REVUE DE STATISTIQUE APPLIQUÉE

G. CARAUX

Réorganisation et représentation visuelle d'une matrice de données numériques : un algorithme itératif

Revue de statistique appliquée, tome 32, n° 4 (1984), p. 5-23

<http://www.numdam.org/item?id=RSA_1984__32_4_5_0>

REORGANISATION ET REPRESENTATION VISUELLE D'UNE MATRICE DE DONNEES NUMERIQUES : UN ALGORITHME ITERATIF

G. CARAUX

Unité de Biométrie
E.N.S.A.M. – I.N.R.A. – U.S.T.L.
9, Place Pierre Viala, 34060 Montpellier Cedex

I. INTRODUCTION

Un tableau rectangulaire de données est souvent l'élément de base d'une analyse statistique. C'est lui qui, soumis à différentes approches méthodologiques, se réflète dans les résultats numériques ou graphiques des calculs.

La grande facilité d'accès aux résultats d'une analyse statistique plus ou moins complexe, par des moyens informatiques, a développé une pratique regrettable qui permet de passer à la résolution numérique d'un modèle sans prendre le temps d'examiner en détail le tableau de données (BRUNET, 1977). A titre d'exemple peu d'auteurs développent autant l'étude préliminaire d'une analyse en composantes principales que CAILLIEZ et PAGES (1976) sur les données des poissons d'"Amiard".

Est-il vraiment raisonnable de faire une analyse des données sur les tableaux de gauche de la figure 1, alors qu'ils peuvent se mettre, après réorganisation, sous la forme de droite ?

La généralisation de l'utilisation par des non spécialistes de certaines analyses de données n'est-elle pas sans danger ? L'utilisateur non rompu aux méthodes statistiques, souligne LEBART (1979) quitte son domaine familier sans percevoir la barrière du calcul.

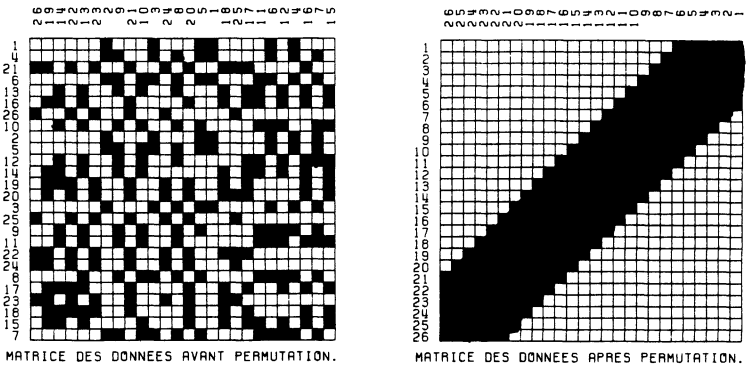
Nous proposons dans cet article l'utilisation des nouveaux moyens d'impression graphique disponibles sur ordinateur, pour représenter visuellement un tableau de données. L'aspect pédagogique de cette représentation devra inciter l'utilisateur d'outil statistique à mieux s'imprégner de ses données et à en rechercher la cohérence par des moyens du domaine de l'intelligence et non du calcul. L'utilisation de modèles au pouvoir de résolution plus puissant n'en sera que plus féconde.

Nous utiliserons pour cela les outils de sémiologie graphique développés par BERTIN (1971) et une technique de réorganisation automatique d'un tableau de données.

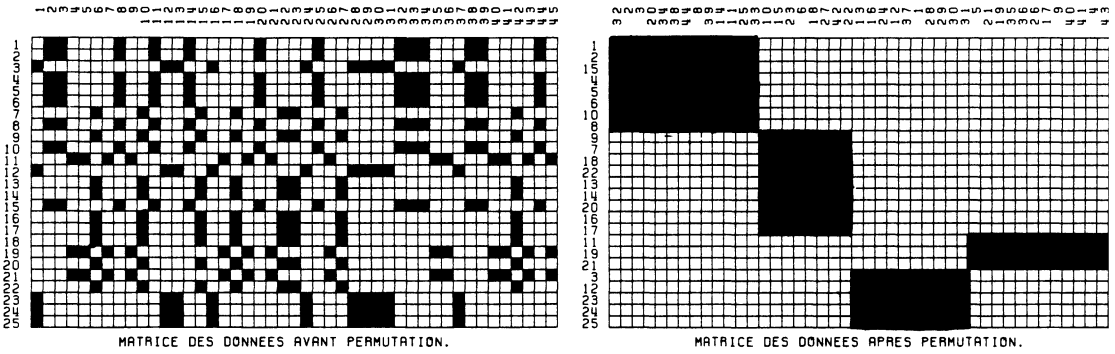
Nous essayerons, en cela, par permutation des lignes et des colonnes, de mettre de l'ordre dans le tableau des données afin d'en faciliter la lecture visuelle.

Après avoir fait une rapide retrospective sur le sujet, nous proposerons un algorithme basé sur la réorganisation de la matrice des distances entre les lignes d'une part, et celle des distances entre les colonnes du tableau des données d'autre part. La méthode avancée tentera, par une recherche itérative de transpositions, de

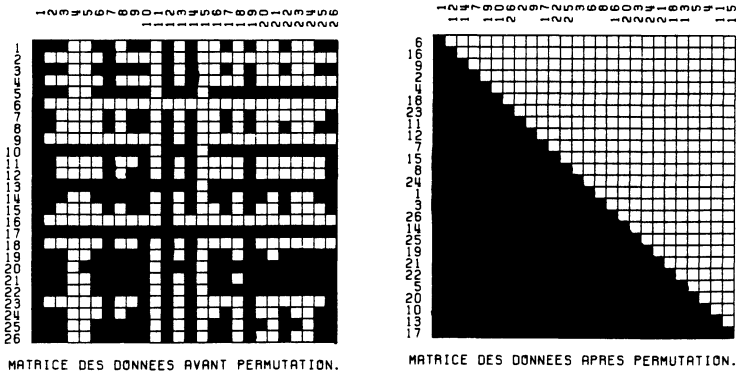
Mots clés : Graphique, Réorganisation de données, Sémiologie graphique, Matrice ordonnable.



a) Exemple Théorique I



b) Exemple Théorique II



c) Exemple Théorique III

Figure 1

faire apparaître des distances faibles le long de la diagonale des matrices des distances et d'en éloigner les distances élevées.

Le critère mesurant le degré de réorganisation aura une expression utilisant le concept de moment d'inertie le long d'un axe. Il aboutira à des simplifications numériques qui seront déterminantes sur la durée des calculs. Pour terminer, nous présenterons plusieurs exemples obtenus à l'aide d'un programme informatique disponible.

II. RETROSPECTIVE BIBLIOGRAPHIQUE SUR CE SUJET

L'idée de réorganiser le tableau des données pour en clarifier la structure n'est pas une idée récente. Certains l'attribuent au polonais CZEKANOWSKI dans un travail d'anthropologie (KULCZYNSKI, 1927). Les écologistes l'utilisent depuis longtemps pour étudier les associations végétales (KULCZYNSKY, 1927 ; GUINOCHET et CAZAL, 1957 ; Mc INTOSH, 1978) ou les associations d'espèces animales (MACFADYEN, 1963). Les archéologues la pratiquent (KENDALL, 1969) pour dégager une chronologie d'objets par des méthodes de sériation (HODSON, KENDALL et TAUTU, 1971), les sociologues pour typer des relations interindividuelles dans un groupe (WHITE, BOORMAN et BREIGER, 1976). En France les travaux de BERTIN (1977) sur ce sujet sont bien connus des géographes (BONIN, 1977). Enfin plusieurs statisticiens connus ont abordé ce problème en marge de leurs travaux (DAGNELIE, 1960 ; SOKAL et SNEATH, 1963 ; BENZECRI *et al.*, 1973).

