Statistique et analyse des données

Robert Sabatier

Analyse factorielle de données structurées et métriques

Statistique et analyse des données, tome 12, n° 3 (1987), p. 75-96

<http://www.numdam.org/item?id=SAD_1987__12_3_75_0>

Statistique et Analyse des Données
1987 - Vol. 12 n° 3 p.75-96

ANALYSE FACTORIELLE DE DONNEES STRUCTUREES ET METRIQUES

Robert SABATIER

Laboratoire de Physique Industrielle Pharmaceutique
Faculté de Pharmacie, Avenue Charles Flahault
34060 MONTPELLIER CEDEX.

Résumé : *Cet article, veut répondre au problème de la prise en compte d'informations concomitantes dans une A.C.P. La méthodologie proposée utilise deux outils fondamentaux de l'Analyse des Données : les opérateurs et les métriques. La décomposition de l'intertie des unités statistiques par projection sur les sous espaces engendrés par les variables concomitantes induit la présence d'un opérateur de produit scalaire d'un type particulier. L'utilisation d'une métrique optimale donne à l'inertie une valeur proportionnelle à la quantité test de Lawley-Hotelling utilisée dans l'approche traditionnelle de l'Analyse de Variance Multidimensionelle.*

Abstract : *This paper aims to suggest a solution of taking account concomitant information in P.C.A. The method use two basic tools of Data Analysis : operators and metrics. Decomposition of inertia by projection on subspaces spanned by the concomitant variables induce an operator of scalar product of a particular type. Utilisation of an optimal metric gave to the inertia a value very close to those obtained by the Lawley-Hotelling test using in Multivariate Analysis of Variance.*

Mots Clés : *A.C.P., A.C.P.V.I., Métrique, Opérateur, Coefficient RV, Facteurs contrôlés.*

Indices de classification STMA : *06-070, 06-010, 06-050, 08-110.*

Manuscrit reçu le 15 novembre 1986
Révisé le17 novembre 1987

0 - INTRODUCTION

Lors de la collecte des données, dans les sciences expérimentales, le praticien après avoir choisi, à priori, quelles sont les variables à mesurer est souvent amené dans une deuxième étape à faire la distinction entre plusieurs types de variables. En particulier, il différencie celles que l'on qualifie de facteurs contrôlés, dans un sens plus ou moins proche de celui de l'analyse de variance, de celles que l'on appelle, toujours dans la même terminologie, variable dépendante ou à expliquer (J.M. LEGAY [10]).

Les analyses de ce type de données sont assez diverses et consistent en deux approches opposées et contradictoires. Les tenant de la statistique classique, utilisant les analyses du type Analyse de Variance Multidimensionnelle (MANOVA). Analyses réalisés toutefois qu'après avoir fait un choix parmi la batterie des tests disponibles (G.A.F. SEBER [16], W.J. KRZANOWSKI [6][7]).L'utilisation dans ce cadre de la statistique inférentielle ne fournit, en général, que des critères globaux sur les effets des variables, mais en contre partie perd l'identité des u.s.,Pour l'Analyse des Données, telle qu'utilisée, en France la démarche rigoureusement inverse s'impose naturellement. Il en résulte ainsi des traitements statistiques qui exacerbent la variabilité inter-u.s mais qui ne sont que très peu appropriés à intégrer la variabilité due aux facteurs. La technique des points suplémentaires en AFC ou ACP, ainsi que le "coloriage" des u.s. appartenant au même ensemble d'une partition semble être le seul apport méthodologique original de la plus grande part des Analystes de Données. Il existe heureusement quelques travaux qui essaient de prendre en compte cette variabilité dans le cadre de l'AFC (P. CAZES [3], D. CHESSEL [4], J.D. LEBRETON [9]). En général on tient compte parfaitement (?) de la structure, si elle est uni-factorielle en utilisant une Analyse Factorielle Discriminante (J.M.ROMEDER · [12]).

Notre propos ici est donc multiple :

. Enrichir l'ACP de façon à tenir compte de la structure des u.s. pour les structures factorielles précédemment définies, en utilisant cette structure comme une information connue à priori.

. Montrer qu'il est possible, dans certains cas, de choisir des ACP qui décomposent les critères traditionnels de MANOVA en choisissant des métriques appropriées.

Dans une première partie, nous allons montrer, sur un exemple, comment une ACP peut être insuffisante pour appréhender l'effet d'une structure factorielle sur les u.s.,La deuxième partie introduit l'Analyse en Composantes Principales par rapport à des Variables Instrumentales. Il est montré comment cette analyse permet d'introduire des opérateurs de produits scalaires entre u.s. pour tenir compte d'une structure. La troisième partie introduit une nouvelle ACP par rapport à un modèle factoriel. Dans les cas où cette structure est bifactorielle, on retrouve des conditions nécessaires et suffisantes, bien connues en Analyse de Variance, pour que ce plan factoriel soit orthogonal. Enfin, le choix d'une métrique rendant minimale l'inertie intra-factorielle permet de trouver une composition de l'inertie du nuage des u.s. décomposant les critères de Lawley-Hotelling utilisés dans MANOVA. La dernière partie applique notre méthodologie à l'exemple initial.

1 - UN EXEMPLE DE DONNEES STRUCTUREES

L'exemple qui servira d'application pour les techniques proposées au paragraphe 3, nous a été soumis par D. Chessel [5]. Il s'agit d'une étude portant sur la pollution de la Meaudret dans le Vercors. Les 10 variables mesurées par les expérimontateurs sont :

1 - La température de l'eau (T) en °C.
2 - Le débit (D) en l/s.
3 - Le pH (PH).
4 - La conductivité (C) en $\mu s/cm/cm^2$.
5 - L'oxygène dissoud (O2) en %.
6 - La demande biochimique en oxygène (DBO) en mg/l d'oxygène.
7 - La demande chimique en oxygène (DCO) en mg/l d'oxygène.
8 - L'azote ammoniacal (NH4) en mg/l.
9 - L'azote nitrique (NO3) en mg/l.
10 - Les orthophosphates (PO4) en mg/l.

De plus les mesures (Y) ont été effectuées sur 6 stations réparties le long du cours d'eau et pour 4 dates : juin, août, novembre, février. Il est donc clair que les données à analyser entrent dans la catégorie des données structurées précédemment définies. Dans la suite de ce travail les

deux facteurs influençant seront appelés facteur station (à 6 modalités) et facteur saison (à 4 modalités). On ne tiendra pas compte des interactions possible , dans un sens qu'il serait bon toutefois de préciser, entre les deux facteurs puisque l'on a pas de répétition. La matrice des données possède donc 10 colonnes (les variables) 24 lignes (les unités statistiques) et deux facteurs influençant. Les données sont fournies en annexe.

