REVUE DE STATISTIQUE APPLIQUÉE

J. PICARD

Classification des profils évolutifs incomplets et asynchrones

Revue de statistique appliquée, tome 35, nº 2 (1987), p. 27-37
<http://www.numdam.org/item?id=RSA_1987__35_2_27_0>

Rev. Statistique Appliquée, 1987, **XXXV** (2), 27-38

CLASSIFICATION DES PROFILS ÉVOLUTIFS INCOMPLETS ET ASYNCHRONES

PICARD J.

CRSSA — Service de Biomathématiques et Informatique Générale
108, boulevard Pinel, 69275 LYON CEDEX 03

RÉSUMÉ

Une méthode de classification de profils évolutifs est proposée. Elle met à profit la notion de produit scalaire et de distance euclidienne entre des opérateurs associés à chaque profil. Les opérateurs utilisés sont de type covariance, mais leur estimation tient compte de la relation d'ordre existant entre les diverses observations d'un sujet. L'introduction du temps parmi les variables utilisées permet de conserver l'information relative à la croissance ou à la décroissance des variables. La constitution de ces opérateurs montre qu'il est possible de classer des profils incomplets et/ou asynchrones, c'est-à-dire dont les dates d'observations sont variables. Un jeu d'essai montre la robustesse de l'approche et ses limites; une application sur des profils évolutifs de nouveaux-nés est présentée.

Mots clés : *Classification, Profils évolutifs, Opérateurs de covariance, Données manquantes.*

Introduction

L'analyse de l'évolution multivariée d'individus se ramène à l'étude des structures d'un tableau tridimensionnel dont les trois indices correspondent respectivement aux individus, aux variables et au temps.

Les méthodes utilisées pour ces études reposent essentiellement sur deux types d'approches factorielles :

— La première approche a été proposée par LEBART [5] sous le nom d'analyse factorielle locale (AFL) qu'il a appliqué à des contiguités spatiales. Indépendamment, LE FOLL [6] introduit l'analyse factorielle des évolutions (AFE) puis en fait une généralisation sous le nom d'analyse factorielle pondérée (AFP). Dans un article récent CARLIER [2] reprend l'AFE et montre que c'est une ACP sur un tableau d'observations associé à un graphe descriptif des contiguités, comme l'avait fait LEBART pour l'AFL.

— La deuxième approche est celle de la méthode STATIS développée par l'HERMIER DES PLANTES et ESCOUFIER [3] dont un prolongement très intéressant de FOUCART [4] aboutit à l'analyse factorielle des opérateurs (AFO).

En plus de ces problèmes structuraux, il se pose fréquemment celui de l'homogénéïté des individus sur le plan de leurs profils évolutifs : existe-il des classes d'évolution parmi les individus ? Lorsque le tableau de données est complet et que toutes les observations successives sont réalisées aux mêmes dates pour tous les individus (observations complètes et synchrones), il est possible d'utiliser les outils classiques de classification sur les profils.

Mais, en biométrie et en particulier en biométrie humaine, il est fréquent que les observations relevées ne soient pas synchrones ou que le nombre d'observations par individu ne soit pas constant : le thérapeute n'est pas maître de la disponibilité du patient. Dans ce cas, les observations sont nécessairement asynchrones. Le simple rejet des sujets incomplets ou asynchrones peut entraîner des pertes de cas très importantes : les outils habituels ne sont plus adaptés.

La méthode proposée permet de répondre à ce genre de situation en limitant la perte d'informations car les individus, mêmes incomplets ou asynchrones, ne sont pas rejetés : seules la première et la dernière observation doivent avoir été réalisées aux mêmes dates sur l'ensemble des individus afin de garantir la même période de référence pour tous les individus. La méthode fait appel aussi bien à la notion d'inertie des trajectoires (profils évolutifs) au sens de LE FOLL qu'à celui de produit scalaire d'HILBERT-SMITH entre opérateurs-individus.

