Statistique et analyse des données

J. Jacq
J. M. Roblin
M. Defayolle

Taxonomie numérique et modèles d'organisation structurelle en biologie

Statistique et analyse des données, tome 2, n° 3 (1977), p. 96-106.
<http://www.numdam.org/item?id=SAD_1977__2_3_96_0>

Statistique et Analyse des Do...
3 - 1977

taxonomie numérique et modèles d'organisation structurelle en biologie

J A C Q J., R O B L I N J.M. , D E F A Y O L L E M.

Centre de Recherches du Service de Santé des Armées, Division de Psychologie, 108 Bd Pinel, F 69003 LYON

SUMMARY.

Numerical Taxonomy refers to those methods which classify sets of individuals, described by a set of variables, into clusters.

The usual models in Numerical Taxonomy often have an implicit and subjective notion of underlying clusters. While, the notion of "Natural classes" is frequently employed by biologists, the concepts of Biologists and Taxonomists do not necessarily cover the same entitres.

On the basis of the data the measures and selection processes used, selection of different algorithms leads to a model of structural organisation, which is most often based on the notion of class with a centroïd tendency. Since these models are exceptional in Biology the results are sometimes difficult to explain and create a certain perplexity among researchers. OUr findings permit us to elaborate a number of different models of structural organization which are better adapted to the study of biological phenomena than existing ones.

Le but de la taxonomie est de regrouper en un certain nombre de classes un ensemble d'individus décrits par un ensemble de variables ou de caractères. Si cette formulation est satisfaisante pour le théoricien, elle ne saurait toutefois recouvrir tous les aspects qui préoccupent l'utilisateur ; ce dernier s'intéresse en effet plus au résultat de la classification qu'à la méthode employée. Cet utilisateur, qu'il soit naturaliste, biologiste, psychologue, ne se contentera pas uniquement de constater qu'un algorithme lui fournit effectivement le meilleur ensemble de classes en fonction de différentes hypothèses, prémisses ou critères sous-jacents parfois implicites. En réalité, son approche se situe dans un espace bien plus vaste que celui induit par le tableau de données "individus X variables". Même s'il a l'assurance que l'ensemble des classes obtenues par une procédure soit le meilleur, du point de vue théorique en fonction de l'échantillonnage des individus et des variables, il cherchera à donner un sens à ces classes et cela à partir des divers éléments qui caractérisent, soit le phénomène étudié, soit les théories sous-jacentes aux différentes disciplines concernées.

Il arrive souvent que les résultats obtenus apportent une certaine confirmation de l'idée sous-jacente au travail demandé, mais il est aussi fréquent qu'ils posent certains problèmes d'interprétation. Les quelques réflexions développées ici résultent précisément de la démarche que nous avons entreprise pour essayer d'apporter une réponse à ces problèmes. Il nous a paru plus direct de les situer à partir des travaux qui ont été à l'origine de ces études et nous aborderons plus particulièrement deux catégories de problèmes qui concernent respectivement la microbiologie d'une part, et la biologie humaine ou les sciences humaines de l'autre.

1 - Quelques aspects des études taxonomiques en microbiologie.

La microbiologie représente un domaine d'application particulièrement fécond pour la taxonomie, et cela n'est certainement pas étranger à la publication, dès 1963, de l'ouvrage de SOKAL et SNEATH "Numerical Taxonomy". En 1969, J. BUISSIERE nous soumis un jeux de données concernant la tribu des KLEBSIELLAE. BUISSIERE avait mis au point une micro-méthode permettant d'effectuer simultanément un nombre de réactions biochimiques relativement important ; en assurant ainsi le recueil systématique de tous les caractères sur toutes les souches étudiées, il devenait possible de constituer des jeux de données susceptibles de se prêter à une approche multivariée. Le traitement de ces données par un programme de classification fit apparaître des classes correspondant aux espèces définies par les microbiologistes. Toutefois, l'année suivante, l'analyse par la même procédure, d'un échantillon comprenant 74 souches de LACTOBACILLE ne devait pas fournir des résultats aussi évidents. Bien que cette analyse fit apparaître une structure en 3 phylums correspondant aux BETABACTERIUM, THERMOBACTERIUM et aux STREPTOBACTERIUM, il fut particulièrement délicat de mettre en évidence des classes correspondant aux espèces notemment pour les THERMOBACTERIUM. Le fait que plusieurs espèces semblaient particulièrement voisines devait nous amener à remettre en cause l'hypothèse de l'existence de classes "centroïdes" constituées par des groupements d'individus séparés par des espaces plus ou moins importants. Par ailleurs, il est bien connu que les microorganismes sont plus sujets aux mutations que les organismes plus évolués, cette cause de variabilité des espèces était donc susceptible d'expliquer la proximité, voir l'imbrication de certaines espèces. La constitution de classes de type centroïde n'était donc pas particulièrement adaptée à la description de phénomène de cette nature.

