REVUE DE STATISTIQUE APPLIQUÉE

J. PONTIER
M. O. PERNIN

Relations entre analyse canonique complète et méthode Longi

Revue de statistique appliquée, tome 37, n° 4 (1989), p. 67-82
<http://www.numdam.org/item?id=RSA_1989__37_4_67_0>

Rev. Statistique Appliquée, 1989, **XXXVII** (4), 67-82

RELATIONS ENTRE ANALYSE CANONIQUE COMPLÈTE ET MÉTHODE LONGI

J. PONTIER & M.O. PERNIN

Laboratoire d'Analyse de Données et Biométrie
Université Claude Bernard, 43 Boulevard du 11 Novembre 1918
69622 VILLEURBANNE CEDEX FRANCE

RÉSUMÉ

Dans l'analyse d'un ensemble de variables, dépendant de deux facteurs structurants, la méthode LONGI produit des fonctions linéaires des variables, à la fois les plus corrélées avec l'un des facteurs structurants, et les moins corrélées avec l'autre facteur structurant. D'autre part, l'analyse canonique complète est un prolongement de l'analyse canonique ordinaire. Le présent article montre que la méthode LONGI est une forme particulière d'utilisation de l'analyse canonique complète, et propose une extension de définition de la méthode LONGI.

ABSTRACT

In the analysis of a set of variables depending on two factors, the so-called LONGI method produces linear functions depending the most on one factor and the least on the other. Complete canonical correlation analysis is an extension of ordinary canonical correlation analysis. In this paper we show that LONGI is a particular use of complete canonical correlation analysis, and we suggest an extension of the definition of this method.

Mots-clés : *Méthode LONGI, Analyse canonique, Analyse factorielle sous contraintes.*
Indices de classification STMA : *06 :040 06 :030 06 :110*

1. Introduction

En 1985, lors des Journées de Statistique (Pau), et lors du Third Symposium on Data Analysis (Bruxelles), nous avions présenté, sous le nom de "méthode LONGI", une méthode d'analyse de données longitudinales multivariées. Nous avions créé cette méthode dans un contexte d'étude de la croissance morphologique chez l'être humain (auxologie). L'idée était de démêler ce qui, dans l'expression de cette croissance (mesures diverses faites sur chaque individu, à divers âges échelonnées entre la naissance et l'âge adulte), revenait au phénomène biologique de croissance indépendamment de l'individu, de ce qui revenait en propre à l'individu indépendamment de la "loi" biologique générale. Nous avons décrit en détails cette méthode, d'une part dans la Thèse de Doctorat de l'un de nous (PERNIN [5]), d'autre part dans les Proceedings of the Symposium on Data Analysis (PONTIER et PERNIN [6]).

Une recherche théorique plus approfondie sur les propriétés de la méthode LONGI nous a amenés à une reformulation de l'analyse canonique, montrant le parti que l'on pouvait tirer d'une décomposition "complète" de chacun des espaces de variables concernés. C'est ce qui nous a conduits à exposer, aux Journées de Statistiques 1987 (Lausanne), l'Analyse Canonique Complète dont la théorie est entièrement décrite dans PONTIER, JOLICOEUR et PERNIN [7]. Notre propos est ici de préciser les liens qui existent entre la méthode LONGI et l'Analyse Canonique Complète, et de proposer une extension naturelle de la méthode LONGI.

2. Rappels méthodologiques

Les notations utilisées ici sont les mêmes que celles utilisées dans PERNIN [5], PONTIER et PERNIN [6], PONTIER et al.[7]. La situation que nous examinons concerne un ensemble $\Omega = \{\omega_1, \omega_2, ..., \omega_n\}$ de n unités statistiques ou unités d'observation, pondérées par les poids π_j =poids de ω_j (strictement positifs, de somme unité), et sur lesquelles ont été observées des variables réparties en trois ensembles : $X = \{x_1, x_2, ..., x_p\}$, $Y = \{y_1, y_2, ..., y_q\}$, $Z = \{z_1, z_2, ..., z_r\}$. L'ensemble des applications de Ω dans $\mathbb{R}$ est un espace vectoriel, noté $[\Omega]$, de dimension n, dont la base naturelle est constituée par les indicatrices des n unités statistiques; il est muni du produit scalaire φ associé à la pondération des unités statistiques (métrique diagonale des poids), soit :

$$x, y \epsilon [\Omega] \longrightarrow \varphi(x,y) = \sum_j \pi_j x(\omega_j)\, y(\omega_j)$$

Dans $[\Omega]$, qui est donc un espace euclidien, les trois ensembles de variables engendrent respectivement trois sous-espaces notés $[X]$ (de dimension $\leq$ p), $[Y]$ (de dimension $\leq$ q), $[Z]$ (de dimension $\leq$ r).

Les données numériques disponibles se présentent donc sous la forme au tableau suivant :

obs	poids	x_1	...	x_p	y_1	...	y_q	z_1	...	z_r
ω_1	π_1	x_{11}	...	x_p1	y_{11}	...	y_q1	z_{11}	...	z_r1
ω_2	π_2	x_{12}	...	x_p2	y_{12}	...	y_q2	z_{12}	...	z_r2
:	:	:		:	:		:	:		:
ω_n	π_n	x_{1n}	...	x_{pn}	y_{1n}	...	y_{qn}	z_{1n}	...	z_{rn}

Notre problématique générale vise à obtenir des fonctions linéaires des variables de l'ensemble X, possédant des propriétés particulières exprimées par des relations de proximité faisant intervenir conjointement les deux ensembles de variables Y et Z. Les moyens d'y parvenir, que nous décrivons ci-après, s'inscrivent donc dans la catégorie des "analyses factorielles sous contraintes". Les fonctions linéaires de variables seront interprétées comme vecteurs dans l'espace vectoriel $[\Omega]$ en général, et dans le sous-espace engendré par ces variables en particulier. Les relations de proximité seront interprétées en termes de cosinus d'angles (ou de corrélations linéaires), et de distances, dans l'espace euclidien $[\Omega]$.