1. Méthodologie

1.1. Le tableau de données

Sur chaque individu indicé par i :

$$i \in I, \quad I = \{1 ,..., N\}$$

on dispose de K_i observations indicées par k_i à des dates diverses sur P variables réelles indicées par j :

$$j \in J, \quad J = \{1 ,..., P\}$$

On rappelle que les dates pour $k_i = 1$ et $k_i = K_i$ sont les mêmes pour tous les individus quel que soit le nombre d'observations K_i du sujet i pour assurer sur tous les individus la même période d'observation.

On peut considérer que chaque individu i est décrit par un sous-tableau X_i de K_i observations sur P variables. L'ensemble des informations permet de construire une suite de N sous-tableaux X_i.

1.2. Les opérateurs de covariance sujet

Chaque individu est représente dans $\mathbb{R}^P$ par un sous-nuage dynamique de K_i points : $\mathcal{N}_i$. Si M est la métrique dans $\mathbb{R}^P$ et D_i la matrice diagonale des poids des observations, c'est-à-dire la métrique dans $\mathbb{R}^{K_i}$, l'opérateur de covariance O_i dans $\mathbb{R}^P$ associé à l'individu i s'écrit :

$$O_i = MV_i$$

ou l'élément générique de V_i est :

$$V_i(j, j') = \sum_{k_i=1}^{K_i} p_{k_i} \, {}^t(x^j_{ik_i} - \bar{x}^j_i)(x^{j'}_{ik_i} - \bar{x}^{j'}_i)$$

$x^j_{ik_i}$ est la valeur pour l'individu i de la variable j pour la $k_i^{\text{ème}}$ observations, $\bar{x}^j_i$ la moyenne des $x^j_{ik_i}$ et p_{k_i} le poids de l'observation k_i. On choisit en général :

$$p_{k_i} = D_i(k_i, k_i) = \frac{1}{k_i}$$

V_i est donc la matrice de covariance intra-sujet et elle est symétrique semi-définie et positive. Dans cette estimation de V_i, on ne tient pas compte de la relation d'ordre entre les observations par le temps.

Pour tenir compte de cette relation, on considère que les sujets ne sont plus représentés par un sous-nuage, mais par une trajectoire $\mathcal{T}_i$ ou profil évolutif. Cela revient à substituer à V_i la matrice U_i qui, en AFE, minimise la déformation de la trajectoire lors de sa projection sur les plans principaux.

$$U_i(j, j') = \sum_{k_i=2}^{K_i} p_{k_i} \, {}^t(x^j_{ik_i} - x^j_{ik_i-1})(x^{j'}_{ik_i} - x^{j'}_{ik_i-1})$$

Le nombre de sommation correspond au nombre d'arcs (intervalles) du profil évolutif : il convient donc de prendre pour p_{k_i} :

$$p_{k_i} = \frac{1}{K_i - 1}$$

U_i est une matrice symétrique semi-définie et positive par construction et O'_i est l'opérateur de type covariance locale correspondant dans $\mathbb{R}^p$:

$$O'_i = MU_i$$

Il est possible de prendre en compte la nature réelle du temps simplement en modifiant le poids de l'observation (ou plutôt de l'arc). Soit t_{k_i} la date de l'observation k_i et t_{K_i} la date de la dernière observation, on pose alors :

$$p_{k_i} = (t_{k_i} - t_{k_i-1})/(t_{K_i} - t_1)$$

Si sur la période de référence et l'ensemble des individus les observations sont synchrones et équi-réparties dans le temps, alors :

$$t_{k_i} = \text{cte}, \qquad k_i = \text{cte}; \qquad \forall i$$

et les deux systèmes de pondération sont équivalents.

On montrerait facilement que les vecteurs propres de O'_i sont les axes d'évolution du sujet i, c'est-à-dire les axes selon lesquels les profils sont les mieux représentés c'est-à-dire moins déformés. L'AFE n'est autre que la recherche des éléments propres de l'opérateur :

$$\sum_{i \in I} p_i O'_i$$

où p_i est le poids de chaque individu.