On peut aussi trouver en recherche opérationnelle des sujets très voisins (Mc CORMICK, SCHWEITZER et WHITE, 1972) et notamment le problème de "l'assignation quadratique" (LAWLER, 1963). Celui-ci trouve des applications dans la localisation optimale des composantes d'un ordinateur (HANAN et KURTZBERG, 1972). On doit citer également les numériciens qui utilisent pour la mémorisation et l'accélération du calcul matriciel, dans le cas de matrices clair-semées (Sparse matrix), des techniques qui s'apparentent tout à fait à notre problème de réorganisation (DUFF, 1977).

Pour finir cette énumération, notons que certains logiciels statistiques de grande diffusion proposent des programmes de réorganisation de données statistiques. On en trouve un dans la bibliothèque de l'ADDAD (JAMBU et LEBEAUX, 1978).

Celui-ci est surtout adapté à la réorganisation d'un tableau de contingence quand un effet Guttman est constaté dans les plans factoriels d'une analyse des correspondances (BENZECRI *et al.*, 1973 ; SCHRIEVER 1982).

Dans BMDP (1981), on trouve aussi, dans le programme d'analyse en composantes principales, à la suite d'une demande de rotation orthogonale des axes (critère varimax ou autre. . .), une édition graphique de la matrice des corrélations, après réorganisation de celle-ci. Cette réorganisation est opérée en fonction des corrélations variable-facteurs après rotation.

Solutions déjà proposées

De nombreuses méthodes de résolution ont été étudiées. Certains articles à la bibliographie imposante (ARABIE, BOORMAN et LEVITT, 1978) sont là pour nous le rappeler.

Dans toute cette littérature le problème est de trouver une permutation des lignes et une permutation des colonnes d'un tableau qui maximisent un certain critère plus ou moins bien défini suivant les auteurs.

La difficulté majeure rencontrée par tous dans la résolution des algorithmes proposés, réside dans le nombre très rapidement incommensurable de permutations envisageables. KENDALL (1971) affirme que l'utilisation d'un des algorithmes (KENDALL, 1963) que lui a inspiré les travaux de PETRI (1899) et qui explore d'une manière exhaustive l'ensemble des permutations, nécessiterait plus d'un milliard d'années (l'âge de l'Univers) de calcul en ordinateur pour un nombre d'objets permutés supérieur à 33.

Les très nombreux algorithmes présentés dans la littérature tentent de s'affranchir de cette contrainte.

Quand la solution optimum exacte veut être trouvée, les méthodes de résolution restent en général lentes et non opérationnelles pour un nombre d'éléments permutés supérieurs à 15.

Dans ces techniques nous trouvons entre autres celles qui utilisent les résultats de la programmation linéaire, non linéaire ou en nombre entier (LAWLER, 1963 ; ELMAGHRABY, 1968 ; LAPORTE, 1975 ; MARCOTORCHINO et MICHAUD, 1981), ou celles qui utilisent la programmation dynamique (HUBERT et COLLEDGE, 1981 ; ELMAGHRABY, 1968 ; ADELSON, NORMAN et LAPORTE, 1976).

Cette dernière approche au lieu d'énumérer n ! permutations, n'envisage que 2^n itérations. Ceci reste encore rapidement inopérant pour $n \geqslant 15$.

Pour trouver des méthodes de résolution rapide, de nombreuses solutions sont avancées. Dans l'abondante littérature sur ce sujet, nous pouvons distinguer plusieurs catégories (non disjointes) de solutions :

celles qui utilisent des méthodes de représentation plane empruntées à l'analyse multidimensionnelle des données (KENDALL, 1971 ; BENZECRI, 1973 ; BMDP, 1981),

celles qui ne retiennent qu'une part des valeurs du tableau pour ranger chaque ligne ou colonne à une place optimale (GELFAUD, 1971 ; RENFREW et STERUD, 1969),

celles qui appliquent les résultats de la théorie des graphes (LENSTRA et RINNOY KAN, 1975 ; HUBERT, 1974 ; BATBEDAT, 1984 ; EYTAN, 1975).

celles qui restreignent le nombre de permutations envisageables en opérant sur les transpositions ou les réallocations par itération successive (HOLE & SHAW, 1967 ; SZCZOTKA, 1972 ; LEDUC, 1982).

Très souvent la performance de ces algorithmes est conditionnée par la structure de la matrice des données à réorganiser ou par des propriétés simplificatrices de l'algorithme lui-même. Par exemple les méthodes de GELFAUD (1971) appliquées à l'archéologie sont optimales si la matrice à réorganiser, a la forme

d'une matrice de ROBINSON (ROBINSON, 1951). De même l'utilisation de l'analyse des correspondances n'est justifiée (BENZECRI *et al.*, 1973 ; SCHRIEVER, 1982) que si les deux transitions f_I^J et f_J^I sont "latéralement croissantes".

III. VISUALISATION D'UN TABLEAU DE DONNEES

Dans le souci pédagogique que nous poursuivons, il nous a semblé très efficace de représenter des données numériques non pas par des chiffres dans un tableau, mais par des ombres sur un damier.

L'oeil est un outil de synthèse qui peut nous livrer instantanément beaucoup d'informations.

Un travail connu a été réalisé en France par BERTIN (1967). Nous nous sommes inspirés de ses travaux en matière de sémiologie graphique. Nous y avons puisé notamment une échelle de symboles graphiques qui permet de représenter visuellement l'intensité d'un nombre. Cette échelle a été utilisée par son auteur pour construire des dominos plastiques qui, une fois rassemblés, peuvent servir à visualiser et à réorganiser comme nous, une matrice de données. Cependant si la manipulation de ces dominos peut avoir un intérêt pédagogique pour certains, cette manipulation est longue et empreinte d'une subjectivité qui lui a souvent été reprochée (BRUNET, 1977).

L'utilisation de ces graphismes en informatique a été freinée par l'impossibilité d'utiliser, pour les imprimer, les moyens d'édition connectés aux ordinateurs. Les traceurs de courbes monoplume étaient inadaptés en raison du temps nécessaire à l'édition de hachures sur ce type de matériel.

Depuis quelques années, les écrans cathodiques au pouvoir de résolution élevé, peuvent être utilisés (LEDUC, 1982 ; GRONOFF, 1982) ainsi que des imprimantes à caractères programmables (de GOLBERG, CHAPPUIS et TAN CHONG CHIN, 1982).

On rencontre maintenant des imprimantes électro-statiques (Benson-Varian par exemple) qui sont capables d'éditer des dessins en continu (ligne par ligne), sans être freinées par la densité du graphisme. Le pouvoir de résolution de ce type de matériel est de 100 points par centimètre. Sa rapidité d'édition est élevée et son coût d'utilisation modique (nous avons utilisé l'imprimante Benson-Varian du CNUSC de MONTPELLIER).

IV. SOLUTION PROPOSEE ICI

Comme nous l'avons annoncé plus haut l'objectif que nous nous posons est de fournir avant toute analyse statistique, une visualisation de la matrice des données réorganisée.

