Chapitre 8

« Sériation » et classification à partir d'une analyse de la variance des proximités

I - *INTRODUCTION.*

Le problème que pose l'archéologue pour la recherche des "sériations chronologiques" est en fait un problème fréquent dans les Sciences de la Nature et de l'Homme. Un des buts du spécialiste est en effet de cerner à partir d'un tableau de données un phénomène qui évolue et la variable qui conditionne cette évolution.

Commençons par formuler de façon intuitive le problème, en nous appuyant sur un exemple tiré de l'Archéologie. On suppose établi pour la description d'un ensemble E de n tombes, un ensemble fini A de m types d'objets. La description est matérialisée par un tableau d'incidence $(\varepsilon_{\ell j})$; $\ell = 1,2,\ldots,m$ et $j = 1,2,\ldots,n$, dont l'ensemble des lignes est indexé par A et celui des colonnes par E. $\varepsilon_{\ell j} = 1$ si le type d'objet ℓ est présent dans la tombe j et 0 sinon.

Au cours du temps des divers types d'objets se succèdent ; une tombe correspondra à un âge d'autant plus reculé qu'elle contiendra les types d'objets les plus vieux. Il s'agit de découvrir sur A l'ordre chronologique. L'hypothèse H du spécialiste qui rend possible la solution du problème est la suivante :

"Un type d'objet donné existe pendant une période continue de temps ; de plus, si a_{t1}, a_{t2} et a_{t3} sont trois types apparus aux dates t_1, t_2 et t_3 avec $t_1 < t_2 < t_3$, le type a_{t2} est plus "proche" de a_{t1} que ne l'est a_{t3}, respectivement, le type a_{t2} est plus "proche" de a_{t3} que ne l'est a_{t1}". La notion de proximité entre deux types sera établie à partir du nombre de tombes possédant simultanément les deux types par rapport à ceux qui possèdent soit l'un soit l'autre.

L'hypothèse H n'est, de l'avis même de l'archéologue qu'une approximation de la réalité. Il se peut en effet qu'au cours des âges, certains types réapparaissent. Il faut toutefois espérer que H est vraie à des fluctuations assez négligeables près pour qu'on puisse appréhender le problème par la statistique. En fait, il nous sera possible de juger de la validité d'une telle hypothèse au moyen d'un test qui mesurera le degré

d'invraisemblance de l'hypothèse d'absence de structure au profit de celle définie par H ; mais il nous faut une longue expérience pour définir le seuil à adopter pour ce test.

On ne restreint en rien la généralité du problème si on suppose la condition C_o suivante.

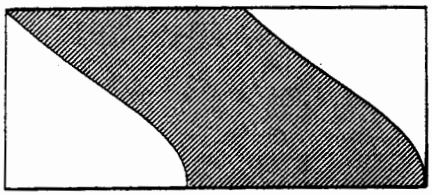
C_o : il n'existe pas deux types d'objets distincts qui soient simultanément présents ou absents de chacune des tombes.

Cette condition exprime qu'il n'existe pas deux lignes identiques de la matrice d'incidence. Si tel n'est pas le cas, on s'y ramène aisément en décrivant la matrice ligne par ligne et en supprimant tout vecteur ligne déjà rencontré. Il résultera en effet de l'analyse mathématique que la répétition d'une ligne ne fournit aucune information supplémentaire pour la solution du problème de la sériation.

La condition suivante C_1 peut paraître très restrictive ; C_1 : le nombre de tombes où un type d'objet donné apparaît est le même quel que soit le type. Cette condition signifie que le nombre de composantes égales à 1 dans un même vecteur ligne de la matrice d'incidence est constant.

Des considérations statistiques nous permettront de ramener le problème général à celui où cette condition C_1 se trouve satisfaite.

Si l'hypothèse H de l'Archéologue est exacte, il est aisé de voir qu'une permutation des lignes et des colonnes du tableau d'incidence pourra l'amener à la forme σ suivante qui tient compte des conditions C_o et C_1 ci-dessus.



La partie hachurée est celle chargée de 1.

Si k désigne le nombre commun de composantes égales à 1 dans une même ligne du tableau l'expression mathématique de cette forme σ peut être une application qui à une ligne donnée de rang ℓ, associe l'entier $c(\ell)$ tel que :

$$\varepsilon_{\ell j} = \begin{cases} 0 & \text{si } j < c(\ell) \text{ ou } j \geqslant c(\ell) + k \\ 1 & \text{si } c(\ell) \leqslant j < c(\ell) + k \ . \end{cases}$$

La fonction $\ell \longrightarrow c(\ell)$ étant strictement croissante, $c(1) = 1$ et $c(m) + k = n$.

Avant de nous engager plus avant, expliquons pourquoi ce travail ; en d'autres termes, par quoi se distingue notre point de vue des autres méthodes d'approche du problème.

Signalons d'abord la technique de J. Bertin (cf. [2]). Cette dernière est essentiellement graphique ; noircissant les cases du tableau d'incidence T où un 1 apparaît, l'expérimentateur opère, au moyen de réglettes, directement par permutation des lignes et des colonnes de façon à recon-

naître visuellement, au mieux, la forme σ sous-jacente à l'hypothèse du spécialiste. Une telle procédure est très liée à l'intuition et à l'expérience de l'opérateur, elle n'est pas automatique et ne peut par conséquent traiter de très grands tableaux.

D'autre part, certaines méthodes connues attaquent le problème de la sériation à partir de techniques complexes élaborées pour d'autres problèmes et rendent compte des résultats obtenus pour des formes particulières du tableau d'incidence. Compte tenu de la complexité de ces techniques, il est difficile d'analyser pourquoi le phénomène de la sériation se manifeste d'une façon plutôt que d'une autre.

Nous venons en fait de présenter notre principale critique relative à la méthode de D. G. Kendall (cf. [3]), que nous allons rapidement esquisser.

Le point de départ de cette méthode est un algorithme de J. B. Kruskal (cf. [4]), MDSCAL, qui tente de réaliser une idée de R. N. Shepard (cf. [9]) : "Etant donné un ensemble de n points dans un espace de dimension m, déterminer une représentation euclidienne de cet ensemble dans un espace de faible dimension de façon à respecter au mieux le système des inégalités entre les distances des points". Bien que consacré par l'expérience, l'algorithme de J. B. Kruskal ne semble pas avoir un fondement mathématique clair. L'ensemble des n points considéré par D. G. Kendall est celui des vecteurs colonnes de la matrice d'incidence ; chaque vecteur colonne qui définit la description d'une tombe est un point dans un espace de dimension m. Les distances entre les points sont donnés par la métrique euclidienne où par conséquent le produit scalaire $\sum_{\ell} \varepsilon_{\ell j} \varepsilon_{\ell k}$ définit la proximité entre les deux points représentant les tombes j et k ; c'est le nombre de types simultanément présents dans les deux tombes qui définit ainsi leur mesure de ressemblance. On applique MDSCAL pour obtenir une configuration des points dans un espace à 3 dimensions. D. G. Kendall établit la matrice d'inertie 3 × 3 du nuage des points obtenu qu'il projette sur le plan des deux premiers vecteurs propres. On constate alors, que pour une forme σ du tableau d'incidence, les points s'organisent sur ce plan en une courbe rappelant la forme du "fer à cheval" ; l'ordre des points sur la courbe définit l'ordre des colonnes ou son inverse pour σ. Cet ordre permet de retrouver celui chronologique sur l'ensemble des tombes.

La symétrie formelle entre la représentation de l'ensemble des attributs et celui de l'ensemble des objets, dans le cas d'un tableau d'incidence, permet de considérer de façon analogue le problème de la "sériation" des objets (e.g. des tombes dans l'exemple précédent) ; ce problème se formule en inversant les rôles de l'attribut et de l'objet. Nous envisageons quant à nous le problème plus direct de la sériation des types ; dans ce cas d'ailleurs, la justification statistique de la mesure de proximité que nous adopterons apparaît plus clairement. Cette mesure de proximité permet de ramener le problème à celui où la condition C_1 ci-dessus est satisfaite. Nous étudierons dans ce cas la question de l'unicité de la solution (cf. § III). Pour une classe assez générale de tableaux d'incidence admettant la forme σ, il existe une représentation, des vecteurs lignes d'un tableau par des points d'un vecteur orienté de longueur 1, telle que les distances entre les points respectent exactement les "écarts" entre les vecteurs lignes du tableau, calculés conformément à la mesure de proximité établie (cf. § IV). Le problème se trouve

ainsi réduit à la détermination d'une distribution d'un ensemble de point, dont on connaît le système des distances, sur un segment de droite orienté. Le lemme 1 du paragraphe IV permet de retrouver cette distribution à partir de l'un de ses deux points extrêmes.

Si la forme σ se manifeste par une forme en "fer à cheval" dans la méthode de D. G. Kendall ; c'est une forme parabolique qui apparait en Analyse Factorielle (en composantes principales et des correspondances). Pour une forme σ très générale qui peut même correspondre à plusieurs "dimensions" relativement "indépendantes" les unes des autres ; on peut prétendre, comme en Analyse Factorielle, en donner une représentation géométrique *qui ne nécessite pas la diagonalisation d'une matrice* et qui est basée sur une analyse simultanée de la moyenne et de la variance des proximités. Dans des cas très généraux, cette représentation a surtout un intérêt local ; elle permet d'organiser géométriquement les éléments d'une même classe, présentant une certaine cohésion, à partir de ses éléments extrémaux les moins liés, que nous appelons "pôles d'attraction". La détermination des "pôles d'attraction" a par contre, un intérêt général ; elle conduit à un riche ensemble d'algorithmes de classification d'une grande rapidité et pouvant embrasser de très gros tableaux de données. Une partie de ces algorithmes est mise en oeuvre dans le cadre d'une thèse de 3e cycle qu'a conduite H. Leredde (cf. [5]).

II - *INDICE DE PROXIMITE ENTRE LIGNES DU TABLEAU.*

Reprenons un instant le tableau d'incidence pour en rappeler les diverses formulations. Ce tableau $(\varepsilon_{\ell j})$, $1 \leq \ell \leq m$ et $1 \leq j \leq n$, formé de zéros et de uns permet le croisement des deux ensembles finis A de cardinal m et E de cardinal n. A chaque élément a de A se trouve associé une ligne du tableau qui est un vecteur logique $\varepsilon_\ell = (\varepsilon_{\ell 1},\ldots,\varepsilon_{\ell j},\ldots,\varepsilon_{\ell n})$ où $\varepsilon_{\ell j} = 1$ ou 0 selon que a "rencontre" ou non l'élément de E codé j. ε_ℓ est un point du cube $\{0,1\}^n$. De façon équivalente a est représenté par la partie E_a dont la fonction indicatrice est définie par le vecteur ε_ℓ. La matrice d'incidence nous transmet donc A comme un échantillon dans $\{0,1\}^n$ ou dans l'ensemble P(E) des parties de E. De façon semblable cette matrice nous transmet E comme un échantillon dans $\{0,1\}^m$ ou, de manière équivalente, dans l'ensemble P(A) des parties de A.

On peut regarder de façon plus symétrique et moins spatiale le tableau d'incidence comme définissant une distribution de masses égales à 1 ou 0 sur le rectangle $[1,2,\ldots,m]\times[1,2,\ldots,n]$ de N^2 où N est l'ensemble des entiers.

Si l'indice de proximité est établi relativement à la première représentation ; la seconde, plus graphique, guidera mieux notre intuition dans cette étude. Etant intéressés par la sériation des lignes du tableau; introduisons, relativement à deux d'entre elles d'indices ℓ et k, les nombres suivants : $s = \sum_{1\leq j<n} \varepsilon_{\ell j}\ \varepsilon_{kj}$ est le nombre d'"associations positives", μ_ℓ (resp. μ_k) est la proportion de composantes égales à 1 dans le vecteur ligne ε_ℓ (resp. ε_k). L'indice de proximité que nous adoptons entre ε_ℓ et ε_k sera, conformément à la formule (31), chapitre 2, paragraphe IV.1,

$$S(\ell,k) = \frac{s-n\mu_\ell\mu_k}{\sqrt{n\mu_\ell\mu_k}} \qquad (1)$$

$S(\ell,k)$ est, dans l'hypothèse N_3 d'absence de lien exprimée dans la référence citée, une réalisation d'une distribution très voisine de la loi normale centrée et réduite.

