REVUE DE STATISTIQUE APPLIQUÉE

P. DAGNELIE

À propos des différentes méthodes de classification numérique

Revue de statistique appliquée, tome 14, n° 3 (1966), p. 55-75

<http://www.numdam.org/item?id=RSA_1966__14_3_55_0>

A PROPOS DES DIFFÉRENTES MÉTHODES DE CLASSIFICATION NUMÉRIQUE

P. DAGNELIE
Faculté des Sciences Agronomiques de Gembloux (Belgique)

I - INTRODUCTION ET RESUME

On confond volontiers, en parlant de classification, deux types différents de problèmes, liés à deux types différents de données. D'une part, on peut considérer un échantillon (ou une population) de n individus, pour chacun desquels on dispose des valeurs de p variables, et on souhaite subdiviser l'ensemble des n individus en un nombre limité, mais souvent mal défini, de groupes d'individus semblables. D'autre part, on peut considérer k + 1 échantillons comprenant respectivement n_1, n_2, ..., n_{k+1} individus, pour chacun desquels on dispose également des valeurs de p variables ; on sait que les k premiers échantillons proviennent de k populations différentes, et on souhaite déterminer en conséquence celle de ces populations dont provient le dernier échantillon. Le premier problème est réellement un problème de classification, en ce sens que l'on souhaite dans ce cas définir ou "faire" des classes ; le second problème est un problème de rangement ou de classement, en ce sens que l'on désire à ce stade ranger un ou plusieurs individus dans des classes préalablement définies.

De tout temps, les statisticiens se sont intéressés surtout au second problème, dont la résolution fait appel notamment à la notion de fonction discriminante. Le cas le plus simple est celui de trois échantillons, d'effectifs n_1, n_2 et 1, que l'on étudie dans le but d'attribuer l'unique individu constituant le troisième échantillon à l'une des deux populations dont proviennent les deux premiers échantillons (Fisher, 1936). On remarquera d'ailleurs que, très souvent, en français comme en anglais, ce problème est considéré comme le seul problème de classification (classification problem), alors qu'il nous semble plus correct de parler ici de classement ou de discrimination (discrimination, allocation ou assignment problem).

Le désintéressement des statisticiens pour le premier problème est d'autant plus surprenant que celui-ci se pose très fréquemment dans de nombreux domaines : en psychologie (classification de tests ou d'individus), en taxonomie (classification des organismes végétaux et animaux), en phytosociologie (définition et classification des communautés végétales), en pédologie (classification des sols), en météorologie (définition de types de temps), etc. Ce désintéressement s'explique vraisemblablement par le fait qu'il s'agit à première vue d'un problème mal défini, le nombre de classes n'étant généralement pas connu a priori, et aussi par le fait que les hypothèses habituelles, de normalité notamment, sont souvent difficilement admissibles ici.

Ce désintéressement des statisticiens pour un problème qui est fondamentalement de nature statistique et qui se pose quotidiennement à de nombreux hommes de science, a provoqué indirectement l'éclosion de nombreuses méthodes de classification plus ou moins acceptables, mais en tout cas élaborées en ordre dispersé. L'utilisation de plus en plus courante des moyens modernes de calcul et de traitement de l'information (machines mécanographiques et calculatrices électroniques) n'a fait qu'accentuer ce mouvement au cours des dernières années. On peut d'ailleurs citer, comme preuves de cette tendance, la publication d'un ouvrage entièrement consacré aux questions de taxonomie numérique (Sokal et Sneath, 1963), la naissance d'un bulletin d'information spécial intitulé Taxometrics et édité par L. R. Hill[1], et la fondation en Angleterre d'une Classification Society, dont le secrétariat est assuré par J. S. L. Gilmour[2], et qui publie depuis peu The Classification Society Bulletin .

Nous nous proposons d'exposer tout d'abord les principes de quelques méthodes de classification numérique, utilisées notamment en sociologie végétale et en taxonomie (paragraphe 2). Nous reprendrons ensuite le problème sous sa forme générale (paragraphe 3), et nous définirons les deux lignes de recherche qui nous paraissent actuellement les plus intéressantes à suivre (paragraphe 4). Enfin, nous illustrerons l'exposé théorique par un exemple (paragraphe 5), et nous terminerons par quelques conclusions (paragraphe 6).

II - QUELQUES METHODES DE CLASSIFICATION NUMERIQUE

2.1 - Méthodes utilisées en sociologie végétale

En matière phytosociologique, les individus considérés sont généralement des stations, en chacune desquelles a été effectué un relevé de végétation, et les variables sont des caractéristiques floristiques, par exemple l'abondance ou le recouvrement, ou tout simplement la présence ou l'absence des différentes espèces végétales. Le but poursuivi est d'établir une classification des différents relevés considérés en un petit nombre de types de végétation aussi homogènes que possible, ou encore, d'établir une classification des différentes espèces considérées en un petit nombre de groupes d'espèces de comportement semblable.

Sørensen (1948) aborde ce problème en procédant à des regroupements successifs, basés sur les degrés de similitude entre relevés ou, d'une manière plus générale, entre individus. Dans ce but, il calcule pour chaque couple de relevés la valeur d'un coefficient de similitude , égal au quotient du nombre d'espèces communes aux deux relevés par le nombre moyen d'espèces présentes dans chacun d'eux. Il regroupe ensuite, à un premier échelon, les différents relevés qui sont liés par des coefficients de similitude supérieurs à une valeur minimum donnée (par exemple 0,7 ou 70 %) ; puis il procède de même pour les différents groupes ainsi définis, en délimitant des unités de plus en plus vastes, correspondant à des coefficients de similitude de plus en plus bas.

Les résultats obtenus de la sorte peuvent être synthétisés graphiquement sous la forme de dendrogrammes. Nous en donnons un exemple

---

(1) Central Public Health Laboratory, Colindale Avenue, London N.W.9 (Grande-Bretagne).

(2) University Botanic Garden, Cambridge (Grande-Bretagne).

(figure 1), relatif à une partie des données considérées par Sørensen[1]. Les relevés 1 et 2 sont les premiers à être regroupés, car leur similitude est de 78 %. Ensuite, à un niveau supérieur à 60 %, on réunit d'une part les relevés 3 et 4 (62 %), et d'autre part les relevés 5 et 6 (61 %). De même, le relevé 7 vient se joindre au groupe déjà formé par les numéros 1 et 2, la similitude moyenne entre ceux-ci étant de 58 % ; et le processus est poursuivi jusqu'à son terme, le dernier regroupement étant effectué à un niveau de similitude de 25 %.

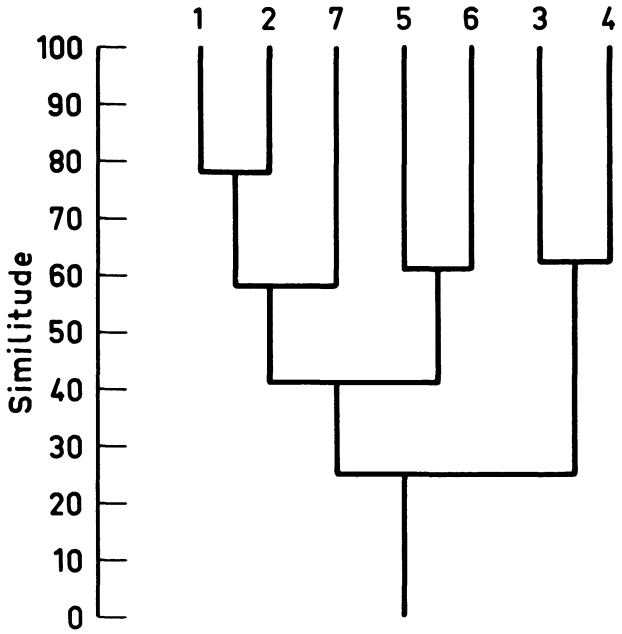

Figure 1 - Exemple de dendrogramme, relatif à des données de Sørensen (1948).