Cette constatation devait nous amener à examiner sous un autre aspect la technique décrite sous le nom de "single linkage"(littéralement simple enchaînement) dans l'ouvrage de SOKAL et SNEATH. Cette technique avait été fréquemment utilisée par les microbiologistes, et différents auteurs lui

reprochait d'être sensible à l'effet de chaîne. Les diverses difficultés d'application et les recherches relatives à sa mise en œuvre par ordinateurs sont sant doute à l'origine des approfondissements théoriques réalisés notamment en France par ROUX et en Angleterre par les chercheurs de ROTHAMSTEAD.

On doit à ces derniers d'avoir mis au point un algorithme d'élaboration des classes par la construction intermédiaire d'un arbre de longueur minimale. Cette étape intermédiaire nous a paru propice à la description et à la mise en évidence de structures chaînées sur la base des considérations précédemment énoncées. La figure N° 1 illustre pour les LACTOBACILLES, l'apport respectif des deux méthodes, l'intérêt de la méthode du chaînage est de faire mieux apparaître la structure du phénomène étudié en mettant notamment en évidence les formes intermédiaires.

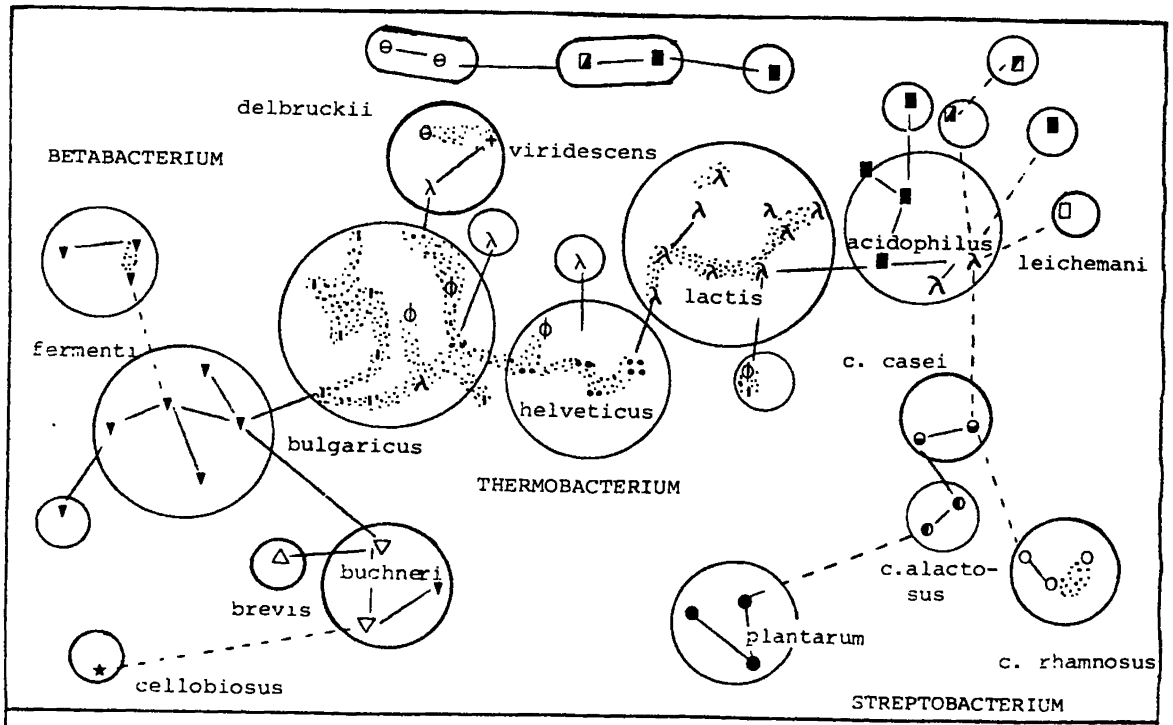

Figure 1 : Représentation comparée des résultats obtenus par la méthode centroïde et la méthode du chainage sur un échantillon de Lactobacilles.

2 - Quelques aspects des études taxonomiques concernant les problèmes humains.

Le classement de sujets en groupes plus ou moins homogènes, en fonction de divers symptômes ou de caractéristiques de la personnalité, représente une démarche du médecin, du psychologue ou du psychiatre, et a donné lieu à plusieurs travaux.

