Statistique et analyse des données

Annie Barré
Bernard Fichet

Analyse des correspondances et rotations procustéennes, représentation hiérarchique et ordres compatibles

Statistique et analyse des données, tome 10, n° 1 (1985), p. 16-26
<http://www.numdam.org/item?id=SAD_1985__10_1_16_0>

Statistiques et Analyse de Données.
1985 Vol. 10 n° 1 pp. 16-26

ANALYSE DES CORRESPONDANCES ET ROTATIONS PROCUSTEENNES REPRESENTATION HIERARCHIQUE ET ORDRES COMPATIBLES

Annie BARRÉ et Bernard FICHET

Laboratoire de Biomathématiques
Faculté de Médecine
13385 MARSEILLE CEDEX 5

Résumé : *Des tableaux de contingence indicés par le temps définissent l'évolution des modalités de deux variables qualitatives. Nous étudions cette évolution à l'aide d'une analyse factorielle des correspondances et d'une classification hiérarchique en chaque temps. En outre, des rotations procustéennes et la recherche d'un ordre compatible sur les hiérarchies favorisent la comparaison des schémas.*

Abstract : *Contingency tables depending on the time define the evolution of items for two qualitative variables. We study this evolution by means of a correspondence factor analysis and a hierarchical classification, at each time. Moreover, Procrustes analysis and the search of a common order on units for the hierarchies help in comparing the graphical representations.*

Mots clés : *Analyse factorielle des correspondances, classification hiérarchique, rotations procustéenes, ordre compatible.*

Indices de classification STMA : *06-110, 06-120.*

0 - INTRODUCTION

Considérons les K(K=4) tableaux de contingence $\{X^k,\ k=1,\ldots,K\}$ dépendant du temps k et introduits dans la présentation générale. Ils sont croisés sur deux variables qualitatives I et J représentant respectivement les catégories socio-professionnelles et les cantons du Languedoc-Roussillon. Conformément à l'optique de l'analyse des données, introduisons pour chaque tableau X^k une métrique d_I^k sur I (ou d_J^k sur J) et une famille de masses $\{m_i,\ i\in I\}$ (ou $\{n_j,\ j\in J\}$); nous supposons ces masses indépendantes du temps. Nous sommes ainsi en présence d'une étude $(I,\ d_I^k,\ \{m_i, i\in I\})$ (ou $(J,\ d_J^k,\ \{n_j, j\in J\})$), indicée par le temps k, k=1,...,K.

Pour approcher de telles structures, on peut :
- Soit réaliser un compromis, en introduisant une métrique sur I (ou J), résumée des métriques d_I^k (ou d_J^k), k=1,...,K.
- Soit étudier l'évolution. Dans ce cas, la création d'une métrique sur l'ensemble des quatre temps, fonction des différences entre les métriques d_I^k (ou d_J^k), k=1,...,K, donne une approche globale. En revanche, une représentation des études pour chaque temps (quel que soit le mode de représentation choisi) et une comparaison de ces représentations offrent une approche plus fine de l'évolution, individu par individu. C'est dans cette dernière optique que nous nous plaçons, avec les deux principaux modes de représentation : factoriel et hiérarchique.

1 - ANALYSE FACTORIELLE

La nature des données (tableaux de contingence) nous conduit en chaque temps à réaliser une analyse des correspondances; et pour comparer les schémas, nous effectuons des rotations procustéennes. Une rotation procustéenne consiste à superposer au mieux un nuage $\{N_i,\ m_i\}_{i\in I}$ sur un nuage $\{M_i,\ m_i\}_{i\in I}$ suivant le critère :
$\min\{\Sigma_i m_i \|\vec{M_i N'_i}\|^2 \mid \{N'_i,\ m_i\}_{i\in I}$ isométrique de $\{N_i,\ m_i\}_{i\in I}\}$.
Pour cela, il convient de centrer les deux nuages, puis de trouver une meilleure rotation à l'aide de la matrice de dispersion mixte (voir par exemple Gower [6], Bourgeois [2], Mouttet [8], Lafosse [7], Reissian [9]).

Notons que l'on ne sait pas résoudre un critère global pour une superposition de tous les nuages, mais que la nature ordinale du temps nous amène assez naturellement à superposer successivement le nuage au temps (k+1) sur le nuage au temps k (après rotation de ce dernier) pour k=1,...,(K-1).