On peut se rendre compte que si on effectue, dans le tableau d'incidence des données, le changement de mesure :

$$\varepsilon_{\ell j} \longrightarrow \frac{\varepsilon_{\ell j}-\mu_\ell}{\sqrt{\mu_\ell\sqrt{n}}} = \varepsilon'_{\ell j} ,$$

la statistique

$$s = \sum_{j=1}^{n} \varepsilon_{\ell j}\ \varepsilon_{kj}$$

devient :

$$S = \sum_{j=1}^{n} \varepsilon'_{\ell j}\ \varepsilon'_{kj} \qquad (2)$$

en vertu de la relation :

$$\sum_{j=1}^{n} (\varepsilon_{\ell j}-\mu_\ell)(\varepsilon_{kj}-\mu_k) = s-n\mu_\ell\mu_k .$$

S correspond ainsi au produit scalaire euclidien sur le tableau transformé. Si μ_ℓ est constant pour tout ℓ dans A, s et S sont équivalents du point de vue de la préordonnance associée (cf. chap.2, § III). Nous établirons les différents résultats dans ce dernier cas. Puisque notre analyse est basée sur l'étude des proximités et que S corrige s lorsque le nombre d'objets où un même attribut est présent n'est pas constant ; nous étendrons au cas général l'algorithme obtenu.

Dans les paragraphes suivants III et IV on a $\mu_\ell = \mu$ pour toute ligne ℓ du tableau d'incidence.

III - *PROBLEME DE L'UNICITE DE LA SOLUTION.*

Nous allons poser quelques définitions qui nous permettront de préciser notre vocabulaire.

Relativement à une forme σ (cf. § I), posons pour tout i, $c_i = c(i)/n$, la fonction $i \longrightarrow c(i)$ a été définie au paragraphe précédent. La forme σ sera dite réduite à la forme parallélogrammique π si pour tout couple de lignes d'indices ℓ et k ($\ell<k$) on a $\frac{c_k-c_\ell}{(k-\ell)}$ = constante. Le tableau sera dit horizontalement enchaîné si pour tout couple de ligne (ℓ,k) est strictement positif l'entier

$$s(\ell,k) = \sum_{j=1}^{n} \varepsilon_{\ell j}\ \varepsilon_{\ell k} .$$

Compte tenu de la signification des lignes, la condition exprime que pour tout couple d'attributs, il existe au moins un objets qui les pos-

sède simultanément.

Le tableau sera dit faiblement horizontalement enchaîné si pour tout couple de lignes (ℓ,k), il existe une suite d'indices $(i_1,i_2,\ldots,i_r)$ telle que

$$\min\{s(\ell,i_1),\ s(i_1,i_2),\ldots,s(i_r,k)\} > 0 \ .$$

Un tableau enchaîné l'est faiblement ; si un tableau n'est pas faiblement enchaîné nous dirons qu'il est disconnexe.

Dans ce cas une composante connexe est définie comme la restriction du tableau à un ensemble maximal de lignes tel que le tableau restreint soit faiblement horizontalement enchaîné.

1. PROPOSITION. *La condition nécessaire et suffisante pour qu'un tableau d'incidence de zéros et de uns puisse être ramené à la forme* σ *est qu'il existe une permutation des colonnes telle que la partie chargée de toute ligne définisse un intervalle de l'ensemble des colonnes.*

La condition est évidemment nécessaire ; elle est aussi suffisante, l'ordre des lignes étant déterminé par la suite croissante des valeurs de $c(i)$; $i = 1,2,\ldots,m$.

A toute solution σ définie pour une permutation donnée des colonnes, il correspond bijectivement une solution σ' définie pour la permutation inverse qui détermine sur les lignes l'ordre inverse de celui correspondant à σ. Dans la pratique, une information extérieure permettra au spécialiste de choisir entre deux solutions σ et σ'.

Relativement à une permutation des colonnes qui définit une forme σ du tableau, l'ordre des lignes est défini de façon unique. Chaque ligne définit sur l'ensemble des colonnes un préordre total à trois classes dont la classe médiane correspond à l'intervalle de la ligne chargé de uns. L'intersection des différents préordres définit sur l'ensemble des colonnes un préordre total π à n classes (n étant le nombre de lignes). En disant qu'une colonne est présente dans une ligne si à leur intersection se trouve un 1 ; une même classe du préordre total π est formé des colonnes simultanément présentes ou absentes de toute ligne du tableau.

Une permutation P des colonnes définit un ordre total O sur l'ensemble des colonnes ; P sera dite compatible avec le préordre total π s'il en est ainsi de O ; c'est-à-dire ($x < y$ pour l'ordre quotient défini par π) $\Longrightarrow$ ($x < y$ pour O). Désignons toujours par π un préordre total sur l'ensemble des colonnes associé à une forme σ du tableau et soit π' le préordre total inverse où l'ordre quotient sur les classes définies par π est inversé. Il est évident que toute permutation des colonnes compatibles avec π ou π' donne au tableau une forme σ ; de plus deux permutations compatibles avec le même préordre total π se déduisent l'une de l'autre par un produit de permutations dont chacune opère sur une même classe de π.

2. PROPOSITION. *Si le tableau est faiblement horizontalement enchaîné, les seules permutations des colonnes qui laissent invariante une forme* σ *sont celles qui sont compatibles avec* π *ou* π'.

Nous allons montrer que pour une forme σ, les classes de π ainsi que l'ordre quotient, à son inverse près, sont déterminés de façon unique. En

effet une classe extrême du préordre est définie par un ensemble maximal de colonnes présentes dans exactement une ligne. Cette classe extrême va définir la première ligne du tableau à partir de laquelle seront déterminées en même temps que l'ordre total sur les lignes, les autres classes du préordre selon π ou π'. La kième ligne est celle, ℓ, parmi les lignes non encore rangées, pour laquelle est maximum le nombre de colonnes simultanément présentes dans les lignes (k-1) et ℓ. L'ordre total sur les lignes permet de définir le préordre total π ou π' selon que la classe extrême initialement considérée est la première ou la dernière de π.

En conclusion, si une forme σ du tableau existe, l'ordre total sur les lignes permettant de l'obtenir est déterminé à l'ordre inverse près ; d'autre part, la proposition précédente nous permet, à partir d'une forme σ obtenue, d'énumérer toutes les permutations de colonnes définissant une forme σ.

IV - *REPRESENTATION SUR UN SEGMENT DE DROITE ORIENTE.*

On ne restreint en rien la généralité de ce qui va suivre en supposant le segment de droite de longueur 1.

1. LEMME. *Soit* $\{c_1, c_2, \ldots, c_m\}$ *l'ensemble des abscisses de* m *points répartis de façon quelconque sur un segment de droite orienté* $\overrightarrow{AB}$ *de longueur* 1 ; $c_1 < c_2 < \ldots < c_m$. *On a* (α) *: celui des* m *points dont la moyenne des distances à l'ensemble des points est la plus grande est nécessairement un point extrême.* (β) *: celui des* m *points dont la variance des distances à l'ensemble des points est la plus grande et aussi nécessairement un point extrême.*

(α) Moyenne des distances d'un point

Désignons chacun des m points par son rang en allant de A vers B, le point i étant ainsi d'abscisse c_i.

Nous allons comparer la moyenne des distances de 1 à celle de ℓ pour $\ell \leq m/2$. La suite des distances de 1 est :

$$0, c_2-c_1, c_3-c_1, \ldots, c_\ell-c_1, c_{\ell+1}-c_1, \ldots, c_m-c_1,$$

d'où la moyenne des distances de 1 :

$$M(1) = \frac{1}{m}\left(\sum_{i=1}^{m} c_i - mc_1\right) .$$

La suite des distances de ℓ est :

$$c_\ell-c_1, c_\ell-c_2, \ldots, c_\ell-c_{(\ell-1)}, 0, c_{(\ell+1)}-c_\ell, c_{\ell+2}-c_\ell, \ldots, c_m-c_\ell.$$

d'où la moyenne des distances de ℓ :

$$M(\ell) = \frac{1}{m}\left((2\ell-m)c_\ell - \sum_{i=1}^{\ell} c_i + \sum_{i=\ell+1}^{m} c_i\right) .$$

La différence :

$$M(1)-M(\ell) = \frac{1}{m}\left((m-2\ell)c_\ell + 2\sum_{i=1}^{\ell} c_i - mc_1\right)$$

$$= \frac{1}{m}((m-2\ell)(c_\ell - c_1) + 2\sum_{i=1}^{\ell}(c_i - c_1))$$

or :

$$m-2\ell \geqslant 0, c_\ell - c_1 > 0 \quad \text{et} \quad c_i - c_1 \geqslant 0 \quad \text{pour} \quad 1 \leqslant i \leqslant \ell ,$$

donc :

$$M(1)-M(\ell) > 0 \quad \text{pour tout } \ell \text{ tel que } 2\ell \leqslant m .$$

L'inégalité est encore vraie si m est impair et si $2\ell = m + 1$; en effet, on alors :

$$M(1)-M(\ell) = \frac{1}{m}((c_\ell - c_1)+2\sum_{i=1}^{(\ell-1)}(c_i - c_1))$$

finalement :

$$M(1) = \max_{1\leqslant\ell\leqslant[(m+1)/2]} M(\ell)$$

où $[(m+1)/2]$ désigne la partie entière de $(m+1)/2$.

Symétriquement :

$$M(m) = \max_{\ell>[(m+1)/2]} M(\ell).$$

La partie (α) du lemme se trouve démontrée. Par conséquent, si la fonction distance est donnée, l'ordre des points, ou son inverse, sur le segment orienté, peut être déterminé à partir du point dont la moyenne des distances est la plus grande.

(β) Variance des distances d'un point.

Comparons la variance des distances de 1 à celle de ℓ. La variance des distances de 1 est :

$$V(1) = \frac{1}{m}\sum_{i=1}^{m}(c_i - c_1)^2 - \frac{1}{m^2}\left(\sum_{i=1}^{m} c_i - mc_1\right)^2 .$$

Prenons le point 1 pour origine et posons $d_i = c_i - c_1$, on a :

$$V(1) = \frac{1}{m}\sum_{i=1}^{m} d_i^2 - \left(\frac{1}{m}\sum_{i=1}^{m} d_i\right)^2 .$$

La variance des distances de ℓ est :

$$V(\ell) = \frac{1}{m}\sum_{i=1}^{m}(c_\ell - c_i)^2 - \frac{1}{m^2}\left[\sum_{i=1}^{\ell}(c_\ell - c_i) + \sum_{i=\ell+1}^{m}(c_i - c_\ell)\right]^2$$

soit :

$$V(\ell) = \frac{1}{m}\sum_{i=1}^{m}(d_\ell - d_i)^2 - \frac{1}{m^2}\left[\sum_{i=1}^{\ell}(d_\ell - d_i) + \sum_{i=\ell+1}^{m}(d_i - d_\ell)\right]^2 .$$

Calculons la différence $V(1)-V(\ell)$,

$$V(1)-V(\ell) = \frac{d_\ell}{m}\sum_{i=1}^{m}(2d_i - d_\ell) - \frac{1}{m^2}\left[2\sum_{i=1}^{\ell} d_i + (m-2\ell)d_\ell\right] \times \left[2\sum_{i=\ell+1}^{m} d_i - (m-2\ell)d_\ell\right] .$$

Un calcul, dont nous ne donnons pas le détail ici, montre que :

$$V(1)-V(\ell) = \frac{4}{m^2} d^2 \left(\sum_{\ell+1}^{m} e_j - (m-\ell) \right) \left(\ell - \sum_{1}^{\ell} e_i \right)$$

où $e_j = d_j / d_\ell$.

Comme $e_i > 1$ (resp. $e_i < 1$) pour $i > \ell$ (resp. $i \leq \ell$) ; on a bien

$$V(1)-V(\ell) \geq 0$$

et la partie (β) du lemme se trouve ainsi démontrée. H. Leredde a établi une propriété analogue lorsqu'on prend pour V le moment absolu d'ordre 2 au lieu de la variance.

Par conséquent, le point pour lequel V est maximum nous permettra de retrouver l'ordre, ou son inverse, des points. En effet, à partir du point extrême on utilisera l'algorithme des "enchaînements successifs" où à chaque pas on détermine le plus voisin du dernier retenu.

2. EXPRESSION DE LA MESURE DE PROXIMITE POUR UNE FORME σ.

Remettons-nous en mémoire le tableau d'incidence et reprenons le problème de la "sériation" où on cherchera à découvrir l'ordre total sur les lignes du tableau. Avec les notations définies précédemment (cf. § III), l'expression de la mesure de proximité entre deux lignes est :

$$S(\ell,k) = \sum_{j=1}^{n} \frac{\varepsilon_{\ell j} - \mu}{\sqrt{\mu \sqrt{n}}} \times \frac{\varepsilon_{kj} - \mu}{\sqrt{\mu \sqrt{n}}} = \frac{s - n\mu^2}{\mu \sqrt{n}}$$

puisqu'on suppose $\mu_\ell = \mu$ pour tout ligne ℓ du tableau.

