

J.-P. BRIANE

J.-J. LAZARE

R. SALANON

## **L'analyse en données groupées de très grands tableaux de correspondance**

*Les cahiers de l'analyse des données*, tome 3, n° 2 (1978),  
p. 167-173

[http://www.numdam.org/item?id=CAD\\_1978\\_\\_3\\_2\\_167\\_0](http://www.numdam.org/item?id=CAD_1978__3_2_167_0)

© Les cahiers de l'analyse des données, Dunod, 1978, tous droits réservés.

L'accès aux archives de la revue « Les cahiers de l'analyse des données » implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme  
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

## L'ANALYSE EN DONNÉES GROUPÉES DE TRÈS GRANDS TABLEAUX DE CORRESPONDANCE [DONNÉES GROUPÉES]

par J.-P. Briane <sup>(1)</sup>

J.-J. Lazare <sup>(2)</sup>

et R. Salanon <sup>(3)</sup>

### 0 Introduction : le traitement de très grands tableaux

Le développement récent des travaux de synthèse phytosociologiques nécessite des analyses multidimensionnelles portant sur des tableaux de données de plus en plus volumineux. Or, les moyens informatiques, pourtant puissants, dont on dispose à l'heure actuelle, ne permettent pas de traiter directement par l'analyse factorielle des correspondances (A. F.C. dans la suite) des tableaux dépassant quelques centaines de relevés floristiques. Pour donner quelques ordres de grandeur : on peut traiter jusqu'à 300 relevés environ sur UNIVAC 1110 et atteindre près de 500 relevés avec l'IBM 370-168 du Centre Inter-Régional de Calcul Electronique d'Orsay, chiffres pouvant paraître relativement confortables, mais vite insuffisants lorsqu'il s'agit, par exemple, de réviser une classe phytosociologique.

Au problème de la place disponible en mémoire centrale s'ajoute celui de la durée des calculs : c'est ainsi que l'analyse des 400 relevés demande 4 minutes environ sur UNIVAC 1110 ; celle de 500 relevés contenant un millier de taxons exige 30 minutes sur IBM 370-168. Notons en outre qu'il se peut qu'au terme de ce calcul, l'on doive recommencer l'analyse (erreurs dans les données, hétérogénéité, etc.).

En tout état de cause, au-delà des limites indiquées ci-dessus, il n'est plus possible de faire débiter le traitement par une A.F.C. traditionnelle ou mieux, "directe". Il apparaît donc hautement souhaitable de disposer de méthodes, certes moins sensibles au sens physique du terme, mais qui allient les avantages de la rapidité et de l'économie tout en autorisant des traitements très volumineux. L'objet du présent article est de décrire une telle méthode ; un deuxième article présente aux non-spécialistes quelques applications dans le domaine de la phytosociologie : nous espérons susciter ainsi dans d'autres domaines des applications nouvelles et des généralisations de la méthode.

### 1 Description de la méthode d'analyse en données groupées (A.D.G.)

1.1 Réduction des données. Jusqu'à présent, tous les procédés employés commençaient par une transformation plus ou moins importante de la matrice des données afin d'en réduire les dimensions. Le plus fréquent consistait à extraire du tableau initial un sous-tableau portant sur un nombre réduit de relevés, ces derniers étant choisis soit systématiquement soit au hasard. La taille de ce sous-tableau permettait alors le traitement direct par l'A.F.C.. Dans la phase terminale, on projetait en éléments supplémentaires les relevés n'ayant pas été pris en compte dans le sous-tableau.

(1) Laboratoire de taxinomie végétale expérimentale et numérique associé au C.N.R.S. n° 121, université de Paris XI, Orsay.

(2) Centre d'écologie montagnarde de Gabas. Parc national des Pyrénées. Université de Bordeaux I. 64440 Laruns.

(3) Laboratoire d'écologie des régions arides, U.E.R. - Domaine méditerranéen, université de Nice, campus Valrose.

