

G. FLAMENBAUM

J. THIERY

J. P. BENZÉCRI

**Agrégation en boules de rayon fixe et
centres optimisés**

Les cahiers de l'analyse des données, tome 4, n° 3 (1979),
p. 357-364

http://www.numdam.org/item?id=CAD_1979__4_3_357_0

© Les cahiers de l'analyse des données, Dunod, 1979, tous droits réservés.

L'accès aux archives de la revue « Les cahiers de l'analyse des données » implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
<http://www.numdam.org/>

AGRÉGATION EN BOULES DE RAYON FIXE ET CENTRES OPTIMISÉS [BOULES OPTIMISÉES]

par G. Flamenbaum ⁽¹⁾, J. Thiery ⁽²⁾
et J. P. Benzécri ⁽³⁾

L'algorithme présenté ici, rentre dans le cadre des algorithmes d'agrégation autour de noyaux variables conçus par E. Diday et ses collaborateurs (Y. Ok ; A. Schroeder ;...). Il diffère de l'algorithme des *nuées dynamiques* (n. d.) en ce que tout point doit être à une distance du centre auquel on l'agrège, inférieure à un rayon R fixé *a priori* (mais dont le choix peut être guidé : cf § 3.3) : c'est pourquoi nous parlerons en bref, de *méthode des boules optimisées* (en abrégé b. o.). Outre qu'il produit des classes de rayon fixé (ce qui semble essentiel dans certaines applications cf § 3.3), il apparaît que l'algorithme a le mérite de converger en un nombre d'itérations moindre que celui requis pour les *nuées dynamiques* ; de plus après recalcul des centres d'agrégation il peut être nécessaire de créer de nouveaux centres (ce n'est jamais le cas avec les n. d.) ce qui semble favoriser une bonne répartition spatiale des classes. Comme celui des n. d., le présent algorithme part d'un ensemble fini I de points (éventuellement affectés de masses), dans un espace métrique quelconque E : toutefois dans cet exposé, on se bornera au cadre de la géométrie euclidienne. Les propriétés classiques d'inertie des nuages de points ont en effet des conséquences intéressantes, dont il est difficile de démontrer un analogue dans le cadre le plus général. Signalons seulement que l'algorithme des b. o. (comme celui des n. d.) s'applique non seulement si l'ensemble I est donné dans un espace métrique ambiant E , mais aussi à tout ensemble muni de masses et distances, pourvu que les centres d'agrégation soient des points de I : on substitue alors à la recherche du centre de gravité d'une classe (cf § 1.1) la recherche du point de cette classe relativement auquel l'écart moyen de la classe est le plus faible (cf § 2.1).

La présente note est divisée en trois § : au § 1, on décrit par la composition d'opérateurs simples les deux algorithmes des n. d. et des boules optimisées ; au § 2 on étudie la convergence des algorithmes ; au § 3 on considère les applications de l'algorithme des boules optimisées : soit pour traduire une partition directement utilisée, soit pour accélérer des traitements ultérieurs. Signalons qu'en vue de choisir le rayon R , on définit un nouveau critère global de la qualité d'une partition.

1 Description de l'algorithme : Nous utilisons des notations qui permettent de décrire parallèlement le présent algorithme et celui des n. d. .

Soit I un ensemble fini de points d'un espace euclidien de dimension finie. Les masses m_i des individus n'intervenant que pour le calcul des centres de gravité, on ne les précise pas, afin d'alléger les notations. L'ensemble I (ou plus généralement le nuage $N(I)$) est fixé une fois pour toutes.

(1) Dr G. Flamenbaum, (2) Dr J. Thiery. C.E.A. CEN/Cadarache St Paul-Lez-Durances.

(3) J.P. Benzécri. Professeur de statistique. Univ. P. & M. Curie.

Les auteurs remercient Monsieur B. Pullman, Professeur de biologie physico-chimique à l'université Pierre et Marie Curie.