Dans le but de décrire le plus simplement possible le corpus des données, nous allons effectuer une ACP sur les données centrées réduites. La matrice de corrélation entre les 10 variables est donnée dans le tableau suivant :

	T	D	PH	C	O2	DBO	DCO	NH4	NO3	PO4
T	1									
D	-0,108	1								
PH	-0,127	0,187	1							
C	-0,079	-0,323	-0,716	1						
O2	-0,215	0,354	0,682	0,568	1					
DBO	0,091	-0,255	-0,638	0,757	-0,697	1				
DCO	0,110	-0,264	-0,594	0,765	-0,713	0,947	1			
NH4	0,143	-0,309	-0,745	0,805	-0,768	0,963	0,914	1		
NO3	-0,116	-0,286	0,078	0,173	0,144	-0,198	-0,250	-0,132	1	
PO4	0,062	-0,330	-0,647	0,855	-0,646	0,875	0,814	0,908	0,229	1

Sans entrer dans trop de détails, on peut constater que la structure des corrélations entre variables est assez complexe. En effet : les coefficients de corrélations varient de -0,768 à 0,963, on constate que la variable T est peu corrélée avec les autres variables, par contre DBO et NH4 sont très corrélées avec toutes les autres.

Les pourcentages d'inertie expliquée étant de 57,4 % pour le premier axe et 14,3 % pour le second (et 10,8 % pour le troisième) nous ne donnerons les représentations des unités statistiques (Figure 1) et des variables (Figure 2) que pour le premier plan (71,7 % d'inertie expliquée). Les unités statistiques sont repérées par un identificateur à deux caractères : le premier identifie la modalité du premier facteur (station de 1 à 6) le deuxième celle du deuxième facteur (saison de 1 à 4). Une lecture simultanée des deux graphiques permet de noter (ainsi que la lecture des aides à l'interprétation traditionnelle en ACP, non fournies ici pour ne pas alourdir inutilement l'exposé).

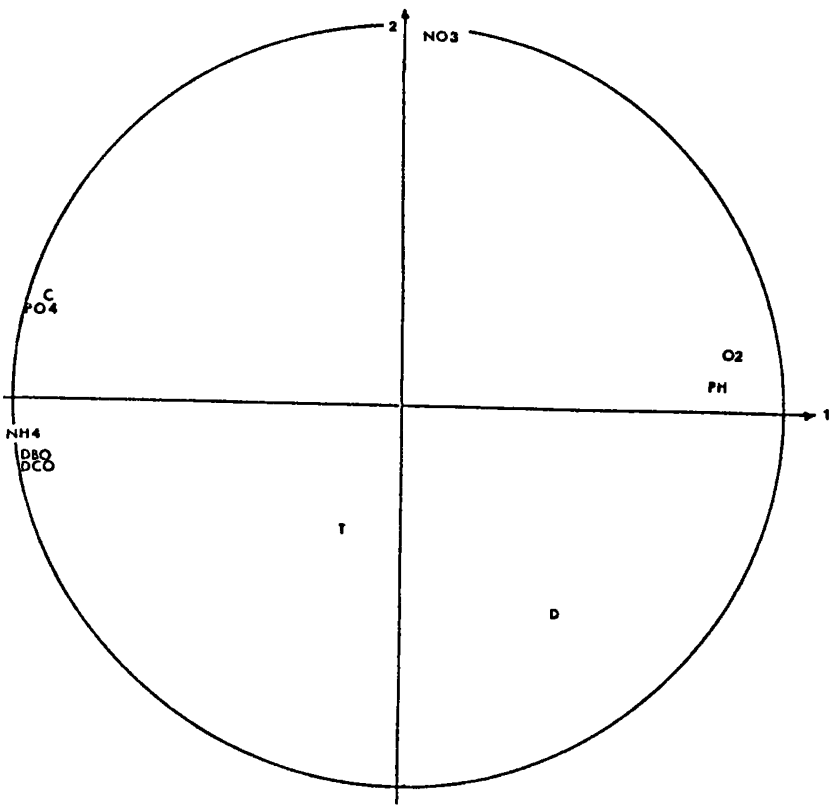


Figure 2 : Représentation des variables dans l'ACP de (Y, Id, D)

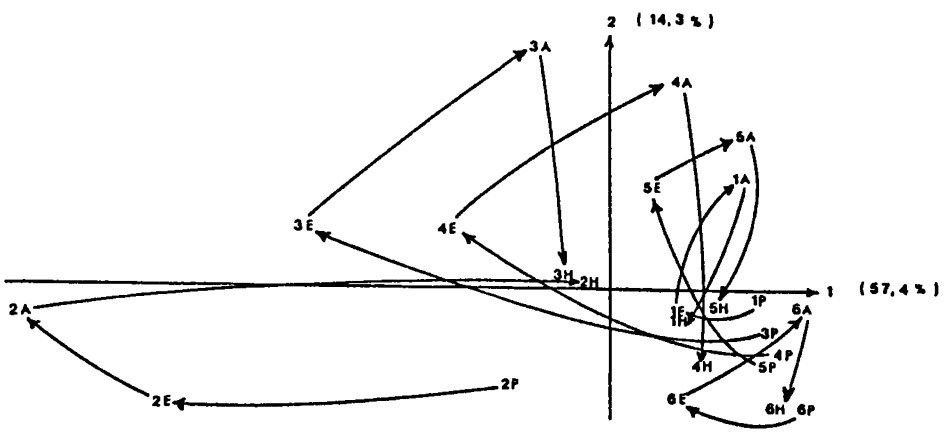


Figure 1 : Représentation des unités statistiques dans l'ACP de (Y, Id, D)

- La station 2 est la plus polluée, particulièrement en été et en automne (NH4, DBO, DCO élevés) .
- Les stations 1 et 6 sont les moins polluées (PH et O2 élevés) .
- Les stations 3 à 5 sont intermédiaires vis à vis de la pollution .
- La variable NO3 semble avoir des valeurs plus fortes en automne ?
- T et D semblent ne pas être influencés par la pollution ?

Toutefois il y a clairement un effet saisonnier pour chaque station (trajectoires visualisées) qu'il serait bon d'éliminer . De même qu'il serait intéressant de déterminer si une partie de la pollution mesurée ne peut être imputée ni à un effet station ni à un effet saison .

En conclusion l'A.C.P. précédente tout en décrivant globalement les influences des facteurs station et saison sur la pollution ne nous permet pas de déceler réellement leur influence . Des analyses plus spécifiques s'imposent .

