REVUE DE STATISTIQUE APPLIQUÉE

J. P. BENZÉCRI

Problèmes et méthodes de la taxinomie

Revue de statistique appliquée, tome 18, nº 4 (1970), p. 73-98

<http://www.numdam.org/item?id=RSA_1970__18_4_73_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/

PROBLÈMES ET MÉTHODES DE LA TAXINOMIE

J. P. BENZÉCRI

Professeur de Statistique à la Faculté de Sciences de Paris

La systématique du règne animal et du règne végétal est sans doute la plus vaste calssification que l'homme ait édifiée. Depuis le siècle de Linné, ce grand arbre n'a cessé de se ramifier et si la prudence ne permet pas souvent d'en rotailler ou d'en pormuter les branches, bien des naturalistes se gardent de prendre les nécessaires conventions de leur science pour l'ordre même de la nature. A la théorie générale des classifications, ces savants peuvent apporter des principes et des exemples. Dans un remarquable ouvrage, R. Sokal et P. Sneath (cités dans la suite : S & S) exposent les "principles of numerical taxonomy" c'est-à-dire l'introduction des méthodes quantitatives dans "the theoritical study of classification, including its bases, principles, procedures and rules (Simpson 1961, cité d'après S & S page 3).

Nous ne ferons guère plus ici que de présenter rapidement ce livre, en le commentant mathématiquement.

I - LES PRINCIPES DE LA TAXINOMIE

Dans les sciences de la nature, les classifications servent à des fins si diverses que la taxinomie doit satisfaire à des exigences quasi-contradictoires. C'est dans ce contexte que nous essaierons de situer l'oeuvre de S & S.

D'une part la taxinomie est la science du système des vivants disposant les individus en une hiérarchie de classes elle cherche à décrire l'équilibre des espèces, leurs ressemblances et aussi leurs parentées. Son objet le plus élevé est (quoique disent S & S qui en sont à la logique d'Aristote) de connaitre l'essence de la vie, c'est-à-dire, en toute généralité, et par delà les différences d'être à être, "ce que c'est que d'être "vivant ; et par quels choix successifs se spécifient (se définissent) les formes individuelles qui s'offrent d'abord seules à l'étude. Pour atteindre ce point de vue global la science doit s'élever du particulier au général, procéder par groupements successifs pour découvrir les grandes lignes derrière les détails.

D'autre part la taxinomie vise, dans la pratique quotidienne à déterminer les individus, c'est-à-dire à en reconnaitre le nom, (la place, "l'adresse", dans la classification).

Ceci peut avoir un intérêt immédiat : retrouver l'information, savoir qu'une plante qu'on vient de ramasser appartient à telle espèce bien étudiée, soit comestible, soit vénéneuse.

Et aussi un intérêt général par la connaissance du système des vivants : il est utile de faire l'inventaire de la faune et de la flore de zones en équilibre, et c'est l'objet de l'écologie que de classer (nouveau problème de taxinomie) les groupements d'espèces en biocénose. Dans la détermination le mouvement est du général au particulier : on considère d'abord un poisson puis on le reconnait aurata-aurata.

Sokal et Sneath ont pour premier intérêt la conception des classifications, non leur emploi quotidien. Là où l'intuition, les habitudes, les conjectures mises au rang de principe, forment avec les faits positifs un tout inanalysable, ils veulent introduire des méthodes si complètement formalisées, qu'on les puisse traduire en programmes pour les machines électroniques (cf. S & S page 270). D'où la vivacité de leurs critiques que nous présenterons d'abord avant leurs méthodes.

Quoiqu'il ne nous appartienne pas d'arbitrer entre partisans et adversaires de la taxinomie numérique, remarquons que celle-ci s'est montrée féconde dans les domaines nouvellement explorés tels que l'écologie ; là où le temps manque pour familiariser l'homme avec son objet, et développer l'intuition, une synthèse automatique peut être précieuse.

I.1 Groupements monothétiques et groupements polythétiques.

Pour diviser un règne en groupes de plus en plus restreints (appelés embranchement, classe, ordre, famille, tribu, genre, espèce...) on procède souvent par dichotomies successives : un groupement est divisé en les individus qui possèdent telle propriété, et ceux qui ne la possède pas. Pour désigner cette distribution suivant un seul critère, S & S (loc. cit p. 13) ont introduit l'adjectif monothétique. Cependant Vicq d'Azyr (cité par S & S ; (1992)) remarquait déjà qu'il serait possible qu'une classe fût très-naturelle et qu'il n'y eût pas un seul caractère commun à toutes les espèces qui le composent ce que Beckner précise ainsi (1959), p. 22, d'après S & S) : "A class is ordinarily defined by reference to a set of properties which are both necessary and sufficient (by stipulation) for membership in the class. It is possible, however, to define a group K in term of a set G of properties f_1, f_2,... f_n in a different manner. Suppose we have an aggregation of individuals (we shall not as yet call them a class) such that :

1/ Each one possesses a large (but unspecified) number of the properties in G.

2/ Each f in G is possessed by large numbers of these individuals and

3/ No f is possessed by every individual in the aggregate..."

les trois conditions définissent une classe proprement polythétique (cf aussi en botanique la notion classique de famille par enchainement, e.g. les rosacées.)

I.2 Classification ascendante et classification descendante

A la méthode monothétique des dichotomies successives qui descend du règne entier jusqu'aux espèces, S & S opposent la démarche polythétique

et ascendante d'Adanson (1757, p xi, d'après S & S) : "Je me contenterai de rapprocher les objets suivant le plus grand nombre des degrés de leurs rapports et de leurs ressemblances... Les objets ainsi réunis formeront plusieurs petites familles que je réunirai encore ensemble, afin d'en faire un tout dont les parties soient unies et liées intimement"). En fait une telle synthèse est le préliminaire indispensable de toute classification : mais la synthèse une fois faite, on peut être assez heureux pour découvrir qu'un petit nombre de caractères suffisent à définir monothétiquement les classes que la considération d'un vaste ensemble de caractères à seule permis de découvrir (S & S p. 17 et p. 275). La reconnaissance d'un individu est tout autre chose que la découverte d'une classification puis d'une méthode de reconnaissance. Il est regrettable que l'étude de machines capables d'apprentissage ait quelque peu fait perdre de vue cette distinction entre algorithmes qui discriminent les formes, et algorithmes qui préparent les recherches de tels algorithmes.

I. 3 Hiérarchie des caractères

Soit un ensemble fini C de caractères $C_1, \ldots, C_i, \ldots, C_n$, et un ensemble E d'espèce animale $e_1, \ldots, e_j, \ldots, e_n$. Supposons que les caractères de C suffisent à distinguer les espèces de E, on pourra reconnaitre l'espèce d'un individu en cherchant successivement s'il possède chacun des C_i. Mais si toutes les combinaisons possibles de caractères de C ne sont pas représentées dans E (e.g. il n'y a pas d'animal qui ait à la fois une carapace et des plumes...), il existe un ordre hiérarchique optimum dans lequel chercher les caractères pour déterminer au plus vite l'espèce (e.g. chercher C_3, et si l'individu possède C_3 chercher C_6, sinon chercher C_{12} etc...). On peut, si l'on connait les fréquences relatives des diverses espèces de E, établir une stratégie de recherche qui minimise l'espérance mathématique du nombre de questions à poser pour déterminer l'espèce d'un individu (voir infra 442). Ainsi certains caractères apparaissent comme hiérarchiquement élevés : il faut toujours les rechercher d'abord pour situer l'individu dans un grand groupe d'espèces (disons dans une classe). D'autres caractères au contraire ne doivent être recherchés qu'à l'intérieur de certains genres pour achever la détermination de l'espèce (ce d'autant plus qu'en dehors du groupe la détermination peut en être impossible, cf infra 212) enfin l'on doit se baser autant que possible sur des caractères faciles à déterminer... Les anciens naturalistes de Candolle, de Jumieu (cf S & S p. 34) avaient un sens inné des principes de théorie de l'information qui régissent la subordination des caractères. Mais S & S leur répondent d'avoir érigé en règles à priori, ce qui était plutôt le fruit d'une vaste expérience de la synthèse.

I. 4 Phylogénie ou phénotypie

S & S notent non sans ironie (p. 18) : "The advent of the theory of evolution changes the practice of systematics very little, although the professed philosophical basis of systematics was radically altered. Natural classifications were considered to be those established on the basis of monophyletic taxa". Seulement il est bien plus facile de voir que deux sous-espèces de chevaux se ressemblent, que de démontrer qu'elles descendent toutes deux de eohippus (ce qui est une cause acceptable de leur ressemblance) ; bien plus, le seul moyen d'établir que equus descend de eohippus et non, comme on l'avait

dit, d'hipparion c'est de comparer avec autant de soins qu'il se peut les apparences, les caractères phénotypiques, des animaux vivants et des fossiles, avant d'en inférer des filiations qui soient cause des ressemblances (cf. S & S pp. 55, 94, 216). Il peut même arriver qu'une classification légitimement reçue contredise ce que l'on présume de la parenté des espèces. On croit les crocodiles plus proches parents des oiseaux que des lézards ou des tortues (i,e. que crocodiles et oiseaux ont des ancètres communs dont ni lézards ni tortues n'en descendent) et cependant, on range les crocodiles avec les autres reptiles (cf. S & S p. 226). Rappelons ici que réagissant contre l'étude exclusive de la linguistique historique, F de Saussure a marqué que les divers éléments d'une langue, quelle qu'en soit l'origine constituent à un moment donné un système unique en équilibre dynamique. Dans le même sens, S & S soulignent que l'unité du groupe des mammifères doit être concilliée avec son origine apparemment multiple, polyphyletique "It is generally agreed among paleontologists that a minimum of two groups of reptiles (and possibly four or five) independently crossed the arbitrary line separating reptiles from mannuals." (p. 104). Ils vont beaucoup plus loin quand ils conjecturent (p. 75) : "it is ... not beyond the bounds of possibility that some unusual features of higher organisms are derived by gene transfer from very dissimilar forms of life". On ne saurait attribuer un plus grand rôle aux rapports synchroniques que de faire des caractères des êtres vivants, des sortes de maladies non seulement héréditaires mais contagieuses !