2.1 Analyse Canonique Complète

L'Analyse Canonique Complète entre deux sous-espaces $[X]$ et $[Y]$ de $[\Omega]$, prolonge l'analyse canonique au sens habituel, en décomposant "complètement" $[X]$ (resp.[Y]) en somme directe φ – orthogonale (symbolisée par $\oplus\perp$) de trois sous-espaces ayant, avec $[Y]$ (resp. $[X]$), des relations bien précises basées sur l'angle formé par les vecteurs de $[X]$ (resp. $[Y]$) et ceux de $[Y]$ (resp. [X]). On obtient ainsi la *décomposition complète de* $[X]$ *canonique par rapport à* $[Y]$, et la *décomposition complète de* $[Y]$ *canonique par rapport à* $[X]$, soit :

$$[X] = [X]_1 \oplus \perp[X]_2 \oplus \perp[X]_3 \quad \text{et} \quad [Y] = [Y]_1 \oplus \perp[Y]_2 \oplus \perp[Y]_3$$

décompositions dont les termes sont précisés ci-après. De plus, à partir de ces deux décompositions, on induit une *décomposition canonique complète* de $[X] + [Y]$ (sous-espace de $[\Omega]$ engendré par l'ensemble $X = Y = \{x_1, ..., x_p, y_1, ..., y_q\}$) en somme directe φ – orthogonale de quatre sous-espaces :

$$[X] + [Y] = ([X] \cap [Y]) \oplus \perp([X]_2 \oplus [Y]_2) \oplus \perp[X]_3 \oplus \perp[Y]_3$$

Les sous-espaces intervenant dans ces décompositions sont obtenus à partir des éléments propres des matrices $M_{XX} = {\Phi_{XX}}^{-1}\Phi_{XY}{\Phi_{YY}}^{-1}\Phi_{YX}$ et $N_{YY} = {\Phi_{YY}}^{-1}\Phi_{YX}{\Phi_{XX}}^{-1}\Phi_{XY}$ (Φ_{UV} note la matrice des produits scalaires croisés entre les élements d'un ensemble U et ceux d'un ensemble V d'éléments de $[\Omega]$). Ces sous-espaces sont :

$[X]_1 = [X] \cap [Y]$ = ensemble des $u \epsilon [X]$ appartenant aussi à $[Y]$

$[X]_2$ = ensemble des $u \epsilon [X]$ obliques par rapport à $[Y]$ (on dit que u est "oblique" par rapport à $[Y]$ s'il n'appartient pas à $[Y]$ et n'est *φ-orthogonal* à aucun vecteur de $[Y]$)

$[X]_3 = [X] \cap [Y]\perp$ =ensemble des $u \epsilon [X]$ φ – orthogonaux à $[Y]$

$[Y]_1 = [Y] \cap [X]$ =ensemble des $v \epsilon [Y]$ appartenant aussi à $[X]$

$[Y]_2$ = ensemble des $v \epsilon [Y]$ obliques par rapport à $[X]$

$[Y]_3 = [Y] \cap [X]\perp$ = ensemble des $v \epsilon [Y]$ φ – orthogonaux à $[X]$

$[X]_2 \oplus [Y]_2$ = sous-espace de $[X] + [Y]$ "interdépendant" de $[X]$ et de $[Y]$

(Cette somme directe n'est pas nécessairement φ – orthogonale)

p' note le rang de X, q' delui de Y

μ note le nombre de valeurs propres strictement positives, communes aux matrices M_{XX} et N_{YY}

ν note le nombre de valeurs propres égales à 1 (elles figurent parmi les μ valeurs propres positives)

$$\begin{aligned}
\dim [X] &= p' \leq p \ ; \ \dim [Y] = q' \leq q \ ; \ \dim([X] + [Y] = p' + q' - \nu) \\
\dim [X]_1 &= \dim [Y]_1 = \nu = \text{ nombre de corrélations canoniques égales à 1} \\
\dim [X]_2 &= \dim [Y]_2 = \mu - \nu \\
&= \text{nombre de corrélations strictement comprises entre 0 et 1}
\end{aligned}$$

$\dim [X]_3 = p' - \mu =$ nombre de corrélations canoniques nulles dans $[X]$
$\dim [Y]_3 = q' - \mu =$ nombre de corrélations canoniques nulles dans $[Y]$
$\dim ([X]_2 \oplus [Y]_2) = 2(\mu - \nu)$

On notera que, dans la décomposition complète de $[X]$ (resp.$[Y]$) canonique par rapport à $[Y]$ (resp. $[X]$), chacun des trois sous-espaces composants $[X]_1, [X]_2, [X]_3$ (resp. $[Y]_1, [Y]_2, [Y]_3$) peut être réduit au vecteur nul; mais ils ne peuvent pas l'être tous les trois en même temps.

Si l'un ou/et l'autre des deux ensembles de variables X, Y, est consitué par l'ensemble des indicatrices des modalités d'une variable qualitative, l'analyse canonique complète devient, selon le cas, une *Analyse Discriminante Complète* (cas d'une seule variable qualitative), ou une *Analyse des Correspondances Complète* (cas de deux variables qualitatives).

La décomposition canonique complète de $[X]$ et celle de$[Y]$, telles que nous les avons décrites ci-dessus, sont exprimées par BENZECRI [1] (p. 179-181) et par CAILLIEZ et PAGES [2] (p. 351-392). Ces auteurs cependant s'en tiennent à une affirmation d'existence, sans expliciter l'usage (propriétés, manière d'obtenir, exemples) des sous-espaces que nous avons notés ci-dessus $[X]_3$ et $[Y]_3$. Ce fut précisément l'objet de notre propre article PONTIER, JOLICOEUR et PERNIN [7], d'attirer l'attention du lecteur sur l'intérêt que pouvaient présenter ces sous-espaces, et de montrer comment pratiquer cette analyse canonique alors qualifiée de "complète".

2.2 Méthode LONGI

Il s'agit du cas où chacun des deux systèmes $[Y]$ et $[Z]$est constitué par les indicatrices des modalités d'une variable qualitative, le système $[X]$ étant lui, constitué par des variables quelconques (en particulier, $[X]$ peut éventuellement être l'ensemble des indicatrices des modalités d'une troisième variable qualitative). Le tableau des données, incluant donc deux tableaux disjonctifs complets, se présente sous la forme suivante, où la notation 01 symbolise la présence soit de 0, soit de 1 :

obs	poids	x_1	...	x_p	y_1	...	y_q	z_1	...	z_r
ω_1	π_1	x_{11}	...	x_{p1}	01	...	01	01	...	01
ω_2	π_2	x_{12}	...	x_{p2}	01	...	01	01	...	01
:	:	:		:	:		:	:		:
ω_n	π_n	x_{1n}	...	x_{pn}	01	...	01	01	...	01

Notre objectif est d'obtenir une fonction linéaire des variables x_i, qui soit la plus "proche" possible de la variable qualitative Y (resp. Z), *et* la plus "éloignée" possible de la variable qualitative Z (resp. Y) (cf l'exemple d'application cité en introduction). La méthode LONGI, traduisant la notion de proximité en termes de corrélation linéaire, résout ce problème de la manière suivante, reformulée ici en terme d'analyse canonique complète :

