Statistique et analyse des données

H. Brisse
G. Grandjouan

Un procédé de classification par agrégations d'un effectif nombreux

Statistique et analyse des données, tome 2, n° 3 (1977), p. 85-95
<http://www.numdam.org/item?id=SAD_1977__2_3_85_0>

Statistique et Analyse des Données
3 - 1977

un procédé de classification par agrégations d'un effectif nombreux

B R I S S E H. & G R A N D J O U A N G.

Institut de Botanique, 28, rue Goethe
67083 STRASBOURG CEDEX

RESUME

Le procédé proposé économise une partie du calcul des distances entre les éléments à classer et, cependant, il aboutit à un résultat pratiquement identique à celui des classifications usuelles par agrégations (du type "W.P.G.M."). L'économie est obtenue par le choix d'éléments - repères successifs et par l'application de l'inégalité triangulaire à la détermination du plus proche voisin de chaque élément. L'économie atteint respectivement 50 %, 70 % et 80 % du tableau complet des distances pour des effectifs de 50, 100 et 450 éléments.

SUMMARY

A method of classification by aggregative clustering for large sets. - The hierarchy produced by this method is quite similar to the usual ones (W.P.G.M., for instance), but it is produced by shorter computations. It uses only a part of the table of distances between the units to be classified. The economy is obtained by choosing successive marks among the units and by using the triangle inequality to determine the nearest neighbour for each unit. When applied to respectively 50, 100 and 450 units, the method saves the computation of 50 %, 70 % and 80 % of the complete table of distances.

1. INTRODUCTION

Les classifications par agrégations ont l'avantage de pouvoir être précises, mais elles ont l'inconvénient d'être coûteuses. Elles peuvent être précises car elles tiennent compte de toutes les distances entre les éléments à classer, et elles peuvent minimiser leurs déformations. Elles sont coûteuses parce que le calcul du tableau des distances est volumineux. Le procédé proposé économise le calcul d'une partie du tableau des distances. Cependant, il aboutit au même résultat que les procédés usuels, ou à des résultats extrêmement similaires. L'économie de calculs est obtenue grâce à l'utilisation de repères successifs et grâce à l'application de l'inégalité triangulaire. Indiquons que les éléments sont classés en fonction de leurs distances euclidiennes dans l'espace défini par leurs coordonnées. Un groupe d'éléments est caractérisé par son centre de gravité. Le procédé de référence, auquel est comparé le procédé proposé, est donc l'agrégation avec pondération des groupes, appelé "W.P.G.M." (SOKAL et SNEATH, 1973).

2. DESCRIPTION DU PROCEDE

2.1. Algorithme d'agrégation

2.1.1. Expression de l'algorithme proposé

Deux éléments I et J sont agrégés lorsque chacun d'eux est le plus proche voisin de l'autre (en abrégé : PPV), ce qui se traduit par les conditions (1) et (2) réalisées simultanément.
(1) I = PPV (J)
(2) J = PPV (I)
Ces conditions entraînent l'agrégation de I et de J quelle que soit la valeur de la distance D (I, J) qui sépare ces deux éléments.

2.1.2. Comparaison avec l'algorithme usuel

L'algorithme usuel agrège les deux éléments I et J les plus proches. Ces deux éléments sont également les plus proches voisins l'un de l'autre, mais l'algorithme usuel ajoute une condition supplémentaire aux deux précédentes :
(3) D (I, J) est la plus petite de toutes les distances.
Il agrège les PPV dans l'ordre de leurs distances croîssantes, tandis que l'algorithme proposé les agrège dans un ordre quelconque. Les deux algorithmes aboutissent cependant à des résultats très similaires, car deux éléments qui sont initialement les PPV l'un de l'autre, le restent généralement jusqu'à ce qu'ils soient agrégés, quel que soit l'ordre des agrégations des autres éléments. En effet, les distances ont tendance à augmenter au cours des agrégations ; la distance D (I, J) est généralement inférieure à la distance D (I, JK), JK étant le groupe constitué par l'agrégation de J et K. Toutefois, le contraire peut se produire, mais il est rare : il peut modifier un couple de PPV et entraîner de faibles différences entre