I. 5 Taxinomie quantitative

Puisque les critères ne s'imposent pas d'eux mêmes, que leur hiérarchie est inconnue a priori, que l'histoire des espèces se déchiffre dans leurs ressemblances, l'étude positive des effets, des ressemblances, doit précéder celles des principes. Pour S & S le recours à des algorithmes rigoureux s'impose.

II - CHOIX ET DESCRIPTION DES UNITÉS

Un algorithme numérique de classification ne peut s'appliquer qu'à un ensemble d'unités, chacune décrite suivant un code numérique. Ces unités ne peuvent guères être les animaux ou les plantes individuelles mais des groupes définis au préalable sans le secours d'algorithme. S & S remarquent, (p. 121), que de toutes les catégories en usage (embranchements, classe, ordre, famille, tribu, genre, espèce... pour ne rien dire des sous-ordres, sous-genre, sous-espèces etc...) les unités inférieures que l'on a appelées espèces dont les mieux définies, à l'exception, peut-être des très grandes classes. C'est donc en général un ensemble d'espèces qu'on se propose de décrire puis de répartir en groupes. Mais les propriétés des espèces ne nous sont connues que par l'examen des individus et c'est encore à reconnaitre les individus que les naturalistes utilisent les classifications. D'où difficultés.

II. 1. Les descriptions sont présentées sous forme d'une matrice, où l'usage attribue aux espèces les colonnes et aux caractères les lignes : l'élément a_{ij}, situé à l'intersection de la ligne i et de la colonne j, décrit l'espèce j du point de vue du caractère i. Comme les caractères sont de natures très diverses, il est difficile que les informations a_{ij} soient codées de façon homogène, ce qui est pourtant à peu près exigé par les algorithmes de classification. Examinons les principaux cas.

II 1.1 Le caractère i est une attribution ("avoir des ailes") ou une privation ("n'en pas avoir"). L'élément de matrice a_{ij} est un élément logique.

On notera :

$$a_{ij} = \text{oui} = + = 1 \ ,$$

s'il y a attribution de i à j, et

$$a_{ij} = \text{non} = - = 0,$$

s'il y a privation. Les naturalistes préfèrent les + et les - ; mais les calculateurs veulent 1 et 0, ce qui permet au besoin de traiter un élément logique comme si c'était une grandeur continue, un nombre réel.

II 1.2 Le caractère i est bivalent = e.g. gastéroptérygien signifie : dont les nageoires ventrales sont situées derrière les pectorales. Il est clair qu'ici "être derrière" n'est ni plus ni moins une privation que l'alternative "être devant". Il n'y a donc aucune raison de poser $a_{ij} = 1$ plutôt que 0 si j est gastéroptérigien : c'est une pure convention. Nous verrons dans la suite (cf. S & S p. 127) qu'il peut être important de s'en souvenir. De plus le caractère considéré exprime une relation entre des éléments dits nageoires que l'espèce est sensée posséder : en l'absence de ces éléments on ne peut rien dire : S & S mettent alors un "N.C.", non comparable, dans la case a . Nous croyons commode d'introduire dans le tableau deux lignes j et $\bar{j}$, correspondant aux deux valeurs ("être devant et être derrière"). Un individu sur lequel le caractère est déterminé aura un 1 et un 0 (l'un dans l'une, l'un dans l'autre ligne) ; aux N.C. correspondent deux zéros. D'où la matrice des trois possibilités :

	i	i	i_o
j	1	O	O
$\bar{j}$	O	1	O

II 1.3 Le caractère est multivalent (peut prendre plus de deux valeurs). On peut scinder le caractère en plusieurs caractères binaires. Supposons qu'il y ait trois valeurs : sodium (Na), potassium (K), et calcium (Ca). On fera trois (ou six, cf ci-dessus) caractères avec trois colonnes de réponses possibles

Na	1	O	O
K	O	1	O
Ca	O	O	1
$\overline{Na}$	O	1	1
$\bar{K}$	1	O	1
$\overline{Ca}$	1	1	O

Mais la plupart des caractères multivalents correspondent à des dénombrements : nombre de pétales, de paires de pattes. Comme les programmes des algorithmes de classification sont insensibles à l'ordre dans lequel sont rangées les lignes, si l'on code chaque valeur entière comme un caractère indépendant (comme Na-K-Ca), la colonne d'un élément à six pattes ne ressemble pas plus à celle d'un quadrupède qu'à celle d'un bipède. Si l'on veut éviter cela, sans donner à a_{ij} la forme d'un entier quelconque mais en lui gardant son caractère d'élément logique, (0 ou 1) il est préférable de prendre pour caractères élémentaires des inégalités. Voici deux solutions présentées avec leur matrice des cas possibles : la première matrice est un scalogramme de Guttmann, la seconde n'a de 1 qu'au voisinage de la diagonale :

$N \leqslant 1$	1	0	0	0	0
$N \leqslant 2$	1	1	0	0	0
$N \leqslant 3$	1	1	1	0	0
$N \leqslant 4$	1	1	1	1	0
$N > 1$	0	1	1	1	1
$N > 2$	0	0	1	1	1
$N > 3$	0	0	0	1	1
$N > 4$	0	0	0	0	1

$N < 3$	1	1	0	0	0
$2 \leqslant N < 4$	0	1	1	0	0
$3 \leqslant N < 5$	0	0	1	1	0
$4 \leqslant N$	0	0	0	1	1
$5 \leqslant N$	0	0	0	0	1

II 1.4 Le caractère est continu : c'est une mesure ou un rapport de mesures. Il semble naturel de noter a_{ij} comme un nombre réel, continu. Se pose alors le problème de l'échelle. Si les variations sont très grandes, on doit choisir une échelle logarithmique. Il est souvent utile de normaliser les mesures, c'est-à-dire de faire sur les a_{ij} d'une même ligne de la matrice une transformation linéaire afin que la moyenne en soit zéro et l'écart-type 1. Mais la normalisation peut être faite lors de l'analyse de la matrice, le naturaliste peut l'abandonner au statisticien...

Si l'on désire que la matrice ne contienne que des éléments logiques on peut diviser en degrés l'échelle continue de variation du caractère et le me-

surer par un entier ; cet entier est ensuite traduit en propriétés élémentaires comme on vient de le voir (cf 213).

II 1.5 Le caractère est aléatoire : rappelons que les caractères des espèces s'étudient sur des individus qui ne sont pas identiques entre eux. Supposons qu'un caractère i ne soit attribué qu'à une partie des individus de l'espèce j. On devra donner à a_{ij} non la valeur logique 1 ou 0, mais une valeur probabiliste comprise entre 0 et 1, (estimée par la fréquence de présence du caractère). S'il s'agit d'un caractère continu et codé comme tel (a_{ij} est un nombre 1) on pourra donner à la grandeur i la valeur d'une moyenne faite sur plusieurs individus de l'espèce). Ici mieux vaut choisir pour caractères des rapports de longueur que des longueurs, pour éviter qu'une trop grande dépression intraspécifique n'ôte tout intéret à la mesure.

II 1.6 Relevés écologiques : l'objet de l'écologie est, nous l'avons dit, l'étude des communautés d'espèces occupant un terrain (prairie, lac...). Ici, les terrains jouent le rôle des individus (ou des espèces) à classer ; les caractères sont les espèces (animales ou végétales) présentes à des degrés variables, ainsi que divers paramètres climatiques, chimiques ... Eventuellement, on peut adopter le point de vue transposé du précédent : l'objet demeure la classification des espèces, mais on considère leur "sociabilité" non leur morphologie. Si l'on élimine paramètres climatiques ou chimiques, la matrice de description revêt la forme homogène d'un tableau de relevé : les colonnes correspondent aux divers échantillons de terrains étudiés, les lignes aux espèces ; a_{ij} peut être soit le nombre d'individus de l'espèce i trouvés dans le terrain j, soit la mesure de l'aire de j recouverte par l'espèce i, soit un caractère logique (1 si l'espèce i est présente dans j, 0 si elle est absente). Pour une étude très complète nous renvoyons à Dagnélie (1960).

En résumé, les matrices de description se présentent sous les formes principales (pour ne rien dire des hybrides que l'on s'efforce, comme on vient de le voir d'éliminer par des artifices) :

- matrices logiques : les éléments sont des 1 et des 0, qui signifient "oui" et "non",

- matrices logiques aléatoires : les éléments sont des nombres continus compris entre 0 et 1, qui signifient : "probabilité de oui".