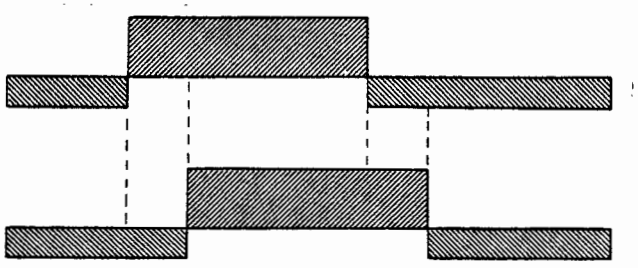
Rappelons le changement de mesure :

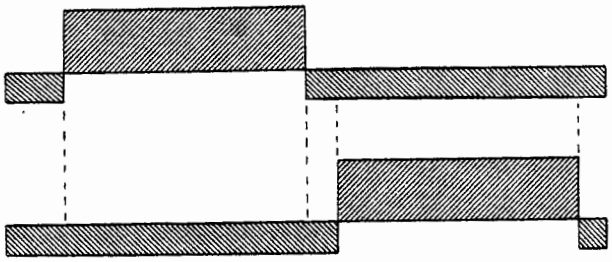
$$\varepsilon_{\ell j} \longrightarrow \varepsilon'_{\ell j} = \frac{\varepsilon_{\ell j} - \mu}{\sqrt{\mu \sqrt{n}}}$$

pour lequel :

$$S(\ell,k) = \sum_{j} \varepsilon'_j \, \varepsilon'_{kj} \ .$$

Supposons que le tableau d'incidence admette la forme σ et désignons par ℓ et k ($\ell < k$), les rangs respectifs de deux lignes pour σ. Les deux lignes peuvent se présenter, après le changement de mesure $\varepsilon_{\ell j} \longrightarrow \varepsilon'_{\ell j}$, sous l'une des deux formes suivantes, où les hachures ////// expriment une charge positive égale à $(1-\mu)/\sqrt{\mu \sqrt{n}}$ et où celles \\\\\\ , une charge négative, égale à $-\sqrt{\mu / \sqrt{n}}$. Le cas a) est caractérisé par $c_k \leq c_\ell + \mu$, le cas b) par $c_k > c_\ell + \mu$.





Les diverses valeurs de $\varepsilon'_{\ell j}\varepsilon'_{kj}$ sont :

$$\begin{aligned}
\mu/\sqrt{n} \quad &\text{si} \quad \varepsilon_{\ell j} = \varepsilon_{kj} = 0 \\
-(1-\mu)/\sqrt{n} \quad &\text{si} \quad \varepsilon_{\ell j} = 0, \varepsilon_{kj} = 1 \\
-(1-\mu)/\sqrt{n} \quad &\text{si} \quad \varepsilon_{\ell j} = 1, \varepsilon_{kj} = 0 \\
(1-\mu)^2/\mu\sqrt{n} \quad &\text{si} \quad \varepsilon_{\ell j} = \varepsilon_{kj} = 1 \ .
\end{aligned}$$

De sorte que l'expression de la mesure de proximité $S(\ell,k)$ est :

a') dans le cas a) où $c_k \leqslant c_\ell + \mu$:

$$S(\ell,k) = \sum_{1\leqslant j\leqslant nc_\ell} \mu/\sqrt{n} - \sum_{nc_\ell+1\leqslant j\leqslant nc_k} (1-\mu)/\sqrt{n} + \sum_{nc_k+1\leqslant j\leqslant n(c_\ell+\mu)} (1-\mu)^2/\mu\sqrt{n}$$

$$- \sum_{n(c_\ell+\mu)+1\leqslant j\leqslant n(c_k+\mu)} (1-\mu)/\sqrt{n} + \sum_{n(c_k+\mu)+1\leqslant j\leqslant n} \mu/\sqrt{n}$$

$$= \frac{1}{\sqrt{n}} \{[nc_\ell+n(1-c_k-\mu)]\,\mu + [n(c_\ell+\mu-c_k)](1-\mu)^2/\mu - 2n(c_k-c_\ell)(1-\mu)\}$$

$$= \frac{\sqrt{n}}{\mu} \{[(c_\ell-c_k)+(1-\mu)]\,\mu^2 + [(c_\ell-c_k)+\mu]\,(1-\mu)^2 + 2n(c_\ell-c_k)(1-\mu)\}$$

d'où en regroupant on obtient :

$$S(\ell,k) = \frac{\sqrt{n}}{\mu}\ [\mu(1-\mu)-(c_k-c_\ell)] \ .$$

La valeur de $S(\ell,k)$ pour $c_k = c_\ell+\mu$ est $-\ n\mu$; d'autre part, $S(\ell,k) > -\ n\mu$ pour $c_k \leqslant c_\ell+\mu$.

b') dans le cas b) où $c_k > c_\ell+\mu$

$$S(\ell,k) = \frac{1}{\sqrt{n}} \{nc_\ell\mu - n\mu(1-\mu) + n(c_k-c_\ell-\mu)\mu - n\mu(1-\mu) + n(1-c_k-\mu)\mu\}$$

$$= -\sqrt{n}\mu$$

finalement :

$$S(\ell,k) = \begin{cases} \frac{\sqrt{n}}{\mu}\,[\mu(1-\mu)-(c_k-c_\ell)] & \text{si} \quad c_k \leqslant c_\ell + \mu \\ -\sqrt{n}\mu & \text{si} \quad c_k \geqslant c_\ell + \mu \end{cases}$$

Introduisons l'"écart" $D(\ell,k)$:

$$D(\ell,k) = -\frac{\mu}{\sqrt{n}}\, S(\ell,k) + \mu(1-\mu) \; ;$$

on a :

$$D(\ell,k) = \begin{cases} c_k - c_\ell & \text{si} \quad c_k \leqslant c_\ell + \mu \\ \mu & \text{si} \quad c_k \geqslant c_\ell + \mu \; . \end{cases}$$

2.1. THEOREME. *Si un tableau d'incidence remplit les conditions*

(i) *le nombre de* 1 *dans toute ligne est le même (i.e.* $\mu_\ell = \mu$ *pour tout* ℓ*),*

(ii) *le tableau est horizontalement enchaîné,*

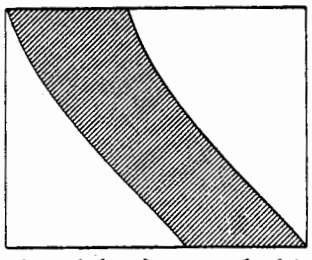
(iii) *le tableau admet la forme* σ*,*

alors, il existe une représentation, sur un segment de droite orienté de longueur 1*, des lignes du tableau par des points du segment dont l'ordre est celui relatif à* σ *et dont les distances sont les écarts* $D(\ell,k)$.

En effet, la condition (ii) est équivalente à $c_m < c_1 + \mu$. Le lemme ci-dessus nous permet de conclure.

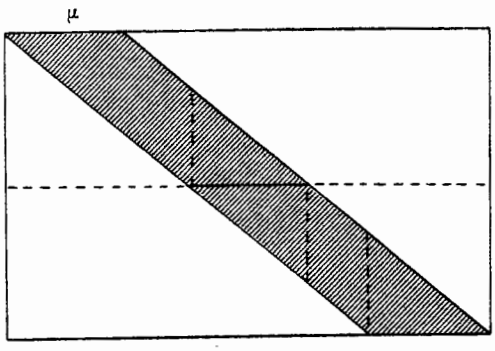
2.2. THEOREME. *Si un tableau d'incidence remplit les conditions* (i), (ii) *et* (iii) *du théorème précédent, on a* (α) *celle des* n *lignes dont la moyenne des proximités* S *à l'ensemble des lignes est la plus faible est nécessairement une ligne extrême pour la forme* σ *du tableau ; de même* (β) *celle des* m *lignes dont la variance des proximités à l'ensemble des lignes est la plus grande est nécessairement une ligne extrême pour la forme* σ.

Dans le cas d'un tableau faiblement enchaîné comme l'indique la figure suivante, où les hachures définissent la partie chargée du tableau ; on peut montrer aisément que la ligne dont la moyenne des écarts D est la plus grande est nécessairement une ligne extrême.



Mais la propriété (β) du théorème précédent n'est plus vérifiée en général dans ce cas. Pour le montrer, nous allons examiner le cas d'un tableau faiblement enchaîné ayant une forme parallélogrammique π suffisamment inclinée pour que l'intervalle chargé de la ligne médiane ait

avec chacun des intervalles chargés des deux lignes extrêmes une intersection vide.



Dans les calculs qui vont suivre, 1 désignera la première ligne du tableau et ℓ, la ligne médiane. Nous allons comparer d'une part les moyennes ; d'autre part, les variances des écarts de 1 et de ℓ à l'ensemble des lignes du tableau.

Rappelons que pour le cas d'une forme π :

$$D(\ell,k) = \begin{cases} c_k - c_\ell = \alpha(k-\ell) & \text{si} \quad c_k \leqslant c_\ell + \mu \\ \mu & \text{si} \quad c_k \geqslant c_\ell + \mu \end{cases}$$

où $k \geqslant \ell$ et où α est une constante.

La suite des valeurs des écarts de 1 est donc :

$$\alpha, 2\alpha, \ldots, (\tfrac{\mu}{\alpha}-1)\,\alpha, \mu, \mu, \ldots, \mu \ ;$$

la suite se termine par $(m - \frac{\mu}{\alpha})$ termes tous égaux à μ.

La suite des valeurs des écarts de ℓ est :

$$\alpha, 2\alpha, \ldots, (\tfrac{\mu}{\alpha} - 1)\alpha, \alpha, 2\alpha, \ldots (\tfrac{\mu}{\alpha} - 1)\alpha, \mu, \mu, \ldots, \mu \ ;$$

la suite comprend deux fois la séquence $\alpha, 2\alpha, \ldots, (\frac{\mu}{\alpha} - 1)\alpha$ et se termine par $(m - 2\frac{\mu}{\alpha} + 1)$ termes tous égaux à μ.

Moyennes des écarts de 1, M(1)

Posons $q = \mu/\alpha$; on a :

$$\begin{aligned} M(1) &= \frac{1}{m}\,[\alpha q(q-1)/2 + (m-q)\mu] \\ &= \frac{1}{m}\,[\mu(m - \tfrac{q+1}{2m})\,] \\ &= \mu(1 - \tfrac{q+1}{2m}) \ . \end{aligned}$$

Moyenne des écarts de ℓ, M(ℓ)

$$\begin{aligned} M(\ell) &= \frac{1}{m}\,[\alpha q(q-1) + (m-2q+1)\mu] \\ &= \mu(1 - \tfrac{q}{m}). \end{aligned}$$

D'où :

$$M(1) > M(\ell).$$

Variance des écarts de 1, V(1) et de ℓ, V(ℓ)

$$V(1) = \frac{1}{m}\left[\sum_{1\leqslant k\leqslant(q-1)} \alpha^2k^2+(m-q)\mu^2\right] - \mu^2\left[1-(q+1)/2m\right]^2$$

$$V(\ell) = \frac{1}{m}\left[2\sum_{1\leqslant k\leqslant(q-1)} \alpha^2k^2+(m-2q+1)\mu^2\right] - \mu^2(1-\frac{q}{m})^2 \ .$$

Etudions le signe de la différence V(ℓ) - V(1).

$$V(\ell)-V(1) = \frac{1}{m}\left[\alpha^2(q-1)q(2q-1)/6-(q-1)\mu^2\right]-\mu^2\left\{(1-\frac{q}{m})^2-[1-(q+1)/2m]^2\right\} .$$

Un calcul élémentaire nous montre que :

$$V(\ell)-V(1) \backsim \frac{\mu^2q}{3m} - \frac{\mu^2q}{m} - \mu^2\left[(1-\frac{q}{m})^2 - (1-\frac{q}{2m})^2\right]$$

l'écart entre le second et le premier nombre étant de l'ordre de $\mu^2/2m$. Le membre de droite peut se mettre sous la forme :

$$\frac{\mu^2q}{m}\left[-\frac{2}{3} + (1-\frac{3}{4}\frac{q}{m})\right]$$

$$= \frac{\mu^2q}{m}(\frac{1}{3} - \frac{3}{4}\frac{q}{m}) \ .$$

Donc : V(ℓ) > V(1) si $\frac{q}{m} > \frac{4}{9}$ c'est-à-dire $\frac{\mu}{\alpha} < \frac{4}{9}m$.

