double analyse factorielle privilégiante (LAFAYE DE MICHEAUX, 1978), ou encore comme une double S.A.C.P. (BOUDOU, 1982).

Compte tenu de ce que $D_I^{-1} Q_X' = Q_X D_I^{-1}$ et $D_J^{-1} Q_Y' = Q_Y D_J^{-1}$, l'ACI est identique à

$$\text{l'ACP}\,(D_I^{-1} P^\circ D_J^{-1}, D_J, D_I)$$

qui correspond à la première ACP du § II.1., mais où P est remplacé par P°. Il en résulte que l'ACI est aussi équivalente aux trois autres ACP définies au § II.1., à condition de remplacer P par P°.

En fait cette analyse revient à comparer le tableau P au tableau modèle S de terme général :

$$s_{ij} = p_{ij} - P^\circ(i,j) = p_{\delta(i)j}\,(p_{i.}/p_{\delta(i)+}) + p_{i\mu(j)}\,(p_{.j}/p_{+\mu(j)}) - p_{\delta(i)\mu(j)}\,(p_{.j}/p_{+\mu(j)})\,(p_{i.}/p_{\delta(i)+})$$

et à analyser le tableau résiduel $P^\circ = P - S$. Le principe général de ce type d'analyse est donné par DE FALGUEROLLES *et al.* (1987).

On peut noter que les tableaux P et S ont mêmes marges sur I et J respectivement, marges qui servent, comme dans l'AFC usuelle à définir les métriques pour analyser $P^\circ = P - S$.

On déduit alors des recherches de ESCOFIER (1983), ainsi que de la seconde remarque située à la fin du § II.1. que l'ACI est équivalente à l'AFC du tableau R de terme général

$$r_{ij} = P^\circ(i,j) + p_{i.}p_{.j}$$

puisque le tableau R^- déduit de R, comme P^- se déduit de P est identique à P°.

L'ACI est donc équivalente aux quatre ACP du § II.1., à condition de remplacer P par R. On peut alors définir « l'ACI-PROGRAMME » par la procédure suivante :

1) Calculer le tableau P° défini par le double centrage, puis le tableau R.

2) Effectuer « l'AFC-PROGRAMME » sur le tableau R.

III. Propriétés de l'ACI

III.1. L'ACI est une AFC sous contraintes

Proposition 1

Supposons que l'on recherche les couples de fonctions (a_1^I, b_1^J) de variance 1, et de corrélation maximale (pour la mesure définie par P) sous la contrainte que a_1^I et b_1^J sont de variance interclasse nulle. Notons que a_1^I et b_1^J sont centrés par blocs, i.e. si I_ℓ et J_k désignent respectivement les éléments de I et J associés à la ℓ^e modalités de X, et à la k^e modalité de Y, on a :

$$\forall \ell = 1,L;\ \forall k = 1,K : \Sigma (p_{i.} a_1^i \mid i \in I_\ell) = \Sigma (p_{.j} b_1^j \mid j \in J_k) = 0$$

Alors (a_1^I, b_1^J) correspond au premier couple de facteurs issu de l'ACI de P, i.e. au premier couple de facteurs non triviaux issu de l'AFC de R.

Si on recherche un couple de fonctions (a_2^I, b_2^J) de variance 1, non corrélé à (a_1^I, b_1^J) et de corrélation maximale, sous la contrainte que a_2^I et b_2^J sont de variance interclasse nulle, on obtient le second couple de facteurs associés issus de l'ACI de P, i.e. de l'AFC de R, et on peut continuer le processus qui revient à effectuer l'AFC de R.

Démonstration

On sait que l'analyse des correspondances usuelles revient à faire l'analyse canonique dans $E^* \otimes F^* = R^I \otimes R^J$ des sous-espaces $E^* = R^I$ et $F^* = R^J$ respectivement engendrés par les fonctions qui ne dépendent que de I ou de J, ce qui revient à chercher les fonctions normées $a^I \in R^I$ et $b^J \in R^J$, maximisant $\underline{a}' P\underline{b} = \Sigma_{i,j}\, p_{ij}\, a_i\, b_j$, qui est la corrélation entre a^I et b^J (une fois retirée la fonction triviale constante qui se trouve dans l'intersection de R^I et R^J).