- matrices de scalaires quelconques : les éléments en sont des résultats de mesure traduits dans une échelle convenable et éventuellement normalisés.

- matrices de relevés écologiques.

Sans présumer ici de l'analyse de ces tableaux de nombres, remarquons dès maintenant que matrices logiques, matrices logiques aléatoires et, particulièrement, matrices de relevés écologiques, sont des matrices de correspondance.

II 2 La construction de matrices à peu près homogènes n'est pas le seul problème d'intéret statistique que la description des espèces pose au naturaliste.

II 2.1 Choix des caractères. Une description exhaustive des espèces est impossible, particulièrement s'il s'agit de fossiles. Mais Musset dit que

"quand on voit le pied la jambe se devine", et les paléontologues, qui sont quelque peu poètes, ne se font pas faute de disserter péremptoirement sur une machoire comme sur un être entier. Qu'un échantillonage assez arbitraire des caractères permette d'avoir une connaissance sûre des espèces repose selon S & S (p. 84 sqq.) sur deux hypothèses qui semblent confirmées :

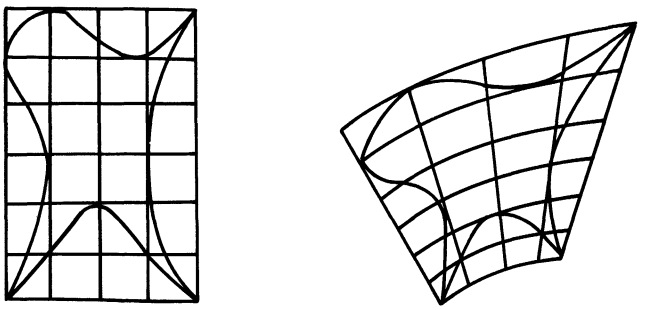
- l'hypothèse du nexus, selon laquelle tout caractère est affecté par plusieurs facteurs génétiques et, réciproquement, la plupart des gènes influencent plusieurs caractères.

- l'hypothèse de la non-spécificité, selon laquelle il n'existe pas de groupe de gènes dont l'influence se limiterait exclusivement à une seule partie, ou une seule fonction de l'organisme.

Selon ces hypothèses, la connaissance d'une partie des caractères d'une partie de l'être vivant (par exemple la connaissance des caractères extérieurs, plus faciles à déterminer), peut nous renseigner sur l'ensemble des facteurs génétiques qui en contrôlent toute la structure. Effectivement, à partir de systèmes très différents de caractères (musculature d'une partie du corps, squelette d'une autre) on a pu calculer des coefficients de proximité entre espèces (cf. infra) fortement mèlés entre eux, et conduisent à des classifications voisines. Chez les insectes (48 espèces de moustiques) des études basées l'une sur les caractères des larves, l'autre sur ceux des adultes ont donné des résultats similaires mais divergents sur des détails. Mais il n'en arrive pas moins parfois que les métamorphoses défient la taxonomie. Ainsi (cf Bohn (1934) p. 64) "axolote et amblystone sont deux formes différentes de la même espèce, et la reproduction peut avoir lieu sous l'une ou l'autre de ces formes : sous la forme infantile, l'axolote, et sous la forme adulte, l'amblystone. Avant qu'on s'en soit rendu compte, le même animal avait été placé dans deux sous-ordres différents : Perennibranches axolote) et Salamandrines (amblystone)". (cf aussi la difficulté de la classification des éphémères, où ressemblances entre larves et entre image ne s'impliquent pas mutuellement).

II 2.2 Correspondances entre espèces. Déterminer les mêmes caractères pour des espèces différentes présuppose qu'une même terminologie puisse leur être appliquée : e.g. que l'on puisse parler des deux côtés du coeur ou du foie, de l'aorte ou de l'acidité gastrique. Faute d'une correspondance assez étendue entre les membres et les fonctions de deux individus d'espèces différentes, le tableau des caractères comporterait surtout des "N.C." (cf. supra 212 et, aussi, 13). Dans les cas des mesures de longueurs, cette correspondance doit souvent être une véritable carte d'un individu sur l'autre, carte faisant se correspondre entre eux les divers points de repère (apophyses, sutures...)

La figure donnée ci après est inspirée de celles que proposent S & S (pp. 82, 83). D'un point de vue mathématique il serait intéressant de rapporter l'un des deux individus à un système de coordonnées cartésiennes (le quadrillage régulier à gauche) et de déterminer une carte différentiable simple (e.g. des polynomes) qui envoie au mieux le premier individu sur le second (le quadrillage régulier de gauche, sur le réseau curviligne de droite). Les coefficients de ces fonctions seraient des paramètres naturels pour mesurer les déviations des formes à partir d'une espèce choisie pour référence. Dans ce sens S&S signalent un travail de Smirnov (1927) où sont calculés les premiers coefficients de Fourrier du contour des élitres de coccinelles.



III - INDICES DE SIMILARITE

La forme matricielle de la description suffit à elle seule à suggérer l'emploi de l'analyse factorielle. Mais la taxinomie quantitative a été fondée par des praticiens qui savaient peu d'algèbre et ne disposaient pas encore d'ordinateurs. Nous présenterons leurs techniques les premières avant de traiter de l'analyse factorielle. Quoique l'analyse des proximitées, due à R. N. Shepard, ne semble pas encore avoir pénétré dans la taxinomie, c'est sur la définition d'indices numériques de similarité que reposent la plupart des techniques de classifications citées par S & S. (Un indice de similarité sur un ensemble K est une fonction $s_{kk'} = s_{k'k}$, à valeur réelle, définie sur les paires, k, k', d'éléments de K ; s n'est pas assujettie aux axiomes d'une distance, on demande simplement que s ait une valeur maxima, v, atteinte seulement pour toute paire formée de deux éléments identiques : $\forall k \in K, s_{kk} = v$. Mais il est possible de définir sur K une vraie distance qui ne soit fonction que de s.) La classification des caractères (appelée R-technique par S & S) étant encore peu étudiée sauf toutefois en écologie (cf. 216), c'est principalement la similitude des individus représentés chacun par une colonne de la matrice de description) qu'on s'attache à chiffrer (Q-technique). Dans ce § nous suivons le chapitre VI de S & S.

III. 1 Coefficients d'association

Ces coefficients se calculent sur les colonnes d'une matrice logique ; les formules peuvent être généralisées aux matrices logiques aléatoires. Notons I l'ensemble des caractères i, et k et l (ou parfois K, L) deux individus de l'ensemble J à classer ; on pose :

n = nombre total des éléments de I

$$n_K = \sum_{i \in I} a_{ik} \quad ; n_k = \sum_{i \in I} (1 - a_{ik})$$

$$n_L = \sum_{i \in I} a_{il} \quad ; n_l = \sum_{i \in I} (1 - a_{il})$$

$$n_{KL} = \sum_{i \in I} a_{ik}\, a_{il} \quad ; n_{kl} = \sum_{i \in I} (1 - a_{ik})\,(1 - a_{il})$$

$$n_{KL} = \sum_{i \in I} a_{ik}\,(1 - a_{il}) \; ; n_{kL} = \sum_{i \in I} (1 - a_{ik})\, a_{il}$$

Les notations sont claires : un indice capital, e.g., K, se réfère aux caractères que possède l'individu k, aux 1 de sa colonne ; un indice minuscule se réfère aux zéros. Ainsi n_{kl} est le nombre total des caractères que k possède et que l ne possède pas.

Dans le cas d'une matrice aléatoire, notre formule donne à n la valeur de l'espérance mathématique du nombre des caractères trouvés sur un k mais non sur un l (le k et le l, étant choisis au hasard dans leur espèce). Si les colonnes k et l de la matrice de description comporte des "N.C." (cf. 212), on ne considère que les caractères déterminés à la fois sur l et k. Les deux sommes :

$$m = n_{KL} + n_{Kl}$$

$$u = n_{Kl} + n_{kL}$$

qui représentent respectivement le total des concordances (m = matched) et des discordances (u = unmatched) entre les colonnes k et l, ont sans doute plus de sens que les bilans partiels, tels que n_{KL} ou n_{Kl}, qui dépendent de ce que l'on a décidé d'appeler "caractères positifs" et "caractères négatifs" ; nous avons discuté plus haut (212) du codage des caractères bivalents dans la matrice de description. Si l'on admet que les a_{ij} égaux à 1 correspondent pour la plupart à des attributions vraies, les zéros à des privations (c'est bien le cas pour une matrice logique d'origine écologique, cf 216), on pourra définir des indices dépendant non seulement de m et u, mais aussi de n_{KL}, n_{Kl} etc... S & S présentent un tableau de nombreuses formules qui ont été utilisées ou seulement proposées. Avec chaque indice ils donnent son "espérance mathématique" exprimée en fonction de n_K, n_L, n, les n_K caractères que possède k et les n de l étant tirés au sort indépendamment dans I : or leurs formules sont, pour la plupart, fausses et semblent avoir été calculées comme si la moyenne du quotient était le quotient des moyennes (x/y = x/y !). Nous avons cru devoir signaler cette erreur, quoiqu'à nos yeux elle ne soit qu'un point noir sur un très beau travail de synthèse. Ceci dit, voici, désignés par les noms des auteurs, divers indices avec leurs intervalles de variation (la proximité maxima correspond à la valeur 1 ou ∞).