2 - L'ANALYSE EN COMPOSANTES PRINCIPALES PAR RAPPORT A DES VARIABLES INSTRUMENTALES

L'analyse en composantes principales par rapport à des variables instrumentales ne sera ici que brièvement rappelée, ce qui nous permettra également d'introduire le formalisme, qui est assez proche de celui de F.CAILLIEZ et J.P.PAGES [2] . Le lecteur voulant en savoir plus sur l'ACPVI consultera avec profit C.R. RAO [11] R.SABATIER [13] et [14] ou L.BONIFAS [1] .

2.1 - Un projecteur dans l'espace vectoriel des opérateurs autoadjoints

Soit σ (F) l'espace vectoriel des opérateurs autoadjoints d'un espace vectoriel euclidien F . On rappelle que σ (F) peut être muni d'une structure euclidienne par le produit scalaire suivant

$$(A,B) \in \sigma(F) \times \sigma(F) \quad ; \quad < A,B > = tr(AB)$$

On notera $||\ ||$ la norme déduite de ce produit scalaire . Un

élément de σ (F) est dit opérateur . Le coefficient Rv entre deux opérateurs A et B de σ (F) est défini par :

$$Rv\ (A,B) = \frac{< A,B >}{||A|| \cdot ||B||} \quad .$$

Définition et proposition 1

Soient G un sous espace vectoriel (s.e.v.) de F, P_G l'opérateur de projection orthogonale sur G . Posons :

$$\pi_G \begin{cases} \sigma\ (F) \longrightarrow \sigma\ (F) \\ A \longrightarrow P_G\ A\ P_G \end{cases}$$

alors :

i) π_G est un projecteur orthogonal de σ (F)

ii) $A \in Ker\ (\pi_G) \iff A\ (G) \subset G^{\perp}$

iii) $A \in Im\ (\pi_G) \iff G^{\perp} \subset Ker\ A$.

Démonstration

i) . Clairement si $A \in \sigma\ (F)$ alors $P_G\ AP_G \in \sigma\ (F)$. Soit $A \in \sigma\ (F)$, $\pi_G^2\ (A) = P_G\ \pi_G\ (A)\ P_G = P_G\ AP_G = \pi_G\ (A)$, donc π_G est un projecteur .

Montrons que π_G est orthogonal

soient $A \in Ker\ (\pi_G) \iff P_G\ AP_G = 0$,

$B \in Im\ (\pi_G) \iff P_G\ BP_G = B$

alors $< A,B > = tr\ (AB) = tr\ (A\ P_G\ BP_G) = tr\ (P_G\ AP_G\ B) = 0$

donc $Ker\ (\pi_G) = (Im\ (\pi_G))^{\perp}$.

ii) ($\Leftarrow$) évident . Montrons ($\Rightarrow$) . Soit $A \in Ker\ (\pi_G) \iff P_G\ AP_G = 0$

soit $x \in A(G) \iff \exists\ y \in G$ tel que $x = A\ (y)$ donc

$P_G\ (x) = P_G\ (\ A(y)\ ) = P_G\ A\ P_G\ (y) = 0$ donc $x \in G^{\perp}$.

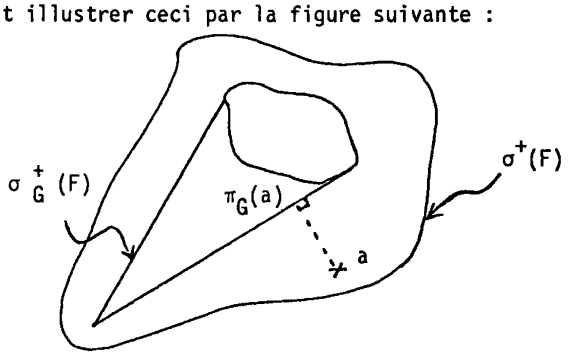
iii) ($\Rightarrow$) soit $A \in \mathrm{Im}(\pi_G) \Leftrightarrow P_G A P_G = A$. Donc $\forall x \in G^{\perp}$, $P_G(x) = 0$
c'est à dire $P_G A P_G = A$ donc $x \in \mathrm{Ker}(A)$
($\Leftarrow$) $G^{\perp} \subset \mathrm{Ker}\, A \Leftrightarrow (\mathrm{Ker}\, A)^{\perp} = \mathrm{Im}\, A \subset G$. Soit $x \in F$
alors $P_G A P_G (x) = A P_G (x)$
si $x \in G$ alors $P_G(x) = x$ donc $P_G A P_G (x) = A(x)$
si $x \in G^{\perp} \subset \mathrm{Ker}\, A$ alors $P_G(x) = 0 = A(x)$ donc $P_G A P_G (x) = A(x) = 0$.

□

Une conséquence triviale de ce théorème est la solution du problème :
Soit $a \in \sigma^{+}(F)$ (sous ensemble de $\sigma(F)$ des opérateurs positifs de F), alors

$$\text{Min}\ ||\, a-b\, || = ||a-\pi_G(a)|| \text{ où } \sigma_G^{+}(F) = \{ b \in \sigma^{+}(F) \text{ et } \mathrm{Im}(b) \subset G \}$$
$b \in \sigma_G^{+}(F)$

On peut illustrer ceci par la figure suivante :



2.2 - Définition de l'A.C.P.V.I.

Soient (X_1, Q_1, D) un triplet statistique réalisé sur n unités statistiques (u.s.) et p variables avec $n > p$. La matrice des données X_1 (n x p) est supposée centrée pour les poids définissant la matrice diagonale des poids des u.s. D . On note $W_1 D = X_1 Q_1 {}^{t} X_1 D$ l'opérateur de $F = \mathbb{R}^n$ des proximités entre u.s. Soit X_2 un tableau (n x q) de mesures réalisées sur les mêmes n u.s. . On suppose X_2 centré pour D . On notera F_2 le s.e.v. de F engendré par les colonnes de X_2 : $F_2 = \mathrm{Im}(X_2)$;

. On appelle A C P V I d'ordre k de X_2 par rapport à (X_1, Q_1, D) l' ACP d'ordre k de (X_2, $\widetilde{Q}_2$, D) où $\widetilde{Q}_2$ est une semi-métrique telle que $|| W_1 D - X_2 Q_2 {}^t X_2 D ||$ soit minimale ($1 \leqslant k \leqslant q$) .

. On a : $|| W_1 D - X_2 Q_2 {}^t X_2 D ||^2 = || W_1 D - \pi_{F_2} (W_1 D) ||^2 + || \pi_{F_2} (W_1 D) - X_2 Q_2 {}^t X_2 D ||^2$ (1)

où $P_{F_2} = X_2 \ S_{22}^- \ {}^t X_2 D$ le projecteur D-symétrique sur F_2 ,

S_{22}^- indique l'inverse généralisé de Moore-Penrose de S_{22} .