Relativement à un ensemble C de centres et à une partition α de I en classes affectées à ces centres nous définirons plusieurs opérateurs d'affectation, ou de recalcul des centres. Ces opérateurs permettent de définir le présent algorithme, ainsi que l'algorithme n. d. de E. Diday.

1.1 L'opérateur de recalcul des centres $REC(\alpha, C)$: On donne d'abord les arguments de cet opérateur ; puis son effet. Les arguments sont (outre l'ensemble I fixé ici) : un ensemble fini C de points de E appelé ensemble des centres ; et une application α de I sur C , appelée application d'affectation : tout point $i \in I$, est affecté à un centre unique $c = \alpha(i) \in C$; de telle sorte qu'à tout centre c de C soit affecté au moins un point i de I (surjectivité de α).

$REC(\alpha, C) = (\alpha', C')$: est un ensemble de centres C' avec une application d'affectation α' , de I à C' . Voici la définition de (α', C') . A tout c de C correspond selon l'affectation α une classe $\alpha^{-1}(c)$: on note $g\alpha(c)$ le centre de gravité de cette classe ; et on pose : $\alpha' = g\alpha \circ \alpha$; $C' = \{g\alpha(c) | c \in C\}$; i.e. les nouveaux centres sont les centres de gravité des anciennes classes ; et chaque individu i est affecté au centre de gravité de la classe à laquelle il appartenait précédemment. Notons deux propriétés de l'opérateur REC :

1°) $REC(\alpha, C)$ ne dépend pas de l'ensemble C des centres, mais seulement de la partition de I définie par l'application α .

2°) Par suite, l'opérateur REC est idempotent, i.e. $REC \circ REC = REC$; de façon précise, si :

$$REC(\alpha, C) = (\alpha', C') ; REC \circ REC(\alpha, C) = REC(\alpha', C') = REC(\alpha, C) ;$$

en effet α et α' définissent la même partition de I .

1.2 L'opérateur d'affectation $AFF(\alpha, C) = AFF(I, C)$: En fait cet opérateur ne dépend pas d'une affectation antérieure éventuelle α , mais seulement de l'ensemble I des individus et d'un ensemble C de centres. Voici son effet :

$AFF(\alpha, C) = AFF(I, C) = (\alpha', C')$: par α' , chaque individu i de I est affecté à celui des points c de C dont il est le plus proche ; et C' est l'ensemble des centres c de C auxquels est ainsi affecté au moins un individu : $\alpha'(C) = C' \subset C$.

1.3 L'opérateur d'affectation et de création $ACR(\alpha, C) = ACR(I, C)$. En fait, ACR , comme AFF ne dépend pas d'une affectation antérieure α , mais seulement des ensembles I et C . Plus précisément comme on le verra, en suivant le calcul de ACR , cet opérateur dépend d'un ordre séquentiel dans lequel sont rangés les éléments i de I ; et de plus d'un paramètre numérique le rayon R , qui dans les §§ 1 et 2 est considéré comme une constante (On reviendra au § 3.3 sur le choix de R). On a :

$ACR(\alpha, I) = ACR(I, C) = (\alpha', C')$: ACR est semblable à AFF , mais en diffère parce que, du fait de la contrainte de rayon, il se peut que de nouveaux centres soient créés (tandis que par AFF , il se peut seulement que des centres soient supprimés). De façon précise on affecte successivement chaque élément i de I au centre c de C dont il est le plus proche ; mais seulement à la condition que la distance (i, c) soit inférieure ou égale à R . Sinon, on crée un nouveau centre qui n'est autre que i lui-même ; et la procédure d'affectation se poursuit en tenant compte non seulement de l'ensemble C des centres donnés initialement ; mais aussi des centres créés, chemin faisant.

