STATISTIQUE ET ANALYSE DES DONNÉES

CHRISTOPHE PERRUCHET

Une analyse bibliographique des épreuves de classifiabilité en analyse des données

Statistique et analyse des données, tome 8, n° 2 (1983), p. 18-41
<http://www.numdam.org/item?id=SAD_1983__8_2_18_0>

Statistiques et Analyse de données .18.
1983 - Vol. 8 n°2 - pp. 18-41

UNE ANALYSE BIBLIOGRAPHIQUE DES EPREUVES DE CLASSIFIABILITE EN ANALYSE DES DONNEES

Christophe PERRUCHET

Département de Mathématiques Appliquées pour les Télécommunications et l'Informatique

Centre National d'Etudes des Télécommunications
92131 ISSY-LES-MOULINEAUX

Résumé : *On présente ici une analyse critique des travaux existant dans le domaine de la classifiabilité, qui, se distinguant par des approches et des outils à priori fort divers : modèles géométriques, mélanges de densités, mesures de la séparation entre classes, graphes aléatoires, types de partition, similarités probabilistes, sont susceptibles d'apporter des réponses aux questions suivantes :*

- les données sont-elles classifiables ?

- comment tester l'existence d'une structure au moyen d'une hypothèse nulle de "non classifiabilité" ?

- comment tester si la structure, nécessairement produite par l'algorithme, n'est pas qu'un artefact résultant de phénomènes purement aléatoires ?

Abstract : *This critical review presents existing published works about validity tests in classification which can bring answers to the following questions :*

- are the data classifiable ?

- how are we to test the existence of a structure by means of a null hypothesis of "non-classifiability ?"

- how are we to test whether the structure, necessarily produced by the algorithm, is only an artifact produced by pure random phenomena ?

- how are we to judge the statistical significance of the results (trees, partitions, clusters) produced ?

These works differ widely in their tools and approaches : geometrical models, mixture of densities, measure of disjunction between clusters, random graphs, type of partition, probabilistic similarities.

Mots-clés : *Classification - Classifiabilité - Validité - Test statistique.*

1. INTRODUCTION

Alors que la production d'articles et d'ouvrages en classification se fait à un rythme élevé et que les méthodes de classification hiérarchique et de partitionnement abondent, jusqu'à rendre parfois bien difficile le choix du praticien, on constate un manque relatif de travaux sur la validité de ces méthodes et la valeur des résultats produits (arbres, partitions, classes).

Cela n'est certes pas étranger à l'absence d'une théorie statistique (au sens classique du terme) de la classification et d'une définition rigoureuse du concept de classe.

Ainsi, les travaux existant dans le domaine de la validité se distinguent par des approches et des outils a priori fort divers : modèles géométriques, mélanges de densités, mesures de la séparation entre classes, graphes aléatoires, types de partition, similarités probabilistes.

Après une période consacrée à la production de méthodes, la classification doit maintenant se consacrer pour une part plus importante à une réflexion critique sur ses acquis si elle veut continuer à s'affirmer devant des utilisateurs potentiels comme une approche privilégiée du traitement statistique des données.

Cette réflexion peut se faire en recherchant d'abord ce qui unit et distingue les travaux achevés, en en tirant ensuite une méthodologie de l'emploi de la classification, en proposant enfin de nouvelles voies de recherche.

En particulier, une méthodologie pertinente de la classification doit poser les questions suivantes comme préalable à toute utilisation des résultats produits par une technique :

- les données sont-elles classifiables ?

Avant l'utilisation d'un algorithme, cela signifie :

- comment tester l'existence d'une structure au moyen d'une hypothèse nulle de "non classifiabilité" ?

Après l'utilisation d'un algorithme :

- comment tester si la structure, nécessairement produite par l'algorithme, n'est pas qu'un artefact produit par des phénomènes purement aléatoires ?

- comment juger de la significativité statistique des résultats produits ?

Ce travail présente les travaux susceptible d'apporter des réponses à ces seules questions. Ainsi ceux traitant de la stabilité des résultats, de la comparaison des méthodes, et plus généralement tous ceux nécessitant l'existence d'une classification pertinente ne sont pas évoqués.

2. MODELES GEOMETRIQUES

L'utilisation de modèles géométriques remonte à une trentaine d'années et s'est d'abord faite dans les domaines de l'écologie, de la biologie, de la botanique. Leur particularité est de se limiter à des modélisations bidimensionnelles, avec parfois des études mono ou tridimensionnelles.

Mack (1954, 1956) traite de particules bidimensionnelles ayant une surface convexe. Sous une hypothèse de localisation et d'orientation uniforme des particules, il étudie les lois exactes et asymptotiques du nombre moyen de groupes (ensemble de particules se chevauchant), du nombre moyen de particules isolées, de l'aire moyenne non couverte par les particules. L'auteur généralise en dimension trois en appliquant ces résultats aux projections bidimensionnelles des particules.

Les tests monodimensionnels n'ont que peu d'intérêt pour nous, sinon qu'ils donnent parfois lieu à des extensions multidimensionnelles.

Ainsi Naus (1966) après avoir étudié deux tests de l'hypothèse nulle d'une répartition uniforme sur un fermé borné de $\mathbb{R}$, basés sur des statistiques de comptage sur une partition arbitraire de l'intervalle de variation, en propose une extension multidimensionnelle à la distribution uniforme sur le cube unité de $\mathbb{R}^p$.

Cette hypothèse d'uniformité est la plus couramment utilisée comme hypothèse nulle d'un test de classifiabilité faisant appel à des modèles géométriques.

En dimension un, pour une densité continue, elle s'exprime de trois manières équivalentes :

- les nombres d'occurence dans des intervalles disjoints de même longueur sont identiquement et indépendamment distribués (i.i.d.) suivant une loi de Poisson de moyenne inconnue.

- les temps d'attente entre deux occurences sont i.i.d. suivant une loi exponentielle de moyenne inconnue.

- dans tout intervalle borné contenant n évènements, ceux-ci sont i.i.d. uniformément.

En dimension un, pour une densité discrète, l'hypothèse nulle peut être exprimée de deux manières équivalentes :

- à chaque essai, la probabilité d'occurence d'un évènement est constante.

- étant donné n occurences et m essais les $\binom{m}{n}$ combinaisons des n occurences sont équiprobables.

En dimension deux, pour une densité continue, on a les formulations équivalentes :

- les nombres d'occurences dans des régions disjointes de même aire sont i.i.d. suivant une loi de Poisson de moyenne inconnue.

- n évènements étant donnés dans le carré unité, leur coordonnées sont i.i.d. uniformément.

Dans ce même article, Naus donne un grand nombre de référence pour l'étude d'alternatives particulières.

Clark et Evans (1954) se réfèrent à la même hypothèse nulle mais utilisent comme statistique de test la somme des distances entre plus proches voisins. De grandes valeurs indiquent une tendance à la régularité, de petites valeurs suggèrent des regroupements. Malheureusement les auteurs ignorent les dépendances entre distances (inégalité triangulaire, ...) et la loi de la statistique ne semble pouvoir être évaluée que par simulation.

Bartlett (1964) évalue la loi de la distance entre deux points i.i.d. uniformément dans un carré et la compare à la loi observée. Lui aussi ignore les dépendances entre distances.

Ces deux approches sont reprises par Besag et Diggle (1977) qui obtiennent les lois des statistiques de test par simulation de type Monte Carlo.

Holgate (1965) utilise aussi la même hypothèse nulle en dimension deux et fait appel à des statistiques basées sur les distances entre les objets à classifier et n points choisis aléatoirement dans une région fixée du plan (distance entre chaque objet et son k^e plus proche point).