Nous avons donc besoin d'un algorithme de réorganisation qui soit rapide excluant ainsi ceux qui fournissent une solution optimum. Le concept d'optimum est relatif à l'étape préliminaire de l'analyse où nous voulons nous situer. Nous

souhaitons fournir une présentation visuellement parlante des données au moindre coût. Un algorithme donnant une solution "quasi-optimale" est suffisant.

Parmi les quatre catégories d'heuristiques énumérées plus haut, nous excluons les méthodes de projection sur un plan qui sont des analyses en elles-même et qui ne sont adaptées qu'à certains tableaux. Nous excluons aussi les méthodes ne retenant qu'une partie des données, leurs robustesses étant faibles dans le cas général.

Appartenant aux deux dernières catégories, le critère proposé par Mc CORMICK, SCHWEITZER et WHITE (1972) repris par LENSTRA et RINNOY KAN (1975) et le critère évoqué par LEDUC (1982) permettent des délais de résolution performants.

Ces deux critères sont construits d'une manière voisine et les heuristiques qui leur sont associées tendent à les minimiser. Le premier est la somme des produits de chaque valeur du tableau des données par ses éléments adjacents, alors que le second est la somme des différences de chaque valeur du tableau avec ses éléments contigus.

Ainsi ces deux critères restituent chaque valeur du tableau dans son environnement immédiat.

Leur optimisation est obtenue par permutation des lignes et des colonnes en utilisant des heuristiques empruntées à la théorie des graphes.

L'absence de référence, par ces deux critères, à des concepts statistiques connus (distances, . . .) ne permet que partiellement de comprendre ce qu'ils conduisent dans la réorganisation du tableau des données. Ceci d'autant plus que les valeurs de ce dernier sont hétérogènes.

Cependant si le tableau des données contient des valeurs 0 ou 1, le premier critère regroupe les valeurs 1, dans le tableau des données, en zone les plus "compactes" possibles, le second rend la frontière entre les valeurs 1 et 0, la plus courte possible.

Ces propriétés, si elles sont esthétiquement satisfaisantes, ne traduisent pas les pratiques empiriques de certains auteurs. Les phytosociologues, par exemple, désirent regrouper les lignes et les colonnes qui se ressemblent (DAGNELIE, 1960), plutôt que de faire apparaître des taches compactes et bien délimitées dans leur tableau des données.

C'est pour ces raisons que notre curiosité nous a incité à explorer une autre approche.

V. METHODE DE REORGANISATION PROPOSEE

Formalisation

Comme beaucoup d'auteurs, nous envisagerons la permutation des lignes ou des colonnes du tableau de données à l'aide de la matrice des distances calculées entre les lignes puis entre les colonnes du tableau.

Soit $I = \{1, \ldots, n\}$ et $J = \{1, \ldots, p\}$ deux ensembles ordonnés d'objets statistiques indiçant la matrice des données que nous noterons X_{IJ}.

Notons P une permutation des objets de I et Q une permutation des objets de J.

P(I) est alors l'ensemble des objets étudiés exprimés dans l'ordre induit par la permutation P (idem pour Q(J)).

$X_{P(I)\,Q(J)}$ est la matrice des données après réorganisation des lignes et des colonnes par P et Q.

Les matrices X_{IJ} et $X_{P(I)\,Q(J)}$ contiennent les mêmes valeurs, seule change leur présentation.

Soit D_{II} et D_{JJ} les matrices des distances entre les lignes et les colonnes du tableau des données X_{IJ}.

Le choix de ces deux matrices est fondamental ; de lui dépendra comme nous le verrons plus loin, la qualité pédagogique de la réorganisation obtenue. Ce choix s'inspirera de la nature des données traitées et des contrastes que l'on veut révéler.

$D_{P(I)\,P(I)}$ et $D_{Q(J)\,Q(J)}$ sont les deux matrices de distances après permutation des lignes et des colonnes.

Matriciellement on peut écrire :

$$D_{P(I)\,P(I)} = P\,D_{II}\,P' \qquad PD_{II}\,P'(i,i') = D_{II}(P(i), P(i'))$$

$$D_{Q(J)\,Q(J)} = Q\,D_{JJ}\,Q' \qquad QD_{JJ}\,Q'(j,j') = D_{JJ}(Q(j), Q(j'))$$

et $$X_{P(I)\,Q(J)} = P\,X\,Q \qquad P\,X\,Q\,(i,j) = X(P(i), Q(i))$$

Admettons, ce qui semble trivial, qu'une permutation P sur I laisse inchangé D_{JJ} (de même pour Q, J et D_{II}). Si, comme nous le souhaitons, P est choisi en fonction de D_{II} et Q en fonction de D_{JJ}, alors la recherche de P et Q pourra se faire en deux étapes disjointes.

Ainsi ce que nous allons dire maintenant sur le choix de P pourra s'appliquer à Q directement.

Recherche de la permutation P

Nous allons essayer par permutation simultanée des lignes (et des colonnes) de D_{II}, de regrouper autour de la diagonale de D_{II}, les distances les plus faibles et de faire apparaître le plus loin possible de cette diagonale, les éléments élevés (cf. Fig. 2).

Ainsi nous regrouperons les éléments de I qui sont voisins et opposerons ceux qui sont éloignés.

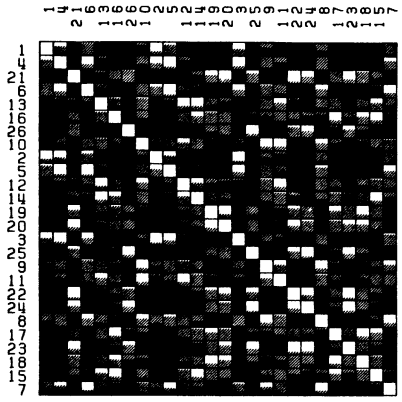
Cette approche reprend l'idée empirique des phytosociologues (DAGNELIE 1960) quand ils réorganisent une matrice d'association entre espèces (cf. exemple 1 au paragraphe VI).

Pour atteindre cet objectif, nous avons besoin d'un indice mesurant la dispersion des valeurs de la matrice D_{II} autour de sa diagonale.

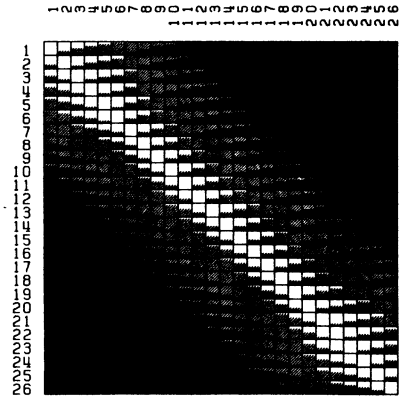
Pour cela, définissons sur $I \times I$, une application δ définie par $\delta(i, i') = |i - i'|$. On peut considérer δ comme l'application distance entre la position d'une valeur dans D_{II} et la diagonale de cette matrice.

Considérons également que les valeurs de D_{II} sont des masses affectées aux éléments de $I \times I$.