1.4 Algorithmes à centres variables : Avec les opérateurs REC et AFF définis ci-dessus l'algorithme de E. Diday peut s'écrire comme une suite de paires (α, C) . On part d'un ensemble initial de centres C_0 (déterminé d'une façon quelconque) et on pose :

$$\begin{aligned}
 (\alpha_1, C_1) &= \text{AFF}(I, C_0); (\alpha_2, C_2) = \text{REC}(\alpha_1, C_1); \\
 (\alpha_3, C_3) &= \text{AFF}(\alpha_2, C_2); (\alpha_4, C_4) = \text{REC}(\alpha_3, C_3); \\
 &\dots\dots\dots \\
 (\alpha_{2n+1}, C_{2n+1}) &= \text{AFF}(\alpha_{2n}, C_{2n}); (\alpha_{2n+2}, C_{2n+2}) = \text{REC}(\alpha_{2n+1}, C_{2n+1});
 \end{aligned}$$

en bref on fait opérer alternativement AFF et REC.

Quant à l'algorithme des *boules optimisées* ; il s'obtient en substituant ACR à AFF : on est ainsi assuré qu'après l'application de ACR, tout individu est affecté à un centre dont il est distant de moins de R (ou de R au maximum) ; autrement dit toute classe $\alpha^{-1}(c)$ est incluse dans la boule de centre c et de rayon R. Plus précisément, cf fig 1, le domaine maximum d'extension de chaque classe $\alpha^{-1}(c)$ est une boule tronquée par les hyperplans médiateurs des segments joignant le centre c aux centres voisins. Tandis qu'avec l'algorithme n. d. le domaine d'une classe est le polyèdre convexe délimité par ces mêmes hyperplans médiateurs. De plus l'initialisation de programme ne requiert pas un ensemble C_0 : car on pose $(\alpha_1, C_1) = \text{ACR}(I, \emptyset)$: on part sans fixer de centres ; et on crée ceux-ci au fur et à mesure que l'on parcourt l'ensemble I. On a donc :

$$\begin{aligned}
 (\alpha_1, C_1) &= \text{ACR}(I, \emptyset) ; (\alpha_2, C_2) = \text{REC}(\alpha_1, C_1) ; \\
 (\alpha_3, C_3) &= \text{ACR}(\alpha_2, C_2); (\alpha_4, C_4) = \text{REC}(\alpha_3, C_3) ; \\
 &\dots\dots\dots \\
 (\alpha_{2n+1}, C_{2n+1}) &= \text{ACR}(\alpha_{2n}, C_{2n}); (\alpha_{2n+2}, C_{2n+2}) = \text{REC}(\alpha_{2n+1}, C_{2n+1}) ;
 \end{aligned}$$

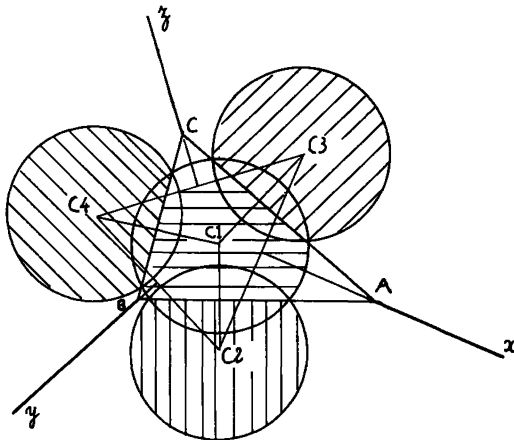


Figure 1 : domaines maxima d'extension des classes : on considère quatre centres c_1, c_2, c_3, c_4 : avec l'algorithme n.d. la classe 1 est un triangle ABC (dont les côtés sont les médiatrices des trois segments $(c_1, c_2), (c_1, c_3), (c_1, c_4)$) ; les classes 2, 3, 4 sont chacune délimitées par un côté du triangle ABC et deux des demi-droites Ax, By, Cz (e.g. Ax est la médiatrice de (c_2, c_3) etc) ; avec l'algorithme des boules optimisées on obtient pour domaines des boules tronquées (ici, en dim. 2 des disques tronqués) que l'on a indiqués par des hachures.