L'auteur étudie pour k = 1 les tests proposés par Hopkins (1954), Piélou (1959), et Moore (1954) et les compare vis-à-vis de diverses alternatives (grille de classes, processus de Thomas). Il constate la difficulté à établir des alternatives non arbitraires pour apprécier la puissance des tests.

Mead (1974) après avoir fait une critique des tests existants qu'ils soient basés sur des statistiques de comptage ou de distances, et en remarquant que ces tests ne peuvent détecter que des classes de même échelle, propose une procédure de détection de "classes de classes".

Cette technique est due originellement à Greig-Smith (1952) et consiste à effectuer des comptages dans une grille de petits quadrats contigus. On construit alors une table des comptages hiérarchisée suivant les dichotomies successives de la grille et usant d'une démarche de type analyse de variance, on arrive à une statistique suivant approximativement une loi de Fisher sous l'hypothèse nulle poissonienne. L'inconvénient de ce test est que le rejet de l'hypothèse nulle à un ou plusieurs niveaux de la grille invalide le modèle poissonien des comptages dans chaque quadrat.

D'autres formes de test sont possibles ; e.g. en testant la probabilité d'affectation des objets aux groupes successifs de quadrats, ou bien en testant les dichotomies successives de la grille au moyen de comptages dans chacun des demi-groupes. Notons que l'alternative la plus convenable pour Mead est le "center satellite process" de Warren (1977).

La démarche de Strauss (1975) est tout à fait différente. Il se propose d'estimer la densité des objets à classifier en fonction des distances inter-objets, l'estimation sera alors invariante par rotation et translation. Plus précisément, la distance entre deux points est remplacée par un indicateur binaire précisant si ces deux objets sont "proches" ou non, cette proximité étant définie à l'aide d'un seuil sur les distances.

Sous les hypothèses, précisées par Kelly et Ripley (1976), que la densité est fonction symétrique des objets et que la loi conditionnelle de X_n sachant $x_1, \ldots, x_{n-1}$ ne dépend que du nombre d'objets situés dans une sphère de rayon r autour de X_n, on montre que la densité de $(X_1, \ldots, X_n)$ ne dépend que de la loi du nombre d'objets "proches" Y sous l'hypothèse nulle de distribution aléatoire et d'un paramètre v indépendant de n et mesurant la "clustering tendency" des données (v = 0 signifiant une absence de structure).

Strauss et Kelly et Ripley donnent en dimension 1, 2 et 3 les résultats permettant l'application du modèle. Hormis le problème de l'estimation de la loi ou des moments du nombre d'objets "proches" Y sous l'hypothèse nulle, celui du choix de la valeur du rayon r n'est pas négligeable. Ainsi pour tester $v = 0$ contre $v > 0$, le test UPP est de région critique $\{y/y>c\}$, mais la valeur de r qui conduit à une puissance asymptotique maximum est généralement trop grande et inadmissible pour les applications.

Saunders et Funk (1977) reprennent et améliorent le travail de Strauss. Sous certaines conditions d'éparpillement des objets (surface de la région proportionnelle au carré du nombre d'objets), la statistique de Strauss converge vers un processus de Poisson. La loi de la statistique est obtenue par simulation, même pour $v = 0$. Les auteurs constatent que le test proposé par Strauss est inapplicable quand la statistique de test est trop petite et proposent des résultats sur l'approximation poissonienne (pour $v = 0$), ainsi qu'un test utilisant, non la valeur précise de r, mais seulement un intervalle de variation. Ce test est convergent, mais asymptotiquement le nombre d'objets proches tend vers l'infini pour $v > 0$, ce qui interdit une approximation poissonienne. Ici encore, les auteurs constatent leur incapacité à définir une alternative pertinente ($v > 0$).

3. MODELES DE MELANGE DE DENSITES

La modélisation par mélange de lois peut s'appliquer de plusieurs manières au problème de la classification.

Il peut s'agir de tester si l'échantillon observé provient d'une population homogène ou d'un mélange de population hétérogènes.

Il peut s'agir de la recherche et de l'estimation des modes de la distribution observée ce qui conduira à la conclusion que la population est homogène ou bien est classifiable.

Il peut s'agir du test de la "réalité" d'une classe ce qui pourra se faire en se ramenant à l'une des deux techniques précédentes ou bien en examinant les meilleurs ajustements en k, puis $k + 1$ composantes.

Toutefois cette approche se heurte à plusieurs difficultés. D'une part, il est extrêmement difficile (sauf en dimension un) de mener à bien les calculs nécessaires à l'estimation des paramètres du modèle ou à l'obtention des lois des statistiques de test. D'autre part, la nécessité d'admettre un modèle fixé conduit à une estimation des modes ou des composantes du mélange dépendant étroitement de la justesse du modèle.

Bock (1977) propose un test permettant de décider si un échantillon provient d'une population homogène ou d'un mélange de populations hétérogènes se distinguant par des paramètres de localisation et d'échelle. I.e. un test de l'hypothèse nulle :

$$H : g(x) = f(x - \mu)$$

où g est la densité inconnue de l'échantillon, f une densité connue, et μ un paramètre de localisation inconnu ($\mu \in \mathbb{R}^p$) ; contre l'hypothèse alternative.

A : $g(x) = \Sigma\{\alpha_r\, f(x-\mu_r)/\sigma_r)/1\leqslant r\leqslant m\}$, (avec $m>1$) où les $\{\alpha_r\}$ sont les fréquences des classes inconnues, $\{\mu_r\}$ et $\{\sigma_r\}$ les paramètres de localisation et d'échelle inconnus (les $\{\mu_r\}$ étant deux à deux distincts).

Bock utilise un résultat dont le fondement est déjà présent chez Clark et Evans (1954) : sous l'hypothèse d'unimodalité, le nombre de distances inter-objets faibles est plus élevé que sous l'hypothèse de multimodalité.

Le test proposé a l'avantage de n'être pas limité au cas de mélanges gaussiens. Sous certaines conditions (e.g. si l'alternative est suffisamment éloignée de l'hypothèse nulle), le test est convergent. Toutefois vu la faible vitesse de convergence, l'auteur note que la reconnaissance de mélange nécessite un grand nombre d'observations. On trouvera dans l'article une application de ce test à des mélanges gaussiens multidimensionnels et exponentiels monodimensionnels.

Certains problèmes restent ouverts, que ce soit la robustesse vis-à-vis des écarts aux hypothèses de modélisation, l'efficacité par rapport au test du rapport des vraisemblances maximum, ou bien le choix optimal des paramètres d'estimation de la densité observée (par la méthode des noyaux de Parzen).

Rohlf et Fisher (1968) étudient la distribution du coefficient de corrélation cophénetique (CPCC) de Sokal et Sneath sous l'hypothèse que les objets sont issus d'une loi gaussienne ou uniforme pour le critère d'agrégation UPGMA (Unweighted Pair-Group Arithmetic Average).

La loi est obtenue par simulation avec très peu d'essais. La valeur moyenne du CPCC semble décroître avec le nombre d'objets et être indépendante du nombre de dimensions pour les deux modèles étudiés.

Engelman et Hartigan (1969) proposent un test de détection du mélange de deux gaussiennes monodimensionnelles de même variance. La statistique de test est le maximum du rapport de l'inertie inter-classe à l'inertie intra-classe pour toutes les partitions en deux classes (ce qui revient au test du rapport des vraisemblances maximum).

On sait depuis Fisher (1958) que la partition optimale est constituée de deux intervalles disjoints. Les auteurs fournissent des tables des valeurs critiques du test obtenues par simulation, ainsi qu'une formule analytique approchée de la valeur critique.