Faisons deux remarques :

Pour une représentation visuelle des nuages ainsi superposés, on peut réaliser une A.C.P. du nuage global; mais, et c'est la démarche que nous adoptons, on peut également effectuer des rotations procustéennes dans le plan, à partir des A.C.P. d'ordre 2 de chaque nuage; sur le plan numérique, les rotations sont obtenues analytiquement (Benzécri et Cazes [1]).

Lorsqu'en chaque temps est réalisée une analyse des correspondances entre I et J, plutôt que d'effectuer les rotations procustéennes pour I, puis pour J, il semble préférable d'effectuer les rotations procustéennes pour les nuages relatifs à I∪J offerts par les représentations simultanées classiques. Une telle démarche, que nous adoptons ici, préserve en effet, pour un temps donné, les positions relatives des éléments de I et J.

L'analyse procustéenne nécessite des masses indépendantes du temps. Dans cette optique, nous avons choisi comme masses sur I et J, les moyennes (pondérées par les effectifs) des marginales en chaque temps. Par souci d'homogénéité, ces mêmes masses ont été utilisées pour les A.F.C. temps par temps; et pour ce faire, c'est une A.F.C. avec marges modifiées (Escofier [4]) qui a été réalisée. Notons que cette A.F.C. est non centrée.

Donnons les expressions analytiques des études $(I, d_I^k, \{m_i, i\in I\})$ et $(J, d_J^k, \{n_j, j\in J\})$ aux temps k=1,...,K.
Définissons les quantités suivantes :
Pour i de I et j de J, X_{ij}^k désigne le terme courant du tableau X^k.
Pour i de I et j de J :

$$x^k_{i.} = \sum_{j \in J} x^k_{ij} \; ; \; x^k_{.j} = \sum_{i \in I} x^k_{ij} \; ; \; x^k_{..} = \sum_{i \in I} x^k_{i.} \; .$$

Les masses m_i et n_j vérifient alors :

$$m_i = \left(\sum_{k=1}^{K} x^k_{i.} \right) / \left(\sum_{k=1}^{K} x^k_{..} \right) \; ; \; n_j = \left(\sum_{k=1}^{K} x^k_{.j} \right) / \left(\sum_{k=1}^{K} x^k_{..} \right) .$$

Notant encore pour i de I et j de J, $f^k_{ij} = x^k_{ij} / x^k_{..}$, l'on a :

$$\forall (i,i') \in I^2, \; d^k_I(i,i') = \left\{ \sum_{j \in J} \frac{1}{n_j} \left[\frac{f^k_{ij}}{m_i} - \frac{f^k_{i'j}}{m_{i'}} \right]^2 \right\}^{1/2}$$

$$\forall (j,j') \in J^2, \; d^k_J(j,j') = \left\{ \sum_{i \in I} \frac{1}{m_i} \left[\frac{f^k_{ij}}{n_j} - \frac{f^k_{ij'}}{n_{j'}} \right]^2 \right\}^{1/2}$$

2 - REPRESENTATION HIERARCHIQUE

L'analyse ne porte que sur un ensemble (ici I représentant les catégories socio-professionnelles), mais la démarche est en tout point semblable à celle recherchée pour l'analyse factorielle : représentation en chaque temps, puis aide pour la comparaison des représentations. Pour l'ensemble des catégories socio-professionnelles la métrique est toujours celle de l'A.F.C. avec marges modifiées; et pour une représentation hiérarchique, cette métrique est approchée par l'ultramétrique sous-dominante.

L'aide pour la comparaison des hiérarchies, découle de la recherche d'un ordre compatible, i.e. un ordre sur les éléments offrant un tracé sans croisement de toutes les hiérarchies. Un algorithme (Diday [3], Gaud [5]) détecte s'il y a existence d'un tel ordre, et si oui, l'exhibe. Dans le cas présent il n'y aura pas d'ordre compatible, mais l'ordre proposé répondra aux deux critères d'optimalité suivants :

- il est compatible pour les trois premiers temps.
- il est compatible pour un sous-ensemble de cardinalité maximun (huit catégories).