2 Optimisation et convergence : Pour démontrer (§ 2.3) la convergence des algorithmes définis au § 1.4 dans le cas d'un ensemble I fini, on suit (§ 2.2) la variation de deux critères de dispersion de classe définis au § 2.1 (on considérera de nouveau la dispersion des classes au § 3.2 pour choisir le rayon R).

2.1 Critères de dispersion des classes : Soit (α, C) un couple formé d'un ensemble C de points (appelés centres) et d'une application d'affectation α de I sur C (cf § 1.1) on définit :

$EX(\alpha, C)$: l'excentrement : on a en notant m_i la masse (éventuellement 1) de l'élément i :

$$EX(\alpha, C) = \sum \{m_i d^2(i, \alpha(i)) \mid i \in I\} ;$$

$VA(\alpha, C)$: la variance intraclasse (i.e. intérieure aux classes) : on a avec la définition de ga donnée au § 1.1 :

$$VA(\alpha, C) = EX(REC(\alpha, C)) = \sum \{m_i d^2(i, ga(\alpha(i))) \mid i \in I\} ;$$

i.e. $VA(\alpha, C)$ est la somme des variances des classes définie par l'application α : cette variance n'est autre que l'excentrement relative aux centres de gravité $ga(c)$ des classes $\alpha^{-1}(c)$; c'est-à-dire relativement aux centres déterminés par l'opérateur REC (cf § 1.1).

2.2 Variation des critères de dispersion : On a les relations suivantes entre les critères, et les opérateurs définis au § 1 :

$$EX(REC(\alpha, C)) = VA(\alpha, C) \leq EX(\alpha, C) ;$$

$$EX(AFF(\alpha, C)) \leq EX(\alpha, C) ;$$

$$EX(ACR(\alpha, C)) \leq EX(\alpha, C) ;$$

La première de ces relations résulte de la définition de VA et EX (cf § 2.1) et de ce que l'excentrement d'une classe est minimum lorsque celle-ci est rapportée à son centre de gravité (th. de Huyghens). Les deux dernières relations résultent de ce que l'opérateur AFF (ou l'opérateur ACR) ne modifie l'affectation $\alpha(i)$ d'un individu i qu'en vue de diminuer la distance $d(i, \alpha(i))$. De plus (à l'exception du cas où un point i équidistant de deux centres peut changer d'affectation sans que EX soit modifié), toute modification de (α, C) par AFF (ou par ACR) entraîne une diminution *stricte* de EX .

2.3 Convergence des algorithmes : Elle résulte des faits suivants qui valent aussi bien pour l'un ou l'autre des deux algorithmes définis au § 1.4.

1°) Il existe un nombre fini de partitions de l'ensemble fini I : donc sur l'ensemble des entiers n le couple $(\alpha_{2n}, C_{2n}) = REC(\alpha_{2n-1}, C_{2n-1})$, qui ne dépend que de la partition définie par α_{2n-1} (cf § 1.1) ne peut prendre qu'un nombre fini d'états différents ; et nécessairement il prendra même valeur pour deux entiers pairs différents, $2n$ et $2(n+p)$:

$$\exists n, p >_s 0 : (\alpha_{2n}, C_{2n}) = (\alpha_{2(n+p)}, C_{2(n+p)}) .$$

2°) La suite de valeurs $\{EX(\alpha_n, C_n) \mid n = 1, 2, \dots\}$ que prend l'excentrement, ne peut croître quand n croît. Il résulte des relations énoncées au § 2.2 que l'on a :

$$EX(1) \geq VA(1) = EX(2) \geq EX(3) \geq VA(3) = EX(4) \geq EX(5) \dots$$

$$\dots EX(2n-1) \geq VA(2n-1) = EX(2n) \geq EX(2n+1) \geq VA(2n+1) = EX(2n+2) \dots$$

Dans ces inégalités, comme désormais dans la suite la notation (n) est employée comme une abréviation de la paire (α_n, C_n) .