L'article de Hartigan (1977), beaucoup moins rigoureux que celui de Bock, fait suite au travail précédent et suggère des procédures de test de la "réalité" des classes. Il y a peu de résultats exacts ou même asymptotiques, mais surtout des considérations sur les lois "probables" des statistiques de test évaluées, dans le cas multidimensionnel, à partir des situations en dimension un. On pourra toutefois en tirer quelques renseignements sur les tests de bimodalité de type F.

Lee (1979) propose quatre critères de test équivalents de l'existence d'un mélange de deux lois gaussiennnes multidimensionnelles de même matrice de covariance V.

Le premier critère est celui bien connu du rapport d'inerties maximum pour toutes les partitions en deux classes : $C = \max |T|/|W|$, mais malgré le résultat de Scott et Simons (1971) prouvant que les deux classes optimales sont séparées par un hyperplan, les calculs restent prohibitifs.

Le deuxième critère consiste à utiliser les tests monodimensionnels pour toutes les projections en dimension un, bien que l'on ne sache rien des propriétés statistiques de ce procédé.

Au contraire, le troisième critère consiste à partitionner avant de projeter en dimension un et est fonction du rapport de la variance interclasses à la variance intraclasses.

Enfin, le quatrième critère consiste à rechercher $\max \mathrm{tr}\,(W^{-1}B)$ sur toutes les partitions possibles en deux classes.

Les lois de ces quatre critères sont difficiles à obtenir bien que des considérations sur le rang de V puissent simplifier le problème. Notons qu'une puissance élevée du test (qui est fonction croissante de la distance de Mahalanobis entre les groupes) peut être produite uniquement par la présence de deux variables fortement corrélées.

Friedman et Rubin (1967) utilisent les critères un et quatre de Lee pour juger de l'existence de k classes ($k \geqslant 2$) bien que l'on ignore tout des comportements de ces tests pour $k > 2$) et en particulier s'ils se comportent comme des tests qui utiliseraient toutes les valeurs propres, ce dont on peut légitimement douter.

Lennington et Flake (1975) font l'hypothèse d'un mélange gaussien multidimensionnel. Leur principal résultat est que si les composantes du modèle sont indépendantes, l'ensemble des distances euclidiennes usuelles entre les points a une distribution asymptotique gaussienne multidimensionnelle dont l'espérance et la matrice de covariance sont fonction des espérances et matrices de covariance des composantes. Les estimateurs de ces paramètres provenant des moyennes et covariances d'échantillonage étant convergents.

Malheureusement l'épreuve de validité qu'ils proposent est conçue pour être appliquée à une famille très particulière de méthodes de classification et semble impossible à appliquer aux méthodes les plus couramment utilisées.

4. TESTS SUR LA SEPARATION DE DEUX CLASSES

Cette rubrique regroupe des procédures fondées sur des représentations géométriques et une modélisation généralement gaussienne sans faire appel explicitement à des mélanges de lois ou au test d'hypothèses nulles poissonniennes.

Les tests de séparation de classes peuvent s'appliquer à des classes fournies par une méthode de partitionnement mais aussi au test de la significativité des noeuds successifs d'un arbre hiérarchique binaire.

Sneath (1977a, 1977b), constatant que ce qui importe dans l'appréciation de la distinction entre deux classes est leur coupure (ou son absence) en liaison avec la dispersion des classes, et non l'écart entre centres de gravité, propose différents indices mesurant le chevauchement et la séparation de deux classes convexes au sens de Fischer et Van Ness (1971). Ces indices sont construits après projection des points des classes sur l'axe reliant les centres de gravité et construction de l'histogramme des fréquences des classes sur cet axe, ou bien de l'histogramme des fréquences des distances des objets aux centres de gravité.

Des test sont proposés sous des hypothèses, soit de distributions gaussiennes sphériques indépendantes des deux classes, soit de distribution uniforme des projections sur l'axe reliant les centres de gravité des classes (modèle poissonien). Ces tests sont approximatifs et des problèmes peuvent se poser quand, e.g., les tailles des classes sont faibles.

Sneath propose aussi d'utiliser le test de Kolmogorov-Smirnov dans le cas où le choix de la statistique de mesure de la séparation entre classes se fait sans ambiguité. Ce test est malheureusement peu puissant et est conservatif comme le montre un exemple d'application à huit cas de figure sur des observations uniformes ou gaussiennes, homogènes ou plus ou moins distinctes.

La démarche de Mountford (1971) est tout à fait différente. Le modèle probabiliste n'est plus posé sur les positions des objets mais sur leurs similarités. La similarité x_{ij} d'une paire $\{i,j\}$ est supposée suivre une loi gaussienne $\mathcal{N}(\mu_{ij}, \sigma^2)$, la corrélation entre similarités étant nulle pour tout quadruplet formant un quadrilatère. Le modèle ainsi posé se rapproche des travaux de Gower (1966).

Deux classes sont considérées comme étant significativement distinctes si les similarités intragroupes sont significativement plus grandes en moyenne que les similarités intergroupes. Sous les conditions que les similarités du premier (resp. deuxième) groupe ont même moyenne μ_{11} (resp. μ_{22}) et que les similarités intergroupes ont même moyenne μ_{12}, Mountford propose un test de l'hypothèse nulle : $\mu_{11} = \mu_{22} = \mu_{12}$

Si le critère de classfication est externe (classification non fondée a priori sur les similarités), le test est du type t. Si les μ_{ij} peuvent être différents, alors ce test sous évalue la significativité de l'écart à l'hypothèse nulle.

Dans le cas où le critère de classification est interne les valeurs du t sont conditionnées par la procédure de classification et seront supérieures à celle attendues. Comme il est impossible de construire un test exact en raison du grand nombre de partitions en deux classes, on se contente de comparer le t observé à la loi du maximum des $2^{n-1} - n - 1$ valeurs obtenues par la décomposition de n éléments en deux groupes d'aux moins deux éléments. Ce test est conservatif au sens ou à chaque fois que la statistique observée n'est pas celle produisant le t maximum, la significativité de l'écart à l'hypothèse nulle est sous évaluée. Il serait bien entendu utile de disposer de tests liés aux critères de classification utilisés.

Une bonne approximation de la loi du maximum est obtenue pour $n < 10$. Pour $n > 10$, l'auteur fournit des approximations basées sur des conditions acceptables quant au comportement asymptotique du modèle.

Cette méthode s'applique de manière privilégiée à des similarités de nature probabiliste (cf § 7).

5. GRAPHES ALEATOIRES

Ling (1973) a proposé le modèle de graphe aléatoire le plus utilisé en classification. L'hypothèse nulle de graphe aléatoire (HGA) consiste à supposer que les $(n(n-1)/2)\,!$ matrices possibles de rang des dissimilarités sont équiprobables.

Ce modèle, dont on trouvera un exposé clair et concis chez Bailey et Dubes (1982), a comme inconvénient d'être invalide si des contraintes existent sur les dissimilarités e.g. si ce sont des distances, car alors certaines des matrices sont impossibles à réaliser. En revanche, ce modèle permet généralement de mener à bien les calculs. Ling estime que ce modèle peut être considéré comme un cas limite, et ainsi qu'une classe jugée non significative pour ce modèle ne le sera pas non plus pour un modèle plus réaliste.

Dans le même article, l'auteur définit l'indice d'isolation d'une classe comme sa durée de vie dans l'arbre hiérarchique et en donne la loi sous HGA (loi hypergéométrique négative) pour le critère d'agrégation du saut minimum (Single Linkage ou SL), un test de la "réalité" d'une classe est donc aisé à construire.

L'auteur propose aussi de tester la compacité d'une partition en utilisant la probabilité, connue sous HGA, pour chaque niveau d'un arbre SL, qu'au plus k objets appartiennent à des classes non réduites à un singleton. De petites valeurs de cette probabilité pour k observé feront rejetter l'hypothèse nulle.