Les inconvénients que comporte un tel procédé apparaissent clairement : omission d'espèces éventuellement significatives ; obligation de recommencer parfois l'analyse pour s'assurer de la stabilité des résultats. Notons une variante de cette procédure, qui consistait purement et simplement à éliminer du tableau, soit un ensemble de relevés, soit des espèces peu fréquentes, dont rien ne permettait de dire *a priori* s'il s'agissait d'"accidentelles" ou, au contraire, de "caractéristiques"(ex.: F. ROMANE, 1972, E. et S. PIGNATTI, 1975). De même, l'introduction dans des analyses partielles de données n'entrant pas initialement dans l'analyse globale (ex.: P. PUTOT, 1975) nous semble être encore plus artificielle. En somme, ces méthodes ne sont pas plus satisfaisantes les unes que les autres en ce qu'elles comportent d'arbitraire à des degrés divers.

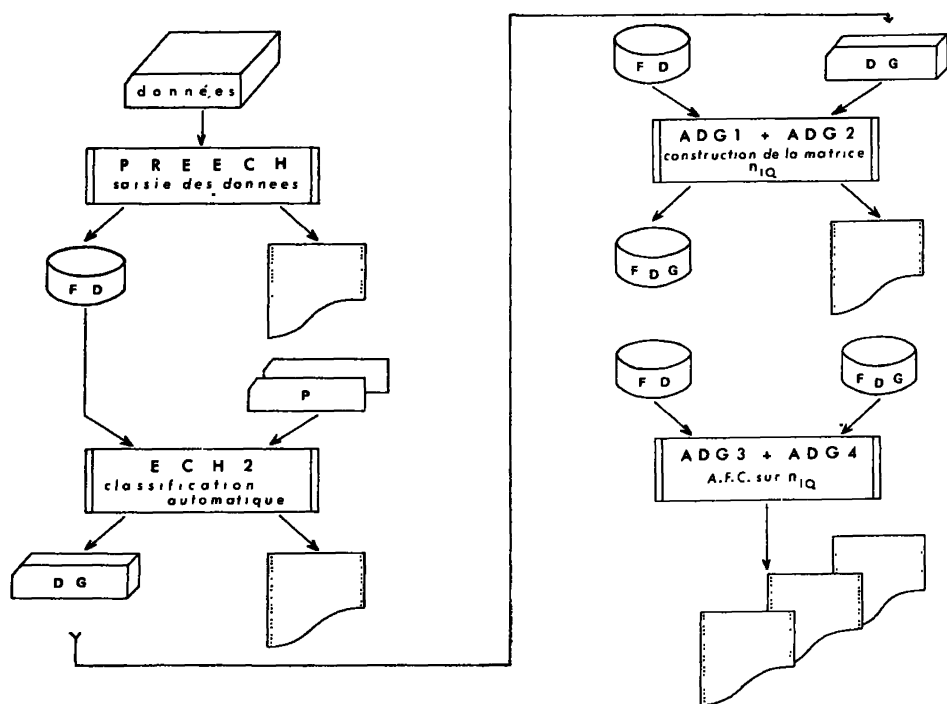


Fig.1.- Organigramme de l'analyse en données groupées (A.D.G.)

### 1.2 Utilisation de la classification automatique : l'analyse en données groupées

(A.D.G.) : Les programmes de classification automatique ont fait au cours de ces dernières années des progrès spectaculaires (ex. : nuées dynamiques). En effet, ces programmes sont désormais rapides et beaucoup moins sensibles qu'autrefois à la taille des données. Le procédé que nous avons élaboré est le suivant :

a - Grâce à un programme de classification rapide (DIDAY ou algorithme d'échange nouvelle version : ECH2, cf Note *in fine*, § 1.4), on découpe l'ensemble des relevés en un nombre fixé de classes (ex. : une dizaine), opération qui, à titre indicatif, requiert dix minutes environ pour le classement de 550 relevés. La partition obtenue peut fournir d'utiles renseignements au phytosociologue, le plus souvent sous la forme d'une confirmation des coupures principales à envisager dans les données. Mais son interprétation de détail reste malaisée, de sorte qu'il est préférable de construire directement le tableau des fréquences des espèces à l'intérieur de chaque classe de relevés : pour chaque espèce  $i$  et chaque groupe  $q$  de relevés,  $n_{iq}$  est le nombre de relevés de  $q$  comportant l'espèce  $i$ .