2.3. PROPOSITION. *Si un tableau d'incidence admet une forme π pour laquelle $\mu/\alpha < 4m/9$, pour cette forme, la variance des écarts de la ligne médiane est supérieure à celle d'une ligne extrême.*

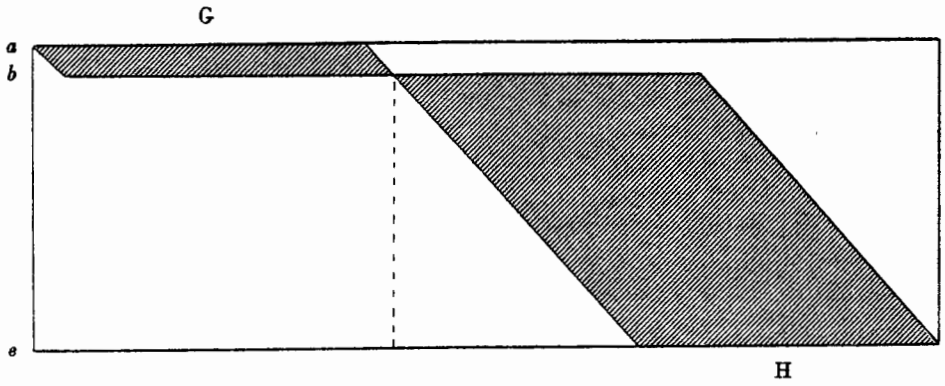
Pour un tableau faiblement horizontalement enchaîné et admettant une forme σ comme l'indique la figure ci-dessus, l'ordre des lignes, ou son inverse, peut être déterminé en utilisant l'algorithme des "enchaînements successifs" à partir d'une ligne extrême ℓ_1, définie comme étant celle dont la moyenne des écarts est la plus grande. Cet algorithme consiste à retenir au kième pas la ligne ℓ_k la plus proche de celle ℓ_{k-1}. Le théorème de représentation 2.1 n'est plus vrai si un tableau remplit les conditions (i), (iii), mais n'est que faiblement horizontalement enchaîné.

V - ANALYSE SIMULTANEE DE LA MOYENNE ET DE LA VARIANCE DES PROXIMITES.

Nous allons examiner quelques exemples où la forme σ fait apparaître des blocs parallélogrammiques ; deux blocs successifs étant tels que la dernière ligne du premier et la première ligne du second aient en commun pour tout au plus un petit intervalle de l'ensemble des colonnes, une charge simultanée égale à 1. Ces exemples, pour lesquels $\mu_\ell = \mu$ pour tout ℓ, nous permettront de préciser un algorithme de représentation géométrique des tableaux d'incidence des données qu'on appliquera dans le cas général.

1. EXEMPLE 1.

Soit le tableau d'incidence ramené à la forme σ comme il est indiqué dans la figure.



La partie hachurée est celle qui est chargée des uns ; elle est constituée de deux blocs G et H ayant la forme de parallélogramme. p est l'indice de la ligne qui termine le bloc G ; on a p = m/10.(p+1) est l'indice de la ligne qui commence le bloc H. On supposera dans les calculs qui vont suivre n grand ; en tout cas assez grand pour que, dans ces calculs on puisse confondre (p-1) ou (p+1) avec p, sans effet sensible sur le résultat.

Soit comme il est indiqué dans la figure, μ = 1/3.

Lorsque les lignes ℓ et k (ℓ < k), appartiennent à un même bloc, on a :

$$\frac{c_k - c_\ell}{(k-\ell)} = \text{constante } \alpha \text{ où } \alpha = 1/3m$$

a est le vecteur représenté par la première ligne du tableau ; b, celui représenté par la (m+1)ème ligne et e , celui représenté par la dernière ligne.

Un calcul analogue à celui qui a abouti à la proposition 2.3 du paragraphe précédent, permet d'établir le tableau des valeurs :

M(a) = 0,302	M(b) = 0,1684
V(a) = 0,00899	V(b) = 0,00989.

M(a) (resp. M(b)) est la moyenne des écarts de a (resp. de b) ;
V(a) (resp. V(b)) est la variance des écarts de a (resp. de b).

Ainsi M(a) est sensiblement plus grand que M(b) alors que V(b) > V(a). Dans le cas de notre talbeau, il est surtout intéressant de découvrir la présence des deux blocs, quitte par la suite à ranger les lignes de chacun d'entre eux. Certes, une classification automatique permet de détecter les deux blocs ; mais ici, nous l'effectuerons plus simplement par une analyse simultanée de la variance et de la moyenne des proximités. En effet, à partir de b dont la variance des proximités est maximum, a peut être obtenu parmi les vecteurs lignes différents de b qui rendent maximum toute fonction du couple (V(x),S(x,b)) croissante par rapport à V(x) et décroissante par rapport à $|S(x,b)|$, où S(x,b) est la mesure de proximité entre x et b. Si le choix d'une telle fonction peut être indifférent lorsque la forme σ, formée de plusieurs blocs, est définie, comme ci-dessus, mathématiquement ; il devient crucial lorsque la forme σ n'est que floue et

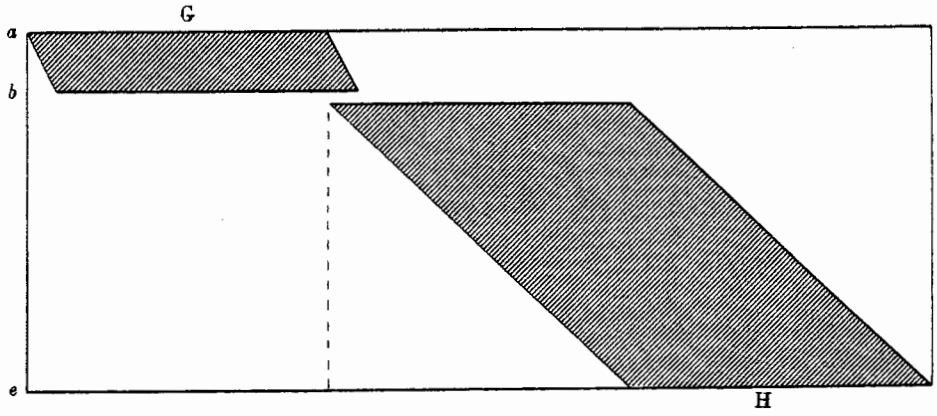
domine statistiquement. Nous avons commencé par proposer comme fonction de discrimination le rapport homogène $V(x)/(S(x,b))^2$; mais une analyse expérimentale nous a permis de nous rendre compte qu'il est plus naturel de proposer une statistique de même dimension que celle qui sert à déterminer le premier pôle b ; c'est-à-dire, la dimension d'une variance. Par conséquent le second pôle d'attraction d'une classe sera défini par

$$\{V(c)/S(c,b)\}^2 = \max_{x \neq b} \{V(x)/S(x,b)\}^2 .$$

Dans notre exemple, ce sont les deux lignes extrêmes du bloc G qui réalisent le maximum de la quantité critère ci-dessus. En appelant, dans l'exemple considéré, I. (resp. J), l'ensemble totalement ordonné des colonnes associées à G (resp. H) ; on remarque que le tableau conserve la forme σ si on place J avant I (cf. § IV). Ce ne sera plus le cas pour le second exemple que nous allons envisager et qui nous permettra d'améliorer notre intuition du problème.

2. EXEMPLE 2.

Le tableau d'incidence se présente comme suit lorsqu'il est ramené à la forme σ.



L'indice p de la ligne qui termine le bloc G est ici égal à m/6 où m est le nombre total de lignes. Les lignes p et (p+1) ont en commun pour une tranche de l'ensemble des colonnes une charge simultanée égale à 1. m est supposé assez grand et on a $\mu = 1/3$.

Lorsque les lignes ℓ et k appartiennent au même bloc G, on a :

$$\frac{c_k - c_\ell}{(k-\ell)} = \text{constante } \alpha, \text{ où } \alpha = 1/5m .$$

D'autre part, lorsque les lignes ℓ et k appartiennent au même bloc H, on a :

$$\frac{c_k - c_\ell}{(k-\ell)} = \text{constante } \beta, \text{ où } \beta = 2/5m$$

a est toujours le point représenté par la première ligne du tableau ;

b celui représenté par la (p+1)ème ligne. Les résultats du calcul,

présentés avec les notations adoptées ci-dessus sont :

$$M(a) = 0,280 \qquad M(b) = 0,192$$
$$V(a) = 0,014 \qquad V(b) = 0,011.$$

On a ici $M(a) > M(b)$ et $V(a) > V(b)$; toutefois le rapport $M(a)/M(b)$ est sensiblement plus élevé que celui $V(a)/V(b)$.

L'algorithme des enchaînements successifs aurait permis ici de reconstituer l'ordre des lignes en partant de a dont la moyenne des écarts M est maximum. Mais cet algorithme ne met pas en évidence la présence des deux blocs G et H ; il faut pour cela recourir à une analyse simultanée de la variance et de la moyenne des proximités comme il a été fait allusion dans l'exemple ci-dessus. En effet, pour le point b il se produit un saut positif brutal de :

$$\{V(x)/S(x,a))\}^2$$

lorsque x parcourt l'ensemble des vecteurs lignes du tableau différents de a et rangés par proximité décroissante à a. Le vecteur ligne pour lequel le rapport précédent est maximum est b ou e (il faut faire le calcul pour e) ; de toute façon l'un quelconque de ces deux points permet de définir la dimension sous-jacente à H ; c'est-à-dire, l'ordre ou son inverse des attributs de description représentés par les lignes de H ramené à la forme σ.

3. FORMULE D'ANALYSE DE LA VARIANCE DES PROXIMITES.

Considérons le tableau carré $m \times m$ des proximités $S_{\ell k}$:

$$S(\ell,k) = \frac{s - n\mu_\ell \mu_k}{\sqrt{n\mu_\ell \mu_k}}$$

Nous allons commencer par effectuer une analyse de la variance globale des proximités selon les lignes du tableau carré auquel on aura ôté le contenu de la diagonale (cette analyse est d'ailleurs identique à celle selon les colonnes du tableau carré).

Posons :

$$\overline{S}_\ell = \frac{1}{(m-1)} \sum_{\{k/k\neq\ell\}} S_{\ell k} = \text{moyenne des proximités de } \ell.$$

$$\overline{S} = \frac{1}{m} \sum_{1\leq\ell\leq m} \overline{S}_\ell = \text{moyenne globale des proximités.}$$

Décomposons la différence $(S_{\ell k} - \overline{S})$ comme suit :

$$(S_{\ell k} - S) = (S_{\ell k} - \overline{S}_\ell) + (\overline{S}_\ell - \overline{S})$$
$$(S_{\ell k} - \overline{S})^2 = (S_{\ell k} - \overline{S})^2 + (\overline{S}_\ell - \overline{S}^2) + 2(S_{\ell k} - \overline{S}_\ell)(\overline{S}_\ell - \overline{S})$$

en sommant par rapport à k et pour $k \neq \ell$, on obtient :

$$\sum_{\{k/k\neq k\}} (S_{\ell k} - \overline{S})^2 = \sum_{\{k/k\neq\ell\}} (S_{\ell k} - \overline{S}_\ell)^2 + (n-1)(\overline{S}_\ell - \overline{S})^2 + 0 .$$

en sommant par rapport à ℓ et en divisant par $m(m-1)$ les deux membres

on a :

$$\frac{1}{m(m-1)} \sum_{\{(\ell,k)/\ell\neq k\}} (S_{\ell k}-\overline{S})^2 = \frac{1}{n} \sum_{1\leq\ell\leq m} \frac{1}{(m-1)} \sum_{\{k/k\neq\ell\}} (S_{\ell k}-\overline{S}_\ell)^2 + \frac{1}{m} \sum_{1\leq\ell\leq m} (S_\ell-\overline{S})^2.$$

Dans cette formule le premier membre définit la dispersion du nuage des points représentant les attributs dans le simplexe P(E) (cf. § II). Cette dispersion se décompose d'une part, en la moyenne des variances des proximités de chacun des éléments descriptifs avec les autres (variances intra-lignes) et d'autre part, la variance inter-lignes ; cette dernière est une mesure de la distorsion par rapport à un état sphérique des données défini par $\overline{S}_\ell$ constant pour tout ℓ, $S(\ell,k)$ est un produit scalaire pour le tableau transformé $\varepsilon_{\ell k} \longrightarrow \varepsilon'_{\ell k}$ (cf. § II).

3.1. Tests.

En se référant à l'hypothèse N (cf. § II), on peut aisément effectuer des tests.