1) L'analyse canonique complète entre $[Y]$ et $[Z]$ (c'est-à-dire, en l'occurence, l'analyse des correspondances complète entre les deux variables qualitatives Y et

Z) nous permet d'obtenir, entre autres, $[Y] \cap [Z]\perp$ (resp.$[Z] \cap [Y]\perp$), ensemble des vecteurs de $[Y]$ (resp.[Z]) φ – orthogonaux à tout vecteur de $[Z]$ (resp.$[Y]$). En d'autres termes, dans la mesure où il n'est pas réduit au vecteur nul, $[Y] \cap [Z]\perp$ (resp.$[Z] \cap [Y]\perp$) est l'ensemble des codages de la variable qualitative Y (resp. Z) ayant une corrélation nulle avec tout codage de la variable qualitative Z (resp. Y). Remarquons que, dans ces conditions, les éléments de $[Y] \cap [Z]\perp$ (resp. $[Z] \cap [Y]\perp$) sont tous φ–orthogonaux au vecteur $1\epsilon[Z]$ (resp. $1\epsilon[Y]$). Ils sont donc des codages *centrés* de la variable qualitative Y (resp. Z).

2) L'analyse canonique ordinaire entre $[X]$ et $[Y] \cap [Z]\perp$ (resp. $[Z] \cap [Y]\perp$) fournit un "premier couple canonique", constitué par :

- d'une part, une fonction linéaire $f_1 = f_{11}x_1 + f_{12}x_2 + ... + f_{1p}x_p$ des variables x_i,
- d'autre part un codage $c_1 = c_{11}y_1 + c_{12}y_2 + ... + c_{1q}y_q$ (resp. $c_1 = c_{11}z_1 + c_{12}z_2 + ... + c_{1r}z_r$) de la variable qualitative Y (resp. Z),ces deux variables f_1 et c_1 ayant entre elles la corrélation maximum (égale à la première corrélation canonique) qui puisse exister entre d'une part les fonctions linéaires des variables $x_1, ..., x_p$, et d'autre part les codages de Y (resp. Z) non corrélés avec Z (resp. Y).

La fonction linéaire f_1 est donc la réponse apportée à notre problème initial, par cette façon de procéder, dans la mesure où le critère de proximité choisi est celui de corrélation maximum. Les éventuelles autres fonctions canoniques $f_2, f_3, ...$, apportent par ailleurs des solutions *complémentaires*, répondant au même problème auquel s'ajoute la contrainte de φ–orthogonalité par rapport aux solutions précédentes.

3. Exemple

Nous reprenons un exemple récemment utilisé par DOLEDEC et CHESSEL [3], THIOULOUSE et CHESSEL [10], et SABATIER [8, 9], dans un contexte de propositions méthodologiques en analyse des données. Les données disponibles sont relatives à une étude, due à PEGAZ-MAUCET [4], de la pollution du ruisseau "le Méaudret", affluent de la Bourne, dans le Vercors. Le Méaudret reçoit les eaux résiduaires des deux localités Autrans et Méaudre. Nous avons affaire à deux variables qualitatives :

Y= le facteur "station d'observation" (6 modalités : les stations d'observation, désignées par un numéro de 1 à 6; les cinq premières sont échelonnées le long du Méaudret, de l'amont vers l'aval; la sixième est située sur la Bourne, peu avant le confluent);

Z = le facteur "saison d'observation" (4 modalités : printemps = juin 1978, été = Août 1978, automne = novembre 1978, hiver = février 1979).

Une "observation" a consisté à mesurer, en une station donnée, en une saison donnée, 10 variables physico-chimiques, témoins de pollution :

x_1 = température de l' eau en ° C

x_2 = débit en l/s

x_3 = pH (x 10)
x_4 = conductivité en μmhos/cm
x_5 = oxygène dissous (O_2) en mg/l
x_6 = demande biochimique en oxygène (DBO_5) en mg/l d'oxygène (x 10)
x_7 = demande chimiques en oxygène (DCO) en mg/l d'oxygène (x 10)
x_8 = azote ammoniacal (NH_4) en mg/l de NH_4^+ (x 100)
x_9 = azote nitrique (NO_3) en mg/l de NO_3^- (x 10)
x_{10} = orthophosphates (PO_4) en mg/l de PO_4^{---} (x 100)

Dans le tableau ci-dessous sont rassemblées les informations concernant toutes les variables. Pour chacune des variables $x_i (i = 1 à 10)$ sont indiquées les valeurs numériques mesurées. A chacune des variables qualitatives Y et Z, on a associé une colonne unique contenant les numéros des modalités (procédé équivalent au tableau disjonctif complet).

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	Y	Z
10	41	85	295	110	23	14	12	34	11	1	1
13	62	83	325	95	23	18	11	30	13	1	2
1	25	84	315	91	16	5	7	64	3	1	3
3	118	80	325	100	16	12	17	18	19	1	4
11	158	83	315	13	76	33	285	27	150	2	1
13	80	76	380	20	210	57	980	8	365	2	2
3	63	80	425	38	360	80	1250	22	650	2	3
3	252	83	360	100	95	29	252	46	160	2	4
11	198	85	290	113	33	15	40	40	10	3	1
15	100	78	385	46	150	25	790	77	450	3	2
2	79	81	350	84	71	19	270	132	370	3	3
3	315	83	370	100	87	28	280	48	285	3	4
12	280	86	290	126	35	15	45	40	73	4	1
16	140	80	360	76	120	26	490	84	345	4	2
3	85	83	330	106	20	14	42	120	160	4	3
3	498	83	330	100	48	16	104	44	82	4	4
13	322	85	285	117	36	16	48	46	84	5	1
15	160	84	345	91	17	19	22	100	174	5	2
2	72	86	305	91	16	9	10	95	125	5	3
2	390	82	330	100	17	12	56	50	60	5	4
11	303	85	245	100	17	9	5	27	16	6	1
13	310	82	285	82	85	16	59	37	60	6	2
4	181	86	270	105	28	5	10	37	43	6	3
3	480	82	290	100	13	8	4	22	13	6	4

(pour des raisons techniques de présentation, nous avons modifié des variables d'origine, en les multipliant par un facteur 10 ou 100, facteur indiqué entre parenthèses dans la nomenclature des variables; les méthodes d'analyse présentées étant à base d'analyse canonique, ne seront pas affectées par ces changements d'unité).

