STRATEGIE DE CLASSIFICATION POUR DE GRANDS ENSEMBLES DE DONNEES

J.L. MOLLIERE

E.D.F.
Direction des Etudes et Recherches
Service IMA
1, avenue du Général de Gaulle
92141 CLAMART
Tél. 47.65.43.21

Résumé : L'objectif principal est ici le choix du nombre "optimal" de classes comme préalable à la solution du problème de classification sur une population de taille importante. Des réponses satisfaisantes, expérimentées en pratique, sont ici apportées, s'appuyant sur l'utilisation de critères adéquats et la mise en oeuvre des algorithmes classiques de classification (hiérarchique et non hiérarchique) pour leur optimisation.

Mots-clés :

Classification automatique, classifiabilité, nombre de classes, critère de la variance, critère de De La Vega, nuées dynamiques, K-means, méthode de Ward, formes fortes.

Introduction

En dépit de la difficulté à définir des critères de choix d'un nombre de classes optimal dans une population à classifier, fondés sur des tests rigoureux se référant à des hypothèses nulles de non classificabilité ou à des hypothèses alternatives d'existence d'une structure de classification [1], les utilisateurs sont bien amenés dans la pratique à s'appuyer sur des critères remplissant pour le moins la fonction d'indicateurs.

Le plus courant d'entre deux, pour des attributs numériques, est le critère R^2 de variance expliquée par une partition. L'intérêt des méthodes de classification hiérarchique est alors d'offrir globalement l'évolution des valeurs d'un tel critère aux différents niveaux.

Nous voulons dans cet article, d'une part montrer l'intérêt pratique de combiner classification non hiérarchique (spécialement par la méthode des nuées dynamiques) et classification hiérarchique, dès que les populations à classifier dépassent quelques centaines d'individus, d'autre part montrer que d'autres critères, ayant de meilleures propriétés que le critère R^2 sont dès à présent utilisables pour apprécier le caractère classifiable d'une population et déterminer le nombre de classes le plus significatif.

Une expérience de plusieurs années d'utilisation du "Cubic Clustering Criterion" (CCC) dans le cadre du logiciel SAS [2], pour des variables quantitatives, nous amènera à proposer une stratégie de classification basée sur l'utilisation de ce critère
Par ailleurs, nous montrerons que cette stratégie générale combinant classification non hiérarchique et classification hiérarchique peut également s'appliquer sur des données non métriques (indices de similarité).

A - Données numériques - Utilisation du critère du CCC

Les méthodes métriques de classification automatique, fondées sur le calcul de distances (dans un espace approprié), mesurent la qualité d'une partition par le critère R^2 de variance expliquée (rapport de la variance inter-classes entre les centres pondérés des classes à la variance totale).

Ce critère classique, compris entre 0 et 1, a évidemment une tendance naturelle à croître avec le nombre de classes, ce qui rend délicat son utilisation pour comparer deux partitions en des nombres de classes différents.

Le critère du CCC dû à Warren S. Sarle [3] et popularisé par le logiciel SAS vise à mesurer le caractère plus ou moins significatif d'une partition en référence à une hypothèse nulle d'absence de classes (une seule distribution statistique uniforme de l'ensemble des unités à classifier).

De cette façon, on peut comparer de manière plus objective différents niveaux de partition de la population à classifier.

D'autre part, ce critère présente l'avantage de mettre en évidence des populations à priori peu classifiables.

On montrera dans ce cas l'intérêt de la notion des formes fortes (attachée à la méthode des nuées dynamiques) pour extraire une sous-population mieux structurée.

La validité du CCC, comparé à d'autres critères, du point de vue de sa capacité à retrouver une structure en classes connue à l'avance, a été prouvée par MILLIGAN et COOPER [4]

I. Le critère du CCC

Le CCC est obtenu en comparant la valeur du R^2 de la partition sur les données réelles à ce que serait la valeur attendue du R^2 (soit $E(R^2)$) pour le même nombre de classes si la population était issue d'une distribution uniforme sur un parallélipipède rectangle de dimension p (p étant le nombre de variables).

On peut trouver une formule d'approximation de $E(R^2)$ dépendant de :

- n = nombre d'unités à classifier
- q = nombre de classes
- p = nombre de variables

Le CCC est alors calculé par la formule :

$$CCC = Ln\left[\frac{1 - E(R^2)}{1 - R^2}\right] \times a$$

a est un coefficient empirique pour stabiliser la variance du CCC lorsque varient les paramètres n, p et q.

La formule exacte de calcul du CCC est donnée en Annexe 1.

Suivant les valeurs positives ou négatives du CCC on concluera au rejet ou à l'acceptation de l'hypothèse nulle d'absence de classes.

Une partition sera jugée d'autant plus significative que la valeur positive du CCC est élevée.

La véritable hypothèse alternative est, pour la population étudiée, d'être issue d'un mélange de distributions multi-normales, de mêmes matrices de variance-covariance, avec des probabilités égales d'échantillonnage pour les différents composants

Il s'agit là des conditions d'utilisation optimales du R^2 dans les problèmes de classification [5], ce critère, comme le CCC, n'étant donc pas adapté à la reconnaissance de classes de forme allongée ou irrégulière.

Deux remarques importantes peuvent être faites.

Remarque 1 :

Le calcul de E (R^2) servant de base au calcul du CCC est effectué à partir des p variables de classification en supposant que celles-ci sont non corrélées. Pour interpréter les valeurs du CCC, il faudra donc en général effectuer préalablement une A.C.P. (analyse en composantes principales) ou une A.F.C. (analyse factorielle des correspondances) à partir des variables initiales et utiliser les facteurs non corrélés obtenus, comme variables de classification.

Remarque 2 :

Une simulation de Monte-Carlo assez complète [3] a permis d'étudier le comportement du CCC pour la distribution uniforme de référence. Les valeurs le plus souvent négatives observées pour le CCC montrent le caractère généralement conservatif du test (trop sévère dans sa décision de rejet de l'hypothèse nulle).

Il est intéressant, pour des conditions voisines de celles dans lesquelles nous utiliserons le CCC (c'est-à-dire en particulier pour un nombre d'individus important) de noter pour la distribution uniforme de référence l'évolution des valeurs (négatives) du CCC en fonction du nombre de classes.

On doit noter sur la figure 1 que pour la distribution uniforme la valeur du CCC de référence croît avec le nombre de classes.

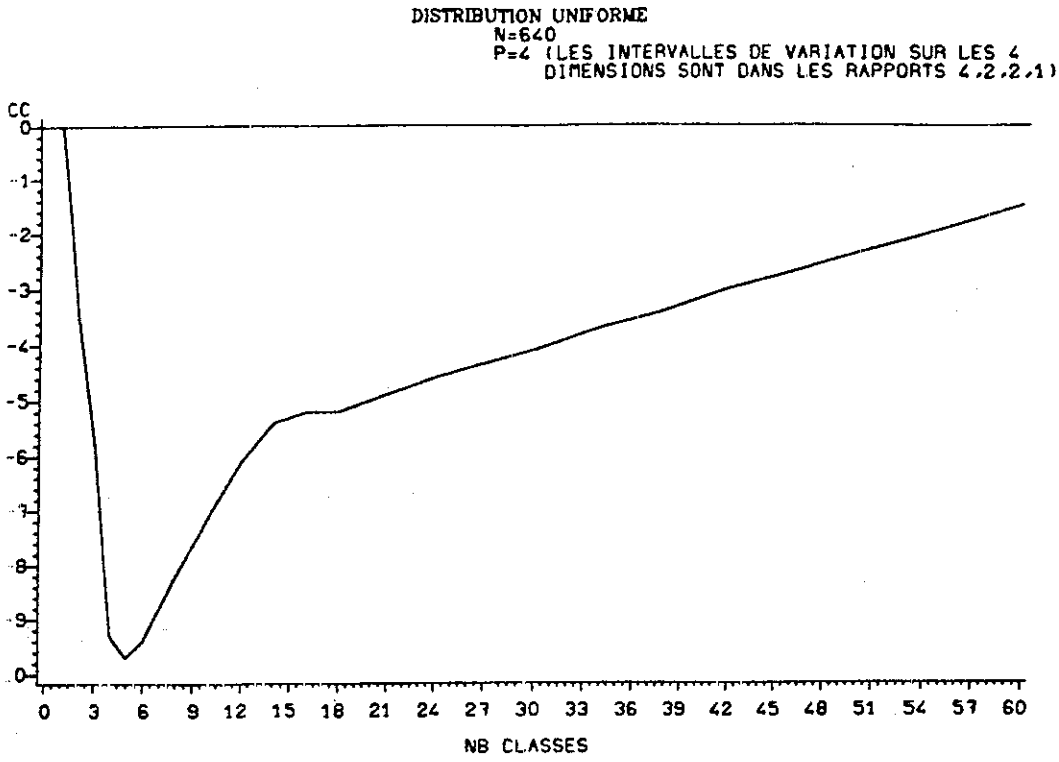


Figure 1

II. Une méthodologie pour classer les populations de grande taille

1- Principes

Deux idées-forces, étayées par des arguments d'ordre pratique et théorique sont à la base de cette méthodologie :

1) L'intérêt des méthodes de classification hiérarchique pour juger des niveaux de partition les plus significatifs.

La courbe d'évolution du critère de classification aux différents niveaux de regroupement de l'arbre hiérarchique peut donner une réponse globale au problème du choix du nombre de classes.

On préfèrera ici évidemment le CCC comme critère au classique R^2, en se rappelant toutefois que l'évolution du CCC sous l'hypothèse nulle n'est pas absolument indépendante du nombre de classes (I - Remarque 2).

2) L'inconvénient d'utiliser une méthode hiérarchique sur un nombre trop important d'unités statistiques à classer.

Au-delà de quelques centaines d'observations, il devient impossible de faire clairement apparaître une structure en classes plus significative dans les derniers niveaux d'agrégation de l'arbre qui sont justement ceux auxquels on s'intéresse en pratique.

Cela résulte du fait que les derniers niveaux de regroupement cumulent les erreurs d'écart à l'optimum des partitions obtenues à tous les niveaux précédents. Il est connu en effet qu'à un niveau donné, la partition n'est jamais optimale pour le critère de classification considéré, mais est d'une certaine manière (dépendant de l'algorithme d'agrégation) seulement localement optimale par rapport au niveau de partition précédent. Sur les courbes d'évolution du critère, pour des arbres agrégeant un très grand nombre d'unités à classer, on observe alors dans les derniers niveaux des valeurs très médiocres du critère de classification ainsi qu'une évolution très monotone de celui-ci, rendant difficile la détection de niveaux plus significatifs.
Les pointes de significavité sont amorties par le bruit important résultant de la somme des erreurs commises dans la suite des très nombreux regroupements.
Un tel type de courbe est donné pour le critère du CCC, sur la figure 2, pour une population de 1000 individus classée par la méthode hiérarchique de Ward (critère de la variance). Ce graphique devra être comparé à ceux obtenus en suivant la méthodologie que nous proposons.

Remarque :

Cette solution combinant classifications non hiérarchique et hiérarchique nous a paru à l'expérience [6] préférable à une méthode consistant à comparer les valeurs du critère pour différents nombres de classes, obtenues par un algorithme de classification non hiérarchique. La raison en est la variabilité des valeurs du critère pour un nombre de classes donné, pour différentes initialisations, rendant délicate et coûteuse en temps calcul la mise en évidence d'un nombre de classes optimal.

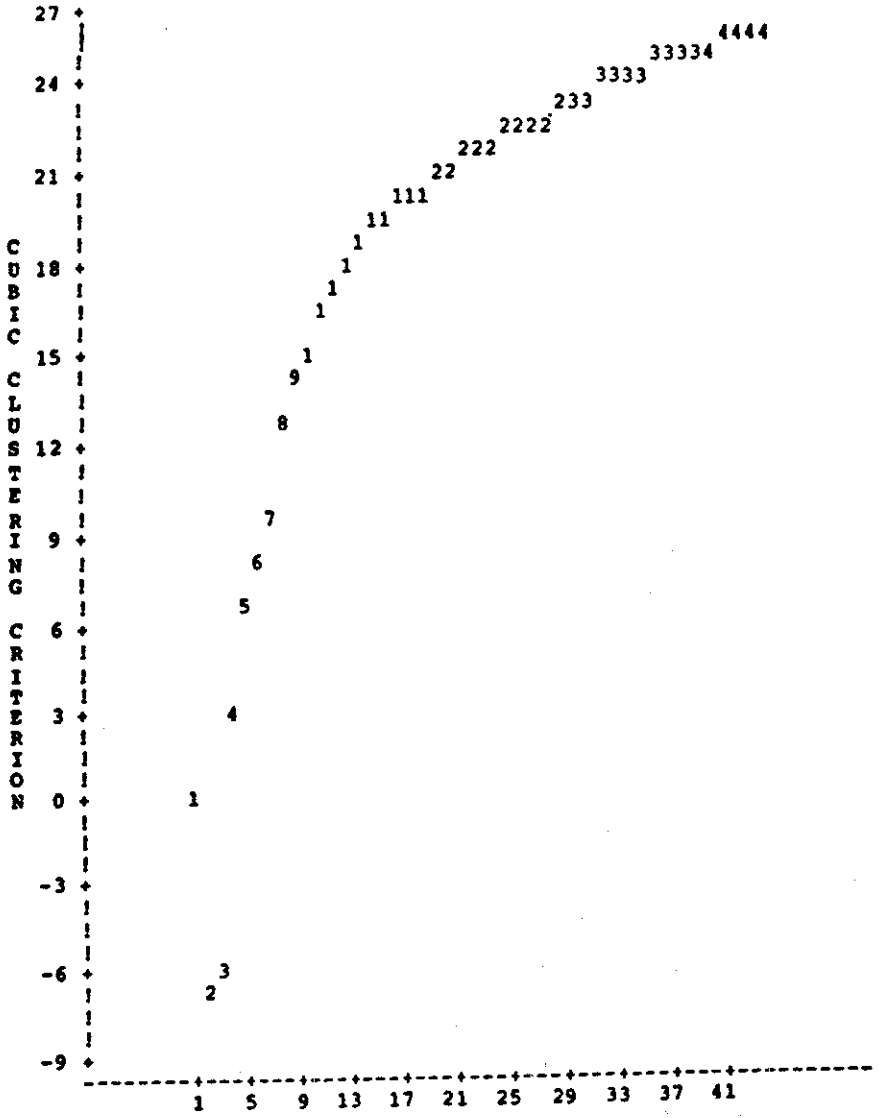


Figure 2

2. Mise en oeuvre

La méthodologie proposée procède en deux phases :

- obtention à partir de la population initiale à classer, d'une partition en un nombre élevé de classes (par exemple de l'ordre de la centaine).