3°) Si (cf 1°) on a $(\alpha_{2n}, C_{2n}) = (\alpha_{2(n+p)}, C_{2(n+p)})$, entre $2n$ et $2(n+p)$ la quantité critère EX est nécessairement constante (cf 2°) : en particulier on a

$$EX(2n+1) = EX(ACR(2n)) = EX(2n).$$

Or, cf § 2.2, on a (sauf exception des points équidistants considérés au § 2.4), l'implication :

$$EX(ACR(\alpha, C)) = EX(\alpha, C) \Rightarrow ACR(\alpha, C) = (\alpha, C) ;$$

l'égalité $EX(2n+1) = EX(2n)$ implique donc $(2n+1) = (2n)$. Compte-tenu de la définition de la suite des $(\alpha_n, C_n) = (n)$, et de la relation d'idempotence REC, $REC = REC$ (cf § 1.1) on a donc :

$$(2n+1) = ACR(2n) = (2n) = REC(2n-1) ; \text{ et de même :}$$

$$(2n+2) = REC(2n+1) = REC \circ REC(2n-1) = REC(2n-1) = (2n).$$

On a ainsi établi que la suite des (α_n, C_n) a trois termes consécutifs égaux : $(2n) = (2n+1) = (2n+2)$: la suite est donc stationnaire à partir du rang $2n$ (car e.g. $(2n+3) = ACR(2n+2) = ACR(2n) = (2n+1) = (2n)$ etc) : on dit que l'algorithme converge.

2.4 Exceptions à la convergence : Le raisonnement du § 2.3, 3°) est en défaut dans le cas de points équidistants de deux centres, car l'opérateur d'affectation ACR (ou AFF) n'est pas défini de façon unique. Il est en effet possible d'imaginer une suite de (α_n, C_n) alternant indéfiniment : la figure 2 en donne un exemple simple. L'ensemble I est formé de 12 points

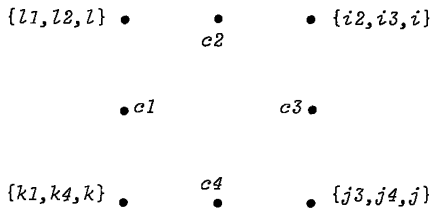


Figure 2 : exemple de couple (I, C) permettant d'alterner indéfiniment entre deux affectations.

disposés trois par trois aux sommets d'un carré de côté $2R$; l'ensemble C des centres n'est autre que l'ensemble des milieux des côtés du carré. Beaucoup d'affectations sont possibles, puisque chaque point de I est situé à la distance R de deux centres distincts (exactement cela fait 2^{12} affectations...). En particulier il est possible d'alterner entre deux affectations α et β sans que varient la quantité critère EX, ni les centres de gravité des classes : on pose :

$$\alpha\{i_1, k_1, k, l\} = c_1 ; \alpha\{i_2, i_2\} = c_2 ; \alpha\{i_3, j_3, i, j\} = c_3 ; \alpha\{j_4, k_4\} = c_4 ;$$

$$\beta\{i_1, k_1\} = c_1 ; \beta\{i_2, i_2, l, i\} = c_2 ; \beta\{i_3, j_3\} = c_3 ; \beta\{j_4, k_4, j, k\} = c_4 ;$$

en bref des points tels que $i_2, j_3, \dots$, affectés d'un indice numérique sont toujours affectés au centre de même indice que lui ; mais les points (i, j, k, l) sont affectés par α à un centre de même abscisse, et par β à un centre de même ordonnée.

Toutefois, dans la pratique, l'opérateur d'affectation ACR est réalisé par un programme de calcul qui ne laisse place à aucune indétermination (e.g. tout individu, en cas d'équidistance, est rattaché au centre de plus faible indice...) il ne semble donc pas qu'une alternance comme celle entre α et β , soit possible. Il y aura donc convergence de l'algorithme.