3 - RESULTATS

Les figures 1 et 2 représentent les A.F.C. d'ordre 2, suivies des rotations procustéennes comme décrites précédemment; pour une meilleure lisibilité, bien que les rotations procustéennes aient pour objet de superposer les nuages, les résultats sont présentés temps par temps, i.e. pour les années 1954 (temps 1), 1962 (temps 2), 1968 (temps 3), et 1975 (temps 4). L'évolution des catégories socio-professionnelles, extraite de ces quatre schémas, est présentée dans la figure 3. Enfin, la figure 4 donne, temps par temps, la classification hiérarchique de ces mêmes catégories, suivant l'ordre optimal mentionné au paragraphe 2.
Les analyses factorielles permettent de dégager pour les catégories socio-professionnelles les conditions suivantes :

- Le secteur agricole (catégories EA et OA) se distingue des autres secteurs, d'une part quant à sa représentation à chaque temps et d'autre part quant à son évolution. Ces catégories sont éloignées des autres catégories et tendent à se rejoindre au temsp 4.

Ce rapprochement n'est semble-t-il pas observé dans les autres analyses décrites dans ce volume. Mis également en évidence dans notre classification hiérarchique, il n'est pas la conséquence des différentes approximations nécessaires aux analyses : en témoigne un retour aux tableaux des distances. On ne peut chercher son origine que dans le choix de la métrique. Revenons aux formules exposées au paragraphe 1. Pour un temps donné k, soient i et i' deux catégories dont les fréquences f^k_{ij} et $f^k_{i'j}$ sont plus faibles qu'aux autres dates (les effectifs du secteur agricole décroissent fortement au temps k=4). Il est clair que cette décroissance n'affecte que faiblement les masses m_i et $m_{i'}$, définies comme des moyennes pondérées sur tous les temps. La distance $d^k_I(i,i')$ tend donc vers zéro. Cette explication étant donnée, nous laissons au lecteur le soin d'apprécier si le choix d'une métrique tendant à rapprocher des catégories évanescentes, s'avère positif.

- Les catégories OU et AC évoluent dans la même direction au cours du temps. Cependant les points de départ sont très éloignés et leurs distances mutuelles diminuent jusqu'au temps 4.

- Les catégories CM, EM, PL et SE ont une évolution similaire, bien que celle-ci soit beaucoup plus accentuée pour PL, moins pour CM et EM, et encore moins pour SE. Elles varient dans le même sens et assez peu

aux temps 1,2 et 3. Puis au temps 4, se produit un changement complet de direction, qui représente un éloignement par rapport aux autres catégories, assez important pour SE, CM, et EM et beaucoup plus accentué pour PL.

Il se dégage donc trois types d'évolution, le premier représentant le secteur agricole, le second le secteur industriel et commercial, le troisième les autres catégories.
Notons également que la catégorie CP a une évolution différente des trois types décrits; toutefois l'interprétation s'avère délicate, puisque CP représente l'ensemble hétérogène des autres catégories.

Si nous considérons maintenant l'évolution des cantons, on peut distinguer les zones urbaines des zones rurales.

- Les villes comme Montpellier (K8), Perpignan (O4), Nimes (F1), Béziers (H6) et Narbonne (C9) qui, au temps 1, sont représentatives des différentes catégories socio-professionnelles, s'éloignent des catégories EA et OA au cours du temps, accentuant ainsi la disparité ville-campagne. Montpellier et Nimes se rapprochent plus particulièrement de PL.
- Les cantons de Lozère, plus ruraux, restent plus proches des catégories EA et OA que des autres.

L'analyse hiérarchique sur les catégories socio-professionnelles permet de retrouver le comportement des catégories EA et OA, éloignées des autres catégories, et se rapprochant l'une de l'autre au temps 4. Les catégories EM, CM, SE et PL restent groupées aux temps 1, 2 et 3. PL s'écarte de l'ensemble des autres catégories au temps 4.

Globalement, on notera une nette rupture entre les temps 3 et 4. Pour les catégories socio-professionnelles, celle-ci se traduit au niveau factoriel par un changement de direction de certaines catégories, et au niveau hiérarchique par l'absence d'un ordre compatible.