En 1971, le professeur COLOBERT, Directeur du laboratoire de Biologie Clinique de l'Hôpital Cardiologique de LYON, nous demanda d'essayer de classer un échantillon de 150 sujets décrits par 21 variables biochimiques. Cet échantillon était composé de trois groupes d'effectifs égaux, correspondant à des sujets "présumés normaux", hypertendus et athéromateux ; ce classement résultant des mentions portées sur les dossiers des malades.

En fait, dans de très nombreux cas, le classement des sujets en hypertendus et athéromateux se révélait assez délicat, il était donc possible que ce classement puisse recouvrir une partition plus fine, correspondant par exemple à 4 ou 5 classes.

L'application des diverses méthodes d'analyse univariée ou multivariée, relatée dans la thèse d'ALBRIEUX, ne devait pas contribuer à faire avancer la question ; si ce n'est en permettant de réaliser une représentation des sujets dans l'espace des 2 ou 3 premiers facteurs d'une analyse factorielle en composantes principales. Il était donc rationnel de recourir à une analyse taxonomique. Les résultats obtenus furent assez inattendus, car l'application du processus conduisait à isoler des groupes composés d'un seul sujet, le reliquat des sujets restant se trouvant dans un groupe unique (ainsi, pour 20 groupes demandés, on obtenait 19 groupes de 1 et un groupe de 131, et ainsi de suite). Il est évident que ces résultats ne coïncidaient pas avec l'hypothèse émise à l'origine de l'étude. Un examen plus attentif, avec étude des dossiers cliniques, permit de remarquer que les sujets isolés en premiers étaient en fait des sujets gravement atteints et que les sujets présumés normaux "se trouvaient toujours dans le groupe le plus important.