1.3. Distances de HILBERT-SMITH entre opérateurs O_i (ou O'_i)

On sait [4] que les opérateurs définis plus haut appartiennent à un espace vectoriel de dimension P^2. En munissant cet espace d'un produit scalaire, il devient euclidien et on pourra calculer la distance euclidienne entre les opérateurs et par voie de conséquence entre les sujets, même si les observations sont incomplètes et asynchrones.

Si u_l est un vecteur d'une base de IR^P, le produit scalaire de HILBERT-SMITH de O_i et $O_{i'}$ sera :

$$\psi_{HS}(O_i, O_{i'}) = \sum_{l \in P} \varphi\,[O_i(u_l), O_{i'}(u_l)]$$

où φ est le produit scalaire dans IR^P.

Les opérateurs O_i étant symétriques, leur produit scalaire de H.S. est égal à la trace de leur produit de composition :

$$\psi_{HS}\,(O_i, O_{i'}) = \mathrm{Tr}\,(O_i \,.\, O_{i'})$$

La matrice des distances euclidiennes entre opérateurs s'obtient classiquement par :

$$d^2_{HS}(O_i, O_{i'}) = \psi_{HS}(O_i, O_i) + \psi_{HS}(O_{i'}, O_{i'}) - 2\psi_{HS}(O_i, O_{i'})$$

Dès lors, toutes les méthodes de classification sur tableau de distances sont applicables.

1.4. Forme des profils évolutifs

La représentation d'un individu i par un opérateur de covariance entraîne une certaine perte de l'information car la covariance entre deux variables renseigne sur leur covariation relative, mais pas sur le caractère croissant ou décroissant de leur variation dans le temps.

Pour une classification efficace des profils évolutifs, il convient d'introduire cette information dans l'opérateur de covariance en y ajoutant une référence dynamique : le temps répond parfaitement à cette nécessité. Les opérateurs O_i ou O'_i sont donc calculés en ajoutant aux variables évolutives à étudier la date de chaque observation comme variable supplémentaire. Les covariances variables évolutives/temps permettront de conserver au mieux l'information de la forme des profils et en particulier du caractère croissant ou décroissant de leurs variables constitutives en fonction du temps.

On notera que l'utilisation du temps comme variable réelle banalise le temps par rapport aux autres variables et permet de tenir compte de l'aspect continu du temps.

1.5. Choix de la métrique M dans IR^P

La métrique M dans IR^P doit être évidemment du même type pour tous les opérateurs à comparer.

La métrique canonique n'est envisageable que si les variables sont homogènes pour une date donnée. Lorsqu'il n'en est pas ainsi, il convient de centrer et de réduire les variables.

Si on considère comme équivalent des individus ayant leurs variables correspondantes proportionnelles, c'est-à-dire si seule la forme des profils est à prendre en considération, alors le centrage se fera par rapport au barycentre de chaque profil et la réduction par l'écart-type des observations du seul sujet. Le calcul de l'opérateur O'_i se fera sur le sous-tableau X'_i dérivé de X_i :

$$x'^{j}_{ik_i} = \frac{x^{j}_{ik_i} - \bar{x}^{j}_{i}}{s^{j}_{i}}$$

où $x^{j}_{ik_i}$ est la valeur de la $j^{ème}$ variable à la date k_i pour le sujet i, $\bar{x}^{j}_{i}$ est la valeur moyenne de la $j^{ème}$ variable pour le sujet i et s^{j}_{i} est l'écart-type de cette même variable pour ce même sujet.

Au contraire, si on veut prendre en compte le niveau des profils, la transformation des variables se fera sur l'ensemble des observations tous sujets confondus :

$$x''^{j}_{ik_i} = \frac{x^{j}_{ik_i} - \bar{x}^{j}}{s^{j}}$$

où $\bar{x}^{j}$ est la moyenne de la variable j sur l'ensemble des sujets et s^{j} l'écart-type correspondant.

