

REVUE DE STATISTIQUE APPLIQUÉE

P. DUCIMETIÈRE

Les méthodes de la classification numérique

Revue de statistique appliquée, tome 18, n° 4 (1970), p. 5-25

[<http://www.numdam.org/item?id=RSA_1970__18_4_5_0>](http://www.numdam.org/item?id=RSA_1970__18_4_5_0)

© Société française de statistique, 1970, tous droits réservés.

L'accès aux archives de la revue « *Revue de statistique appliquée* » (<http://www.sfds.asso.fr/publicat/rsa.htm>) 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/>

LES MÉTHODES DE LA CLASSIFICATION NUMÉRIQUE

P. DUCIMETIÈRE

Unité de Recherches Statistiques de l'INSERM

INTRODUCTION

Dans le cadre de l'application des méthodes multivariées à l'analyse de données nombreuses, la distinction entre les méthodes de classification et de classement est aujourd'hui bien précisée (DAGNELIE, 1966).

Le classement d'objets dans des catégories définies a priori doit être envisagé comme un problème de décision statistique et les méthodes de discrimination sont d'ores et déjà utilement appliquées.

Le problème de la classification d'objets, c'est à dire de leur regroupement en classes est généralement abordé par la recherche de représentations simplifiées de l'ensemble des données (analyses en composantes principales, analyses factorielles...). L'observation des représentations géométriques obtenues peut alors mettre en évidence une classification éventuelle des données. De telles méthodes, descriptives et par là universelles ne s'attachent pas à résoudre directement le problème du regroupement d'objets en classes.

La "systématique", étude scientifique des différents types d'organismes et de leurs relations (SIMPSON, 1961), se préoccupe depuis fort longtemps de la classification des êtres vivants. Les tentatives de formulation théorique des problèmes de classification et les diverses techniques proposées pour les résoudre (ou "taxonomie") se sont développées dans ce domaine (SOKAL & SNEATH, 1963). Depuis quelques années l'intérêt des chercheurs pour les méthodes taxonomiques se manifeste dans d'autres disciplines comme l'écologie (WILLIAMS & LAMBERT, 1959), la géologie (GOWER, 1970), la linguistique (NEEDHAM, 1964), la recherche médicale (FRASER & BARON, 1968)... Quelques exemples empruntés à ces divers domaines montrent que la notion de classification est très imprécise.

Exemple 1 :

Sur un ensemble de sujets atteints d'une même maladie, une série de symptômes sont relevés. Le chercheur fait l'hypothèse que cette population est peut-être "hétérogène" et que plusieurs "types de maladies" sont représentés. Comment classer les sujets de manière à mettre en évidence ces différentes formes ?

Exemple 2 :

Un botaniste recueille un ensemble de plantes appartenant à la même famille. Il souhaite regrouper ces espèces en genres, sous-genres et autres niveaux.

Exemple 3 :

Dans le but d'établir un système bibliographique, on étudie l'ensemble des mots utilisés dans une même discipline. Comment établir un système de mots-clefs, chacun regroupant un certain nombre de mots ?

Des remarques de trois types peuvent être faites.

a) La nature des objets à classer

Dans l'exemple 1, l'ensemble des sujets peut être considéré comme des échantillons tirés d'un certain nombre de populations à définir. Le problème est de regrouper les sujets de la manière la plus "vraisemblable" pour obtenir une représentation des univers de départ. Dans les exemples 2 et 3 les objets forment un ensemble exhaustif (même dans l'exemple 2 car la variabilité intra-espèce est négligée). Alors qu'une classification des espèces ou des mots est recherchée de toute façon, une classification des maladies n'a de sens que si l'hypothèse de l'existence de plusieurs types de maladies est vraie. Cette hypothèse mériterait d'être testée, mais nous ne connaissons pas à l'heure actuelle de test permettant de résoudre ce problème sous sa forme générale.

b) Le nombre de regroupements à effectuer

Dans l'exemple 1, si l'hypothèse de l'existence de plusieurs groupes est vraie, la seule classification intéressante à effectuer est celle qui isole le mieux sur l'échantillon les différents types. De même dans l'exemple 3, le regroupement en mots-clefs à utiliser est celui qui rend maximum une certaine fonction d'intérêt, fonction du nombre de mots-clefs construits, du "bruit" moyen introduit lors d'une recherche bibliographique...

Par contre dans l'exemple 2, les espèces seront regroupées de manière assez lâche en genres, puis à l'intérieur d'un genre en sous-genres. Ainsi chaque espèce n'appartient pas à un seul groupe mais à une hiérarchie de groupes.

c) Les conséquences de la classification

Alors que dans l'exemple 3 l'intérêt de la classification ne relève que de l'action, il se place sur le plan de la connaissance dans les exemples 1 et 2. Dans ces conditions, la difficulté réside dans le fait qu'une classification n'est jamais unique, elle dépend de la description des objets mais aussi de la méthode pratique de classification utilisée. Seule une certaine stabilité de la classification par rapport à diverses descriptions des objets et divers algorithmes permet une justification a posteriori.

Aucune théorie de la classification ne paraît être établie actuellement en tenant compte des divers problèmes évoqués sur ces exemples. Une introduction élémentaire aux principales méthodes existantes de classification numérique et quelques résultats originaux sont proposés dans ce travail. Ainsi que le souhaite P. DAGNELIE, une comparaison des résultats fournis par ces méthodes sur des exemples nombreux devrait permettre de juger plus clairement leur utilité.

La notion de ressemblance, la construction des classifications hiérarchiques puis non hiérarchiques et les méthodes de segmentation seront successivement étudiées.

1. - LA MESURE DE LA RESSEMBLANCE

1.1 - La ressemblance des objets

Soit E un ensemble de n objets sur chacun desquels un ensemble de p observations sont effectuées. Nous chercherons dans ce paragraphe à définir une ressemblance des objets pris deux à deux à partir de leur description.

1.1.1 - Définitions et notations

- Soit x un objet de E sur lequel une observation est effectuée, par exemple la mesure d'un paramètre, la présence d'un attribut, l'appréciation qualitative d'un phénomène. Soit A l'ensemble des modalités que peut prendre l'observation.

Nous supposons défini l'écart entre deux observations a_i et a_j du même phénomène. Il s'agit d'une fonction numérique non négative définie sur $A \times A$ vérifiant les conditions suivantes :

- $e(a_i, a_j) = e(a_j, a_i)$
- $e(a_i, a_i) = 0$

En principe il n'est pas nécessaire de supposer que e soit une distance sur A . Le plus souvent, cependant, les observations sont repérées sur une même échelle et alors $e(a_i, a_j)$ sera pris égal à $|a_i - a_j|$ avec des notations évidentes.