Une autre statistique de test, qui fait suite aux travaux de Fillenbaum et Rapoport (1971, 1972) est le nombre d'arêtes minimum nécessaire pour rendre connexe un graphe aléatoire. Cette statistique peut être appliquée à tous les niveaux d'un arbre hiérarchique mais ne permet de juger que d'un niveau à la fois. Connaissant la loi de cette statistique, on peut juger du nombre d'arêtes devant être observées pour rejetter l'hypothèse de données aléatoires de type HGA.

Pour estimer cette loi, les auteurs utilisent une formule asymptotique de Erdos et Renyi (1959, 1960) dont Schultz et Hubert (1973) montrent la mauvaise qualité pour de petits échantillons.

Ling (1975) donne la loi exacte de cette statistique et constate la bonne qualité des simulations de Schultz et Hubert.

Puis Ling et Killough (1976) fournissent la fonction de répartition exacte de cette statistique pour les différents niveaux d'un arbre hiérarchique construit par SL sous HGA. Le même article donne le nombre moyen de composantes connexes d'un graphe aléatoire (sous HGA). Les valeurs sont exactes pour $n \leqslant 30$ et approchées (d'après Erdos et Renyi) pour $30 \leqslant n \leqslant 100$, cette approximation sous estimant légèrement la valeur la valeur exacte. Notons que sur trois exemples où les auteurs testent l'hypothèse HGA en utilisant simultanément les deux statistiques, deux donnent des résultats contradictoires.

Fillenbaum et Rapoport proposent d'autres statistiques de "clustering tendency" telles que le degré des noeuds dans les graphes aléatoires, ou le nombre de cycles d'ordre fixé. Cependant ces statistiques ne font, tout comme les précédentes, que tester l'arbre à un seul niveau (à l'exception de l'indice d'isolation de Ling qui utilise deux niveaux mais ne juge que de la validité d'une classe). D'autre part toutes ces statistiques ne sont valable que pour les hiérarchies SL.

Frank (1978) représente la structure vraie des données comme un graphe transitif non orienté caractérisé par différents paramètres tels que : le nombre de classes de tailles différentes, le nombre total de classes, les moyennes et variance des tailles des classes, etc. Sous l'hypothèse HGA de Ling, il propose une estimation de ces paramètres basées sur des comptages de sous-graphes de deux ou trois sommets ayant zéro, une, deux ou trois (pour les triades) arêtes.

L'article de Frank et Svensson (1981) se limite à l'étude des hiérarchies SL et propose un algorithme de calcul du nombre d'arbre SL à n terminaux. Sous HGA, les auteurs donnent la distribution de ces hiérarchies. Ils étudient aussi pour n = 4 des hypothèses alternatives où la structure vraie et définie comme un sous graphe transitif non orienté.

En ce qui concerne les hiérarchies obtenues par le critère d'agrégation du saut maximum, ou diamètre (Complete Linkage ou CL), Matula (1977) propose d'utiliser la taille de la plus grande clique (on sait que toute classe produite par CL est une clique) dans un graphe aléatoire.

Dans l'hypothèse de graphe aléatoire de Matula ou "random edge graph", les arêtes existent indépendamment avec une probabilité p. Pour une valeur particulière de p, on retrouve le modèle de Ling à un niveau de l'arbre.

La loi de la statistique étant de forme pointue, l'obtention d'une clique de taille même peu différente de la taille espérée est donc assez invraisemblable. Notons que le critère de Matula est surtout un critère de compacité de la partition testée.

Bailey et Dubes (1982) présentent une revue détaillée et critique des épreuves de validité basées sur les graphes aléatoires et proposent une technique destinée à combler les lacunes mises en évidence.

Les auteurs définissent un indice de compacité et un indice d'isolation de classe à chaque niveau de distance, fonction des nombres de distances tances inter et intraclasses inférieures au niveau fixé. Une classe compacte (resp. isolée) aura un indice de compacité (resp. d'isolation) élevé (resp. faible) pour plusieurs niveaux consécutifs. L'introduction d'une échelle de probabilité liée, calculée sous une certaine hypothèse nulle de "non classifiabilité", permettant la construction de la probabilité sous l'hypothèse nulle, qu'une classe est aussi compacte ou isolée que la classe testée.

Le premier modèle de non classifiabilité revient à considérer les classes de même cardinal que la classe testée équiprobables pour chacun des niveaux. Les critères d'isolation et de compacité associés auront tendance à être surévalués si la classe testée est produite par un algorithme de classification basé sur les distances. Pour pallier à cet inconvénient, le second modèle revient à affecter d'une probabilité nulle les classes de même cardinal que la classes testée ne pouvant être produites par l'algorithme utilisé ; la probabilité étant répartie de manière égale sur les autres classes. Dans ce cas, l'absence d'expression analytique pour les critères d'isolation et de compacité nécessite le recours à des simulations.

Le troisième modèle, dit "du meilleur cas" consiste à répartir uniformément la probabilité, d'une part sur l'ensemble des classes de même cardinal que la classe testée ayant un indice d'isolation minimum (i.e. les plus isolées), d'autre part sur l'ensemble des classes de même cardinal que la classe testée ayant un indice de compacité maximum (i.e. les plus compactes). Le critère d'isolation est calculé conditionnellement à la première distribution, et le critère de compacité conditionnellement à la deuxième.

Enfin, le quatrième modèle évalue le critère d'isolation conditionnellement à la population des classes ayant même taille et même compacité que la classe testée, et le critère de compacité conditionnellement à la population des classes ayant même taille et même isolation que la classe testée.

Dans chaque cas les profils probabilistes utilisés sont les graphes des critères d'isolation et de compacité en fonction des niveaux. Malheureusement l'influence et l'importance des niveaux successifs de distance, auxquels une classe est jugée significative, ne sont pas évoqués.

Les auteurs proposent d'utiliser le premier modèle pour des classes a priori. Dans les autres cas, le troisième modèle est plus attractif que le deuxième modèle en raison de son indépendance vis-à-vis de la méthode de classification, ce qui n'oblige pas à des simulations pour établir les valeurs critiques des tests. Notons toutefois que les tests du modèle trois sont conservatifs. Le quatrième modèle, quant à lui, peut s'appliquer à toutes les situations, que les classes soient produites par un algorithme, ou non. Ils proposent enfin une stratégie de test basée sur l'utilisation des critères des modèles trois et quatre, et recommandent de faire précéder le test de classes individuelles par un test global sur la hiérarchie (cf. Ling et Killough, et Hubert).

Citons enfin Baker et Hubert (1976) qui utilisent une matrice de proximité ordinale et la notion de clique isolée. Le saut maximum génère des cliques et leur travail concerne ce seul critère. La classification parfaite est assimilée à un graphe à seuil dont chaque noeud appartient à un sous graphe isolé.

La différence entre le nombre d'arêtes du graphe observé et le nombre d'arêtes minimum pour produire la structure idéale est fonction croissante de la qualité de la partition produite par une hiérarchie CL. La loi de cette statistique dépend du nombre d'objets et n'a pu être que simulée.

6. TYPES DE PARTITIONS

Les travaux suivant sont basés sur le concept de type d'une partition. Le type d'une partition est la donnée du nombre de classes et de la suite des cardinaux des différentes classes. Les techniques évoquées peuvent donc s'appliquer à toutes les structures de classification mais ont l'inconvénient de ne faire intervenir qu'un seul niveau des arbres hiérarchiques.