- application d'une méthode de classification hiérarchique ascendante (suivant la méthode de WARD) sur les classes précédentes et analyse de la courbe d'évolution du critère de classification (CCC) aux différents niveaux.

La première phase a pour but de réduire le nombre d'observations soumises à la classification hiérarchique.

Pour cette première phase on peut soit :

- procéder à une classification non hiérarchique en un nombre fixé assez élevé de classes (standard égal à 100).
  Toutes nos classifications non hiérarchiques sont éffectuées avec une méthode de type K-Means (ou "centres mobiles")

  On obtient alors, avec la phase 2, la procédure que nous nommons MCLAS1

- réaliser une partition en "formes-fortes" [7] de la population initiale. Cette partition s'obtient en réalisant différentes classifications non hiérarchiques en un nombre de classes fixé (ce nombre compris en général entre 5 et 10 classes peut dépendre de l'ordre de grandeur du nombre de classes souhaité pour la partition finale). En standard, on réalisera 5 expériences avec des initialisations différentes, conduisant à 5 classifications correspondant chacune à un optimum local du critère de la variance.

  L'intersection de ces partitions (en général voisines) donnera un nombre plus ou moins élevé de "formes-fortes". Celles-ci réunissent des observations qui se sont toujours trouvées dans la même classe au cours des diverses expériences.

  La raison théorique (voir théorème 3 dans [8], énoncé dans le chapitre B) pour préférer les "formes-fortes" à une partition quelconque est qu'elles doivent normalement conduire dans la phase 2 (agrégation hiérarchique) à de meilleures valeurs du critère de classification aux différents niveaux, ce qui sera vérifié en pratique.

  On obtient ainsi la procédure que nous nommons MCLAS, un peu plus coûteuse en temps calcul que la procédure MCLAS1, mais fournissant de meilleures partitions, et parfois seule capable de donner une indication sûre du nombre de classes le plus significatif.

3- Obtention des classes finales

Les procédures MCLAS1 et MCLAS doivent permettre à l'utilisateur de déterminer le nombre de classes jugé le plus significatif, correspondant à un niveau de la classification hiérarchique (obtenue dans la phase finale de chacune de ces deux procédures).
Il est conseillé à l'utilisateur d'optimiser (du point de vue du critère de la variance) la partition correspondante en enchaînant un algorithme du type K-Means ou un algorithme d'échange susceptible de donner encore une meilleure valeur du critère.

III. Mode d'emploi des procédures MCLAS1 et MCLAS en fonction de la nature des données à classifier

1- Vérification de la stabilité des résultats

Le résultat recherché par la mise en oeuvre de ces deux procédures est l'indication d'un nombre de classes jugé le plus significatif Or dans chacune de ces procédures, la méthode suivie pour arriver au résultat dépend, dans sa première phase (partitionnement de la population en un nombre de classes élevé) de l'initialisation fournie à l'algorithme de classification non hiérarchique. Celui-ci est utilisé une fois dans la procédure MCLAS1 et cinq fois successives (pour le standard de 5 expériences) dans la procédure MCLAS. Il est donc important de vérifier, tous les autres paramètres de la méthode étant fixés, que l'indication finale obtenue pour le nombre de classes est stable vis-à-vis d'un changement d'initialisation de l'algorithme de classification.

Dans l'algorithme de K-Means utilisé pour nos essais, l'initialisation ne pouvait être modifiée que par un changement de l'ordre de rangement des observations sur le fichier initial. Une procédure de tirage aléatoire de cet ordre, paramétrée soit par l'heure de l'horloge (et donc non reproductible) soit par une valeur entière (et donc reproductible) figurant en paramètre de l'algorithme nous a permis très aisément de répéter l'exécution de l'algorithme considéré pour différentes conditions initiales.

2- Première analyse par la procédure MCLAS1

Nous conseillons de commencer l'analyse par l'emploi de la procédure MCLAS1.

D'une part cette procédure est peu coûteuse en temps calcul, d'autre part son utilisation est simple car ne dépendant que d'un seul paramètre : le nombre de classes de la partition initiale (standard égal à 100).

Nous suggérons dans un premier temps de garder ce standard et de répéter plusieurs fois l'exécution de MCLAS1 (avec des initialisations différentes évidemment).

Deux résultats sont attendus de ces premiers essais, à lire sur la courbe d'évolution du CCC en fonction du nombre de classes : le caractère classifiable ou non de la population, et si oui l'indication du nombre de classes à retenir.

Une courbe du type de la figure 3 correspond à une population non classifiable. Le CCC reste très négatif dans toute la zone d'évolution du nombre de classes représentée sur le graphique et particulièrement entre 5 et 10 classes. Pour traiter ce type de population il faut se reporter au paragraphe 4.

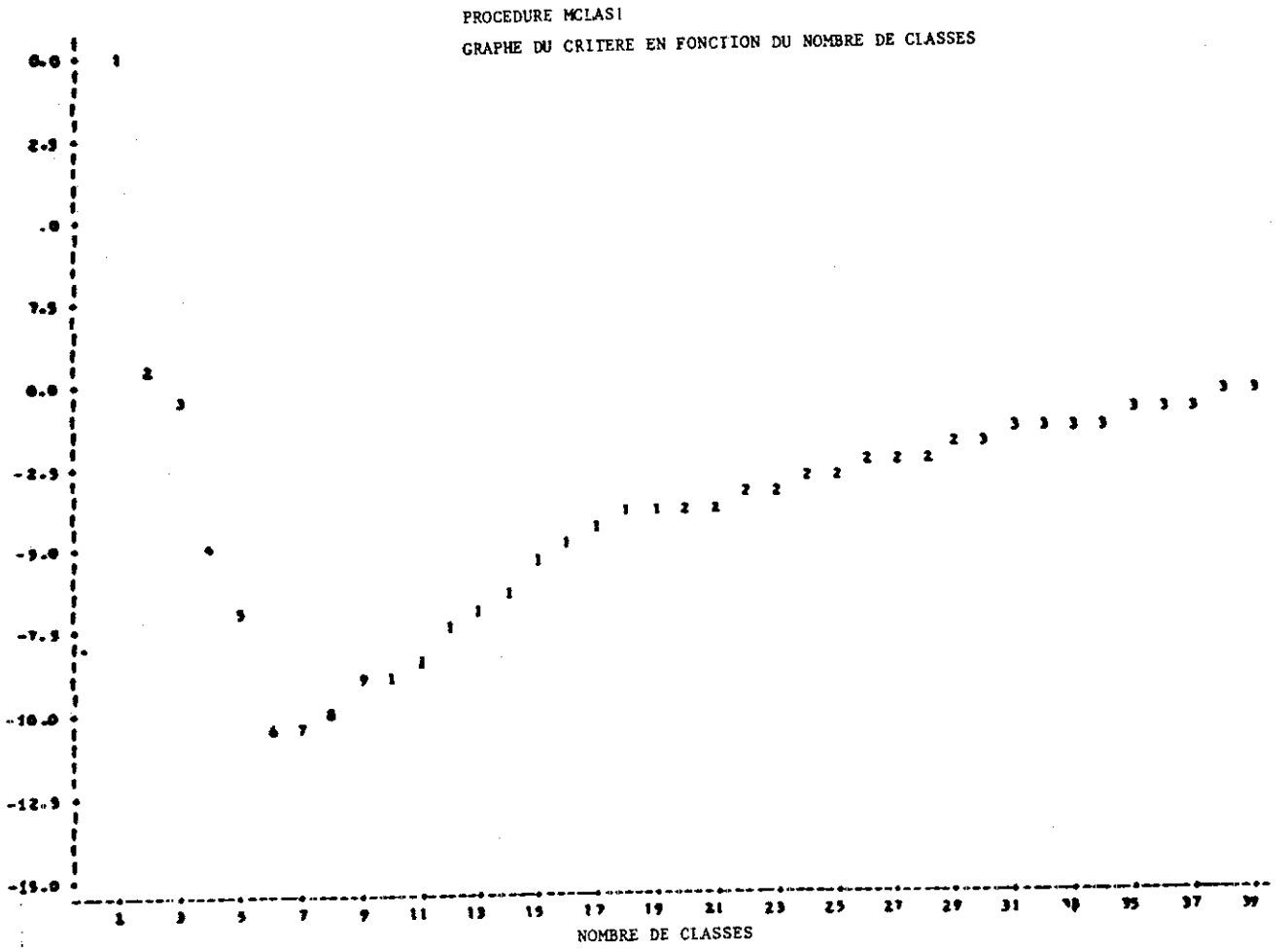


Figure 3

Des courbes du type de la figure 4 (valeurs positives élevées du CCC) ou du type de la figure 5, généralement rencontré (CCC prenant des valeurs positives seulement à partir d'un certain nombre de classes, même élevé, et continuant une croissance positive), révèlent une population classifiable.

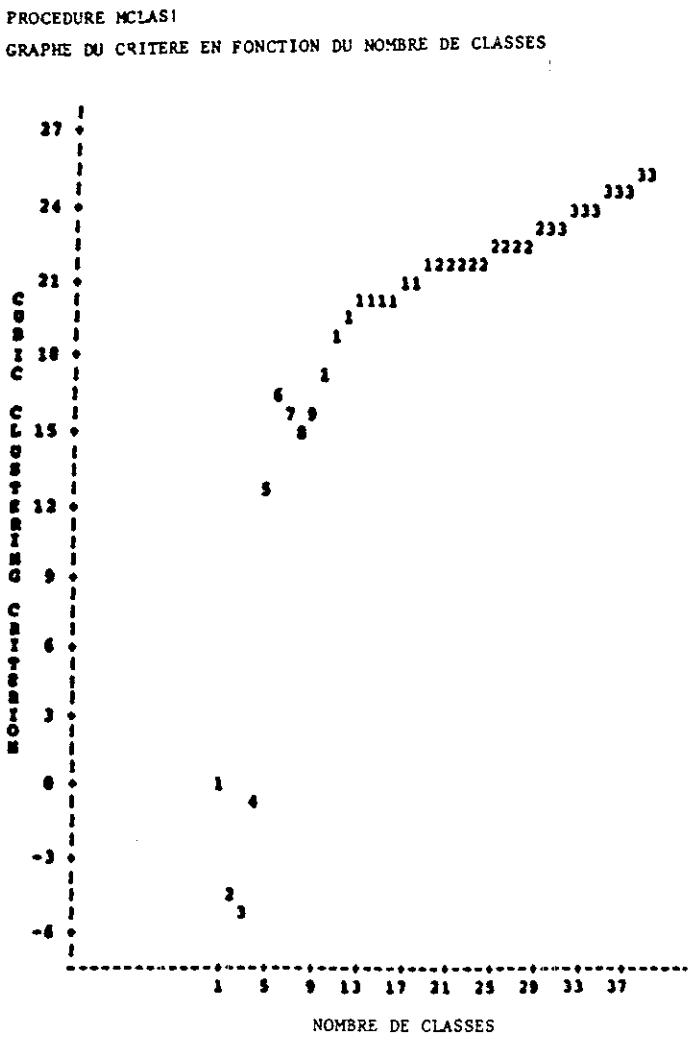


Figure 4

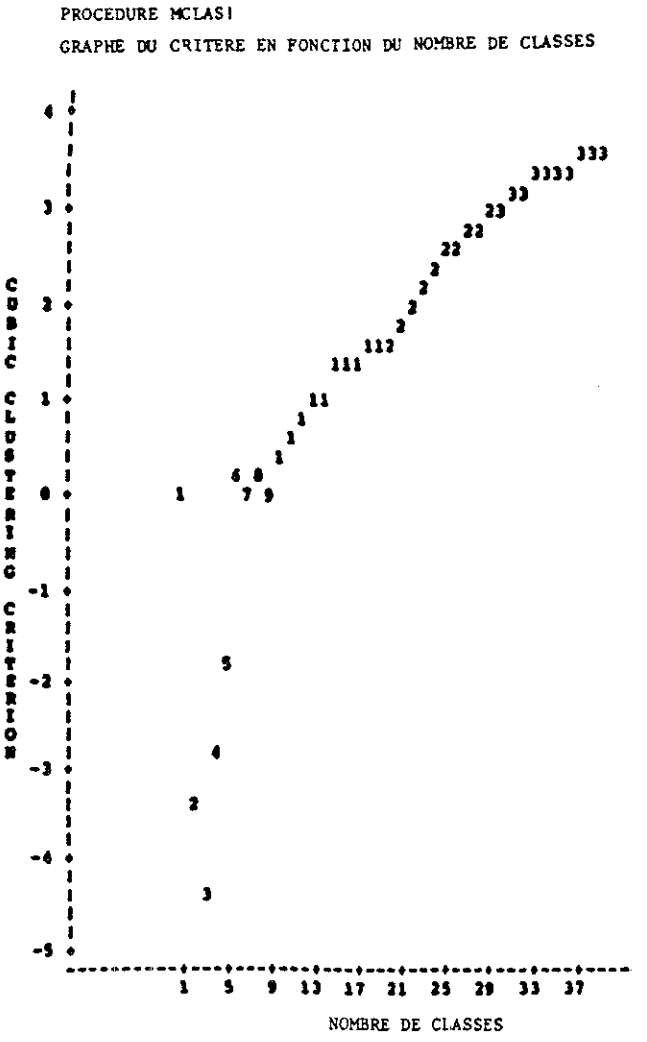


Figure 5

Dans ce cas on doit répéter des essais pour essayer de mettre en évidence soit :

- un maximum local stable du CCC tel qu'on le voit apparaître dans 4 essais différents sur la figure 6.
- un accroissement très significatif et stable de la valeur du CCC pour un certain nombre de classes, tel qu'il apparaît pour 5 classes dans les 4 essais de la figure 7. Un tel accroissement significatif peut souvent donner lieu à un palier sur la courbe d'évolution du CCC. On note ici, compte tenu du caractère plus subjectif de l'indication, l'importance de la validation par une observation répétée sur des essais différents.