1) Test d'absence de structure vis-à-vis de l'hypothèse de la "sériation" où le tableau peut être, au contenu d'un petit nombre de cases près, être ramené à la forme σ.

2) Test d'absence de structure par rapport à celle définie par une disposition sphérique des données pour la métrique qui nous intéresse.

Le premier test sera basé sur la plus grande valeur observée de $\sum_{\{k/k\neq\ell\}} (S_{\ell k}-\overline{S}_\ell)^2$ et le second sur la valeur de $\sum_{1\leq\ell\leq m} (S_\ell-S)^2$.

Si le tableau des données n'est pas de grande dimension, on peut effectuer les tests à partir de simulations du tableau d'incidence dans l'hypothèse N , un nombre suffisant de fois ; nous disposons d'un programme qui permet de le faire. On comparera les valeurs observées des statistiques avec leurs distributions empiriques.

Autrement, on peut constater que pour p grand, dans l'hypothèse N, la suite des valeurs :

$$S(\ell,1),S(\ell,2),\ldots,S(\ell,\ell-1),S(\ell,\ell+1),\ldots,S(\ell,m) \qquad (*)$$

des proximités d'une ligne ℓ fixée avec les autres lignes du tableau, peut être considérée comme une suite de (m-1) réalisations indépendantes d'une variable aléatoire normale centrée réduite ; de sorte que $\sum_{\{k/k\neq\ell\}} (S_{\ell k}-\overline{S})^2$ est pour ℓ fixé, la réalisation d'un χ^2 à (m-2) degrés de liberté. Les m valeurs de cette statistique obtenues pour ℓ variant de 1 à m, ne sont pas rigoureusement indépendantes dans l'hypothèse N ; en effet, pour les n suites telles que (*), chaque $S(\ell,k)$ pour ℓ différent de k, se retrouve dans exactement deux suites différentes. Toutefois ce degré de dépendance est faible et l'est d'autant que m est grand. Par conséquent, on se référera à la distribution de la plus grande valeur de m statistiques indépendantes du χ^2 à (m-2) degrés de liberté pour juger de l'importance relative de la plus grande valeur observée de $\sum_{\{k/k\neq\ell\}} (S_{\ell k}-\overline{S}_\ell)^2$. De même

on se référera à la loi du χ^2 à (m-1) degrés de libertés pour juger de la relative petitesse de $\sum_{1\leq \ell \leq m} (S_\ell - \overline{S})^2$.

Comme nous l'avons annoncé dans l'introduction ces tests ne peuvent acquérir un intérêt pratique qu'après une longue expérience ; en effet, diverses structures de lien, autres que celles définies par une sériation ou une disposition sphérique, pour la métrique considérée, peuvent se manifester par une élévation (resp. baisse) sensible de $\max_{\ell} \sum_{\{k/k\neq\ell\}} (S_{\ell k} - \overline{S}_\ell)^2$ (resp. de $\sum_{\ell} (\overline{S}_\ell - S)^2$). Ces diverses structures de lien sont en général compatibles avec une structure en classes ; d'où l'intérêt du test de classificabilité (cf. chap. 3 § IV).

4. DETERMINATION D'UN PLAN DE REPRESENTATION GEOMETRIQUE.

Algorithme.

L'indice ℓ_1 pour lequel est maximum la variance $\frac{1}{(n-1)} \sum_{\{k/k\neq\ell\}} (S_{\ell k} - \overline{S}_\ell)^2$ définira le premier axe A_{x1} du plan de représentation ; A_{x1} sera pris horizontal et orienté de gauche à droite.

L'élément ℓ_2 qui définira le second axe A_{x2} doit satisfaire deux conditions :

a) Une valeur de $S(\ell_1,\ell_2)$ voisine de 0 ; ℓ_2 devant être assez indépendant de ℓ_1. $S(\ell_1,\ell_2)$ est le produit scalaire de deux vecteurs lignes ℓ_1 et ℓ_2 de la matrice $(\varepsilon'_{\ell j})$ transformée de celle d'incidence $(\varepsilon_{\ell j})$.

b) Une valeur élevée de $V(\ell_2)$ qui nous assurera du caractère discriminant du second axe.

Par conséquent, nous prendrons pour l'indice ℓ_2, qui détermine A_{x2}, celui qui rend maximum le rapport

$$(V(k)/S(\ell_1,k))^2 .$$

Le point d'intersection des deux axes aura comme abscisse commune $S(\ell_1,\ell_2)$ sur chacun des deux axes.

Si t est la valeur de $S(\ell_1,\ell_2)$, l'angle α des deux axes peut être défini par $\pi(t)\pi$

$$\text{où} \quad \pi(t) = \Pr^N\{S(\ell,k) < t\} = \frac{1}{2\pi} \int_{-\infty}^{t} e^{-x^2/2} dx .$$

Ainsi, l'angle des deux axes définirait le degré d'indépendance des deux attributs de description qu'ils représentent respectivement. Cette condition n'a pas été rigoureusement respectée dans le programme ; afin de mieux visualiser la représentation, l'angle entre les deux axes a été pris assez grand même lorsque les deux classes, entraînées respectivement le long de chacun des deux axes, ont un lien non tout à fait négligeable.

Sur le plan des deux axes, un objet k sera représenté par le point de coordonnée $(S(\ell_1,k),S(\ell_2,k))$.

On peut de proche en proche déterminer de nouveaux axes de référence définissant, lorsqu'elles existent, de nouvelles dimensions ; ainsi le troisième axe A_{x3} sera défini à partir de l'indice ℓ_3 qui satisfait :

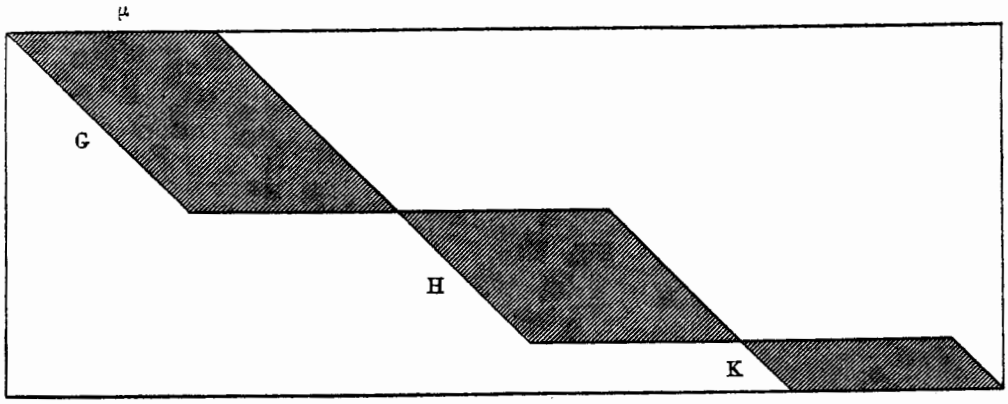
$$\max\{\min(\lambda\frac{V(k)}{S(\ell_1,k)}\}^2 \ , \ \{\frac{V(k)}{S(\ell_2,k)}\}^2)\} \ ;$$

l'élément ℓ_3 doit en effet, d'une part avoir la plus faible liaison possible avec chacun des pôles ℓ_1 et ℓ_2 déjà extraits et, d'autre part, être le plus discriminant possible ; c'est-à-dire, avoir une forte valeur de la variance des proximités ; d'où le critère qui a été adopté. Mais, on peut également proposer de maximiser $\{(V(k))^2/|S(\ell_1,k)S(\ell_2,k)|\}$ pour la détermination de ℓ_3 ; une étude de comparative des deux critères, au niveau expérimental, s'imposera dans ces conditions.

Examinons dans le plan des deux premiers axes A_{x1} et A_{x2}, les exemples 1) et 2) du paragraphe précédent. Pour l'exemple 1), les points représentant les lignes du bloc H (resp. G) se trouvent tous situés dans l'ordre défini par la forme σ du tableau d'incidence, sur A_{x1} (resp. A_{x2}).

Pour l'exemple 2), les points représentant les lignes du bloc G (resp. H) se trouvent rangés selon A_{x1} (resp. A_{x2}) dans l'ordre défini par la forme σ. La plupart des points relatifs au bloc G (resp. H) se trouvent sur A_{x1} (resp. A_{x2}), ceux qui n'y sont pas sont au voisinage de l'origine.

Considérons un troisième exemple où le tableau d'incidence peut prendre la forme ci-dessous.



Les points relatifs aux lignes de G (resp. H) sont situés, dans le plan des deux premiers axes, sur A_{x1} (resp. A_{x2}) selon l'ordre défini par la forme σ. Les diverses lignes du bloc K sont représentés en un même point d'égales coordonnées $(-\sqrt{p\mu},-\sqrt{p\mu})$. Un troisième axe A_{x3} définira la dimension sous-jacente à K.

Avant d'exprimer les résultats de l'analyse expérimentale dans le cas où la forme σ, formée d'un ou plusieurs blocs, n'est que floue et domine statistiquement ; nous allons chercher à situer rapidement cette approche par rapport à l'analyse factorielle la plus voisine qui est l'analyse en composantes principales (cf. chap. 6).

VI - *COMPARAISON AVEC L'ANALYSE FACTORIELLE.*

Discutons la recherche du premier axe puisque les autres s'en déduisent de proche en proche.

En ce qui nous concerne, le premier axe de discrimination A_{x1} est défini à partir d'un vecteur ligne du tableau d'incidence transformé. Si $\vec{n}_k$ désigne un vecteur ligne courant

$$\vec{n}_k = (((\varepsilon_{kj}-\mu_k)/\sqrt{\mu_k}\sqrt{p})/j = 1,2,\ldots,p)$$

l'indice ℓ_1 qui définit le premier axe est, rappelons le, celui qui rend maximum par rapport à ℓ.

$$\sum_{k=1}^{n} \{\vec{n}_\ell(\vec{n}_k-\vec{n}_\ell(\frac{1}{n}\sum_{k=1}^{n}\vec{n}_k))\}^2 .$$

Si les lignes représentent par exemple des attributs de description, c'est un attribut effectivement présent qui sera le point extrême droite du premier axe qu'il caractérise. La suite des projections, au sens de notre métrique, des autres attributs sur cet axe définira le premier facteur.

Pour l'analyse factorielle le plus proche de la méthode, le premier axe est défini par un vecteur unitaire $\vec{v}$, celui qui rend maximum

$$\sum_{k=1}^{n} (\vec{n}_k.\vec{v})^2$$

$\vec{n}_k.\vec{v}$ étant le produit scalaire euclidien.

Par conséquent, dans notre méthode il s'agit, intuitivement parlant, d'un jugement des données de l'intérieur. La technique ne nécessite pas la diagonalisation d'une matrice.

Signalons qu'un des points de départ de ce travail a été une remarque pratique : cherchant à découvrir une classification sur des données pour lesquelles une analyse factorielle des correspondances avait été appliquée, nous avons commencé par déterminer au moyen d'une technique analysée au chapitre 3, et dont ce texte présente une généralisation systématique, les éléments les plus neutres et ceux, les plus discriminants. Nous avons observé, dans le plan des deux premiers axes factoriels, les éléments les plus neutres se grouper autour de l'origine et ceux les plus discriminants aux extrémités du premier axe dont l'importance était d'ailleurs sensiblement supérieure au second.

En fait, c'est au niveau d'une classe bien cohérente, résultant d'une classification automatique, que nous envisageons d'appliquer notre méthode pour étudier la position relative des divers éléments de la classe.

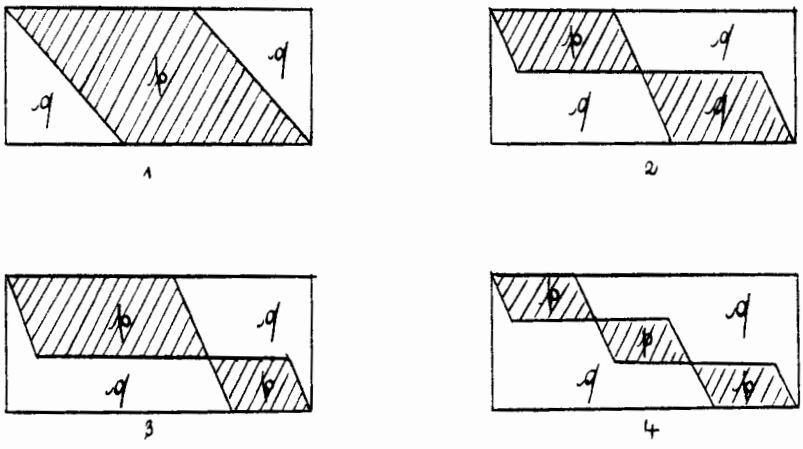
Ce traitement s'applique à n'importe quel tableau de données A×E pourvu que les variables descriptives de A définissent toutes le même type de structure algébrique sur E. L'analyse n'est en effet basée que sur les proximités et nous avons étendu, en le précisant, le principe de défini-

tion de la mesure de proximité entre deux variables indicatrices de parties sur E (cf. § II), au cas d'un couple de variables définissant, soit un couple de partitions, soit un couple de préordres totaux, soit un couple d'ordres totaux, soit enfin, un couple de mesures positives sur E (cf. chap. 2).