Si maintenant, on impose à a^I et b^J d'être de variance interclasse nulle, on doit avoir $a^I \in Q_X(R^I)$ et $b^J \in Q_Y(R^J)$, auquel cas, on a :

$$a^I = Q_X(a^I) \text{ et } b^J = Q_Y(b^J)$$

et maximiser $\underline{a}' P\underline{b}$ sous la contrainte précédente revient à maximiser $\underline{a}' Q_X' P Q_Y \underline{b} = \underline{a}' P^\circ \underline{b}$, ce qui revient encore (après élimination de la solution triviale constante) à maximiser $\underline{a}' R\underline{b}$. R ayant mêmes marges que P, on obtient bien le premier couple de facteurs non triviaux issu de l'AFC de R.

Si on réitère la procédure, on extrait la suite des facteurs issus de l'AFC de R.

c.q.f.d.

On peut trouver une étude plus détaillée sur la recherche simultanée de facteurs de variance interclasse nulle (qui correspond à l'ACI, i.e. à l'AFC de R) et de variance intraclasse nulle (AFC de $P_X' P P_Y$) dans CAZES (1986a) et CAZES (1986b).

On déduit de la proposition 1 et de sa démonstration les propositions suivantes :

Proposition 2

L'ACI est l'analyse canonique dans $R^I \otimes R^J$ des sous-espaces $Q_X(R^I)$ et $Q_Y(R^J)$.

Proposition 3

Les coordonnées factorielles F^I_α et G^J_α obtenues dans l'ACI sont centrées, centrées par classe, et de variances égales à la valeur propre correspondante λ_α. De plus F_α et G_α sont de corrélation (corrélation intraclasse) égale à $\sqrt{\lambda_\alpha}$, tandis que pour $\alpha \neq \beta$, le couple (F_α, G_α) est non corrélé au couple (F_β, G_β).

La démonstration de cette proposition est immédiate si on se rappelle que les coordonnées factorielles sont obtenues en multipliant les facteurs de variance 1 par la racine carrée de la valeur propre associée.

On sait que l'analyse des correspondances porte souvent le nom de double codage (**dual scaling**) (NISHISATO 1980) du fait qu'elle fournit des paires de codages lignes et colonnes maximisant successivement la corrélation calculée sur la table de contingence P. On a un résultat similaire pour l'ACI, résultat qui découle immédiatement de la proposition 2 et qui est donné par la proposition suivante :

Proposition 4

Les facteurs de variance unité issus de l'ACI sont des codes numériques des lignes et des colonnes centrés, réduits et centrés par classes qui maximisent successivement la corrélation calculée sur la table de contingence. Les maxima successifs sont les racines carrées des valeurs propres (corrélations canoniques) et l'ACI est une méthode de double codage (dual scaling) sous contraintes de centrage par blocs.

Remarque :

Cas où l'on a une partition que sur un seul des ensembles I et J.

Nous supposerons pour fixer les idées, que l'on a une partition sur l'ensemble I seulement, ce qui revient encore à considérer que la partition sur J est la partition triviale ne comportant qu'une seule classe, à savoir l'ensemble J tout entier. Dans ce cas, qui a été en particulier étudié par BENZECRI (1983) et ESCOFIER (1983), le tableau P° a pour terme général :

$$P^\circ(i, j) = p_{ij} - p_{\delta(i)j}(p_{i.}/p_{\delta(i)+})$$

et l'ACI revient à décomposer de façon optimale l'inertie intraclasse du nuage des profils des lignes du tableau P.