Si on ne peut pas mettre en évidence une telle indication, on passera alors à l'utilisation de la procédure MCLAS décrite dans le paragraphe suivant. Cette utilisation peut rester souhaitable même si une indication a été trouvée par la procédure MCLAS1, pour confirmation et en raison de la meilleure qualité des partitions fournies par la procédure MCLAS.

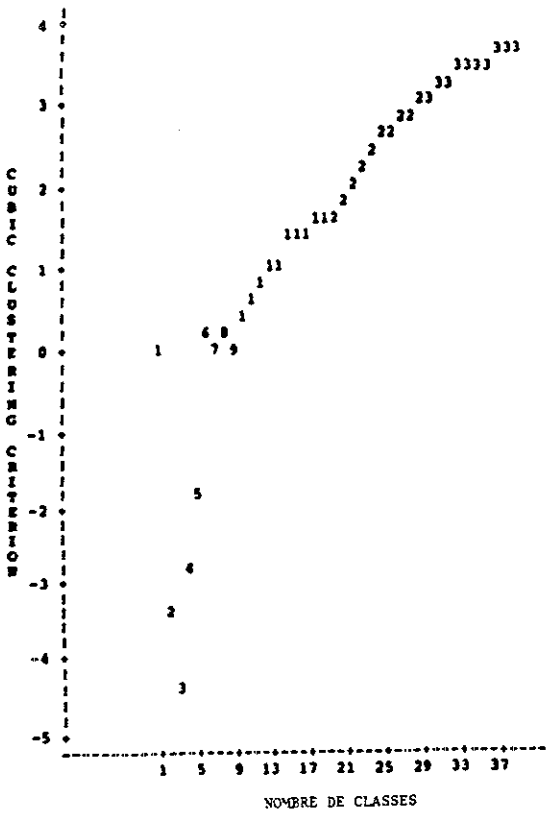


Figure 6a

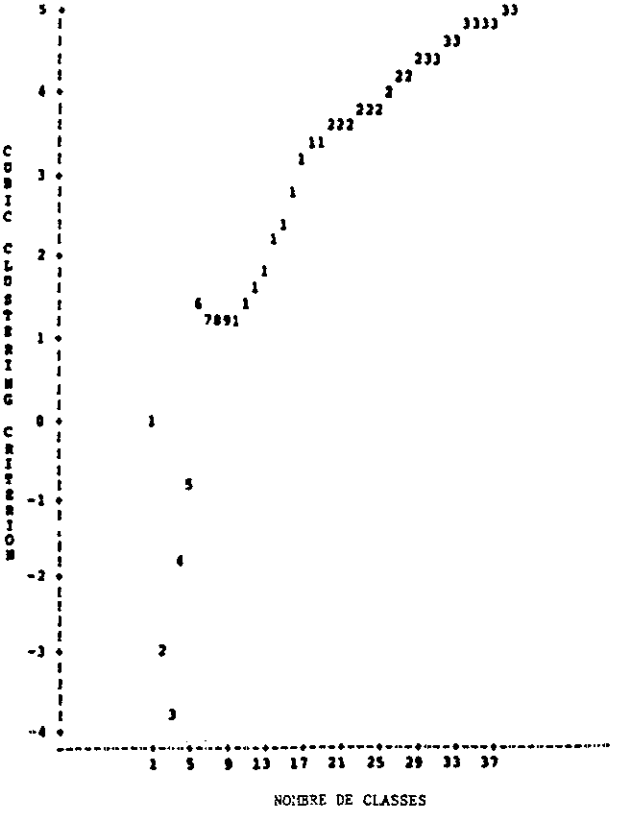


Figure 6b

GRAPHE DU CRITERE EN FONCTION DU NOMBRE DE CLASSES

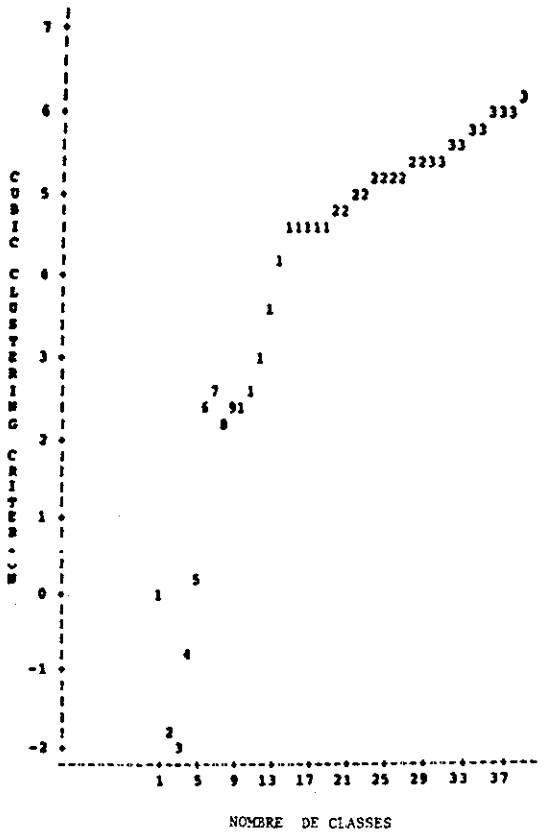


Figure 6c

GRAPHE DU CRITERE EN FONCTION DU NOMBRE DE CLASSES

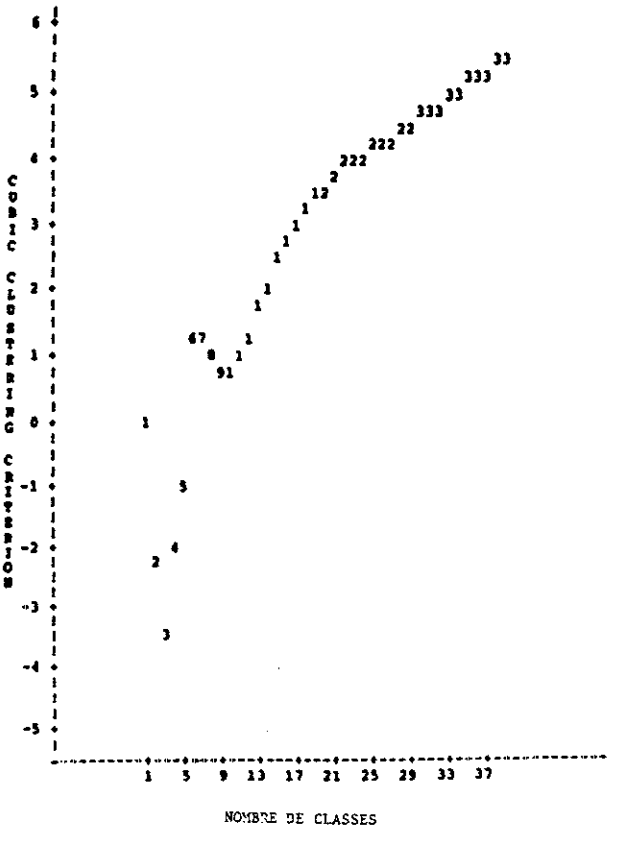
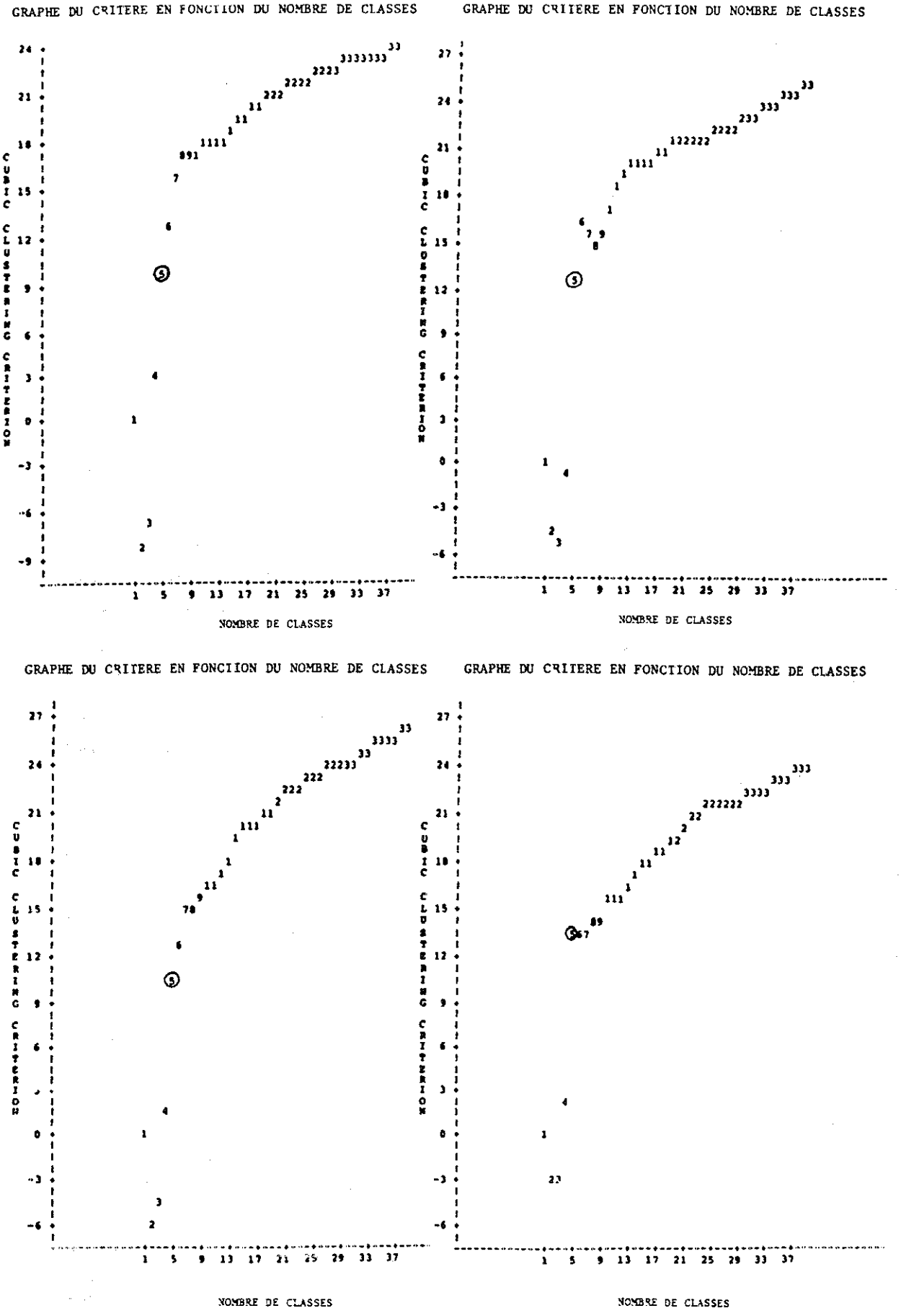


Figure 6d

EXECUTIONS PROCEDURE MCLAS1



EXECUTIONS PROCEDURE MCLAS1

Figure 7

3- Utilisation de la procédure MCLAS (population classifiable)

a) Généralités

La procédure MCLAS dépend de deux paramètres :

- le nombre d'expériences (correspondant à différentes initialisations de l'algorithme de classification non hiérarchique) dont la valeur par défaut est égale à 5.
- le nombre de classes fixé pour ces expériences.

De ces deux paramètres dépend le nombre de formes-fortes, intersection des différentes partitions, qui seront soumises à la classification hiérarchique.

L'intérêt des formes-fortes par rapport aux classes d'une partition a priori en un nombre élevé de classes (procédure MCLAS1) est qu'elles sont de taille très inégale ; en raison de la stratégie d'agrégation de l'algorithme hiérarchique utilisé (méthode de WARD) les plus grosses d'entre elles correspondant à des regroupements très cohérents ne s'agrègeront qu'en haut de l'arbre, ce qui se traduira nécessairement par une chute de la valeur du critère de classification. On est donc assuré par cette méthode de mettre en évidence un maximum du critère ou au moins une indication significative dans la zone du haut de l'arbre où nous recherchons le nombre de classes optimal. Moins les formes-fortes seront nombreuses, plus elles seront, pour certaines d'entre elles, d'effectif important, d'où un maximum plus accentué de la courbe du CCC.

