REVUE DE STATISTIQUE APPLIQUÉE

J. P. BENZÉCRI

Sur les algorithmes de classification

Revue de statistique appliquée, tome 19, n° 1 (1971), p. 17-26

<http://www.numdam.org/item?id=RSA_1971__19_1_17_0>

SUR LES ALGORITHMES DE CLASSIFICATION

J. P. BENZÉCRI

Professeur de Statistique à la Faculté des Sciences de Paris

Les nominalistes contemporains vont bien vite, quand ils parlent de formation de concepts à propos de subdivisions où l'arbitraire tient autant de place que dans la plupart de nos classifications. Aussi, préférant par prudence le continu au discontinu, les gradations aux oppositions, accordons-nous un intérêt particulier aux méthodes qui, comme l'analyse factorielle et ses variantes, cherchent à rendre accessibles à l'intuition des masses de données, en les représentant dans un espace de faible dimension. Mais cette représentation même peut révéler que les objets étudiés se répartissent en classes distinctes, visibles sur le diagramme spatial comme des îlots séparés les uns des autres, ou du moins reliés par des zones de faible densité. D'ailleurs si l'on ne savait reconnaître le discontinu dans le continu, il n'y aurait ni parole ni écriture. Tant donc la synthèse cognitive des surabondantes données des sciences, que l'analyse et la reconnaissance des formes des objets individuels, ne bénéficieront pleinement des nouveaux moyens de calculs que par la conception d'algorithmes qui permettent à une machine d'élaborer sans l'homme une classification.

Si les seules classifications utiles étaient celles qui sont inscrites avec une absolue netteté dans la nature des choses, on pourrait sans doute recourir à des algorithmes bien plus simples que ceux que l'on essaye aujourd'hui de programmer. Mais ne faut-il pas souvent déchiffrer une écriture brouillée ? Ne faut-il pas notamment dans les sciences de l'homme, ramener autant que possible la riche complexité du réel aux médiocres proportions de ce que nous pouvons concevoir ? D'où le choix difficile de critères qui selon les fins poursuivies structurent au mieux un domaine confus.

1 - ORDONNANCE ET PARTITION

Rappelons que l'on dit qu'on s'est donné une ordonnance sur un ensemble I si on a totalement ordonné l'ensemble des paires d'éléments de I : i.e. pour quatre éléments i, i', i'', i'''∈I, est posée l'une ou l'autre des deux inégalités :

$$\{i , i'\} < \{i'' , i'''\}, \quad \text{ou} : \{i'' , i'''\} < \{i , i'\} ;$$

inégalité qui intuitivement correspond à ceci : la distance entre i et i' est inférieure (ou supérieure) à celle entre i'' et i'''. Seulement, en donnant une ordonnance, on ne spécifie pas de distance, se bornant à des inégalités, et

. c'est justement le problème de l'analyse des proximités que de chercher, parmi les distances compatibles avec telle ordonnance, celles qui sont, en un certain sens, les plus simples (e.g. ; distances sur I, réalisé comme une partie d'un espace euclidien de dimension aussi faible que possible ; cf. Shepard (1962) et Benzécri (1964).

Nous présentons ici un autre problème, d'après de belles recherches de S. REGNER : étant donné sur I une ordonnance ω, quelles sont les partitions de I que ω suggère, en classes d'éléments deux à deux proches ? Il faut remarquer que ce problème est complémentaire du précédent : c'est justement quand une nette division de I en deux (ou un petit nombre de) classes, s'impose que la réalisation euclidienne de I est la plus indéterminée (cf Shepard (1962) I, p. 240)