Goodall (1953) envisage le même problème en effectuant des subdivisions successives, et cela à partir des liaisons entre espèces ou, d'une manière plus générale, entre variables. Quatre méthodes distinctes sont en réalité proposées par Goodall : toutes nécessitent, pour chaque couple d'espèces, le calcul préalable d'un coefficient de liaison, qui peut être par exemple un coefficient de corrélation. Si certaines des valeurs ainsi obtenues sont significatives, à un niveau de probabilité préalablement fixé en fonction notamment du volume des données, on considère plus particulièrement les deux espèces correspondant au coefficient le plus élevé, et on effectue un premier partage en tenant compte d'une manière ou d'une autre de la présence de ces espèces. La première méthode proposée par Goodall consiste à réunir en un premier groupe les relevés contenant celle de ces deux espèces qui est la plus fréquente. Les coefficients de liaison interspécifique sont alors recalculés pour ce premier sous-ensemble de relevés et traités de la même manière jusqu'à l'obtention d'un sous-ensemble homogène, c'est-à-dire d'un groupe de relevés ne présentant plus de liaison significative entre les espèces. Tous les relevés qui ont été écartés sont ensuite repris et analysés progressive-

---

(1) Il s'agit en réalité des sept derniers relevés de végétation analysés par Sørensen : nos numéros 1 à 7 correspondent aux numéros 18, 19, 45, 46, 47, 48 et 50 du travail original (Sørensen, 1948).

ment de la même façon. Enfin, certains regroupements ultérieurs peuvent éventuellement être effectués.

Les autres méthodes proposées par Goodall ont pour principe de réunir chaque fois, au lieu des relevés contenant une des deux espèces les plus fortement corrélées, soit des relevés ne contenant pas cette espèce, soit les relevés contenant simultanément ces deux espèces, soit ceux qui ne contiennent aucune de ces deux espèces. Il semble d'ailleurs que, dans certaines limites, ces différentes méthodes conduisent à des résultats très semblables.

Tout comme Goodall, Williams et ses collaborateurs proposent de classer les relevés de végétation en partant des coefficients de liaison interspécifique (Williams et Lance, 1958 ; Williams et Lambert, 1959 et 1960). Après avoir réalisé différents essais, ces auteurs suggèrent d'additionner pour chaque espèce les valeurs absolues de tous les coefficients correspondants, et d'effectuer une première subdivision en tenant compte de la présence et de l'absence de l'espèce pour laquelle la somme ainsi obtenue est la plus élevée. Chacun des deux groupes de relevés ainsi définis est ensuite subdivisé de la même manière, en tenant compte des valeurs des coefficients de liaison observés à l'intérieur de ces groupes ; et ce processus est appliqué jusqu'à ce que toutes ou presque toutes les liaisons significatives aient disparu.

Dans une publication plus récente, les mêmes auteurs suggèrent également d'utiliser cette méthode pour définir, au lieu de groupes de relevés, des groupes d'espèces de comportement semblable (Williams et Lambert, 1961). Il faut alors partir des coefficients de similitude ou de corrélation entre les relevés, considérés comme les caractéristiques ou les variables, les différentes espèces étant considérées comme les individus à classer.

D'autres méthodes, analogues à celles de Sørensen, de Goodall et de Williams et Lambert ont été proposées, notamment par Fager (1957) et par Hopkins (1957). Ces dernières méthodes ont plus particulièrement pour but d'établir, à partir des liaisons interspécifiques, des groupes d'espèces étroitement corrélées. La valeur de ces diverses méthodes a été discutée notamment par Harberd (1960).

2.2 - Méthodes utilisées en taxonomie

Comme nous l'avons déjà signalé, les méthodes de taxonomie numérique ont fait l'objet d'un ouvrage particulier, dû à Sokal et Sneath. D'une manière générale, les méthodes en question ont pour but d'évaluer les similitudes existant entre différentes unités taxonomiques (des espèces animales ou végétales par exemple), et de classer ces unités en fonction des similitudes observées (Sneath et Sokal, 1962 ; Sokal et Sneath, 1963). Les méthodes proposées et décrites par Sneath et Sokal sont par ailleurs très semblables à celles utilisées en sociologie végétale.

S'intéressant particulièrement à la classification de microorganismes, Sneath (1957) détermine tout d'abord les pourcentages de caractères communs aux différents types de microbes, considérés deux à deux. A partir des coefficients de similitude ainsi obtenus, il procède ensuite à des regroupements successifs, selon un processus analogue à celui adopté par Sørensen en sociologie végétale. Toutefois, alors que Sørensen n'introduit dans un groupe que les individus ou les types qui sont suffisamment voisins de tous les membres de ce groupe, Sneath procède à un tel regroupement dès qu'un individu ou un type est très sem-

blable à un membre quelconque du groupe considéré. La classification ainsi réalisée peut également être présentée sous la forme d'un dendrogramme mais, pour un même degré de similitude, on doit s'attendre à définir de cette manière des groupes plus hétérogènes que ceux de Sørensen.

Les diverses méthodes proposées par Sokal et ses collaborateurs procèdent également par regroupements successifs, la base de départ habituelle étant ici la matrice des coefficients de corrélation entre caractères (Sokal et Michener, 1958). D'autre part, les règles adoptées par ces auteurs sont assez semblables à celles de Sørensen, en ce sens que l'introduction d'un individu dans un groupe dépend essentiellement du degré de similitude moyen de cet individu avec tous les membres du groupe.

Enfin, d'autres méthodes de taxonomie numérique ont également été proposées, notamment par Rogers et ses collaborateurs (Rogers et Tanimoto, 1960 ; Rogers et Fleming, 1964). Certaines de ces méthodes ont d'ailleurs été utilisées dans d'autres domaines que la classification des êtres vivants, telle que par exemple la classification des sols (Bidwell et Hole, 1964a et 1964b).

2.3 - Quelques méthodes utilisées dans d'autres domaines

Des problèmes de classification numérique se posent également en psychologie expérimentale, notamment en ce qui concerne le choix des tests. Souvent, il importe en effet d'introduire dans une même expérience, pour chacune des aptitudes humaines que l'on veut étudier, un groupe de plusieurs tests psychologiques donnant chacun une mesure de cette aptitude (Holzinger et Harman, 1941 ; Harman, 1960). La constitution de ces groupes peut être réalisée à l'aide d'un coefficient d'appartenance (coefficient of belonging, B-coefficient), qui est défini en fonction du rapport de la moyenne de tous les coefficients de corrélation des variables appartenant au groupe considéré, à la moyenne des coefficients de corrélation de ces variables avec les autres variables. En considérant comme point de départ du premier groupe le couple de variables de corrélation maximum, on peut calculer une première valeur du coefficient d'appartenance, et on doit s'attendre en général à ce que l'introduction de toute nouvelle variable dans ce groupe provoque une réduction de cette valeur. On procède alors à de telles introductions de variables supplémentaires tant que la diminution du coefficient d'appartenance n'est pas trop importante ; et on constitue ensuite d'autres groupes de la même manière, en partant des variables restantes qui sont les plus étroitement corrélées.