1.6. Biais de l'estimation des opérateurs

1.6.1. Profils incomplets

On peut montrer que l'estimation de la covariance locale de deux variables corrélées avec le temps est dans le cas d'observations manquantes, affectée d'un biais multiplicatif α de l'ordre de :

$$\alpha \leqslant \frac{1 + 2\,m}{1 - m}$$

où m est le nombre relatif d'observations manquantes. On voit que pour $m = 1/2$, α peut atteindre la valeur 4 ce qui est énorme.

Si on introduit l'opérateur de corrélation locale O''_i tel que :

$$O''_i\,(j, j') = \frac{O'_i\,(j, j')}{\sqrt{O'_i\,(j, j)\; O'_i\,(j', j')}}$$

où $O'_i\,(j, j')$ est la covariance locale pour les variables j et j' et $O'_i\,(j, j)$ la variance locale de la variable j, alors le biais devient négligeable car les estimations du numérateur et du dénominateur sont affectées d'un biais du même ordre de grandeur.

1.6.2. Asynchronisme

Sous les mêmes conditions de corrélation des variables avec le temps, si d est le nombre relatif d'observations décalées du temps relatif θ, il y a introduction d'un biais relatif β de la forme :

$$\beta \leqslant 1 + 2\, d\theta^2$$

Ici β reste en général faible, ainsi pour $d = 1/2$ et $\theta = 1/5$, β vaut 1.04. En fait, pour la robustesse de la classification, ce sont les éventuels biais sur les distances euclidiennes qui interviennent et selon les ψ_{HS}-angles entre les opérateurs, les biais ont tendance à diminuer plus ou moins.

En pratique, dès que le nombre relatif d'observations manquantes dépasse 20 %, on devra choisir soit l'opérateur de covariance globale avec perte de la notion de chronocité du temps, soit l'opérateur de corrélation locale.

1.7. Evaluation de la méthode

Un jeu de données artificielles a été constitué pour tester le logiciel et évaluer le comportement de la méthode. Douze sujets décrits par trois variables réelles à six dates se répartissent a priori en deux classes : l'une regroupant les individus de 1 à 6 possède des variables croissantes dans le temps, l'autre regroupant les individus de 7 à 12, des variables décroissantes.

Les opérateurs O'_i de type covariance sur les trois variables sont calculés sur les douze sous-tableaux individus complets, les distances inter-opérateurs sont estimées et la matrice de distances soumise à une classification ascendante hiérarchique avec, comme dans la suite, minimisation de la variance intra-classes; le dendrogramme est donné figure 1.a. La classification obtenue n'est pas celle attendue mais si on la reprend en ajoutant la date des observations comme quatrième variable, on retrouve bien la classification attendue : figure 1.b. Dans la suite, les quatre variables sont conservées.

Afin de tester la robustesse de la technique, nous avons utilisé la même méthode sur des sous-tableaux à quatre observations fortement asynchrones avec $\beta \simeq 2$ (figure 2), puis sur des sous-tableaux ayant trois ou six observations (figure 3), c'est-à-dire sur des profils très incomplets avec $\alpha \simeq 4$. Dans les deux cas, on a pratiqué des classifications sur les opérateurs O'_i de covariance locale (a) et sur les opérateurs O''_i de corrélation locale (b).

Si on considère les classifications obtenues avec les covariances locales, on constate que l'asynchronisme ne modifie pas de façon importante les partitions, par contre, dans le cas des sous-tableaux fortement incomplets, la classification échoue totalement. En reprenant les classifications avec les opérateurs de corrélation locale dans tous les cas on obtient la répartition théorique en deux classes mais on constate néanmoins que les deux classes de la figure 3.b (profils incomplets) sont moins bien séparées que dans la figure 2.b (profils asynchrones).

Comme le prévoit l'étude des biais, la méthode semble relativement robuste mais se révèle plus sensible aux observations manquantes qu'aux observations asynchrones.