- Soit deux objets x et y de E et $x_1, \dots, x_p$; $y_1, \dots, y_p$ les observations de p variables effectuées sur chacun d'eux. L'ensemble E des objets munis de cette description peut alors être considéré comme un sous-ensemble E_p de $A_1 \times A_2 \times \dots \times A_p$ où A_i est l'ensemble de modalités de la i ème variable.

L'écart entre les deux objets x et y est une fonction numérique, non négative définie sur $E_p \times E_p$ vérifiant les conditions suivantes :

- $D(x, y) = D(y, x)$

- Quel que soit i , les deux fonctions $e(x_i, y_i)$ et la restriction de $D(x, y)$ à l'ensemble $A \times A$ sont simultanément croissantes ou décroissantes au sens large.

La ressemblance $S(x, y)$ des objets est définie de la même manière, les fonctions $e(x_i, y_i)$ et la restriction de $S(x, y)$ à $A_i \times A_i$ variant en sens inverse quel que soit i .

Par convention nous n'utiliserons l'écart $D(x, y)$ entre objets que si $D(x, y)$ définit une distance sur E , dans tous les autres cas nous préférons parler en termes de ressemblance.

De nombreux coefficients ou distances ont été proposés :

1.1.2 - Les variables correspondent à la présence ou l'absence d'un attribut :

Ce cas est particulièrement étudié dans les ouvrages de taxonomie (SOKAL & SNEATH, 1963), (LERMAN, 1970), car les données peuvent toujours être mises sous cette forme au prix de certains regroupements.

Si la même importance est donnée à chaque attribut dans la définition d'une ressemblance entre deux objets, $S(x, y)$ n'est fonction que du nombre a d'attributs possédés en commun par x et y , du nombre b d'attributs possédés par x et non par y et du nombre c d'attributs non possédés par x et possédés par y par exemple. Cette hypothèse signifie que la ressemblance entre x et y est résumé par le tableau suivant :

		Objet x		
		+	-	
Objet y	+	a	c	
	-	b	d	avec $a + b + c + d = p$ nombre d'attributs

LERMAN a montré que tout coefficient de ressemblance issu de ce tableau est une fonction décroissante symétrique par rapport à b et c et croissante par rapport à a . Si chaque objet x est représenté par un point dans l'espace $[0, 1]^p$, il faut remarquer que la plupart des coefficients décrits par SOKAL & SNEATH sont fonctionnellement reliés à deux caractéristiques géométriques de cet espace :

- le carré de la distance euclidienne entre les points x et y

$$\sum (x_i - y_i)^2 = \sum |x_i - y_i| = b + c$$

- le produit scalaire des vecteurs x et y :

$$\sum x_i y_i = a$$

auxquelles nous pouvons ajouter le cosinus de l'angle des vecteurs x et y (coefficient d'OCHIAI)

$$\frac{a}{\sqrt{a+b} \sqrt{a+c}}$$

Si, de plus, une normalisation est effectuée sur les objets (et non sur les variables) de manière que la somme des coordonnées du point x soit nulle et que la longueur du vecteur x soit unitaire, le carré de la distance euclidienne entre deux points x et y s'écrit :

$$2p(1 - S_\phi) \text{ où } S_\phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} \text{ est le coefficient}$$

de ressemblance proposé par PEARSON.

- Une pondération "objective" des différents attributs dans le coefficient de ressemblance peut être souhaitable, ce dernier n'est plus fonction seulement des quantités a, b, c, d. Il est possible par exemple de donner un poids à la présence d'un attribut plus important s'il est rare dans l'ensemble E que s'il est fréquent (coefficient de SMIRNOF décrit dans l'ouvrage de SOKAL & SNEATH).

1.1.3 - Les variables sont quantitatives :

Il est naturel dans ce cas de choisir une distance entre objets plutôt qu'un coefficient de ressemblance et son choix peut être guidé si nous imposons la propriété que cette distance soit invariante par certains changements de variables.

La propriété la plus générale d'invariance pour une distance $\varphi(x, y)$ peut s'écrire :

$$\varphi(x, y) = \varphi[f(x), f(y)] \text{ où } f(x) \text{ représente } \begin{matrix} (f_1(x_1 \dots x_p) \\ (f_p(x_1 \dots x_p) \end{matrix}$$

Suivant la nature des variables $x_1 \dots x_p$, il est parfois possible d'effectuer les changements de variables qui assurent les invariances souhaitées. Par exemple, le choix d'une distance fonction seulement des $|x_1 - y_1|$ assure l'invariance par translation, la normalisation des écarts $\frac{|x_1 - y_1|}{s_1}$ où s_1 est l'écart type sur E de la ième variable permet les changements d'unité. La distance généralisée (MAHALANOBIS), définie sur E est invariante par toute transformation linéaire non singulière des variables et le choix de cette distance s'impose particulièrement lorsque les mesures faites sont homogènes (un ensemble de longueurs par exemple). Peu de travaux utilisent cette distance bien qu'elle soit très souvent recommandée (RUBIN, 1967).

1.2 - La ressemblance des variables

Le problème de la classification de n objets à partir d'une description formée de p variables est historiquement relié au problème de la classification de n variables observées sur un ensemble de p objets (Méthodes Q et Méthodes R dans la terminologie de SOKAL & SNEATH).

Les deux problèmes ne sont qu'apparemment symétriques en particulier dans le cadre de l'exemple 1 de l'introduction. Alors que l'ensemble des objets est aléatoire, les variables mesurées sont à la disposition du chercheur et nous ne pouvons dire que leur ensemble forme un échantillon représentatif de l'ensemble des variables qu'il serait possible de mesurer. Cette différence profonde de nature a plusieurs conséquences.

Les définitions du paragraphe 1.1.1 ne peuvent être étendues. En effet il faudrait définir un écart $e(x_i, y_i)$ entre les observations de deux variables x et y effectuées sur le même ième sujet. En général cette notion n'a aucun sens précis. Une certaine stabilité de la ressemblance de deux variables est assurée si nous changeons l'échantillon des sujets, stabilité qui ne peut être supposée pour la ressemblance de deux objets lorsque leur description est modifiée.

La pondération éventuelle des diverses variables entrant dans la description d'objets a regrouper a un sens très clair, un coefficient de ressemblance entre variables ne peut être que symétrique par rapport aux objets.

Dans le cas de variables qualitatives à deux classes, cette dernière remarque entraîne qu'un coefficient de ressemblance entre variables n'est fonction que des quantités a, b, c, d , du tableau de contingence habituel à 4 cases où a, b, c, d , sont de nombres d'objets.

		variable x		
		+	-	
variable y	+	a	b	
	-	c	d	

avec $a + b + c + d = p$ nombre d'objets

De nombreux coefficients d'association ont été étudiés (KRUSKAL & GOODMAN, 1954) en particulier ceux reliés au χ^2 d'association entre x et y comme S_ϕ défini plus haut.