Cherchons à exprimer l'accord entre une partition

$$\mathcal{P} = (I_1, \ldots, I_j, \ldots, I_p)$$

de I (en sous ensembles distincts I) et une ordonnance ω. Deux critères se présentent. Ou bien, (A), on voudra que les classes soient bien ramassées ; ou bien (B), on tiendra surtout à reconnaître les composantes connexes de I (i.e. les I_j devront être d'un seul tenant, et de plus si un I'_j contient i il devra contenir les éléments i' qui en sont assez proches). Ici, (B), l'Europe et l'Asie ne formeront qu'un tout dont on séparera la Corse et le Japon ; là (A) on distinguera en Europe les péninsules ibérique, italienne, balcanique, scandinave... Entre A et B le choix, n'est pas si facile qu'il le semble sur nos exemples, non plus que l'élaboration d'un compromis, ce qui invite à une étude mathématique. Auparavant, notons la parenté entre notre problème de la partition de I (en sous-ensembles disjoints $I_1, \ldots, I_p$), et celui du recouvrement de I pour des sous-ensembles I'_j non-nécessairement deux à deux d'intersection vide. D'un recouvrement il est possible de demander simultanément, (A'), que les I'_j soient des classes bien ramassées et, (B'), que deux éléments i, i' de I suffisamment voisins appartiennent à une même classe. Pour construire un recouvrement satisfaisant à (A') et (B'), on pourra construire une partition satisfaisant à (A) puis élargir les I en adjoignant à chacun d'eux les éléments de I qui sont proches de l'un de ses éléments.

Mathématiquement A et B se formulent ainsi : $\mathcal{P}$ sépare les couples d'éléments de I en deux classes C^-, C^+ : d'une part, C^-, les couples d'éléments situés dans une même partie I_j ; d'autre part, C^+, les couples d'éléments situés dans deux parties distinctes. Du point de vue B, $\mathcal{P}$ sera compatible avec ω si aussi peu que possible des couples de C^+ ne se rencontrent dans les derniers rangs. Du point de vue de A, on visera à ce que tout couple de C^- soit inférieur à tout couple de C^+.

Mais cherchons à être précis : on pourra demander (B) que moins de tant % des couples de C ne se rencontre dans les derniers 10 % de l'ensemble de tous les couples ; ou, (A), que des inégalités qui expriment que tout couple de C^- est inférieur à tout couple de C^+, le plus grand nombre possible soient vraies dans ω ; ou, encore, ce qui n'est pas la même chose, que le plus grand pourcentage possible de ces inégalités soient vraies dans ω. Ce que l'on peut craindre entre autre, c'est qu'un critère mathématique mal choisi tende toujours à émietter I en presque autant de parties qu'il a d'éléments ; ou au contraire interdise de le diviser. De plus la partition doit être d'un bon rendement informationnel : si l'on a fait p classes, l'infor-

mation fournie par la donnée de la classe d'un élément doit être aussi voisine que possible du maximum $\mathrm{Log}_2 p$, c'est-à-dire que les classes doivent être à peu près égales. Pour juger de la valeur des divers critères possibles, il nous paraît utile d'expérimenter en les appliquant à des exemples géométriques simples : c'est l'objet du présent travail de traiter quelques cas. On se placera du point de vue A, et dénombrera parmi les inégalités du type $C^- < C^+$, celles qui sont dans ω.

2 - OPPOSITION ET GRADATION

Soit I une suite de points étagés en progression arithmétique sur le segment (0,1) de la droite numérique : la distance usuelle munit I d'une ordonnance ω. Nous étudierons les partitions en deux classes (ou oppositions , comme bon-mauvais ou sauvage-domestique...) le mieux compatibles avec l'ordonnance ω définie par la gradation continue des points sur le segment. Ces partitions se présentent naturellement. Tout nombre $q \in (0,1)$ définit une partition de I en deux classes $\{I_q^-, I_q^+\}$, les éléments à gauche de q, et ceux à sa droite : cette partition (on l'appellera : partition q) est certainement, (avec la partition 1-q), la meilleure de toutes celles dont les deux classes sont dans le rapport q/1-q. Dans les calculs, (afin de faire des intégrales plutôt que des sommes), I, uniformément dense (0,1), sera remplacé par ce segment même, que nous noterons aussi (0,1). Au lieu de parler du nombre des inégalités vérifiées - on parlera du volume du domaine des inégalités vérifiées...