3.1 Analyse sans donnée manquante

Le plan d'observation tel qu'il se présente est un "plan complet orthogonal" : à chaque couple de modalités (y_r, z_s) des facteurs structurants Y= "Station" et Z = "Saison" est associée une et une seule observation. C'est là une situation *particulière*, qui a des conséquences simplificatrices sur l'analyse canonique complète entre $[Y]$ et [Z]. En effet, dans ce cas, l'analyse canonique complète entre les sous-espaces $[Y]$ et $[Z]$ nous donne :

- une seule valeur propre égale à 1, correspondant au vecteur propre **1** commun aux deux sous-espaces; c'est la solution triviale, attendue : $[Y] \cap [Z] = [1]$;
- toutes les autres valeurs propres sont nulles; il n'existe aucun sous-espace interdépendant, on a $[Y] \cap [Z]\perp = [Y] \cap [1]\perp$(ensemble des codages centrés de Y), et $[Z] \cap [Y]\perp = [Z] \cap [1]\perp$ (ensemble des codages centrés de Z), relations également attendues car le plan d'observation est orthogonal.

L'analyse canonique complète entre [X] et $[Y] \cap [Z]\perp$ est alors, dans ce cas particulier "sans donnée manquante", l'analyse canonique complète entre $[X]$ et $[Y] \cap [1]\perp$. La partie "ordinaire" de cette analyse canonique s'identifie par conséquent à une *analyse discriminante ordinaire* (les variables actives sont les 10 variables x_i, les groupes sont les 6 stations). Il en sera bien entendu de même avec l'analyse canonique entre [X] et $[Z] \cap [Y]\perp$, identique ici à l'analyse discriminante ordinaire entre les 4 saisons.

Bien que cette relation entre la méthode LONGI et l'analyse discriminante soit intéressante à noter sur le plan théorique, nous n'avons pas jugé indispensable de faire état ici de sa réalisation numérique, l'analyse discriminante étant par ailleurs une méthode très classique. Nous avons préféré développer dans le détail l'application de la méthode LONGI au cas avec données manquantes, puisque c'est précisément dans ce cas, très répandu, qu'apparaît vraiment l'originalité et l'efficacité de la méthode.

3.2 Cas avec données manquantes

Nous avons créé artificiellement une situation "avec données manquantes" en supprimant, dans le tableau initial, les quatre observations numéros 2, 11, 16, 18 (choix arbitraire). L'espace des observations est donc de dimension 20. La pondération des observations est supposée uniforme, soit π_j= 1/20, j = 1 à 20.

L'analyse canonique complète entre $[Y]$ et $[Z]$, première étape de la méthode, nous donne une unique valeur propre égale à 1 (c'est la valeur propre triviale), et trois autres valeurs propres positives (corrélations canoniques égales respectivement à 0.35076, 0.25820, 0.08675).

La décomposition complète de $[Y]$ canonique par rapport à $[Z]$ est la suivante :

$[Y]_1 = [Y] \cap [Z] = [Z] \cap [Y] = [Z]_1 = [1]$, de dimension 1, comme précédemment;

$[Y]_2$ est de dimension 3, engendré par la base φ – orthonormée :

$u_1 = 0.28746\ \mathbf{1} - 1.73403\ \mathbf{y}_1 + 0.77584\ \mathbf{y}_3 + 0.77584\ \mathbf{y}_4 - 1.73403\ \mathbf{y}_5$
$u_2 = -1.82574\ \mathbf{y}_3 + 1.82574\ \mathbf{y}_4$
$u_3 = 1.19053\ \mathbf{1} - 1.68121\ \mathbf{y}_1 - 2.28723\ \mathbf{y}_3 - 2.28723\ \mathbf{y}_4 - 1.68121\ \mathbf{y}_5$

$[Y]_3 = [Y] \cap [Z]\perp$ est de dimension 2, engendré par la base φ – orthonormée (par exemple) :

$u_4 = 1.82574\ \mathbf{y}_1 - 1.82574\ \mathbf{y}_5$
$u_5 = -1.58114\ \mathbf{1} + 1.58114\ \mathbf{y}_1 + 1.58114\ \mathbf{y}_3 + 1.58114\ \mathbf{y}_4 + 1.58114\ \mathbf{y}_5 + 3.16228\ \mathbf{y}_6$

La décomposition complète de $[Z]$ canonique par rapport à $[Y]$ est la suivante :

$[Z]_1 = [Z] \cap [Y] = [Y] \cap [Z] = [Y]_1 = [1]$, de dimension 1;

$[Z]_2$ est de dimension 3, engendré par la base φ – orthonormée :

$v_1 = -0.09106\ \mathbf{1} + 2.01655\ \mathbf{z}_2 - 0.62450\ \mathbf{z}_3 - 0.62450\ \mathbf{z}_4$
$v_2 = 1.41421\ \mathbf{z}_3 - 1.41421\ \mathbf{z}_4$
$v_3 = -1.52481\ \mathbf{1} + 2.06564\ \mathbf{z}_2 + 2.22336\ \mathbf{z}_3 + 2.22336\ \mathbf{z}_4$

$[Z]_3 = [Z] \cap [Y]\perp = [0]$, sous-espace réduit au vecteur nul.

La recherche de fonctions linéaires des variables x_i optimales pour les stations (c'est-à-dire corrélées le plus possible avec le facteur Station, et le moins possible avec le facteur Saison), deuxième étape de la méthode, consiste en l'analyse canonique ordinaire entre $[X]$ et $[Y] \cap [Z]\perp$. Cette analyse canonique nous fournit deux "fonctions optimales-stations" f_1, f_2, d'expressions respectives :

$f_1 = 27.65006\ \mathbf{1} - 0.05543\ \mathbf{x}_1 + 0.00185\ \mathbf{x}_2 - 0.20995\ \mathbf{x}_3 - 0.03406\ \mathbf{x}_4 + 0.00653\ \mathbf{x}_5 + 0.01235\ \mathbf{x}_6 - 0.07107\ \mathbf{x}_7 - 0.00160\ \mathbf{x}_8 - 0.00079\ \mathbf{x}_9 + 0.00796\ \mathbf{x}_{10}$
$f_2 = 19.90677\ \mathbf{1} + 0.01730\ \mathbf{x}_1 - 0.00843\ \mathbf{x}_2 - 0.22847\ \mathbf{x}_3 + 0.00901\ \mathbf{x}_4 + 0.00762\ \mathbf{x}_5 + 0.01782\ \mathbf{x}_6 - 0.07889\ \mathbf{x}_7 - 0.00657\ \mathbf{x}_8 - 0.04044\ \mathbf{x}_9 + 0.00505\ \mathbf{x}_{10}$

Les codages canoniques des stations (codages optimaux-stations), qui leur sont associés, sont :