. Si l'on pose $\widetilde{Q}_2 = S_{22}^- \ S_{21} Q_1 S_{12} \ S_{22}^-$ alors $\pi_{F_2} (W_1 D) = X_2 \widetilde{Q}_2 {}^t X_2 D$ donc l'ACP de (X_2, $\widetilde{Q}_2$, D) est équivalente à celle de ($P_2 X_1$, Q_1, D) et $|| W_1 D - \pi_{F_2} (W_1 D) ||^2 = ||W_1 D ||^2 \ (1 - Rv^2 (W_1 D, \pi_{F_2} (W_1 D)))$.

. On note que rg $(\widetilde{Q}_2) \leqslant$ min (p,q) . Si S_{22} est inversible $\widetilde{Q}_2$ est solution unique, si S_{22} n'est pas inversible toute métrique $Q_2 = \widetilde{Q}_2 + Q^*$ avec $Q^* \subset$ Ker (X_2) vérifie aussi la propriété de minimum .

3 - A C P ET MODELE FACTORIEL

Dans cette partie nous allons utiliser directement les résultats du paragraphe précédent, pour définir une A C P par rapport à un " modèle factoriel " . Le terme factoriel n'est présent que pour rappeler que nous faisons une A C P par rapport à un modèle défini par un ou des facteurs .

Définition

Soient F_M un s.e.v. de F, (Y, Q, D) un triplet statistique où Y est une matrice de données (n x p) (avec n > p) centrée pour la matrice D des poids des n u.s. . On appelle A C P d'ordre k ($1 \leqslant k \leqslant$ min (dim F_M, p))

de (Y,Q,D) par rapport à F_M (noté ACP de $(Y,Q,D)/F_M$), l'ACP d'ordre k de Π_{F_M} (WD) (noté par abus Π_{F_M}) avec $WD = XQ^tXD$.

. De la même façon, on peut définir l'ACP par rapport à $F_M^{\perp}$, où $F_M^{\perp}$ est le s.e.v. de F, D-orthogonal à F_M.

. En fait, on peut donner une autre définition d'une ACP par rapport à un sous espace : l'ACP d'ordre k de $\Pi_{Y}(P_{F_M})$ (ou de façon symétrique celle de $\Pi_{F_M}(P_Y)$. Clairement cette ACP est l'Analyse Canonique de F_M et F_Y. Dans le cadre d'une étude sur les données structurées, nous pensons qu'une telle ACP n'est pas intéressante car nous recherchons l'influence de la structure sur les données. Ainsi les méthodes capables d'apporter, nous semble t-il, des résultats pertinents semblent être des analyses faisant intervenir le tableau de données et la structure de façon "non symétrique" : l'ACPVI par exemple.

Dans la suite de ce travail, nous allons nous intéresser aux choix particuliers de F_M et de Q, selon le type de structure. Nous allons énoncer dans une première étape un lemme technique qui, appliqué dans la proposition suivante va nous montrer que l'orthogonalité d'un plan à deux facteurs structurants est une simple propriété géométrique. La propriété suivante exhibera les propriétés de notre A.C.P., par rapport aux différents sous espaces dans le cas de deux facteurs orthogonaux. Nous conclurons enfin par un lemme qui nous fournira une métrique donnant une décomposition optimale de l'inertie.

. On peut noter que cette analyse, dans le cas où Y ne possède qu'une seule variable permet de retrouver tous les résultats de l'Analyse de Variance unidimensionnelle exprimée en termes de RV.

. Dans le cas où la structure n'est pas de type factoriel mais de contiguité (i.e. les u.s. sont situées au sommet d'un graphe non orienté) on peut montrer que notre approche rejoint celle proposée par L. LEBART [8] (voir R. SABATIER [15]).

Lemme 1

Soient F_1 et F_2 deux s.e.v. de F, P_1 et P_2 les projecteurs D symétriques sur F_1 et F_2 . Alors on a les équivalences suivantes :

i) $F_1 \subset F_2$

ii) $P_1 P_2 = P_2 P_1 = P_1$

iii) $P = P_2 - P_1$ est le projecteur D-symétrique sur $F_2 \cap F_1^{\perp}$.

La démonstration est laissée au soin du lecteur.

Notations

Soient A et B deux "facteurs" (partition d'u.s.) à I (I > 1) et J (J > 1) modalités (sous ensembles), A_k (resp B_ℓ) la $k^{\text{ième}}$ (resp $\ell^{\text{ième}}$) modalité de A (resp B), n_k (resp n_ℓ ; $n_{k,\ell}$) le poids (supposé non nul) de la modalité A_k (resp B_ℓ ; $A_k \cap B_\ell$) , U_A (resp U_B) le tableau des indicatrices des I (resp J) modalités du facteur A (resp B) mesuré sur les n u.s., $F_A = I\,m\,(U_A)$ et $F_B = I\,m\,(U_B)$.

Dans le cas où le modèle est défini par un seul facteur, alors on peut écrire $F = \Delta_1 \dot{\oplus} (F_A \cap \Delta_1^{\perp}) \dot{\oplus} F_A^{\perp}$ où $\dot{\oplus}$ signifie somme directe D-orthogonale, Δ_1 droite de F engendrée par $1_F = \sum_{i=1}^{n} f_i$ avec $B_F = \{ f_i / i = 1, 2, \ldots, n \}$ base naturelle de F . Il est alors simple de définir les deux ACP par rapport aux deux modèles : $F_A \cap \Delta_1^{\perp}$ et $F_A^{\perp}$.

Proposition 2

Si l'on pose $F = F_0 \oplus F_0^{\perp}$ où $F_0 = \Delta_1 \oplus (F_A \cap \Delta_1^{\perp}) \oplus (F_B \cap \Delta_1^{\perp})$

Alors on a les équivalences suivantes :

i) $F_0 = \Delta_1 \dot{\oplus} (F_A \cap \Delta_1^{\perp}) \dot{\oplus} (F_B \cap \Delta_1^{\perp})$,

ii) $P_A P_B = P_B P_A = P_1$,

iii) $<U_A^k, U_B^l>_D = \| U_A^k \|_D \| U_B^l \|_D$ où U_A^k (resp U_B^l) est la $k^{ième}$ (resp $l^{ième}$) variable de U_A (resp U_B) et $<U_A^k, U_B^l>_D$ le produit scalaire dans F, au sens de D, des vecteurs U_A^k et U_B^l,

Démonstration

i) $\Leftrightarrow$ ii) Il est clair que la démonstration repose sur l'équivalence suivante $(F_A \cap \Delta_1^{\perp}) \perp (F_B \cap \Delta_1^{\perp}) \Leftrightarrow (P_A - P_1)(P_B - P_1) = 0$. L'équation précédente développée nous donne $P_A P_B = P_1$, car $1 \in F_A \cap F_B$ implique $P_1 P_A = P_B P_1 = P_1$. L'égalité $P_B P_A = P_1$ se démontre de façon analogue .