Il arrive souvent dans des études pour le développement économique et social que A soit formé d'échelles (chaque variable détermine sur E un préordre total). Si une classification automatique sur A permet de dégager les principales "dimensions" du développement, on peut espérer que l'analyse d'une même classe de variables par cette méthode, suivie d'une recherche plus précise, étudiée au chapitre , établira pour la "dimension" étudiée, un enchaînement entre les différents états du développement.

VII - *ANALYSE DES FORMES σ FLOUES.*

Nous allons exprimer ici, de façon shématique, les résultats de la mise en oeuvre et de l'expérimentation de l'algorithme précédent telle qu'elle a été conduite par H. Leredde (cf. [5]) sur des formes σ "floues", obtenues par simulation, comme réalisation d'un modèle aléatoire que nous allons préciser. A une forme σ, formée d'un ou plusieurs blocs, associons un tableau d'incidence aléatoire $(\varepsilon'_{\ell j} / 1 \leqslant \ell \leqslant m, 1 \leqslant j \leqslant n)$ où les différents $\varepsilon'_{\ell j}$ sont des variables aléatoires indicatrices ($\varepsilon'_{\ell j} = 0$ ou 1) indépendantes. La probabilité d'une valeur 1 dépend de la position de (ℓ, j) par rapport à la partie chargée du tableau de référence où la forme σ est mise en évidence. Cette probabilité sera ici p ou q selon que (ℓ, j) tombe à l'intérieur d'un bloc chargé ou non. Le tableau à m lignes et n colonnes est ainsi un élément aléatoire de $\{0,1\}^{mn}$ muni d'une mesure de probabilité qui résulte du produit de mn aléas indépendants construits sur $\{0,1\}$. On suppose évidemment $p > q$; la grandeur de p et la petitesse de q indiqueront la netteté statistique avec laquelle on peut espérer voir apparaître la forme σ, après simulation. L'hypothèse définissant le talbeau d'incidence aléatoire peut être considérée comme étant celle de la liaison la plus faible compatible avec la configuration σ fixée du tableau. A ce sujet, les configurations étudiées expérimentalement sont de l'un des types suivants,



formées d'un ou plusieurs blocs parallélogrammiques de même importance ou d'importance inégale. Les dimensions du tableau considérées sont 50 lignes × 500 colonnes ; pour une configuration σ déterminée de ce tableau, on commence par fixer les deux nombres p et q, probabilités d'apparition d'un 1 à l'intérieur et à l'extérieur des blocs ; par exemple p = 0,8 et q = 0,3 qui définissent des densités de chargement. Pratiquement, on considère une fonction engendrant des nombres aléatoires compris entre 0 et 1 ; à une case (ℓ,j) du tableau d'incidence, on associe une réalisation $x_{\ell j}$ d'une telle fonction ; si (ℓ,j) est à l'intérieur d'un bloc et si $x_{\ell j} \leqslant p$ (resp. $x_{\ell j} > p$), on pose $\varepsilon_{\ell j} = 1$ (resp. 0) ; si par contre (ℓ,j) est à l'extérieur de la réunion des blocs et si $x_{\ell j} \leqslant q$ (resp. $x_{\ell j} > q$), on pose $\varepsilon_{\ell j} = 1$ (resp. 0).

Pour l'analyse expérimentale, on a fait varier, simultanément ou indépendamment, les paramètres suivants :

a) nombre de blocs ;

b) caractère "flou" du chargement (modification de p et de q) ;

c) épaisseur des blocs : nombre de composantes où la probabilité d'apparition d'un 1 est p ;

d) importance relative de chaque bloc (par exemple, un gros et un petit bloc comme dans la troisième configuration ci-dessus).

Nous allons à présent indiquer, pour chacun des cas, les résultats observés les plus significatifs.

a') Dans l'exemple d'une forme σ à un seul bloc, on retrouve, au niveau de la représentation géométrique, la forme en "fer à cheval" mise en évidence dans l'algorithme de D. G. Kendall. Pour des formes σ à deux ou trois blocs, on retrouve clairement dans la représentation les différents blocs ; chaque pôle sort d'un bloc différent qu'il entraîne. Ce résultat ne tient que dans la mesure où le nombre de pôles qu'on se fixe de déterminer n'excède pas le nombre de blocs.

b') Lorsque les densités p et q de chargement, qui expriment le caractère "flou", varient, la représentation apparait plus ou moins discriminante. En particulier, q restant constant, lorsque p augmente (forme σ de plus en plus prononcée), la représentation s'en ressent par l'allongement du nuage associé à un bloc le long de l'axe qu'il entraîne ; la dispersion des points représentatifs devient plus grande.

c') Lorsque les différents blocs s'amincissent ; alors cet amincissement se réflète au niveau de la représentation et ceci d'autant plus nettement que p (resp. q) est grand (resp. petit).

d') En faisant varier l'importance relative des blocs ; on remarque que le premier pôle extrait provient presque toujours du plus petit agrégat. La dissymétrie entre blocs peut certes provenir de la forme σ ; mais, elle peut également être dûe à la différence entre les densités de chargement des différents blocs.

VIII - *NOUVEAUX ALGORITHMES DE CLASSIFICATION.*

1. INTRODUCTION

Nous allons présenter ici une nouvelle classe d'algorithmes de classification dont l'analyse est au coeur des travaux qu'a conduit H. Leredde (Université Paris-Nord) dans le cadre d'une thèse (cf. [5]). Nous présenterons dans la partie II , l'organisation shématique des programmes.

L'importance de ces algorithmes aurait certainement mérité un développement sur tout un chapitre ; mais l'historique de leur apparition leur confine pour le moment ce paragraphe. Pour une étude plus détaillée on se refèrera à la thèse précitée.

La première application de l'algorithme de représentation euclidienne autour des deux premiers pôles a porté sur une restriction de la matrice d'incidence de description des personnages enfants de la littérature enfantine, à un sous ensemble de quelques soixante dix attributs parmi les plus représentatifs (cf. partie II, chap.). La disposition relative des points attributs sur le plan des deux premiers axes portant les deux premiers pôles était d'une part conforme à la typologie définie par notre principale méthode de classification hiérarchique basée sur la vraisemblance des liens (A.V.L.) et d'autre part, tout à fait voisine de celle définie par la représentation sur le premier plan factoriel issu d'une analyse des correspondances. Ce résultat, compte tenu du caractère élémentaire de la technique où il n'est pas fait appel à la diagonalisation d'une matrice, est d'un intérêt certain. Toutefois, l'aspect local de la représentation nous interdit de procéder à une telle organisation géométrique d'un vaste ensemble, recouvrant de multiples "dimensions", par rapport à quelques uns seulement de ses éléments, quel que soit le rôle privilégié que jouent ces derniers. Nous ne mentionnerons donc l'intérêt d'une telle représentation qu'au niveau d'une ou de deux "bonnes" classes, pouvant d'ailleurs être sous-tendues par des noeuds significatifs de l'arbre des classifications (cf. chap. 5, § V).

Nous avons pu nous rendre compte, dans l'étude expérimentale, que les différents pôles extraits appartenaient respectivement à des blocs différents, dans la mesure où le nombre de pôles déterminés est inférieur ou égal au nombre de blocs que comporte la forme simulée σ ; et ce, même si σ est très "floue". Chacun des pôles sorti pouvait ne pas se situer à l'extrémité du bloc auquel il appartient, en raison sans doute du caractère imprécis de la forme du bloc, lequel étant néanmoins entraîné par son pôle. Il en résulte la possibilité de la découverte d'une classification à partir de la détermination a priori de pôles d'attraction et de la répartition des différents objets de l'ensemble à classifier entre les différents pôles, en attribuant chaque élément au pôle le plus proche, au sens de la proximité établie. D'ores et déjà on voit que ce type d'algorithme, qu'il s'agit d'ailleurs de préciser et peut-être d'étendre à la recherche d'une hiérarchie de classifications, peut être d'une part, rapide et efficace dans le traitement des gros tableaux de données et pose d'autre part, des problèmes méthodologiques dans ses rapports avec les autres algorithmes de classification.

2. CRITERES DANS LA DETERMINATION DES "POLES" ET LA FORMATION DES CLASSES.

L'analyse statistique et informatique des données réelles a conduit à préciser trois types d'algorithmes distincts que nous nous contenterons de présenter shématiquement ici ; renvoyant à la thèse de Leredde pour discuter de leurs mérites respectifs.

- Le premier algorithme, le plus directement lié à ce qui précède par le critère qu'il utilise, détermine une suite de classifications ; la c-ème comporte c classes et s'obtient à partir de la détermination de c "pôles d'attraction" par affectation de chaque élément de l'ensemble à classifier à exactement l'un des pôles : celui, le plus proche, au sens de l'indice de proximité établi. La détermination de la suite des pôles est conforme à celle adoptée dans la représentation géométrique. Par conséquent, la première est triviale, en une classe ; la seconde, en deux classes, s'obtient autour des deux premiers pôles ℓ_1 et ℓ_2 ; rappelons que si le premier ℓ_1 est déterminé de façon à maximiser la variance des proximités à k, V(k), le second ℓ_2 maximise le critère $\{V(k)/S(\ell_1,k)\}^2$ sur $\{k/k\neq\ell_1\}$. La troisième classification réorganise l'ensemble, par proximité, autour des trois premiers pôles, dont le troisième est solution de

$$\max \{\min(\{\frac{V(k)}{S(\ell_1,k)}\}^2 , \{\frac{V(k)}{S(\ell_2,k)}\}^2)\} \qquad \text{(cf. § V.4)}$$

sur $\{k/k\neq\ell_1$ et $k\neq\ell_2\}$; et ainsi de suite, la (c+1)ème classification en (c+1) classes s'obtient après la détermination du (c+1)ème pôle, lequel satisfaisant le critère

$$\max \{\min(\{\frac{V(k)}{S(\ell_1,k)}\}^2 , \{\frac{V(k)}{S(\ell_2,k)}\}^2 ,\ldots, \{\frac{V(k)}{S(\ell_c,k)}\}^2) \} \qquad (1)$$

sur $\{k/k\neq\ell_j$ pour $1 \leqslant j \leqslant c\}$.

La (c+1)-ème classification s'obtient à partir de la c-ème par la création autour d'un nouveau pôle d'une nouvelle classe qui emprunte ses éléments aux diverses classes déjà établies ; ce "raffinement" de la classification est d'une nature tout à fait différente de celui relatif à un arbre de classification, qu'on regarde dans le sens racine $\longrightarrow$ terminaux, où le raffinement est défini par l'éclatement d'une classe en deux.

Une autre option peut être proposée ou le (c+1)ème pôle est choisi de façon, non pas à être le plus "neutre" ou 'indépendant" ; mais, le plus "éloigné" des c précédents pôles. Ceci peut se faire avec la même règle (1) que ci dessus, au moyen d'un changement monotone de l'échelle des similarités, ramenant à 0 la valeur de l'indice dans le cas d'un "éloignement" maximal. Plus précisément, S_o désignant la valeur maximale de la similarité sur l'ensemble F des paires, on travaillera avec le tableau d'indices

$$\{S^*(x,y) = S(x,y)-S_o+\varepsilon_s/\{x,y\} \in F\} \qquad (2)$$

où ε_s est un nombre positif très petit dont le seul rôle est de ne pas

provoquer de division par 0.

L'ensemble des algorithmes de représentation euclidienne et de classification où on travaille avec les similarités, ont été regroupés dans un même programme sous le titre "Méthode des Pôles d'Attraction utilisant les Similarités" (MPATS).