Cette méthode a été utilisée également en anthropologie, par Clements (1954).

Toujours dans le domaine psychologique, Thorndike (1953) envisage les questions de classification d'un certain nombre d'individus en un nombre restreint de groupes, à partir des différences ou des distances existant entre ces individus. Il propose de prendre comme points de départ des différents groupes les individus les plus dissemblables, et d'attribuer à tour de rôle à chacun des groupes l'individu qui est le plus proche de ceux dejà contenus dans le groupe considéré. On obtient ainsi, à une unité près, des groupes de même effectif ; et on peut procéder ensuite à des transferts d'un groupe à l'autre, de manière à réduire autant que possible la variabilité existant à l'intérieur des groupes.

La diversité des méthodes de classification numérique ressort nettement des paragraphes précédents, qui n'en donnent cependant qu'une idée fort incomplète. Nous aurions en effet pu citer également certains travaux récents, tels ceux de Bonner (1964), d'Edwards et Cavalli-Sforza (1965), de Fortier et Solomon (1965), de Gyllenberg (1963), de Hill et al. (1965), de Macnaughton-Smith et al. (1964), de Schnell (1964) et de Van Den Driessche (1965), ou d'autres travaux plus anciens, antérieurs même à la notion de taxonomie numérique, tels ceux de Cattell (1944) et de Tryon (1939). Nous aurions pu parler aussi des nombreuses méthodes proposées par Mc Quitty, et de leurs différentes versions : agreement analysis, elementary linkage analysis, hierarchical linkage analysis, hierarchical syndrome analysis, comprehensive hierarchical analysis, typal analysis, rank order typal analysis, linkage analysis, etc ; nous ne citerons à ce propos qu'une seule référence, à savoir Mc Quitty (1964).

Cette énumération, même très incomplète, des méthodes de classification numérique, met en évidence l'intérêt considérable porté aux problèmes de classification par de très nombreux chercheurs, appartenant à des disciplines elles-mêmes très variées. En fait, il s'avère souvent indispensable, dans de nombreux domaines, d'établir une classification des individus étudiés, même si cette classification est arbitraire et, de ce fait, éminemment contestable. Tel est le cas par exemple en matière d'étude de la végétation ou des sols, lorsque le but poursuivi est l'établissement d'une carte de la végétation ou d'une carte des sols. Que les différences existant d'un type de végétation à l'autre ou d'un type de sol à l'autre soient très importantes ou très réduites, que l'on se trouve ou non en présence de nombreux individus intermédiaires entre les types principaux, il est indispensable, avant toute cartographie, de procéder à la définition des types à cartographier.

III - DISCUSSION GENERALE DU PROBLEME

3.1 - Exposé du problème

La diversité des méthodes de classification résulte évidemment du fait que des chercheurs appartenant à des disciplines différentes ont oeuvré indépendamment les uns des autres ; mais cette diversité provient aussi de ce que les problèmes considérés ne sont pas identiquement les mêmes dans tous les cas. Aussi nous paraît-il intéressant de préciser ces problèmes, et leurs différentes variantes, avant de comparer les méthodes.

D'une manière générale, on considère un ensemble de n individus, pour chacun desquels on a observé p variables. Dans certains cas, les individus constituent, ou sont considérés comme constituant un échantillon prélevé au hasard dans une population plus étendue : c'est généralement le cas pour des relevés de végétation, des organismes végétaux ou animaux, ou des individus soumis à des tests psychologiques. Mais les individus étudiés peuvent également être parfaitement spécifiés, et constituer l'ensemble de la population qui est prise en considération : c'est le cas pour des types de végétation préalablement définis, pour des variétés ou des espèces végétales ou animales connues, pour des groupes sociologiques donnés, etc.

Cette distinction, qui s'apparente à celle intervenant dans la définition des modèles fixes et aléatoires de l'analyse de la variance, est fondamentale. Lorsque les individus sont aléatoires, la définition des classes, ou groupes d'individus, se confond simplement avec la subdivision d'un

espace à p dimensions, dans lequel on peut représenter les n individus, en autant de régions qu'il y a de classes. Par contre, lorsque les individus sont fixes, c'est-à-dire lorsque l'on a affaire à des types préalablement définis, le problème se complique généralement par le fait que l'on ne recherche pas seulement une classification des types, mais aussi des informations relatives aux relations existant entre ces types, et impliquant généralement une certaine hiérarchisation, ou une certaine interprétation des critères de classification. Les problèmes de classification de types ou de groupes d'individus viennent donc logiquement après ceux de classification des individus eux-mêmes, en groupes ou en types : en taxonomie par exemple, la classification des espèces ou des genres en familles, en ordres, etc, ne peut se faire qu'après la classification des espèces ou des genres.

D'autre part, nous avons vu également que le but poursuivi pouvait être d'établir une classification des caractères, et non pas des individus, ainsi que cela se présente notamment en sociologie végétale (classification des espèces) et en psychologie (classification des tests).

3.2 - Comparaison théorique de quelques méthodes

En ce qui concerne plus particulièrement les méthodes, les différents problèmes peuvent être abordés soit en partant des liaisons entre variables, soit en partant des similitudes entre relevés. Dans chaque cas, on est amené aussi à choisir un des nombreux coefficients de liaison et de similitude proposés : ce choix dépend notamment du caractère qualitatif ou quantitatif des données considérées (Dagnelie, 1960 ; Goodman et Kruskal, 1959 ; Sokal et Sneath, 1963). Enfin, on doit distinguer les méthodes procédant par regroupements progressifs, à partir de noyaux peu importants, de celles procédant par subdivisions successives, conduisant à la formation de groupes de plus en plus restreints et de plus en plus nombreux.

En tenant compte de ces diverses alternatives (classification des individus ou des caractères, individus fixes ou aléatoires, etc.), nous avons condensé en un tableau les caractéristiques essentielles de quelques méthodes, choisies parmi les plus importantes (tableau 1). Ces informations doivent cependant être interprétées avec prudence, comme nous allons le souligner en passant en revue une nouvelle fois les différentes caractéristiques.

Déjà au sujet de la première alternative, qui concerne la classification des individus (I) ou des caractères (C), des doutes peuvent se présenter. La distinction entre individus et caractères est en effet assez conventionnelle. En sociologie végétale par exemple, on peut admettre que la présence de telle espèce dans telle station permet aussi bien de caractériser les qualités de cette station, l'espèce étant alors le caractère, que les exigences de l'espèce considérée, la station étant alors le caractère. Nous avons en réalité adopté chaque fois le point de vue qui nous paraissait le plus logique ou le plus important.

En ce qui concerne le caractère fixe (F) ou aléatoire (A) des individus, nous avons noté autant que possible le point de vue adopté par les auteurs au cours de leurs premiers travaux, bien que de nombreuses méthodes puissent être utilisées indifféremment dans les deux cas. D'une manière générale, on remarquera qu'en taxonomie (paragraphe 2.2), les individus sont pratiquement toujours considérés comme fixes, le problème étant essentiellement d'établir une classification ou une hiérarchisation de races, de variétés ou d'espèces préalablement définies.

Tableau 1

Principales caractéristiques de quelques méthodes de classification numérique.