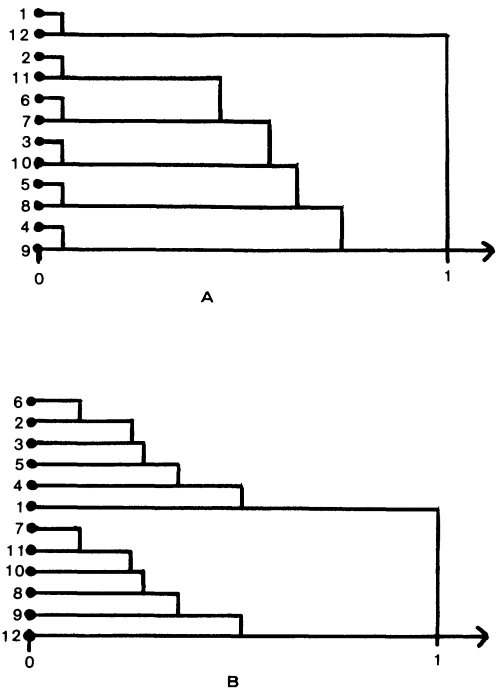

FIGURE 1

Hiérarchie sur données artificielles. Opérateurs de covariances locales O'_i.

A) Sans la date comme variable.

B) Avec la date comme variable.

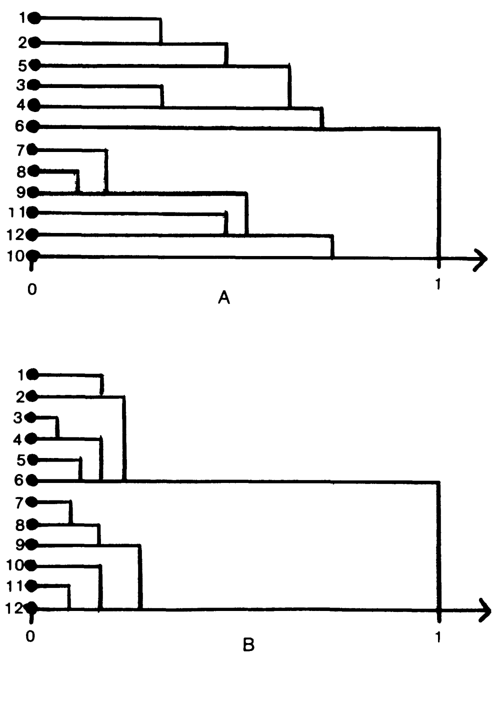

FIGURE 2

Hiérarchie sur données artificielles. Quatre observations asynchrones.

A) Opérateurs de covariances locales O'_i.

B) Opérateurs de corrélations locales O''_i.

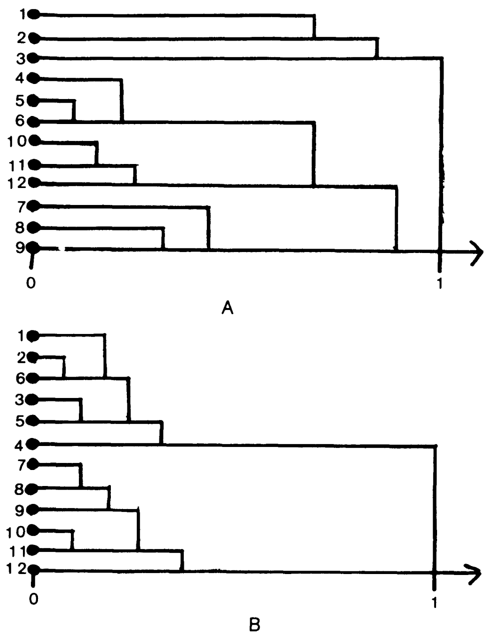

FIGURE 3
Hiérarchie sur données artificielles. Profils incomplets (six et trois observations).
A) Opérateurs de covariances locales O'_i.
B) Opérateurs de corrélations locales O''_i.