3 Application de l'algorithme des boules optimisées : Le présent algorithme (b.o.) comme celui des nuées dynamiques (n.d.) munit un ensemble I d'une partition en classes c dont l'étude directe peut être intéressante (§ 3.1). Du fait de la rapidité des calculs, on peut essayer plusieurs valeurs du rayon R et choisir d'après divers critères mesurant la qualité des partitions construites (§§ 3.2 & 3.3). De plus l'algorithme des b. o. semble particulièrement désigné pour préparer l'action d'un algorithme accéléré de classification ascendante hiérarchique (§ 3.4).

3.1 Typologie des conformations d'une molécule : Comme beaucoup d'autres méthodes générales d'analyse des données, l'algorithme des b. o. a été proposé à propos d'un cas particulier. Il s'agissait ici de physico-chimie ; les individus à classer étant un ensemble I de conformations qu'est susceptible de prendre la molécule d'un polypeptide (l'angiotensine II ; et une deuxième molécule semblable). A l'analyse conformationnelle, nous consacrons dans ces *Cahiers* un autre article (cf ce *Cahier* p 339) et G. F. : Thèse) : on se borne ici à rendre compte de l'utilisation du programme de calcul. Il est apparu que la convergence de l'algorithme des b. o. requiert sur nos données un nombre d'itérations (e.g. 4) encore plus faible que celui des n. d. . C'est ce qui a permis de multiplier les expériences d'une part en utilisant de nombreuses formules de calcul des distances entre conformations (sur ces formules, cf *Cahiers loc. cit.*) et d'autre part en faisant varier pas-à-pas le rayon R (ou distance maxima d'un individu i au centre auquel il est affecté). Ces essais suggèrent de choisir R (§ 3.3) d'après des critères globaux que l'on définit ci-dessous (§ 3.2).

3.2 Critères globaux de la qualité d'une partition : D'une partition d'un ensemble I en classes, on requiert deux qualités contradictoires : d'une part, que le nombre des classes soit peu élevé, en sorte que la partition donne de l'ensemble un schéma simple ; d'autre part, que chacune des classes soit aussi peu dispersée que possible, en sorte que le schéma obtenu en assimilant l'ensemble à quelques classes ponctuelles, soit fidèle à la réalité. Il est facile d'assurer une seule de ces qualités aux dépens de l'autre : soit en prenant une seule classe égale à I (simplicité) soit en faisant de chaque point une classe ($\text{Card } C = \text{Card } I$; fidélité au détail des données) : mais seule une solution de compromis peut être à peu près satisfaisante des deux points de vue. Sans prétendre donner de solution absolue, on indiquera la voie suivie dans la thèse. On utilise quatre quantités critères (liées entre elles) : $E(\alpha, C)$; $H(\alpha, C)$; $G(\alpha, C)$; $Q(\alpha, C)$. On a noté (α, C) pour conserver l'écriture adoptée au § 1 : bien que les quatre critères E , H , G , Q dépendent non de la place des centres mais seulement de la partition en classes :

$I = \cup \{ \alpha^{-1}(c) \mid c \in C \}$. Voici les définitions.

a). $E(\alpha, C)$: rapport de la variance interclasse à la variance totale de I . On rappelle que la variance totale VI de I

$(\sum \{ \sum_i d^2(g, i) \mid i \in I \})$, où g est centre de gravité de I) est la somme de la variance intraclasse (intérieure aux classes) $VA(\alpha, C)$ (cf § 2.1) et de la variance interclasse (i.e. entre les classes) ou variance VC du nuage des centres de gravité des classes. On a donc

$$E(\alpha, C) = (VI - VA(\alpha, C)) / VI = 1 - (VA(\alpha, C) / VI) = VC / VI \in (0, 1)$$

Quant à la fidélité de la représentation, l'optimum est $E(\alpha, C) = 1$: alors $VA = 0$: les classes ont chacune variance nulle, elles sont concentrées en un point.

b) $H(\alpha; C)$: information apportée par la connaissance de la classe d'un individu (cf Shannon, Brillouin, Kolmogorov ; etc). Continuons de noter :

$m_I = \sum \{m_i | i \in I\}$: masse totale de I ;

$m_c = \sum \{m_i | i \in I : \alpha(i) = c\}$: masse de la classe c ;

$p_c = m_c / m_I$: masse relative (ou probabilité) de la classe c .