Il faut remarquer à ce sujet une ambiguïté qu'il est nécessaire de lever dans les applications. La ressemblance entre x et y est clairement maximum si $b = c = 0$ ($S_\phi = 1$) c'est à dire si les deux attributs sont simultanément présents ou absents chez tous les sujets. Cependant, si $a = d = 0$ ($S_\phi = 1$) nous pouvons soit considérer que les attributs x et y ont une ressemblance maximum car ils ne sont jamais associés chez un même objet ou au contraire qu'elle est maximum, en changeant la définition de l'un des attributs.

Selon le choix fait S_ϕ ou $|S_\phi|$ peut être préféré comme coefficient de ressemblance.

Lorsque les variables sont quantitatives, une propriété générale d'invariance d'un coefficient de ressemblance peut s'écrire :

$$\phi(x, y) = \phi(f(x), g(y)) \text{ où } f(x) \text{ par exemple représente } [f(x_1) \dots f(x_p)]$$

La ressemblance entre x et y étant maximum si ces variables sont reliées fonctionnellement. Le coefficient de corrélation habituel entre x et y a cette propriété dans le cadre des relations fonctionnelles linéaires. Il est invariant pour des fonctions f et g linéaires simultanément croissantes ou décroissantes et symétrique par rapport aux objets.

Un problème identique à celui posé plus haut demeure quant au choix de r ou $|r|$.

Plus généralement, si nous restreignons l'ensemble des relations fonctionnelles entre x et y aux fonctions monotones, le coefficient de corrélation des rangs (ou sa valeur absolue) exprime l'intensité d'une telle relation et donc une ressemblance entre les variables.

Désormais, le problème de la classification concernera un ensemble E de n objets (pouvant être des variables), ensemble muni d'un coefficient de ressemblance ou d'une distance.

1.3 - L'ordonnance associée à une matrice de ressemblance

L'ensemble $(E \times E)$ des couples d'objets est ordonné (ordonnance sur E) et nous écrirons $(x, y) < (z, t)$ si la ressemblance (resp. la distance) entre x et y est supérieure (resp. inférieure) à celle entre z et t . Pour simplifier nous supposons dans la suite que l'ordre défini sur $E \times E$ est total. Ainsi nous faisons correspondre à la matrice de ressemblance, la matrice formée des rangs des couples d'objets dans l'ordonnance sur E .

Si deux coefficients sont reliés par une fonction croissante au sens strict, la matrice des rangs est identique.

C'est ainsi que, dans le cas des variables qualitatives les coefficients de SOKAL & MICHENER, HARMAN, ROGERS & TANIMOTO par exemple (voir SOKAL & SNEATH, 1963) fournissent la même matrice de rang. De plus, si le nombre d'attributs possédés par tous les objets est le même, la position respective des points x et y dans l'espace $[0, 1]^p$ ne dépend que d'un seul paramètre et les coefficients reliés aux caractéristiques géométriques dans cet espace conduisent à la même matrice de rang.

Les méthodes de classification construites à partir de la matrice de rang sont développées dans (LERMAN, 1970), certaines d'entre elles sont introduites dans ce travail sous une forme différente.

2. - LA CLASSIFICATION HIERARCHIQUE

Nous donnerons une définition élémentaire de la classification et nous montrerons à partir d'exemples la nécessité d'étendre cette définition.

2.1 - Classer un ensemble E d'objets équivaut à définir sur E une partition telle que les objets qui se ressemblent soient regroupés et les objets qui ne se ressemblent pas appartiennent à des sous-groupes différents (Définition 1)

Nous pouvons remarquer que cette définition conduit à rechercher des partitions de E , à l'exclusion des recouvrements de E qui en un certain sens les généralisent (NEEDHAM, 1961 ; JARDINE & SIBSON, 1968). D'autre part si une partition de E possède cette propriété, elle est unique.

Considérons l'exemple suivant d'une matrice de ressemblance entre 3 objets, le coefficient étant supposé par définition compris entre 0 et 1.

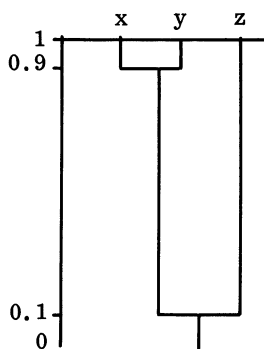
	x	y
y	0.9	
z	0.1	0.1

Les objets x et y se ressemblent "beaucoup" alors que z leur ressemble "très peu". Il semble donc que la partition (x, y) ; z vérifie la définition 1. Ce raisonnement intuitif doit être nuancé ; si nous sommes très exigeants, x et y peuvent être considérés comme peu ressemblants et la partition x ; y ;

z est la seule en accord avec la définition. Il en est de même pour la partition (x, y, z) si nous sommes très peu exigeants.

La solution au problème posé par la définition n'est pas une partition de E mais un ensemble de partitions, correspondant à divers choix d'un seuil de ressemblance S a priori arbitraire.

Si pour toute valeur possible de S , il est possible de définir la partition de E qui vérifie la définition 1, l'ensemble des partitions obtenues constitue une hiérarchie. En effet si les objets x et y sont tels que $S(x, y) \geq S$, (nous dirons que les objets x et y se ressemblent au seuil S), d'après la définition 1, x et y appartiennent à un même sous-groupe de la partition correspondant au seuil S . Au seuil $S' < S$ x et y se ressemblent également, ils appartiennent donc encore à un même sous-groupe dans la partition correspondant au seuil S' . Tout sous-groupe de la partition $S' < S$ est donc la réunion de sous-groupes de la partition S . L'ensemble des partitions forme une hiérarchie représentée classiquement sous la forme suivante :



Un raisonnement identique pourrait être fait à partir de distances entre les objets. Réciproquement la donnée de cette hiérarchie est équivalente à la matrice de ressemblance, le niveau de regroupement ou de fusion de deux objets est égal à leur ressemblance.

Nous dirons que la hiérarchie réalise la classification de l'ensemble xyz . Il est important de mettre en évidence deux notions complémentaires dans cette classification :

- La hiérarchie de partitions ne dépend que de l'ordonnance sur E ; en effet la matrice de ressemblance suivante conduit à la même hiérarchie quels que soient a et b :

	x	y	
y	a		
z	b	b	

avec $1 < a < b < 0$

- Il est bien clair cependant que la seule hiérarchie n'exprime pas

toute l'information, le niveau de chaque fusion permet d'opposer la forte ressemblance entre x et y à la ressemblance faible de z avec x et y.