ii) $\Leftrightarrow$ iii) Un calcul simple permet de vérifier que :

$$P_A(f_j) = \sum_{k=1}^{I} \frac{1}{n_k} <U_A^k, f_i>_D U_A^k = \frac{p_i}{n_k} U_A^k \text{ si } i \text{ appartient à } A_k$$

Donc $<P_A P_B (f_i), f_i>_D = <P_B(f_i), P_A(f_i)>_D =$

$$\frac{p_i^2}{n_k n_l} <U_A^k, U_B^l>_D \text{ et } <P_1(f_i), f_i>_D = p_i^2 .$$

□

Quand l'une de ces trois conditions est vérifiée on dit que " le plan factoriel A, B" est orthogonal (d'après i) . En général les n u.s. ont le même poids $p_i = \frac{1}{n}$, alors la condition iii) précédente prend la forme plus connue :

$$\text{Card}(A_k \cap B_l) = \frac{\text{Card}(A_k) \cdot \text{Card}(B_l)}{n}$$ que l'on prend en général comme définition de l'orthogonalité d'un plan factoriel . Il est clair que, dès que l'une des trois conditions de la proposition est vérifiée, F_0 est le s.e.v. de F engendré par $F_A \cup F_B$ et d'après le lemme $P_{F_0} = P_A + P_B - P_1$.

On peut montrer aussi que les trois conditions précédentes sont équivalentes à $Rv\left(P_{F_A \cap \Delta_1^{\perp}} , P_{F_B \cap \Delta_1^{\perp}} \right) = 0$.

Proposition 3

Sous les hypothèses de la définition précédente, on a :

i) $Rv\left(\pi_{F_M} , \pi_{F_M^{\perp}} \right) = 0$ et $||\pi_{F_M} - \pi_{F_M^{\perp}}||^2 = ||\pi_{F_M}||^2 + ||\pi_{F_M^{\perp}}||^2$.

ii) Si $I(F')$ est l'inertie du nuage des u.s. par rapport à F' s.e.v. de F, alors : $I(F_M) + I(F_M^{\perp}) = I(F)$.

$I(F_0) = I(F_A) + I(F_B)$ si le plan A,B est orthogonal.

iii) l'ACP d'ordre k de (Y, S_{yy}^{-1} , D) / F_A ou $S_{yy} = {}^tYDY$, est équivalente à l'analyse discriminante de Y/A .

iv) Dans le cas où le plan A,B est orthogonal, l'ACP d'ordre k de (Y, Q, D)/ $F_0^{\perp}$ est l'ACP d'ordre k de $(Y - U_A G_A - U_B G_B, Q, D)$ où G_A (resp G_B) est la matrice dont la $j^{ème}$ ligne est le c.d.g. des u.s. qui prennent la modalité A_j (resp B_j) du facteur A (resp B) .

Démonstration

i) $Rv\left(\pi_{F_M} , \pi_{F_M^{\perp}} \right) = \dfrac{< \pi_{F_M} , \pi_{F_M^{\perp}} >}{||\pi_{F_M}||.||\pi_{F_M^{\perp}}||}$, $<\pi_{F_M} , \pi_{F_M^{\perp}}> =$

$tr\left(P_{F_M} WD P_{F_M} P_{F_M^{\perp}} WD P_{F_M^{\perp}} \right)$ et $P_{F_M} P_{F_M^{\perp}} = 0$, la deuxième égalité en découle également .

ii) $I(F_M) = tr(\pi_{F_M}) = tr(P_{F_M} WD P_{F_M})$. Or $I_F = P_{F_M} + P_{F_M^{\perp}}$

donc $I(F) = tr(WD) = tr((P_{F_M} + P_{F_M^{\perp}}) WD (P_{F_M} + P_{F_M^{\perp}}) =$

$I(F_M) + I(F_M^{\perp})$ car $< \pi_{F_M} , \pi_{F_M^{\perp}} > = 0$. Si le plan A,B est

orthogonal, on a vu que $P_{F_0} = P_A + P_B - P_1$, donc
$I(\Gamma_0) = tr\,(\,(P_A + P_B - P_1)\,WD\,(P_A + P_B - P_1)\,) = tr\,(\,P_A\,WD\,P_A\,) +$
$tr\,(\,P_A\,WD\,P_B\,) + tr\,(\,P_B\,WD\,P_A\,) + tr\,(\,P_B\,WD\,P_B\,)$, car Y centré est équivalent à $Im\,(Y) \subset \Delta_1^{\perp}$. Or $tr\,(\,P_A\,WD\,P_B\,) = tr\,(\,P_B\,P_A\,WD\,) = tr\,(\,P_1\,WD\,) = 0$

iii) L'analyse discriminante de Y/Λ est l'ACP de $(\,G_A,\ S_{yy}^{-1},\ D_A\,)$ (Cailliez-Pages [2]) avec $G_A = D_A^{-1}\,{}^t U_A\,DY$. La recherche des facteurs associés à ce triplet s'effectue par la recherche des valeurs propres de $S_{yy}^{-1}\,{}^t G_A D_A G_A = S_{yy}^{-1}\,Y\,D\,P_A Y$ qui est identique à celle de $P_A P_Y$ où $\pi_{F_A}(P_Y)$

iv) évident car $G_A = D_A^{-1}\,{}^t U_A DY$ donc $U_A G_A = P_A Y$ c'est à dire que l'opérateur de produit scalaire est égal à π_{F_0} . □

Le point ii) montre donc que si le plan A,B est orthogonal alors, on a $I(F) = I(F_A) + I(F_B) + I(F_0^{\perp})$ (2), c'est à dire que l'on peut décomposer l'inertie totale du nuage selon chaque facteur et selon $F_0^{\perp}$ (i.e. "le résidu") ce qui fourni une aide à l'interprétation supplémentaire .

A la suite de la proposition 3, il ne reste qu'à choisir la métrique Q . Le choix le plus simple est $Q = Id_p$, avec cette métrique l'ACP par rapport à F_A est l'ACP sur les centres de gravités définis par les classes de A, celle par rapport à $F_A^{\perp}$ l'ACP sur les écarts aux centres de gravités, et ainsi de même pour les autres ACP .

En fait, un choix raisonnable de Q peut être effectué de la façon suivante : dans la décomposition (2) précédente l'inertie résiduelle $I(F_0^{\perp})$, c'est à dire orthogonale au plan A,B peut etre choisie, sous une contrainte, minimale.