Dans un cadre appelé "quadratic assignment", Hubert et Schultz (1976) proposent de choisir une structure idéale a priori en fixant un type de partition. Ils calculent ensuite un moment produit Λ entre la matrice de dissimilarité donnée et la structure idéale représentée par une fonction binaire. Ce coefficient mesure la qualité (en un certain sens) de l'ajustement des deux structures.

Sous l'hypothèse de réarrangements simultanés et équiprobables des lignes et des colonnes de la matrice de dissimilarité, l'espérance et la variance de Λ sont connus. Sa loi n'a pu être évaluée que par simulations.

Mielke, Berry et Johnson (1976) fixent aussi le type de la partition. Mais ici l'hypothèse nulle est que les affectations des objets aux classes (relativement à leur cardinal), se font de manière indépendante et équiprobable. La statistique de test est la somme des distances intra-groupes.

Sous l'hypothèse nulle, on connait l'espérance, la variance et le coefficient d'asymétrie de la statistique. Malheureusement le test exact est impossible à mettre en oeuvre, même pour de petits ensembles de données, aussi les auteurs proposent-ils deux tests approchés. Le premier, fondé sur une approximation par loi β fait intervenir un paramètre dont le choix est souvent quelque peu empirique et non unique. Le second, fondé sur une approximation gaussienne n'est applicable que pour un coefficient d'asymétrie proche de zéro (e.g. compris entre -0,1 et 0,1).

Dans une toute autre optique, Baker et Hubert (1975) proposent un test de classifiabilité en classification hiérarchique ainsi que le calcul de sa puissance pour une famille d'alternatives pour les stratégies d'agrégation SL et CL.

L'hypothèse nulle est la nullité de la matrice des rangs des dissimilarités. Cette matrice est bruitée par une variable aléatoire centrée de variance arbitrairement grande (ce qui peut se rapprocher du modèle de Ling). On calcule ensuite les hiérarchies SL et CL associées pour mille réalisations de la matrice bruitée. Pour chacune de ces réalisations une corrélation de rangs (γ de Goodman-Kruskal) est calculée entre les partitions à un niveau fixé des deux hiérarchies et la matrice des rangs bruitée. On obtient ainsi une distribution empirique de γ sous l'hypothèse nulle.

Les auteurs limitent donc le test à un seul niveau de la hiérarchie. De plus la statistique γ utilisée se rattache linéairement à un indice d'adéquation de partition ne favorisant guère les classes allongées, ce qui explique certains résultats quantitatifs de l'étude.

La caractérisation de l'alternative se fait par le choix d'un type de partition au niveau testé. Cette structure est définie par une fonction sur les paires d'objets, égale à zéro si les objets sont dans la même classe, et à un sinon, et est perturbée par l'addition d'un bruit gaussien. Cela fournit une matrice de rangs de dissimilarités qui est classifiée par SL et CL. Comme dans le cas de l'hypothèse nulle, on calcule ensuite la corrélation γ entre la partition extraite au niveau testé et la matrice de rangs bruitée. Un calcul exact par dénombrement étant impossible, on fait encore appel à des simulations.

Baker et Hubert proposent aussi le calcul direct de γ entre la partition testée et la partition donnée par la matrice bruitée comme une mesure de l'adéquation à une structure vraie. Il pourrait être intéressant de comparer plus précisément ces deux méthodes.

Les auteurs observent que quand la taille maximum d'une classe est inférieure (resp. supérieure) à 7, CL (resp. SL) donne une partition plus proche de la structure vraie, il en est de même du comportement de la puissance du test.

7. SIMILARITES PROBABILISTES

Ce paragraphe présente des travaux sur la classifiabilité qui découlent de la signification probabiliste attachée à la similarité utilisée.

Goodall (1964, 1966, 1970) définit la similarité d'une paire d'objets comme un indice probabiliste égal au complément de la probabilité que ceux-ci ne soient pas moins semblables en ayant des attributs assignés uniformément et indépendamment.

Cette hypothèse d'indépendance et d'uniformité de répartition des attributs est le principal obstacle à l'utilisation de l'indice de Goodall.

L'application à la classifiabilité peut se faire en testant le maximum des similarités par rapport à des simulations faites sous l'hypothèse nulle de Goodall : attributs distribués suivant des fréquences fixées. Le nombre d'attributs croissant, les probabilités d'erreur de type I simulées se rapprochent de celles calculées sous l'hypothèse d'indépendance, et la loi de la similarité maximum semble en pratique indépendante des fréquences associées aux attributs.

Ce test peut être adopté aux triplets et quadruplets en utilisant le maximum des similarités minimum dans chaque groupe de trois ou quatre objets. La loi théorique n'est toujours pas calculable, et dans ce cas les simulations tendent à prouver qu'elle n'est pas indépendante des fréquences assignées aux attributs.

Lerman (1980, 1981) propose aussi un indice de similarité ayant une signification probabiliste, et qui peut être défini aussi bien pour les attributs que pour les objets.

En ce qui concerne les attributs, Lerman adopte comme indice de similarité brut le nombre d'objets possédant simultanément les deux attributs. Puis il introduit une "hypothèse d'absence de lien" permettant d'éviter que la valeur de la similarité soit biaisée par la fréquence trop importante, ou trop faible, de l'un ou l'autre des attributs. Cette hypothèse consiste à fixer l'un des attributs et à associer à l'autre une distribution uniforme sur l'ensemble des parties de l'ensemble des objets ayant pour cardinal la masse de ce dernier attribut. L'auteur arrive ainsi à un indice de similarité probabiliste bien approché par la loi normale.

L'hypothèse nulle de non-classifiabilité revient à munir l'ensemble des tableaux croisant objets et attributs, et respectant la proportion d'attributs possédés par chaque objet dans le cas des données étudiées, d'une distribution uniforme.

La statistique de test H est définie comme une mesure de l'écart à la nature ultramétrique de la similarité observée (H = 0 si la similarité est ultramétrique). On teste l'hypothèse nulle : $H \neq 0$, contre l'alternative générale : H = 0. Ce test exact est généralement impossible à mettre en oeuvre, on se contente donc de simulations. Une exception est le cas où le nombre d'attributs possédés par chaque objet est constant (e.g. un questionnaire), on connaît alors la loi exacte de H.

8. CONCLUSIONS

La première conclusion est l'impossibilité actuelle à définir une hypothèse nulle de non-classifiabilité, ou même un petit nombre d'hypothèses nulles, qui satisfasse d'un point de vue pratique les utilisateurs des méthodes de classfication. Les hypothèses nulles ayant donné lieu aux travaux théoriquement les plus féconds (e.g. graphes aléatoires) ne sont pas pertinentes en pratique. En revanche, les hypothèses nulles ayant donné lieu à des applications ont été conçues dans un cadre limité (écologie, botanique), et ont donné lieu à des travaux théoriques de peu d'ampleur.

La seconde conclusion est la constatation d'une situation encore pire pour les hypothèses alternatives d'existence d'une structure de classification. Même dans les cas où une hypothèse nulle satisfaisante a pu être définie, le choix d'une,ou d'une famille, d'alternatives laisse une grande place à l'arbitraire. En conséquence, il est extrêmement difficile de comparer, en terme de puissance statistique, les tests proposés pour des hypothèses nulles semblables, et ainsi de sélectionner le, ou les meilleurs tests indépendamment de considérations sur les facilités de calcul théorique ou automatique.

Troisièmement, s'il existe de nombreux tests sur les partitions, i.e. permettant de tester la classifiabilité de données pour une méthode de partitionnement, il existe très peu de tests usant simultanément de tous les niveaux d'un arbre hiérarchique. Certes, il est toujours possible de tester séparément les partitions aux différents niveaux, mais l'utilisation qui en est faite (cf. e.g. Bailey et Dubes (1982)) montre bien les difficultés à apprécier les interactions existant entre les niveaux successifs.