Cependant, il est à craindre si le nombre de classes et le nombre d'expériences ne sont pas assez élevés que l'indication trouvée soit dépendante du nombre de classes choisi comme paramètre. En effet, les formes-fortes importantes correspondront alors aux classes des expériences et on risque de retrouver ce nombre de classes (ou plus souvent ce nombre majoré de 1) comme niveau significatif de l'arbre.

Il est donc souhaitable d'obtenir un nombre de formes-fortes assez élevé et nous préférons pour ce faire répéter un nombre plus grand nombre d'expériences (en majorant le standard égal à 5 ou en enchaînant plusieurs exécutions de la procédure MCLAS, ce qui est possible et permet defaire croître le nombre d'expériences de 5 en 5) plutôt que d'augmenter trop le nombe de classes initial : on risque en effet de fragmenter, de manière artificielle, a population en un nombre élevé de formes-fortes.
Savoir où s'arrêter dans l'accroissement du nombre de formes-fortes par la répétition des expériences n'est pas une question simple ; un critère reste peut être à définir mais dès à présent il est possible de suivre l'évolution des effectifs des formes-fortes au fil des expériences et de noter quand une certaine stabilité est atteinte, au moins concernant les formes-fortes les plus importantes.

b) Conseils pratiques

Nous recommandons de faire différentes exécutions de MCLAS pour différentes valeurs du nombre de classes, par exemple entre 6 et 10 classes. Dans chaque cas, 5 expériences (valeur standard) peuvent suffire si la stabilité des formes- fortes importantes paraît établie (population bien classifiable). On comparera alors les courbes d'évolution du CCC, en portant attention pour les différents niveaux significatifs aux plus fortes valeurs obtenues pour ce critère. Un maximum absolu, entre toutes les courbes, observé pour un certain nombre de classes peut être un indice très significatif.

Une illustration de cette méthode est donnée par les graphiques de la figure 8 (10 expériences pour 6, 7, 8, 9 classes) La conclusion du choix d'une partition en 7 classes provient à la fois des figures 8-c et 8-d (saut significatif à 7 classes pour deux partititons différentes des formes-fortes) mais aussi de la figure 8-a (le maximum à 8 est sans doute dû au nombre initial de classes égal à 7, mais le saut à 7 est plus significatif et correspond surtout à la plus forte valeur observée du CCC, de l'ordre de 20) et dans une moindre mesure de la figure 8-b (le maximum observé à 7 dépendant pour une part du nombre initial de 6 classes, ce qui sera d'autant moins vrai cependant que le nombre d'expériences sera élevé)
Une partition en 5 classes paraît également assez significative au vu des 4 graphiques précédents.