les résultats des deux algorithmes. Par exemple, imaginons quatre éléments IJKL situés approximativement aux sommets d'un losange, IJ et KL étant deux couples de PPV réciproques. L'agrégation de IJ peut créer un groupe qui devienne le PPV de K à la place de L. Le phénomène symétrique pourrait se produire, si l'agrégation de KL avait lieu la première. Le résultat dépend donc de l'ordre dans lequel se produisent les agrégations. D'ailleurs, il ne semble pas que l'ordre défini par les distances croîssantes soit meilleur ni moins bon qu'un autre. Nous donnerons quelques exemples numériques au paragraphe 3.

2.1.3. Programmation de l'algorithme proposé

Dans le programme, l'algorithme proposé est traduit sous une forme légèrement différente, de façon à éviter un blocage, dans le cas d'une chaîne de PPV équidistants. C'est le cas, par exemple, de trois éléments I, J, K, situés aux sommets d'un triangle équilatéral. I pourrait être attribué comme PPV à J, J à K et K à I. Les conditions (1) et (2) provoqueraient un blocage de ces agrégations, alors que les conditions (4) et (5) les exécutent dans l'ordre où les éléments sont examinés. Dans la condition (4) DMIN (I) désigne la distance de I à son PPV.

(4) DMIN (I) = DMIN (J) $\dot{+}$ ε

(5) I = PPV (J) ou J = PPV (I)

2.2. Utilisation de l'inégalité triangulaire

2.2.1. Etablissement d'un ordre parmi les éléments à classer

L'algorithme proposé consiste donc à déterminer les PPV des éléments, pris un par un, et à agréger les PPV réciproques dès qu'ils sont déterminés. Cet algorithme est appliqué aux éléments pris dans l'ordre de leurs distances croîssantes à l'un des éléments, choisi arbitrairement comme repère. L'ordre des éléments défini par leurs distances de repérage, est utilisé à deux fins :

- c'est l'ordre de l'examen des éléments i dont on recherche le PPV ;
- pour chaque élément I, c'est aussi l'ordre de l'examen des éléments K susceptibles d'être le PPV de I.

On calcul les distances D (I, K) pour savoir quelle est la plus petite. Le rangement des éléments K permet d'arrêter la recherche du PPV dès que la distance D (R, K) dépasse une borne ; c'est la condition (6) dans laquelle J est l'élément le plus voisin de I, parmi les éléments K déjà considérés.

(6) D (R, K) > D (R, I) + D (I, J)

En effet, cette condition (6) a pour conséquence que D (I, K) est supérieur à D (I, J), c'est-à-dire que K ne peut pas être le PPV de I. Cette conséquence se déduit de l'inégalité (7) dans le triangle RIK.

(7) D (I, K) > D (R, K) - D (R, I)

Les figures 1 et 2 schématisent la recherche des PPV puis l'agrégation des PPV réciproques. La condition (6) entraîne une grande économie de

Figure 1
SCHEMA DE LA RECHERCHE
DES PPV (données fictives)

Un point représente un élément dans l'espace des variables. R est le repère. Les autres éléments sont numérotés dans l'ordre où ils sont examinés, en s'éloignant de R. Le petit cercle a pour rayon D (R, 3) + D (3, 1) ; le grand cercle, D (R, 6) + D (6, 7). La recherche du PPV de 3 est limitée par le petit cercle, celle de 6 par le grand cercle. R suffit à montrer, par exemple, que 6 ne peut pas être le PPV de 3, car D (R, 6) - D (R, 3) > D (3, 1).