b - C'est sur ce tableau que l'on effectue une analyse factorielle. La dernière étape consiste à projeter en éléments supplémentaires, sur les axes factoriels du nuage des classes, les relevés eux-mêmes.

### 1.3 Description organique de la chaîne (UNIVAC, FORTRAN V)

Nous donnons ici une description simplifiée de la chaîne des programmes employés. Chaque rectangle (symbolisant un traitement) peut être composé de plusieurs programmes séparés. Nous n'avons noté ici que les fichiers principaux (fig. 1).

#### 1 - PREECH

Chargement des données sur disque (Fichier données : FD) durée : 10 à 30 secondes d'unité centrale.

#### 2 - ECH2

Classification automatique (algorithme d'échange).

- en entrée : ° le fichier FD
  - ° les paramètres (nombre de classes désiré, etc.)
  - en sortie : ° cartes perforées donnant le contenu de chaque classe (données groupées : DG)
  - ° listing des résultats : composition des classes.
- durée : suivant taille ; ex. : pour 550 relevés et 10 classes : env. 10mn.

#### 3 - ADG1 + ADG2

Construction de la matrice synthétique (fréquence de chaque espèce  $i$  dans chaque groupe  $q$ ) :

$$n_{iq} = \sum \{n_{ij} \mid j \in q\}$$

- en entrée : ° le fichier FD et les cartes DG
- en sortie : ° le fichier des données groupées (FDG) : données réorganisées (triées par numéros d'espèces) et augmentées des données de groupage.
- ° le listing de la matrice  $n_{IQ}$ .

duré : de une à deux minutes au maximum.

#### 4 - ADG3 + ADG4

Analyse factorielle de  $n_{IQ}$  :  $(\lambda_\alpha, F_\alpha(q), G_\alpha(i))$  (1) projection des relevés sur les axes factoriels en éléments supplémentaires :

$$F_\alpha^*(j) = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{i \in I} \left\{ \frac{n_{ij}}{n_j} G_\alpha(i) \right\}$$

$j \in J$  : ensemble de relevés ;  $i \in I$  ensemble des espèces

- en entrée : FD et FDG
  - en sortie : les cartes factorielles des groupes, des espèces et des relevés.
- durée : de une à deux minutes.

(1)  $\lambda_\alpha$  = moment d'inertie par rapport au  $\alpha$ ième axe.  $F_\alpha(q)$  = valeur du  $\alpha$ ième facteur pour le centre de gravité  $q$ .  $G_\alpha(i)$  = valeur du  $\alpha$ ième facteur pour l'espèce  $i$ .

1.4 Note : Le programme de classification rapide ECH2 : Fondé sur l'algorithme d'échange, ce programme s'applique à la recherche d'une partition d'un ensemble d'objets en un nombre fixé de classes. Le critère optimisé est le moment centré d'ordre 2 de la partition Q :

$$M^2(Q) = \sum_{q \in Q} \{ p_q \cdot \| p_J^q - p_J \|_{P_J}^2 \}, \text{ avec les notations usuelles.}$$

Le programme est écrit en FORTRAN pour l'ordinateur UNIVAC 1110. En voici les caractéristiques principales :

- Les données traitées sont des tableaux de présence-absence, en (0-1). (Il en existe une version pouvant traiter des tableaux quelconques de nombres positifs).

- La capacité actuelle du programme (pour les données en (0-1) et le matériel considéré) est limité à environ 700 relevés (objets), pour un nombre plus grand d'espèces (attributs). Dans les applications phytosociologiques volumineuses ce nombre est de l'ordre de 1000.