$H(\alpha; C) = - \sum \{p_c \log_2 p_c | c \in C\}$.

Pour un nombre donné de classes $N = \text{Card } C$, l'information est maximale lorsque les classes ont toutes même probabilité (qui vaut donc $1/N$) ; on a alors : $H = \log_2 N$.

c) $G(\alpha; C) = H(\alpha; C) / (\log_2 \text{Card } C)$: quotient de l'information par le maximum compatible avec le nombre de classes ; ce rapport (comme $E(\alpha; C)$) est compris entre zéro et 1 : la valeur 1 étant un optimum, dont on s'approche d'autant plus que les individus sont plus également répartis entre les classes. On a dit ci-dessus que pour la simplicité du schéma, il est souhaitable que $\text{Card } C = N$ soit aussi faible que possible ; d'autre part l'information apportée par la partition croît avec N , mais n'est pas directement mesurée par N : le quotient $G(\alpha; C)$ est pour l'information comme un rendement, (d'autant plus voisin de 1 que la dépense, ici un nombre N de classes, a été mieux utilisée).

d) $Q(\alpha; C) = E(\alpha; C) \times G(\alpha; C)$: Il est inutile de diviser l'ensemble I en classes d'égal effectif (afin de rendre G maximum) si d'autre part l'information apportée par la partition n'a pas un sens géométrique direct, c'est-à-dire si les classes ne sont pas faiblement dispersées. On rejoint ici un paradoxe de l'information signalé par L. Brillouin : donner (pour un nombre réel compris entre 0 et 1) la 3^{ème} décimale sans les deux premières, c'est donner une information que la théorie mesure par $\log_2 10$, mais qui est en général inutilisable : car c'est diviser le segment (0,1) en 10 classes dont chacune compte 100 intervalles équidistants répartis de 0 à 1... Le produit Q rend compte simultanément du rendement en information (par G) et de la compacité des classes (par E).

3.3 Choix du rayon R : Comme on l'a dit au § 3.1, la rapidité de l'algorithme des b. o. a permis de faire de nombreux essais. Pour un même ensemble I de données on peut définir plusieurs codages aboutissant à différentes formules de distance entre éléments de I . La formule de distance étant fixée, on a appliqué l'algorithme des b. o. avec 12 valeurs différentes de R , définies relativement au diamètre D de I , (i.e. au maximum de la distance entre deux points de I) :

$R = 0,05 D$; $R = 0,10 D$; ... ; $R = 0,70 D$.

(il eût été, à la réflexion, préférable de prendre pour étalon de longueur non le diamètre D de I ; mais plutôt le maximum Ray, de la distance d'un point i de I au centre de gravité g de I ...). Il est dès lors possible de tracer les courbes donnant en fonction de R la variation des divers critères globaux caractérisant la partition optimale. Voici des exemples de telles courbes. On constate (ce n'est pas toujours le cas !) que les graphiques présentent des points de rupture (marqués RE, RQ sur la fig. 3) qui peuvent suggérer le choix de R .

Remarque : Le problème du recouvrement non d'un ensemble fini I , mais d'un espace continu, par un ensemble de boules est associé en analyse aux noms de Borel et de Lebesgue. Plus récemment Kolmogorov et ses élèves (Vitouchkine etc.) ont étudié pour des espaces de dimension infinie la variation en fonction de R du nombre de boules de rayon R requis pour recouvrir une boule de rayon 1. Sans se prêter à des applications directes, leurs méthodes pourraient inspirer des recherches statistiques.