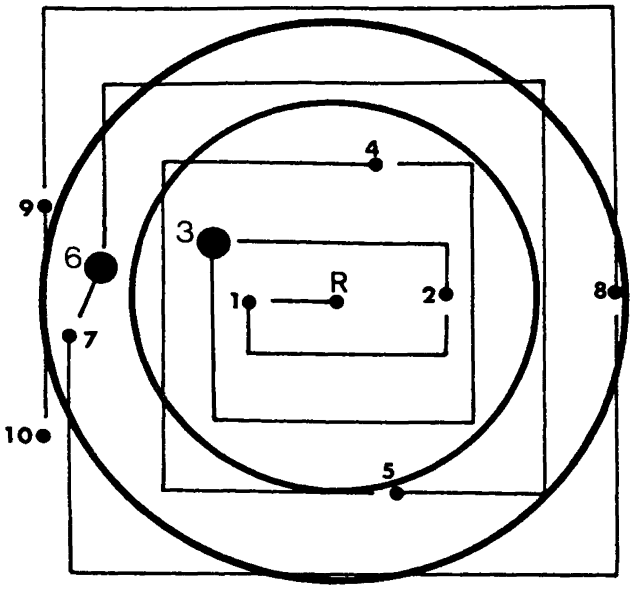


Figure 2
SCHEMA DE L'AGREGATION
DES PPV RECIPROQUES
au fur et à mesure de la détermination des PPV. (données fictives)

L'agrégation de deux éléments ou groupes d'éléments est schématisée par un contour.

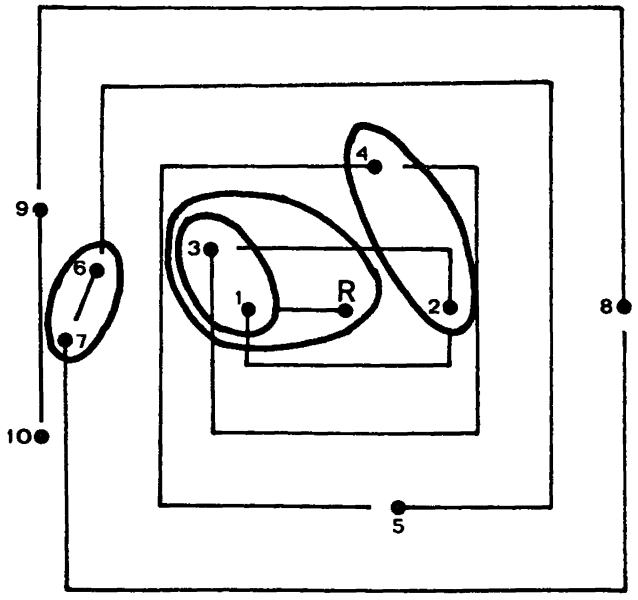
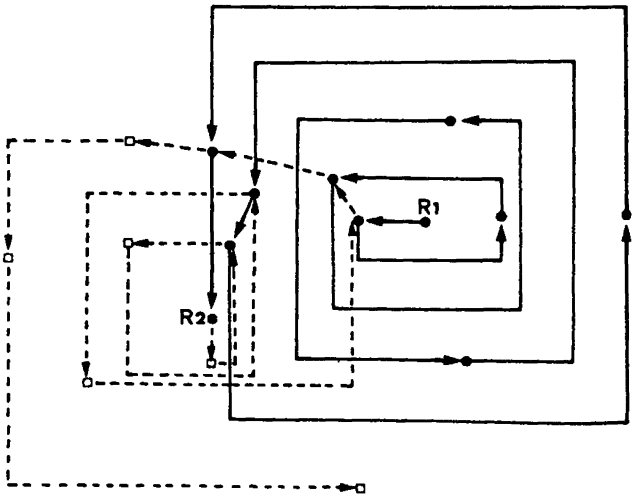


Figure 3
SCHEMA DE L'ORDRE
D'EXAMEN DES ELEMENTS
par deux repères successifs.
(données fictives)

Les points noirs représentent les éléments repérés par R1, les points blancs ceux qui le sont par R2. Les flèches indiquent l'ordre d'examen des éléments.



calculs, par rapport aux procédés usuels, dans lesquels il faut calculer les distances entre I et tous les autres éléments K, pris dans un ordre quelconque, pour trouver quelle est la distance la plus petite.