III.2. Relations de transition en ACI

Soit (a^I, b^J) un couple non trivial de facteurs associés, de variance 1 et relatifs à la valeur propre λ issu de l'AFC de R. La première formule de transition s'écrit (en notations matricielles) :

$$\underline{b} = (1/\sqrt{\lambda})\, D_J^{-1}\, R'\underline{a}$$

Compte tenu de ce que $\underline{a}$ est centré, et que $\underline{a} = Q_X\underline{a}$, on a, puisque $Q_Y\, D_J^{-1} = D_J^{-1}\, Q_Y'$:

$$\underline{b} = (1/\sqrt{\lambda})\, D_J^{-1}\, P^{o\prime}\, \underline{a} = (1/\sqrt{\lambda})\, D_J^{-1}\, Q_Y'\, P'Q_X\, \underline{a}$$
$$= (1/\sqrt{\lambda})\, Q_Y D_J^{-1}\, P'\, \underline{a} = (1/\sqrt{\lambda})\, Q_Y\, T\underline{a} \quad (1)$$

où $T = D_J^{-1}\, P'$ désigne dans $R^I \otimes R^J$ l'opérateur de projection de R^I vers R^J.

Pour obtenir $\underline{b}$ à partir de $\underline{a}$, il suffit donc d'appliquer la formule de transition usuelle $(1/\sqrt{\lambda})\, T\underline{a}$ dans l'AFC de P, puis de projeter le résultat sur $Q_Y\,(R^J)$.

On a de même :

$$\underline{a} = (1/\sqrt{\lambda})\, D_I^{-1}\, R\, \underline{b} = (1/\sqrt{\lambda})\, Q_X D_I^{-1}\, P\, \underline{b} = (1/\sqrt{\lambda})\, Q_X\, U\, \underline{b} \quad (2)$$

où $U = D_I^{-1}\, P$ désigne l'opérateur de projection de R^J vers R^I.

Les relations $\underline{a} = (1/\sqrt{\lambda})\, Q_X D_I^{-1}\, P\underline{b}$ et $\underline{b} = (1/\sqrt{\lambda})\, Q_Y\, D_J^{-1}\, P'\, \underline{a}$ sont caractéristiques de l'ACI.

Notons qu'on obtient bien sur les mêmes relations (1) et (2) si on remplace $\underline{a}$ et $\underline{b}$ par les coordonnées factorielles $\underline{F} = \underline{a}\sqrt{\lambda}$ et $\underline{G} = \underline{b}\sqrt{\lambda}$. Les relations correspondantes expriment qu'à la déformation minimale $1/\sqrt{\lambda}$ près les lignes sont positionnées à la moyenne conditionnelle des coordonnées des colonnes **recentrée par groupe de lignes,** la propriété symétrique étant vraie pour les colonnes.

Les modes de lecture habituels des cartes factorielles sont donc valides, en particulier :

— pour des valeurs propres proches de l'unité les superpositions des cartes sont des lectures par moyennes réciproques.
— dans les autres cas, la lecture des directions d'éloignement à l'origine est acceptable mais on ne verra que l'éloignement à la moyenne du groupe auquel appartient le point, ce qui est l'objectif choisi.

IV. Illustration

La figure 5 contient une première comparaison entre AFC et ACI du tableau. Le plan 1-2 de l'AFC décrit l'évolution temporelle du contenu faunistique, évolution globalement cyclique (rythme saisonnier) et globalement commune à toutes les stations.

Comme prévu l'ACI élimine ces deux premières valeurs propres et décrit une typologie stationnelle. Mais elle le fait simultanément pour toutes les dates et met directement en évidence trois éléments :

1) La station B ($F_2 > 0$) située sur l'affluent Chassezac est la plus originale; elle s'oppose en permanence aux stations A ($F_2 < 0$) et surtout D ($F_2 < 0$ et $F_1 > 0$), la dernière ayant une identité propre plus marquée (écotone à richesse spécifique élevée). On a là la partie constante de la typologie stationnelle à laquelle il faut ajouter la composition très moyenne et peu originale de la station C, station de pollution insidieuse.