- Pour classer 700 relevés en 5 classes il faut environ 10' à 15' ; on a un certain nombre d'options :

- Pour un nombre fixé de classes on peut demander d'itérer le processus à partir de tirages différents de la partition initiale. Dans ce cas on peut demander en sortie le tableau des groupes d'objets toujours classés ensemble ( $\sim$  "formes fortes" dans la terminologie de E. Diday).

- Au lieu de tirer au hasard la partition initiale on peut l'imposer. On peut aussi fixer en partie le contenu des classes.

## 2 Application de l'analyse en données groupées en phytosociologie - Discussion.

Le procédé de réduction employé ici est comparable à celui qui consiste à remplacer un échantillon d'une variable aléatoire continue par un histogramme de fréquence en classes de valeurs. Il présente l'avantage d'opérer une régularisation des données en atténuant l'hétérogénéité des tableaux (présence de relevés plus ou moins aberrants ou comportant un nombre trop élevé d'espèces accidentelles), et ceci sans procéder à des amputations arbitraires. Nous avons également noté au niveau des cartes factorielles des espèces qu'il se produisait un phénomène analogue à celui provoqué par le dédoublement des relevés (BRIANE et al., in *A-dansonia* Série 2, 14 (1) pp 111-137 ; 1974) : les espèces de fréquences moyennes - parmi lesquelles se trouvent les espèces caractéristiques - se dissocient beaucoup mieux qu'en A.F.C. directe et, de surcroît, les espèces accidentelles ne perturbent plus l'analyse.

Nous appuyant dans ce chapitre sur les résultats d'analyses effectuées dans le cadre des recherches :

- tant analytiques que synthétiques sur les peuplements de *Pinus sylvestris* et de *Pinus uncinata* d'Europe austro-occidentale, ainsi que sur un certain nombre de groupements de l'étage montagnard, tels que groupements de landes, de garides et pelouses, de hêtraies, chênaies pubescentes et même sapinières (R.S.),

- synthétiques sur les pelouses orophiles du *Caricion curvulae* des Alpes centrales et occidentales (J.J. L.), nous nous sommes donc attaché à mettre en évidence, à chaque étape de l'interprétation, les avantages l'A. D.G. que nous proposons, par rapport à l'analyse directe (1).

Dans le premier cas ("exemple pineraies" par la suite), la matrice de 746 relevés x 1027 espèces ne permet évidemment pas une analyse factorielle directe ; le second cas ("exemple *Caricion curvulae*"), dont la matrice de données se limite, dans un premier stade de travail, à 240 relevés x 313 espèces, se prête en raison de sa taille relativement modeste, à des traitements comparatifs par les deux méthodes. Enfin, si dans l'exemple pineraies, l'ensemble des relevés émane du même auteur, ceux de l'exemple *Caricion curvulae* ont par contre été réalisés par différents auteurs.

(1) Signalons qu'à la suite des recherches entreprises depuis quelques années par notre équipe, des remarques méthodologiques de même nature apparaissent dans un essai de synthèse phytosociologique sur les groupements forestiers du bassin parisien (G. DUME, 1975).

### 2.1 Répartition des données en classes.

Cette première étape de l'A.D.G. consiste en une partition des données en plusieurs classes, chaque classe regroupant les relevés de plus forte parenté floristique. Le nombre des classes peut être choisi arbitrairement, ce qui semble n'avoir qu'une incidence réduite sur la configuration des cartes factorielles ultérieures (ex. *Caricion curvulae*, où des cartes identiques ont été obtenues après partition en 7, 10 puis 12 classes). Il est clair toutefois que la recherche du nombre optimum de classes conduit dès ce stade à la mise en évidence des grandes coupures qui, sinon, n'apparaissent qu'après la projection des données sur les différents axes. Ainsi, dans l'exemple pineraie, où les partitions ne sauraient être décelées intuitivement en raison du nombre considérable de données, nous avons utilisé dans ce but la méthode dite d'agrégation autour des centres variables (cf. en particulier E. Diday, *Revue de Statistique Appliquée* 1970 ; J.P. BENZECRI et al., 1973. L'analyse des données, Tome I : 293 sqq, auxquels nous renvoyons, ne pouvant entrer ici dans le détail). Cette méthode nous a permis d'obtenir un découpage optimum de nos relevés en 34 classes à l'issue d'une série d'opérations dont nous donnons un résumé ci-dessous.