Sokal et Michener (1958)	: m/n	; de 0 à 1
S & S	: $2m/(m+n)$	; de 0 à 1
Rogers et Tanimoto (1960)	: $m/(m+u)$	; de 0 à 1
S & S	: m/u	; de 0 à ∞
Dice (1915) ; Sorensen (1948)	: $2n_{JK}/(2n_{JK}+u)$	; de 0 à 1
Juccard (1908)	: $n_{JK}/(n_{JK}+u)$	; de 0 à 1
Russel et Rao (1940)	: n_{JK}/n	; de 0 à 1
S & S	: $n_{JK}/(n_{JK}+2u)$	; de 0 à 1
Kulczynski (1927)	: n_{JK}/u	; de 0 à ∞
Ibid.	: $n_{JK}\frac{1}{2}\left[\frac{1}{n_J}+\frac{1}{n_K}\right]$	; de 0 à 1
Ochiai (1957)	: $n_{JK}/\sqrt{n_J n_K}$	; de 0 à 1

Notons que de nombreux auteurs, (Kulczynsli, Sørensen etc...) ont en vue les applications écologiques (cf 216).

III.2 Coefficients de corrélation

Ces coefficients se calculent sur les colonnes d'une matrice de scalaires quelconques. On traite ces colonnes comme des fonctions numériques de la variable i I, et on a :

$$r_{kl} = \frac{\sum_{i \in I} (a_{ik} - \overline{a_k})\ (a_{il} - \overline{a_l})}{\sqrt{\left[\sum_{i \in I} (a_{ik} = a_k)^2\right] \left[\sum_{i \in I} (a_{il} - a_l)^2\right]}}$$

Où $\overline{a_k}$ (resp. a_l) est la moyenne arithmétique des a_{ik} (resp. a_{il}). Certains auteurs travaillent sur des matrices dont les lignes ont été normalisées (amenées à avoir pour moyenne 0 et pour écart type 1, cf 214). L'interprétation géométrique est classique. A chaque individu $k \in J$, correspond dans R^n le vecteur $\vec{a}_k$ de composantes les a_{ik} ; le vecteur $\vec{a}'_k$, de composantes les $\overline{a'_{ik}} = a_{ik} - \overline{a_k}$, et la projection orthogonale de $\vec{a}_k$ sur l'hyperplan perpendiculaire à la diagonale principale, hyperplan lieu des vecteurs de R^n dont la somme des composantes est nulle. On a

$$r_{kl} = \cos\ (\vec{a}_k\ ,\ \vec{a}'_l)$$

(les angles sont définis pour la norme euclidienne usuelle de R^n, racine carrée de la somme des carrés des coordonnées). Le coefficient de corrélation vaut 1 si $\vec{a}'_k$ et $\vec{a}'_l$ sont parallèles et de même sens ; dans le cas particulier où la matrice de corrélation est une matrice logique (des 1 et des zéros) cela implique que les vecteurs colonnes $\vec{a}_k$ et $\vec{a}_l$ sont identiques. En général on admet que la ressemblance est d'autant plus grande que le coefficient de corrélation est plus voisin de 1. Selon S & S, les valeurs fortement négatives sont rares (si les lignes n'ont pas été normalisées) : pour eux cela s'explique parce que la notion d'antiespèce est inconcevable (cf p. 141). Considérons le cas particulier d'une matrice logique, et reprenons les notations du n° 31. On aura $r_{kl} = -1$ si et seulement si :

$$n_K = n_{Kl} = n_l$$

$$n_k = n_{kL} = n_l\ ,$$

c'est-à-dire si les deux espèces k et l s'opposent sous tous les rapports : or en prenant l'antithèse d'un être vivant k, on a peu de chances de définir un être viable l ! Cette remarque intéresse la logique, la sémantique et la reconnaissance des formes. En logique formelle (e.g. calcul des prédicats) une propriété et sa négation sont sur le même plan, comme le sont un ensemble et son complémentaire (ou, cf 212, les deux états d'un caractère bivalent). Mais une notion définie comme une limite idéale (être vivant, être bon) n'a rien de même nature qui s'oppose à elle ; disons, pour user d'une image mathématique, que, si une partie admet un complémentaire, un filtre n'a pas d'opposé (soit $\mathcal{F} \subset \mathcal{B}(E)$ un filtre sur E ; le complémentaire de $\mathcal{F}$ dans $\mathcal{B}(E)$ n'est pas un filtre ; $\{g \mid Cg \in \mathcal{F}\}$, ensemble des complémentaires des parties de E, éléments de $\mathcal{F}$, n'est pas un filtre non plus). Un modèle ma-

thématique de ces limites idéales pourrait se présenter ainsi : soit E un ensemble infini de qualités sur lequel est défini une mesure positive μ (e.g. par une densité de masse...), supposons donnée pour tout nombre x de l'intervalle (0,1) une partie mesurable $Q_x \subset E$ (on peut postuler que les Q_x forment une famille décroissante : $x < y \Longrightarrow Q_y \subset Q_x$, et que Q_1 est vide) ; un individu i sera défini comme une fonction f sur E prenant ses valeurs dans (0,1), si q est une qualité, $q \in E$, f(q) est la probabilité que i manifeste q. Maintenant, on dira que i possède la qualité idéale Q, définie par φQ , si :

$$\forall x \in 0,1 = x \leftarrow \frac{\int_{Q_x} f(q)\ d\mu(q)}{\int_{Q_x} d\mu(q)} ;$$

autrement dit si i possède une fraction supérieure à x des qualités de chaque Q_x. Avec un tel modèle, la qualité idéale Q et sa négation se présentent si différemment, qu'on peut dire que Q n'a de contraire qui soit sur le même plan qu'elle. De même une forme ; "être un A", n'a pas de contraire : l'ensemble des dessins qui sont des A n'est défini que par rapport à une structure idéale, à laquelle aucune structure opposée ne correspond, qui puissent définir les "non-A". Dans un message télégraphique, le "non-A" peut se définir par "B ou C... ou Z... ou 9", mais en général l'ensemble des objets informes est le contexte inévitable de tout problème de reconnaissance des formes (or ce contexte est méconnu par ceux qui parlent d'une partition de tous les objets entre les formes). De toutes les combinaisons concevables de caractères, bien peu sont assez conformes à l'idéal de la vie pour définir un être viable d'une forme ou d'une autre. C'est justement cette distribution sporadique des vivants qui rend possible et utile leur séparation en classes.

A titre d'exercice de géométrie, examinons comment, si les lignes sont normalisées, peuvent apparaître des coefficients de corrélation voisins de -1. Si les lignes ont été normalisées, les vecteurs $\vec{a}_k$, et donc aussi les a'_k, ont pour somme zéro :

$$\sum_{k \in J} \vec{a}_k = \sum_{k \in J} \vec{a}'_k = 0 ;$$

ils sont répartis à peu près symétriquement autour de l'origine. Il ne doit pas être rare que deux d'entre eux $\vec{a}'_k$ et $\vec{a}'_l$ aient approximativement des directions opposées ; on a alors :

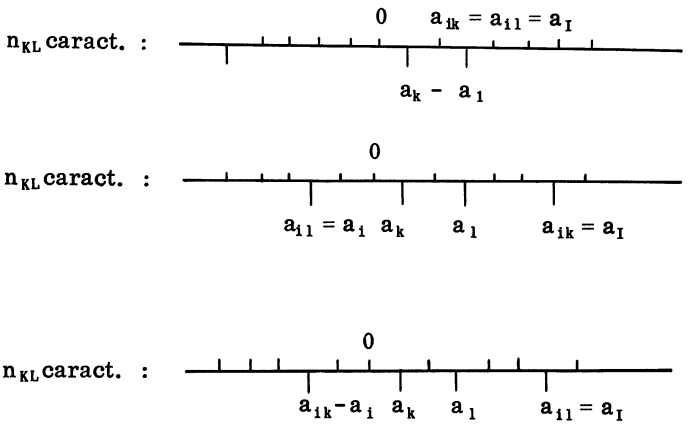
$$r_{kl} = \cos(\vec{a}_k , \vec{a}_l) \# -1 .$$

Pareille situation peut se rencontrer, e.g., si la matrice a_{ij} est obtenue par normalisation d'une matrice logique. Voici un exemple artificiel où $\vec{a}_k = -\vec{a}_l$, cependant que k et l s'accordent sur plus de 70 % des caractères, situation vraisemblable et, en tout cas, fort éloignée de celle des anti-espèces. Remarquons d'abord quelle est la forme, après normalisation, d'une ligne de matrice logique : s'il y a m_I 1 et m zéros, les 1 sont remplacés par a_I, les zéros par a_i avec :

$$a_I = \sqrt{m_i/m_I} \quad ; \quad a_i = -\sqrt{m_I/m_i} = -1/a_I .$$

Supposons que les caractères se répartissent en trois classes d'effectifs n_{KL}, n_{kL}, n_{KL} (cf notations 31), pour lesquelles les valeurs de deux in-

dividus k et l se placent après normalisation comme indiqué sur les graphiques ci-dessous (où la graduation vaut $1/\sqrt{12}$ afin que les $a_I \times a_i$ vaillent 1)



On voit que si les valeurs moyennes $\bar{a}_k$ et $\bar{a}_l$ sont bien à la place figurée, on a $\vec{a}_k^{\,l} + \vec{a}_l^{\,l} = 0$.