Auteurs et références	Classifications des individus (I) ou des caractères (C)	Individus fixes (F) ou aléatoires (A)	Données qualitatives (0/1) ou quantitatives (X)	Liaisons entre caractères (C) ou similitudes entre individus (I)	Coefficient de corrélation (R), distances (D) ou autres paramètres (A)	Regroupements (R) ou subdivisions (S)
BONNER (1964)	I	A	0/1	I	A	R
EDWARDS et al. (1965)	I	F	0/1	I	D	S
FAGER (1957)	C	A	0/1	C	A	R
GOODALL (1953)	I	A	0/1	C	R	S
HOLZINGER et al. (1941)	C	A	X	C	R	R
HOPKINS (1957)	C	A	0/1	C	R	R
MACNAUGHTON-SMITH et al. (1964)	I	A	0/1	I	D	R
ROGERS et al. (1960, 1964)	I	A	0/1	I	D	R
SNEATH (1957)	I	F	0/1	I	A	R
SOKAL et al. (1958)	I	F	X	I	R	R
SØRENSEN (1948)	I	A	0/1	I	A	R
THORNDIKE (1953)	I	F	X	I	D	R
VAN DEN DRIESSCHE (1965)	I	F	X	I	D	R
WILLIAMS et al. (1958, 1959, 1960)	I	A	0/1	C	R	S
WILLIAMS et al. (1961)	C	A	0/1	I	R	S

Quant à la nature des données, nous avons tenu compte dans chaque cas des données qui interviennent réellement dans l'analyse, après une éventuelle codification. Il est en effet fréquent que des variables de nature fondamentalement quantitative soient transformées en variables alternatives, ne pouvant prendre que les valeurs conventionnelles 0 et 1. De telles variables ont été groupées avec les données qualitatives proprement dites et, comme elles, désignées par le symbole "0/1". Par contre, les variables de nature quantitative, intervenant comme telles dans l'analyse, ont été marquées d'un X. Il s'agit évidemment, ici aussi, des données considérées par les auteurs au cours de leurs premiers travaux, la plupart des méthodes pouvant être adaptées aux différents types de données.

Au sujet des différents paramètres qui peuvent être calculés pour mesurer soit les liaisons ou les corrélations entre caractères (C), soit les similitudes, les distances ou les corrélations entre individus (I), nous avons désigné par R le coefficient de corrélation et ses dérivés immédiats (coefficient de corrélation de point et variable χ^2, pour des données qualitatives), par D les différentes notions de distance (diffé-

rences moyennes, distances généralisées au sens de Mahalanobis, etc.), et par A les autres paramètres utilisés.

Enfin, nous avons désigné respectivement par R et par S les méthodes procédant respectivement par regroupements et par subdivisions. Il faut noter cependant que certaines méthodes basées essentiellement sur le principe des subdivisions successives peuvent se terminer par quelques regroupements de classes voisines.

Bien qu'il ne fasse intervenir que six caractéristiques, le tableau ainsi dressé montre que les méthodes envisagées sont dans l'ensemble très différentes les unes des autres. Seules quelques méthodes sont identiques aux six points de vue considérés : tel est le cas par exemple pour les méthodes de Goodall et de Williams et al. d'une part, de Thorndike et de Van Den Driessche d'autre part.

IV - LES PRINCIPALES POSSIBILITES DE RECHERCHE

4.1 - Deux lignes de recherche

En présence de ces nombreuses méthodes de classification numérique, il nous semble indispensable, pour progresser réellement dans ce domaine, de commencer par en débrouiller l'écheveau. Deux lignes de recherche différentes peuvent être suivies dans ce but : l'une consiste à comparer les méthodes existantes, l'autre à rechercher immédiatement une méthode optimale.

De toute façon, il s'avère nécessaire de se fixer préalablement un ou plusieurs critères d'optimalité, sans qu'il soit cependant possible de choisir un critère unique, valable dans toutes les situations rencontrées. Le but poursuivi le plus souvent étant de définir des classes d'individus (ou de caractères) aussi homogènes que possible, il nous semble justifié d'utiliser surtout des critères d'homogénéité, basés par exemple sur les notions de distance généralisée ou de variance généralisée (Anderson, 1958 ; Kendall, 1957). Une classification optimale serait celle qui assure, pour un nombre donné de classes, soit une valeur moyenne minimum des distances généralisées entre individus à l'intérieur des classes, soit une valeur moyenne maximum des distances généralisées entre classes, soit une valeur minimum de la variance généralisée dans les classes, soit une valeur maximum de la variance généralisée entre classes, ces quatre critères étant étroitement liés les uns aux autres.

En relation avec une analyse de la variance à plusieurs variables (analyse de la dispersion), le critère d'optimalité pourrait être également un rapport de deux variances généralisées : soit le rapport de la variance factorielle (entre classes) à la variance résiduelle (dans les classes), soit le rapport de la variance factorielle ou de la variance résiduelle à la variance totale.

Certains auteurs, dont Edwards et Cavalli-Sforza (1965), ont déjà utilisé certains de ces critères. D'autres, notamment Sneath (1957), ont proposé des critères différents, basés par exemple sur l'emploi de l'un ou l'autre coefficient de similitude, mais de tels critères ont l'inconvénient de ne pas tenir compte des liaisons pouvant exister entre les caractères étudiés.

Il peut être utile également, notamment si l'on utilise la notion de distance généralisée, d'effectuer avant toute analyse une standardisation des variables. Les résultats obtenus, et notamment les distances calcu-

lées, dépendent en effet des unités utilisées pour les différentes variables, et cet inconvénient peut être éliminé notamment en partant des variables réduites. Au lieu d'effectuer cette réduction variable par variable, on peut également opérer une transformation globale des variables, en soumettant la matrice de corrélation à l'analyse des composantes. Cette façon de faire a, en outre, le double avantage de conduire d'une part à l'utilisation de variables non corrélées, et de permettre d'autre part une réduction souvent importante du nombre de variables prises en considération.

4.2 - La comparaison des méthodes existantes

Quelques publications ont été consacrées, totalement ou partiellement, à la comparaison de diverses méthodes de classification numérique. Il en est ainsi par exemple des travaux de Bonner (1964) et de Goodall (1953), qui présentent, en les comparant, plusieurs méthodes de classification. De même, Beers et al. (1962), ainsi que Hill et al. (1965), ont comparé diverses méthodes de classification.

D'autre part, Sokal et Rohlf (1962) ont proposé une méthode de comparaison de différents dendrogrammes provenant des mêmes données. Toutefois, si cette méthode permet d'évaluer la similitude existant entre deux ou plusieurs dendrogrammes, elle ne fournit aucune information quant au fait que certains des dendrogrammes considérés sont "meilleurs" que les autres.

Enfin, Sokal et Sneath (1963) signalent plusieurs publications relatives à la comparaison des paramètres, et discutent eux-mêmes les mérites de plusieurs méthodes de classification. Rohlf et Sokal (1965) ont d'ailleurs consacré un travail récent au même sujet.

Toutes ces recherches sont cependant de portée relativement restreinte, en ce sens que les comparaisons réalisées portent seulement sur un petit nombre d'exemples, souvent un seul, et ne font intervenir qu'un petit nombre de méthodes. Il serait utile en réalité de comparer entre elles, autant que possible, toutes les méthodes qui poursuivent un même but, en traitant par ces différentes méthodes un nombre aussi élevé que possible d'ensembles de données, réelles ou hypothétiques. La comparaison devrait porter non seulement sur l'un ou l'autre critère d'optimalité, tel que ceux qui ont été cités ci-dessus, mais aussi sur d'autres considérations, comme l'importance des calculs nécessités par les diverses méthodes. En effet, les moyens modernes de calcul permettent l'emploi de méthodes chaque jour plus complexes, mais le coût d'utilisation de ces moyens de calcul ne peut en aucune façon être négligé.