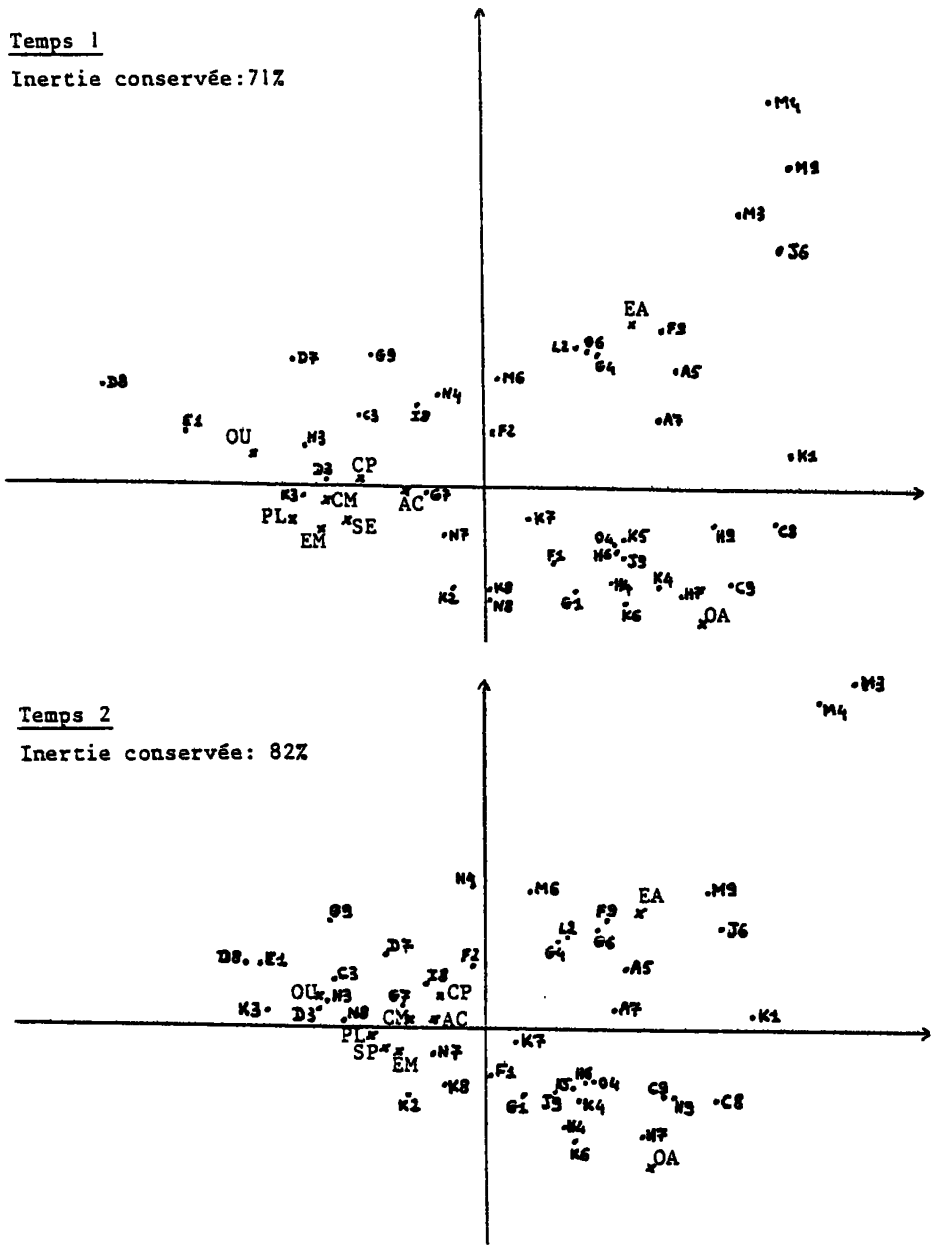

Temps 1
Inertie conservée:71%
EA
OU
CP
CM
AC
PL
SE
EM
OA
Temps 2
Inertie conservée: 82%
EA
OU
CP
CM
AC
PL
SP
EM
OA