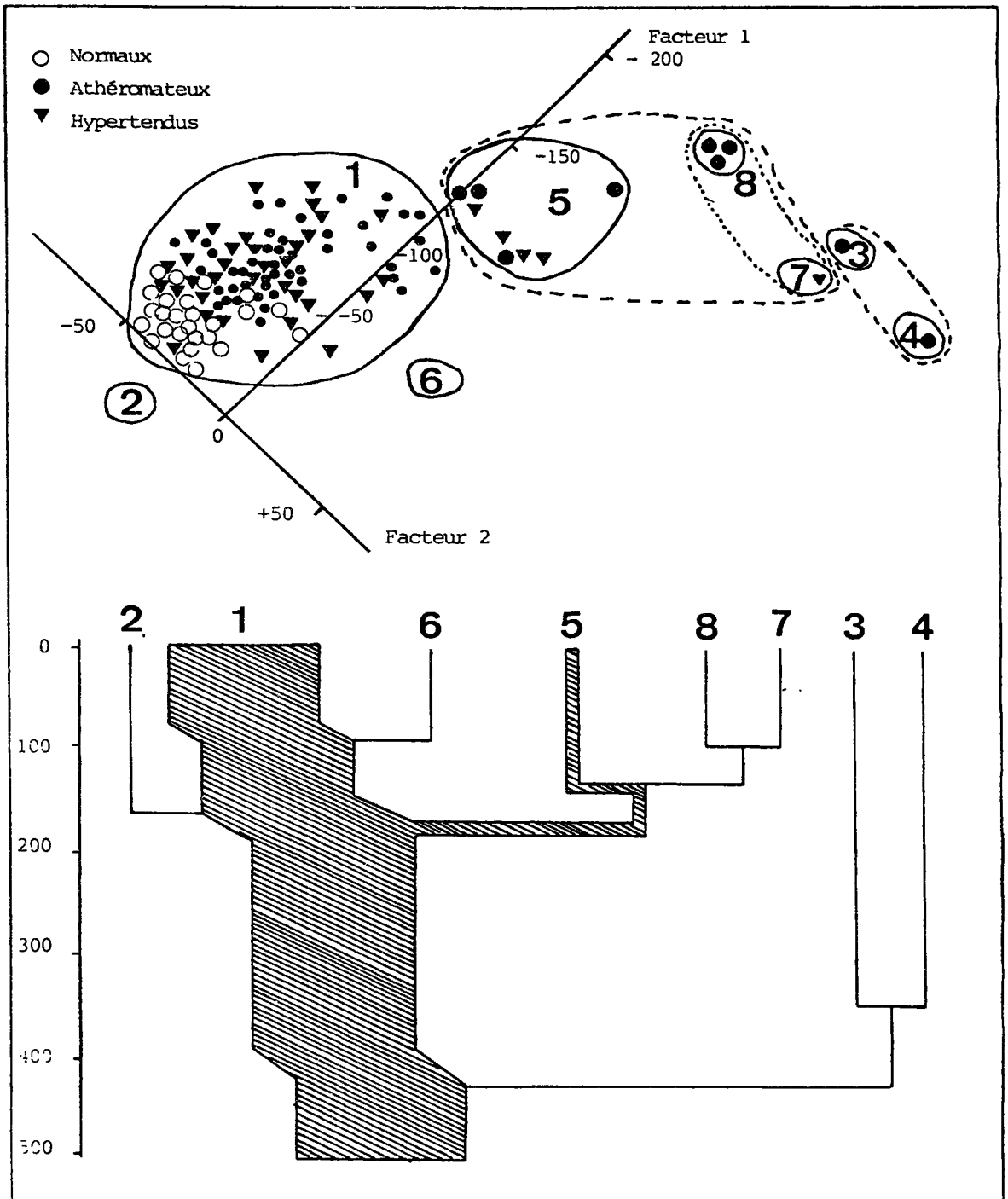

Figure 2 : Classification de 150 sujets en fonction des notes en facteur calculées à partir de 21 variables biochimiques.

A - Représentation des sujets dans l'espace des deux premiers facteurs.

B - Classification obtenue à partir des cinq premiers facteurs. Les groupes correspondant aux branches du dendrogramme sont représentés sur le schéma A.

Les représentations effectuées à partir des résultats de l'analyse factorielle permirent de mieux "visualiser" la structure en faisant apparaître un groupe de sujets "présumés normaux" relativement compact à partir duquel s'étirent les sujets pathologiques (figure N°2). Dès lors, l'objectif de l'étude fut décomposé en deux phases : d'abord comment séparer des sujets pathologiques de sujet "normaux", ensuite comment classer des sujets pathologiques. La première étape conduit à une approche du dépistage susceptible d'être appliquée dans divers domaines (la médecine du travail, par exemple) ; la deuxième concerne plus spécifiquement l'aide au diagnostic. Pour tenir compte des aspects particuliers concernant une maladie évolutive, telle que l'hypertension, il était indispensable de construire un modèle dans lequel les sujets "normaux" constituent une classe centroïde et où les sujets pathologiques se caractérisent par l'écart de cette classe ; le type de pathologie se traduisant par la direction de l'écart et la gravité par son importance. C'est ce que nous avons essayé de faire avec la classification directionnelle. Cette méthode fut appliquée en 1973 à l'étude des maladies caractérisant les gaz du sang (acidose, alcalose, etc...).

On pourrait penser qu'une telle structure à tendance centroïde soit un cas d'espèce, mais divers travaux devaient nous amener à constater qu'elle se retrouvait fréquemment dans les problèmes humains, qu'ils relèvent de la biologie humaine ou des sciences humaines.

Pour ces dernières, l'application de tels modèles implique que l'on ait construit au préalable un espace métrique convenable, ce qui pose divers problèmes dans le cas des données qualitatives inhérentes aux sciences humaines. Une telle approche réalisée récemment dans une étude relative à la conduite suicidaire (JACQ 1976) semble toutefois mettre en évidence l'existence de cette structure centroïde.

3 - DISCUSSION.

La formulation intuitive de BECKNER (1959) du concept de classe naturelle : "une telle classe se réfère à un ensemble fini d'attributs tel que chaque élément de la classe possède une proportion importante, mais non spécifiée, de ces attributs et réciproquement chaque attribut est possédé par une proportion importante des éléments de la classe, sans qu'il soit nécessaire qu'un des attributs soit possédé par tous les éléments de la classe" (LERMAN 1970). Correspond-elle à la majorité des structures biologiques ? Lorsque le modèle ne s'adapte pas à la réalité, le théoricien incrimine le choix des variables ou des individus, le biologiste a tendance à ne prendre que les aspects qui confirment ses hypothèses à priori ; n'est-il pas plus rationnel de chercher les raisons de ces divergences ? Comme le rappellent CARLES J. et CASSAGNES P. dans leur ouvrage sur l'origine des espèces : "Les individus existent dans la nature, les classes n'existent que dans notre esprit". Certes, une vision manichéenne du monde et une tendance naturelle de simplification des phénomènes complexes, nous conduit à concevoir et à construire des classes bien disjointes, combien plus satisfaisantes pour notre esprit. Il faut toutefois noter une tendance à élaborer des modèles ou des théories susceptibles de permettre une meilleure description de la réalité biologique, il en est ainsi de la méthode des nuées dynamique qui introduit la notion de formes fortes et faibles. Le développement de la théorie des sous-ensembles flous élaborée par ZADEH aux Etats Unis et diffusée par les travaux de KAUFMAN en France va également dans ce sens. Certes les "capacités opératoires" actuelles de cette théorie paraîssent faibles par rapport à d'autres modèles, mais il faut tenir compte du fait qu'elle inclut dans son corps d'hypothèse des éléments fondamentaux de la biologie et des sciences humaines et qu'elle est donc, par conséquent, plus générale que la théorie des ensembles. Il arrive, en effet, fréquemment que l'appartenance d'un élément à un ensemble ne puisse être donnée en tout ou rien.