Une étude comparative telle que celle que nous préconisons de réaliser représenterait évidemment un travail considérable, qu'il serait utile de répartir entre différents chercheurs ou différents groupes de chercheurs : dans ce domaine, une collaboration organisée serait certainement préférable à l'élaboration désordonnée de méthodes de plus en plus nombreuses.

4.3. - La recherche d'une solution optimale

Diverses tentatives de recherche d'une solution optimale, au sens défini ci-dessus, ont été réalisées. Edwards et Cavalli-Sforza (1965), par exemple, proposent d'effectuer la première subdivision de l'ensemble des individus considérés de manière à assurer une variance maximum entre classes : d'une façon générale, l'application de ce principe nécessite

cependant, pour n individus, la comparaison des variances correspondant aux $2^{n-1} - 1$ solutions possibles. Ces auteurs suggèrent également de poursuivre de la sorte, par subdivisions successives, tout en reconnaissant qu'il ne faut pas s'attendre à obtenir ainsi une classification optimale, même si chacune des subdivisions, considérée individuellement, est optimale.

Le principe adopté par Edwards et Cavalli-Sforza dans le cas d'une dichotomie simple, pourrait théoriquement être étendu à une subdivision plus complexe, en un nombre quelconque de classes. Toutefois, le nombre de solutions à comparer augmente très rapidement en fonction du nombre d'individus n et du nombre de classes m, de telle sorte que, pratiquement, le problème devient rapidement insoluble. Le tableau 2 donne les nombres de solutions, ou leur ordre de grandeur, pour quelques valeurs de n et de m. Ces résultats ont été obtenus à l'aide de règles simples d'analyse combinatoire, et contrôlés en utilisant notamment les propriétés données indépendamment par Edwards et Cavalli-Sforza (1965) et par Fortier et Solomon (1965).

Tableau 2

Nombre de solutions théoriquement possibles (ou ordre de grandeur de ce nombre de solutions), pour différents nombres d'individus (n) et différents nombres de classes (m).

n \ m	2	3	4	5
5	15	25	10	1
10	511	9.330	34.105	42.525
20	524.287	580.606.446	45.232.115.901	749.206.090.500
50	10^{15}	10^{23}	10^{29}	10^{33}
100	10^{30}	10^{47}	10^{59}	10^{68}

Selon Edwards et Cavalli-Sforza, le nombre de solutions $S_{n,m}$ est égal à $n!/m!$ fois le coefficient de x^n dans le développement en série de $(e^x - 1)^m$, c'est-à-dire d'une manière générale :

$$S_{n,m} = \frac{1}{m!} \sum_{i=0}^{m-1} (-1)^i \binom{m}{i} (m - i)^n,$$

et en particulier, pour m = 3 :

$$S_{n,3} = (3^{n-1} - 2^n + 1)/2,$$

pour m = 4 :

$$S_{n,4} = [4^{n-1} - 3^n + 3(2^{n-1}) - 1]/6,$$

et pour m = 5 :

$$S_{n,5} = [5^{n-1} - 4^n + 2(3^n) - 2^{n+1} + 1]/24.$$

Pour n suffisamment grand par rapport à m, l'ordre de grandeur du nombre de solutions est, d'une manière générale :

$m^n/m!$

Fortier et Solomon donnent d'autre part une formule générale plus complexe, conduisant aux mêmes résultats, ainsi que la formule de récurrence suivante :

$$S_{n,m} = m\, S_{n-1,m} + S_{n-1,m-1}.$$

S'il s'avère donc vite impossible, ou pratiquement impossible de comparer toutes les solutions, on peut cependant envisager la possibilité de comparer celles de ces solutions qui sont réellement admissibles, compte tenu de la disposition des points dans l'espace à p dimensions. Dans le cas d'une seule variable (p = 1), les différentes solutions admissibles peuvent être obtenues en choisissant, sur l'axe correspondant, autant de points que l'on désire former de classes, moins un. Ces points peuvent être choisis indifféremment, mais à raison d'un seul par intervalle, dans les n – 1 intervalles existant entre les individus : le nombre de solutions admissibles est donc égal à $\binom{n-1}{m-1}$. Toute autre solution est inadmissible, car elle supposerait qu'au moins un individu appartenant à une classe donnée est compris entre deux individus appartenant à une même autre classe. Par comparaison avec le tableau 2, le tableau 3 montre que la réduction ainsi opérée est considérable.

Tableau 3

Nombre de solutions admissibles, pour différents nombres d'individus (n) et différents nombres de classes (m), dans le cas d'une variable (p = 1).

n \ m	2	3	4	5
5	4	6	4	1
10	9	36	84	126
20	19	171	969	3.876
50	49	1.176	18.424	211.876
100	99	4.851	156.849	3.764.376

Comme nous le verrons plus loin par un exemple (paragraphe 5.4), ce principe peut être étendu facilement aux cas les plus simples à deux variables (p = 2). Dans un espace à deux dimensions, on peut en effet écarter a priori toutes les solutions telles qu'un individu d'une classe se trouve situé dans un triangle dont les trois sommets correspondent à des individus d'une même autre classe. Pour cinq individus (n = 5) et deux classes (m = 2), par exemple, que les points formant le contour extérieur de l'ensemble soient les sommets d'un triangle, d'un quadrilatère ou d'un pentagone, on peut toujours montrer que seules 10 des 15 solutions à envisager sont réellement admissibles. Par extension, il pourrait être possible de limiter d'une manière générale le nombre de solutions à prendre en considération et, peut-être, d'aboutir à un algorithme utilisable. Tout comme la méthode du simplexe en programmation linéaire, cet algorithme pourrait être basé essentiellement sur l'étude d'ensembles convexes : d'une manière générale, on doit en effet écarter a priori toute solution telle qu'un individu d'une classe se trouve à l'intérieur d'un

polyèdre convexe dont les sommets correspondent à des individus d'une même autre classe.

En l'absence de solution optimale, il pourrait être utile de rechercher des solutions suboptimales, en utilisant l'une ou l'autre des méthodes citées ci-dessus. On peut penser par exemple que la méthode de Thorndike (paragraphe 2.3) doit donner des résultats assez voisins de l'optimum, mais cette méthode devrait pouvoir être encore améliorée en procédant comme suit. Après avoir choisi, pour définir m classes, les m points les plus éloignés les uns des autres, c'est-à-dire les m individus associés par exemple à la plus grande variance, on pourrait s'efforcer d'introduire chacun des n - m individus restants dans chacun des m embryons de groupes ainsi formés, et on adopterait la solution la meilleure. On procèderait ensuite de même avec les n - m - 1 individus non classés, puis avec les n - m - 2 individus toujours disponibles, etc. Une telle procédure nécessiterait, au départ, la comparaison de $\binom{n}{m}$ solutions et, au cours des étapes successives, de m(n - m), m(n - m - 1), m(n - m - 2), ... solutions. Le nombre total de solutions à comparer est donc :

$$\binom{n}{m} + m\binom{n - m + 1}{2} ;$$

et le tableau 4 montre que les nombres de solutions ainsi obtenus deviennent rapidement très inférieurs aux nombres de solutions existant a priori (tableau 2).