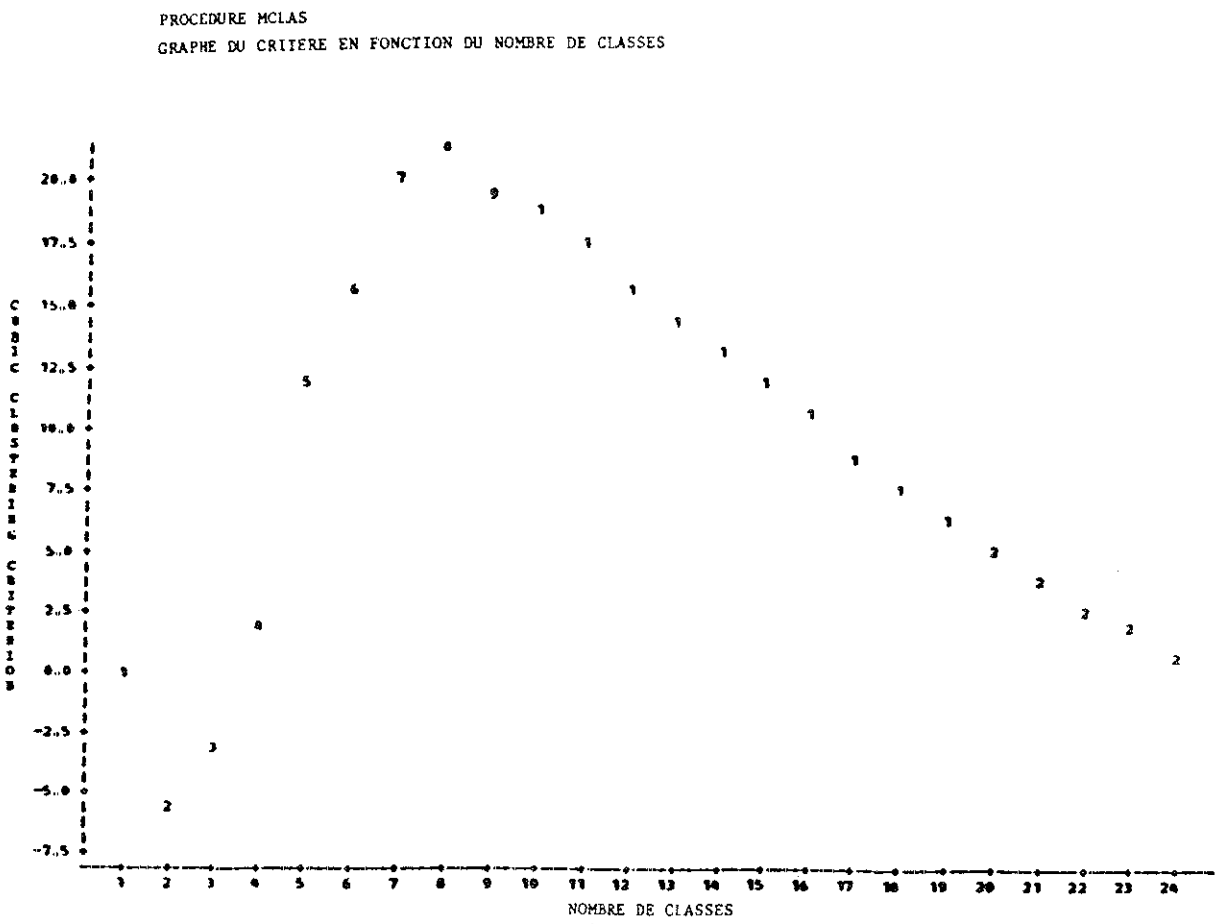


Figure 8a

EXPERIENCES AVEC 7 CLASSES

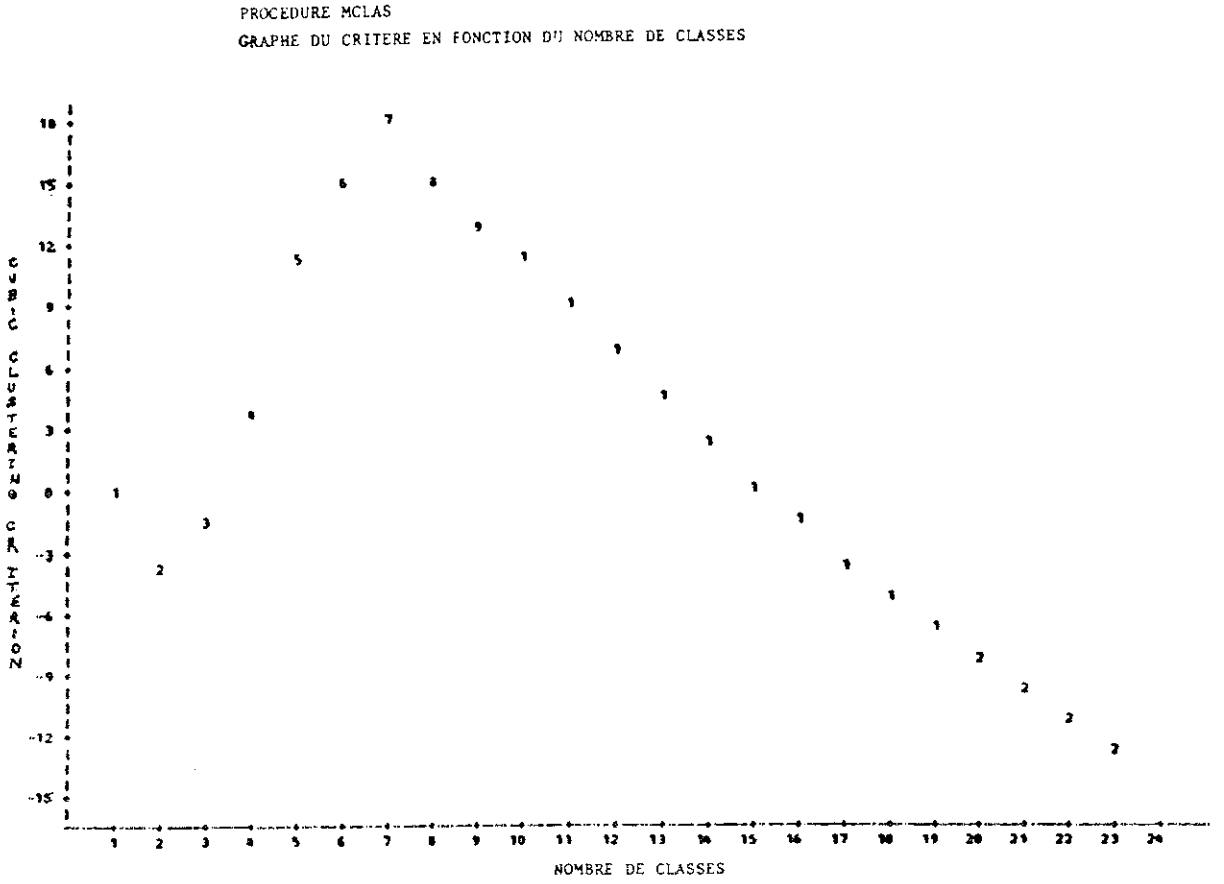


Figure 8b

EXPERIENCES AVEC 6 CLASSES

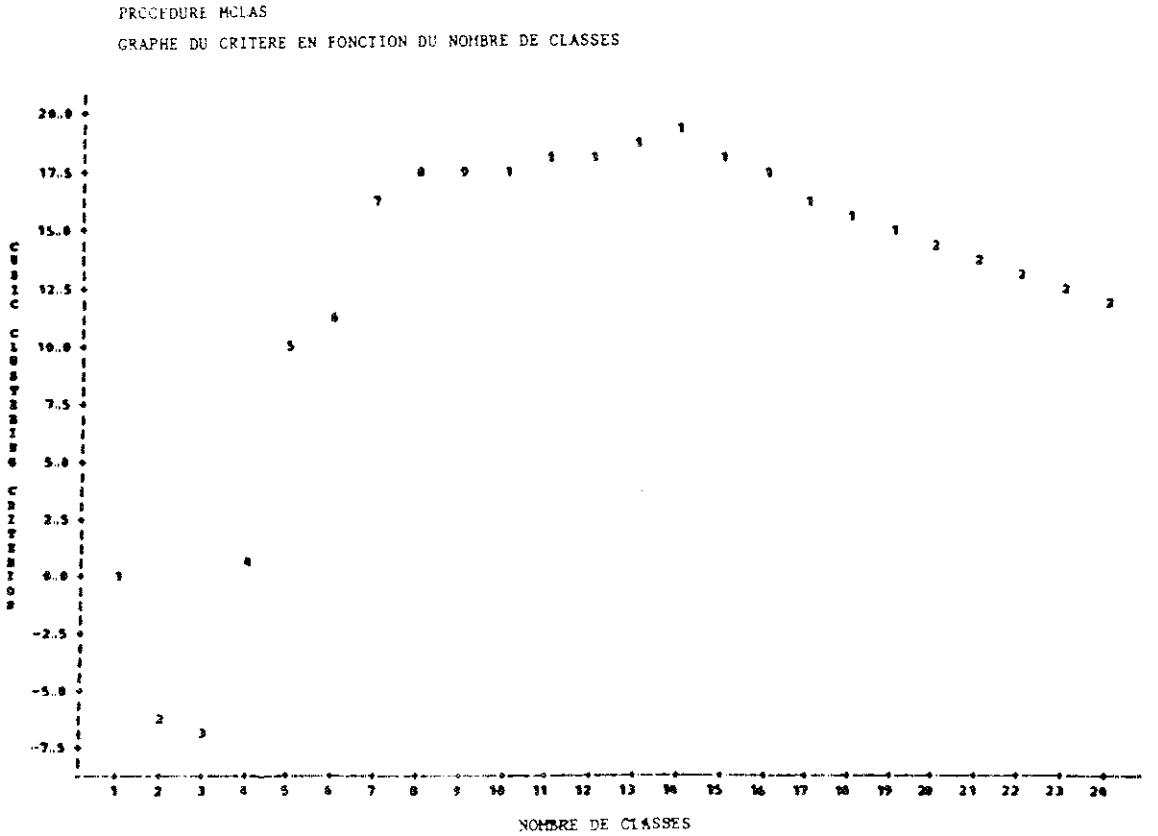


Figure 8c

EXPERIENCES AVEC 8 CLASSES

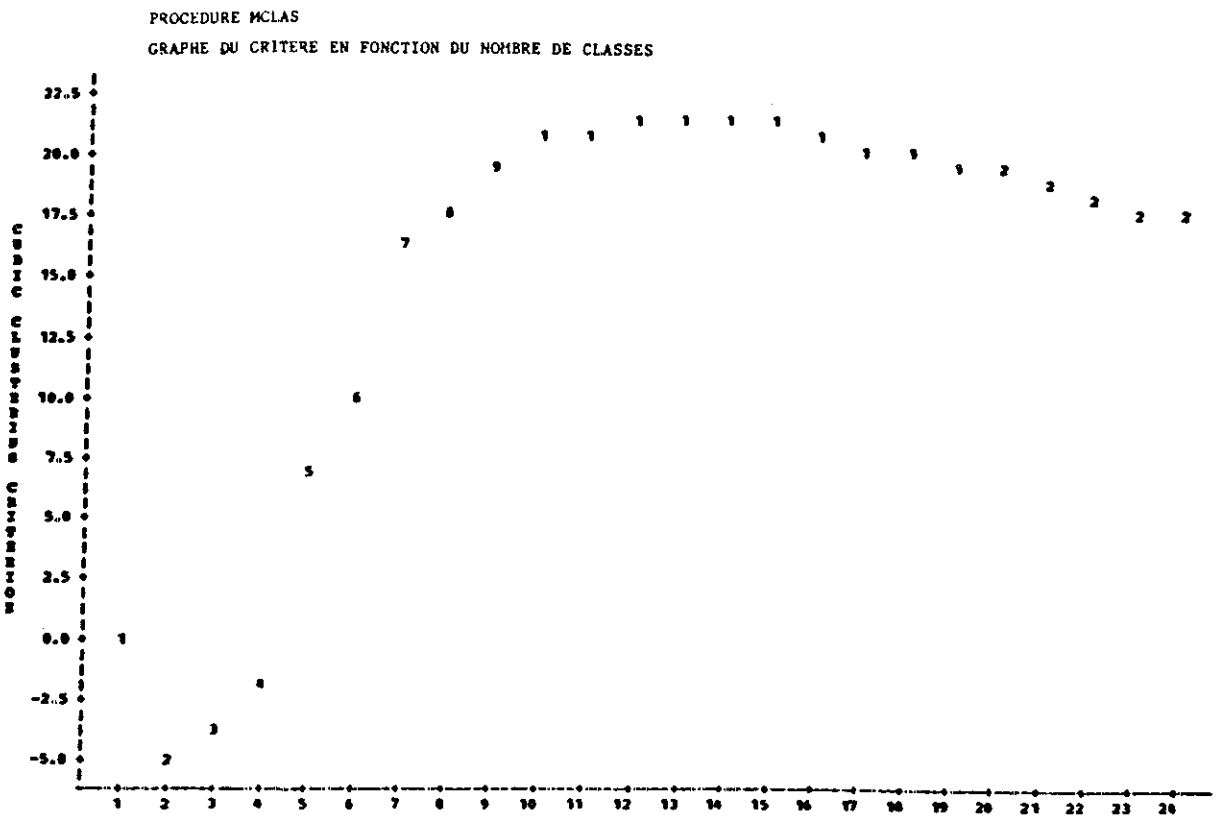


Figure 8d

EXPERIENCES AVEC 9 CLASSES

4. Utilisation de MCLAS pour des populations "non classifiables"

a) Méthodologie

Une population non classifiable se reconnait aux valeurs très négatives du CCC sur la courbe d'évolution de ce critère en fonction du nombre de classes après utilisation de la procédure MCLAS1. L'utilisation de la procédure MCLAS fournira des valeurs moins négatives du critère, mais la courbe demeurera nettement en dessous de zéro. Un exemple d'une telle courbe obtenue par MCLAS sur ce type de population est donnée sur la figure 9.
En face de cette situation, nous proposons de définir une sous-population plus classifiable. Celle-ci sera obtenue en éliminant de la population initiale des objets ou groupes d'objets ayant montré une grande instabilité pour se rattacher à une classe caractéristique de la population initiale. La mise en évidence de ces classes caractéristiques (formes-fortes) comme celle des éléments instables se fera en augmentant le nombre d'expériences dans la procédure MCLAS.

Pratiquement on enchaînera différentes exécutions de cette procédure, et on analysera l'évolution des formes-fortes obtenues, après chaque nouvelle série de 5 expériences.

On s'arrêtera quand les formes-fortes les plus importantes apparaîtront stables. Le nombre total de celles-ci sera évidemment très important, un grand nombre d'entre elles ne contenant qu'un seul élément. Sur ce type de population,il peut être nécessaire d'aller jusqu'à 20 expériences pour voir apparaître des formes- fortes stables.

Pour obtenir, à partir de la population initiale des sous-populations de plus en plus structurées, on éliminera les formes-fortes d'un seul élément, puis celles de deux éléments, etc..., jusqu'à obtenir pour les sous-populations sélectionnées des valeurs positives du critère CCC, lors de l'application de la procédure MCLAS1 par exemple. On arrivera ainsi sur une sous-population à mettre en évidence un nombre significatif de classes.

L'utilisation finale de l'algorithme non hiérarchique pour optimiser la partition, servira également ici pour allouer les individus restants (éliminés dans la première phase) aux classes caractéristiques et fournir la partition finale de l'ensemble de la population.

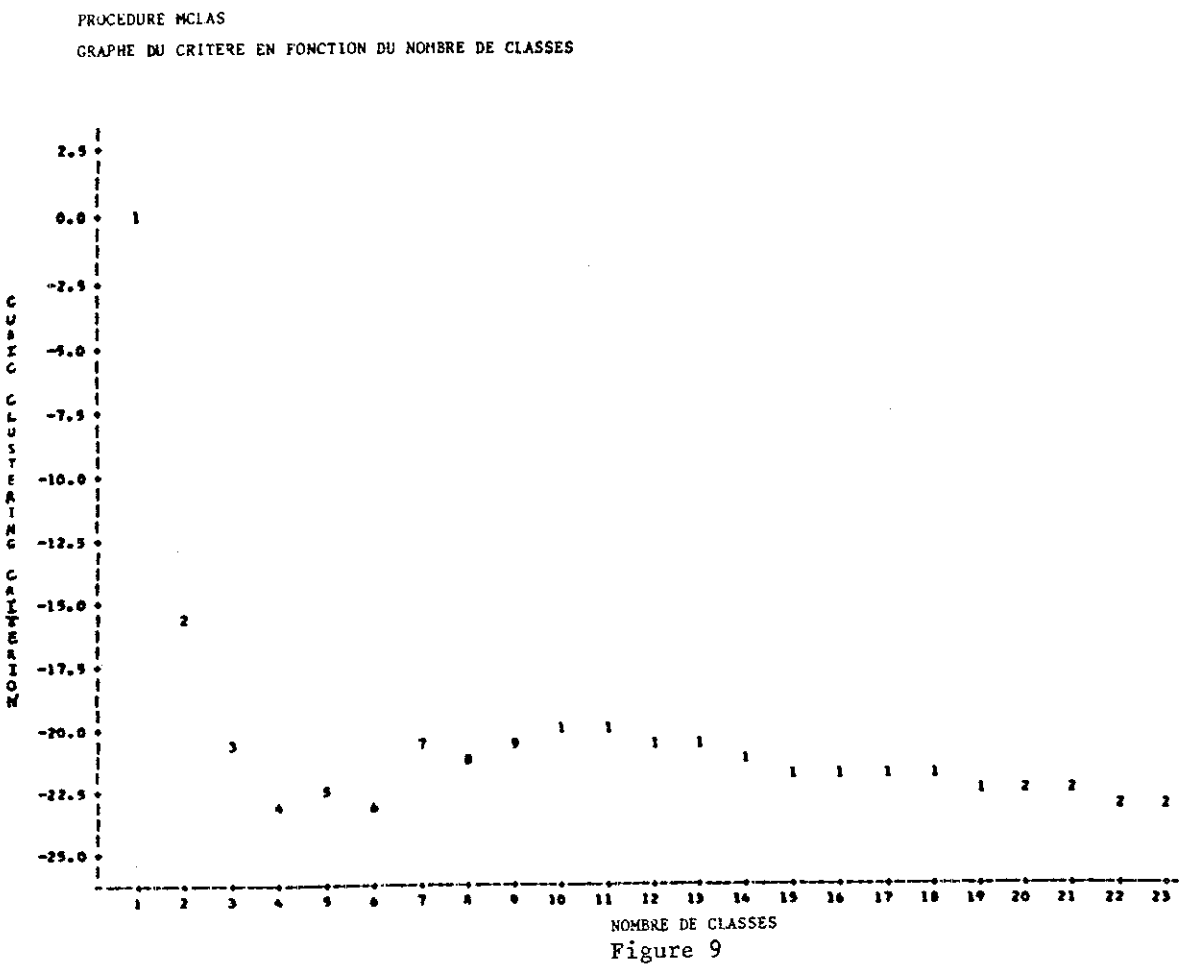


Figure 9

Cette méthode a été mise en oeuvre sur la population dont le caractère très peu classifiable a été mis en évidence sur la figure 9. Soumettant cette population d'environ mille individus à 20 expériences (4 exécutions enchaînées de la procédure MCLAS) de classification non hiérarchique en 8 classes, nous avons obtenu 571 formes-fortes, 467 d'entre elles étant formées d'un seul individu.

Les quatre graphiques de la figure 5 représentent le résultat de l'application de MCLAS1 à trois sous-populations différentes, la première obtenue en supprimant les formes-fortes de un et deux individus (deux essais différents), population réduite à 435 individus, la seconde excluant les formes-fortes de trois individus et moins (387 individus), la dernière celles de quatre individus et moins (352 individus).

Cette dernière sous-population résultant de l'élimination la plus sévère des individus atypiques donne des valeurs plus élevées du critère (figure 6-c). Les quatre graphiques concluent nettement à une partition en 6 classes.

Enfin, un essai par la procédure MCLAS (figure 10) a été réalisé sur une population un peu plus nombreuse, n'excluant que les seules formes-fortes d'un individu.

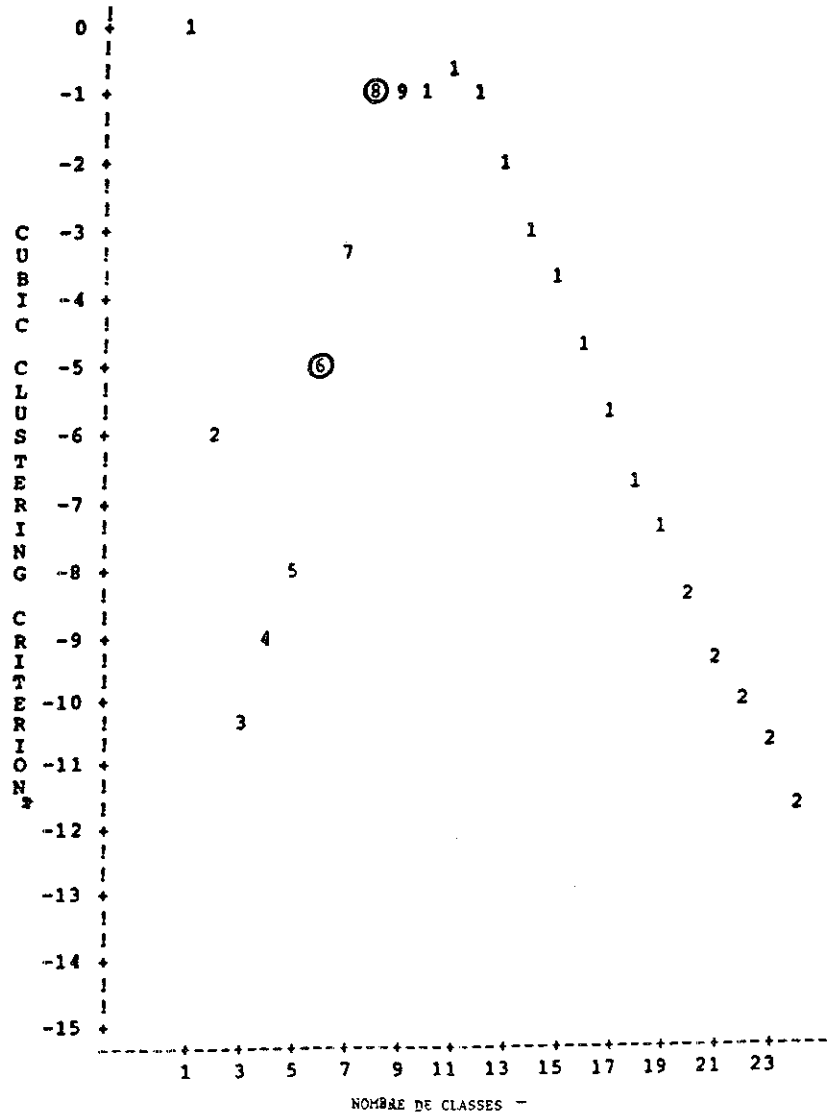


Figure 10 GRAPHE DU CRITERE EN FONCTION DU NOMBRE DE CLASSES

Le CCC redevient négatif, ce qui n'empêche pas cependant (vu la forme de la courbe) de prendre en compte les deux indications fournies : confirmation d'une partition significative en 6 classes, et mise en évidence d'un autre niveau significatif à 8 classes qui sera finalement retenu.

Un compromis est donc à trouver entre des courbes significatives du CCC (dont les valeurs ne soient pas trop négatives) et une sous-population d'effectif suffisant pour ne pas trop perdre en finesse de description

5- Un paramètre supplémentaire : le nombre de facteurs

La condition d'utilisation du CCC étant la non corrélation des variables (voir I) on fera donc précéder la classification d'une analyse en composantes principales ou d'une analyse des correspondances, suivant la nature des variables initiales (quantitatives ou qualitatives). Un problème important est donc celui du choix du nombre de facteurs à retenir pour la classification.

Nous conseillons de retenir un nombre de facteurs légèrement supérieur au nombre de ceux qui sont bien interprétés dans l'analyse factorielle.