### 2.2 Volume des données

L'A.D.G. ne perturbe pas les positions relatives des grands ensembles de relevés par rapport à ce qu'elles sont en analyse directe, dans la mesure toutefois où cette dernière les fait apparaître : c'est notamment le cas dans l'exemple pineraies, où l'A.D.G. concerne la totalité des données, alors que l'analyse directe doit se limiter au quart des relevés en raison de la capacité de l'ordinateur.

Il est en outre possible d'avoir une représentation très synthétique des rapports entre les ensembles de relevés grâce à la construction d'un bloc diagramme combinant les plans (1-2) et (1-3) à l'issue d'une analyse effectuée sur le nombre optimum de classes.

### 2.3 Hétérogénéité des données

La plupart du temps, comme les données comportent quelques relevés floristiquement pauvres ou renfermant des espèces accidentelles, l'analyse directe provoque leur très grand éloignement le long des premiers axes factoriels (1), ce qui entraîne corrélativement une agglutination des centres relevés autour de l'origine des axes. Ultérieurement, au cours des analyses partielles successives, la même probabilité de voir s'éloigner à nouveau un ou deux relevés le long d'un axe demeure élevée, entraînant par là le risque d'une grande perte de temps. Il est à déplorer que cet écueil ait été trop souvent contourné en épurant au préalable les données de façon intuitive, ce qui nous semble évidemment en contradiction avec le recours ultérieur à une technique dont l'objectivité n'est plus à démontrer ! Ceci a pour conséquence de déformer d'une certaine manière la réalité en atténuant la visualisation de la dynamique entre les groupements et en donnant ainsi une image abusivement schématisée de la configuration du tapis végétal.

Pour pallier cet inconvénient majeur, nous avons suggéré précédemment (J.P. BRIANE et al., in *Adansonia* série 2 ; 14(1) ; pp 111-137 ; 1974) de traiter ensemble les relevés et leurs "anti-relevés" correspondants (procédé de duplication des relevés au moyen du programme DEDOU ; programme expliqué dans deux articles de F. Nakhlé publié dans le Vol I de ces *Cahiers* : pp 243 sqq et 367 sqq). Cette technique rend les cartes d'analyses plus équilibrées en faisant immédiatement apparaître les grandes directions du nuage de points, aussi bien dans le traitement de données floristiquement affines que dans celui de groupements variés pris simultanément en considération (J.J. L., non publié). De plus, son intérêt s'est trouvé accru dans l'interprétation des cartes d'espèces par le fait que les accidentelles migrent de la périphérie vers le centre du nuage au profit d'espèces qui, mises au contraire en évidence par un déplacement

(1) Par contre, dans une recherche volontaire de ces anomalies, il est évidemment préférable d'utiliser l'A.F.C. directe (cf. fig. 2).

vers les marges, s'avèrent être, soit des compagnes de haute présence, soit les véritables espèces caractéristiques de la classe, d'ordre, d'alliance ou d'association suivant les types de groupements analysés. Ce phénomène facilite évidemment la recherche de la signification écologique des axes factoriels.

Cependant, le programme DEDOU ne permettant pas d'analyser de plus grandes matrices que l'A.F.C. directe, il s'est révélé nécessaire de trouver un moyen qui, tout en gardant les mêmes avantages, permette d'augmenter considérablement les dimensions de la matrice des données : ces qualités sont réunies par l'A.D.G., dont nous commentons l'utilisation et les résultats dans un autre article, destiné aux statisticiens non spécialistes d'écologie et de botanique.