Tableau 4

Nombre de solutions à comparer pour établir une classification "suboptimale", pour différents nombres d'individus (n) et différents nombres de classes (m).

n \ m	2	3	4	5
5	22	19	9	1
10	117	204	294	327
20	532	1,599	5.389	16.104
50	3.577	22,984	234.624	2.123.935
100	14.652	175.959	3.939.849	75.310.320

Dans le même ordre d'idée, il faut signaler l'essai réalisé par Solomon et Fortier (1965). Utilisant le coefficient d'appartenance dont il a été question au cours du paragraphe 2.3, ces auteurs ont tenté de déterminer une solution suboptimale en procédant à un échantillonnage de l'ensemble des solutions possibles. Mais ce procédé s'est avéré assez décevant, en raison du fait que les solutions optimales et suboptimales sont nécessairement excentriques.

La recherche d'une solution optimale ou suboptimale soulève également le problème de la détermination du nombre de classes à établir . En réalité, il faut s'attendre à ce que ce nombre de classes dépende souvent de considérations pratiques, plus ou moins arbitraires. On peut envisager néanmoins la possibilité de fixer d'une manière relativement

objective la fin du processus de classification, par exemple en procédant à diverses classifications successives, correspondant à des nombres de classes croissants, jusqu'à ce que l'accentuation de l'homogénéité des classes (réduction de variance dans les classes ou augmentation de variance entre les classes) ne soit plus suffisante. Thorndike (1953) donne quelques informations à ce sujet.

D'autre part, quand on dispose de plusieurs observations par individu, ou quand on considère des types d'individus et que l'on dispose de plusieurs individus par type, il est possible de déterminer de façon objective le nombre de subdivisions à effectuer, en procédant à des tests d'hypothèses, tels que ceux fournis par l'analyse de la variance à plusieurs variables. Rao (1952) présente un problème de ce type.

V - EXEMPLE

5.1 - Données

Nous illustrerons les principes qui viennent d'être exposés par un exemple particulièrement simple, choisi de manière à bien mettre en évidence les principaux points saillants. Cet exemple ne concerne en réalité que cinq individus et trois variables ($n = 5$ et $p = 3$).

Le tableau 5 présente tout d'abord les données initiales : elles sont relatives à la composition granulométrique des sols de cinq stations, appartenant à un ensemble plus vaste que nous avons déjà étudié par ailleurs (Dagnelie, 1965). Le même tableau contient aussi les données réduites correspondantes, et les données transformées orthogonales, déduites de l'analyse des composantes principales. Les variances attachées à des différentes composantes sont respectivement 2,294, 0,654, et 0,052, soit 76,5, 21,8 et 1,7 % du total.

Tableau 5.

Données initiales, données réduites et composantes principales.

Individus	Variables initiales			Variables réduites			Composantes principales		
	1	2	3	1	2	3	1	2	3
1	63,0	16,5	9,0	0,212	0,374	0,138	0,188	0,190	0,365
2	48,8	23,0	14,0	-0,984	1,340	1,287	2,074	0,339	0,048
3	50,3	14,2	2,5	-0,857	0,033	-1,356	-0,151	-1,596	-0,071
4	58,5	14,0	12,0	-0,167	0,003	0,827	0,507	0,580	-0,343
5	81,8	2,2	4,5	1,796	-1,750	-0,896	-2,618	0,487	0,001

5.2 - Classifications optimales obtenues à partir des données réduites

Le tableau 2 montre que, pour identifier la subdivision optimale en deux classes, il faut en principe prendre en considération les quinze solutions possibles, que l'on peut désigner de la manière suivante :

(1)(2, 3, 4, 5), (2)(1, 3, 4, 5), (3)(1, 2, 4, 5), (4)(1, 2, 3, 5), (5)(1, 2, 3, 4),
(1, 2)(3, 4, 5), (1, 3)(2, 4, 5), (1, 4)(2, 3, 5), (1, 5)(2, 3, 4), (2, 3)(1, 4, 5),
(2, 4)(1, 3, 5), (2, 5)(1, 3, 4), (3, 4)(1, 2, 5), (3, 5)(1, 2, 4), (4, 5)(1, 2, 3),

La comparaison des sommes de carrés d'écarts entre classes,

relatives à ces quinze solutions et calculées à partir des trois variables réduites, donne comme valeur maximum :

$$SCE_{max} = (0,212 - 0,984 - 0,857 - 0,167)^2/4 + 1,796^2 + (0,374 + 1,340 + 0,033 + 0,003)^2/4 + (-1,750)^2 + (0,138 + 1,287 - 1,356 + 0,827)^2/4 + (-0,896)^2$$

$$= \frac{5}{4}\,(1,796^2 + 1,750^2 + 0,896^2) = 8,86,$$

et ce maximum correspond à la solution :

$$(1,2,3,4)\ (5).$$

Cette dernière subdivision peut donc être considérée comme optimale, au sens défini précédemment.

De même, pour trois classes, en considérant les 25 solutions possibles, on obtient la valeur maximale :

$$SCE_{max} = (0,212 - 0,984 - 0,167)^2/3 + (-0,857)^2 + 1,796^2 + (0,374 + 1,340 + 0,003)^2/3 + 0,033^2 + (-1,750)^2 + (0,138 + 1,287 + 0,827)^2 + (-1,356)^2 + (-0,896)^2 = 12,63,$$

correspondant à la subdivision :

$$(1,2,4)\ (3)\ (5)$$

Enfin, pour quatre classes, on obtient aussi :

$$SCE_{max} = 14,62,$$

et la solution :

$$(1,4)\ (2)\ (3)\ (5)\ ;$$

tandis que la subdivision en cinq classes conduit évidemment, pour $n = 5$ et $p = 3$, à la valeur limite

$$np = 15.$$

Ces différents résultats sont résumés dans le tableau 6 et la figure 2, les sommes de carrés d'écarts entre classes y étant exprimées en pour-cent de cette valeur limite.

5.3 - Classifications optimales et suboptimales obtenues à partir des composantes principales

En tenant compte seulement de la première composante principale, à laquelle est attachée 76,5 % de la variation totale, on obtient pour deux classes la même solution que ci-dessus, avec une somme de carrés d'écarts légèrement inférieurs :

$$SCE_{max} = 8,57.$$

Tableau 6.

Résultats de différents processus de classification : solutions et sommes des carrés des écarts (en % du total), en fonction du nombre de classes, pour trois variables (ou trois composantes principales), pour la première composante principale, et pour les deux premières composantes principales.

Nombre de classes (m)	Variables réd. Solution	1 à 3 SCE (%)	Compos. princ. Solution	1 SCE (%)	Compos. princ. Solution	1 et 2 SCE (%)
1	(1,2,3,4,5)	0	(1,2,3,4,5)	0	(1,2,3,4,5)	0
2	(1,2,3,4)(5)	59,1	(1,2,3,4)(5)	57,1	(1,2,3,4)(5)	59,1
3	(1,2,4)(3)(5)	84,2	(1,3,4)(2)(5)	75,0	(1,2,4)(3)(5)	84,2
4	(1,4)(2)(3)(5)	97,5	(1,4)(2)(3)(5)	76,1	(1,4)(2)(3)(5)	97,4
5	(1)(2)(3)(4)(5)	100	(1)(2)(3)(4)(5)	76,5	(1)(2)(3)(4)(5)	98,3