CONCLUSION.

Il ne suffit pas qu'un biologiste constate un phénomène pour que le degré de connaissance évolué, il faut encore que la culture lui fournisse des techniques et des outils adaptés pour l'interprétation de ce phénomène. L'étude des structures, commencée au début de ce siècle avec les factorialistes et poursuivie plus récemment par les travaux des taxonomistes, représente une part importante de l'analyse des données. L'attention des biologistes doit toutefois être attirée sur le fait que les modèles, quels qu'ils soient, ne permettent jamais que la mise en évidence plus ou moins parfaite de structures sous jacentes inhérentes à leurs corps d'hypothèses. Ainsi, un modèle de classification de type centroïde fournira des classes centroïdes ; il incombe au biologiste de déterminer si ces dernières sont artificielles ou correspondent à une réalité biologique. Comme l'écrit TOMASSONE, les transformations détruisent la structure de référence, et on pourrait même ajouter qu'elles reconstruisent une structure à leur image. Pour toutes ces raisons, une approche, réalisée à partir de différents modèles, est seule susceptible de traduire la complexité inhérente du monde biologique. Il arrive parfois, à un moment donné, que les modèles disponibles ne prennent pas en compte certains aspects primordiaux des phénomènes étudiés ; il est alors nécessaire de s'en dégager en essayant d'arriver à une formulation plus concrète et plus réaliste, c'est ce que nous avons essayé de faire pour les problèmes qui se sont posés à nous.

REFERENCES BIBLIOGRAPHIQUES

- CARLES J. ; CASSAGNES P.

  L'origine des espèces - 1972 - P.U.F.

- CROCQ L. ; DUCOTTET F. ; DUSSUYER J. ; JACQ J. ; MAIGROT J.C. ; VEDRINE J. ; QUENARD

  Methods of evaluating individual and collective factors related to suicid behavior. IX INTERNATIONAL CONGRESS On Suicid Prevention and Crisis Intervention - HELSINKI Juin 1977, 15 pages.

- DIDAY E. ;

  Nouvelles méthodes et nouveaux concepts en classification automatique et reconnaissance de formes. Thèse de doctorat d'Etat es Sciences Mathématiques - Université PARIS VI - 1972

- JACQ J.

  Contribution à l'étude de la formation des classes en biologie. Doctorat de Spécialité de Sciences Biologiques - Université Claude Bernard 1972 LYON 1.59

- JACQ J. ; ALBRIEUX M.J. ; LALARD J.M. ; COLOBERT L. ; DEFAYOLLE M.

  Elaboration d'un modèle de classification directionnelle - Application à létude des maladies à caractère évolutif. Colloque IRIA - 1973, 18 p.

- JACQ J. ; DEFAYOLLE M.

  La classification directionnelle - Application aux problèmes de sélection du personnel et d'aide au diagnostic. Travaux Scientifiques CRSSA 1973 p. 136-138

- JACQ J. ; BUISSIERE J. ; SOUM S. ; DEFAYOLLE M.

  Méthodes de classification appliquées à l'étude de l'organisation structurelle des microorganismes. Travaux Scientifiques CRSSA 1974, p.231-233

- JACQ J.

  Evaluation des facteurs latents de la conduite suicidaire. Mémoire de Maîtrise de Recherches du Service de Santé des Armées (76 pages)

- KAUFMAN A.

  Introduction à la théorie des sous-ensembles flous (4 vol.) MASSON 1975, 1976, 1977

- ROUX M.

  Deux algorithmes de recherches en classification automatique. Revue de Statistique Appliquée - 1970 Vol. XVIII, N° 4

- SOKAL R.M. ; SNEATH P.H.A.

Principles of Numerical Taxonomy W.H. Freeman - San Francisco-London 1963.

- TOMASSONE R.

Analyse multidimensionnelle et classification - Revue de Statistique Appliquée - 1970 - Vol. XVIII N° 4