L'agrégation des PPV réciproques dans un ordre indépendant de leurs distances mutuelles a un intérêt pratique. Elle permet de comparer les éléments voisins, de proche en proche, et de les agréger par des opérations souvent consécutives, ce qui diminue rapidement l'effectif des éléments qui restent à classer. D'une façon imagée, on pourrait baptiser l'algorithme proposé "algorithme de la chèvre", car il traite les éléments situés tout autour d'un repère, de même que la chèvre broute tout autour de son piquet, le piquet étant déplacé, de proche en proche, dans toute l'étendue du pré. Par contre, l'algorithme usuel pourrait être baptisé "algorithme de la puce", car il agrège les PPV dans l'ordre des distances croîssantes et que, par conséquent, il traite, l'un après l'autre, des éléments souvent éloignés. Il est comparable à une puce qui, à chaque agrégation, saute d'un bout du nuage à l'autre.

2.2.2. Changement de l'élément - repère (figure 3)

a) Critère du changement de repère

Un repère R est d'autant plus efficace qu'il est plus proche de l'élément I dont on détermine le PPV. L'efficacité se mesure au nombre d'éléments K qui sont éliminés par la condition (6), sans qu'on ait besoin de calculer la distance D (I, K). Par contre, les éléments H situés à la même distance du repère que I, ou à une distance inférieure, sont indissociables de I par la condition (6), c'est-à-dire qu'il faut calculer les distances D (I, H). Le nombre de ces éléments H augmente comme le volume de la sphère de rayon D (R, I), dans l'espace des coordonnées. Par conséquent, plus on s'éloigne du repère, plus le nombre de distances à calculer augmente. C'est pourquoi, le repère est changé, après avoir servi pour un certain nombre d'éléments, afin qu'il reste proche des éléments considérés. Le repère est changé lorsque la condition (8) est satisfaite.

(8) $NV > NQ \times NR$

NV désigne le nombre de distances D (I, K) calculées pour déterminer des PPV ; NR désigne le nombre de distances D (R, I) calculées pour le tableau de repérage. La condition (8) traduit une proportion à respecter entre NR et NV, NR exprimant le coût du rangement des éléments et NV exprimant le coût des opérations effectuées sur la base de ce rangement. La valeur NQ est fixée d'une façon empirique ; c'est celle qui permet les meilleures performances. La condition (8) influe sur l'économie du procédé, mais non sur l'exactitude des résultats.

b) Calcul du nouveau tableau de repérage

Le calcul du nouveau tableau des distances de repérage est abrégé par une seconde utilisation de l'inégalité triangulaire. Elle consiste à remplacer les distances de repérage par des bornes minorantes, pour les élé-

ments les plus éloignés du repère. Elle s'applique à un élément K lorsque sa précédente distance de repérage, DR (K), est supérieure à un seuil, choisi empiriquement (condition (9)). La borne minorante, BM, est alors calculée par la relation (10), qui exprime une inégalité du triangle constitué par l'élément K, le repère précédent Q et le nouveau repère R.

(9) DR (K) > 2 x D (Q, R)

(10) BM = DR (K) - D (Q, R)

(11) BM < D (R, K)

La relation (10) entraîne une économie, car elle est plus rapide à calculer qu'une distance. La condition (9) et la relation (10) ont pour effet de ranger correctement les éléments proches du nouveau repère, et de sous-estimer les distances de repérage des éléments éloignés. Ces distances, quoique sous- estimées, permettent quand même d'éliminer beaucoup d'éléments de la recherche des PPV, puisque cette recherche est cantonnée dans le voisinage du repère, et que la distance de deux repères successifs est généralement faible. La borne minorante d'une distance diminue à chaque changement de repère ; lorsqu'elle devient inférieure au seuil choisi, la distance de repérage est calculée exactement et l'élément correspondant reprend son rang correct. Comme le précédent critère empirique (condition (8)), le seuil choisi influe sur l'économie du procédé mais non sur l'exactitude des résultats. L'exactitude est garantie par la condition (6) qui permet de limiter la recherche d'un PPV à une partie des éléments et qui donne cependant l'assurance de déterminer le vrai PPV de chaque élément. La condition (6) donne la même assurance si on y remplace D (R, K) par sa borne minorante DR (K).