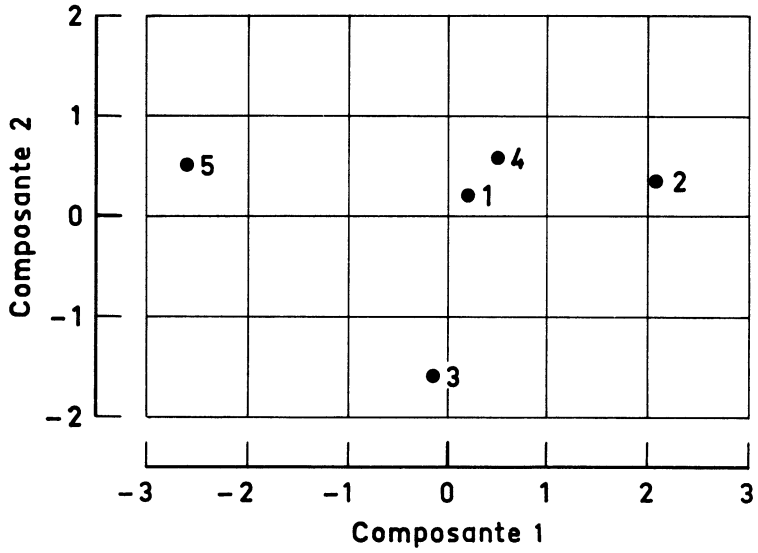

Figure 2 - Représentation graphique des résultats de différents processus de classification : sommes des carrés des écarts (en % du total), en fonction du nombre de classes (m), pour trois variables ou trois composantes (3), pour deux composantes (2) et pour une composante (1).

Par contre, pour trois classes, la solution optimale relative à la première composante principale serait :

$$(1,3,4)\ (2)\ (5),$$

avec une somme de carrés d'écarts :

$$SCE_{max} = 11,25.$$

Pour l'ensemble des trois variables, cette solution vient en réalité immédiatement après la subdivision optimale :

$$(1,2,4)\ (3)\ (5),$$

avec une somme de carrés d'écarts :

$$SCE = 11,84,$$

à peine inférieure au maximum correspondant :

$$SCE_{max} = 12,63.$$

On observe ainsi, à propos d'un exemple, qu'une solution optimale relative à un nombre réduit de composantes principales peut être suboptimale relativement à l'ensemble des variables.

Enfin, pour quatre classes, et toujours pour la première composante seulement, on retrouve également la solution rencontrée précédemment :

$$(1,4)(2)(3)(5),$$

avec une somme de carrés d'écarts :

$$SCE_{max} = 11,42.$$

D'autre part, si on considère simultanément les deux premières composantes principales, les résultats obtenus sont, à tous les stades, pratiquement identiques à ceux déduits de l'ensemble des trois variables (tableau 6 et figure 2). Cette constatation n'est pas surprenante si l'on se souvient que les deux premières composantes justifient la quasi-totalité de la variation :

$$76,5 + 21,8 = 98,3\ \%.$$

Enfin, pour l'ensemble des trois composantes principales, les résultats obtenus sont nécessairement identiques à ceux déduits des trois variables initiales.

5.4 - Utilisation de la notion de solution admissible

Les résultats relatifs à la première ou aux deux premières composantes principales peuvent être obtenus plus rapidement en ne comparant entre elles que les solutions réellement admissibles, et en tenant compte dans ce but des positions respectives des différents individus. Ces positions sont données par la figure 3, dans le plan des deux premières composantes.

Pour une seule variable, ou une seule composante, quatre solutions seulement sont admissibles (tableau 3). Ce sont ici :

$$(1,2,3,4)(5),\quad (1,2,4)(3,5),\quad (2,4)(1,3,5)\quad \text{et}\quad (2)(1,3,4,5),$$

la solution (1, 2)(3, 4, 5), par exemple, étant inadmissible du fait que l'individu 1 est compris entre les individus 3 et 4, le long du premier axe principal.

Pour deux variables ou deux composantes, dix solutions seulement doivent être comparées. Ce sont les solutions :

(2)(1, 3, 4, 5), (3)(1, 2, 4, 5), (4)(1, 2, 3, 5), (5)(1, 2, 3, 4), (2, 3)(1, 4, 5), etc.,

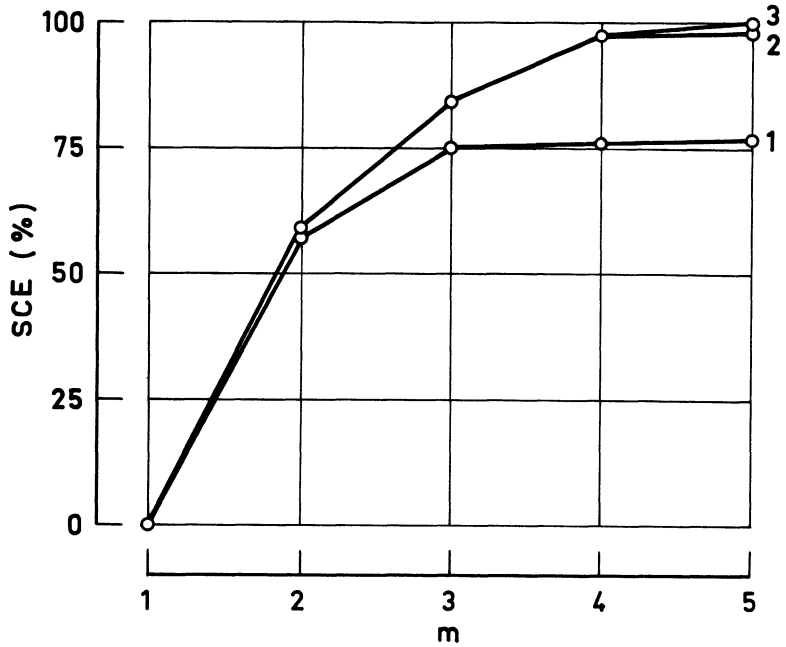

Figure 3 - Répartition des cinq points observés dans le plan des deux premières composantes principales.

des solutions telles que :

$$(1)(2,3,4,5) \quad \text{et} \quad (1,2)(3,4,5)$$

étant inadmissibles du fait que le point 1 se trouve dans le triangle de sommets 3, 4 et 5.

5.5 - Application de la méthode de Thorndike modifiée

La méthode de Thorndike, modifiée comme nous l'avons dit au paragraphe 4.3, ne réduit pas sensiblement le travail dans un cas aussi simple que celui qui est considéré ici. Aussi ne ferons-nous qu'en exposer le principe, dans le cas d'une subdivision en trois classes par exemple.

Au premier stade, il faut choisir les trois points les plus éloignés les uns des autres. Si l'on ne tient pas compte des positions relatives des différents individus, ce choix nécessite le calcul et la comparaison des sommes de carrés d'écarts relatives aux dix combinaisons des points pris 3 à 3 :

$$\binom{n}{m} = \binom{5}{3} = 10.$$

Au deuxième stade, on tente d'ajouter chacun des deux points restants à chacun des trois noyaux qui viennent d'être définis et on adopte la solution la plus favorable : cette deuxième étape nécessite, d'une manière générale, le calcul et la comparaison de six sommes de carrés d'écarts. Enfin, il ne reste qu'un point à ajouter, de la manière la plus favorable, à l'un des trois noyaux, ce qui demande encore la comparaison de trois solutions. Au total, ce processus nécessite donc bien le calcul et la comparaison de 19 sommes de carrés d'écarts, au lieu de 25 (tableaux 2 et 4).

VI - CONCLUSIONS

En conclusion, nous tenons à souligner les points essentiels suivants :

1/ l'importance réelle des méthodes de classification numérique, pour de très nombreux chercheurs oeuvrant dans des domaines fort différents ;

2/ la diversité des problèmes rencontrés, et la diversité encore plus grande des méthodes de classification qui ont été proposées ;

3/ les inconvénients qui résultent de cette diversité des méthodes, et d'une telle dispersion des efforts ;

4/ la nécessité de procéder à une comparaison plus systématique de ces méthodes ;

5/ l'intérêt que présenterait l'obtention d'un processus de détermination des solutions optimales, ou suboptimales.