Prenons un autre exemple de matrice de ressemblance :

	x	y	z
y	0.9		
z	0.8	0.4	
t	0.2	0.6	0.7

Essayons de rechercher pour toute valeur de S, la partition vérifiant la définition 1

$1 \geq S > 0.9$	x ; y ; z ; t
$0.9 \geq S > 0.8$	(x, y) ; z ; t
$0.8 \geq S > 0.2$	aucune partition
$0.2 \geq S > 0$	(x, y, z, t)

La définition 1, sur laquelle un accord pouvait être établi, ne peut être utilisée ; cela implique que nous ne pouvons trouver une hiérarchie strictement équivalente à la matrice de ressemblance.

Un résultat classique fournit une condition nécessaire et suffisante pour qu'une matrice de ressemblance soit exactement représentable par une hiérarchie :

Quel que soit x, y, z $S(x, y) \geq \min(S(x, z), S(y, z))$

Si les objets peuvent être représentés dans un espace métrique, cette inégalité signifie que tout triangle est isocèle à base inférieure ou égal aux côtés égaux et que nous pouvons écrire :

$D(x, y) \geq \max(D(x, z), D(z, y))$ inégalité de l'ultramétrie

La définition 1 de la classification impose une condition stricte pour que deux objets soient réunis, mais aussi pour qu'ils demeurent séparés. Les définitions 2 et 3 qui suivent, proposent une condition moins stricte pour l'une ou l'autre des possibilités.

2.2 - Pour tout seuil de vraisemblance S, pour tout couple d'objets x, y, s'il existe au moins un sous-ensemble $x_1, \dots, x_r$ d'objets de E tel que $x \in x_1, x_1 \in x_2, \dots, x_r \in y$ où $x_\alpha \in x_\beta$ signifie que $S(x_\alpha, x_\beta) \geq S$, alors x et y appartiennent à un même groupe. Dans le cas contraire x et y n'appartiennent pas à un même groupe. (Définition 2)

Dans cette définition, la condition stricte entre objets de groupes différents est conservée car leur ressemblance ne peut être supérieure à S,

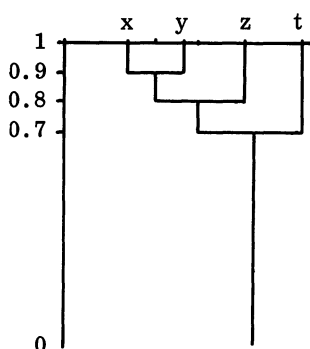
cependant la condition concernant les objets d'un même groupe est remplacée par une notion de ressemblance "transitive" évidemment plus faible.

Si nous établissons la liste des couples x, y tels que $x R y$, pour un seuil S donné, la construction d'une partition obtenue est unique.

Par exemple pour $S = 0.7$: (xRy) la partition cherchée
 (xRz) est (x, y, z, t)
 (zRt)

Il est aisé de montrer que l'ensemble des partitions forme une hiérarchie, car si $x R y$ au seuil S , $x R y$ est vrai pour tout seuil $S' < S$

La classification obtenue est la suivante :



D'après les remarques faites plus haut, cette hiérarchie est équivalente à une matrice de ressemblance ultramétrique qui bien entendu n'est pas identique à la matrice de départ. La hiérarchie obtenue ne dépend que de l'ordonnance définie sur E et ce fait est important.

- Mathématiquement, la classification donnée par la définition 2 peut-être introduite comme étant optimale pour certains critères :

Ainsi pour ROUX (1967), la distance ultramétrique équivalente à la classification est la borne supérieure des distances ultramétriques inférieures à la distance définie dans la matrice originale (au sens d'un certain ordre partiel sur les distances). Pour LERMAN (1970) la classification est optimale au sens du critère dit lexicographique défini seulement à partir de l'ordonnance sur E .

- Sur le plan pratique, cette méthode de classification est connue sous le nom de méthode "single linkage" (SNEATH, 1957). Elle est en règle générale non satisfaisante car elle entraîne le regroupement d'objets souvent fort dissemblables ; une ressemblance ne pouvant être supposée transitive. Les conséquences sont de deux types :

a) Les diverses partitions de E obtenues lorsque le seuil S de ressemblance diminue sont souvent créées par fusion successive d'objets dans le même sous-groupe (chaînes). En effet, par construction un objet a d'autant plus de chance de s'unir à un groupe que celui-ci est de grande

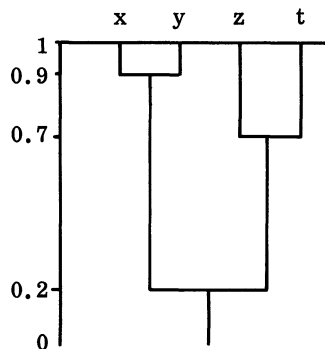
taille. Certaines méthodes heuristiques (WISHART, 1969) se proposent d'éviter la création de chaînes au cours d'une classification "single linkage".

b) Les niveaux de fusion représentent très mal la matrice de ressemblance originale et le résultat proposé par ROUX montre que les ressemblances entre objets sont systématiquement trop élevées lorsqu'elles sont estimées sur la hiérarchie. Il serait éventuellement possible, à partir de la hiérarchie obtenue par la seule ordonnance sur E, de placer les niveaux des diverses fusions de manière à rendre minimum une distance à définir entre la matrice originale et la matrice ultramétrique. La définition 2 de la classification n'aurait évidemment plus de sens précis dans ce cas.

2.3 - Pour un seuil de ressemblance S et pour tout couple d'objets (x, y) appartenant à un même groupe, $S(x, y) > S$ et dans chaque groupe il existe au moins un objet qui ne ressemble pas à au moins un objet de tout autre groupe (Définition 3)

Dans l'exemple étudié plus haut, pour $S = 0.4$, deux partitions vérifient cette définition : $(x, y, z) ; t$ et $x ; (y, z, t)$.

La définition 3 n'assure pas l'unicité de la classification (JARDINE & SIBSON, 1968). Pour obtenir une classification unique, il est nécessaire d'ajouter la condition que si deux objets sont réunis dans un même groupe au seuil S , ils le sont encore pour tout seuil $S' < S$, c'est-à-dire que l'ensemble des partitions forme une hiérarchie. La méthode de classification est appelée "méthode complète linkage" dans l'ouvrage de SOKAL & SNEATH, elle ne possède pas les propriétés d'optimalité de la méthode précédente. Les différentes partitions obtenues en fonction de S ne dépendent là aussi que de l'ordonnance définie sur E.