Enfin, il faut noter l'usage très important des simulations de type Monte Carlo. Cela s'explique en grande partie par les difficultés à obtenir les lois exactes des statistiques de test, mais aussi par les facilités a définir des hypothèses nulles ou alternatives très particulières et pour lesquelles les calculs théoriques sont impossibles.

Cependant, un inconvénient de cette pratique est la nécessité, ou bien de prendre en compte toutes les valeurs possibles des paramètres (nombre d'objets, de variables, etc.) afin de construire des tables numériques des valeurs critiques, ou bien de faire une simulation numérique à l'occasion de chaque application.

Il est légitime de penser que les puissances des ordinateurs continuant à augmenter, l'usage de ces simulations sera de plus en plus important (sur la pratique et l'intérêt des simulations de Monte Carlo, voir Besag et Diggle (1977), et Milligan (1981)).

Il faut donc espérer que l'analyse des résultats de classification de données concrètes provenant de domaines variés pourra permettre la définition d'hypothèses nulles pertinentes pour un grand nombre d'applications.

Dans un premier temps la définition d'hypothèses alternatives, ou de modèles de structuresde classification, doit pouvoir se faire en s'inspirant soit de structure mathématiques fortes et bien connues (sphères, cubes, polyedres convexes, ellipsoïdes, etc.) soit de structures reflétant l'organisation de données concrètes définies à l'aide de critères externes aux méthodes de classification mathématique.

REFERENCES

ADAMS E.N. : Consensus techniques and the comparison of taxonomic trees Syst. Zool., 1972, 21, 390-397.

BAILEY, T.A. : Cluster validity and intrinsic dimensionality. Thesis, Michigan State Univ., Dep. of Computer Science, East Lanning, 1978.

BAILEY T.A. & DUBES R : Cluster validity profiles. Patt. Recogn., 1982, 15, 2, 61-83.

BAKER F.B. : Stability of two hierarchical grouping techniques. Case 1 : sensitivity to data errors. J. Am. Stat. Assoc., 1974, 69, 346, 440-445.

BAKER F.B. & HUBERT L.J. : Measuring the power of hierarchical cluster analysis. J. Am. Stat. Assoc., 1975, 70, 349, 31-38.

BAKER F.B. & HUBERT L.J. : A graph theoretic approach to goodness of fit in complete-link hierarchical clustering. J. Am. Stat. Assoc., 1976, 71, 870-878.

BARTLETT M. : The spectral analysis of two-dimensional point processes. Biometrika, 1964, 51, 299-311.

BAKER F.B. : Sensitivity of the complete-link clustering technique to missing individuals. J. Educ. Stat. 1978, 3, 233-252.

BAYNE C.K., BEAUCHAMP J.J., BEGOVICH C.L. & KANE V.E. : Monte-Carlo comparisons of selected clustering methods. Patt. Recogn., 12, 2, 51-62, 1980.

BENZECRI J.P. : L'Analyse des Données (Tome 1). Dunod, 1980.
BESAG J. & DIGGLE P.J. : Simple Monte Carlo tests for spatial pattern. Appl. Stat., 1977, 26, 327-333.
BLASHFIELD R.K. : Mixture model tests of cluster analysis : accuracy of four agglomerative hierarchical methods. Psychol. Bull., 1976, 83, 377-388.
BLASHFIELD R.K. & ALDENDERFER M. : Cluster analysis literature on validation. Abstract in Classif. Soc. Bull., 1978, 4, 2, 30, and unpublished paper.
BOCK H.H. : Statistische modelle und bayessche verfahren zur bestimmung einer unbekannten klassifikation normalverteilter zufälliger vektoren. Metrika, 1972, 18, 120-132.
BOCK H.H. : On tests concerning the existence of a classification. Analyse de Données et Informatique, IRIA, 1977, 177-195.
BASTIN C., BENZECRI J.P., BOURGARIT C., CAZES P. : Pratique de l'Analyse des Données (Tome 2). Dunod, 1980.
CARMICHAEL J.W., GEORGE J.A. & JULIUS R.S. : Finding natural clusters. Syst. Zool., 1968, 8, 17, 144-150.
CHANDRASEKARAN B., JAIN A.K. : Independence, measurement complexity and classification performance. IEEE Trans. Syst. Man. Cybern, 1975, 5, 240-244, and 1977, 7 564-566.
CHRISTENSON A., READ O.W. : Numerical Taxonomy, R-mode factor analysis and archeological classification. American Antiquity, 1977, 42, 163-179.
CLARK P. & EVANS F. : Distance to nearest neighbour as a measure of spatial relationships in populations Ecology, 1954, 35, 445-453.
COLEMAN R. : Random paths trough convex bodies. J. Appl. Prob., 1969, 6, 430-441.
CORMACK R.M. : Classification : An overview. Analyse de Données & Informatique, INRIA-CCE, 1980, 125-147.
CRESSIE N.A. : Testing for uniformity against a clustering alternative Thesis, 1975, Dep. of Stat., Princeton Univ.
CSORGO M. & GUTTMAN I. : On the empty cell test. Technometrics, 1962, 4, 235-247.
CUNNINGHAM K.M. & OGILVIE J.C. : Evaluation of hierarchical grouping techniques : a preliminary study. The Comp. J., 1972, 15, 3, 209-213.
DAVIES D.L. & BOULDIN D.W. : A cluster separation measure. IEEE Trans. PAMI, 1979, 1, 224-227.
DAY W.H.E. : Validity of clusters formed by graph-theoretic cluster methods. Math. Biosc, 1977, 36, 299-317.
DIDAY E. et coll. : Optimisation en classification automatique (Tomes 1 et 2). INRIA, 1979.
DIGGLE P.J. : On parameter estimation and goodness of fit testing for spatial point patterns. Biometrics, 1979, 35, 87-101.
DUBES R. & JAIN A.K. : Clustering techniques : the user's dilemna. Patt. Recogn, 1976, 8, 247-260.
DUBES R. & JAIN A.K. : Validity studies in clustering methodologies. Patt. Recogn. 1979, 11, 235-254.
DUBES R.C. & SMITH S.P. : Stability of a hierarchical clustering. Patt. Recogn, 1980, 12, 177-187.
DUBES R., JAIN A.K. : Clustering methodologies in exploratory data analysis. Adv. Comp., 1980, 19, 113-228.