2.2.3. Utilisation de repères accessoires

Certains éléments peuvent être utilisés comme repères accessoires. Ce sont ceux dont on a déjà déterminé les PPV et dont on a calculé et stocké les distances avec un certain nombre d'éléments. Un repère accessoire S peut épargner le calcul de la distance D (I, K) à deux conditions :

(12) D (S, I) et D (S, K) ont été calculées toutes les deux ;

(13) D (S, I) - D (S, K) > D (I, J)

La condition (13) utilise une inégalité du triangle SIK, et détermine une borne minorante de D (I, K). Si cette borne est supérieure à D (I, J) qui est la plus petite des distances déjà calculées pour I, K ne peut pas être le PPV de I.

En pratique, on stocke en mémoire centrale les distances calculées pour les derniers éléments examinés, à concurrence de la place disponible. Beaucoup de ces éléments sont proches de l'élément dont on recherche le PPV, puisqu'ils sont, les uns et les autres, proches du repère principal en fonction. Un repère accessoire épargne le calcul d'un certain nombre de distances, mais il ne permet pas de mettre un terme à la recherche d'un PPV. Seul le repère principal le permet, car c'est lui qui détermine l'ordre d'examen des éléments.

3. RESULTATS NUMERIQUES

Il existe deux versions du programme de classification utilisant l'inégalité triangulaire.

- La première version, la plus simple, stocke les coordonnées en mémoire centrale. Elle ne stocke pas les distances calculées et, par conséquent, elle n'utilise pas de repères accessoires. Elle s'applique à un tableau de coordonnées ayant une taille maximale de 64 K - mots (sur UNIVAC 1110). Par exemple, elle classe un tableau de 521 éléments à 36 coordonnées en 6 mn (dont 12 s d'entrées et sorties), en occupant 33 K - mots en mémoire centrale.
- La deuxième version du programme stocke les coordonnées sur tambour; elle s'applique donc à un tableau de taille théoriquement illimitée. Elle convient aux éléments caractérisés par des coordonnées nombreuses. Elle stocke aussi sur tambour les distances calculées pour chaque élément, car il est plus rapide de les relire que de les recalculer. Les résultats suivants, obtenus avec la deuxième version du programme donnent des informations sur l'économie et sur l'exactitude du procédé.

3.1. Economie du procédé

L'économie est mesurée à partir du nombre de distances calculées, comparé au nombre de distances qu'il faudrait calculer avec le procédé usuel, basé sur le tableau complet des distances. Pour des effectifs de 50, 100 et 450 éléments, l'économie atteint respectivement 50 %, 70 % et 80 %. Elle augmente avec l'effectif à classer, mais elle augmente de moins en moins vite. Pour 450 éléments, caractérisés chacun par 1020 variables, la classification demande 33 minutes et elle occupe environ 60 K - mots en mémoire centrale avec un ordinateur UNIVAC 1110.

3.2. Exactitude du procédé

Pour comparer deux hiérarchies obtenues par des procédés différents appliqués aux mêmes données, on déduit de chacune des hiérarchies un tableau de distances entre les éléments pris deux par deux. La concordance C de deux hiérarchies est exprimée par la similitude globale des deux tableaux de distances correspondants. Dans la formule (14), D et D' désignent respectivement les distances de l'un et l'autre tableau, et les sommes portent sur toutes les distances.

$$(14)\quad C = 100 \times (\sum D^2 - \sum (D - D')^2) / \sum D^2$$

Les comparaisons suivantes portent sur un même lot de cent éléments.

3.2.1. Influence du changement de repère

Le premier repère est choisi arbitrairement. Il conditionne la détermination, en chaîne, des repères ultérieurs. Le procédé proposé a été appliqué en prenant successivement comme premier repère, l'élément numéro 1, puis numéro 50, puis numéro 100. Les concordances entre la première hiérarchie et les deux autres sont respectivement égales à 99.9 et 99.8 %. Le changement de repère initial ne modifie pratiquement

pas les résultats, comme le prévoyait le principe du calcul.

3.2.2. Comparaison avec un procédé de référence (W.P.G.M.)

La concordance entre les distances initiales et les distances déduites de la hiérarchie vaut 80 % pour le procédé proposé comme pour le procédé de référence. Les agrégations faites dans l'ordre des distances croîssantes n'aboutissent donc pas à une meilleure justesse que dans un ordre quelconque.