$c_1 = -0.19675\ \mathbf{u}_4 + 0.98045\ \mathbf{u}_5$
$= -1.55023\ \mathbf{1} + 1.19102\ \mathbf{y}_1 + 1.55023\ \mathbf{y}_3 + 1.55023\ \mathbf{y}_4 + 1.90944\ \mathbf{y}_5 + 3.10046\ \mathbf{y}_6$
$c_2 = 0.98045\ \mathbf{u}_4 + 0.19675\ \mathbf{u}_5$
$= -0.31109\ \mathbf{1} + 2.10114\ \mathbf{y}_1 + 0.31109\ \mathbf{y}_3 + 0.31109\ \mathbf{y}_4 - 1.47896\ \mathbf{y}_5 + 0.62218\ \mathbf{y}_6$

Les corrélations canoniques sont respectivement égales à 0.97604 et 0.79872. Les résultats ci-dessus nous permettent d'obtenir d'une part toutes les autres corrélations utiles, d'autre part les valeurs individuelles des fonctions f_1, f_2, c_1, c_2.Nous ne donnons pas ici tous ces chiffres, faciles à retrouver; nous préférons aller directement aux représentations graphiques qui en découlent.

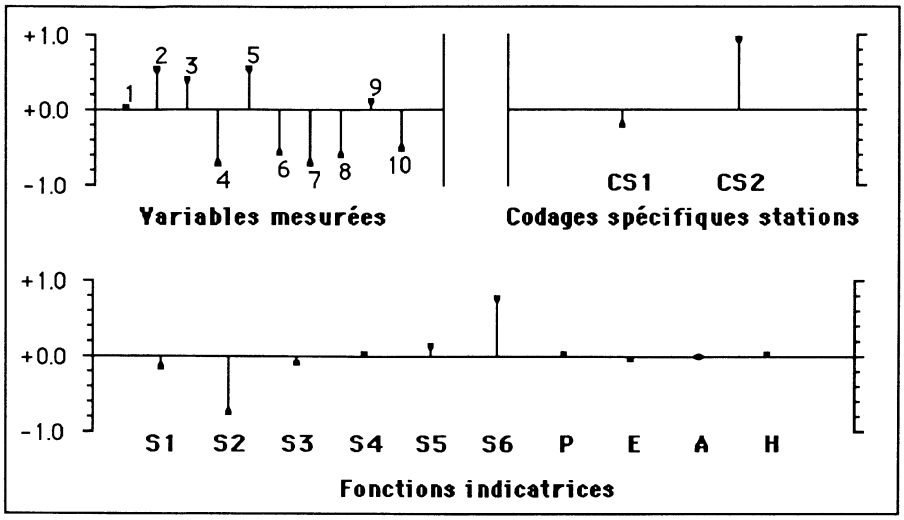

Fig. 1. Représentation des corrélations entre la fonction optimale-stations f_1, et diverses variables intervenant dans cette analyse; les variables mesurées x_1 à x_{10}, les codages CS1 et CS2 spécifiques des stations, les indicatrices des stations S1 à S6, les indicatrices des saisons P, E, A, H.

La figure 1 permet de préciser les liens (linéaires) entre f_1 d'une part, et les 10 variables mesurées d'autre part. La fonction f_1, corrélée négativement avec les variables "de pollution" x_6, x_7, x_8, x_{10}, et positivement avec les variables "de non pollution" x_3, x_5, peut être considérée comme un indice global de pollution (une augmentation de l'indice traduit une diminution de la pollution). Notons par ailleurs la corrélation positive entre f_1 et x_2 : une augmentation du débit est liée à une diminution de la pollution.

La fonction f_2(fig 2) n'est corrélée positivement qu'avec x_5 (oxygène dissous). Par ailleurs, f_1 et f_2 sont pratiquement sans corrélation avec x_1 : c'est normal, car la température de l'eau est très liée à la saison, alors que f_1 et f_2 sont par construction les moins corrélées possible avec la saison. Les variables supplémentaires sont d'une part les codages spécifiques des stations, u_4 et u_5 (notés ici **CS1** et **CS2**), d'autre part les indicatrices des stations y_1 à y_6 (notées **S1** à **S6**) et les indicatrices des saisons z_1 à z_4 (notées **P, E, A, H**). Les corrélations de f_1 et f_2 avec ces quatre dernières sont pratiquement nulles (but recherché); quant aux corrélations avec les indicatrices de stations, elles font ressortir : pour f_1, la forte pollution de la station S2 et la faible pollution de S6, et pour f_2, le fort taux d'oxygène de la station S1.

Ces corrélations sont reprises en représentation dans le plan f_1, f_2 (fig 3), ce qui va nous permettre d'interpréter, dans la carte factorielle (fig 4), les positions des observations, entre elles et par rapport aux axes. Dans cette carte factorielle, on a relié entre elles les observations relatives à une même station, par une ligne polygonale ("trajectoire") passant par **p** (printemps), **e** (été), **a** (automne),

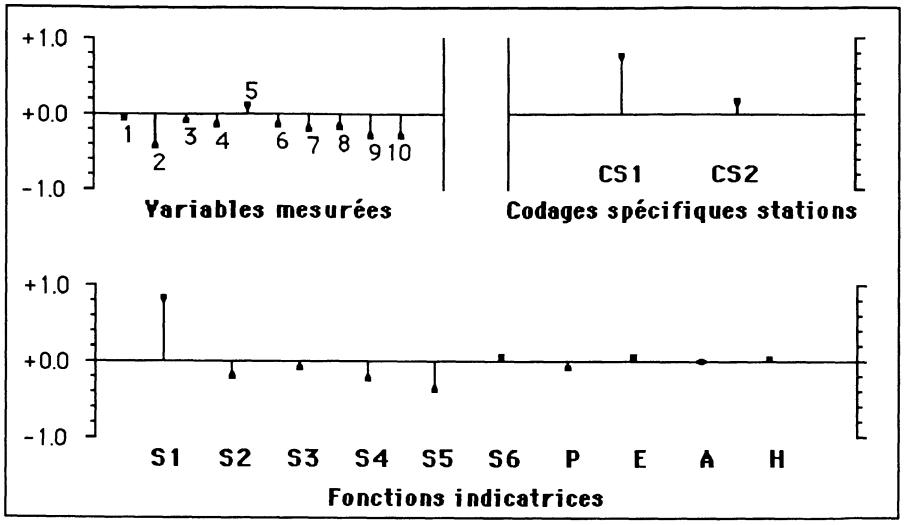

Fig. 2. Représentation des corrélations entre la fonction optimale-stations f_2, et les diverses variables intervenant dans cette analyse (voir légende de la figure 1).