DUDA R.O. & HART P.E. : Pattern classification and scene analysis. Wiley, 1973.
EDELBROCK C. : Mixture model test of hierarchical clustering algorithms : the problem of classifying everybody. Multiv. Behav. Res., 1979, 14, 367-384.
ENGELMAN L. & HARTIGAN J.A. : Percentage points of a test for clusters. J. Am. Stat. Assoc., 1969, 64, 1647-1648.
ERDOS P. & RENYI A. : On random graphs. Publ. Math. Debreen., 1959, 6, 290-297.
ERDOS P. & RENYI A. : On the evolution of random graphs Math. Inst. Hung. Acad. Sci, 1960, 5, 17-61.
FILLENBAUM S. & RAPOPORT A. : Structures in the subjective lexicon, Academic Press, 1971.
FILLENBAUM S. & RAPOPORT A. : An experimental study of semantic structures. In "Multidimensional Scaling", Vol II, Seminar Press, Romney-Shepard-Nerlove eds, 1972.
FARRIS J.S. : On the cophenetic correlation coefficient Syst. Zool., 1969, 18, 279-285.
FISHER W.D : On grouping for maximum homogenerty. J. Am. Stat. Assoc., 1958, 53, 789-798.
FISHER L. & VAN NESS J.W. : Admissible clustering procedures. Biometrika, 1971, 58, 91-104.
FLORY A., GUNTHER J., KOULOUMDJIAN J. : Etude de performance d'algorithmes d'affectation. Séminaires IRIA de Classif. Autom., 1975.
FRANK O. : Estimation of the number of connected components in a graph by using a sampled subgraph. Scand. J. Stat., 1978, 5, 177-188.
FRANK O. : Inferences concerning cluster structures. COMPSTAT 1978, Physica-Verlag, 259-265.
FRANK O. & SVENSSON K. : On probability distributions of single linkage dendrograms. J. Stat. Comp. Sim., 1981, 12, 121-132.
FRIEDMAN H.P. & RUBIN J. : On some invariant criteria for grouping data. J. Am. Stat. Assoc., 1967, 62, 1159-1178.
GEARY R.C. : The contiguity ratio and statistical mapping. The Incorp. Stat., 1954, 5, 115-145.
GENGERELLI J.A. : A method for detecting subgroups in a population and specifying their membership. J. Psychol., 1963, 55, 457-468.
GLICK N. : Seperation and probability of correct classification among two or more distributions. Ann. Inst. Stat. Math. Jap., 1973, 25, 373-382.
GNANADESIKAN R. & KETTENRING J.R. : Interpreting and assessing the results of cluster analysis. Bull. Int. Stat. Inst., 1977, 43, 451-463.
GOOD I.J. : The serial test for sampling numbers and others tests for randomness. Proc. Camb. Phil. Soc., 1953, 49, 276-284.
GOODALL D.W. : A probabilistic similarity index. Nature, 1964, 203, 1098.
GOODALL D.W. : A new similarity index based on probability. Biometrics, 1966, 22, 882-907.
GOODALL D.W. : Cluster analysis using similarity and dissimilarity. Biometrie-Praximetrie, 1970, XI, 1, 34-41.
GORDON A.D. : On the assessment and comparison of classifications. Analyse de Données & Informatique, INRIA-CCE, 1980, 149-159.
GOWER J.C. : Some distance properties of latent roots and vector methods used in multivanate analysis. Biometrika 1966, 53, 325-338.

GOWER J.C. : A comparison of some methods of cluster analysis. Biometrics, 1967, 23, 623-637.
GREIG-SMITH P. : The use of random and contiguous quadrats in the study of the structure of plant communities. Ann. Bot. N.S., 1952, 16, 293.
HAFNER R. : The asymptotic distribution of random clumps. Computing, 1972, 10, 335-351.
HARTIGAN J.A. : Representation of similarity matrices by trees. J. Am. Stat. Assoc., 1967, 62, 1140-1158.
HARTIGAN J.A. : Distribution problems in clustering. In Classification and Clustering, J. Van Ryzin ed., Academic Press, 1977, 45-71.
HARTIGAN J.A. : Asymptotic distributions for clustering criteria. Ann. Stat., 1978, 6, 1, 117-131.
HEALY J.D. : The effects of misclassification error on the estimation of several population proportions. The Bell Syst. Tech. J., 1981, 60, 5.
HOLGATE P. : Tests of randomness based on distance methods. Biometrika, 1965, 52, 345-353.
HOPKINS B. : A new method of determining the type of distribution of plant individuals. Ann. Bot., 1954, 18, 213-226.
HORN J.L. : Significance tests for use with rp an related profils characteristics. Educ. Psychol. Measur., 1961, 21, 363-370.
HUBERT L. : Approximate evaluation techniques for the single-link and complete-link hierarchical clustering procedures. J. Am. Stat. Assoc., 1974, 69, 347, 698-704.
HUBERT L. & SCHULTZ J. : Hierarchical clustering and the concept of space distortion. Br. J. Math. Stat. Psychol., 1975, 28, 121-133.
HUBERT L. & SCHULTZ J. : Quadratic assignment as a general data analysis strategy, Br. J. Math. Stat. Psychol., 1976, 29, 190-241.
HUBERT L. & LEVIN J.R. : A general statistical framework for assessing categorical clustering in free recall. Psychol. Bull., 1976, 83, 1072-1080.
HUBERT L. & LEVIN J.R. : Evaluating object set partitions : free sort analysis and some generalisations. J. Verb. Learn. Verb. Behav., 1976, 15, 459-470.
HUBERT L.J. & BAKER F.B. : An empirical comparison of baseline models for goodness-of-fit in r-diameter hierarchical clustering. In Classification & Clustering, J. Van Ryzin ed., Academic Press, 1977, 131-153.
HUIZINGA D. : Are there any clusters ? The evaluation of numerically constructed typologies. Techn. Report 78-2, 1978, Behavioral Research Institute, Boulder, Colorado.
HUNTINGTON R.J. : Distribution for clusters in continuous and discrete cases. Northeast Sci. Rev., 1974, 4, 153-161.
JAIN A.K. & WALLER W.G. : On the optimal number of features in the classification of multivate gaussian data. Patt. Recogn., 1978, 10, 365-374.
JAIN A.K., DUBES R. : Feature definition in pattern recognition with small sample size. Patt. Recogn., 1978, 10, 85-97.
JAMBU M. : Classification automatique pour l'Analyse des Données (Tome 1). Dunod, 1976.
JARDINE N. & SIBSON R. : The construction of hierarchic and non hierarchic classifications. The Comp. J., 1968, 11, 177-183.
JARDINE N. & SIBSON R. : Mathematical taxonomy. Wiley, 1971.
JOHNSON R.L. & WALL D.D. : Cluster analysis of semantic differential data. Educ. Psychol. Measur., 1969, 29, 769-780.
JOHNSTON B., BAILEY T. & DUBES R. : A variation on a non parametric clustering method. IEEE Trans. PAMI, 1979, 1, 4, 400-408.

KARONSKI M. : On a definition of cluster and pseudo cluster for multivariate normal population. Bull. Int. Stat. Int., 1973, 45, 593-598. KELLY F.P. & RIPLEY B.D. : A note on Strauss's model for clustering Biometrika, 1976, 63, 2, 357-360.
KNOX E.G. : The detection of space-time interactions. Appl. Stat., 1964, XIII, 1, 25-29.
KUIPER F.K. & FISHER L. : A Monte-Carlo comparison of six clustering procedures. Biometrics, 1975, 31, 777-783.
LEBART L., MORINEAU A., TABARD N. : Techniques de la description statistique : Méthodes et logiciels pour l'analyse des grands tableaux. Dunod, 1977.
LEE K.L. : Multivariate tests for clusters. J. Am. Stat. Assoc., 1979, 74, 1708-714.
LENNINGTON R.F. & FLAKE RH : Statistical evaluation of a family of clustering methods. Proceedings of the Eigth International Conference on Numerical Taxonomy, Eastabrook ed, 1975, Freeman.
LEVENE H. : A test of randomness in two dimensions. Abstract in Ann. Math. Stat., 1946, 17, 500.
LERMAN I.C : Combinatorial analysis in the statistical treatment of behavioral data. Quality and Quantity, 1980, 14, 431-469.
LERMAN I.C : Classification et analyse ordinale des données. Dunod, 1981.
LING R.F. : On the theory and construction of K-clusters. The Comp. J., 1972, 15, 4, 326-332.
LING R.F. : A probability theory of cluster analysis. J.Am. Stat. Assoc., 1973, 68, 341, 159-164.
LING R.F. : An exact probability distribution on the connectivity of random graphs. J. Math. Psychol., 1975, 12, 90-98.
LING R.F. & KILLOUGH G.G. : Probability tables for cluster analysis based on a theory of random graphs. J. Am. Stat. Assoc., 1976, 71, 354, 293-300.
MACK C. : The expected number of clumps when convex laminae are placed at random and with random orientation in a plane area. Proc. Camb. Phil. Soc., 1954, 50, 581-585.
MACK C. : On clumps formed when convex laminae or bodies are placed at random in two or three dimensions. Proc. Camb. Phil. Soc., 1956, 52, 246-250.
Mc QUITTY L.L. : A comparative study of some selected methods of pattern analysis. Educ. Psychol. Meas., 1971, 31, 607-626.
Mc QUITY L.L. & FRARY J.M. : Reliable and valid hierarchical classification. Educ. Psychol. Meas., 1971, 31, 321-346.
MARONNA R. & JACOVKIS P.M. : Multivariate clustring procedures with variable metrics. Biometrics, 1974, 30, 499.
MATULA D.W. : Graph theoretic techniques for cluster analysis algorithms. In "Classification and Clustering", Van Ryzin ed, Academic Press, 1977.
MEAD R. : A test for spatial pattern at several scales using data from a grid of contiguous quadrats. Biometrics, 1974, 30, 295-307.
MEZZICH J. & SOLOMON H. : Taxonomy and Behavioral Science : Comparative Performance of Grouping Methods. Academic Press, 1980.
MIELKE P.W. & BERRY K.J. : Multi-response permutation procedures for a priori classifications. Communic. in Stat. (Theor. Meth.), 1976, 5A, 14.