FIGURE 1

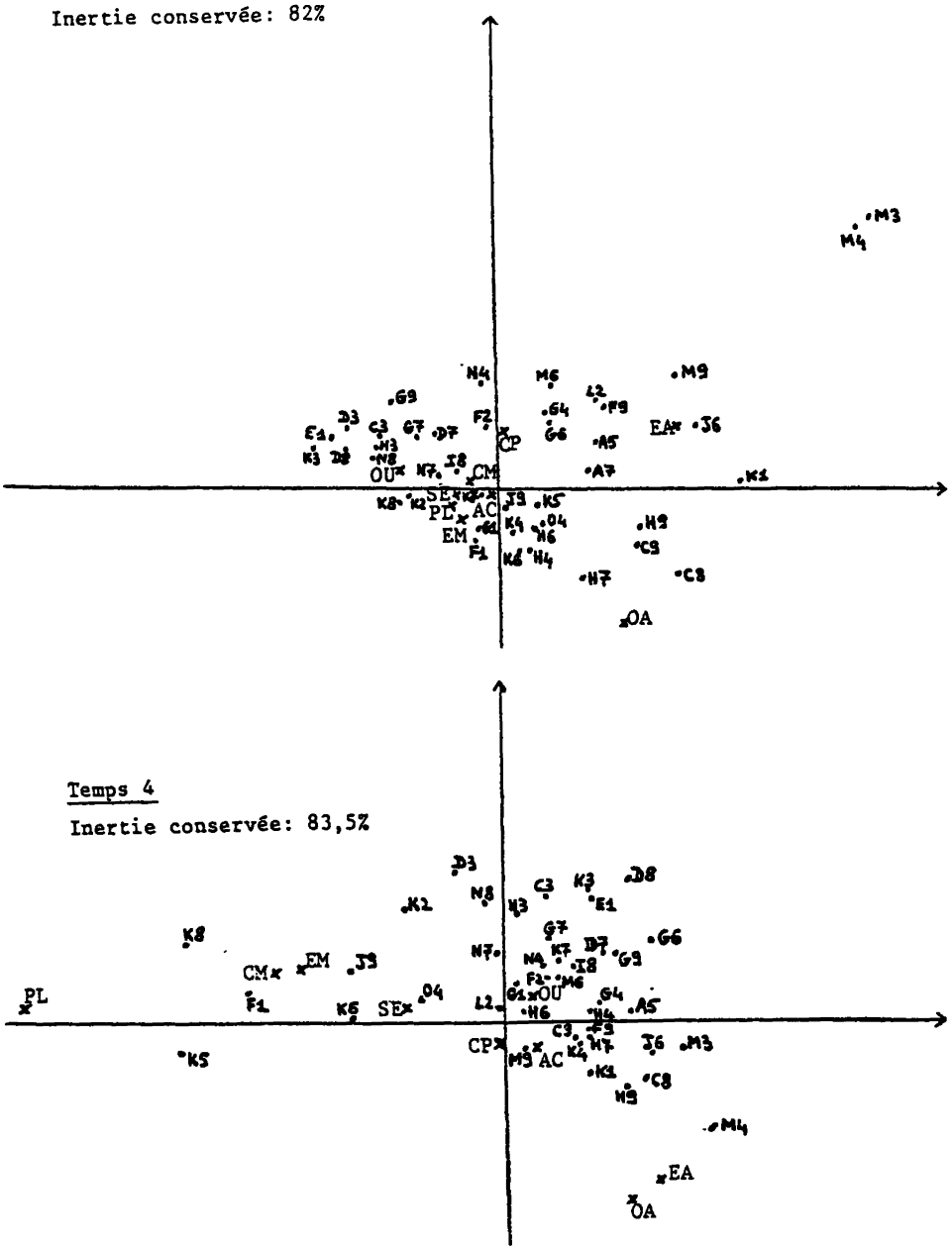
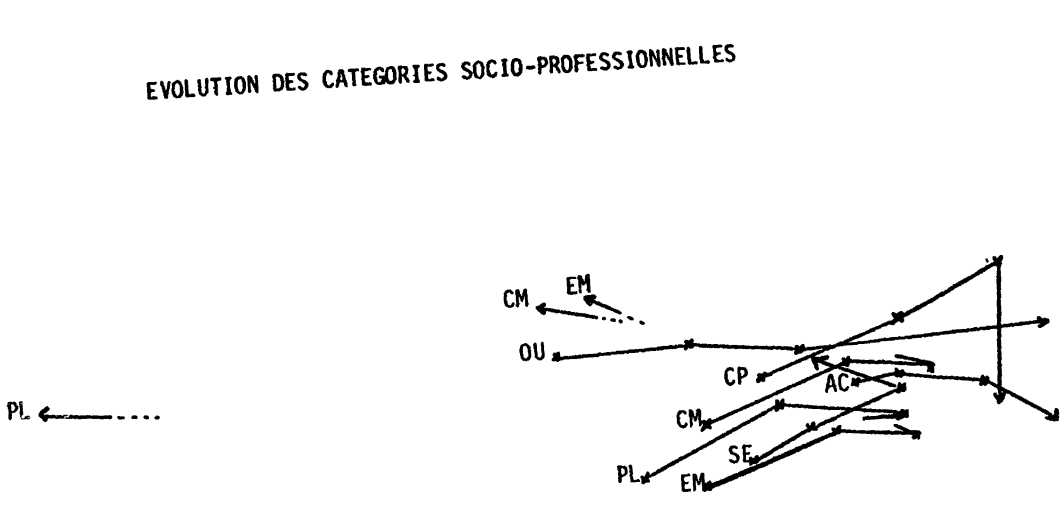