Pour cela il faut et il suffit que :

$$n_{KL} + 5n_{Kl} - 3n_{kL} = 0,$$

ce qui est réalisé, e.g. si :

$$n_{KL} = 110 \ ; \ n_{Kl} = 2 \ ; \ n_{kL} = 40.$$

III. 3 Distances

S & S, s'adressant à un public peu versé dans les mathématiques commentent soigneusement des formules classiques.

$$\Delta_{lk} = \frac{1}{n} \sum_{i \in I} |a_{il} - a_{ik}| \quad , \text{ ou}$$

$$\Delta_{lk} = \sqrt{\sum_{i \in I} (a_{il} - a_{ik})^2}$$

Ils signalent aussi le coefficient de divergence de Clark (1952) qui présente l'intérêt de varier entre 0 et 1, si le a_{ik} sont tous positifs :

$$D_{lk} - \left[\frac{1}{n} \sum_{i=1}^{n} \left(\frac{a_{il} - a_{ik}}{a_{il} + a_{ik}} \right)^2 \right]^{1/2}$$

Et rappellent le coefficient de similitude entre races défini par K. Pearson (1926) :

$$S_{lk} = \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[\frac{(a_{ik} - a_{il})^2}{\frac{s_{ik}^2}{n_k} + \frac{s_{il}^2}{s_l}} \right] \right\} - \frac{2}{n}$$

où a_{ik} est une valeur expérimentale du caractère i calculée comme moyenne sur un échantillon de n_k individus de l'espèce k ; et s_{ik} est l'écart-type du caractère i sur l'espèce k.

L'analyse factorielle des correspondances utilise, on le sait, une distance définie par une forme quadratique particulière. Nous retrouverons donc les problèmes de distances au § 6.

III.4 Indices réduits

Au paragraphe suivant, nous présentons quelques algorithmes indépendants de l'analyse factorielle. Les algorithmes utilisent pour données des indices de similarité calculés à partir de la matrice de description suivant l'une des techniques exposées ci-dessus. Il est généralement commode de supposer que l'indice de similarité varie de 0 (similarité minimale à 1 (identité). D'où la necessité de réduire les indices qui ne satisfont pas à cette condition. Si on a calculé d'abord une distance d on la convertira en indice s par une formule telle que :

$$s_{kl} = [10 - d_{kl}]/10 ,$$

(où [x] désigne x si x est positif et o autrement), deux unités séparées par une distance $\geqslant$ à un seuil arbitrairement fixé à 10 ont une similarité nulle. De même, à partir d'un coefficient de corrélation on peut poser :

$$s_{kl} = (1 + r_{kl})/2 .$$

Parfois S & S écrivent : "k et l sont similaires à x %". Cela n'a pas de sens précis indépendamment de la convention qui régit le calcul d'un indice s, et signifie seulement : $s_{kl} = x/100$. Une terminologie plus précise est : k et l sont similaires, pour s, à x (resp. à x %, resp. à plus de x, ce qui équivaut à la formule : $s_{kl} = x$ (resp. $= x/100$, resp $> x$). On peut cependant tenter de donner de l'indice réduit une définition à peu près indépendante du point de vue adopté (association corrélation, distance pour élaborer la matrice de description. Les recherches de R. N. Shepard ont montré l'intérêt de considérer, non les distances d' (ou les similarités s'), mais les inégalités qu'elles vérifient et il est vraisemblable que ces inégalités ne sont pas trop sensibles aux divers modes de calcul de d ou de s. D'où une nouvelle définition de l'indice réduit s : s_{kl} est la probabilité que deux individus pris au hasard dans J soient moins semblables (moins corrélés, plus éloignés) que k et l. (Pratiquement, on pourrait se borner à ne considérer dans les calculs qu'une dizaine de degrés de similarité distincts également espacés). Il importe de remarquer que si s est l'indice réduit (1-s) n'est pas nécessairement une distance satisfaisant à l'inégalité du triangle.

IV - UN PROBLÈME D'OPTIMISATION

Afin de préciser le terme idéal vers lequel nous paraissent tendre beaucoup de recherches (cf S & S ch VII), nous donnerons une définition. Nous proposerons ensuite quelques remarques et généralisations.

IV.1. Soit J un ensemble fini (l'ensemble des unités, cf supra). On dit que l'ensemble H de parties de J ($H \subset \mathcal{B}(J)$) est une hiérarchie de parties de J si sont satisfaites les conditions suivantes :

H_0 : $J \in H$, H contient la partie de J constituée par J lui-même.

H_1 : $\forall j \in J : \{j\} \in H$: H contient les parties formées d'un seul élément.

H_2 : Soit h, $h' \in H$; si $h \cap h' = \emptyset$, h et h' sont comparable pour la relation d'inclusion (i.e. l'un des deux est inclus dans l'autre).

Des axiomes, résulte la propriété suivante : Soit $h \in H$; notons P (h) la partie de H ensemble de successeurs immédiats de h dans H ordonné par inclusion (i.e. P (h) est l'ensemble des h' tels que $h' \in H$, $h' \subset h$, et qu'il n'existe pas de h" avec $h'' \in H$ $h' \subset h'' \subset h$). P (h), considéré comme famille de parties de J est une partition de h.

Une hiérarchie H est dite <u>indicée</u> si pour tout $h \in H$ est donné un indice x_h, nombre compris entre 0 et 1, avec les conditions suivantes :

I_1 : si la partie h compte un seul élément ($h = \{j\}$), $x_h = 1,00$.

I_2 : si $h \subset h'$ (inclusion stricte), $x_{h'} < x_h$ (inégalité stricte).

A partir d'une hiérarchie indicée, on définit <u>un indice de similarité</u> s sur J ; soit k, $l \in J$; s_{kl} est le plus grand nombre tel qu'il existe $h \in H$ avec :

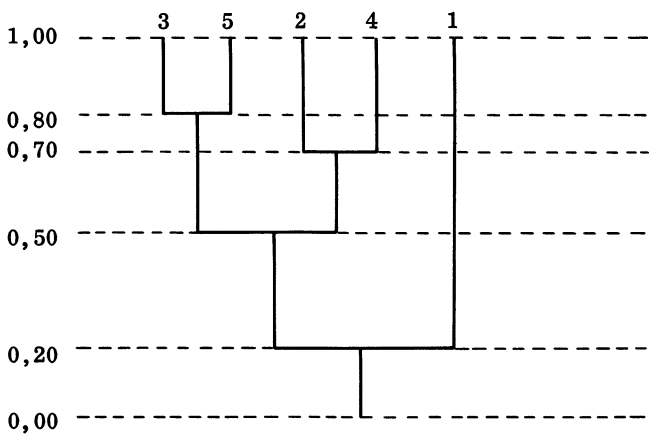
$$k \in h \ ; \ l \in h \ ; \ x_h = s_{kl} .$$

On déduit facilement des axiomes la propriété suivante de l'indice s : soit x un nombre entre zéro et 1, la relation :

$$s_{kl} \geqslant x$$

(entre deux éléments k, $l \in J$) est une relation d'équivalence sur J. On peut noter P_x la partition correspondante de J : les classes de P_x ne sont autres que les $h \in H$ maximaux parmi ceux qui satisfont à $x_h \geqslant x$. Il est clair que si $x > y$, la partition P_x est plus fine que P_y.

IV.2. Voici un exemple de hiérarchie indicée, représentée comme un arbre dont les noeuds sont les $h \in H$, ont pour cote x_h, et sont reliés par des branches à leurs successeurs immédiats :



Les parties h de la hiérarchie sont données avec leurs indices :

{1, 2, 3, 4, 5} : indice 0,20

{5, 2, 3, 4} : indice 0,50

{2, 4} : indice 0,70

{3, 5} : indice 0,80

{1}, {2}, {3}, {4}, {5} : indice 1,00.

A H est associée la matrice de similarité :

	1	2	3	4	5
1	1,00	0,2	0,2	0,2	0,2
2	0,2	1,00	0,5	0,7	0,5
3	0,2	0,5	1,00	0,7	0,8
4	0,2	0,7	0,5	1,00	0,5
5	0,2	0,5	0,8	0,5	1,00

Les partitions successives sont les sections horizontales de l'arbre.

$P_{1,00} = (\{1\}, \{2\}, \{3\}, \{4\}, \{5\} = P_x \; ; \; 1 \geqslant x \geqslant 0,8$

$P_{0,80} = (\{3,5\}, \{1\}, \{2\}, \{4\}) = P_x \; ; \; 0,8 \geqslant x \geqslant 0,7$

$P_{0,70} = (\{3,5\}, \{2,4\}, \{1\}) = P_x \; ; \; 0,7 \geqslant x \geqslant 0,5$

$P_{0,50} = (\{2,3,4,5\}, \{1\}) = P_x \; ; \; 0,5 \geqslant x \geqslant 0,2$

$P_{0,20} = (\{1,2,3,4,5\}) = P_x \; ; \; 0,2 \geqslant x \geqslant 0$

IV. 3. Ceci dit, l'idéal des taxinomistes semble être de définir une hiérarchie indicée H de parties de J, telle que l'indice de similarité associé à H ait une corrélation maxima avec celui calculé comme indiqué ci-dessus § 34) à partir de la matrice de description. Une classe $h \in H$ sera appelée non "ordre", "famille", "tribu"... selon la terminologie en usage en taxinomie non-numérique, mais (selon S-15) "x -taxon" "x - .vut", terme qui a une valeur quantitative précise. Certains voudraient qu'un x-taxon soit formé par la descendance d'une espèce unique, ayant vécu à une époque d'autant plus reculée que x est plus petit, l'arbre des ressemblances, associé à la hiérarchie H serait ainsi l'arbre généalogique. C'est là un point de vue que la glottochronologie a rendu familier aux linguistes. Mais (cf 14) plus de prudence s'impose dans l'établissement des généalogies.