Lemme 2

Soit $(\,X_1,\ Q_1,\ D\,)$ un triplet statistique défini sur p variables, alors :

$$\text{Min } tr\,(V\,Q_1) = p\,(\det V)^{\frac{1}{p}}$$
$$\det(Q_1) = 1$$
$$\text{et le minimum est atteint pour } Q^* = (\det V)^{\frac{1}{p}}\,V^{-1}$$

Démonstration

l'application $\begin{cases} \ell(E, E^*) \longrightarrow R^* \\ Q_1 \longrightarrow tr(VQ_1) \end{cases}$ est continue,

l'ensemble des métriques de déterminant 1 est compact donc le minimum existe et est unique . On sait que $VQ \in \ell(E)$ est diagonalisable de valeurs propres réelles $\lambda_1 > \lambda_2 > \ldots > \lambda_p$. Or $tr(VQ_1) = \sum_{i=1}^{p} \lambda_i$ et $\det(VQ_1) = \det(V) \det(Q_1) = \det(V)$ qui sera posé égal à α donc $\det(VQ_1) = \prod_{i=1}^{p} \lambda_i = \alpha$. La recherche du minimum de $\sum_{i=1}^{p} \lambda_i$ sans la contrainte $\prod_{i=1}^{p} \lambda_i = \alpha$ est obtenue pour $\lambda_i = \alpha^{1/p}$; $i = 1,2, \ldots, p$. Donc $VQ = \alpha^{\frac{1}{p}} Id_p$ d'où $Q^* = (\det V)^{\frac{1}{p}} V^{-1}$.

□

On a vu que l'ACP de (Y,Q,D) / $F_0^{\perp}$ était équivalente à celle de $(Y - U_A G_A - U_B G_B, Q, D)$, dont on notera V_0 la matrice de variance associée. La métrique Q^* optimale déduite du lemne précédent est donc $Q^* = (\det V_0)^{\frac{1}{p}} V_0^{-1}$. Nous allons voir que cette métrique nous permet de retrouver des résultats classiques de MANOVA . Dans ce contexte de statistique inférentielle, les tests d'égalité des moyennes, par exemple, pour le facteur A, s'effectuent (entre-autre) par le critère de Lawley-Hotelling (voir SEBER [16]) : $tr(V_A V_0^{-1})$ où V_A est la matrice de variance ${}^t G_A D_A G_A$. C'est à dire, au coefficient $(\det V_0)^{\frac{1}{p}}$ près, l'ACP de (Y, V_0^{-1}, D) / F_A (ou de $(U_A G_A, V_0^{-1}, D)$) est une décomposition du critère de Lawley-Hotelling . C'est sous cette forme mais dans un autre contexte que G.S. SEBESTYEN [17] a proposé cette métrique.

4 - RETOUR A L'EXEMPLE INITIAL

Pour illustrer ce qui précéde nous allons effectuer les ACP qui décomposent le critère de Lawley-Hotelling, en reprenant l'exemple initial.

* ACP de (Y, V_0^{-1}, D) / F_{ST} .

L'inertie associée à cette ACP est de 38,6, la probabilité de rejet de l'hypothèse nulle de non influence du facteur station est de p = 0,0014 . Les pourcentages d'inertie associés aux deux premiers axes sont de 57,7 % et 28,5 % . La figure 3 représente les u.s. (stations) dans le premier plan principal . On constate que le rejet global de l'hypothèse nulle est dû à deux effets : opposition station 1 à station 6 (axe 1) et opposition stations 1 et 6 et station 2 . La représentation des variables que l'on obtient par cette ACP ne sera pas fournie car elle est très difficilement interprétable, comme pour toutes les ACP avec métriques quelconques. Par contre, la figure 4, nous fournit le cercle des corrélations entre les variables (moyennes par station) et les composantes principales . On note que le premier axe est dû essentiellement aux variables température et débit, le deuxième axe met bien en évidence les variables témoins de pollution .

* ACP de (Y, V_0^{-1}, D) / F_{SA} .

L'inertie associée à cette ACP est de 184,2, la probabilité de rejet de l'hypothèse nulle de non influence du facteur saison est inférieure à 0,0001 . Les pourcentages d'inertie associés aux deux premiers axes sont de 82,9 % et 12,7 % . La figure 5 représente les u.s. (saisons) dans le premier plan principal . On note que le premier axe est une opposition entre le printemps et l'été d'une part et l'hiver d'autre part . La représentation des variables, (corrélations composantes, variables) montre que le premier axe est très corrélé avec la variable Température, et que celle-ci s'oppose à l'oxygène dissou (figure 6) .

* Conclusions .

Une telle analyse, pour être menée à bien, devrait considérer en plus les quatre ACP suivantes : par rapport à $F_{ST}^{\perp}$, $F_{SA}^{\perp}$, F_0 et $F_0^{\perp}$. Nous ne donnerons, pas faute de place les sorties des ACP mais nous résumerons le tout par le tableau 2 suivant, qui décompose l'inertie totale .

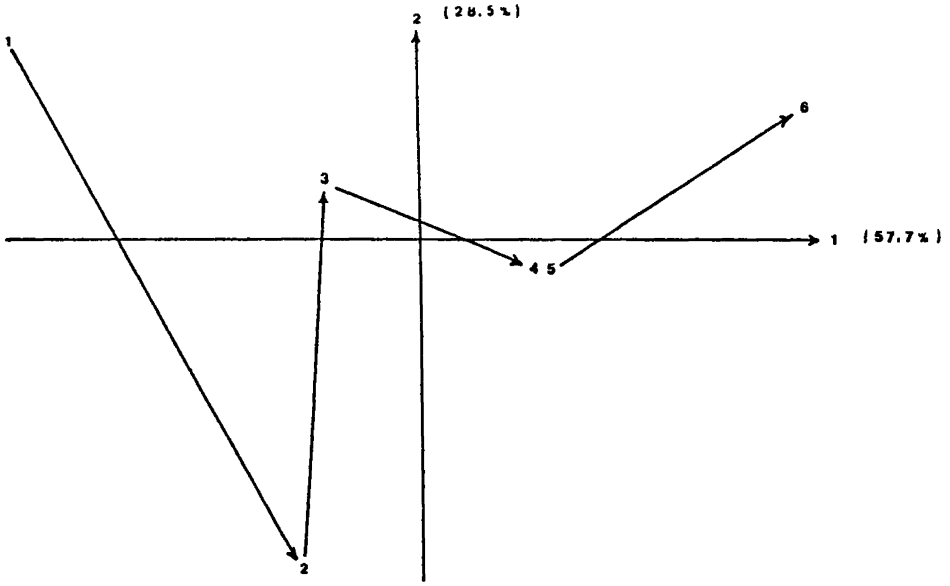


Figure 3 : Representation des u.s. (stations) dans l'ACP de $(Y, V_0^{-1}, D) / F_{ST}$