L'ensemble des quadruples (i_1, i_2, i_3, i_4) est un hypercube I^4 de volume 1 ; l'ordonnance ω divise I^4 en deux parties de même volume :

$$I(<) = \{(i_1, i_2, i_3, i_4) \mid |i_2 - i_1| < |i_4 - i_3|\},$$

et :

$$I(>) = \{(i_1, i_2, i_3, i_4) \mid |i_2 - i_1| > |i_4 - i_3|\}.$$

A chaque inégalité de l'ordonnance ω

$$\{i_1, i_2\} < \{i_3, i_4\},$$

correspondent quatre points de I (<) : car on ne change pas d'inégalité en permutant entre eux i_1 et i_2 ou i_3 et i_4. En sorte qu'on peut dire que l'ensemble des inégalités spécifiées par ω a pour volume ou mesure 1/8. (Si l'on traite I comme un ensemble de N points et non comme un continuum, il y a N^4 quadruples, N(N-1) k couples, et

$$\frac{1}{2} \cdot \left(\frac{N(N-1)}{2}\right) \cdot \left(\frac{N(N-1)}{2} - 1\right) \approx N^4/8$$

inégalités dans l'ordonnance ω, on retrouve ce rapport 1/8).

Les inégalités qu'implique la partition q sont les suivantes :

$$\forall i_1, i_2, i_3 \in (0,q) \; ; \; \forall j \in (q, 1) :$$

$$\{i_1, i_2\} < \{i_3, j\}$$

$$\forall i \in (0, q) ; \forall j_j, j_2, j_3 \in (q, 1)$$

$$\{j_1, j_2\} < \{j_3, i\}$$

L'ensemble des points de $I(<)$ concerné par ces inégalités a pour volume (en tenant compte de ce qu'à une même inégalité correspondent plusieurs points) :

$$2[q^3(1-q) + (1-q)^3 q] = 2q(1-q)[q^2 + (1-q)^2]$$

Le volume des inégalités correspondantes est quatre fois moindre, soit :

$$\frac{1}{2} q(1-q)[q^2 + (1-q)^2],$$

quantité dont le maximum est atteint pour $q = 1/2$, et vaut $1/16$.

D'où un premier résultat : ce sont les oppositions de rendement maximum qui impliquent le plus d'inégalités, la mesure en est $1/16$ soit la moitié de la mesure, $1/8$, de toutes celles que spécifie une ordonnance).

La mesure de l'ensemble des inégalités impliquées par la partition q, et compatibles avec ω est donnée par la demi-somme des volumes des deux domaines V et W de R^4, ci-dessous définis :

$$V_q = \{(x,y,z,t) \mid \{x,y,z\} \in (0,q) ; t \in (q,1) ; |x-y| < t-z\}$$

$$W_q = \{(x,y,z,t) \mid x \in (0,q) ; \{y,z,t\} \in (q,1) ; |y-z| < t-x\}$$

Ces volumes peuvent se calculer par des intégrales multiples ; cette méthode-là étant bien connue, nous préférons donner ici une méthode géométrique : V et W sont des parallélépipèdes, d'où on a enlevé certaines régions de forme assez simple pour que la seule formule suivante en donne le volume. Soit

$$D_h = \{(x,y,z,t) \mid \{x,y,z,t\} \in (0,\infty) ; x+y+z+t \leqslant h\}$$

$$\text{Vol}(D_h) = h^4/24$$

Dans la suite de l'étude, il apparaîtra clairement que :

$$\text{Vol}(V_{1-q}) = \text{Vol}(W_q)$$

c'est pourquoi nous bornerons nos considérations géométriques à V. V est un parallélépipède de volume $q^3(1-q)$, d'où on a ôté deux parties polyédrales V^1, V^2, d'intersection vide, et égales entre elles en volume.

$$V_q^1 = \{(x,y,z,t) \mid \{x,y,z\} \in (0,q) ; t \in (q,1) ; x-y > t-z\}$$

$$V_q^2 = \{(x,y,z,t) \mid \{x,y,z\} \in (0,q) ; t \in (q,1) ; y-x > t-z\}$$

En posant :

$$x' = q - x \; ; \; y' = y \; ; \; z' = q - z \; ; \; t' = t - q$$

Le domaine V^1 s'écrit :

$$V_q^1 = \{(x', y', z', t') \mid x', y', z' \in (0, q) \; ; \; t' \in (0, 1-q) \; ; \; x' + y' + z' + t' < q\}$$

Ici deux cas se présentent (voir figures 1 et 2, où l'on a dû omettre l'une des quatre dimension g') :