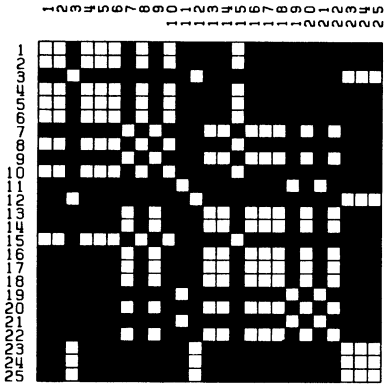
Nous pouvons alors proposer deux indices de dispersion des valeurs de D_{II} autour de sa diagonale.



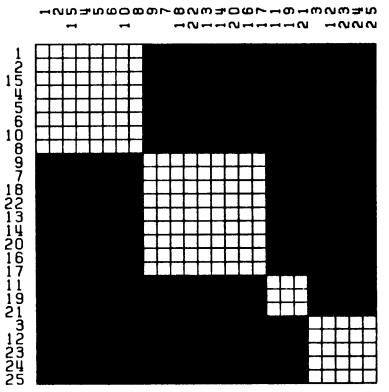
EXEMPLE THEORIQUE (1) - INDICE DE JACCARD
MATRICE DES DISTANCES ENTRE LES LIGNES
AVANT PERMUTATION.



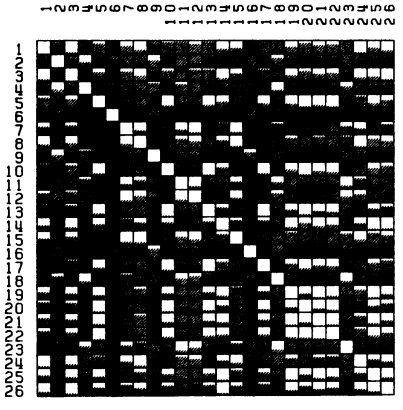
EXEMPLE THEORIQUE (1) - INDICE DE JACCARD
MATRICE DES DISTANCES ENTRE LES LIGNES
APRES PERMUTATION.



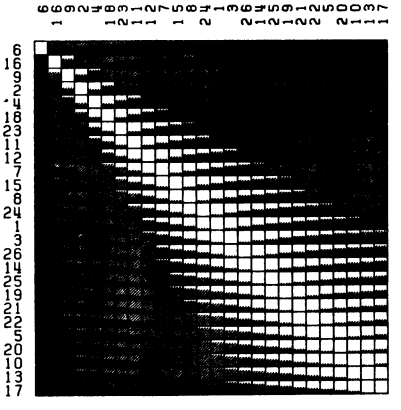
EXEMPLE THEORIQUE (2) - INDICE DE JACCARD
MATRICE DES DISTANCES ENTRE LES LIGNES
AVANT PERMUTATION.



EXEMPLE THEORIQUE (2) - INDICE DE JACCARD
MATRICE DES DISTANCES ENTRE LES LIGNES
APRES PERMUTATION.



EXEMPLE THEORIQUE (3)
MATRICE DES DISTANCES ENTRE LES LIGNES
AVANT PERMUTATION.



EXEMPLE THEORIQUE (3)
MATRICE DES DISTANCES ENTRE LES LIGNES
APRES PERMUTATION.

Figure 2

$$\mathfrak{L}_D = \sum_i \sum_{i'} \delta_{ii'} d_{ii'} = \sum_i \sum_{i'} |i - i'| \, d_{ii'}$$

$$\mathfrak{M}_D \quad \sum_i \sum_{i'} \delta^2_{ii'} d_{ii'} = \sum_i \sum_{i'} (i - i')^2 d_{ii'}$$

Le problème posé plus haut est de trouver la permutation P qui maximise soit $\mathfrak{L}_{PDP'}$ soit $\mathfrak{M}_{PDP'}$.

En effet, plus les valeurs faibles de D seront placées proches de la diagonale et correlativement, plus les valeurs fortes en seront éloignées, plus les critères $\mathfrak{L}$ et $\mathfrak{M}$ seront élevés.

Comme toujours en statistique, les critères de dispersion d'expression quadratique sont plus efficaces numériquement que ceux utilisant les valeurs absolues. Ici l'utilisation de $\mathfrak{M}$ va aboutir à des simplifications numériques qui seront déterminantes pour la durée des calculs. C'est pourquoi nous n'avons pas retenu le critère $\mathfrak{L}$, bien que SZCZOTKA (1972) l'ait utilisé avec succès.

Recherche de P maximisant $\mathfrak{M}_{PDP'}$

Plusieurs solutions se présentent à nous dans la littérature.

Dans le cadre des "problèmes d'assignation quadratique" (quadratic assignment problem) par exemple, plusieurs algorithmes optimaux ou sous optimaux ont été proposés pour la résolution de problèmes connexes au nôtre (GILMORE, 1962 ; LAWLER, 1962 ; HILLIER et CONNERS, 1966).

On peut également mettre notre problème sous une forme solvable par les techniques itératives de la programmation dynamique (HUBERT et GOLLEDGE, 1981 ; ELMAGHRABY, 1968).

Aucun de ces algorithmes ne permet de tenir compte du caractère très particulier de $\delta(i, i') = |i - i'|$ qui dans notre cas, aboutit à des simplifications numériques importantes.

La construction de P que nous proposons, se fera par la recherche itérative d'un produit de transpositions qui à chaque pas maximise le gradient de $\mathfrak{M}$. On trouve cette technique utilisée par plusieurs auteurs pour d'autres critères (HOL et SCHAW, 1967 ; SZCZOTKA, 1972).

Nous pouvons définir notre algorithme par récurrence. Supposons qu'à l'étape $k - 1$ ($k \geqslant 1$), la permutation proposée soit

$$P_{k-1} = T_{k-1} \ldots T_r \ldots T_2 \, T_1 \qquad (P_0 = I)$$

où T_r est la transposition qui, à l'étape r, échange $P_{r-1}(i_r)$ et $P_{r-1}(i'_r)$ (Rappelons que $P_{r-1}(i_r)$ est l'image de i_r dans la permutation P_{r-1}).

A l'étape k, nous améliorerons la solution P_{k-1} en proposant $P_k = T_k \, P_{k-1}$

Choix de T_k

Pour simplifier l'écriture posons :

$$P_{k-1} = \rho$$

$$\mathfrak{M}_k = \mathfrak{M}_{P_k D P'_k}$$

et soit (i_k, i'_k) le couple de I x I transposé à l'étape k.

Pour choisir T_k nous calculerons $\mathcal{E}_k$ la matrice symétrique des gradients de $\mathfrak{M}_{k-1}$ définie sur les couples de I x I

$$\mathcal{E}_k(\rho(i_k), \rho(i'_k)) = \mathfrak{M}_k - \mathfrak{M}_{k-1} \quad (i_k, i'_k) \in I \times I$$

Cette matrice est symétrique et d'éléments diagonaux nuls.

Si $\mathcal{E}_k(\rho(i_k), \rho(i'_k)) < 0 \quad \forall (i_k, i'_k) \in I \times I$

la valeur de $\mathfrak{M}_{k-1}$ ne peut être augmentée par une transposition supplémentaire à l'étape k.

Dans ce cas $P = P_{k-1}$ sera considérée comme solution de notre algorithme.

Sinon nous choisirons la transposition T_k correspondant à la valeur maximum de la matrice des gradiants $\mathcal{E}_k$.