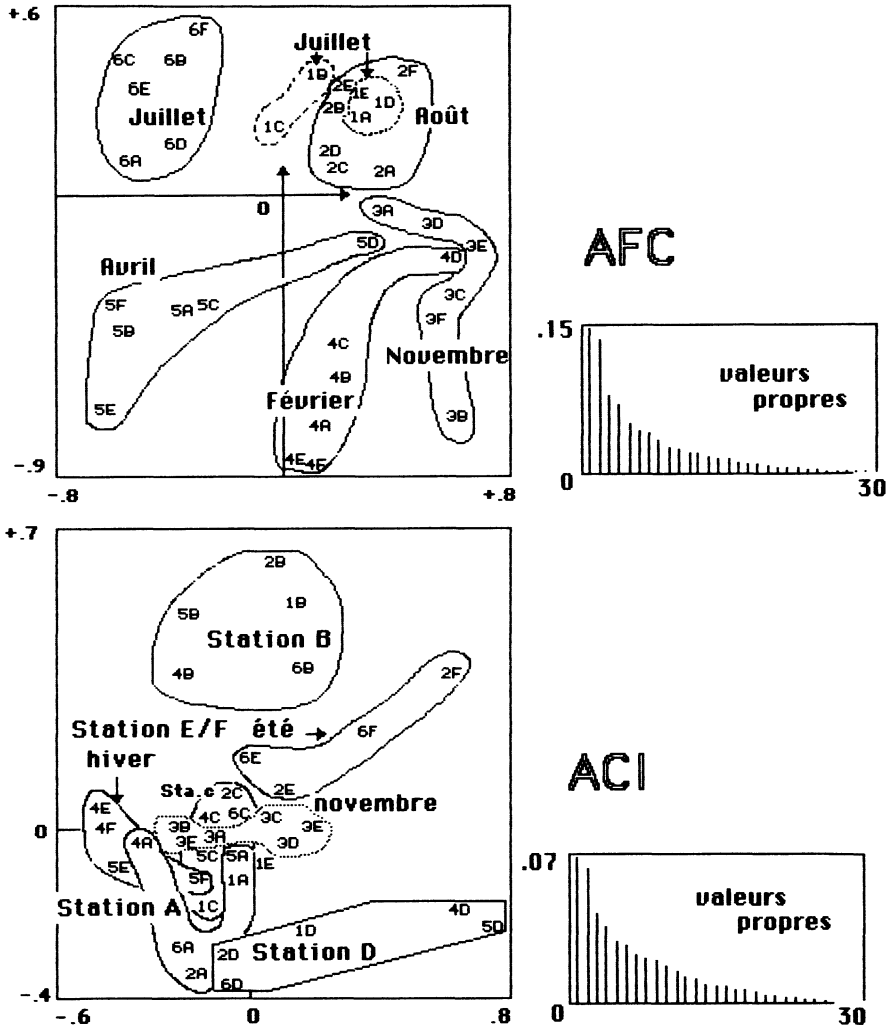


FIGURE 5

Cartes factorielles 1-2 des analyses AFC et ACI du tableau traité. Hormis la position hivernale de la station D, l'AFC fait une typologie des rythmes saisonniers des espèces. L'ACI élimine ce point de vue a priori. *On obtient une typologie stationnelle dans le genre d'un compromis d'une analyse multi-tableaux. L'image de la destruction de la structure faunistique par la première crue de novembre montre la difficulté de l'analyse expérimentale des hydrosystèmes d'eaux courantes.*

2) Les stations E et F présentent ensemble deux états indiquant une partie instable de la typologie stationnelle (conditions d'étiage défavorable en été dans le canyon des Gorges de l'Ardèche).