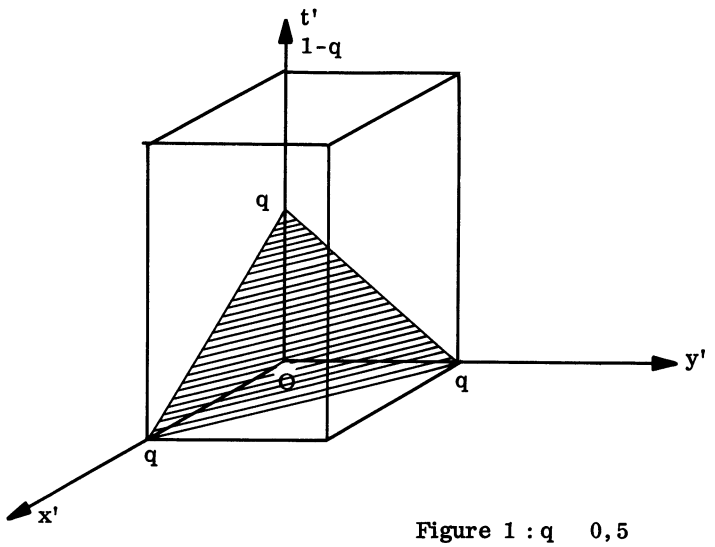


Figure 1 : q 0,5

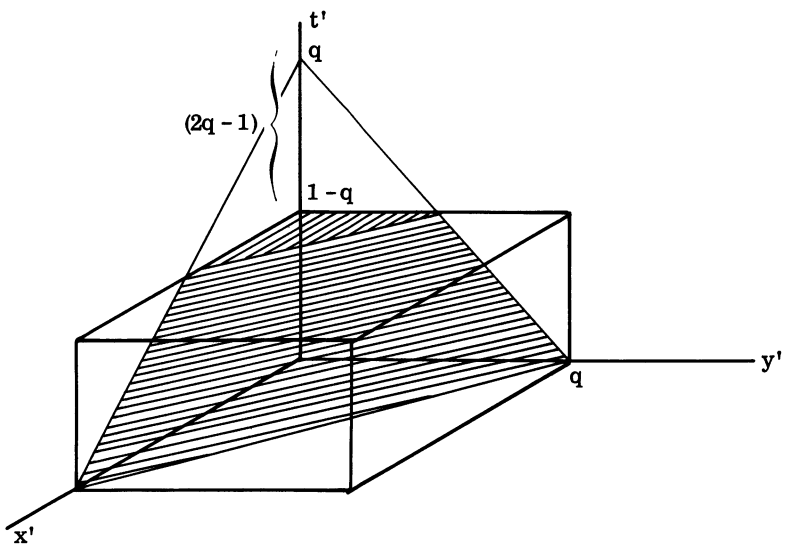


Figure 2 : q 0,5

a) $q < 1-q$, i.e. : $q \in (0,1/2)$

V_q^1 est la pyramide D_q définie ci-dessus ;

b) $q > 1-q$, i.e. $q \in (1/2,1)$

la pyramide D_q déborde du parallélipipède $(0,q)^3 \times (0,1-q)$, donc de V_q^1 d'une pyramide égale à D_{2q-1} et translatée de celle-ci parallèlement au 4° axe.

D'où les formules :

$$q \in (0,\ 1/2) : \text{Vol}\ (V_q^1) = q^4/24$$

$$q \in (1/2,1) : \text{Vol}\ (V_q^1) = q^4/24 - (2q-1)^4/24$$

L'on en déduit immédiatement le volume de V_q et W_q d'où la mesure de l'ensemble des inégalités spécifiées par la partition q et compatibles avec ω :

$$\mathcal{C}(q) = 1/2q\ (1-q)\ [2/3\ (q^2 + (1-q)^2) + 1/2\ (1-q)\ q]$$

ainsi que le taux ou rapport du nombre des inégalités vérifiées au total des inégalités spécifiées :

$$\mathcal{T}(q) = \frac{2}{3} + \frac{1}{2}\,\frac{q(1-q)}{q^2 + (1-q)^2}$$

Les courbes de $\mathcal{C}(q)$ et $\mathcal{T}(q)$ sont jointes (fig. 3,4) : ces deux fonctions sont maxima pour $q = 1/2$, mais le maximum de $\mathcal{C}(q)$ est beaucoup plus net.