La population choisie pour illustrer l'emploi de la procédure MCLAS dans le paragraphe III.3 (figure 8) résultait du choix des cinq premiers facteurs d'une analyse factorielle des correspondances. Le choix s'était arrêté sur 7 classes. Une autre approche pour noter l'évolution du CCC à différents niveaux de partition consiste à réaliser des essais directement au moyen de l'algorithme de classification non hiérarchique.

Nous avons signalé en remarque au paragraphe II.1 les difficultés de cette approche. Néanmoins, on peut tirer une information très significative de la comparaison de deux séries de valeurs du CCC (obtenues par l'algorithme de classification non hiérarchique pour différents nombres de classes) correspondant respectivement au choix des cinq puis des six premiers facteurs de l'analyse des correspondances :

5 facteurs

Nombre de classes	4	5	6	7	8	9	10
CCC	9.89	22.39	23.07	24.22	22.67	23.37	26.02

6 facteurs

Nombre de classes	4	5	6	7	8	9	10
CCC	9.30	20.48	21.06	22.08	28.12	27.42	24.66

Malgré la difficulté à conclure avec cette approche, en particulier pour 5 facteurs, à cause de la variabilité des résultats [6], il semble établi par contre que l'importante variation du CCC entre 7 et 8 classes, pour 6 facteurs, est l'indication d'un niveau significatif à 8 classes.

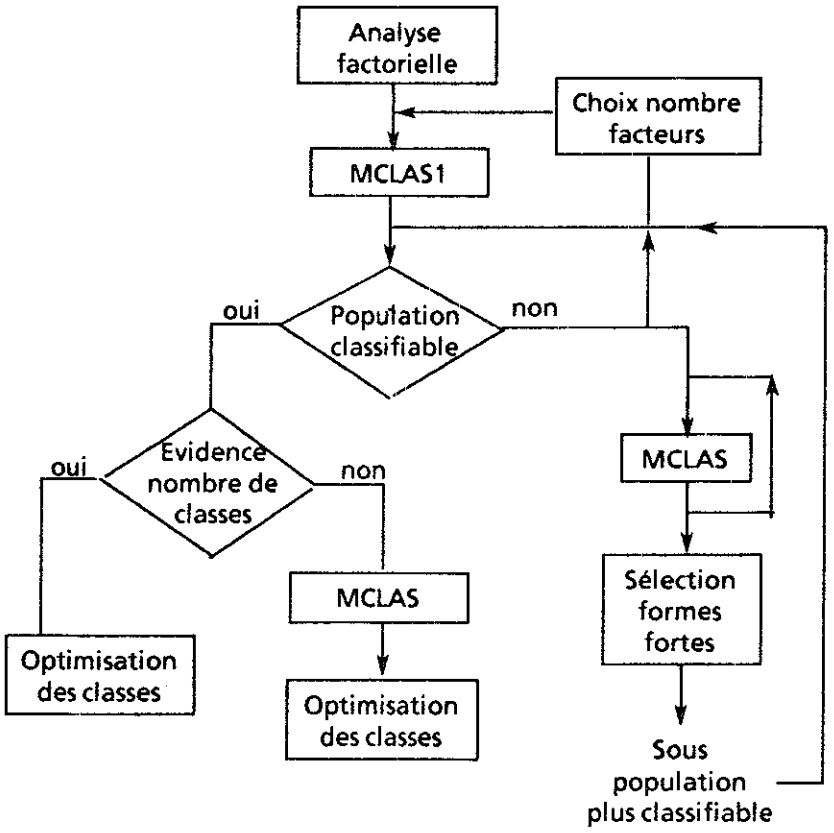
Le simple ajout d'un facteur supplémentaire fait apparaître une structure beaucoup plus nette dans les données.

A contrario, le choix d'un nombre trop élevé de facteurs (ne correspondant pas à une interprétation claire) introduit un bruit de fond qui empêche de voir des indications significatives.

Cet effet a été vérifié clairement, sur d'autres données, en faisant croître le nombre de facteurs de 6 à 15. Pour un nombre trop élevé de facteurs la courbe du CCC montre une évolution très monotone sans aucune irrégularité pouvant être l'indice d'une indication.

Le choix du nombre de facteurs doit se faire dès le début de l'analyse en essayant différentes valeurs au moyen de la procédure MCLAS1, et compte tenu des résultats de l'analyse factorielle.

6. Organigramme d'utilisation des différentes procédures



B - Indices de similarité - critère de La Véga

Nous allons montrer que la stratégie d'utilisation conjointe d'une méthode non hiérarchique (méthode des "nuées dynamiques") et d'une méthode hiérarchique est également applicable en respectant la nature initiale des données si celles-ci sont qualitatives (indices de similarité ou de dissimilarité).

En première étape nous définirons un critère nous permettant, comme dans le cas quantitatif, de comparer entre elles des partitions de manière intrinsèque, c'est-à-dire en faisant abstraction du nombre de classes ou de leurs effectifs.

Ensuite nous montrerons comment l'algorithme des nuées dynamiques, de même qu'une méthode de classification hiérarchique adaptée peuvent contribuer à optimiser ce critère et grâce à son caractère intrinsèque, à déterminer un nombre de classes optimal.

L'intérêt de l'ensemble de la procédure sera également ici d'utiliser la classification hiérarchique sur un nombre pas trop élevé de classes initiales, qui seront en l'occurence les formes fortes obtenues par les nuées dynamiques.

La méthode est limitée à des tailles de population pour lesquelles la matrice des similarités peut être stockée en mémoire ; elle a été appliquée à un jeu de données de mille individus d'une enquête, précédemment classifiés suivant la procédure du chapitre A, à partir d'une représentation métrique des données (par une description des individus sur les premiers facteurs d'une analyse des correspondances). La comparaison des approches métrique et non métrique s'est révélée tout à fait intéressante, pouvant plaider en faveur de cette dernière même au prix d'une implémentation informatique plus lourde (temps calcul et place mémoire).

Cette méthodologie originale a été présentée en [9].

I - Critère de De La Véga

1. Définition initiale [10]

Il s'agit d'un critère d'adéquation entre une partition $\Pi(D)$ d'un ensemble D d'objets à classifier, et une structure de proximité $P(D)$ sur ces objets;

Celle-ci est définie par une préordonnance sur D, c'est-à-dire un préordre total sur l'ensemble de tous les couples d'objets de D, où le rang d'une paire d'objets est une fonction strictement monotone de la ressemblance de ces deux objets

Plus précisément cette ressemblance est mesurée par l'indice de similarité s défini sur toutes les paires d'objets
Pour la définition du critère de De La Véga, on définit :

F ensemble de toutes les paires d'objets de D
ω l'ordre total tel que pour deux paires p et q de F :

$$p < q \Leftrightarrow s(p) > s(q)$$

et le graphe correspondant à ω dans $F \times F$:

$$gr(\omega) = \{(p,q) \in F \times F / p < q \text{ pour } \omega\}$$

Toute partition π peut aussi être représentée dans $F \times F$ par le produit cartésien $R \times S$ où :
R est l'ensemble des paires réunies par la partition
S est l'ensemble des paires séparées par la partition

Le cardinal de l'intersection dans $F \times F$ de ces deux sous-ensembles représentant l'un la structure de proximité, l'autre la partition, sera à la base du critère, soit :

$$\text{card}\,(gr(\omega) \cap (R \times S))$$

Le critère lui même, dans sa définition initiale, est égal au cardinal du complémentaire, dans $F \times F$, de la différence symétrique entre les ensembles $gr(\omega)$ et $R \times S$

Dénotant le cardinal de A par |A| et la différence symétrique entre les ensembles A et B par $A \Delta B$, le critère de De La Véga s'écrit :

$$|F \times F - gr\,(\omega) \Delta (R \times S)| \qquad (1)$$

Le problème de classification se résume donc, avec cette définition, à rechercher parmi toutes les partitions celle qui maximise le critère d'adéquation exprimé en (1).

2. Résultats de LERMAN [11]

LERMAN a montré que maximiser l'expression (1) équivaut à maximiser :

$$|gr(\omega) \cap (R \times S)| - \tfrac{1}{2} |(R \times S)| = |gr(\omega) \cap (R \times S)| - \tfrac{1}{2} |R| \times |S| \qquad (2)$$

A partir de cette expression, LERMAN a prouvé le résultat suivant :
sur la distribution de toutes les partitions d'un type donné (c'est-à-dire en K classes d'effectifs n_1, n_2, ... n_k), la moyenne statistique de l'expression (2) du critère de De la Vega est nulle, sa variance est égale à :

$$|R| \times |S| \times (|R| + |S| + 1) / 12,$$

et sa distributrion asymptotique est Gaussienne.

L'importance de ce résultat est de permettre par la normalisation du critère, de lui donner un caractère intrinsèque, en éliminant l'influence de la taille de la population, du nombre de classes, et de leurs effectifs respectifs, sur sa valeur absolue.

3. Linéarisation du critère de De la Vega

Said CHAH [12] a donné une nouvelle formulation, très simple, du problème de maximisation de l'expression (2).
On définit les fonctions Y(p) et O(p,q), p et q étant des paires d'objets, donc éléments de F :

$$p \in F \qquad Y(p) = \begin{cases} 1 \text{ si } p \in R \\ 0 \text{ si } p \in S \end{cases}$$

$$p,q \in F \qquad O(p,q) = \begin{cases} 1 \text{ si } s(p) > s(q) \\ 0 \text{ sinon} \end{cases}$$

Alors le critère de De la Vega peut s'écrire :

$$\sum_{p \in F} Y(p) \sum_{q \in F} \Big(O(p,q) - O(q,p) \Big)$$

ou encore, en revenant aux objets i, j de l'ensemble D à classifier :

$$\sum_{i \neq j} Y(i,j) \times W(i,j) \qquad i,j \in D \qquad (3)$$

W(i,j) étant égal à la différence entre le nombre de paires d'objets de D ayant une similarité inférieure à s(i,j) et le nombre de paires d'objets ayant une similarité supérieure à s(i,j)

Cette définition reste vraie même s'il y a des paires ex-aequo dans l'ordre sur les valeurs des similarités s(i,j).

En l'absence d'ex-aequo, la quantité W(i,j) à la propriété remarquable suivante :

$$\sum_{i \neq j} W(i,j) = 0$$

Le formalisme des comparaisons par paires [13] permet donc de donner au problème de la recherche de la partition maximisant le critère de De la Vega une formulation très simple, équivalente à la "règle de la majorité", du critère de Condorcet [14] :

on décidera de réunir dans une même classe deux objets i et j (Y(i,j) = 1), si le nombre de couples d'objets moins ressemblants que i et j l'emporte sur le nombre de couples d'objets plus ressemblants que i et j.

4. Interprétation statistique du critère

Si l'on définit : $Z(p,q) = Y(p) - Y(q) \quad p,q \in F$
et se rappelant la définition de O (p,q) (voir paragraphe 3), alors S CHAH a montré [12] que maximiser le critère de De la Vega parmi toutes les partitions de l'ensemble D équivaut à rechercher Z, O étant donné, qui maximise :

$$Cov(Z, O)$$

La covariance est donc calculée sur l'ensemble F × F de tous les couples (p,q) de paires d'objets.

Se rappelant par ailleurs la valeur, calculée par LERMAN, de la variance du critère de De la Vega, on peut ramener la maximisation du critère de De la Vega normalisé, parmi toutes les partitions, à celle du coefficient de corrélation entre Z et O :

$$\text{Max Corr}(Z, O) \Leftrightarrow \text{Max} \frac{|gr(\omega) \cap (R \times S)| - \frac{1}{2}|R| \times |S|}{\sqrt{|R| \times |S| \times (|R| + |S| + 1)}}$$

Nous retiendrons finalement ce critère normalisé, dû à LERMAN, que nous appelerons pour simplifier "critère de LERMAN", comme critère de choix d'une partition.

II - L'utilisation de la méthode des Nuées Dynamiques

Deux points seront soulignés : l'adéquation de la méthode, grâce à sa notion de noyau constitué de plusieurs éléments, à l'optimisation du critère de De la Véga; et l'intérêt de la notion de formes fortes, pour une classification hiérarchique ultérieure, en référence à l'optimisation de ce même critère.

1. Critère optimisé par les nuées dynamiques [7]

Utilisé dans la premier phase de notre stratégie globale de classification, cet algorithme non hiérarchique est du type "centres-mobiles" ou "allocation-réallocation". Il procède alternativement par la définition de classes, puis de représentants pour chacune de ces classes. Ces représentants pour une classe peuvent être soit un ensemble, appelé noyau, d'un nombre fixé d'objets les plus représentatifs de la classe, soit dans le cas métrique le centre de gravité de la classe, les nuées dynamiques se ramenant dans ce cas à une méthode de "K-Means".

La construction des classes et des noyaux nécessite seulement une fonction R(A,B) définissant la proximité entre deux ensembles d'objets A et B à partir de l'indice de proximité (ou de similarité) s entre les objets, par la formule suivante :

$$R(A,B) = \sum_{a \in A} R(a,B) = \sum_{a \in A} \sum_{b \in B} s(a,b)$$

A l'itération de numéro j de la procédure globale, nous avons un ensemble N^J de noyaux, soit (N_1^j ... N_k^j) pour les k classes et un ensemble P^J de classes, soit (P_1^j, ... P_k^j).

La convergence de la procédure alternative de définition des classes et des noyaux, à partir d'une initialisation aléatoire (des classes ou des noyaux), pour un nombre de classes fixé, peut être prouvée sous un certain nombre de conditions générales.