Même des résultats partiels pourraient être utiles dans ce domaine. La mise au point d'une méthode de résolution qui ne serait applicable que pour un petit nombre de variables, par exemple, pourrait rendre de grands services, dans la mesure où l'on envisage la possibilité de réduire, au départ, le nombre de variables par l'analyse des composantes : une solution optimale pour les variables transformées pourrait être considérée comme suboptimale pour les variables initiales. De même, l'obtention d'une solution optimale pour un nombre restreint de cas réels permettrait de préciser la valeur des différentes méthodes proposées. A ce point de vue, il serait donc particulièrement justifié de poursuivre simultanément des recherches selon les deux lignes que nous avons suggérées.

REMERCIEMENTS

L'essentiel de ce travail a pu être réalisé au cours d'un séjour de dix semaines à l'Université d'Aberdeen (Ecosse). A l'occasion de cette publication, nous tenons à remercier l'University Court et le Professeur D. J. FINNEY, de leur aimable invitation et de leur excellent accueil.

REFERENCES

ANDERSON, T. W. - An introduction to multivariate statistical analysis. New York, Wiley, 1958, 374 p.

BEERS, R. J., FISHER, J., MEGRAW, S. et LOCKHART, W. R. - A comparison of methods for computer taxonomy. Jour. Gen. Microbiol. 28, 641-652, 1962.

BIDWELL, O. W. et HOLE, F. D. - Numerical taxonomy and soil classification. Soil Sci. 58-62, 1964a.

BIDWELL, O. W. et HOLE, F. D. - An experiment in the numerical classification of some Kansas soils. Proc. Soil Sci. Soc. Amer. 28, 263-268, 1964b.

BONNER, R. E. - On some clustering techniques. I. B. M. Jour. Res. Devel. 8, 22-32, 1964.

CATTELL, R. B. - A note on correlation clusters and cluster search methods. Psychometrika 9, 169-184, 1944.

CLEMENTS, F. E. - Use of cluster analysis with anthropological data. Amer. Anthropologist 56, 180-199, 1954.

DAGNELIE, P. - Contribution à l'étude des communautés végétales par l'analyse factorielle. Bull. Serv. Carte Phytogéogr., sér. B, 5, 7-71, 1960.

DAGNELIE, P. - L'étude des communautés végétales par l'analyse statistique des liaisons entre les espèces et les variables écologiques ; un exemple. Biometrics 21, 890-907, 1965.

EDWARDS, A. W. F. et CAVALLI-SFORZA, L. L. - A method for cluster analysis. Biometrics 21, 362-375, 1965.

FAGER, E. W. - Determination and analysis of recurrent groups. Ecology 38, 586-595, 1957.

FISHER, R. A. - The use of multiple measurements in taxonomic problems. Ann. Eugen. 7, 179-188, 1936.

FORTIER, J. J. et SOLOMON, H. - Clustering procedures. International Symposium on Multivariate Analysis, Dayton (Ohio), 1965, 24 p.

GOODALL, D. W. - Objective methods for the classification of vegetation. I. The use of positive interspecific correlation. Austral. Jour. Bot. 1, 39-63, 1953.

GOODMAN, L. A. et KRUSKAL, W. H. - Measures of association for cross classifications. II. Further discussion and references. Jour. Amer. Stat. Assoc. 54, 123-163, 1959.

GYLLENBERG, H. - A general method for deriving determination schemes for random collections of microbial isolates. Ann. Acad. Scient. Fenn., Ser. A IV, 69, 1963, 23 p.

HARBERD, D. J. - Association-analysis in plant communities. Nature 185, 53-54, 1960.

HARMAN, H. H. - Modern factor analysis. Chicago, Univ. Press, 1960, 469 p.

HILL, L. R., SILVESTRI, L. G., IHM, P., FARCHI, G. et LANCIANI, P. - Automatic classification of staphylococci by principal component analysis and a gradient method. Jour. Bacteriol. 89, 1393-1401, 1965.

HOLZINGER, K. J. et HARMAN, H. H. - Factor analysis : a synthesis of factorial methods. Chicago, Univ. Press, 1941, 417 p.

HOPKINS, B. - Pattern in the plant community. Jour. Ecol. 45, 451-463, 1957.

KENDALL, M. G. - A course in multivariate analysis. London, Griffin, 1957, 185 p.

MACNAUGTON-SMITH, P., WILLIAMS, W. T., DALE, M. B. et MOCKETT, L. G. - Dissimilarity analysis : a new technique of hierarchical subdivision. Nature 202, 1034-1035, 1965.

Mc QUITTY, L. L. - Capabilities and improvements of linkage analysis as a clustering method. Educ. Psychol. Measurement 24, 441-456, 1964.

RAO, C. R. - Advanced statistical methods in biometric research. New York, Wiley, 1952, 390 p.

ROGERS, D. J. et FLEMING, H. - A computer program for classifying plants. II. A numerical handling of non-numerical data. Bioscience 14, 15-28, 1964.

ROGERS, D. J. et TAMIMOTO, T. T. - A computer program for classifying plants. Science 132, 1115-1118, 1960.

ROHLF, F. J. et SOKAL, R. R. - Coefficients of correlation and distance in numerical taxonomy. Univ. Kansas Sci. Bull. 45, 1-27, 1965.

SCHNELL, P. - Eine Methode zur Auffindung von Gruppen. Biom. Zeitschr. 6, 47-48, 1964.

SNEATH, P. H. A. - The application of computers to taxonomy. Jour. Gen. Microbiol. 17, 201-226, 1957.

SNEATH, P. H. A. et SOKAL, R. R. - Numerical taxonomy. Nature 193, 855-860, 1962.

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

SOKAL, R. R. et ROHLF, F. J. - The comparison of dendrograms by objective methods. Taxon 11, 33-40, 1962.

SOKAL, R. R. et SNEATH, P. H. A. - Principles of numerical taxonomy. San Francisco and London, Freeman and C°, 1963, 359 p.

SØRENSEN, T. - A method of establishing groups of equal amplitude in plant sociology based on similarity of species content. Konge Dan. Vidensk. Selsk., Biol. Skr. 5, 1-34, 1948.

THORNDIKE, R. L. - Who belongs in the family ? Psychometrika 18, 267-276, 1953.

TRYON, R. C. - Cluster analysis. Ann Arbor. Edwards, 1939, 122 p.

VAN DEN DRIESSCHE, R. - La recherche des constellations de groupes à partir des distances généralisées D^2 de MAHALANOBIS. Biom. Praxim. 6, 36-47, 1965.

WILLIAMS, W. T. et LAMBERT, J. M. - Multivariate methods in plant ecology. I. Association-analysis in plant communities. Jour. Ecol. 47, 83-101, 1959.

WILLIAMS, W. T. et LAMBERT, J. M. - Multivariate methods in plant ecology. II. The use of an electronic digital computer for association-analysis. Jour. Ecol. 48, 689-710, 1960.

WILLIAMS, W. T. et LAMBERT, J. M. - Multivariate methods in plant ecology. III. Inverse association-analysis. Jour. Ecol. 49, 717-729, 1961.

WILLIAMS, W. T. et LANCE, G. N. - Automatic subdivision of associated populations. Nature 182, 1755, 1958.