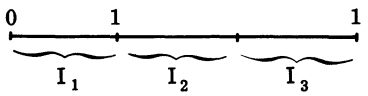
On pourrait envisager comment la présence de "trous" déplace le maximum de $\mathcal{T}$ et de $\mathcal{C}$. De façon précise, il faudrait ôter du segment (0,1) un (ou plusieurs) sous-intervalle J et reprendre le calcul de V, W etc. en interdisant à x,y,z,t de prendre des valeurs dans J.

3 - QUELQUES PARTITIONS EN PLUS DE DEUX CLASSES

Le but de ce § est de calculer $\mathcal{C}$ et $\mathcal{T}$ pour quelques partitions qui ne sont pas nécessairement des oppositions en deux classes.

3.1. Un segment en trois segments consécutifs égaux

La situation est la même qu'en 2 : I est assimilé au segment (0,1), et le volume total de l'ensemble des inégalités de l'ordonnance ω est 1/8. Mais on considère une partition en trois segments égaux, comme figuré ci-dessous:



Des calculs, analogues à ceux de 2, donnent :

$$\mathcal{C} = \frac{17}{4 \times 81}$$

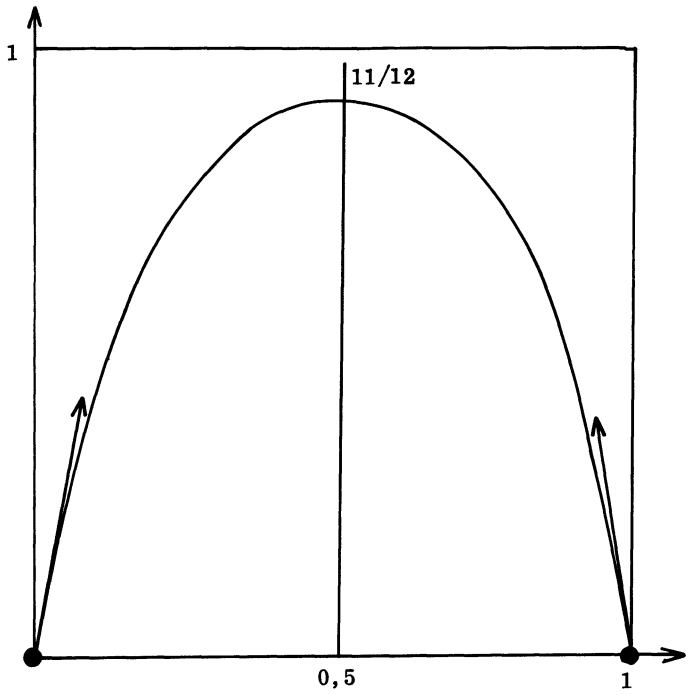


Figure 3

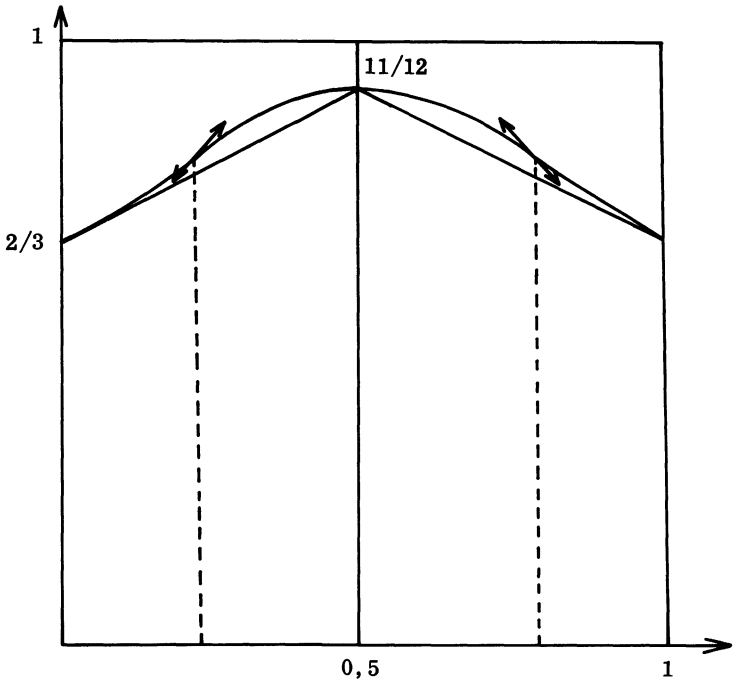


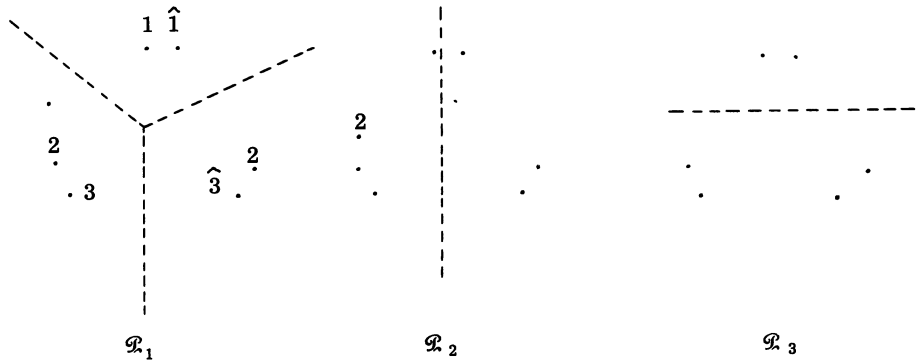
Figure 4

$$\mathfrak{C} = \frac{17}{18}$$

Comparons ces résultats à ceux pour une partition en deux segments égaux : on voit que la multiplication des classes a amélioré $\mathfrak{C}$ (qui se rapproche de 1, en passant de 11/12 à 17/18) ; mais $\mathfrak{c}$ a décru. En divisant I en n segments consécutifs égaux, on aurait encore accroissement de $\mathfrak{C}$ et décroissance de $\mathfrak{c}$.

3.2. Trois groupes de deux îlots