Pour la raison inverse de la précédente (2.a)), cette méthode empêche la création de chaînes et met en évidence les couples d'objets très ressemblants plutôt que les sous-ensembles réellement homogènes. D'autre part la matrice de ressemblance équivalente à la hiérarchie est très différente de la matrice originale, les niveaux de fusion sont en effet systématiquement inférieurs aux ressemblances. Une correction des niveaux de fusion identique à celle proposée pour la méthode précédente peut être envisagée.

2.4 - La construction des hiérarchies

Les paragraphes précédents ont montré l'importance d'une hiérarchie de partitions pour représenter une classification d'objets. Pour les définitions 1 et 2, cette représentation est imposée par la notion de seuil de ressemblance, pour la définition 3 elle permet d'obtenir une classification unique. Dans ce chapitre nous conserverons cette représentation, qui, rappelons-le, exige que des objets réunis pour un seuil S de ressemblance, le soient encore pour une définition moins exigeante de la ressemblance. Un algorithme simple permet de généraliser la construction des hiérarchies sur E . Supposons qu'il soit possible de calculer un coefficient de ressemblance (ou une distance) entre deux groupes disjoints quelconques d'objets de E .

Soit une partition $G_1 \dots G_s$ de E et G_p et G_q les deux groupes dont la ressemblance S_{pq} est maximum. Au seuil de ressemblance $S = S_{pq}$ les deux groupes G_p et G_q sont fusionnés en un nouveau groupe $G_r = G_p \cup G_q$. La ressemblance de tout autre groupe G_i de la partition avec G_r est calculée, soit S_{ir} . Si, quel que soit i , $S_{ir} \leq S_{pq}$, le même processus peut être répété sur la partition $G_1 \dots, G_r = G_p \cup G_q, \dots G_s$, une nouvelle fusion étant acquise à un seuil de ressemblance $S' < S$.

En général, S_{ir} est une fonction quelconque des ressemblances entre les objets des trois groupes G_i , G_p et G_q . LANCE & WILLIAMS (1967) énoncent certaines propriétés souhaitables d'un algorithme de classification hiérarchique en plus de la condition suffisante d'existence définie plus haut ($S_{ir} \leq S_{pq}$ quel que soit i)

1/ La ressemblance entre groupes doit être de même "nature" que la ressemblance entre objets.

2/ La matrice ultramétrique reconstruite à partir de la hiérarchie doit être aussi "voisine" que possible de la matrice de ressemblance.

3/ D'autre part si $S_{ir} = f(S_{ip}, S_{iq}, S_{pq})$ une méthode très simple de calcul permet d'obtenir les ressemblances après chaque fusion.

LANCE & WILLIAMS proposent la classe suivante de fonctions f permettant de construire la plupart des hiérarchies décrites dans la littérature :

$$S_{ir} = \alpha S_{ip} + \beta S_{iq} + \gamma S_{pq} + \delta |S_{ip} - S_{iq}| \quad \alpha, \beta > 0$$

(la relation est identique pour les distances)

La propriété (3) est vérifiée ; la relation linéaire assure sinon une identité de "nature" entre ressemblances de groupes et d'objets, une homogénéité de dimension entre ces coefficients (propriété (1)).

Les principales méthodes sont passées en revue :

a) "Single linkage" $\alpha = \beta = 1/2 \quad \gamma = 0 \quad \delta = 1/2$

b) "Complete linkage" $\alpha = \beta = 1/2 \quad \gamma = 0 \quad \delta = -1/2$

Rappelons que pour ces deux méthodes la condition d'existence est vérifiée mais en aucune façon la condition (2).

c) "Average linkage" (SOKAL & MICHENER, 1958)

La ressemblance entre deux groupes est la moyenne de ressemblances entre objets de chaque groupe, ce qui implique :

$$\alpha = \frac{k_p}{k_p + k_q} \quad \beta = \frac{k_q}{k_p + k_q} \quad \gamma = \delta = 0$$

où k_p et k_q sont respectivement les effectifs des groupes G_p et G_q .

Remarquons que $\alpha + \beta + \gamma = 1$ et montrons que cette relation est suffisante pour que la condition d'existence soit vérifiée : en effet $S_{pq} \geq S_{ip}$ et $S_{pq} \geq S_{iq}$ car G_p et G_q sont fusionnés au seuil $S = S_{pq}$

donc $S_{ir} \leq \alpha S_{pq} + \beta S_{pq} + \gamma S_{pq} = S_{pq}$ quel que soit i .

Dans le cas particulier où de plus $\gamma = 0$, si une distance entre objets a été choisie, l'écart entre groupes ainsi défini est encore une distance (propriété (1)).

En effet si $G_r = G_p \cup G_q$, et G_i et G_j deux autres groupes :

$$D_{ir} + D_{jr} = \alpha D_{ip} + \beta D_{iq} + \alpha D_{jp} + \beta D_{jq}$$

$$D_{ir} + D_{jr} = \alpha (D_{ip} + D_{jp}) + \beta (D_{iq} + D_{jq})$$

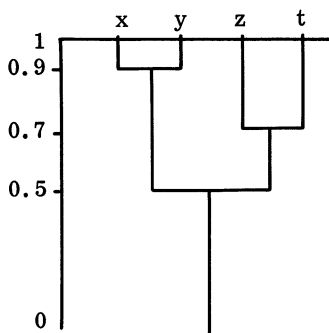
$$D_{ir} + D_{jr} \geq \alpha D_{ij} + \beta D_{ij} = D_{ij}$$

L'inégalité du triangle est vérifiée.

D'autre part une certaine formulation de la propriété (2) est vérifiée par cette méthode. Si à une certaine étape les deux groupes G_p et G_q sont fusionnés au seuil S , la ressemblance ultra métrique entre les objets de G_p et de G_q est égale à S . Si pour ces objets nous définissons "l'écart" entre

les coefficients S_{ij} de départ et la ressemblance S par $\sum_{\substack{i \in G_p \\ j \in G_q}} (S_{ij} - S)^2$.