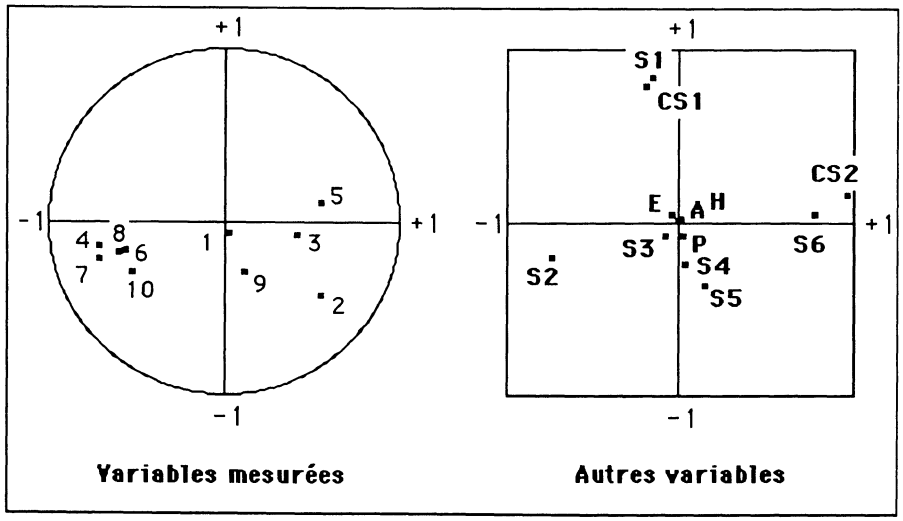

Fig. 3. Représentation des corrélations des diverses variables avec les deux fonctions optimales-stations f_1 et f_2 (axe horizontal : f_1, axe vertical : f_2). Pour la clarté du graphique, on a présenté séparément les variables mesurées $x_1, ..., x_{10}$, participant au calcul de f_1 et f_2, et dont les représentations sont dans le cercle de corrélation, et les autres variables, supplémentaires, dont les représentations sont dans le carré de corrélation.

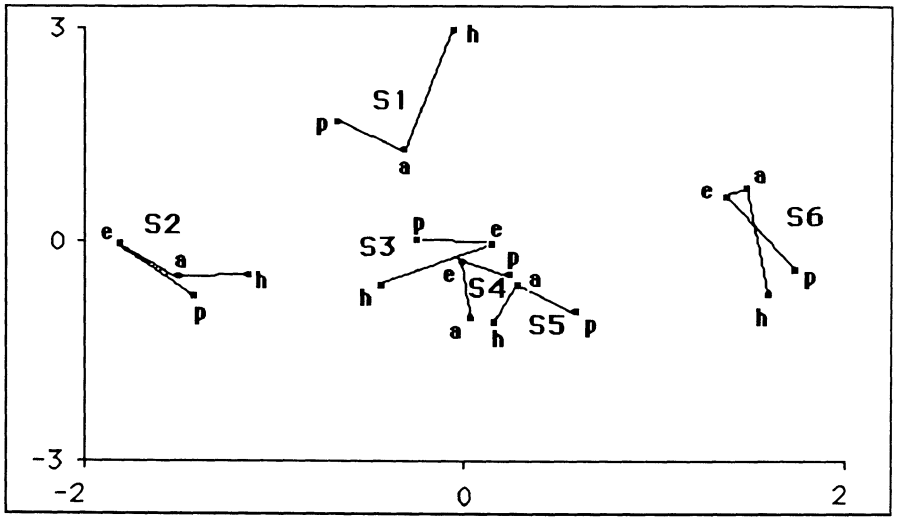

Fig. 4. Représentation des observations dans le plan des deux fonctions optimales-stations f_1 (axe horizontal) et f_2 (axe vertical). Les observations relatives à une même station ont été reliées entre elles par une "trajectoire" le long de laquelle les saisons sont désignées par **p**, **e**, **a**, **h** : l'absence éventuelle de l'une de ces quatre lettres traduit l'absence d'observation (donnée manquante). Commentaire dans le texte.

h (hiver), avec possibilité d'absence de quelques points intermédiaires (cas de données manquantes).

Sur cette carte factorielle, nous ferons les commentaires suivants. Les 6 stations sont très bien individualisées dans ce plan : les trois ou quatre points associés à une station sont très regroupés, et les groupes de points associés à des stations différentes sont nettement séparés; rappelons que cette séparation des stations était l'un des buts recherchés (construction de fonctions les plus corrélées possible avec les stations). D'autre part, la configuration des 6 trajectoires ne permet pas de déceler une disposition systématique particulière attachée à chacune des saisons : l'éventuel "effet saison" global a bien été éliminé (c'est l'autre but recherché : corrélation la plus faible possible avec les saisons), seules subsistent les variations saisonnières particulières à chaque station.

La disposition des 6 stations, interprétée en fonction des corrélations (voir plus haut), fait ressortir S2 comme la station la plus polluée (surtout en été, le point **e** de cette station étant le plus à gauche dans la trajectoire correspondante), et S6 comme la moins polluée (à peu près également en toute saison). Les quatre autres stations occupent des positions intermédiaires, relativement à la pollution, avec un gradient S3→S4→S5 allant vers une moindre pollution. Enfin, la station

S1 occupe une position particulière, ayant les plus fortes valeurs de f_2 (surtout en hiver), ce qui semble lié notamment à un taux élevé d'oxygène dissous.

La recherche de fonctions linéaires des variables x_i optimales pour les saisons, troisième étape de la méthode, consisterait en l'analyse canonique ordinaire entre $[X]$ et $[Z] \cap [Y]\perp$, et serait en tout point semblable à l'étape précédente. Cependant, nous sommes ici en présence d'un cas particulier, celui où $[Z] \cap [Y]\perp$ est réduit au vecteur nul. L'analyse canonique envisagée n'a donc aucun sens, et si l'on s'en tient strictement à la recherche de fonctions φ – orthogonales à tout codage des stations, le problème n'admet aucune solution : il est impossible de trouver une fonction linéaire des 10 variables mesurées, qui soit non corrélée avec certains codages des stations. Nous ne pouvons donc pas nous débarrasser complètement de l'"effet stations", dans l'étude de l'"effet saisons" (phénomène qui nous est imposé ici par les données manquantes). Dans PONTIER et PERNIN [6], nous avions évoqué l'éventualité de cette situation, et envisagé une façon d'y faire face, sans toutefois passer à l'application numérique. L'occasion nous est donnée ici d'expliquer notre démarche.