On suppose maintenant que I, d'effectif élevé, est réparti dans le plan en six îlots, groupés par paires aux sommets d'un triangle équilatéral, comme l'indique la figure suivante, où sont aussi notées trois partitions envisagées :



la partition qui paraît s'imposer est $\mathfrak{P}_1$; mais il est intéressant de vérifier au nom de quels critères numériques on la peut préférer à $\mathfrak{P}_2$, $\mathfrak{P}_3$, et aussi à la partition $\mathfrak{P}_4$ en six îlots. Dans les calculs on suppose comme précédemment qu'il s'agit d'une distribution continue de masse (ainsi l'ensemble des inégalités de ω a pour mesure 1/8). La figure particulière a été choisie pour que le calcul de $\mathfrak{c}$ et $\mathfrak{C}$ soit simple : on remarque que la distribution des distances deux à deux des points de I est pentamodale, se répartissant au voisinage de cinq valeurs principales qui sont : le rayon d'un îlot, la distance $1-\hat{1}$, la distance $1-2$, la distance $1-3$, la distance $1-\hat{3}$.

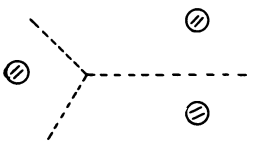
Les résultats sont présentés sur le tableau suivant :

	$\mathfrak{C}$	16 $\mathfrak{c}$
$\mathfrak{P}_1$	1	16/18 = 0,89
$\mathfrak{P}_2$	23/27 = 0,85	23/27 = 0,85
$\mathfrak{P}_3$	4/5 = 0,8	64/81 = 0,79
$\mathfrak{P}_4$	1	5/9 = 0,55

La partition $\mathfrak{P}_1$, la plus naturelle, se distingue à la fois par les meilleures valeurs de $\mathfrak{C}$ et de $\mathfrak{c}$. Mais la partition $\mathfrak{P}_4$, qui semble trop fine a aussi $\mathfrak{C}$ égal à 1.