Cette quantité est précisément minimum pour $S = \frac{\sum S_{ij}}{k_p k_q}$, ce que réalise l'algorithme. Contrairement aux méthodes a) et b), la hiérarchie de partitions obtenues ne dépend pas seulement de l'ordonnance sur E . Dans l'exemple précédent la classification suivante est obtenue :



d) Méthode centróide (SOKAL & MICHENER, 1958)

L'algorithme est établi dans le cas de distances euclidiennes, chaque objet étant représenté par un point de R^p . La distance de deux groupes est choisie comme la distance de leurs centróides, ce qui conduit à écrire

$$\alpha = \frac{k_p}{k_p + k_q} \quad \beta = \frac{k_q}{k_p + k_q} \quad \gamma = \frac{-k_p k_q}{(k_p + k_q)^2} \quad \delta = 0$$

La condition d'existence de l'algorithme n'étant pas toujours remplie même dans le cas particulier de distances euclidiennes, nous ne pouvons proposer cette méthode. Il en est de même si la distance entre G_1 et $G_r = G_p \cup G_q$ est choisie comme la distance entre le centróide de G_1 et le point milieu du segment joignant les centróides de G_p et G_q , algorithme pour lequel $\alpha = \beta = 1/2$, $\gamma = -1/4$ $\delta = 0$ (GOWER, 1967).

e) Méthode de WARD (1963)

WARD mesure la perte "d'information" entraînée par la fusion de deux groupes G_p et G_q par la différence entre la somme des carrés des distances de chaque point au centróide du groupe $G_r = G_p \cup G_q$ et la somme des carrés de distances de chaque point au centróide du groupe auquel il appartient avant la fusion. Cette dernière concerne alors les groupes pour lesquels cette perte d'information est minimum.

WISHART (1969) a montré que cet algorithme est un cas particulier de celui étudié ici avec :

$$\alpha = \frac{k_p + k_1}{k_1 + k_p + k_q} \quad \beta = \frac{k_q + k_1}{k_1 + k_p + k_q} \quad \gamma = \frac{-k_1}{k_p + k_q + k_1}$$

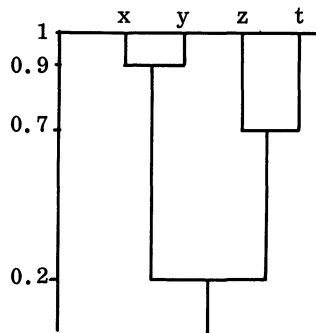
La condition d'existence est toujours vérifiée car $\alpha + \beta + \gamma = 1$

Il faut remarquer que :

$$D_{1r} = \frac{k_1 (k_p + k_q)}{k_1 + k_p + k_q} \quad d_{1r}^2 \quad \text{où} \quad d_{1r}^2$$

est le carré de la distance euclidienne entre les centróides de G_1 et de G_r , en conséquence D_{1r} ne peut plus être considéré comme le carré d'une distance euclidienne.

Exemple :



f) Méthode "flexible" (LANCE & WILLIAMS, 1967)

Elle dépend d'un paramètre a : $\alpha = \beta = \frac{1-a}{2}$ $\gamma = a$.

La condition d'existence est vérifiée car $\alpha + \beta + \gamma = 1$. Les auteurs proposent pour des raisons empiriques de choisir une valeur de a faiblement négative. Montrons que l'écart entre groupes est encore une distance, si une distance entre objets est choisie au départ et si a est négatif supérieur à -1 . [propriété (1)]

$$D_{ir} + D_{jr} = \frac{1-a}{2} (D_{ip} + D_{iq} + D_{jp} + D_{jq}) + 2a D_{pq}$$

Or $D_{pq} \leq D_{ip}$, $D_{pq} \leq D_{jp}$ car G_p et G_q sont regroupés au seuil $S = D_{pq}$ donc si a est négatif :

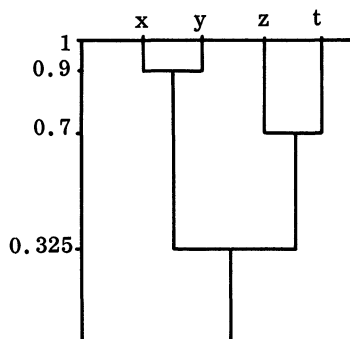
$$D_{ir} + D_{jr} \geq \frac{1-a}{2} (D_{ip} + D_{iq} + D_{jp} + D_{jq}) + a (D_{ip} + D_{jp})$$

$$D_{ir} + D_{jr} \geq \frac{1+a}{2} (D_{ip} + D_{jp}) + \frac{1-a}{2} (D_{jp} + D_{iq})$$

Si $a > -1$ alors $\frac{1+a}{2} > 0$

$$D_{ir} + D_{jr} \geq \frac{1+a}{2} D_{ij} + \frac{1-a}{2} D_{ij} = D_{ij}$$

Exemple avec $a = -0.25$, valeur suggérée par les auteurs.



2.5 - Le problème de la représentation optimale d'une matrice de ressemblance par une hiérarchie (HARTIGAN, 1967)

HARTIGAN s'est proposé de rechercher la hiérarchie la plus proche en un certain sens de la matrice de ressemblance, indépendamment de la notion de seuil de ressemblance. Il a étudié le critère suivant :

$\rho (S_1, S_2) = \sum_{i,j} W_{ij} [S_1(i, j) - S_2(i, j)]^2$ où $S_1(i, j)$ est la ressemblance donnée entre les objets i et j , $S_2(i, j)$ la ressemblance ultramétrique définie par une hiérarchie et W_{ij} un coefficient de pondération. La solution

théorique de ce problème indique que lorsque deux groupes G_p et G_q fusionnent dans la hiérarchie le niveau de fusion est la moyenne pondérée des ressemblances entre objets des deux groupes soit :

$$\sum_{\substack{i \in G_p \\ j \in G_q}} \frac{W_{ij} S_i(i, j)}{\sum W_{ij}} \quad (1)$$

Cependant un algorithme permettant de construire les divers groupes supposerait le calcul du critère ρ pour toutes les hiérarchies satisfaisant (1) et le choix de celle qui offre un minimum absolu. Il faut remarquer que la méthode "average linkage" satisfait la condition (1) avec $W_{ij} = 1$ mais n'est qu'une solution possible au problème posé. Le choix ne pouvant guère être effectué pratiquement, HARTIGAN cherche à mettre en évidence un algorithme permettant d'obtenir des minima "locaux" du critère ρ dont on peut penser qu'ils donnent des solutions voisines de la solution théorique (solutions sub-optimales).

En conclusion, si nous désirons utiliser une méthode de classification hiérarchique simple, la méthode "average linkage" nous paraît réunir un certain nombre de propriétés intéressantes, ainsi que le remarquent LANCE & WILLIAMS (1967) elle mérite d'être plus souvent étudiée. Si une représentation euclidienne des objets est choisie, la méthode de WARD possède des avantages théoriques certains.

La relation de LANCE & WILLIAMS permet cependant de programmer facilement sur un ordinateur l'ensemble des méthodes citées au chapitre 2.1 ; ces dernières peuvent ainsi être comparées pour chaque application (WISHART, 1968).