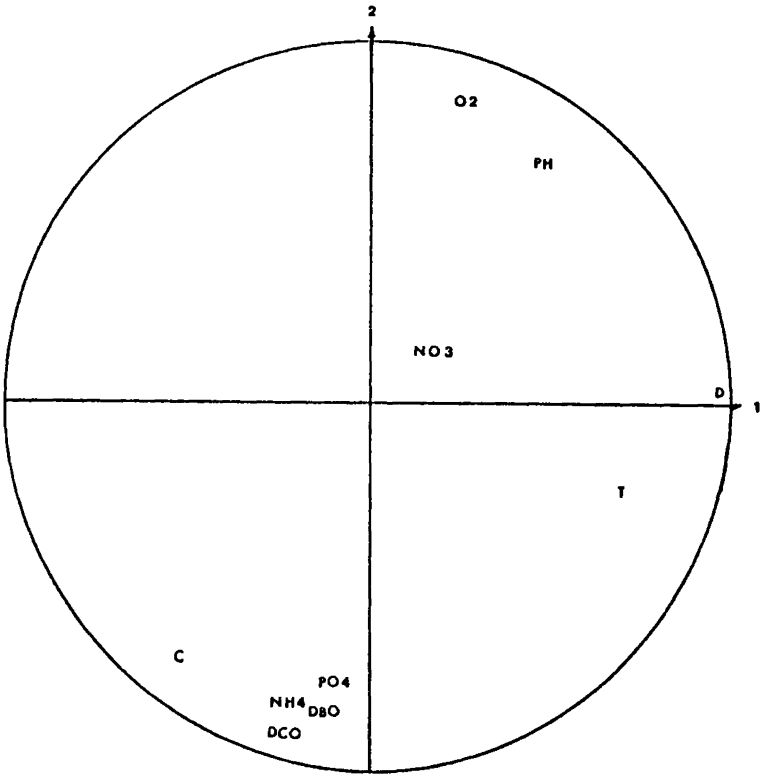


Figure 4 : Représentation des variables dans l'ACP de $(Y, V_0^{-1}, D) / F_{ST}$

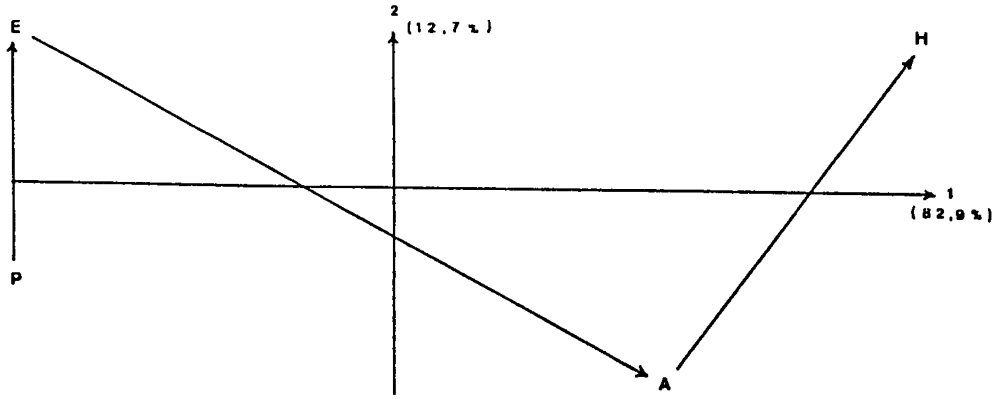


Figure 5 : Représentation des u.s. (saisons) dans l'ACP de (Y, V_0^{-1}, D) / F_{SA}

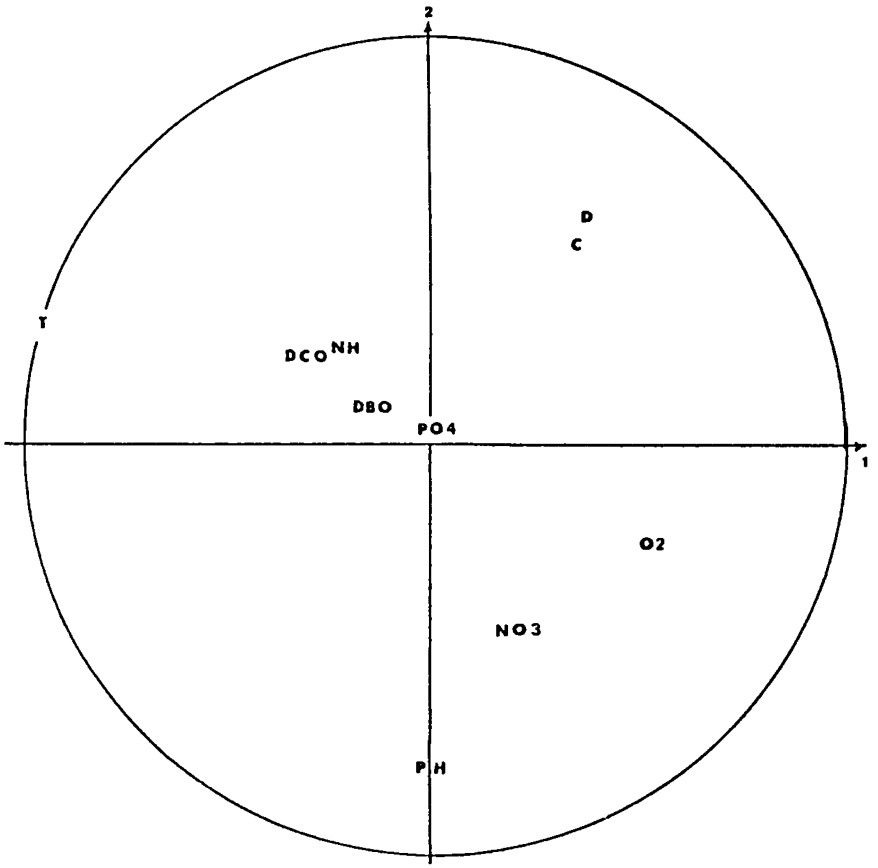


Figure 6 : Représentation des variables dans l'ACP de (Y, V_0^{-1}, D) / F_{SA}

TABLEAU 2

Décomposition de l'inertie

	Totale	F_{ST}	$F_{ST}^{\perp}$	F_{SA}	$F_{SA}^{\perp}$	F_0	$F_0^{\perp}$
Inertie	232,8	36,8	194	184,2	48,6	222,8	10
Pourcentage d'inertie	100 %	16,6 %	83,3 %	79,1 %	20,9 %	95,7 %	4,3 %

Ce tableau montre que la variabilité la plus importante est dûe au facteur saison. Le choix de la métrique de Mahalanabis associée au résidu pose un problème pour sa décomposition, puisque l'ACP fournit 10 composantes de même inertie, c'est-à-dire une isotropie totale de l'espace $F_0^{\perp}$ (le seul argument positif est que son inertie relative est de 4,3 %).