3.3. Autres exemples typiques

Il semblerait utile de poursuivre l'expérimentation sur des cas analogues à ceux déjà traités. Outre un segment, considérer une boule, un cube... Après les trois sommets d'un triangle équilatéral, passer aux quatre sommets d'un tétraèdre régulier. Mais faisons ici une remarque sur le cas (e.g.) d'un triangle isocèle :



Du point de vue de l'ordonnance, cette figure est équivalente à :

(où les taches rondes ont été rapprochées jusqu'à une distance égale à leur diamètre) ; ce qui suggère une partition en deux classes, au lieu de trois auxquelles on songeait d'abord. Le calcul donne :

pour 3 classes : $\mathfrak{T} = 1$; $\mathfrak{C} = 1/18$.

pour 2 classes : $\mathfrak{T} = 1$; $\mathfrak{C} = 1/16$.

ce qui conduit à préférer la partition en deux classes.

4 - CONCLUSIONS

4.1. Le choix d'un critère

En l'état de notre expérience, il apparait que plus la partition est fine, (Plus les classes sont petites) plus il est généralement facile d'obtenir un $\mathfrak{T}$ voisin du maximum : maximiser $\mathfrak{T}$, est donc un critère dangereux qui risque de conduire à émietter l'ensemble I. En revanche maximiser $\mathfrak{C}$ (le nombre des inégalités spécifiées par la partition et compatibles avec l'ordonnance) semble conduire à des résultats naturels.

4.2. Complexité des algorithmes.

S. Regnier utilise un algorithme basé sur le critère suivant. L'ensemble I à n éléments étant muni d'une distance (ou d'un indice de proximité) d(i , i') variant entre 0 et 1, on lui associe le point de $R^{N(N-1)/2}$ dont les coordonnées sont les d(i , i'). Une partition de I définit aussi une telle distance (d(i , i') = 0 si i et i' sont de la même classe, 1 autrement) donc aussi un point de $R^{N(N-1)/2}$. La partition la mieux compatible avec la distance donnée sera celle qui comme point de $R^{N(N-1)/2}$, est la plus proche. L'algorithme d'optimisation est une suite d'essai : à chaque essai on déplace un seul élément i d'une classe à une autre de la partition (éventuellement on crée pour lui une classe nouvelle) et on calcule si cette modification a amélioré l'ac-

cord entre métrique et partition : comme déplacer i affecte au plus (N-1) des coordonnées du point qui dans $R^{N(N-1)/2}$ représente la partition, l'ordre de complexité d'un essai peut être mesuré par N. L'expérience a montré que cette complexité n'est pas excessive si N est, par exemple de l'ordre de 200.

Pour optimiser $\mathcal{C}$, il est naturel de procéder de même par essais successifs portant chacun sur un seul élément i de I. Mais à déplacer i on affecte un nombre d'inégalités qui est de l'ordre de N^3, et une telle complexité paraît prohibitive. Il faut donc se contenter d'un calcul approché de $\mathcal{C}$.

Divisons l'ensemble C des N(N-1)/2 couples d'éléments de I en 10 classes c_j d'égal effectif de telle sorte que si $j < j'$, tout couple c_j soit inférieur à tout couple de c_j pour l'ordonnance ω, (les classes c_j sont des segments consécutifs égaux de C pour l'ordre ω). Pour toute partition $\mathcal{P}$ on définit : (où si E est ensemble on note $\overline{\overline{E}}$ le nombre de ces éléments) :

$$C_j^+ = \overline{\overline{C^+ \wedge C_j}}$$

$$C_j^- = \overline{\overline{C^- \wedge C_j}}$$

$$C_j^- = C_1^- + C_2^- + \ldots + C_{j-1}^- + \frac{1}{2} C_j^-$$

Et cette formule nous paraît à la fois assez simple et assez précise pour être satisfaisante.

Notons enfin qu'il serait possible pour N très grand, de simplifier les calculs par une première classification préalable identifiant les points séparés par des distances dont leur rang dans l'ordonnance indique qu'elles sont très faibles.

BIBLIOGRAPHIE

J. P. BENZECRI - Sur l'analyse factorielle des proximités. Publications de l'Institut de Statistique de l'Université de Paris, 1964-1965.

R. N. SHEPARD - The analysis of proximities : scaling with an unknown distance function. I and II ; Psychometrika 1962.