Faute de pouvoir trouver une fonction linéaire des variables, ayant une corrélation maximum avec les codages des saisons, tout en étant non corrélée avec les codages des stations, nous nous montrerons un peu moins exigeants sur les relations avec les stations, en demandant seulement une fonction linéaire des variables, ayant une corrélation maximum avec les codages des saisons, et une corrélation *minimum* avec les codages des stations (étant admis que nous ne pouvons espérer la nullité de ce minimum). Nous résolvons ce problème en utilisant $[Z]_2$,sous-espace de $[Z]$ constitué par les codages des saisons "obliques" par rapport aux codages des stations. En règle générale, ce sous-espace $[Z]_2$ est engendré par les vecteurs propres (supposés $\varphi-$normés) associés aux valeurs propres strictement comprises entre 0 et 1, dans l'analyse canonique entre $[Y]$ et [Z]. Parmi ces vecteurs propres, sélectionnons celui (ou ceux) associés(s) à la plus petite (ou aux plus petites) valeur(s) propre(s). Les vecteurs propres ainsi sélectionnés engendrent un sous-espace $[Z]'_2$ de $[Z]_2$. Dans ces conditions, l'analyse canonique entre $[X]$ et $[Z]'_2$ nous fournira les fonctions linéaires des x_i répondant à notre objectif initial : corrélation minimum avec les stations, et corrélation maximum avec les saisons.

Ainsi, dans l'exemple numérique étudié, le vecteur v_3 est "presque orthogonal" à $[Y]$, avec une corrélation égale à 0.08675. Si nous choisissons pour $[Z]'_2$ le sous-espace engendré par ce vecteur v_3, alors l'analyse canonique entre $[X]$ et $[Z]'_2$ est, dans ce cas particulier où $[Z]'_2$ est unidimensionnel, la régression linéaire multiple de v_3,variable dépendante, sur $\{x_1, ...x_{10}\}$, régresseurs (la fonction de régression ainsi obtenue doit ensuite être normée à l'unité). Nous pourrions éventuellement considérer aussi le vecteur v_2, mais ce serait moins justifié que pour le vecteur v_3 (la corrélation canonique associée à v_2 étant égale à 0.25820, non négligeable). Nous obtiendrions alors le sous-espace $[Z]'_2 = [v_3, v_2]$,de dimension 2; l'analyse canonique entre $[X]$ et ce $[Z]'_2$ nous fournirait donc des fonctions linéaires des variables mesurées, dont la corrélation avec les codages des stations ne dépasserait pas 0.26 (en valeur absolue), mais serait de toutes façons non nulle.

Dans la poursuite de l'application numérique, nous ne ferons que l'analyse canonique entre $[X]$ et $[v_3]$. Nous obtenons donc un seul couple canonique, de corrélation canonique égale à 0.94073, constitué par la fonction canonique g et le

codage canonique correspondant, qui n'est autre que v_3 :

$g = 25.25111\ \mathbf{1} - 0.13868\ \mathbf{x_1} + 0.00066\ \mathbf{x_2} - 0.29357\ \mathbf{x_3} + 0.00070\ \mathbf{x_4}$
$- 0.00328\ \mathbf{x_5} + 0.02027\ \mathbf{x_6} - 0.06964\ \mathbf{x_7} - 0.00216\ \mathbf{x_8} + 0.00927\ \mathbf{x_9}$
$- 0.00063\ \mathbf{x_{10}}$

L'examen des corrélations entre g et les autre variables intervenant dans cette analyse (voir fig 5), indique des corrélations négatives importantes avec x_1(température) et x_3(pH), une corrélation positive importante avec x_4 (conductivité), et des corrélations relativement faibles avec les autres variables. On notera que les trois variables x_1, x_3, x_4 ne dépendent pas de la pollution; les deux dernières notamment "dépendent essentiellement de la nature du terrain traversé par le cours d'eau" (PEGAZ-MAUCET [4], p. 21).

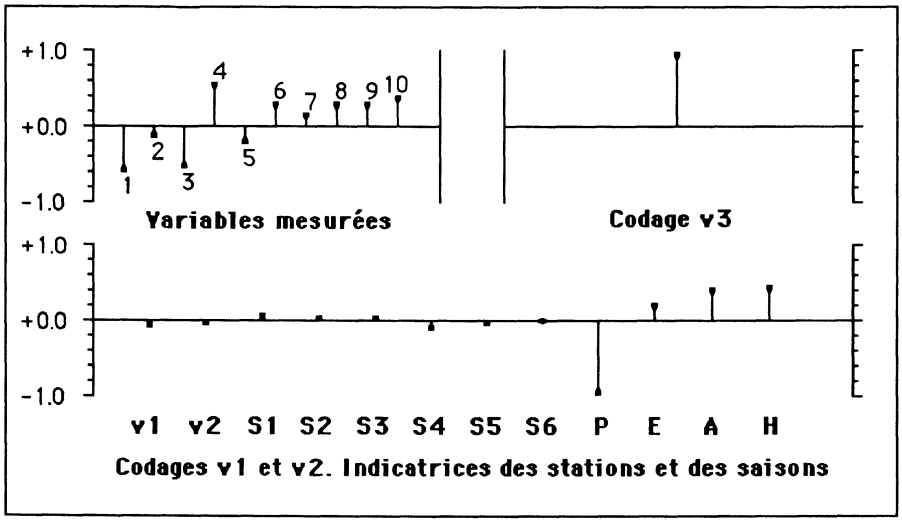

Fig. 5. Représentation des corrélations entre la fonction optimale-saisons **g**, et les diverses variables intervenant dans cette analyse : les 10 variables mesurées $x1$ à $x10$, le codage $v3$, les deux autres codages $v1$ et $v2$, les indicatrices des 6 stations S1 à S6, les indicatrices des 4 saisons P, E, A, H. Commentaire dans le texte.

La corrélation avec le codage v_3 est forte (0.94), alors qu'elle est pratiquement nulle avec chacun des deux autres codages des saisons v_1 et v_2 (ce qui est normal puisque v_3 n'est corrélé ni avec v_1 ni avec v_2). Les corrélations entre g et les indicatrices des 6 stations sont quasi-nulles (but recherché); les corrélations entre g et les indicatrices des 4 saisons sont non nulles, fortement négative pour le printemps, moyennement positives pour les 3 autres saisons. Nous avons d'autre part associé, dans un même graphique (fig 6), la fonction optimale-saisons g, et la première fonction optimale-stations f_1. Pour toutes les stations, le printemps se distingue par sa position nettement différenciée, vers les valeurs fortement négatives de g (en abscisses). Quant aux trois autres saisons, elles sont assez imbriquées