3. - LES METHODES NON HIERARCHIQUES DE CLASSIFICATION

Les méthodes hiérarchiques introduisent une contrainte sur la notion de classification que nous avons rappelée au début du chapitre 2.1 et qui peut ne pas être imposée par la nature du problème. L'avantage réside essentiellement dans le fait qu'elles facilitent l'interprétation d'une classification. Nous nous intéressons dans ce chapitre à la recherche d'une suite de partitions (mais non obligatoirement hiérarchique) ou d'une partition possédant des propriétés optimales interprétables en termes de fonction de décision. Supposons que sur l'ensemble P des partitions de E soit définie une fonction numérique exprimant la "valeur" d'une partition. Par exemple, si $f(P_1) > f(P_2)$ alors la partition P_1 sera dite préférable à la partition P_2 .

Le problème de trouver la partition optimale (ou la suite de partitions optimales si f dépend d'un paramètre) est théoriquement résolu : il suffit de rechercher la partition P telle que $f(P)$ soit maximum.

Le calcul du nombre de partitions possibles (DAGNELIE, 1966) montre que dès que n est supérieur à quelques unités, une telle recherche est impraticable. Le choix d'une fonction f ne suffit pas à résoudre pratiquement le problème de classification ainsi défini.

3.1 - La "valeur" d'une partition

RUBIN (1967) a formulé un ensemble d'axiomes qu'une fonction $f(P)$ doit vérifier, ensemble d'ailleurs insuffisant pour permettre l'édification d'une fonction f unique. La caractéristique essentielle d'une recherche de classification au sens de RUBIN est qu'une partition de E n'a de sens que pour une certaine exigence quant à la ressemblance des objets, c'est-à-dire pour le choix d'un seuil particulier de ressemblance S . La classification s'exprime alors par une suite de partitions optimales.

Essayons de construire "naturellement" quelques fonctions f .

- Pour une valeur donnée de S , si $S_{ij} > S$ il est souhaitable que les objets i et j soient regroupés et ceci d'autant plus que la quantité $S_{ij} - S$ est grande, de même si $S_{ij} < S$ il est souhaitable qu'ils soient séparés (cf Définition 1). La quantité $S_{ij} - S$ peut être considérée comme le gain de réunir des objets vérifiant $S_{ij} > S$ ou de séparer des objets vérifiant $S_{ij} < S$.

Le gain total associé au choix d'une partition P s'écrit alors :

$$f(P) = \sum_{i,j \in E} g_{ij} (S_{ij} - S)$$

où $g_{ij} = 1$ si i et j appartiennent à un même groupe et
- 1 dans le cas contraire.

Il est aisé de montrer que la maximisation de f revient à rendre maximum $(P) = \sum_{ij} (S_{ij} - S)$ où la sommation est étendue aux seuls couples

d'objets regroupés. Nous pouvons remarquer que si la recherche des partitions optimales est effectuée séquentiellement pour les valeurs décroissantes de S en imposant à chaque étape la condition que l'ensemble des partitions déjà obtenues forment une hiérarchie, rendre $f(P)$ maximum au seuil S revient à fusionner les deux groupes précédemment construits tels que la moyenne des ressemblances des objets pris dans chacun d'eux soit maximum.

La méthode "average linkage" définie précédemment est optimale pour $f(P)$ pour l'ensemble des hiérarchie sur E .

- Il est possible d'envisager une fonction de décision indépendante de la valeur absolue des quantités $S_{ij} - S$, c'est-à-dire d'écrire :

$$f(P) = \sum_{i,j \in E} g_{ij} \operatorname{sgn} (S_{ij} - S)$$

où g_{ij} a le même sens que précédemment.

Dans ce cas, la suite des partitions obtenues ne dépend que de l'ordonnance définie sur E par la matrice de ressemblance.

Si nous imposons une construction séquentielle de la suite de partitions optimales de manière à ce qu'elle soit une hiérarchie nous obtenons une méthode analogue à la méthode "average linkage" pour laquelle la res-

semblance entre deux groupes est la médiane des ressemblances des objets pris dans chacun d'eux. Il s'agit d'un exemple de méthode hiérarchique ne satisfaisant pas à la propriété (3) de LANCE ET WILLIAMS.

En abandonnant les axiomes proposés par RUBIN, il est possible de rechercher la partition P optimale indépendamment de tout seuil de ressemblance.

Les exemples précédents nous suggèrent deux critères, le second n'ayant à notre connaissance jamais été proposé.

- Considérons deux couples d'objets ($i \neq j$ et $k \neq l$) de E

Posons deux conditions relatives aux deux couples d'objets :

1/ Les objets i et j sont regroupés dans P
Les objets k et l sont séparés dans P

2/ $S_{ij} \geq S_{kl}$

Le "gain" introduit par les deux couples d'objets dans la définition de la valeur de P est le suivant :

- si les conditions (1) et (2) sont vérifiées $g(i, j; k, l) = 1$
- si la condition (1) est vérifiée mais non la condition (2) $g(i, j; k, l) = -1$
- si la condition (1) est non vérifiée $g(i, j; k, l) = 0$

La fonction f (P) est alors définie en posant pour simplifier $g_{\alpha\beta} = g(i, j; k, l)$

$$f(P) = \sum_{1 \leq \alpha < \beta}^{n(n-1)/2} g_{\alpha\beta}$$

Cette fonction est identique au critère de LA VEGA décrit et étudié dans (LERMAN, 1970) et ne dépend que de l'ordonnance sur E.

- Une seconde fonction f dérive tout naturellement de la précédente :

$$f(P) = \sum_{1 \leq \alpha < \beta}^{n(n-1)/2} \epsilon_{\alpha\beta} (S_{\alpha} - S_{\beta})$$

où $\epsilon_{\alpha\beta} = 0$ si la α ème paire et la β ème paire d'objets sont toutes deux regroupées ou séparées.

$\epsilon_{\alpha\beta} = 1$ si la α ème paire est regroupée et la β ème séparée.

$\epsilon_{\alpha\beta} = -1$ dans l'autre cas.

et où S_{α} , S_{β} sont respectivement la ressemblance des objets de la α ème et de la β ème paire.

Sans doute plus efficace que le critère précédent, il ne possède pas ses propriétés de stabilité par rapport à l'ordonnance sur E.

3.2 - Algorithmes de recherche de solutions suboptimales

DAGNELIE (1966) remarque que si le nombre de partitions possibles d'un ensemble E est très grand, toutes ne sont pas admissibles si les objets sont représentés dans R^p .

Les partitions admissibles ne peuvent être pratiquement définies que pour p petit (≤ 3). La réduction du nombre de variables par une analyse en composantes principales peut permettre de rechercher plus facilement par la suite les solutions suboptimales admissibles.