On montre facilement :

$$\mathcal{E}_k(\rho(i), \rho(i')) = (i^2 - i'^2)\left(\sum_r d_{r\rho(i')} - \sum_r d_{r\rho(i)}\right)$$

$$- 2(i - i')\left(\sum_r r\, d_{r\rho(i')} - \sum_r r\, d_{r\rho(i)}\right) + 2(i - i')^2\, d_{\rho(i)\,\rho(i')}$$

Ainsi à chaque étape le calcul des $\dfrac{n(n-1)}{2}$ termes distincts de $\mathcal{E}_k$ nécessite

a) la connaissance des sommes par ligne des éléments de la matrice de distances D_{II}. A chaque étape ces valeurs sont identiques, seul change leur ordre d'énumération,

b) la connaissance des sommes pondérées

$$S_k(\rho(i)) = \sum_r r\, d_{r\rho(i)}$$

Or on peut démontrer que :

$$S_k(\rho(i)) = S_{k-1}(\rho(i)) + (i_{k-1} - i'_{k-1})\,(d_{i'_{k-1}\,\rho(i)} - d_{i_{k-1}\,\rho(i)})$$

Ainsi les sommes S_k se calculent à partir des sommes S_{k-1} utilisées à l'étape précédente, par une opération simple indépendante de n.

Donc le temps de calcul des termes de $\mathcal{E}_k$ est indépendant de n sauf pour $\mathcal{E}_1$.

Le nombre d'opérations arithmétiques nécessaires au choix d'une transposition à l'étape k est donc une fonction en $n(n-1)/2$.

Ainsi si le nombre de transpositions nécessaires pour trouver P n'est pas trop important, le temps de convergence de l'algorithme sera rapide. Nous n'avons jamais eu besoin de plus de 2n transpositions pour trouver P.

On peut envisager d'améliorer le temps de calcul en partant non pas du tableau tel qu'il est fourni, mais tel que le réorganiserait un algorithme grossier, mais très rapide. Exemple le premier des deux algorithmes présentés par GELFAUD (1971).

Autre présentation du critère $\mathfrak{M}$

Le coefficient $\mathfrak{M}$ n'est pas normalisé. Il est fonction de l'échelle dans laquelle sont exprimées les valeurs de D et d'autre part est sensible à la valeur de n.

Nous proposons un coefficient R_D normalisé variant de -1 à $+1$ qui s'apparente à un coefficient de corrélation empirique.

$$R_D = \frac{\sum_i \sum_{i'} d_{ii'}(i-i')^2 - n^2 \frac{n^2-1}{12} \bar{D}}{\left(\sum_i \sum_{i'} (i-i')^4 - n^2 \frac{n^2-1}{12}\right)\left(\sum_i \sum_{i'} d_{ii'} - n^2 \bar{D}^2\right)}$$

où $$\bar{D} = \sum_i \sum_{i'} \frac{d_{ii'}}{n^2}$$

Le dénominateur et le deuxième terme du numérateur étant constants dans une permutation simultanée des lignes et des colonnes de D, on peut donc affirmer

Maximiser $\mathfrak{M}_{PDP'}$ par rapport à P revient à maximiser $R_{PDP'}$

Nous utiliserons ce coefficient pour évaluer le degré de réorganisation des lignes ou des colonnes d'un tableau.

VI. EXEMPLES

A. Réorganisation d'un tableau symétrique (ici P = Q)

a) *Exemple 1* : Nous avons emprunté à un article d'écologie (TOURNIER et LEBRETON, 1979) le tableau ci-dessous.

Il regroupe les coefficients d'affinités entre 25 milieux écologiques. Ce tableau est en d'autres termes, la matrice des similitudes (indice de Jaccard) calculé sur un tableau à double entrée : Espèces d'oiseaux x milieu écologique.

Notre algorithme reprend l'idée des phytosociologues (KULCZYNSKI, 1927 ; AGREL, 1945 ; CULBERSON, 1955) qui est de modifier l'ordre des milieux de telle sorte que les valeurs les plus élevées du coefficient de liaison se trouvent à proximité de la diagonale. Ceci a pour effet de rapprocher les uns des autres, les milieux qui sont fortement liés.

La réorganisation du tableau et sa représentation visuelle fournissent un classement satisfaisant des relevés et permettent d'identifier facilement plusieurs types de milieux écologiques.

b) *Exemple 2* : Nous avons soumis à notre algorithme la matrice des corrélations entre variables de l'analyse des "poissons d'Amiard" (CAILLIEZ et PAGES,

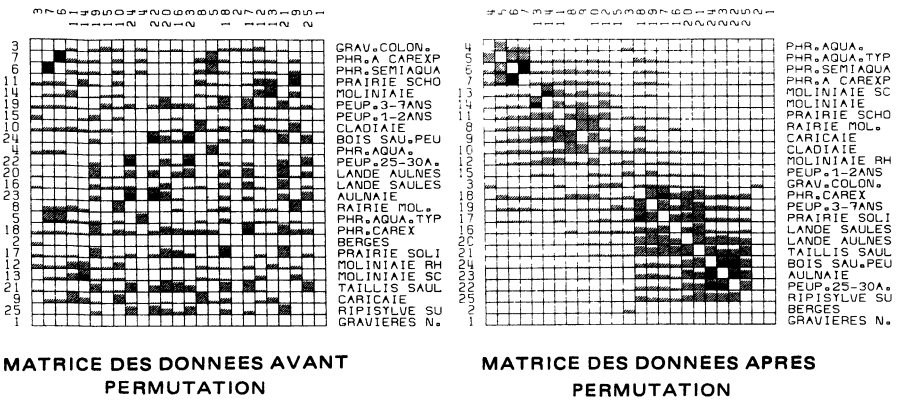


Figure 3. – Exemple article TOURNIER et LEBRETON (79). Etude du coefficient d'affinité.

1976). La matrice de départ est déjà bien structurée. Après permutation (Fig. 4), nous décelons encore mieux les orthogonalités entre variables. La marginalité de la variable 7 est également perceptible.

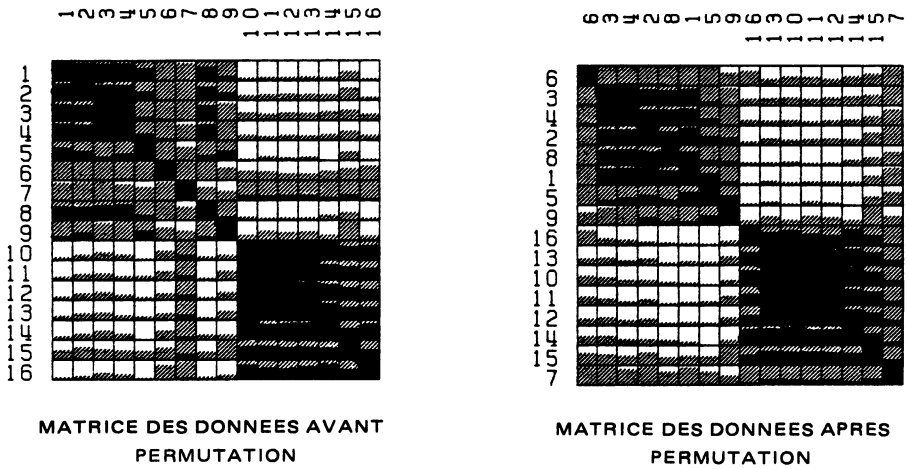


Figure 4. – Matrice des corrélations. Les poissons d'AMIARD. CAILLIEZ, PAGES (1976)

B. Exemples sur des données en 0,1

Exemple 3 : Les graphiques présentés à la figure 1 ont été obtenus par notre algorithme.