A la convergence, la solution finale (N,P) est associée à un maximum local (dans le cas d'un indice de similarité) du critère suivant :

$$R(N,P) = \sum_{i=1}^{k} R(N_i, P_i)$$

2. Application au critère de De la Vega

Dans le cas particulier d'un indice de dissimilarité correspondant à une vraie distance dans un espace métrique, et à condition de représenter chaque classe par son centre de gravité, le critère optimisé par les nuées dynamiques est la somme pondérée des variances intra-classes, c'est-à-dire le classique critère R^2 (voir chapitre A).

A l'opposé, et c'est là un intérêt de la notion de noyau, si l'on impose que chaque classe soit représentée par un nombre d'objets égal à l'effectif de la classe, alors il est facile de voir qu'à la convergence, le critère optimisé sera de la forme :

$$\sum_{i=1}^{k} \sum_{(j,l) \in P_i} s(j,l)$$

Or le critère de De La Vega, sous son expression (3) peut encore s'écrire :

$$\sum_{j \neq l} Y(j,l) \times W(j,l) = \sum_{i=1}^{k} \sum_{(j,l) \in P_i} W(j,l)$$

puisque Y(j,l) vaut 1 ou 0 suivant que les objets j et l appartiennent ou non à la même classe.

L'algorithme des nuées dynamiques peut donc optimiser le critère de De la Vega, si l'on choisit comme indice de similarité entre objets la quantité W définie en I.3.

Cette utilisation de la méthode des nuées dynamiques avec des noyaux de taille variable n'est pas classique. Bien que vérifiée en pratique la convergence reste à prouver théoriquement.

En réalité l'optimisation du critère de De la Vega, à nombre de classes fixé, comporte l'inconvénient de favoriser des classes d'effectif inégal. Nous montrons à l'Annexe 2 que pour un nombre de classes fixé, la variance du critère de De La Vega croît lorsqu'on augmente (jusqu'à un certain point) le déséquilibre des effectifs des classes.

Nous vérifierons également ce point lorsque nous chercherons au chapitre suivant III un algorithme hiérarchique optimisant le critère de De La Vega.

Nous serons amenés, pour ne pas obtenir des classes d'effectifs trop déséquilibrés, à chercher un équivalent du critère de la variance, dans le cas non métrique. Cela équivaut ici, pour l'algorithme des nuées dynamiques, à préférer des noyaux formés seulement de quelques objets représentatifs. On se rapproche ainsi de la notion de centre conduisant dans le cas métrique à des classes équilibrées [5]

Dans la pratique, utilisée de cette façon, la méthode des nuées dynamiques a permis d'améliorer substantiellement les valeurs du critère de De la Vega, sous la contrainte de classes d'effectifs non déséquilibrés.
C'est là une manière de prendre en compte la normalisation nécessaire de ce critère.

3. Pourquoi s'intéresser aux formes fortes ?

Celles-ci résultent de l'intersection de plusieurs partitions en un même nombre K de classes, partitions obtenues à partir d'initialisations différentes de la méthode des nuées dynamiques et correspondant à différents optimums locaux du critère optimisé par cette méthode.

Dans notre stratégie exposée dans le chapitre A, combinant classification non hiérarchique puis classification ascendante hiérarchique la première phase a pour but, quand la population est de taille importante, d'obtenir une partition en un nombre assez élevé de classes, mais bien inférieur au nombre initial d'objets à classifier, qui seront agrégées dans la seconde phase par l'algorithme hiérarchique.

D'autre part, l'ensemble de la procédure vise à l'optimisation d'un critère de classification, critère de la variance dans le cas métrique, critère de De la Vega (normalisé) dans le cas non métrique.

La raison de préférer les formes fortes à une partition différente en un même nombre de classes, qui pourrait être obtenue directement par un algorithme non hiérarchique (comme dans la procédure MCLAS1 du chapitre A), est que la partition des formes fortes fournira de meilleurs valeurs du critère quand on recherchera un regroupement de ces classes initiales par la classification hiérarchique (ou d'ailleurs par une autre méthode) pour trouver la partition optimale maximisant globalement le critère

Ce résultat a été vérifié dans la pratique (par comparaison des procédures MCLAS1 et MCLAS décrites dans le chapitre A) et résulte du théorème suivant montré par E. DIDAY [8].

Théorème :

Si les formes fortes ne varient plus, quand on augmente le nombre de partitions différentes en K classes dont elles résultent, à partir d'un nombre n de partitions strictement supérieur à q, n étant strictement inférieur au nombre maximum N d'optimums locaux du critère, alors, la partition des formes fortes résultant des q premières partitions est plus fine que la partition correspondant à l'otimum global en K classes

Cela revient à dire que la partition globalement optimale en K classes pourra être obtenue comme résultat d'une agrégation des formes fortes.

Dans le cas des données non métriques, examiné dans ce paragraphe, notre stratégie développée sur un exemple, s'est appuyée uniquement sur les formes fortes pour réaliser la classification hiérarchique de la seconde phase.

III - Phase de classification ascendante hiérarchique

Nous allons chercher dans un premier temps une méthode d'agrégation permettant lors des regroupement ascendants d'optimiser localement le critère de De la Vega, dont nous rappelons l'expression ci-dessous :

$$\sum_{i \neq j} Y(i,j) \times W(i,j)$$

$Y(i,j) = 1$ si i et j sont réunis dans la même classe,

$Y(i,j) = 0$ sinon

l'indice $W(i,j)$ défini à partir des similarités $s(i,j)$ (voir I.3), a la propriété remarquable d'être de somme nulle sur tous les couples i,j (à condition qu'il n'y ait pas d'ex-aequo parmi les similarités) :

$$\sum_{i \neq j} W(i,j) = 0$$

L'algorithme de classificatoin hiérarchique est défini par la fonction d'agrégation R(A,B) de deux classes A et B à un niveau donné.
Souhaitant optimiser le critère de De La Vega nous avons essayé les fonctions suivantes :

1) $R(A,B) = \sum_{a \in A} \sum_{b \in B} W(a,b)$

Il s'agit d'agréger les deux classes, qui après fusion, maximisent le critère.
En haut de l'arbre on obtient quelques très grosses classes et de nombreuses classes d'effectif très faible.

2) $R(A,B) = \frac{1}{n_A n_B} \sum_{a \in A} \sum_{b \in B} W(a,b)$ ("average linkage method")

Les classes en haut de l'arbre sont un peu plus équilibrées mais encore de tailles très différentes. Les valeurs du critère de De la Vega, dans les derniers niveaux, sont comparables à celles obtenues avec la fonction d'agrégation précédente

3) Optimisation locale d'un critère différent du critère de De la Vega.
Les inconvénients observés, du point de vue de l'équilibre des classes, avec les fonctions précédentes résultent de la tendance naturelle du critère de De la Vega, déjà signalée plus haut (II 2), à favoriser des classes d'effectifs inégaux.

Nous pouvons le comprendre d'une autre manière, en examinant la forme que prendrait le critère sur des dissimilarités correspondant à de vraies distances. Maximiser le critère de De la Vega équivaudrait alors, en terme de variance, à minimiser la quantité :

$$\sum_{i=1}^{k} n_i^2 (Var_i - Var)$$

Var_i : variance de P_i
V_{ar} : variance totale
n_i : effectif de P_i

ce qui est encore équivalent à maximiser :

$$\sum_{i=1}^{k} n_i^2 (Var - Var_i)$$

On constate que le coefficient n_i^2 favorisera des classes d'effectif important, même si leurs variances sont plus grandes que celles de classes plus petites.
Par contre le critère de Ward, connu pour donner des classes plus équilibrées, maximise la quantité :

$$\sum_{i=1}^{k} n_i (Var - Var_i)$$

Partant de la forme du critère de Ward et exprimant les distances en terme de dissimilarités, nous en déduisons la fonction d'agrégation correspondante dans le cas non métrique :

$$R(A,B) = 2 \sum_{a \in A} \sum_{b \in B} W(a,b) - n_A n_B \left(\frac{\sum_{i,j \in A} W(i,j)}{n_A^2} + \frac{\sum_{i,j \in B} W(i,j)}{n_B^2} \right)$$

L'utilisation de cette fonction ne donnera pas d'aussi bonnes valeurs du critère de De la Vega que les deux fonctions définies en 1) et 2) mais satisfaira la contrainte de classes d'effectifs équilibrées. C'est cette fonction que nous retiendrons.

En réalité cette dernière fonction n'optimise plus le critère de De La Vega, ayant la forme suivante :

$$\sum_{i=1}^{k} n_i^2 \frac{\sum_{(j,l) \in P_i} W(j,l)}{n_i^2} \quad \text{mais le critère} \quad : \quad \sum_{i=1}^{k} n_i \frac{\sum_{(j,l) \in P_i} W(j,l)}{n_i^2} \qquad (4)$$

Ce critère, prenant en compte une contrainte de normalisation sur les effectifs des classes, se rapproche d'une certaine façon du critère de De la Vega normalisé.

Or c'est ce dernier (que nous avons appelé "critère de LERMAN") qui demeure notre critère de choix du nombre de classes.

$$\text{La quantité} \sum_{(j,l) \in P_i} W(j,l)/n_i^2$$

sera appelée <u>cohésion interne de la classe P_i</u>, et l'expression (4) constituera ce que nous nommerons la <u>cohésion globale de la partition</u>.

Cette quantité, qui peut être négative car

$$\sum_{i \neq j} W(i,j) = 0,$$

sera un indicateur très commode pour analyser la qualité d'une partition.

IV. <u>Application à des données réelles</u>

1) Résumé sur le critère optimisé

Si nous résumons à ce niveau les résultats précédents, le critère de De la Vega peut être optimisé par la méthode des nuées dynamiques grâce à sa forme suivante :

$$\sum_{j \neq l} Y(j,l)\, W(j,l) = \sum_{i=1}^{k} \sum_{(j,l) \in P_i} W(j,l)$$

W(j,l) étant un indice construit à partir des similarités initiales s(j,l).

Le besoin de normaliser ce critère conduira, aussi bien pour les nuées dynamiques que pour la phase de classification hiérarchique, à se rapprocher des conditions d'utilisation optimales de ces méthodes pour le cas métrique :

- noyaux réduits à quelques objets représentatifs (pour retrouver la notion de centre)
- méthode ascendante hiérarchique équivalente à la méthode de WARD.

Il s'agit là seulement de moyens d'optimiser le critère de De la Vega sous la contrainte de classes équilibrées, ou d'une autre façon de prendre en compte sa normalisation, le critère normalisé appelé "critère de LERMAN" ne pouvant être maximisé directement, mais restant le critère final utilisé.

Dans sa nature le critère de De la Vega reste fondamentalement différent du critère de la variance comme les résultats ci-après le montreront, un intérêt de son caractère non métrique étant ses propriétés intrinsèques (grâce à sa facilité de normalisation) pour comparer entre elles des partitions différentes.

2) Mise en oeuvre pratique de la stratégie de classification.

Le jeu de données utilisé correspondait à une population de mille individus décrits par des réponses à 33 questions qualitatives (à modalités exclusives) d'une enquête d'opinion, population déjà analysée par une méthode de classification sur variables quantitatives (utilisant les cinq premiers facteurs d'une analyse factorielle des correspondances préalable), et ayant montré son caractère difficilement classifiable.
C'est sur cette population que fut appliquée avec succès la méthodologie décrite au chapitre A.II 4, qui consuisit à une classification en 8 classes (figure 10).
Nous avons ici, pour une approche non métrique avec le critère de De La Vega suivi les étapes suivantes :

a) Calcul des similarités entre individus

Nous avons choisi pour mesurer la similarité entre deux individus la quantité très simple égale au nombre de modalités de toutes les variables considérées, possédées en commun par les deux individus.
Les rangs de ces similarités devront ensuite être calculés par un algorithme de tri sur toutes les similarités pour calculer l'indice W, nécessaire au critère de la Vega, pour toute paire (i,j) d'individus.

b) Obtention des formes-fortes par les nuées-dynamiques

15 expériences des nuées dynamiques donnant des partitions en 8 classes ont été réalisées, chacune optimisant donc localement un critère proche du critère de de La Vega normalisé. Résultant de l'intersection des 10 meilleures expériences (du point de vue du critère), environ cinq cent formes fortes ont été obtenues (sur une polulation initiale d'environ mille individus) : ce nombre élevé est un critère du caractère peu classifiable de la population.

c) Sélection d'une sous-population plus classifiable

Suivant la méthode présentée au chapitre A.II.4, nous éliminons les formes fortes formées d'un seul individu ou montrant une faible cohésion (voir plus haut la définition de la cohésion interne). Nous conservons ainsi 97 formes fortes composées de 438 individus sur les 1012 initiaux.

d) Classification hiérarchique

Les deux premières fonctions d'agrégation essayées (formules données en 1) et 2) du paragraphe A.III furent abandonnées au profit de la fonction correspondant à la méthode de Ward dans le cas non métrique, qui conduisit à des classes bien meilleures (du point de vue de l'équilibre de leurs effectifs) en haut de l'arbre.

La courbe d'évolution de l'accroissement du critère de LERMAN aux différents niveaux de la hiérarchie montre une partition significative en 11 classe (figure 11).

e) Phase d'allocation des individus restants et d'optimisation

La méthode des nuées dynamiques est utilisée à la fois pour allouer, en une itération, les individus restants autour de noyaux correspondant aux 11 classes trouvées précédemment, puis pour optimiser cette partition en 11 classes de la population totale.

Une analyse des correspondances montre très peu de changement du point de vue de la signification des classes, entre celles obtenues sur la sous-population classifiable et celles obtenues après réallocation des individus restants, ce qui confirme la pertinence de la sous-population sélectionnée, du point de vue du problème de classification posé.

3) Comparaison des classifications obtenues par les deux approches (métrique et non métrique).