La concordance entre les distances déduites des deux hiérarchies vaut 99.8 %. Les deux dendrogrammes correspondants montrent, en effet, des subdivisions presque toutes identiques, notamment toutes celles qui sont bien individualisées. Une différence se produit lorsque l'augmentation générale des distances, dans les agrégations successives, est faible, et lorsqu'une diminution de distance, provoquée par une agrégation particulière, suffit à modifier un PPV. Cela entraîne l'interversion de deux agrégations, à des distances voisines et, par conséquent, des différences minimes dans le tableau des distances déduites de la hiérarchie.

Précisons enfin que ces différences se produisent seulement pour les algorithmes dans lesquels la distance d'un élément à son PPV est susceptible de diminuer, si celui-ci s'agrège à un autre élément. C'est le cas de la distance des centres de gravité. Ce n'est pas le cas pour d'autres algorithmes, tels que la distance moyenne, ou tout autre moment des distances entre les éléments d'un groupe et ceux de l'autre. Pour ces algorithmes-là, deux PPV réciproques le restent jusqu'à ce qu'ils soient agrégés et. par conséquent, les résultats du procédé proposé seraient totalement identiques à ceux du procédé de référence.

4. DISCUSSION DU PROCEDE

4.1. Limitation des effectifs classables par agrégations

Les procédés usuels s'appliquent à un effectif limité par le nombre de distances à calculer et par la dimension du tableau des distances à stocker. Le nombre de distances à calculer augmente comme le carré de l'effectif ou même, parfois, comme le cube, si la place en mémoire ne permet pas de stocker le tableau des distances et si, par conséquent, il faut le recalculer à chaque agrégation. Rappelons que le calcul de tout un tableau de distances a pour seul but de trouver une valeur minimale ; il y a là du gaspillage. Cette limitation due au volume des calculs a été souvent signalée comme étant un handicap essentiel des classifications par agrégations. BERTIN (1973) constate que "le mur de la combinatoire est vite atteint" ; ANDERBERG (1973) signale que les classifications par agrégations ne peuvent pas dépasser quelques centaines d'éléments. SOKAL et SNEATH (1973 concluent qu'elles ne sont pas d'application générale en taxinomie.

4.2. Versions successives du procédé

Le procédé proposé est la synthèse de procédés partiels utilisés initialement pour limiter le calcul des distances. Trois opérations avaient lieu l'une après l'autre :

- repérage des éléments,
- établissement du voisinage de chaque élément,
- agrégations successives des éléments et des voisinages.

Le voisinage d'un élément est défini comme l'ensemble des "N" éléments les plus proches de lui, N étant constant. La taille d'un voisinage varie avec la densité des éléments. Les calculs sont ensuite limités aux distances entre un élément et ses voisins. Deux éléments sont agrégés lorsqu'ils sont séparés par la plus petite distance (condition (3)). Lorsque deux éléments s'agrègent, leurs voisinages s'additionnent. Il est possible de contrôler les résultats de ce procédé sans recourir au calcul complet que les voisinages ont justement pour but d'éviter. Le contrôle consiste à classer un ensemble plusieurs fois de suite, en augmentant à chaque fois l'effectif des voisinages. Lorsque les résultats ne changent plus, on peut montrer qu'ils sont identiques à ceux du procédé usuel. Les calculs ont porté sur un tableau de 450 éléments avec 1020 coordonnées. L'établissement des voisinages de dix éléments a demandé 35 minutes ; la classification avec des voisinages de 2, 4 et 6 éléments a demandé respectivement 7, 11 et 18 minutes.

La séparation des trois opérations précédentes laissait subsister plusieurs causes de gaspillage dans les calculs, principalement la fixation arbitraire du nombre d'éléments par voisinage. C'est pourquoi l'effectif du voisinage a été ramené à un seul élément (c'est le PPV) et un nouvel algorithme a été utilisé (l'agrégation des PPV réciproques).