Il est important ici de montrer l'influence du choix d'une matrice de distance D_{II} (respectivement D_{JJ}) sur le résultat d'une réorganisation.

Prenons pour cela deux indices de distance que nous utiliserons successivement pour calculer D_{II} et D_{JJ} dans la réorganisation du tableau carré mais non symétrique de la figure 1. Choisissons ici la distance euclidienne et l'indice de distance de Jaccard.

Ce dernier indice introduit pour mesurer l'association d'espèces végétales (Jaccard 1908) s'exprime par l'expression

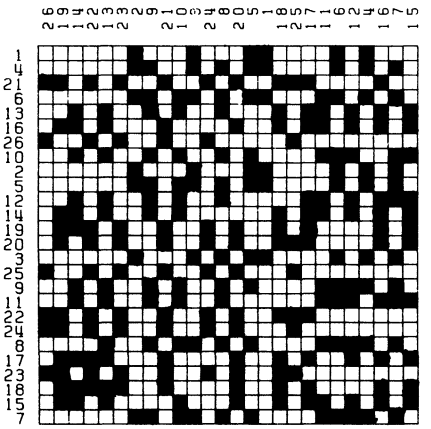
$$d(i, i') \doteq \frac{n_{ii'}}{n_i + n_{i'} - n_{ii'}} \qquad \forall (i, i') \in I \times I$$

où $n_{ii'}$ est le nombre de présences simultanées de 1 sur les lignes i et i'

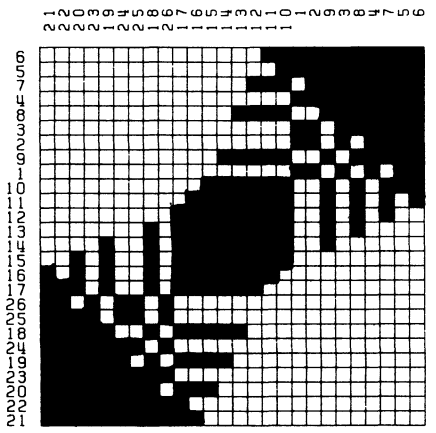
n_i est le nombre de 1 sur la ligne i

$n_{i'}$ est le nombre de 1 sur la ligne i'

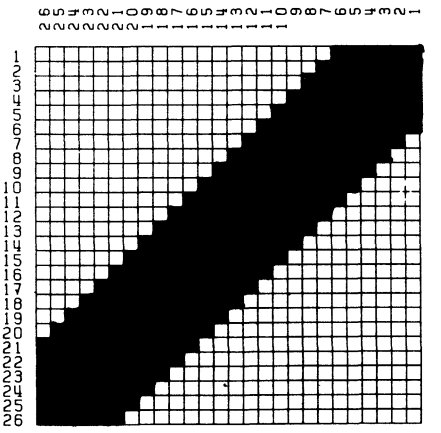
Dans les deux cas la matrice réorganisée apparaît en présentant une symétrie (Fig. 5). Dans le premier (distance euclidienne) les valeurs 0 et 1 ont joué un rôle symétrique. Il est donc normal de trouver une image du tableau réorganisé qui soit équivalente à son négatif (en terme de qualité de réorganisation) (Fig. 5).



MATRICE DES DONNEES AVANT PERMUTATION.



UTILISATION DE LA DISTANCE EUCLIDIENNE
MATRICE DES DONNEES APRES PERMUTATION



UTILISATION DE L'INDICE DE JACCARD
MATRICE DES DONNEES APRES PERMUTATION

Figure 5. – Exemple théorique I

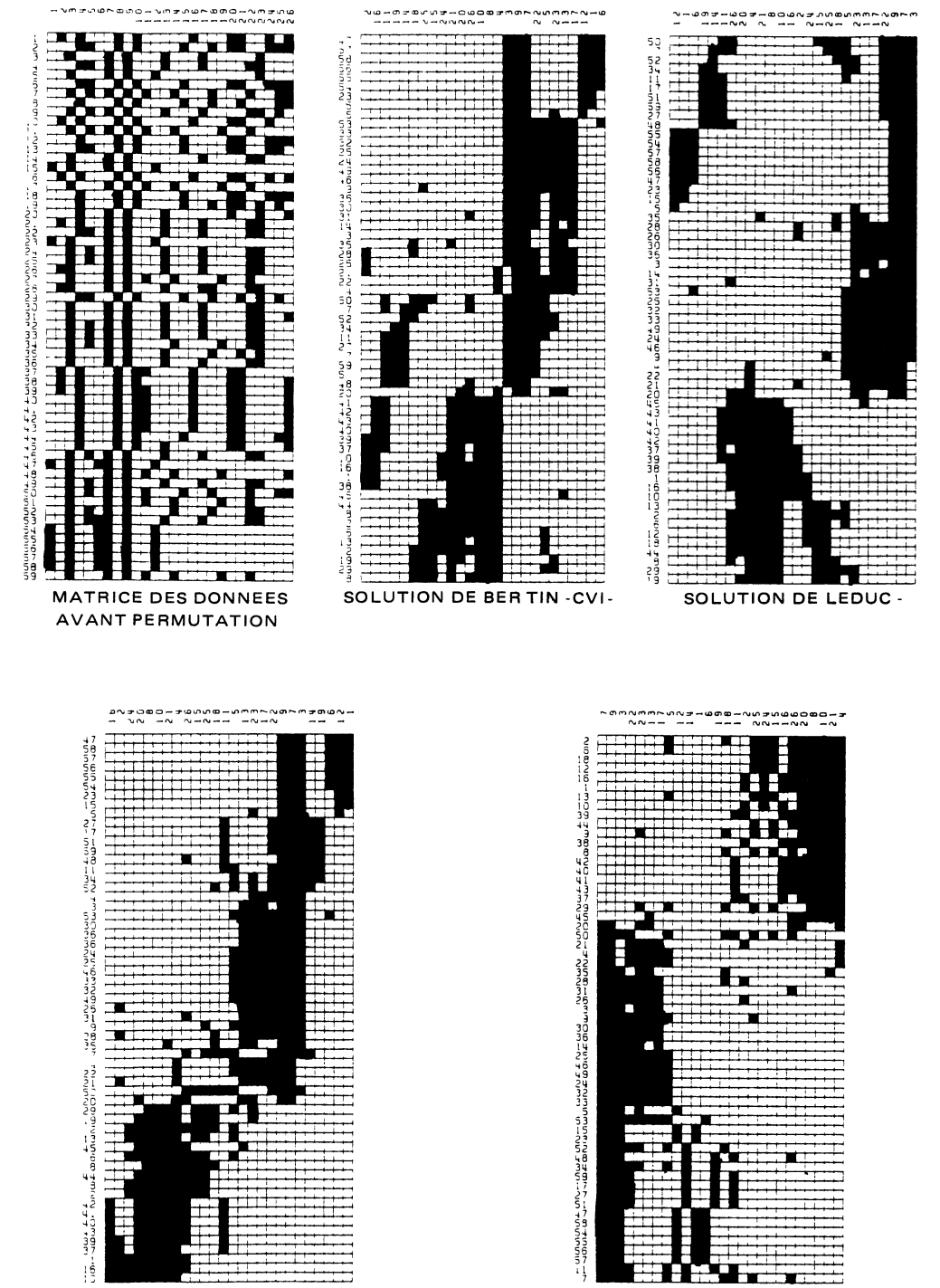


Figure 6. – Exemple article BERTIN (1980) Etude de plaques-boucles Mérovingiennes.