Six classes sont identifiées de la même façon par les deux méthodes. Les différences proviennent à la fois des individus extrêmes et des individus moyens qui paraissent moins bien reconnus par la méthode adoptant le critère de la variance ; en particulier cette méthode ne découvre pas une importante classe centrale, et pourtant de forte cohésion, rassemblant des individus d'opinion moyenne sur la plupart des questions. D'autre part la méthode non métrique distingue plusieurs classes atypiques de très faibles effectifs, qui sont d'un faible intérêt pratique.

Au total, la méthode métrique révèle quelques défauts inhérents au critère de la variance qui tend à refuser de constituer une classe centrale, et d'autre part tend à agréger plus facilement des individus assez différents dès lors qu'ils sont éloignés du profil moyen.

Sur ce jeu de données la méthode non métrique a apporté une information précieuse par la mise en évidence d'une importante classe non typée (18% de la population).

CONCLUSION

Nous avons voulu montrer dans cet article, qu'il est possible de résoudre de manière satisfaisante le problème jugé difficile en théorie du choix du nombre de classes dans une population de taille importante.

En présentant, dans le cas quantitatif, une méthode efficace fondée sur l'utilisation du critère du CCC introduit pour la première fois dans le logiciel SAS, notre intention est de promouvoir l'utilisation de ce critère dont nous donnons l'expression explicite en annexe. La méthode proposée, utilisée couramment dans le cadre de ce logiciel, suppose de disposer de procédures faciles à mettre en oeuvre, pour exécuter les essais successifs requis ou les enchaînements de procédures. Les logiciels d'analyse de données se doivent aujourd'hui d'offrir une telle souplesse d'utilisation.

La seconde partie veut ouvrir une voie concurrente à la trop classique utilisation du critère de la variance, en montrant que dans le cas qualitatif, il est possible aussi d'optimiser, avec le critère de De La Vega, un critère de choix du nombre optimal de classes. Seulement utilisée à titre expérimental, cette seconde méthode aujourd'hui d'implémentation informatique plus lourde, pourrait, avec les progrès des moyens de calcul, faire valoir ses qualités de robustesse et de référence à un critère beaucoup plus intrinsèque.

Dans les deux cas, notre philosophie reste fondée sur l'optimisation d'un critère pour le problème de classification et l'utilisation conjointe à cette fin, pour les populations de grande taille, d'une méthode non hiérarchique et d'une méthode hiérarchique.

Références

[1] C. PERRUCHET
Une Analyse Bibliographique des épreuves de classifiabilité en Analyse des Données
Statistique et Analyse des Données - 1983 - Vol. 8 n° 2 - pp. 18-41
Association des Statisticiens Universitaires

[2] J.L. MOLLIERE
The best mode of use for the CCC
SEUGI'85 - Proceedings of the third annual Conference in COLOGNE
SAS Software Limited

[3] Warren S. SARLE
Cubic Clustering Criterion - SAS Technical Report A - 108
SAS Institute Inc. Box 8000 CARY, NC 27 511 - 8000 - USA

[4] MILLIGAN, G.W., et COOPER, M.C. (1985)
An examination of procedures for determining the number of clusters in a data set.
PSYCHOMETRIKA, 50, 159 - 179

[5] G. CELEUX
Classification et modèles
Revue Statistique Appliquée, 1988, XXXVI (4), 43 - 58

[6] J.L. MOLLIERE - B. GERARDIN
Macro for determining the optimal clusters in large data sets
SEUGI'84 - Proceedings - AMSTERDAM - April 4-6 1984

[7] Edwin DIDAY
Optimisation en Classification Automatique et Reconnaissance des Formes
Revue Française d'Automatique Informatique Recherche Opérationnelle (R.A.I.R.O.)
6ème année - Novembre 1972 - Vol. 3, pp. 61 - 96 - DUNOD - Paris

[8] Edwin DIDAY
Optimization in non hierarchical clustering
Pattern Recognition 1974 - Vol. 6, pp. 17 - 33 - PERGAMON Press.

[9] Jean Luc MOLLIERE
What's the real number of clusters
Classification as a tool of research
W. GAUL and M. SCHADER (Editors) - NORTH HOLLAND - 1986

[10] BENZECRI et Collaborateurs
L'Analyse des Données - 1 - la Taxinomie - DUNOD - 1973

[11] I.C. LERMAN
Classification et analyse ordinale des données - DUNOD - 1981

[12] S. CHAH
Calcul des partitions optimales d'un critère d'adéquation à une préordonnance.
Data Analysis and Informatics, III - North Holland 1984.

[13] F. MARCOTORCHINO - Utilisation des comparaisons par paires en statistique des Contingences - Etude IBM n° F 069 - 1984

[14] P. MICHAUD Agrégation à la majorité : Hommage à Condorcet - Etude IBM n° F 051 - 1982

Annexe 1

Formule de calcul du CCC (Cubic Clustering Criterion)

a) Calcul de R^2 pour une population uniformément distribuée dans un hypercube de dimension p.

Soit s_j la longueur du côté de l'hypercube dans la j-ème dimension Nous supposerons que les s_j sont classés par ordre de valeur décroissante.

Le volume de l'hypercube est :

$$v = \prod_{j=1}^{p} s_j$$

Pour q classes choisies dans cette population de distribution uniforme, on peut supposer celles-ci contenues dans q hyper-cubes inclus dans l'hyper-cube initial, dont la longueur moyenne c du côté sur chaque dimension sera telle que :

$$c = (\frac{v}{q})^{1/p}$$

Le rapport $U_j = S_j/c$ inférieur à l'unité représente le nombre d'hypercubes contenus le long de la j-ème dimension de l'hypercube initial.

La variance totale de la population dans la direction j est proportionnelle à s_j^2 tandis que la variance intra-classe est proportionnelle à C^2.

$$\text{D'où : } R^2 = 1 - \frac{\sum_{j=1}^{p} C^2}{\sum_{j=1}^{p} s_j^2} = 1 - \frac{p}{\sum_{j=1}^{p} u_j^2}$$

b) Prise en compte d'une dimension inter-classes inférieure à p.

Cette dimension inter-classes, soit p*, sera obtenue comme le plus grand entier, plus petit que q (nombre de classes), tel que $u_j = S_j/C$ soit plus petit que 1.

On supposera alors que les classes sont des hypercubes de côté égal à C sur les p* premières dimensions et de côté égal à s_j sur les dimensions restantes (rang supérieur à p*).

On en déduit alors :

$$R^2 = 1 - \frac{p^* + \sum_{j=p^*+1}^{p} u_j^2}{\sum_{j=1}^{p} u_j^2}$$

c) Correction pour les échantillons de taille n faible.

Basée sur des simulations, cette correction conduit à la formule suivante, pour la valeur moyenne

$$E(R^2) = 1 - \left[\frac{\sum_{j=1}^{p^*} \frac{1}{n+u_j} + \sum_{j=p^*+1}^{p} \frac{u_j^2}{n+u_j}}{\sum_{j=1}^{p} u_j^2} \right] \left[\frac{(n-q^2)}{n} \right] \left[1 + \frac{4}{n} \right]$$

d) Formule finale du CCC.

$$CCC = \ln \left[\frac{1 - E(R^2)}{1 - R^2} \right] \frac{\sqrt{\frac{np^*}{2}}}{(0\ 001 + E(R^2))^{1,2}}$$

Cette formule répond au besoin de stabiliser la variance du CCC pour différentes valeurs du nombre d'observations, de variables et de classes. Elle a été obtenue empiriquement.

Annexe 2

Etude de la variance du critère de De La Vega en fonction des effectifs des classes

LERMAN a trouvé pour cette variance l'expression suivante :

$$|R| \times |S| \times (|R| \times |S| + 1)/12 \qquad (1)$$

$|R|$: nombre de paires d'objets réunis par la partition

$|S|$: nombre de paires d'objets séparés par la partition.

Rappelons que la distribution considérée pour le calcul de la variance est la distribution de toutes les partitions d'un type donné, c'est-à-dire des partitions en K classes d'effectifs fixés n_i , sur laquelle la moyenne du critère est nulle.

Dans l'expression (1) la quantité entre parenthèse étant constante (égale au nombre total de couples d'objets plus un), il reste à étudier la variation avec les effectifs des classes de la quantité : $|R| \times |S|$

On appelle n le P.G.C.D. des effectifs des différentes classes ; on a donc :

$$n_i = J_i\, n \qquad J_i \text{ entier}$$

Si nécessaire, on multiplie tous les effectifs par un facteur entier tel que J soit le plus petit entier tel que :

$$\sum_{i=1}^{k} J_i = KJ$$

On notera $|R|_e$ et $|S|_e$ les valeurs de $|R|$ et $|S|$ correspondant à une partition en K classes d'effectifs égaux (à J n) de la population d'effectif total N.

Pour la partition considérée, on a alors :

$|R| = |S|_e + \Delta$ et $|S| = |S|_e - \Delta$

$$\Delta = \frac{\sum_{i=1}^{K} J_i^2 - KJ^2}{2} \times n^2 \qquad (2)$$

$$\text{et} \qquad \sum_{i=1}^{k} J_i = KJ$$

Appelons E la quantité suivante :

$E = |R| \times |S| - |R|_e \times |S|_e$

$E = \Delta \times [|S|_e - |R|_e - \Delta]$

Variation de E en fonction des effectifs des classes :

a) Δ est toujours positif ou nul et d'autant plus grand que les effectifs sont déséquilibrés.
Le signe de E dépend donc du signe de la quantité : $|S|_e - |R|_e - \Delta$ qui s'exprime :

$$|S|e - |R|e - \Delta = \frac{K(K-1)J^2 - \sum_{i=1}^{K} J_i^2}{2} \times n^2 + \frac{KJ}{2} n \qquad (3)$$

b) Etude du signe de (3)

La quantité (3) peut s'écrire :

$$\frac{K^2J^2n^2}{2} [1 - \frac{1}{K} - \sum_{I=1}^{K} (\frac{J_i}{KJ})^2] + \frac{KJ}{2} n$$

En posant

$$\frac{J_i}{KJ} = X_i$$

l'expression (3) devient (en se rappelant que : $KJn = N$) :

$$\frac{N^2}{2} [1 + \frac{1}{N} - \frac{1}{K} - \sum_{i=1}^{K} X_i^2] \text{ avec } \sum_{i=1}^{K} X_i = 1$$

Sans le terme $\frac{1}{K}$ l'expression ci-dessus entre crochets serait toujours positive ou nulle.

Hors mis le cas de deux classes où l'expression (3) sera négative (dès qu'on s'éloignera un tant soit peu du cas de classes équilibrées, si N est grand), cette expression ne deviendra négative que dans un domaine de valeurs de X_i proches de 0 ou de 1, et d'autant plus restreint que le nombre de classes K est élevé.

On mesurera l'écart à la situation de classe équilibrées ($x_1 = x_2 = \ldots = x_k = \frac{1}{k}$) par l'angle Θ du vecteur $(x_1, x_2, \ldots, x_k)$, dont l'extrêmité est dans le plan d'équation

$$\sum_{i=1}^{K} X_i = 1,$$

avec le vecteur

$$(X_1 = X_2 = \ldots = X_K = \frac{1}{K})$$

normal à ce plan.

La distance (au carré) de l'origine à ce plan étant égale à $\frac{1}{K}$, on a :

$$\sum_{i=1}^{K} X_i^2 = \frac{1}{\cos^2\Theta}\,\frac{1}{K} \qquad \text{avec} \sum_{i=1}^{K} X_i = 1$$

d'où la nouvelle forme de l'expression (3) égale à :

$$\frac{N^2}{2}\left[1 + \frac{1}{N} - \frac{1}{K}\,\frac{1}{\cos^2\Theta}\right]$$

c) L'écart E est le produit de Δ, positif et croissant avec Θ, et de l'expression (3) initialement positive et ne changeant de signe qu'à partir d'une certaine valeur Θ_1 de Θ.

Calculant la dérivée de E par rapport à Θ, on trouve :

$$\frac{dE}{d\Theta} = \frac{N^4}{2}\,\sin\Theta\left[\left(1 + \frac{1}{N}\right)K\cos^2\Theta - 2\right]$$

Il existe donc un domaine de valeurs de Xi (d'autant plus grand que K est élevé) où la valeur de la variance croît quand on s'écarte du cas de classes équilibrées.

Précisément, cette croissance est assurée jusqu'à une valeur Θo de Θ telle que :

$$\cos^2\Theta_o = \frac{2}{K}\,\frac{1}{1 + \frac{1}{N}}$$

Ensuite la variance décroît et redevient inférieure à la valeur du cas équilibré pour $\Theta 1$ tel que :

$$\cos^2\Theta_1 = \frac{1}{K-1}$$

d) Exemples :

Pour K = 3, la variance croît jusqu'à Θo tel que :

$$\cos^2\Theta o \simeq \sqrt{\frac{2}{3}},$$

ce qui correspond au profil :

$$(0, \frac{1}{2}, \frac{1}{2}), \text{ou } (\frac{1}{6}, \frac{1}{6}, \frac{2}{3})$$

Pour K = 6, la variance croît jusqu'à Θo tel que :

$$\cos^2\Theta o \simeq \sqrt{\frac{1}{3}},$$

ce qui correspond au profil :

$$(0,0,0,0,\frac{1}{2},\frac{1}{2}),\ \text{ou}\ (0,0,0,\frac{1}{6},\frac{1}{6},\frac{2}{3}),\ \text{ou}\ (\frac{1}{48},\frac{1}{48},\frac{1}{16},\frac{1}{12},\frac{1}{8},\frac{11}{16})$$