Temps 3
Inertie conservée: 82%
M3
M4
N4
M6
M9
L2
G9
F9
G4
D3
C3
G7
D7
F2
CP
G6
EA
J6
E1
K3
D8
N8
OU
K7
I8
CM
A5
A7
K1
SE
J9
K5
PL
AC
K1
O4
H2
EM
F1
K6
H4
C9
H7
C8
OA
Temps 4
Inertie conservée: 83,5%
D3
K2
N8
C3
K3
D8
M3
E1
K8
G7
G6
CM
EM
J9
N7
N4
K7
D7
G9
I8
F2
M6
PL
F1
K6
SE
O4
L2
G1
OU
G4
A5
H6
H4
C3
F3
CP
K5
M5
AC
K4
H7
J6
M3
K1
C8
H9
M4
EA
OA

FIGURE 2

EVOLUTION DES CATEGORIES SOCIO-PROFESSIONNELLES



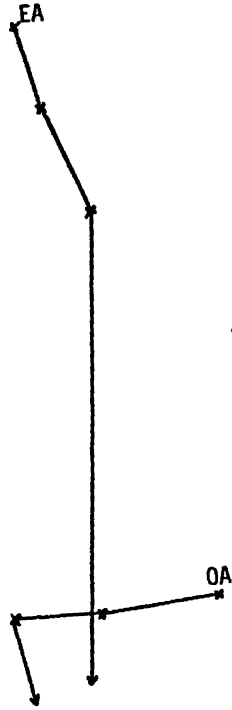


FIGURE 3

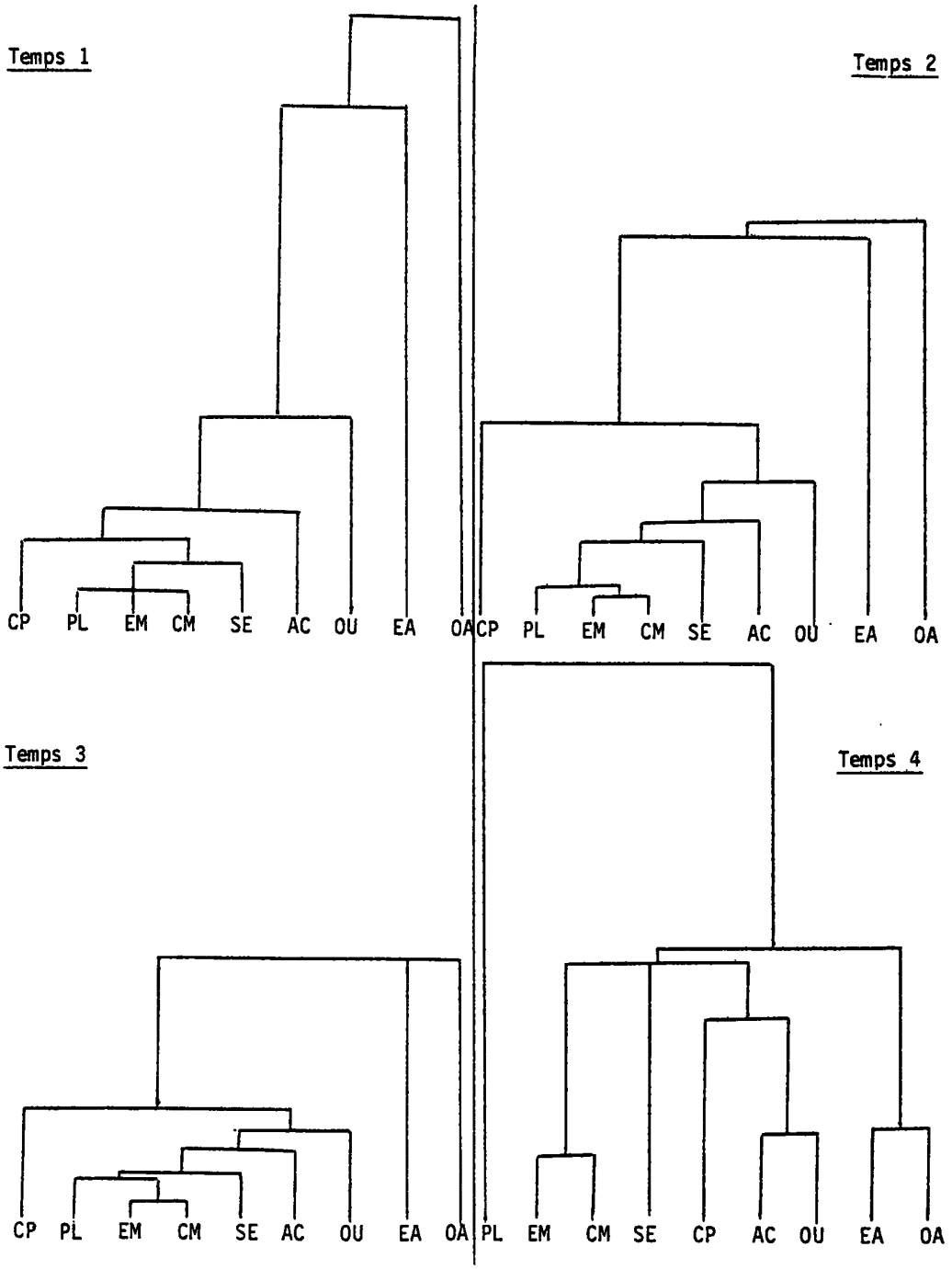

Temps 1
Temps 2
CP PL EM CM SE AC OU EA OA
CP PL EM CM SE AC OU EA OA
Temps 3
Temps 4
CP PL EM CM SE AC OU EA OA
PL EM CM SE CP AC OU EA OA

FIGURE 4

4 - REFERENCES

1. BENZECRI J-P, CAZES P. "Recherche du déplacement minimisant la distance entre deux ensembles de points homologues situés dans un plan". Les Cahiers de l'Analyse des Données, 1978, Vol III, n°4, p.435-439.

2. BOURGEOIS Ph. "Recherche du déplacement minimisant la distance entre deux configurations de points indicés par un même ensemble fini. Méthode et applications en reconnaissance des formes et en analyse des données cubiques". Thèse de 3ème cycle. Université Pierre et Marie Curie. Paris. 1980.