Par contre l'indice de Jaccard donne un rôle prépondérant à la valeur 1. On trouve alors une réorganisation qui privilégie la place des 1.

Exemple 4 : Nous présentons ici un exemple développé par BERTIN (1980) et étudié par LERMAN(1981) et LEDUC (1982). Nous avons comme à l'exemple 3, utilisé deux indices de distance et retrouvons l'effet de ce choix comme décrit plus haut.

Nous avons également représenté (Fig. 6) la solution visuelle proposée par BERTIN (1980) et celle de LEDUC (1982).

Cet exemple montre l'homogénéité des "taches" obtenues par réorganisation du critère de LEDUC. Souvenons-nous que l'objectif de cette méthode est de minimiser la longueur de la frontière des "taches noires".

Notre critère ne cherche pas l'homogénéité des taches, mais à restituer les lignes et les colonnes suivant leur ressemblance. Ainsi la col. 11 apparaît dans notre solution nettement comme une colonne intermédiaire entre deux groupes de colonnes. Ceci est moins perceptible visuellement dans la solution de LEDUC.

C. Exemple sur des données quantitatives quelconques

Exemple 5 : Les données traitées ici viennent d'un travail de J.B. DENIS (1979). Celui-ci présente plusieurs méthodes de structuration de l'interaction, dans un modèle d'analyse de la variance, par l'examen des résidus du modèle additif.

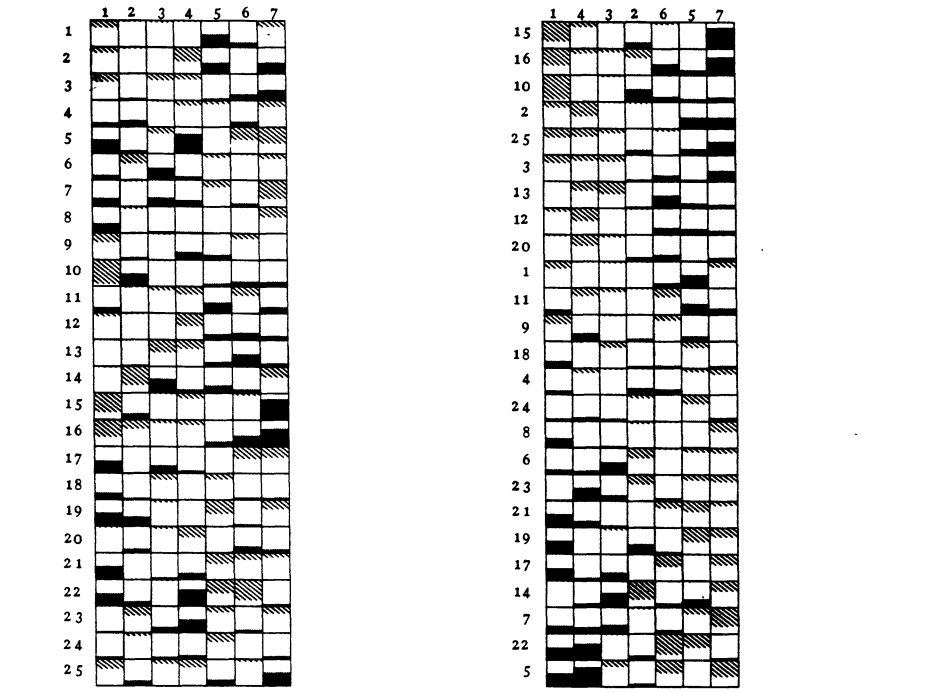


Figure 7. – Structuration de l'interaction J.B. DENIS (1979).

L'objectif fixé est de faire apparaître un partitionnement de la matrice des résidus sur lequel sera basée la structuration.

L'exemple présenté dans cet article se propose d'analyser l'interaction de 25 variétés d'orge de printemps, expérimentées dans 7 milieux répartis en Belgique, France et Grande-Bretagne. La variable étudiée est le rendement. L'interaction milieu-variété est très significative.

La matrice des résidus du modèle additif est visualisée graphiquement à la figure 7. Les parties hachurées visualisent l'intensité des résidus négatifs. Les cases noires celle des résidus positifs.

La distance retenue pour mesurer le degré de ressemblance de deux niveaux de facteurs est la distance euclidienne. Celle-ci (si le modèle est orthogonal) permet de regrouper les niveaux de facteurs ayant un comportement interactif semblable.

La réorganisation du tableau des résidus fait apparaître une structure par bloc.

On perçoit visuellement la prédominance de résidus négatifs (en grisé) en haut et à gauche et en bas à droite de la figure.

On décèle également sur ce dessin l'originalité du comportement de la variété 14 dans l'interaction aux milieux.

VII. CONCLUSION

Les exemples ont été traités par un programme d'ordinateur rédigé en FORTRAN et utilisant les sous programmes de traitement graphique GPGS.

Les temps de calculs obtenus par ce programme (cf. Fig. 8), ou mieux la clarté des graphiques obtenus, nous semblent un argument militant pour l'utilisation de cette méthode comme étude préliminaire d'un tableau de données statistiques.

	Temps de calcul des permutations (en sec.)	Temps de mise en forme des sorties graphiques sur Benson-Varian (en sec.)
Exemple 1	0,057	2,17
Exemple 2	0,02	1,45
Exemple 4	1,223	5,15
Exemple 5	0,103	1,10

Figure 8. – Statistiques obtenues sur l'ordinateur IBM 3033 du CNUSC à Montpellier.

BIBLIOGRAPHIE

R.M. ADELSON, J.M. NORMAN, G. LAPORTE (1976). – A dynamic programming formulation with diverses applications. *Operational Research Quarterly*, 1976, n° 27, pp. 119-121.

I AGRELL (1945). – The Collemboles in nests of warm-blooded animals with a method for sociological analysis. *Lund. Unio. Arsskr.* NF., n° 41, pp. 1-19.

Ph. ARABIE (1978). – Constructing Blockmodels: How and Why. *Journal of Mathematical Psychology*, n° 11, pp. 21-63.

A. BATBEDAT (1984). – Parties exogènes ou homogènes pour les graphes values et les tableaux. Cahier de DEA, n° G. Université des Sciences et Techniques du Languedoc. UER Mathématiques. Montpellier France.

J.P. BENZECRI *et Col.* (1973). – L'analyse des données. Tome 1 : *La Taxinomie* Paris, Dunod 2è éd. 1976).

B.M.D.P. (1981). – *Statistical software.* University of California Press.

J. BERTIN (1967). – *Sémiologie graphique.* Mouton-Gauthier Villars, Paris.

J. BERTIN (1971). – Article graphique. *Encyclopedia Universalis.* Mars 1971.

J. BERTIN (1977). – *La graphique et le traitement graphique* de l'information. Flammarion, Paris.

J. BERTIN (1980). – Traitements graphiques et Mathématiques. Différence fondamentale et complémentarité. *Mathématiques et Sciences Humaines*, n° 72, pp. 60-71.

S. BONIN (1977). – Les problèmes rencontrés dans l'utilisation d'une matrice ordonnable. *L'espace géographique*, n° 4, pp. 218-232.