2. Application

L'application présentée concerne l'évolution anthropométrique de nouveaux-nés caucasiens sur un an. Sur 63 nouveaux-nés, on a mesuré neuf variables anthropométriques : poids, taille, périmètres du crâne, du thorax et du bras, plis cutanés tricipitaux et sous-scapulaires à 15 et 60 secondes. De plus, 13 variables fonctionnelles ont été calculées à partir des variables mesurées : indice pondéral, plis dynamique absolus et relatifs, tricipitaux et sous-scapulaires, compartiments du bras : surface de muscle, de graisse et d'eau et enfin, le liquide extra-cellulaire et la graisse du corps.

Ces déterminations ont été faites à des dates variant de la naissance au premier anniversaire et le nombre d'observations était compris entre 5 et 10 observations (le mode du nombre d'observations était de 7). Les dates (approximatives) les plus fréquentes étaient : naissance, 15 jours, 1, 4, 6, 9, 12 mois, mais avec des variations parfois assez grandes pour les observations intermédiaires.

Une analyse factorielle des évolutions a été pratiquée au préalable permettant de dégager quatre axes d'évolution expliquant ensemble 84 % de l'inertie totale. Nous dirons simplement que ces facteurs représentent : E1 la croissance

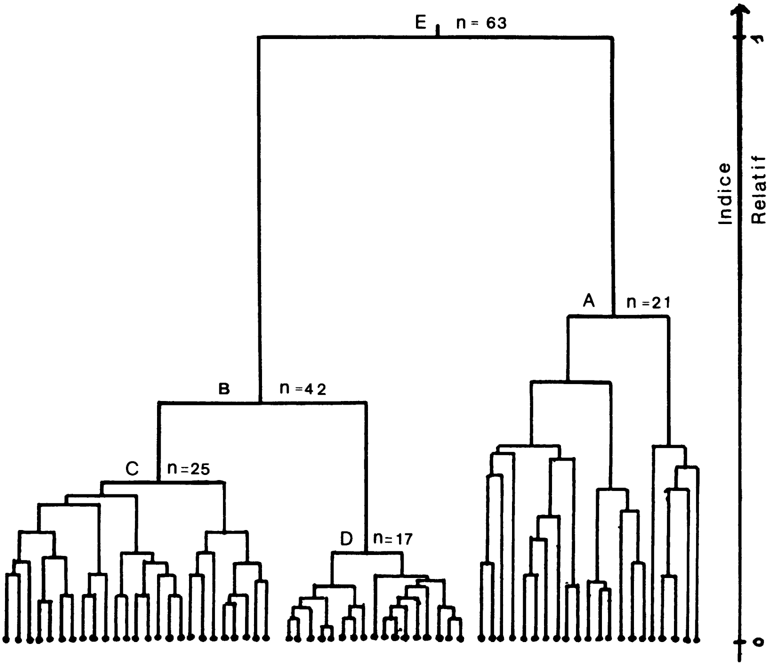

FIGURE 4
Dendrogramme de la hiérarchie de partitions obtenue par classification d'opérateurs de covariances locales O'_i sur 63 enfants de 0 à 1 an.

générale, E2 la déshydratation des tissus sous-cutanés, E3 un facteur différentiel entre les plis tricipitaux et sous-scapulaires et E4 la maigreur. Nous avons voulu savoir s'il existait des classes d'évolution, c'est-à-dire si la population était homogène vis-à-vis de leur évolution anthropométrique.

Ici, la forme et le niveau des profils évolutifs étaient à prendre en compte, nous avons donc procédé à une normalisation des variables sur l'ensemble des observations des 63 sujets. Comme indiqué plus haut, la date de chaque observation était introduite comme variable de référence et les arcs entre les observations étaient pondérés par l'intervalle de temps relatif correspondant. Chaque individu était représenté par l'opérateur de covariance locale O'_i.

La hiérarchie de partition obtenue (Fig. 4) par une classification hiérarchique ascendante avec minimisation de la variance intra-classe montre deux classes distinctes A et B comprenant respectivement 21 et 42 sujets. La classe B semble hétérogène et pouvoir se séparer en deux classes C et D de 17 et 25 sujets. On peut noter l'apparante homogénéité de la classe C sur le critère des indices relatifs de regroupements.