4.3. Intérêt d'une classification par agrégations

Une classification par agrégation permet une bonne concordance entre les distances initiales et les distances déduites de la hiérarchie. Cette concordance est obtenue pour les petites comme pour les grandes distances, puisqu'à chaque agrégation, la distance de deux groupes d'éléments est aussi voisine que possible des distances qu'elle remplace. Cette concordance est incomplètement exprimée par la concordance globale de deux tableaux de distances (formule (1)), car celle-ci dépend principalement des distances les plus grandes. Imaginons, par exemple, qu'à un ensemble de cent plantes des régions tempérées, on ajoute une plante tropicale, située très loin des autres, dans un espace climatique. Toute classification sépare, à coup sûr, la plante tropicale des 100 autres, et peut atteindre une concordance globale élevée, même si les 100 plantes tempérées sont mal classées. Il serait utile de compléter la valeur de la concordance globale par la distribution des écarts entre les distances, dans laquelle chaque couple de distances ait le même poids, par exemple la distribution des écarts moyens. Dans la comparaison entre le procédé

usuel et le procédé proposé (par. 3.2.1.) 80 % des écarts sont inférieurs à 1 %, et 90 % des écarts sont inférieurs à 10 %. Un autre critère de justesse pourrait être la corrélation des rangs des distances D et D', de la formule (14), car ce calcul donne le même poids à toutes les valeurs.

4.4. Comparaison avec d'autres procédés de classification

La classification d'un effectif nombreux est possible par divisions (DIDAY, 1975). Une classification par divisions utilise quelques éléments comme centres, et elle se base sur les distances entre les centres et les éléments. Elle ne peut donc minimiser que la déformation de ces distances-là. Elle aboutit donc à une concordance avec les données initiales qui est probablement moins bonne que par agrégations ; mais elle est moins coûteuse. D'autres procédés commencent par une classification par divisions afin de réduire l'effectif à étudier (ANDERBERG, 1973 ; BENZECRI, 1973 ; BRIANE et al., 1977). D'autres, enfin, négligent une partie des données, en considérant comme nulles les distances inférieures à un seuil arbitraire (ROSS, 1969 ; PARKER-RHODES et JACKSON, 1969).

Tous ces procédés ont en commun d'utiliser, comme critères essentiels, les distances les plus grandes. Celles-ci sont aussi bien conservées que possible. Les autres distances sont, dans l'ensemble, d'autant moins bien conservées qu'elles sont plus petites. Dans ce cas, la distribution des écarts pourrait montrer les discordances parmi les distances inférieures. L'augmentation du domaine d'application des classifications est ainsi obtenue au prix d'une perte de précision dans le détail des résultats. Il y a là un risque de retrouver des phénomènes majeurs, déjà connus, et de brouiller l'expression de phénomènes subordonnés originaux. Ce risque existe non seulement dans la réduction du nombre d'éléments par une classification, mais aussi dans la réduction du nombre de variables, par une analyse factorielle.

REFERENCES BIBLIOGRAPHIQUES

ANDERBERG, M.R., 1973 - Cluster analysis for applications. New-York and London, Academic Press, 359 p., 130 ref., index.

BENZECRI, J.P. et al., 1973 - L'analyse des données. Paris, Dunod, 615 p.

BERTIN, J., 1973 - Le traitement graphique de l'information. Actes du 1er colloque Informatique et biosphère, Paris, Informatique et biosphère, 2-48, 49 fig.

BRIANE, J.P., LAZARE, J.J. et SALANON, R., 1977 - Le traitement des très grands ensembles de données en analyse factorielle des correspondandes. Proposition d'une méthodologie appliquée à la phytosociologie. Université de Paris XI, 91405 Orsay, laboratoire de taxonomie végétale, 63 p.

DIDAY, E., 1975 - Classification automatique séquentielle pour grands tableaux. Revue française d'automatique informatique et recherche opérationnelle, B-1 : 29-61.

PARKER-RHODES, A.F. et JACKSON, D.M., 1969 - Automatic classification in the ecology of the higher Fungi. Numerical taxonomy (COLE, edit.), Academic Press, 181-215.

ROSS, G.J.S., 1969 - Classification techniques for large sets of data. Numerical taxonomy (A.J. COLE, edit.), Academic Press, 224-233.

SOKAL, R.R. et SNEATH, P.H.E., 1973 (1ère édition 1963). - Principles of numerical taxonomy. San Francisco and London, FREEMAN W.H. and Co.