3) En novembre, après une forte crue caractéristique de la rivière, la déstabilisation du substrat entraîne la mise en mouvement des organismes dans le cours d'eau et donc une uniformisation du contenu faunistique. La valeur typologique des relevés à cette date est quasiment nulle. En particulier l'originalité des stations 2 et 4 disparaît complètement.

Le tableau I précise la décomposition des valeurs propres. On peut appeler contribution d'un bloc à l'inertie d'un facteur la part de la variance de la

TABLEAU I

Décomposition des valeurs propres par groupes des partitions imposées au tableau traité. On peut ainsi mesurer la part des groupes faunistiques à la description des structures communes et discuter le coût des déterminations et l'information qu'elles procurent.

BLOCS COLONNES

	Facteur 1		Facteur2		Poids
	Variance	%	Variance	%	
juillet 82	.0109	3	.0706	21	0.186
août 82	.0413	11	.1064	30	0.180
novembre 82	.0170	3	.0007	0	0.105
février 83	.1917	38	.0320	7	0.136
avril 83	.1385	38	.0613	18	0.190
juillet 83	.0243	7	.0736	24	0.203

BLOCS LIGNES

	Facteur 1		Facteur2		Poids
	Variance	%	Variance	%	
Ephéméroptères	.0511	21	.0563	26	0.288
Plécoptères	.1154	11	.0263	3	0.065
Coléoptères	.0204	11	.0709	41	0.370
Trichoptères	1403	57	.0690	30	0.277
Total/moyenne	0686	100	0633	100	1.000

coordonnée associée aux lignes, respectivement aux colonnes, formant un bloc de lignes, respectivement de colonnes.

Ce descriptif confirme la présence d'une typologie estivale (facteur 2) principalement définie par les Coléoptères, d'une typologie hivernale (facteur 1), principalement associée aux Trichoptères et se résumant à l'originalité de la station D et de l'uniformisation produite par la crue. Le tout est résumé dans la figure 6 qui juxtapose les cartes factorielles par blocs. Cet exemple, volontairement limité à une illustration, permet d'entrevoir que l'analyse des hydrosystèmes devrait beaucoup gagner à l'utilisation de méthodes de description multivariée distinctes des techniques de base (ACP, AFC, ACM).

V. Perspectives

On peut souligner pour conclure que l'opération qui introduit deux projecteurs dans une analyse est « canonique » au sens de la proposition 5, qui suit une idée de SABATIER (1987).

Proposition 5

Si **E** est un espace euclidien de dimension p et de norme M_p, si R est un M_p-projecteur de **E**, si **F** est un espace euclidien de dimension n et de norme M_n, si S est un M_n-projecteur de **F**, on sait que l'espace $\mathbf{L}(E^*, F)$ est un espace euclidien pour la forme bilinéaire

$$X \cdot Y = \mathrm{Tr}\,(M_p X' M_n Y)$$

où Tr désigne la trace d'un endomorphisme (JAFFRENOU, 1978). L'application qui à X associe

$$\pi(X) = SXR'$$

est un projecteur orthogonal dans $\mathbf{L}(E^*, F)$ pour cette norme.

En effet R′ appartient à $\mathbf{L}(E^*, E^*)$ et $\pi(X)$ à $\mathbf{L}(E^*, F)$. De plus

$$\pi(\pi(X)) = S\,SXR'R' = \pi(X)$$

et

$$X \cdot \pi(X) = \mathrm{Tr}\,(M_p X' M_n SXR')$$

Or

$$\pi(X) \cdot \pi(X) = \mathrm{Tr}\,(M_p RX'S'M_n SXR') = \mathrm{Tr}\,(R'M_p X'M_n SSXR') = \mathrm{Tr}\,(M_p X'M_n SXR')$$

soit :

$$\pi(X) \cdot \pi(X) = X \cdot \pi(X)$$

On pourra donc mesurer avec un coefficient RV (ROBERT et ESCOUFIER, 1976) la prise en compte des projecteurs dans l'analyse choisie. L'ACI est alors la plus simple des analyses à deux projecteurs. La double projection ne préjuge pas totalement du choix des normes, ce qui laisse une grande liberté dans l'introduction de contraintes destinées à intégrer des objectifs expérimentaux.