C'est un critère de type minimax qui permet de déterminer, à partir du deuxième pôle, la suite des autres pôles ; une telle règle permettant d'éviter le calcul linéaire. Toutefois, si on désire déterminer la suite des autres pôles avec un critère exactement de même nature que celui qui a prévalu à l'extraction des deux premiers pôles, on commencera par remarquer que, pour un codage réduit tel que celui défini au paragraphe II précédent, $S(\ell_1,k)$ n'est autre que la mesure de projection de $(k-0)$ sur (ℓ_1-0) où 0 désigne l'origine de l'espace de représentation. On procèdera ensuite comme suit :

K désignant ici l'ensemble à représenter ; le troisième pôle ℓ_3 est défini par le point j de $(K-\{\ell_1,\ell_2\})$ pour lequel

$$\{V(j)/\text{proj.}(j-0) \text{ sur le plan engendré par } (\ell_1-0,\ell_2-0)\}^2 \qquad (3)$$

est maximum.

Si $\hat{\ell}_1$ (resp. $\hat{\ell}_2$) est le vecteur colonne à n composantes dont la suite des coordonnées est celle de (ℓ_1-0) (resp. (ℓ_2-0)) et si $\hat{\ell} = (\hat{\ell}_1,\hat{\ell}_2)$ désigne la matrice n×2 résultante ; le projecteur P_{12} sur le plan engendré par (ℓ_1-0,ℓ_2-0) est défini par

$$P_{12} = \hat{\ell}(\hat{\ell}'\hat{\ell})^{-1}\hat{\ell} \qquad (4)$$

où $\hat{\ell}'$ désigne la transposée de $\hat{\ell}$. La détermination de P_{12} suppose donc l'inversion d'une matrice 2×2 qui est généralement de rang 2 car il est exceptionnel que (ℓ_2-0) ne soit pas distinct de (ℓ_1-0).

Plus généralement, si $\{\ell_j/1\leqslant j\leqslant c\}$ est l'ensemble des c premiers pôles extraits ; le (c+1)ème pôle sera défini comme le point j de $(K-\{\ell_j/1\leqslant j\leqslant c\})$ pour lequel est maximum

$$\{V(j)/\text{proj.}(j-0) \text{ sur le sous espace engendré par } \{(\ell_j-0)/1\leqslant j\leqslant c\}\}^2 \qquad (5)$$

Le projecteur définissant le dénominateur de l'expression précédente est donné par

$$P_{12\ldots c} = \hat{\ell}(\hat{\ell}'\hat{\ell})^{-1}\hat{\ell} \qquad (6)$$

où cette fois-ci

$$\hat{\ell} = (\hat{\ell}_1,\hat{\ell}_2,\ldots,\hat{\ell}_c) \qquad (7)$$

est une matrice (n,c) qu'on peut espérer de rang c car les pôles sont choisis d'une certaine façon "aussi indépendants que possible" les uns des autres. Dans ces conditions, la détermination de $P_{12\ldots c}$ suppose l'inversion d'une matrice c×c ; donc relativement "petite", et il y a d'excellents programmes à ce sujet.

Cette dernière direction des travaux doit prochainement être testée par rapport aux résultats déjà tout à fait intéressants obtenus par MPATS et dont des exemples seront fournis dans la partie II.

- Un deuxième type d'algorithme consiste à déterminer une classification, classe après classe ; une même étape de l'algorithme est définie par la constitution d'une classe qu'on entraîne autour d'un pôle d'attraction par l'affectation à ce dernier de tous les éléments dont la distance est inférieure à un certain seuil δ. δ qui définit le rayon d'agrégation doit être fixé à partir de considérations statistiques. Une classe étant formée ; le nouveau couple (pôle d'attraction, seuil d'agrégation) s'obtient sur l'ensemble restant (non encore classé) au moyen des mêmes critères qui ont servi à la détermination de ce même premier couple sur l'ensemble plein à classifier.

On peut certes, pour la détermination d'un même pôle, utiliser la même quantité critère que ci-dessus (i.e. variance V des proximités). Ce critère, surtout intéressant pour une classification des variables, a davantage un caractère projectif que classificatoire (cf. chap. 3). Nous inspirant des critères utilisés en classification "des moindres carrés" ; nous avons été conduits, avec d'ailleurs de meilleurs résultats, à travailler avec la distance associée à la métrique que suppose notre indice de proximité et à adopter comme quantité critère le moment absolu d'ordre 2. $M_2(x)$ des distances à l'objet x. De la sorte, on traite avec autant de souplesse, aussi bien l'ensemble des variables que celui des objets. De façon précise, si D est l'ensemble à classifier et si C est la partie déjà organisée en classes par l'algorithme ; le nouveau pôle d'attraction d'une nouvelle classe est l'élément p qui réalise

$$\max_{x\in(D-C)} M_2(x) \quad ; \text{ où } M_2(x) \text{ est le moment d'ordre 2 de } (D-C) \text{ par rapport à x.} \qquad (8)$$

A partir de p, on détermine la classe en agrégeant au pôle tous les sommets dont la distance au pôle est inférieure à un certain seuil qu'il s'agit de déterminer de la façon la moins arbitraire possible. Pour cela, nous considérons la distribution des point du nuage rectiligne $\{d(p,x)/x\in(D-C)\}$; soit la suite croissante

$$d(p,x_{(1)}) \leqslant d(p,x_{(2)}) \leqslant \ldots \leqslant d(p,x_{(m)}) \ , \qquad (9)$$

où nous avons noté m le cardinal de (D-C) et où $x_{(j)}$ est le j-ème point de (D-C) par éloignement relatif à p. Le préordre (9) précédent est en général très fin s'il n'est pas un ordre total strict. Il s'agit, intuitivement parlant, d'arrêter la formation de la classe à $x_{(j)}$ si le point d'abscisse $d(p,x_{(j)})$ peut apparaître comme centre d'un "petit" intervalle où la densité des points du nuage à droite est "sensiblement" plus faible que celle, à gauche. La technique utilisée, qui fait référence à une loi d'adéquation gaussienne, est la suivante :

A la suite (9) précédente des distances à p, on associe la suite des nombres

$$\phi(\Delta_{(\ell)}) = \phi[(d(p,x_{(\ell)}) - \mu_\ell)/\sigma_\ell] \ , \qquad (10)$$

où μ_ℓ et σ_ℓ sont, respectivement, la moyenne et l'écart-type de la suite des nombres $\{d(p,x_{(j)})/1\leqslant j\leqslant \ell\}$ et où ϕ est la fonction de répartition de la loi normale centrée réduite N(0,1). On décide de l'arrêt de la formation de la classe à $\{p = x_{(1)},x_{(2)},\ldots,x_{(j)}\}$ si un saut brutal se produit entre $\phi(\Delta_{(j)})$ et $\phi(\Delta_{(j+1)})$.

Bien que des difficultés peuvent apparaître dans l'algorithme d'appréciation d'une rupture entre $\phi(\Delta_{(j)})$ et $\phi(\Delta_{(j+1)})$; cette méthode de classification, remarquable par sa simplicité, a donné dans certaines situations des résultats très intéressants. D'autre part, on verra ci-dessous qu'on peut espérer une généralisation à la recherche d'une hiérarchie des classifications.

L'arrêt dans cet algorithme de la formation des classes, est liée à la valeur d'un critère de classification (cf. paragraphe suivant).

Ce deuxième algorithme est l'argument d'un programme intitulé : "Méthode des Pôles d'Agrégation utilisant les Distances" (MPAGD).

- Le troisième algorithme de classification nous apparait comme le plus consistant. Qu'il s'agisse de la classifcation des variables descriptives (attributs ou variables numériques) ou bien des objets ; on a ici un traitement unique, moyennant le remplacement des mesures "brutes" par celles, "centrées, réduites" dans le tableau de données, comme nous avons eu à le faire pour interpréter certaines formules en Analyse en Composantes Principales (cf. chap. 6). La métrique sous-jacente correspond à adopter l'indice de K. Pearson lorsqu'il s'agit de la comparaison d'un couple d'attributs descriptifs. La quantité critère utilisée pour la détermination des pôles d'attraction ainsi d'ailleurs que pour la répartition autour de ces derniers des différents éléments de l'ensemble D à classifier, est basée sur le moment d'inertie. Ainsi le premier pôle est l'élément qui réalise

$$\max_{x \in D} M_2(x) , \qquad (11)$$

où $M_2(x)$ est le moment absolu d'ordre 2 par rapport à x ; soit

$$M_2(x) = \frac{1}{\mathrm{card}(D)} \sum_{y \in D} d^2(y,x) , \qquad (12)$$

où $d^2(y,x)$ est le carré de la distance (au sens euclidien ordinaire) entre les deux vecteurs de $\mathbb{R}^k$ représentant x et y dans le tableau des données réduites.

Soit $\mathfrak{P}'$ l'ensemble des pôles déjà extraits. Si, dans le premier algorithme, le nouveau pôle à déterminer est choisi, pour un égal de discrimination, de la façon la plus "neutre" par rapport aux différents éléments de $\mathfrak{P}'$; on choisira ici le nouveau pôle de la façon la plus "éloignée" des différents points de $\mathfrak{P}'$. Le critère de détermination est de façon précise

$$\max_{x \in (D - \mathfrak{P}')} \{ \min_{p \in \mathfrak{P}'} M_2(x) D^2(x,p) \} , \qquad (13)$$

Si $\mathfrak{P}$ est l'ensemble des pôles retenus et si $C\ell^*(q)$ désigne la classe déjà formée autour du pôle q appartenant à $\mathfrak{P}$; on affectera l'objet x de

(D-C), où $C = U\{C\ell^*(q)/q \in \mathfrak{P}\}$, au pôle p de $\mathfrak{P}$, si le couple (x,p) réalise

$$\min_{(y,q) \in (D-C)\times \mathfrak{P}} \{\frac{1}{card(C\ell^*(q))} \sum_{x \in C\ell^*(q)} d^2(x,y)\} ; \quad (14)$$

en d'autres termes, l'objet x est affecté à la classe $C\ell^*(p)$ si le moment d'inertie de $C\ell^*(p)$ par rapport à x est le plus petit.

La classification étant achevée ; on détermine pour chaque classe le meilleur représentant ; lequel est défini comme étant l'élément de la classe qui réalise le minimum de la somme des carrés des distances aux différents éléments de cette dernière.

On offre dans cet algorithme l'option de reformer les classes autour des meilleurs représentants en réaffectant l'ensemble conformément à l'algorithme de la distance minimum d'un point aux différents sommets d'un sous ensemble disjoint ; cette dernière classification n'est retenue que si notre critère d'adéquation, que nous présenterons bientôt, la juge meilleure.

Le programme est ici intitulé "Méthode des Pôles d'Attraction utilisant les Distances" (MPATD).

Malgré la qualité des résultats obtenus (cf. partie II), comme pour MPATS, le critère de détection des pôles semble devoir être mis en concurrence avec un critère de nature géométrique plus explicite et que nous exprimerons dans le cadre général d'un nuage pondéré N(I) de points dans un espace euclidien. Pour fixer les idées, I représente l'ensemble précédemment noté D.

Si la règle de détermination du premier pôle q_1 est celle (11) définie ci-dessus ; celle d'extraction du second pôle q_2 consiste à rendre maximum le moment produit du nuage N(I) par rapport à q_1 et à i décrivant $(I-\{q_1\})$. C'est-à-dire q_2 sera défini tel que soit maximum

$$M(q_1,i) = \sum_{i' \in I} \mu_{i'} d(i',q_1)d(i',i) \quad (15)$$

où $\mu_{i'}$ est la masse affectée à i' de I.

De façon plus générale, si $\{q_j/1 \leqslant j \leqslant c\}$ est l'ensemble des c premiers pôles extraits ; le (c+1)ème sera défini de façon à rendre maximal le moment produit du nuage ; en d'autres termes q_{c+1} sera défini de façon à rendre maximum

$$M(q_1,q_2,\ldots,q_c,i) = \sum_{i' \in I} \mu_{i'} d(i',q_1)\ldots d(i',q_c)d(i',i). \quad (16)$$

Cette nouvelle procédure doit prochainement être testée par rapport à la précédente (cf. (13)) qui a fait ses preuves.

Terminons ce paragraphe en précisant que cette famille de méthodes de classification s'applique directement à trois types de tableaux de données : les tableaux d'incidence de zéros et de uns, les tableaux de mesures numériques et les tableaux de contingence.

3. SUR LE NOMBRE DE POLES A EXTRAIRE ET LA QUALITE DES DIFFERENTES CLASSIFICATIONS.

Les algorithmes précédents demeurent imprécis sur deux points ; d'ailleurs liés : comment juger de chacune des partitions produites et jusqu'à combien de pôles extraire ? Relativement aux expériences sur les formes σ simulées ; il s'agit, on l'a vu, que le nombre de pôles ne dépasse pas le nombre de blocs que comporte la forme σ, afin qu'une même classe ne puisse se subdiviser artificiellement au cours de l'algorithme de répartition.