L'interprétation de ces classes a été recherchée globalement par le calcul de la participation des variables à la variance des classes (8) et, localement à date fixe par des analyses discriminatives de Fisher tant sur les variables évolutives que sur les facteurs E1 ... E4.

De ces analyses, il ressort que les classes présentent les mêmes phénomènes évolutifs, mais avec des intensités différentes : croissance générale continue, fabrication de graisse assez rapide jusqu'à un mois, un ralentissement au-delà avec diminution de l'indice pondéral et une perte relativement continue de l'eau sous-cutanée.

Les individus de la classe D sont de stature plus grande que ceux des autres classes à partir de 1 mois et leur indice pondéral se maintient à une valeur plus faible. L'infiltration des tissus sous-cutanés en eau diminue dans le temps, mais moins vite que pour les sujets de la classe A. Ils représentent des sujets « grands, maigres et hydratés ».

La classe C rassemble des sujets plus homogènes : leur stature est plus petite et leur indice pondéral plus grand. C'est dans cette classe que l'infiltration d'eau dans les tissus sous-cutanés reste la plus faible. Ceux sont des « petits gras et peu hydratés ».

La classe A est intermédiaire sur le plan croissance générale et production de graisse tandis que la déperdition d'eau sous-cutanée est plus rapide que dans les autres classes. Malgré la nette séparation de cette classe sur le dendrogramme, on note son hétérogénéité en terme d'indice de regroupement.

Les courbes d'évolution des facteurs comme les résultats des analyses discriminantes montrent que la séparation des classes est la plus nette de 1 à 4 mois et diminue pour devenir inexistante à 1 an.

Une interprétation plus approfondie est en cours avec la participation des pédiatres : la présentation de cette application n'ayant eu pour but que d'illustrer la méthode utilisée.

3. Conclusion

Les diverses publications sur l'analyse des structures des tableaux multiples et en particulier sur l'étude des évolutions montrent que les outils ne manquent pas dans ce domaine. Par contre, les méthodes de classification existantes ne permettaient pas de prendre en compte les tableaux incomplets et/ou asynchrones. La méthode proposée est une technique possible et représente de ce point de vue un traitement, en classification, des données manquantes. Nous nous sommes attachés à prendre en compte le plus d'information possible : structure des tableaux, relations ordinales ou continues entre les observations et covariation des variables évolutives avec le temps. Ce dernier aspect rapproché du calcul du produit scalaire entre opérateurs montre que nous avons intégré les principes généraux de la méthode STATIS et de l'analyse factorielle des évolutions.

Rappelons que la méthode, sur un jeu d'essai artificiel, est apparue robuste même dans des conditions très défavorables.

Bibliographie

[1] CAILLIEZ F. et PAGES J.P. — Introduction à l'analyse des données. *SMASH*, Paris, 1976.

[2] CARLIER A. — Application de l'analyse factorielle des évolutions et de l'analyse intra-période. *Stat. Ana. Donn.*, 1985, Vol. 10, 1, 27-53.

[3] ESCOUFIER Y. et L'HERMIER DES PLANTES H. — A propos de la comparaison graphique des matrices de variance. *Biometrical* J., 1978, Vol. 20, 5, 491-497.

[4] FOUCART T. — Une nouvelle approche de la méthode STATIS. *Rev. Stat. Appl.*, 1983, Vol. 31, 2, 61-75.

[5] LEBART L. — Analyse statistique de la contiguïté. *Thèse de 3e cycle*, Paris, 1966.

[6] LE FOLL Y. — Pondération des distances en analyse factorielle. *Stat. Ana. Donn.*, 1982, 1, 13-31.

[7] L'HERMIER DES PLANTES H. — Structuration des tableaux à trois indices de la statistique : théorie et application d'une méthode d'analyse conjointe. *Thèse 3e cycle*, Montpellier USTL, 1976.

[8] ROUX M. — Algorithmes de classification. *MASSON*, Paris, 1985.