IV. 4. Il importe de remarquer que la transitivité de la similarité (i. e. de la relation binaire $s_{kl} \geqslant x$) est loin d'être une propriété générale des indices de similarité, mais résulte de ce que, du fait des axiomes des hiérarchies, la fonction

$$d_{kl} = (1 - S_{kl}) \ ,$$

(définie sur $J \times J$) est une distance ultramétrique sur J, i.e. une distance pour laquelle tout triangle est isocèle, ou encore :

$$\forall j, k, l \in J \ : d_{kl} \leqslant \sup \ (d_{jk}, d_{jl}) \ .$$

Nous dirons en général qu'un indice de similarité s est ultra-métrique, si (1-s) est une distance ultramétrique. Soit s un indice ultramétrique sur J, on peut montrer que, réciproquement, il y a une hiérarchie indicée unique H sur J dont s soit l'indice associé. Pout tout $x \in (0,1)$ on définit d'abord une partition P de J, par la relation d'équivalence ($s_{kl} \geqslant x$) ; une partie h de J appartient à la hiérarchie H si et seulement si il existe un x tel que h soit une des composantes de la partition P , ce qu'on peut écrire :

$$H = \bigcup_{x \in (0,1)} P_x$$

(où P_x est considéré comme un ensemble de parties de J) ; et l'indice x_h de h est le plus grand x tel que $h \in P_x$

$$x_h = \sup_{h \in P_x} x$$

IV. 5. Du point de l'analyse des proximités (R. N. Shepard cf supra 34) seules importent les inégalités satisfaites par les indices des parties h d'une hiérarchie H. Les valeurs mêmes données à ces indices sont irrelevantes. On peut les choisir telles que l'indice de similarité réduit de la hiérarchie indicée soit un indice réduit au sens probabiliste que nous avons donné plus haut à ce terme. Mais il semble intéressant d'éliminer les nombres en définissant la notion de hiérarchie stratifiée. Une hiérarchie stratifiée sur J est une hiérarchie H de parties de J sur laquelle est définie un préordre total (i.e. une relation notée $h \prec h'$, transitive, telle que pour tout couple $h, h' \in H$, soit vraie l'une au moins des deux relations :

$$h \prec h' \quad , \quad \text{ou } h' \prec h \ ,$$

est telle que soit une relation d'équivalence la relation $\sim$ définie par :

$$h \sim h' \Longleftrightarrow (h \prec h') \wedge (h' \prec h) \ ;$$

autrement dit, $\prec$ induit une relation d'ordre total sur le quotient de H par $\sim$) satisfaisant aux deux axiomes (substituts de I_1 et I_2)

S_1 : $\forall j, j' \in J$, $\{j\} \prec \{j'\}$ et $\{j'\} \prec \{j\}$ (i.e. $\{j\} \sim \{j'\}$)

S_2 : si $h' \subset h$ (inclusion stricte), on a : $h' \prec h$, et l'on n'a pas : $h \prec h'$.

A toute hiérarchie indicée est associée une hiérarchie stratifiée : il suffit de poser :

$$(h \prec h') \Longleftrightarrow (x_{h'} > x_h)$$

Une hiérarchie stratifiée H sur J définit une ordonnance sur J.

Soit j, j', k, k', quatre éléments de J : pour que j soit plus similaire à j' que ne l'est k à k', il faut que :

$$h_j \prec h_k,$$

où h_j (resp. h_k) est la plus petite partie de la hiérarchie H contenant à la fois j et j' (resp. k et k').

Ceci suggère un nouveau problème d'optimisation (apparenté à celui étudié dans notre première leçon "sur les algorithmes de classification"):

Etant donné un ensemble J muni d'une ordonnance ω, définir sur J une hiérarchie sratifiée H, de telle sorte que soit maximum le nombre $\mathcal{C}$ des inégalités spécifiées par ω et impliquées par H.

V. Algorithmes de Classification :

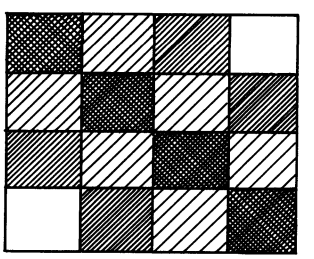
Laissant au § suivant l'application de l'analyse factorielle, nous présentons ici des méthodes diverses sous trois titres : méthodes visuelles, algorithmes ascendants, clefs et algorithmes descendants, non sans les rapporter au passage à ce qui nous paraît l'idéal de la taxinomie numérique.

V.1. Méthodes visuelles

On peut représenter de façon sensible la matrice de similarité (matrice des S_{hl}) en noircissant les cases avec une intensité proportionnelle à l'indice de similarité qu'on y devrait inscrire : e.g.

S_{11}	S_{12}	S_{13}	S_{14}
S_{21}	S_{22}	S_{23}	S_{24}
S_{31}	S_{32}	S_{33}	S_{34}
S_{41}	S_{42}	S_{43}	S_{44}

1.00	.05	.80	.00
.05	1.00	.05	.80
.80	.05	1.0	.05
.00	.80	.05	1.00



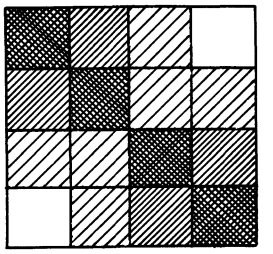
Si les diverses espèces sont rangées successivement dans un ordre tel que les espèces d'une même classe soient groupées, apparaissent le long de la diagonale principale de la matrice de similarité des cases, d'un noir soutenu, correspondant aux classes. D'où une première méthode de classification (Robinson, 1951) : permuter les unités en tendant grouper les colonnes qui se ressemblent, jusqu'à faire apparaître autant que possible des carrés noirs le long de la diagonale.

Voici ce que donne l'exemple que nous avons figuré quand on regroupe les classes 1-3 et 2-4.

Les permutations de colonnes sont d'une pratique connue en statistique appliquée : c'est ainsi qu'on construisait à l'origine les scalogrammes de Guttman. Mais ici, comme le remarquent S & S, on ne peut se borner à manipuler des languettes de papier portant les colonnes, car on doit simultanément permuter de même lignes et colonnes. La méthode de permutation de visu n'est donc guère commode pour plus d'une dizaine d'espèces.

S_{11}	S_{13}	S_{12}	S_{14}
S_{31}	S_{33}	S_{32}	S_{34}
S_{21}	S_{23}	S_{22}	S_{24}
S_{41}	S_{43}	S_{42}	S_{44}

1.00	.80	.05	.00
.80	1.00	.05	.05
.05	.05	1.00	.80
.00	.05	.80	1.00



Une autre méthode visuelle est de figurer un graphe de similarité : les unités sont des points reliés par des liens d'autant plus épais que la similarité est plus grande entre les extrémités ; on tente de grouper les points en des classes dont tous les membres soient liés par des liens épais (cf. e.g. Bocke 1942). Dagnelie remarque judicieusement (1960 p. 33) que la figuration matricielle tente une répartition des individus suivant une dimension (matérialisée par la diagonale), tandis que le graphe plan se place dans un continuum à deux dimensions ; en ce sens, les méthodes visuelles sont une introduction à l'analyse factorielle, méthode spatiale par excellence. Mais les méthodes visuelles sont inaptes à traiter un grand nombre d'espèces et ne permettent pas de recherche systématique de l'optimum.

V.2. Quelques algorithmes ascendants (d'après S & S, chap. VII)

Comme le problème d'optimisation posé ci-dessus (43) est fort complexe, et que d'ailleurs aucun taxinomiste ne l'a formulé en toute rigueur, c'est par des approximations plus ou moins confuses que les recherches quantitatives tendent vers ce que nous croyons pouvoir appeler l'idéal. Mais quoique d'inspiration diverses, les méthodes présentées ci-dessous se trouvent, selon S & S donner des résultats fort comparables entre eux et avec ceux des méthodes non-numériques usuelles ; d'où leur intérêt.

V.3. Méthode de Sneath (1957 b)

Soit s un indice de similarité donné sur J ; x un nombre quelconque entre 0 et 1. Si s n'est pas ultramétrique (cf 424) la relation binaire ρ_x sur J :

$$s_{kl} \geqslant x \ ,$$

n'est pas, en général une relation d'équivalence. Mais il est facile de construire à partir de s un nouvel indice s' qui soit ultramétrique en saturant la relation binaire ci-dessus. Nous définirons s'_{kl} comme le plus petit nombre x, tel que l'on puisse trouver dans J une suite de points k_i (i = 1,..., p) satisfaisant aux conditions ci-dessous :

$$k_1 = k \ ; \ k_p = 1$$

$$\forall\, i < p \ ; \ s_{k_i k_{i+1}} \geqslant x \ .$$

Ainsi la relation ρ'_x

$$s'_{kl} \geqslant x$$

(k et l sont similaires pour s' à plus de x) est bien la relation d'équivalence saturée de ρ_x, relation d'équivalence qui s'énonce : on peut relier k à l par une chaine de points dont chacun est similaire au précédent, pour s, à plus de x. L'indice s' peut encore être défini par les deux propriétés suivantes :

- s' est ultramétrique
- $\forall k,\ l \in J,\ s'_{kl} \geqslant s_{kl}$.