R. BRUNET (1977). – Perception et calcul dans l'analyse typologique. *L'espace géographique*, n° 4, p. 260.

F. CAILLIEZ, J.P. PAGES (1976). – *Introduction à l'anlayse des données.* Ed. SMASH, Paris.

W.L. CULBERSON (1955). – The Corticolons communitien of lickens and bryophytes in the upland. Communities of Northerm Wisconsin. *Ecol. Monogr.*, n° 25, pp. 215-231.

P. DAGNELIE (1960). – Contribution à l'étude des communautés végétales par l'analyse factorielle. *Bull. Serv. Carte Phytogéogr.* Série B, t 5, pp. 7-71 et 93-195.

J.B. DENIS (1979). – Structuration de l'interaction. *Biom. Praxim.*, n° 19, pp. 15-34.

I.S. DUFF (1977). – A survey of sparse matrix research. *Proceedings of the IEEE*, n° 65, pp. 500-535.

S.E. ELMAGHRABY (1968). – The sequencing of "related" Jobs. *Naval Research Logistics Quaterly*, Vol. 15, pp. 23-32.

M. EYTAN (1975). – Matrices ordonnables. Une étude algébrique. Mathématiques et Sciences Humaines, n° 50, pp. 15-22.

A.E. GELFAUD (1971). – Rapid seriation methods with archaelogical application. In F.R. HODSON *et al.* (Eds) *Mathematics in the Archaeological and Historical Sciences.* Edimburgh, University Press, Edimburgh.

P.C. GILMORE (1962). – Optimal and suboptimal algorithms for the quadratic assignment problem. *Journal of Society for Industrial and applied Maths.* Vol. 10, n° 2.

L. de GOLBERG, A. CHAPPUIS, TAN CHUONG CHIN (1972). – Cartographie automatique et informatique : essai d'application de la sémiologie graphique à la sortie de cartes sur imprimantes à caractères programmables. *2è Colloque de Micro-Info-Graphique.* Université de Rouen Haute-Normandie. Sept. 1972, pp. L1-L23.

GRONOFF (1982). – Eurista, logiciel d'aide à l'interprétation de données en sciences humaines. *2è Colloque de Micro-Info-Graphique.* Université de Rouen Haute-Normandie. Sept. 1982.

M. GUINOCHET, P. CASAL (1957). – Sur l'analyse différentielle de CZEKANOWSKI et son application à la phytosociologie. *Bull. Serv. des Cartes Phytogéogr.* Série B, n° 2, pp. 25-33.

M. HANAN, J.M. KURTZBERG (1972). – A review of the placement and quadratic assignment problems. *SIAM Review,* Vol. 14, n° 2, April 1972, pp. 324-342.

S. HILLIER, M.M. CONNORS (1966). – Quadratic assgnment problem algorithms and the location of indivisible facilities. *Management Sciences,* Vol. 13, n° 1, pp. 42-57.

HODSON, KENDALL et TAUTU (1971). – *Mathematics in the archaeological and historical sciences.* Edimburg University Press, Edimburg.

F. HOL, M. SHAW (1967). – Computer analysis of chronological seriation. *Rice Univ. Studies,* Vol. 53, pp. 1-166.

L. HUBERT (1974). – Some application of graph theory and related non metric techniques to problems of approximate seriation : The cas of symmetric proximity mesures. *The Br. Jour. of Math. & Stat. Psychology,* Vol. 27, Part 2, pp. 133-153.

L.J. HUBERT, R.G. GOLLEDGE (1981). – Matrix reorganization and dynamic programming: Applications to paired comparaisons and unidimensional seriation. *Psychometrika,* Vol. 46, n° 4, pp. 420-441.

P. JACCARD (1908). – Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaud. Sci. Nat.,* t. 44, pp. 223-270.

M. JAMBU, M.O. LEBEAUX (1978). – *Classification automatique pour l'analyse des données.* Tome 2 : Logiciels. Dunod, Paris.

D.G. KENDALL (1963). – A statistical approach to Flinders Petrie's sequence dating. *Bulletin of the ISI, 34th Session,* Ohawa, pp. 657-680.

D.G. KENDALL (1969). – *Pacific journal of mathematics,* Vol. 28, pp. 565-570.

D.G. KENDALL (1971). – Seriation from abundance matrices. *Mathematics in the Archaeological and Historical Sciences.* Eds F.R. Hodson, Edimburg.

S. KULEZYNSKI (1927). – Die pflanzenassoziation der pieninen. *Bull. Intern. Acad. Polon. Sci. Lett., Classe Sci. Math. et Nat.,* Série B Sci. Nat., Suppl. 2, pp. 57-203.

G. LAPORTE (1975). – *Permutation programming: Problems, Methods and applications.* Thesis University of London.

E.L. LAWLER (1963). – The quadratic assignmenet problem. *Management Science,* Vol. 9, n° 4, pp. 586-599.

L. LEBART (1979). – L'analyse des données dans les sciences humaines. Quelques critiques et réserves. *Informatique et Sciences Humaines,* n° 40-41.

A. LEDUC (1982). – Chainage automatique des matrices ordonnables. *2è Colloque de Micro-Info-Graphique,* Université de Rouen Haute Normandie, Sept. 1982, pp. G1-G38.

J.K. LENSTRA, A.H.G. RINNOY KAN (1975). – Some simple applications of the travelling salesman problem. *Operational research quaterly,* Vol. 26, n° 4, pp. 717-733.

I.C. LERMAN (1981). – *Classification et analyse ordinale des données.* Paris, Dunod.

W.T. Jr. Mc CORMICK, P.J. SCHWEITZER, T.W. WHITE (1972). – Problem decomposition and data reorganization by a clustering technique. *Operations Research,* n° 20, pp. 993-1009.

A. MACFADYEN (1963). – *Animal Ecology.* Edition Sir Issac Pitman & Sons LTD, London.

J.F. MARCOTORCHINO et P. MICHAUD (1981). – *Agrégation de similarités en classification automatique.* IBM France, Etude n° F 003.

Mc INTOSH (1978). – *Matrix and plexus techniques. Ordination of plant communities.* Whittaker R.H. (Ed), pp. 151-184. Editeur Junk, The Hague.

W.M.F. PETRIE (1899). – Sequences in prehistoric remains. *J. Anthropol. Inst.,* n° 29, pp. 295-301.

C. RENFREW, G. STERUD (1969). – Close proximity analysis: A rapid method for the ordering of archaeological materials. *American Antiquity,* Vol. 34, pp. 265-277.

W.S. ROBINSON (1951). – A method for chornologically ordering archaeological deposits. *American Antiquity,* n° 16, pp. 293-301.

B.F. SCHRIEVER (1982). – Scaling of order dependent categorical variables with correspondence analysis. Stichting mathématisch centrum, Amsterdam.

R. SOKAL, P. SNEATH (1963). – *Principles of Numerical Taxonomy.* W.H. Freeman and Company, San Francisco.

F.A. SZCZOTKA (1972). – On a method of ordering and clustering of objects. *Zastosowania Matematyki,* Vol. 13, pp. 23-33.

H. TOURNIER, P. LEBRETON (1979). – Une approche scynécologique des milieux humides savoyards et de leurs avifaunes. *La Terre et la Vie.* Vol. 33, n° 2, pp. 275-305.

H.C. WHITE, S. BOORMAN, R.L. BREIGER (1976). – Social structure from multiple networks. I. Blockmodels of roles and positions. *American Journal of Sociology,* n° 81, pp. 730-790.