5 - CONCLUSIONS

Nous avons montré que l'utilisation de l'ACP avec une métrique adéquate permet de décomposer l'inertie d'un nuage d'unités statistiques suivant plusieurs facteurs structurant, dans le cas où ces facteurs étaient orthogonaux, au sens de l'analyse de la variance. De plus cette métrique définie sur les variables qualitatives définissant chaque facteur est dépendante de celle choisie à priori sur les variables quantitatives mesurées, ce qui permet dans certains cas de montrer que l'ACP décompose des critères statistiques connus.

On peut trouver dans la thèse de R. SABATIER [15], d'autres applications de cette méthodologie. Les cas de structure envisagés sont les suivants : facteurs non orthogonaux, facteurs avec interaction, facteurs hiérarchiques ainsi que prise en compte de covariables. Des applications à l'AFC ont aussi été proposées.

ANNEXE

Tableau des Données

(d'après DOLEDEC et al. [5])

T	D	PH	C	O2	DBO	DCO	NH4	NO3	PO4	ST	SA
10.00	41.00	8.50	295.00	110.00	2.30	1.40	0.12	3.40	0.11	1	P
13.00	62.00	8.30	325.00	95.00	2.30	1.80	0.11	3.00	0.13	1	E
1.00	25.00	8.40	315.00	91.00	1.60	0.50	0.07	6.40	0.03	1	A
3.00	118.00	8.00	325.00	100.00	1.60	1.20	0.17	1.80	0.19	1	H
11.00	158.00	8.30	315.00	13.00	7.60	3.30	2.85	2.70	1.50	2	P
13.00	80.00	7.60	380.00	20.00	21.00	5.70	9.80	0.80	3.65	2	E
3.00	63.00	8.00	425.00	38.00	36.00	8.00	12.50	2.20	6.50	2	A
3.00	252.00	8.30	360.00	100.00	9.50	2.90	2.52	4.60	1.60	2	H
11.00	198.00	8.50	290.00	113.00	3.30	1.50	0.40	4.00	0.10	3	P
15.00	100.00	7.80	385.00	46.00	15.00	2.50	7.90	7.70	4.50	3	E
2.00	79.00	8.10	350.00	84.00	7.10	1.90	2.70	13.20	3.70	3	A
3.00	315.00	8.30	370.00	100.00	8.70	2.80	2.80	4.80	2.85	3	H
12.00	280.00	8.60	290.00	126.00	3.50	1.50	0.45	4.00	0.73	4	P
16.00	140.00	8.00	360.00	76.00	12.00	2.60	4.90	8.40	3.45	4	E
3.00	85.00	8.30	330.00	106.00	2.00	1.40	0.42	12.00	1.60	4	A
3.00	498.00	8.30	330.00	100.00	4.80	1.60	1.04	4.40	0.82	4	H
13.00	322.00	8.50	285.00	117.00	3.60	1.60	0.48	4.60	0.84	5	P
15.00	160.00	8.40	345.00	91.00	1.70	1.90	0.22	10.00	1.74	5	E
2.00	72.00	8.60	305.00	91.00	1.60	0.90	0.10	9.50	1.25	5	A
2.00	390.00	8.20	330.00	100.00	1.70	1.20	0.56	5.00	0.60	5	H
11.00	303.00	8.50	245.00	100.00	1.70	0.90	0.05	2.70	0.16	6	P
13.00	310.00	8.20	285.00	82.00	8.50	1.60	0.59	3.70	0.60	6	E
4.00	181.00	8.60	270.00	105.00	2.80	0.50	0.10	3.66	0.43	6	A
3.00	480.00	8.20	290.00	100.00	1.30	0.80	0.04	2.20	0.13	6	H

BIBLIOGRAPHIE

[1] BONIFAS, L. ; ESCOUFIER, Y. ; GONZALES, PL. ; SABATIER, R. ; "Choix de variables en analyse en composantes principales" , R.S.A. XXXII, 1984, n°2, p 5-15.

[2] CAILLIEZ, F. ; PAGES, JP. ; Introduction à l'analyse des données. Smash, 9 rue Duban, 75016 PARIS, 1979.

[3] CAZES, P. ; CHESSEL, D. ; DOLEDEC, S. ; "l'Analyse des correspondances internes d'un tableau partitionné : son usage en hydrobiologie". R.S.A. sous presse.

[4] CHESSEL, D. ; LEBRETON, JD. ; YOCCOZ, N. ; "Propriétés de l'analyse canonique des correspondances ; une illustration en hydrobiologie". R.S.A. sous presse.

[5] DOLEDEC, S. ; CHESSEL, D. ; "Rythmes saisonniers et composantes stationnelles en milieu aquatique. 1 - Description d'un plan d'observation complet par projection de variables". Oecal. Gener. Sous presse.

[6] KRZANOWSKI, WJ. ; "Between-groups comparison of principal components", JASA, Vol 74, 1979, n°367, p 703-707.

[7] KRZANOWSKI, WJ. ; "Between-groups comparison of principal components", Some sampling results". J. Statist. Comput. Simul., 1982, Vol 15, p 141-154.

[8] LEBART, L. ; "Analyse statistique de la contiguité", Publ. Inst. Stat. Univ. Paris, 1969, XVIII, p 81-112.

[9] LEBRETON, JD. ; CHESSEL, D. ; PRODON, R. ; YOCCOZ, N ; "L'analyse des relations espèces-milieu par l'analyse canonique des correspondances. I. Variables de milieu quantitatives". Soumis pour publication à Oecologia Generalis.

[10] LEGAY, JM. ; "Quelques réflexions sur le plan expérimental". S.A.D. Vol 11, n°4, 1986, p 51-57.

[11] RAO, CR. ; "The use and the interpretation of principal component analysis in applied research". Sankya. Ser. a, 26, 1964, p 320-359.

[12] ROMEDER, JM. ; Méthodes et programmes d'analyse discriminante. Dunod. 1973. Paris.

[13] SABATIER, R. ; "Approximation d'un tableau de données, application à la reconstitution des polioéléments". Thèse de 3ême cycle. USTL. Montpellier 1983.

[14] SABATIER, R. ; "Quelques généralisations de l'analyse en composantes principales de variables instrumentales". S.A.D. Vol 9, n°3, 1984, p 75-103.

[15] SABATIER, R. ; "Méthodes factorielles en Analyse de Données : approximations et prise en compte de variables concomitantes". Thèse d'Etat. USTL. Montpellier 1987.

[16] SEBER, G.A.F. ; Multivariate observations ; John Wiley 1984.

[17] SEBESTYEN, GS. ; Decision making process in pattern recognitian. 19, The Mac Millan Company.