Les algorithmes utilisés par les différents auteurs (REGNIER 1965, RUBIN 1967) reposent sur la recherche de "maxima locaux". A partir d'une partition donnée, les objets sont successivement échangés entre les divers groupes, ou constituent de nouveaux groupes et la fonction $f(P)$ est calculée chaque fois. Dès que $f(P)$ est supérieure à la valeur précédemment calculée, la partition P est conservée jusqu'à ce qu'aucun échange d'objets n'augmente f . Il est alors possible soit de reprendre l'algorithme avec de nouvelles partitions de départ soit d'utiliser des méthodes heuristiques pour tenter de dépasser le maximum local (RUBIN).

4. - LES METHODES DE SEGMENTATION

De même qu'il a été possible (voir 2.1) de définir une perte d'information résultant de la fusion de deux groupes, un gain d'information résultant de la division d'un groupe en deux sous groupes peut être établi. A chaque étape, chaque sous groupe existant peut être divisé de manière que le gain d'information soit maximum.

Si le problème de la non unicité de la définition d'un gain d'information entraîné par une partition est le même que pour les méthodes hiérarchiques, un problème supplémentaire se pose dans ce cas : comment obtenir la partition optimale parmi les $2^{n-1} - 1$ partitions possibles d'un ensemble de taille n en deux sous groupes ?

Deux cas particuliers sont étudiés par EDWARDS & CAVALLI-SFORZA (1965) puis par GOWER (1967) lorsque les objets sont des points de R^p . Le gain d'information s'exprime soit par la distance D_{pq}^2 des centroïdes des groupes G_p et G_q à créer soit par une quantité analogue à celle utilisée par WARD. GOWER démontre certaines conditions nécessaires devant être vérifiées par un algorithme permettant d'obtenir la division optimale d'un groupe, mais elles ne sont pas suffisantes.

Deux algorithmes heuristiques simples ont été proposés par GOWER (1967) et par MAC MANGHTON-SMITH (1964) ; bien que raisonnables, ils ne prétendent pas fournir une solution même suboptimale au problème de la segmentation. Les algorithmes d'échange cités plus haut sont vraisemblablement plus efficaces.

CONCLUSION

Les méthodes de segmentation présentent l'inconvénient majeur d'effectuer par principe des partitions en deux sous-groupes. Elles peuvent ainsi

ne pas mettre en évidence certains "groupements" (clusters) présents dans l'ensemble E s'ils sont divisés dès les premières segmentations.

Dans le choix entre les méthodes hiérarchiques et non hiérarchiques, un avantage certain des premières tient à la simplicité des calculs qu'elles nécessitent.

Cependant dans le cas de l'exemple 3, la recherche d'une partition optimale s'impose et une fonction $f(P)$ doit être établie au mieux (problème de décision).

La classification d'un ensemble d'espèces végétales ne peut être représentée complètement par une partition car les relations dites "phylogénétiques" entre espèces doivent être mises en évidence et seule une hiérarchie ou une classification par niveaux au sens de RUBIN permet de le faire dans le cadre des méthodes étudiées ici.

Le choix est moins clair dans le cas de l'exemple 1 car la question posée paraît pouvoir être abordée sous l'angle décisionnel mais aussi en recherchant d'éventuels "niveaux de ressemblance" entre sujets. La formulation précise d'un test de l'hypothèse d'existence de plusieurs populations dans l'ensemble E nous paraît représenter le problème fondamental.

L'établissement d'une théorie de la classification considérée comme branche de la statistique mathématique annoncée par BOLSHEV (1970) permettra sans doute de mieux définir les méthodes devant être appliquées. Une telle discipline est appelée à recevoir d'importants développements.

REFERENCES

- BOLSHEV L.N. - Cluster analysis. Bull. I.S.I. 43, 411-425, 1969.
- BONNER R.E. - On some clustering techniques. IBM Jour. Res. Devel. 8, 22-32, 1964.
- DAGNELIE P. - A propos de différentes méthodes de classification numérique. R.S.A. 14, 55-75, 1966.
- EDWARDS A.W.F., CAVALLI-SFORZA L.L. - A method for cluster analysis. Biometrics 21, 890-907, 1965.
- FRASER P.M., BARON D.N. - Taxonomic procedures applied to liver disease. Proc. Roy. Soc. Med. 61, 23-26, 1968.
- GOWER J.C. - Classification and geology. Review of the I.S.I. 38, 35-40, 1970.
- GOWER J.C. - A comparison of some methods of cluster analysis. Biometrics 23, 623-637, 1967.
- HARTIGAN J.A. - Representation of similarity matrices by trees. J.A.S.A. 62, 1140-1158, 1967.
- JARDINE N., SIBSON R. - The construction of hierarchic and non hierarchic classifications. Comp. J. 11, 177-184, 1968.

- KRUSKAL W.H., GOODMAN L.A. - Measures of association for cross classifications. J.A.S.A. 49, 734-764, 1954.
- LANCE G.N., WILLIAMS W.T. - A general theory of classificatory sorting strategies. I. Hierarchical systems. Comp. J. 9, 373-380, 1967.
- LERMAN I.C. - Les bases de la classification automatique. Gauthier-Villars Paris. Coll. Programmation. 1970, 117 p.
- MAGNAUGHTON-SMITH P., WILLIAMS W.T., DALE M.B., MOCKETT L.G. - Dissimilarity analysis : a new technique for hierarchical subdivisions. Nature 202, 1034-1035, 1964.
- NEEDHAM R.M., JONES K.S. - Keywords and clumps. J. Docum. 20, 5, 1964.
- REGNIER S. - Sur quelques aspects mathématiques des problèmes de la classification automatique. I.C.C. Bull. 4, 175-191, 1965.
- ROUX M. - Algorithme pour construire une hiérarchie particulière. Thèse de 3ème Cycle Université de Paris. 1968.
- RUBIN J. - Optimal classification into groups : an approach for solving the taxonomy problem. J. Theoret. Biol. 15, 103-144, 1967.
- SIMPSON G.G. - Principles of animal taxonomy. Columbia University Press, New-York. 247 p.
- SOKAL R.R. ; MICHENER D. - A statistical method for evaluating systematic relationships. Univ. Kansas Sci. Bull. 38, 1409-1438, 1968.
- SOKAL R.P., SNETH P.H.A. - Principles of numerical taxonomy. W. H. Freeman and C° San Francisco and London 1963.
- WARD J.H. - Hierarchical grouping to optimize an objective function. J.A. S.A. 58, 236-244, 1963.
- WILLIAMS W.T., LAMBERT J.M. - Multivariate methods in plant ecology I. J. Ecol. 47, 83, 1959.
- WISHART D. - A FORTRAN II program for numerical classification. Computing Laboratory University of St Andrews, Scotland. 1968.
- WISHART D. - Numerical classification method for deriving natural classes. Nature. 221, 97-98, 1969.
- WISHART D. - An algorithm for hierarchical classifications. Biometrics. 22, 165-170, 1969.