Pour construire les classes de la hiérarchie H' définie par s' Sneath utilise une méthode ascendante ("Adansonienne", cf 12) : il définit successivement les partitions P'_x pour x variant par degrés de 1 à 0. Les classes de la partition $P'_{0,99}$ sont construites par agglutination de sorte que si une classe possède un élément, elle possède tous ceux qui lui sont similaires pour s à plus de 0,99. Les classes de $P'_{0,98}$ sont construites à partir de celles de $P'_{0,99}$, une classe h s'agglutinant à une de h' si l'on trouve $k \in h$, et $k' \in h'$ similaires pour s à plus de 0,98. Et ainsi de suite...

La méthode de Sneath a pour elle sa clarté : la procédure de formation des classes ne fait intervenir aucune règle confuse d'optimisation, comme ce sera le cas pour les méthodes que nous présenterons ensuite. Mais il s'en faut souvent de beaucoup que s' soit l'indice ultramétrique le plus corrélé à s. Par enchainements successifs, on constitue souvent des classes où seul un lien ténu relie des points fort éloignés.

V.3.2. Méthode de Sørensen (1948) : Comme Sneath, Sørensen définit successivement les partitions P_x, pour x variant de 1 à 0 : mais le critère d'agglutination est différent. Soit à construire la partition $P'_{0,99}$ de J : on demande d'abord que deux éléments quelconques d'une même classe de P' aient un indice de similarité $\geqslant 0,99$, et l'on cherche une partition P' satisfaisant à cette condition, et comptant le moins de classes possible. Or ce problème n'admet pas en général de solution unique. Comme une discussion détaillée est pratiquement impossible, on se borne à procéder pas à pas : des agrégats en formation s'adjoignent les éléments encore isolés : quand un élément peut s'ajouter à deux agrégats on l'attribue de préférence au plus gros. Pour construire la partition $P'_{0,98}$, on opère sur $P'_{0,99}$ comme on a opéré sur J pour construire $P'_{0,99}$: les classes de $P'_{0,99}$ sont traitées comme des éléments, l'indice de similarité entre deux classes étant par définition la moyenne des indices de similarité entre leurs éléments. La méthode de Sørensen a l'inconvénient de n'être pas univoque ; mais elle a l'avantage de conduire à des classes très compactes et non filiformes, comme avec la méthode de Sneath.

V.3.3. Méthode de Sokal et Michener (1958) : On construit une hiérarchie indicée H', en commençant par les classes de plus haut indice. Ici, encore, on procède par agglutination, mais sans faire varier un indice x par degrés égaux. Diverses formules ont été proposées pour calculer comme une moyenne l'indice d'une classe, ou l'indice de similarité entre deux classes (certains calculs, assez complexes, ont recours à une formule donnant la corrélation entre deux sommes, en fonction des corrélations entre les termes). Nous croyons qu'il est convenable d'appeler indice d'une classe la moyenne des indices entre paires d'individus de la classe ; et indice de similarité entre deux classes la moyenne des indices entre paires d'individus l'un d'une classe, l'autre de l'autre. Eventuellement, on peut dans ces

calculs considérer une classe comme formée non d'éléments de l'ensemble J initial où l'on veut définir une hiérarchie, mais de sous-classes antérieurement constituées. Ceci dit, supposons qu'au cours du processus ascendant d'agglutination on ait déjà constitué q classes $h_1, \ldots, h_q$, d'indices $x_1, \ldots, x_q$: l'agglutination ne portera plus que sur celles de ces classes qui sont (à ce moment) maximales (i.e. qui ne sont contenues dans aucune autre déjà formée. L'agglutination peut se faire soit par paires, soit par groupes variables. Dans le premier cas, on forme à chaque étape une seule classe en rassemblant les deux classes maximales qui donnent une réunion de plus haut indice. Dans le second cas on réunit d'abord les paires les plus proches, mais on cherche à leur adjoindre d'autres classes, en imposant seulement la condition que l'indice de similarité du nouveau venu avec le groupe déjà formé ne soit pas inférieur de plus d'un certain seuil (e.g. 0,03) à l'indice de similarité intervenant dans la précédente adjonction.

La méthode de Sokal et Michener n'est pas plus exempte d'ambiguités que la méthode de Sørensen. Mais du point de vue de la compacité des classes elle semble occuper une position moyenne entre les deux méthodes déjà citées : la première de celles-ci se contente d'un lien unique, pour rattacher un élément à une classe ; la seconde veut que toute les paires de constituants d'une classe soient solidement liées ; ici, l'on considère la moyenne des liens. Toutefois, comme dans la méthode de Sørensen la proximité entre deux sous-classes est elle-même calculée comme une moyenne, il est prudent de dire que la méthode de S. et M. diffère surtout de celle de S. par des contingences techniques, telles que l'usage de coefficients de corrélation.

V.3.4. La méthode de Rogers et Tanimoto (1960)

Voici le principe de cette méthode : rechercher dans l'ensemble J les points de densité, c'est-à-dire les points qui sont entourés d'un aussi grand nombre possible de proches voisins) pour constituer autour d'eux les classes d'une première partition ; traiter alors l'ensemble de ces classes comme l'ensemble initial J pour constituer un nouvel étage de la hiérarchie ; et ainsi de suite, d'étage en étage. La méthode comporte des calculs assez complexes pour choisir les points de densité, et vérifier l'homogénéité des classes que l'on construit autour d'eux. Cette méthode semble assurer au taxinomiste un bon contrôle de la cohérence de ses constructions, mais elle laisse la place à beaucoup de variabilité dans l'application. Nous ne la décrirons pas en détails.

V.4. Clefs et algorithmes descendants.

V.4.1. Situer un individu à sa place dans la hiérarchie H des espèces, c'est déterminer la suite des sous-ensembles h de H qui le contiennent. Pratiquement la détermination se fait à partir des classes les plus grandes, e.g. on reconnaît un vertébré, qui est un poisson, de l'ordre des téléostéens etc. Il convient d'avoir une stratégie systématique pour rechercher des caractères qui permettent de décider de la place de l'individu : ces caractères sur lesquels se concentre l'attention sont les clefs. Si un sous-ensemble h et de H est monothétique on peut le définir par quelques caractères communs aux espèces de h, et totalement ou partiellement absents des es-

pèces qui ne sont pas dans h. Mais on sait qu'il existe des classes polythétiques (cf supra 11) qu'on ne peut définir mieux que par la présence de la plupart des caractères d'un certain sous-ensemble clef. Pour établir rationnellement un système de clefs, il convient donc de relever, pour chaque sous groupe de la hiérarchie, non seulement les caractères qui y sont toujours présents, mais aussi ceux qui y sont très fréquemment présents (e.g. pour les mammifères "avoir des poils ou des piquants", quoique les baleines soient glabres (conduit auditif excepté !) ; ou pour les vertébrés "avoir un système- urogénital", quoique ce ne soit pas le cas pour les poissons téléostéens). Il faut noter que dans certains sous-genres (spectres continus de bactéries et cas d'hybridation) les espèces forment un continuum ; le problème de la détermination n'est plus de trouver la vraie classe, mais de chiffrer la similarité relative de l'individu considéré à quelques espèces, points de repère bien étudiés (cf S & S p. 254 et 262). Mais dans l'ensemble, sur l'espace de toutes les combinaisons possibles de caractères, les êtres vivants apparaissent plutôt comme des îlots sporadiques, et c'est justement ce qui fait la possibilité et l'intérêt d'une classification hiérarchique (cf. 32)

V.4.2. Au contraire des algorithmes étudiés au n° 43, tant la détermination d'un individu que l'établissement d'un système de clefs suivent un mouvement descendant (cf 12, 13). La question se pose donc : est-il possible de construire aussi en descendant une hiérarchie de classification ? Quoique la réponse de S & S soit résolument négative, nous présenterons ici deux remarques en faveur des algorithmes descendants.

V.4.2.1. Un programme fonctionne à la Bell Corporation, qui fournit simultanément un système de clefs et une hiérarchie de classes. On suppose connues les probabilités relatives des espèces à classer et à reconnaître, ainsi qu'une matrice logique de description de l'ensemble J des espèces. Le programme donne (cf supra 13) l'ordre hiérarchique optimum dans lequel déterminer les caractères pour minimiser l'espérance mathématique du nombre de questions à poser pour reconnaître une espèce : ipso facto se trouve définie une classification par dichotomies successives de l'ensemble J. L'inconvénient de ce programme est qu'il vise à établir une classification purement monothétique ce qui (cf 11) n'est pas toujours dans l'ordre naturel ; de plus il traite une matrice logique, forme à laquelle toutes les descriptions ne se laissent pas réduire sans artifice (cf 2). Mais il montre qu'il est possible d'établir en descendant une hiérarchie de caractères et d'espèces.

V.4.2.2. Reprenons le problème d'optimisation posé en 53 : on peut tenter d'utiliser les algorithmes de S. Regnier pour en donner une solution approchée. La démarche serait la suivante :

1/ Déterminer une partition optima P de l'ensemble J.