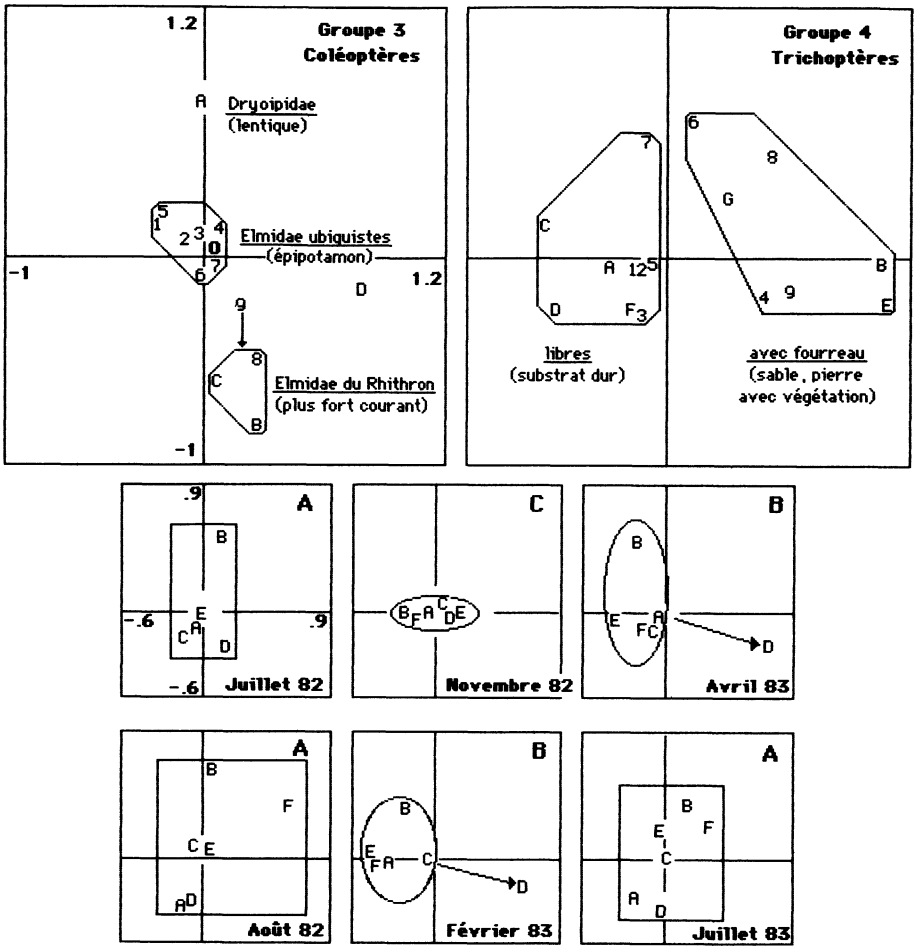


FIGURE 6

Cartes factorielles juxtaposées par groupes de lignes et de colonnes. A : structure estivale, B : originalité de la station D en hiver, C : regroupement des stations en novembre. Le type A est surtout défini par les Coléoptères, le type B par les Trichoptères. Le choix fondamental de VERNEAUX (1973) de définir une biotypologie des eaux courantes à l'étiage est ici illustré.

Références

J.P. BENZECRI (1982). — L'analyse des données, Tome 2 : L'analyse des correspondances, Dunod, 4e édition, 620 pp.

J.P. BENZECRI (1983). — Analyse de l'inertie intraclasse par l'analyse d'un tableau de correspondances, [ANA. INTRACLASSE], *Cah. de l'An. des Données*, 8, 3, 351-358.

L. BONIFAS, Y. ESCOUFIER, P.L. GONZALES et R. SABATIER (1984). — Choix de variables en analyse en composantes principales, *Revue de Statistique Appliquée*, 32, 2, 5-15.