Il est par conséquent naturel d'attacher la valeur d'un critère qui permettrait de mesurer l'adéquation de chacune des classifications produite à chaque étape de l'algorithme. En tenant compte de l'ordre de grandeur du nombre de classes qu'on désire pour la dernière partition produite, le critère permet de définir un test d'arrêt par l'examen de sa distribution sur la suite des classifications ; on arrêtera l'algorithme dès qu'on aura atteint une classification qui correspond à un sensible maximum local du critère et dont le nombre de classes tombe dans l'intervalle que définit l'ordre de grandeur voulu.

Nous commencerons par proposer notre critère de classification, directement basé sur le tableau des Similarités (centrées et réduites (cf. formule (2'), § II A. 3.3.2, chap. 4) dont l'expression, rappelons le, est la suivante

$$\frac{1}{\sqrt{r.s/(f-1)}} \sum_{p \in F} \varepsilon(p)c(p) \ , \qquad (17)$$

où F est l'ensemble des paires ($f = \mathrm{card}(F)$), $\varepsilon(p)$ est la fonction indicatrice de la partition dont l'ensemble R (resp. S) des paires réunies (resp. séparées) a pour cardinal r (resp. s). $c(p) = (S(p)-\alpha)/\lambda$, où α et λ^2 sont respectivement la moyenne et la variance de distribution $\{S(p)/p \in F\}$ des Similarités.

Un autre critère plus classique, qu'on peut proposer de façon concurrentielle, est l'"inertie expliquée" par la classification. On supposera, pour présenter ici ce critère, que le tableau T des données a été transformé en celui T' des "mesures centrées réduites" (cf. formule (2) § II et § II.3, chap. 6), de telle sorte que l'indice de proximité corresponde au produit scalaire ordinaire entre vecteurs lignes (resp. colonnes) du tableau T' ; selon qu'on désire traiter l'ensemble des objets ou celui des variables. D désignant l'ensemble à classifier ; l'expression de ce critère pour une partition $\{D_j/1 \leqslant j \leqslant k\}$ de D, est la suivante

$$I(\pi) = \sum_{1 \leqslant j \leqslant k} d_j(g_j-g)^2 \Big/ \sum_{x \in D} (x-g)^2, \qquad (18)$$

rapport de l'inertie expliquée par la classification sur celle, totale. Dans cette formule, d_j est le cardinal de la classe D_j ; g_j (resp. g) est le centre de gravité de D_j (resp. D) ; soit

$$g_j = \frac{1}{d_j} \sum_{x \in D_j} x \quad \text{et} \quad g = \frac{1}{d} \sum_{x \in D} x \ , \qquad (19)$$

où on a posé $d = \mathrm{card}(D)$.

Le développement des précédents algorithmes rend nécessaire une expression du critère $I(\pi)$ qui utilise le plus directement possible, le tableau S carré, indexé par D×D, des Similarités S. Dans ces conditions, H. Leredde a établi la formule suivante

$$I(\pi) = \frac{d \times \sum_{1 \leq j \leq k} \{\frac{1}{d_j} \text{Som.}(S(D_j)) - 2 \sum_{x \in D_j} \overline{S}_x\} + \text{Som.}(S)}{d \times \text{tr.}(S) - \text{Som.}(S)} , \qquad (20)$$

où nous avons noté S le tableau des similarités, et $S(D_j)$ sa restriction à $D_j \times D_j$. Som. désigne l'opérateur somme de toutes les valeurs du tableau et celui tr. (abréviation de trace) concerne la sommation des éléments diagonaux du tableau.

La mise en oeuvre de la règle d'arrêt de l'algorithme, ci-dessus exprimée, peut être difficile dans certaines situations où, autour de la classification au nombre de classes voulu, le taux d'accroissement du critère peut apparaître quasiment constant. Toutefois, le comportement de chacun des deux critères (17) et (18) sur la suite des partitions est décisif pour une interprétation plus sûre et le choix que fera le spécialiste de quelques unes des classifications dans la suite.

4 - APPORT DANS L'ALGORITHME DES "NUEES DYNAMIQUES".

La détermination a priori d'une suite de "pôles d'attraction" permet de contribuer efficacement à divers types d'algorithmes en enlevant à ces derniers une part importante de leur degré d'arbitraire. C'est ainsi que dans la méthode dite des "nuées dynamiques" (cf. chap. 1), dans la mesure, non aussi restrictive qu'il semble a priori, ou on admet la représentation d'une classe par un de ses éléments, on peut prendre comme système initial $L^{(o)}$ de noyaux, une suite de pôles d'attractions. Dans ces conditions, une question se pose relativement à la qualité des résultats et à la rapidité de l'algorithme lorsqu'on remplace le choix au "hasard" par celui qu'on vient de définir du système initial de noyaux. Relativement au deuxième point, il s'est avéré au cours des nombreuses expériences menées dans le cadre du troisième algorithme ci-dessus considéré (cf. § 2), qu'il y avait, pour ainsi dire, toujours stabilité, directement autour des meilleurs représentants des classes obtenues à partir du système de pôles ; en d'autres termes, les classes et leurs représentants obtenus par réallocation autour de ces premiers meilleurs représentants, restaient invariables. Cet algorithme nous a par ailleurs fourni d'excellents résultats.

Le choix du système initial de noyaux que nous proposons suppose, bien entendu, le calcul préalable du tableau des Similarités ou Distances ; mais, comme nous l'avons déjà fait remarquer (cf. § III.3, chap.1), ce calcul peut être avantageux pour certaines configurations du tableau des données où, notamment, l'ensemble à classifier indéxe le côté du tableau le plus petit ; et ce, même si le système initial de noyaux nous est donné a priori, sans aucun calcul statistique.

5. EXTENSION AU PROBLEME DE LA RECHERCHE D'UNE HIERARCHIE DES CLASSIFICATIONS.

L'algorithme susceptible d'une généralisation à la recherche d'une chaîne, ordonnée par finesse, de classifications, est celui de deuxième type, présenté au paragraphe 2 ci-dessus. Le shéma du nouvel algorithme peut être le suivant :

On commence par déterminer le point p_1 de D par rapport auquel le moment d'inertie total $M_2(p_1)$ est maximum et on précise autour de p_1 une suite de couronnes circulaires $(C_1,K_2,\ldots,K_\ell)$, dont la première est un cercle, qui correspond à la suite $(\delta_1,\delta_2,\ldots,\delta_\ell)$ des seuils de distance qui doivent être définis par l'analyse du nuage rectiligne dont la suite des abscisses sur l'axe supportant le nuage est $\{d(p_1,x_{(j)})/1\leqslant j\leqslant m\}$ où d est la distance adoptée sur D et où $x_{(j)}$ est le j-ème point de D par éloignement relatif à p_1. K_j est défini par $\{x\in D/\delta_{j-1}\leqslant d(x,p_1)<\delta_j\}$. Un même δ_j peut correspondre à un sommet du nuage rectiligne, centre d'un "petit" intervalle où la densité à droite est "sensiblement" plus faible que la densité à gauche. Pour déterminer la suite des nombres δ_j ; on attachera à chacun des sommets du nuage rectiligne un intervalle de centre le sommet, dont l'amplitude dépend de sa position par rapport au nuage : il est clair en effet que la longueur de l'intervalle doit être d'autant plus petite qu'on se trouve dans une région à forte densité. Il reste à déterminer, à la lumière d'une analyse expérimentale, l'amplitude d'un tel intervalle ainsi qu'un seuil significatif pour la différence entre les deux densités de points dans les deux moitiés de l'intervalle dont l'abscisse du milieu est susceptible de définir le borne d'une couronne circulaire K_j.

L'étape suivante consiste à appliquer l'algorithme précédent dans chacune des couronnes circulaires $C_1,K_2,\ldots,K_\ell$; décomposant un même ensemble K_j en une suite de couronnes $(C_{j1},K_{j2},\ldots,K_{j\ell_j})$. On poursuit ainsi la décomposition de la suite de couronnes en sous couronnes jusqu'à ce qu'aucun affinage puisse être justifié statistiquement.

Une même couronne circulaire définira une classe, une couronne de cette dernière, une de ses sous-classes. La construction de l'arbre des classifications se fera, à partir de la partition grossière, de manière "descendante", par "segmentations" successives de chaque classe en sous-classes. A la partition grossière en une classe succède celle en ℓ classes dont la j-ème est définie par la j-ème couronne K_j ; à cette dernière succède celle, résultant de la décomposition de chaque classe K_j en la partition $\{C_{j1},K_{j2},\ldots,K_{j\ell_j}\}$; et, ainsi de suite...

Cet algorithme qui produit directement un arbre condensé des classifications ne fait qu'habiller statistiquement celui (cf. chap. O, § IV.2) qui permet de ramener une matrice de distances ultramétrique à la forme caractéristique précisée par le théorème de ce paragraphe. La définition statistique du centre et des rayons de la suite des couronnes permet à

l'algorithme de conduire une matrice de proximités, associée à un ensemble "classifiable" (cf. chap. 3, § III), à un profil le "plus voisin" de celui caractérisé par le théorème cité (chap. 0, § IV.2).

Certains aspects de cet algorithme restent à être précisés avant programmation ; toutefois, d'ores et déjà, on se rend compte de son grand intérêt informatique en raison de la petitesse de la place mémoire que nécessite le calcul de la décomposition d'une classe en sous-classe (i.e. couronne en sous-couronnes) et de la rapidité de ce calcul. Par conséquent cet algorithme est susceptible de traiter, dans des temps raisonnables et avec un emplacement mémoire limité, de très importants tableaux de données. Dans notre méthode principale, orientée d'abord vers la classification des variables (proximité sous la forme d'un indice de vraisemblance, algorithme A.V.L. de construction ascendante d'un arbre détaillé de classifications, condensation de l'arbre à ses noeuds les plus significatifs et interprétation dynamique de la suite des classifications à partir du comportement simultané de chacune des deux statistiques globale et locale des niveaux ; ainsi d'ailleurs, qu'à l'aide de la distribution, sur l'ensemble à classifier, de la statistique V (variance des proximités) et là, nous rejoignons ici...), on traite aisément quelques centaines de variables mesurées sur quelques milliers d'individus. A titre d'exemple, dans une récente expérience, à partir d'un tableau d'incidence A×E 250×4000, on a établit le tableau symétrique de contingence A×A, 250×250, des valeurs de s (s(a,b) est le nombre de sujets possédant chacun des deux attributs a et b) ; le traitement de ce dernier tableau par notre méthode a pris, sur IBM 370-165, près des trois quarts du temps d'une analyse factorielle par les algorithmes de diagonalisation les plus rapides. On peut espérer atteindre, sans trop de difficulté, une dimension de l'ordre de 10^6 pour le produit des deux côtés du tableau des données. Il reste certes à comparer la richesse et les formes de l'interprétation au niveau des données.

BIBLIOGRAPHIE

[1] J.-P. BENZECRI, *"Ordre latéral entre lois de probabilités sur un ensemble ordonné" in "l'Analyse des Données"* tome I, partie B, Dunod, Paris, 1973.

[2] J. BERTIN, *"La graphique et le traitement graphique de l'information"*, Flammarion, Paris, 1977.

[3] D.-G. KENDALL, *"Seriation from abundances matrices"*, Proc. Conference on mathematical methods in the archaeological and historical sciences, Mamaïa Roumanie, Sept. 1970.

[4] J.-B. KURSKAL, *"Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis"*, Psychometrika, vol. 29, 1964.

[5] H. LEREDDE, *"La méthode des pôles d'attraction ; la méthode des pôles d'agrégation : deux nouvelles familles d'algorithmes en classification automatique et sériation"*, Volume I : méthodes et exemples réels, Volume II : programmes. Thèse de 3ème cycle, Univ. Paris VI, 10 Oct. 1979.

[6] I. C. LERMAN, *"Sur l'analyse des données préalable à une classification automatique"*, Rev. Math. & Sc. Hum. n° 32, Paris, 1970.

[7] I. C. LERMAN, *"Analyse du phénomène de la "sériation"*, Rev. Math. & Sc. Hum. n° 38, Paris, 1972.

[8] I. C. LERMAN, H. LEREDDE, *"La méthode des pôles d'attraction"*, Actes du colloque *"Journées Analyse des Données et Informatique"*, I.R.I.A., Versailles, Sept. 77.

[9] R. N. SHEPARD, *"The analysis of proximities : multidimensional scaling with an unkown distance function"*, Psychometrika, vol. 27, 1962.

[10] W. F. DE LA VEGA, *"Sur deux techniques de sériation"*, note, Centre d'Analyse Documentaire pour l'Archéologie (CNRS), Marseille, 1971.