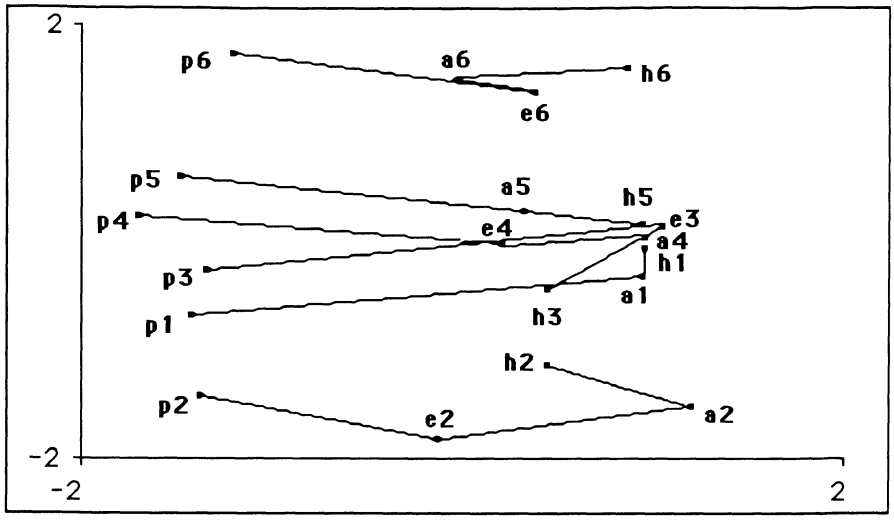

Fig. 6. Représentation cartésienne des 20 observations, selon la fonction optimale-saisons **g** (en abscisses) et la première fonction optimale-stations f_1 (en ordonnées). Chaque observation est désignée par la lettre correspondant à la saison (p, e, a ou h), et par le numéro de la station (1 à 6). Commentaire dans le texte.

les unes dans les autres, leurs positions respectives dépendant des stations. La fonction optimale-saisons g distingue donc bien le printemps de l'ensemble des autres saisons, mais c'est tout; c'est la traduction du fait que la structure factorielle des données (observations avec données manquantes) ne permet pas d'obtenir de codage spécifique des saisons, et donc de fonction optimale-saisons au sens strict.

4 Perspectives : vers une méthode LONGI étendue

La méthode LONGI telle qu'elle a été créée initialement, est un enchaînement de deux analyses canoniques *particulières*, enchaînement appliqué à une situation elle-même *particulière* :

- la particularité de la situation est liée au fait que sur les trois blocs X, Y, Z, deux d'entre eux sont constitués par les indicatrices des modalités de variables qualitatives;
- la particularité de la première analyse canonique est qu'elle est une analyse canonique complète (plus précisément : une analyse des correspondances complètes, selon la définition qui en a été donnée plus haut en 2.1), mais qu'une partie seulement de ses résultats sont utilisés par la deuxième analyse canonique;
- la particularité de la deuxième analyse canonique, est de ne pas être une analyse canonique complète ...

Compte tenu de l'intérêt des propriétés des sous-espaces issus d'une analyse canonique complète, nous proposons, dépassant les particularités soulignées ci-dessus, d'étendre la définition de la méthode LONGI, désignant alors par ce terme la méthode décrite ci-après.

Il s'agit d'une méthode d'analyse de trois ensemble de variables, X, Y, Z observées sur les mêmes unités statistiques. Cette méthode n'est pas symétrique, c'est-à-dire n'est pas indépendante de l'ordre dans lequel sont envisagés les trois ensembles de variables. Ainsi, on notera LONGI(X, Y, Z) la procédure consistant à pratiquer successivement :

1) L'analyse canonique complète entre $[Y]$ et [Z], et obtenir ainsi les sous-espaces $[Y] \cap [Z]$, $[Y]_2, [Z]_2$, $[Y] \cap [Z]\perp$, $[Z] \cap [Y]\perp$, ... Cette première opération, indépendante de l'ensemble de variables $\{X\}$, a pour but de connaître les inter-relations entre les deux ensembles de variables $\{Y\}$ et $\{Z\}$.

2) L'analyse canonique complète entre $[X]$ et les (ou certains des) sous-espaces obtenus précédemment (dans la mesure où ces sous-espaces ne sont pas réduits au vecteur nul). Cette deuxième opération a pour but de préciser les relations entre l'ensemble de variables $\{X\}$ d'une part, et d'autre part les variables des ensembles $\{Y\}$ et $\{Z\}$, compte tenu de certaines relations (par exemple d'orthogonalité) susceptibles d'exister entre elles.

Références

[1] BENZECRI J.P. (1973) L'analyse des données : 2. L'analyse des correspondances (*Paris, Dunod éd.*, 619 p.)

[2] CAILLIEZ F. et PAGES J.P. (1976) Introduction à l'analyse des données (*Paris, S.M.A.S.H. éd.*, 616 p.)

[3] DOLEDEC S., CHESSEL D. (1987) Rythmes saisonniers et composantes stationnelles en milieu aquatiques I. Description d'un plan d'observation complet par projection de variables (*Acta OEcologica/OEcol. Gener.*, **8**, 3, 403-426

[4] PEGAZ-MAUCET D. (1980) Impact d'une perturbation d'origine organique sur la dérive des macro-invertébrés d'un cours d'eau. Comparaison avec le benthos. *Thèse 3e Cycle, Université Claude Bernard, Lyon,* 130 p.

[5] PERNIN M.-O. (1986) Contribution à la méthodologie d'analyse de données longitudinales. Exemple de la croissance chez l'être humain (Auxologie). *Thèse Dipl. Doct., Université Claude Bernard, Lyon,* 264 p

[6] PONTIER J., PERNIN M.-O. (1987) Multivariate and longitudinal data on growing children : solution using LONGI. *Proceedings of the Third Symposium on Data Analysis : the ins and outs of solving real problems, held june 10-12, 1985, in Brussels, Belgium. London, Plenum ed.*, p. 49-65

[7] PONTIER J., JOLICOEUR P., PERNIN M.-O. (1987) Analyse canonique complète. *Statistique et Analyse des Données,* **12,** 1-2, 124-148

[8] SABATIER R. (1987) Méthodes factorielles en analyse des données : approximations et prise en compte de variables concomitantes. *Thèse Doct. ès-Sciences, Univ. Sciences et Techniques du Languedoc, Montpellier,* 242 p

[9] SABATIER R. (1987) Analyse factorielle de données structurées et métriques. *Statistique et Analyse des Données*, 12, 3, 75-96

[10] THIOULOUSE J., CHESSEL D. (1987) Les analyses multitableaux en écologie factorielle. I.- De la typologie d'état à la typologie de fonctionnement par l'analyse triadique. *Acta OEcologica/OEcol. Gener.*, **8**, 4, 463-480