3. DIDAY E. "Croisements,ordres et ultramétriques : application à la recherche de consensus en classification automatique". Rapport I.N.R.I.A. 1982, n°144.

4. ESCOFIER B. "Traitement des questionnaires avec non réponse, analyse des correspondances avec marge modifiée et analyse multicanonique avec contrainte". Rapport I.N.R.I.A. 1981, n°82.

5. GAUD E. "Représentation d'une préordonnance. Etude de ses images euclidiennes. Problèmes de graphes dans sa représentation hiérarchique". Thèse de 3ème cycle. Université de Provence-Marseille. 1983.

6. GOWER J-C. "Generalized Procrustes Analysis". Psychometrika, 1975, Vol 40, p. 33-51.

7. LAFOSSE R. "Analyses procustéennes de deux tableaux. Proposition d'une technique visant à la détection de points originaux. Essai de présentation synthétique d'analyses de deux tableaux". Thèse de 3ème cycle. Université Paul Sabatier. Toulouse. 1985.

8. MOUTTET F. "Comparaison de tableaux par la méthode Procuste". Thèse de 3ème cycle. Université Pierre et Marie Curie. Paris. 1981.

9. REISSIAN S. "Sur l'analyse factorielle de plusieurs tableaux de données". Mémoire de D.E.A. Université de Provence-Marseille. 1981.