MILLIGAN G.W. : An examination of the effect of six types of error perturbation on fifteen clustering algorithms. Psychometrika, 1980, 45, 325-342.

MILLIGAN G.W. & ISAAC P.D. : The validation of four ultrametric clustering algorithms. Patt. Recogn., 1980, 12, 2, 41-50.

MILLIGAN G.W. : A Monte-Carlo study of thirty internal criterion measures for cluster analysis. Psychometrika, 1981, 46, 2, 187-199.

MILLIGAN G.W. : A review of Monte Carlo tests of cluster analysis. Multiv. Behav. Res., 1981, 16, 3.

MOJENA R. : Hierarchical grouping methods and stopping rules : an evaluation. The Comp. J., 1977, 20, 359-363.

MOJENA R., WISHART D. : Stopping rules for Ward's clustering method. In Compstat. 1980, Physica-Verlag, 426-432.

MOORE P.G. : Spacing in plant populations. Ecology, 1954, 35, 222-227.

MORF M.E., MILLER D.E. & SYROTRICK J.M. : A comparison of cluster analysis and Q-factor analysis. J. Clinic. Psychol., 1976, 32, 59-64.

MOUNTFORD M.D. : A test of the difference between clusters. Stat. Ecol., 1971, 3, 237-257.

MURTAGH F. : New results in the comparison and validation of hierarchic clustering. A paraître dans IEEE Trans. PAMI.

NAUS J.I. : A power comparison of two tests of non-random clustering. Technometrics, 1966, 8, 493-517.

PEARSON E.S. : Comparison of tests for randomness of points on a line. Biometrika, 1963, 50, 315-325.

PIELOU E.C : The use of point to plant distances in the study of the pattern of plant populations. J. Ecol. 1959, 47, 607-613.

POLLARD D. : Strong consistency of k-means clustering. Ann. Stat., 1981, 9, 135-140.

POLSON P.G. & HUIZINGA D. : Statistical methods for absorbing Markov-chains models for learning : Estimation and identification. Psychometrika, 1974, 39, 3-21.

RAGHAVAN V.V. & YU C.T. : A comparison of the stability characteristics of some graph theoretic clustering methods. IEEE Trans. PAMI, 1981, 4, 393-402.

RAND W.M. : Objective criteria for the evaluation of clustering methods. J. Am. Stat. Assoc., 1971, 66, 336, 846-850.

REGNIER S. : Stabilité d'un opérateur de classification. In Seminaires IRIA, 1976, 35-48.

RIPLEY B.D. : Modelling spatial patterns (with discussions). J.R. Stat. Soc. (B), 1977, 39, 172-212.

ROGERS G. & LINDEN J.D. : Use of multiple discriminant function analysis in the evaluation of three multivariate grouping techniques. Educ. Psychol. Measur., 1973, 33, 787-802.

ROHLF F.J. & FISHER D.R. : Tests for hierarchical structure in random data sets. Syst. Zool., 1968, 7, 407-412.

ROHLF F.J. : Methods of comparing classifications. Annu. Rev. Ecol. Syst., 1974, 5, 101-113.

ROHLF F.J. & SOKAL R.R. : Comparison numerical taxonomic studies. IBM report n° 8535, 1980.

ROHLF F.J. : Consensus indices for comparing classifications. IBM report, 1981.

RUBIN J. : Optimal classification into groups : An approach for solving the taxonomy problem. J. Theoret. Biol., 1967, 15, 103-144. SAUNDERS R. & FUNK G.M. : Poisson limits for a clustering model of Strauss. J. Appl. Prob., 1977, 14, 776-784.
SCHULTZ J. & HUBERT L.J. : Data analysis and connectivity of random graph J. Math. Psychol., 1975, 12, 90-98.
SCOTT A.J. & SYMONS M.J. On the Edwards and Cavalli-Sforza method of cluster analysis. Biometrics, 1971, 27, 217-219.
SILVERMAN B. : Using kernel density estimates to investigate multimodality. J.R. Stat. Soc (B),1981, 43, 97-99.
SMITH S.P. & DUBES R. : Stability of a hierarchical clustering. Patt. Recogn., 1980, 12, 3, 177-187.
SNEATH P.H.A. : Some statistical problems in numerical taxonomy. The Statistician, 1967, 17, 1-12.
SNEATH P.H.A. & SOKAL R.R. : Numerical taxonomy. Freeman, 1973.
SNEATH P.H.A. : A method for testing the distinctness of clusters : A test of the disjunction of two clusters in euclidean space as measu red by their overlap. Math. Geol., 1977 (a), 9, 2, 123-143. SNEATH P.H.A. : Cluster significance tests and their relation to measures of overlap. 1ères journées Analyse de Données et Informatique, IRIA, 1977 (b), 15-36.
SNEATH P.H.A. : Significance test for clusters in UPGMA phenograms obtained from squared euclidean distances. Classif. Soc. Bull., 1977 (c), 4, 2-14.
SOKAL R.R. & ROHLF F.J. : The comparison of dendrograms by objective methods. Taxon, 1962, XI, 2, 33-40.
STRAUSS D.J. : A model for clustering. Biometrika, 1975, 62, 2, 467-475.
SWED F.S. & EISENHART C. : Tables for testing randomness of grouping in a sequence of alternatives. Ann. Math. Stat., 1943, 14, 66-87. TURNER M.E.: Credibility and cluster. Ann. New York Sci. Acad., 1969, 161, 680-688.
WALTER F.D. : Testing for clusters of disease within households. Bull. Int. Stat. Inst., 1973, 4, 577-579.
WARREN W.G. : The centre - satellite concepts as a basis for ecological sampling. Stat. Ecol., 1971, II, Pensylv. State Univ. Press.
WILLIAMS W.T. & CLIFFORD H.T. : On the comparison of two classifications of the same set of elements. Taxon, 1971, 20, 4, 519-522.
WILLIAMS W.T., CLIFFORD H.T., & LANCE G.N. : Group-size dependance : a rationale for choice between numerical classifications. The Comp. J., 1971, 14, 157-162.
WISHART D. : Treatment of missing values in cluster analysis. COMPSTAT 1980, Physica-Verlag, 281-287.
WOLFE J.H. : Pattern clustering by multivariate mixture analysis. Multiv. Behav. Res., 1970, 5, 329-350.
ZAHL.S. : A comparison of three methods for the analysis of spatial pattern. Biometrics, 1977, 33, 681-692.