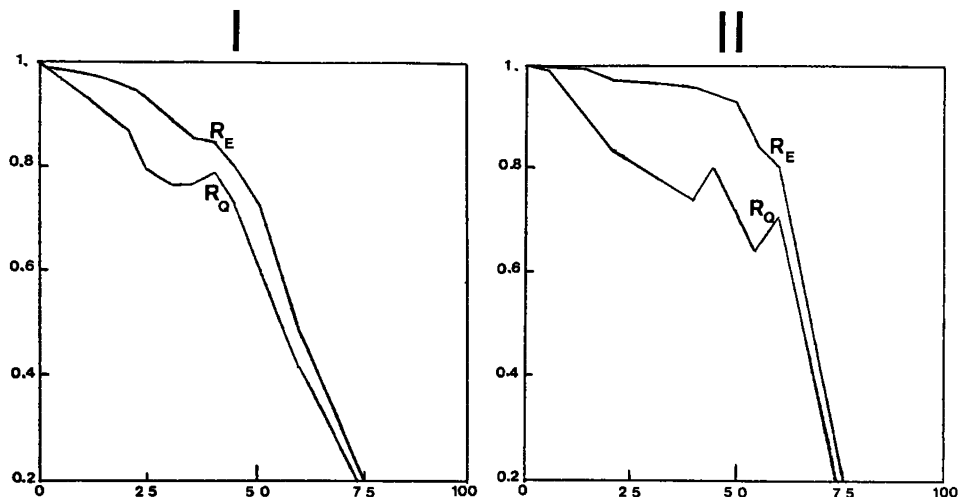


Figure 3 : variation des critères $E(\alpha;C)$ (courbe supérieure) et $Q(\alpha;C)$ (courbe inférieure) en fonction du rayon R choisi ; l'axe des abscisses (axe des R) est gradué en centièmes du diamètre D de I ; chaque graphique correspond à une formule de distance particulière, appliquée aux mêmes données.

3.4 Préparation des très grands ensembles de données : Par sa rapidité, l'algorithme des b. o. permet de réduire un tel ensemble (e.g. 10,000) à un nombre de classes (e.g. quelques centaines) auxquelles on applique un autre traitement, tel que classification ascendante hiérarchique (C.A.H.), ou analyse factorielle...

Ici, la contrainte de rayon des classes a un réel intérêt : puisqu'elle interdit que les domaines des classes ne s'étendent à l'infini (ce qui est possible par les n. d. : cf fig. 1), l'algorithme des b. o. reconnaîtra la présence d'individus excentriques qui formeront une classe à eux-seuls : ces individus sont généralement (e.g. en analyse factorielle) écartés de l'étude principale.

De plus il existe aujourd'hui des programmes accélérés de C.A.H., aptes à traiter en un temps acceptable de très grands ensembles de données (cf M. Bruynooghe ; *Cahiers* 1978 Vol III n° 1 ; et M. Jambu & M.O. Lebeaux ; Dunod 1979), à condition toutefois que soient triés les couples (i, i') séparés par une distance inférieure à un seuil R convenablement choisi. Présentement ce tri ne peut être effectué qu'en calculant effectivement toutes les distances entre points de I , i.e. si

$\text{Card } I = n$, en un temps de calcul de l'ordre de n^2 . Mais si l'algorithme des b. o. (dont le temps de calcul croît comme n) a été appliqué à I , avec une valeur R du rayon, on sait que si $d(i, i') < R$, on a nécessairement entre les centres $\alpha(i)$ et $\alpha(i')$ auxquels sont affectés i et i' une distance $d(\alpha(i), \alpha(i')) < 3R$ (car $d(i, \alpha(i))$ et $d(i', \alpha(i'))$ sont tous deux bornés par R) : on aura donc seulement à calculer effectivement les distances séparant des couples de points i, i' affectés à des centres $\alpha(i)$, $\alpha(i')$ distants de moins de $3R$. Enfin la vue globale simplifiée que l'algorithme des b. o. donne de l'ensemble I , aidera à choisir le seuil R pour l'algorithme accéléré de C.A.H. (e.g. on calculera rapidement, une estimation du nombre des paires (i, i') telles que $d(i, i')$ soit inférieure à R , à $2R$ etc.).