A. BOUDOU (1982). — Analyse en composantes principales partielle, *Stat. et An. des Données*, 7, 2, 1-21.

M. BUYSE (1983). — Les différentes approches conduisant à l'analyse des correspondances. *Biom. Praxim.* XXIII, 1, 1-26.

F. CAILLIEZ et J.P. PAGES (1976). — Introduction à l'analyse des données. SMASH, 9 rue Duban, 75016 Paris, 616 pp.

P. CAZES (1986a). — Correspondance entre deux ensembles et partition de ces deux ensembles [CORR. PART.], *Cah. de l'An. des Données,* 11, 3, 335-340.

P. CAZES (1986b). — Une généralisation des correspondances multiples et des correspondances hiérarchiques, *Cahiers du BURO,* n° 46-47, 37-64.

M. DAKKI (1985). — Sur le choix des données en biotypologie des eaux courantes par l'analyse factorielle des correspondances. *Bull. Ecol.,* 16, 4, 285-296.

M. DAKKI (1986). — Biotypologie et gradient thermique spatio-temporel, étude sur un cours d'eau du moyen-atlas (Maroc). *Bull. Ecol.,* 17, 2, 79-85.

S. DOLEDEC (1986). — Les peuplements de macroinvertébrés benthiques du cours inférieur de l'Ardèche. Dynamique spatio-temporelle. Thèse de doctorat, Université Lyon 1, 244 pp.

B. ESCOFIER (1983). — Analyse de la différence entre deux mesures définies sur le produit de deux mêmes ensembles, [ANA. DIFF. PROD.], *Cah. de l'An. des Données,* 8, 3, 325-329.

A. de FALGUEROLLES et P.G.M. van der HEIJDEN (1987). — Sur l'analyse factorielle des correspondances et quelques-unes de ses variantes, *Revue de Statistique Appliquée,* 35, 3, 7-12.

M.J. GREENACRE (1984). — Theory and applications of correspondence analysis. Academic Press, London, 364 pp.

M.O. HILL (1973). — Reciprocal averaging : An eigenvector method of ordination. *J. Ecol.,* 61, 237-249.

M.O. HILL (1974). — Correspondence analysis : A neglected multivariate method. *J. Roy. Stat. Soc., Séries C,* 23, 340-354.

P.A. JAFFRENOU (1978). — Sur l'analyse des familles finies de variables vectorielles. Bases algébriques et applications à la description statistique. Thèse de 3e cycle, Université Lyon 1, 97 pp.

D. LAFAYE DE MICHEAUX (1978). — Approximations d'analyses canoniques non linéaires de variables aléatoires. Analyses factorielles privilégiantes, Thèse de Docteur-Ingénieur, Université de Nice, 215 pp.

C. LEVEQUE et M. GABORIT (1972). — Utilisation de l'analyse factorielle des correspondances pour l'étude des peuplements en Mollusques benthiques du lac Tchad. *Cah. ORSTOM Hydrobiol.,* 4, 1, 47-66.

S. NISHISATO (1980). — Analysis of categorical data : dual scaling and its applications. University of Toronto Press, London, 276 pp.

P. ROBERT et Y. ESCOUFIER (1976). — A unifying tool for linear multivariate statistical methods : the RV coefficient. *Applied Statistics,* 25, 3, 257-265.

G. ROUX et M. ROUX (1967). — A propos de quelques méthodes de classification en phytosociologie. *Revue de Statistique Appliquée,* 15, 2, 59-72.

R. SABATIER (1987). — Analyse factorielle de données structurées et métriques, (proposé pour publication) 17 pp.

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

J. VERNEAUX (1973). — Cours d'eau de Franche-Comté (Massif du Jura). Recherches écologiques sur le réseau hydrographique du Doubs. Essai de biotypologie. Thèse d'état, Besançon, 257 pp.