2/ Organiser les classes de P en une hiérarchie stratifiée H : H comprend les parties à un seul élément $\{j\}$, les classes de P, certaines réunions finies de classes P ; ces réunions et leur stratification dans la hiérarchie H sont choisies de sorte que parmi toutes les hiérarchies de la forme particulière considérée ici (à partir de P) ce soit H qui réalise le meilleur accord avec l'ordonnance ω donnée sur J.

3/ A chacune des classes de P, appliquer l'algorithme de Regnier.

4/ Organiser au mieux, comme en 2), en une hiérarchie stratifiée, les nouveaux atomes obtenus.

5/ Partitionner les sous-classes construites en 5... etc.

Appliquée à l'exemple du n° 32 de notre leçon "sur les algorithmes de classification", cette méthode donnerait des résultats satisfaisants, reconnaissant d'abord les paires d'îlots, puis séparant les îlots à l'intérieur des paires.

Comme on le voit, le mouvement est ici principalement descendant (cf 1), 3), 5), etc) ; avec, à chaque pas, un retour (cf 2), 4)...) ou les classes que l'on vient de découper sont groupées en unités hiérarchiques d'ordre supérieur (2) ou seulement intermédiaire (4),...)

VI - L'ANALYSE FACTORIELLE EN TAXONOMIE

Quoique le règne animal ou le règne végétal soient des domaines d'élection pour les classifications discontinues, une représentation géométrique des relations de similarité entre espèce n'en est pas moins souvent souhaitable. Ainsi S & S reproduisent (p. 203) un modèle taxinomique des Enterobacteriacède, modèle dû à Lysenko et Sneath (1959), où les espèces sont représentées par des sphères que relient des tiges dont les longueurs indiquent les dissimilarités. Or, représenter géométriquement, dans un espace de faible dimension, des situations multidimensionnelles est l'objet propre de l'analyse factorielle, d'où le présent §.

VI.1. Analyse d'une matrice de scalaires quelconques

Comme nous l'avons dit au n° 32, les colonnes (individus) de la matrice de description peuvent être traités comme des fonctions numériques sur I, entre lesquelles on calcule des coefficients de corrélation. On cherche dans R^n un sous-espace F de dimension aussi faible que possible avec lequel tous les vecteurs représentant les individus fassent un petit angle : la situation sera représentée par projection orthogonale sur ce sous-espace F. Deux techniques principales servent à définir F et, remarquent S & S, (p. 195) "The principal components method is largely employed by British factor analysts, while multiple factor analysis with relation to simple structure is widely accepted in the United States". Là, F est défini par des vecteurs orthogonaux et unitaires que la méthode des moindres carrés détermine sans ambiguité. Ici, F est rapporté à des axes obliques, dont la recherche n'est pas sans artifice, mais qui ont l'avantage d'être choisis pour faire chacun un angle faible avec les vecteurs représentatifs d'un petit groupe d'individus. C'est pourquoi Rohlf et Sokal (1962), ont pu utiliser cette méthode d'analyse factorielle pour repartir 40 espèces d'abeilles en huit classes, dont chacune correspond à un axe auquel on agrège les vecteurs qui en sont le plus proche. Pour une discussion très complète des applications de l'analyse factorielle en écologie, citons Dagnelie (1960).

VI.2. Analyse des proximités : vu l'attention accordée aux indices de similarité (cf § 3) par les taxinomistes, il n'est que normal que les méthodes de R.N. Shepard y doivent trouver bientôt leur application ; mais à notre connaissance aucun travail n'a encore été publié qui trace ainsi la carte des espèces.

VI. 3. Analyse des correspondances : Pour l'analyse des matrices logiques ou des matrices logiques aléatoires, cette méthode nous paraît promettre plusieurs avantages.

VI. 3. 1. Caractères et espèces seraient représentés sur un même diagramme : l'analyse des correspondances est donc à la fois, pour user de la terminologie de S & S, (cf p. 124 et supra n° 5) une Q-technique et une R-technique.

VI. 3. 2. Le même diagramme donnerait à la fois les classes (îlots d'espèces) et les clefs qui permettent de les distinguer (caractères voisins de ces îlots). Toutefois, ici, comme d'ailleurs dans l'analyse factorielle usuelle, la reconnaissance d'un individu pourrait s'effectuer commodément même si une classification polythétique s'impose ou que l'individu lui-même est détérioré : On calculerait les valeurs des facteurs, qui ne sont que des sommes pondérées des caractères reconnus sur l'individu et celui-ci se trouverait situé dans le diagramme. Là où les objets à classer forment un continuum (écologie, bactériologie) pareille méthode est particulièrement souhaitable. On peut parler ici de diagnostic topologique, par opposition au dianostic logique. L'intuition du praticien ressemble sans doute plus à celui-ci qu'à celui-là !

VI. 3. 3. Les divers caractères seraient affectés de poids, (les valeurs des facteurs), fournis non par des considérations intuitives ou a priori mais par un calcul systématique. C'est là d'ailleurs un des avantages que S & S attendent de l'analyse des similarités entre caractères (cf p. 209).

VI. 3. 4. Représenter le détail de toutes les espèces dans un même espace euclidien n'est pas un objectif raisonnable : cela revient à rapporter aux mêmes axes des différences entre poissons et des différences entre mammifères : or rien chez ceux-ci ne correspond, par exemple à l'opposition sélaciens - téléostéens. Mais il est souhaitable, e.g. de situer spatialement les embranchements et les classes puis de définir pour chaque classe son espace propre et ainsi de suite... L'analyse des correspondances permet de construire une telle hiérarchie de plans : on extrait d'abord pour l'ensemble de la correspondance $I \times J$, un petit nombre de facteurs, et l'on en fait soit visuellement soit par un algorithme (cf supra § 5) une répartition en classe J_1, J_2,... des points figuratifs de diverses espèces. On analyse alors des correspondances partielles $I \times J_1$, $I \times J_2$..., les caractères sur lesquels tous les membres d'une classe s'accordent (e.g. pour les poissons, avoir des branchies) pouvant être éliminés de I sans rien changer au calcul des facteurs. C'est là une technique descendante.

Notons pour terminer qu'il peut être intéressant d'analyser non la matrice de description elle-même, mais la matrice qui s'en déduit en multipliant chaque colonne par un poids proportionnel à la fréquence de l'espèce qu'elle décrit.

BIBLIOGRAPHIE :

M. ADANSON - Histoire naturelle du Sénégal, Coquillages etc. Bauche ; Paris ; 1757.

M. BECKNER - The biological way of thought ; Columbia University Press ; New York ; 1959.

G. BOHN - Vertébrés inférieurs ; Actualités Sc et Ind. n° 183 ; Hermann ; Paris ; 1934.

P.J. CLARK - An extension of the coefficient of divergence for use with multiple characters. Cop , T2 pp. 61-64 ; 1952.

P. DAGNELIE - Contribution à l'étude des communautés végétales par l'analyse factorielle ; Bull. Serv. Carte Phytogéogr. (B). T5, pp. 7-71, 93-195 ; 1960.

L.R. DICE - Measures of the amount of ecologie association between species ; Ecology ; T.26) pp. 297-302 ; 1945.

P. JACCARD - Nouvelles recherches sur la distribution florale, Bull Soc. Vaud. Sci. Nat. T.44, pp. 223-270 ; 1908.

S. KULCZYNSKI - Die Pflanzenassoziationnen der Pieninen ; Bull. Inter. Acad. Pol. Sci. Lett. Cl. Sci. Math. Nat., B (SCi. Nat.) pp. 57-203, suppl. 2 ; 1927.

O. LYSENKO et P.H.A. SNEATH - The use of models in bacterial classification ; J. Gen. Microbiol. ? t.20, pp. 284-290 ; (1959).

A. OCHIAI - Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions : Bull. Jap. Soc. Sci. Fish. ; T 22, pp. 526-530 ; 1957.

K. PEARSON - On the coefficient of racial likeliness, Biometrika ; T. 18, pp. 105-117 ; 1926.

D.J. ROGERS et T.T. TANIMOTO - A computer program for classifying plants ; Science ; T 132 pp. 1115-1118 ; 1960.

P.F. RUSSEL et T.R. RAO - On habitat and association of species of anopheline larvae in south-eastern Madras ; J. Malar. Inst. India ; T 3, pp. 153-178, 1940.

G.G. SIMPSON - Principles of Animal Taxonomy ; Columbia Univerty Press ; New York ; 1961.

E. SMIRNOV - Matematische Studien über individuelle und Kongregationenvariabilität ; Verth. 5. Intern. Kong. Vorerbungswiss. T 2 pp. 1373-1392 ; 1927.

P.H.A. SNEATH - The application of computers to taxonomy ; J. Gen. Microbiol. T 17. pp. 201-226 ; 1957.

R.R. SOKAL et C.D. MICHENER - A statistical method for evaluating systematic relationships ; Univ. Kansas Sci. Bull. T 38, pp. 1409-1438 ; 1958.

R.R. SOKAL et P.H.A. SNEATH - Principles of numerical taxonomy ; W. H. Freeman et Cie ; San Francisco et London ; 1963.

T. SØRENSON - A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons ; Biol. Skr, T. 5(4) pp. 1-34 ; 1948.

F. VICQ D'AZYR - Quadrupèdes, Discours préliminaire. In Encyclopédie méthodique, vol 2 ; Panckoucke ; Paris ; 1792.