L'analyse en données groupées permet la même démarche que l'analyse directe, c'est-à-dire une succession d'analyses partielles. Mais l'expérience a montré qu'il fallait logiquement s'arrêter à un certain stade de réduction de la matrice, en l'occurrence lorsque cette dernière se résume à un effectif d'environ trois ou quatre classes. A partir de ce stade, et malgré les réserves émises à leur sujet, le recours aux analyses directes nous paraît, d'une part largement suffisant, mais d'autre part tout à fait nécessaire si l'on veut que se dégagent finalement des ensembles offrant chacun un maximum d'homogénéité. Nous étions d'ailleurs parvenu à une conclusion similaire à propos de l'utilisation du programme DEDOU. Il est en effet aussi logique - grâce à l'emploi de l'A.D.G. - de ne pas éliminer arbitrairement, en début d'analyse, des relevés supposés aberrants ou "intermédiaires" que d'utiliser *in fine* l'A.F.C. classique qui, seule, opère en toute objectivité un rejet des éléments non intégrables aux groupements discriminés.

#### 2.4 Apport de l'A.D.G. dans les travaux de phytosociologie synthétique

Une synthèse envisagée à partir des tableaux phytosociologiques de différents auteurs sur divers territoires est donc réalisable, dans la mesure toutefois où leurs relevés furent élaborés dans le même esprit, puis publiés *in extenso*. C'est dès le premier axe qu'une discrimination d'ordre chorologique apparaît clairement entre les données, ce qui permet la mise en évidence simultanée des vicariances taxonomiques et phytosociologiques correspondantes d'un territoire à l'autre. [Pour comprendre le mot de vicariance, il faut revenir à la définition même de l'espèce végétale. Etroitement dépendantes de leur habitat, dépourvues de la mobilité propre aux animaux, les plantes présentent dans les divers lieux une infinie variété de formes. Parfois, il s'agit seulement de variations non héréditaires, manifestant une plasticité de l'espèce compatible avec la constance de son capital génétique ; ce qu'on reconnaît depuis A. Jordan en transplantant des individus dans des conditions homogènes, où apparaît leur identité foncière. Mais lors même qu'il existe entre deux formes séparées dans l'espace géographique, une véritable différence génotypique, il importe au botaniste - et singulièrement à l'écologiste - de garder présent à l'esprit qu'il s'agit au fond de deux acteurs très semblables et qui jouent presque le même rôle : on parle alors d'espèces *vicariantes* ; le même terme est employé pour des groupements constitués d'espèces soit identiques soit vicariantes, qui se correspondent à plus ou moins grande distance. Sur cette notion de vicariance on se reportera à M. Guinochet (Phytosociologie ; Masson 1973 ; pp 73, 138) ; et aussi M. Bidault (Variation et spéciation chez les végétaux supérieurs ; Doin 1971 ; partout où est cité le nom de A. Jordan!).]

#### 3 Conclusion : le schéma méthodologique

Nous nous sommes efforcé de montrer l'intérêt de l'analyse en données groupées sur le plan scientifique, avant tout par le fait qu'elle permet la réalisation d'analyses objectives d'un très grand nombre de données. Il est désormais possible de traiter plusieurs milliers de relevés, volume qui s'impose, rappelons-le, dans la plupart des travaux de synthèse phytosociologique opérant à grande échelle. Ajoutons que les performances sont encore accrues, par rapport à l'analyse directe, en raison de la réduction considérable, à la fois de la durée des calculs, et du nombre de traitements partiels à effectuer.

Et puisque, à l'heure du choix de tout outil scientifique, se pose aussi la question de sa rentabilité sur le plan économique, nous ne pouvons que vivement souhaiter l'extension de l'utilisation de l'A.D.G., non seulement en phytosociologie, mais aussi dans de nombreux domaines où des problèmes similaires sont à résoudre.

Il est évident que le schéma méthodologique décrit ci-après (fig 2) est transposable sans difficultés à de multiples secteurs de la recherche.

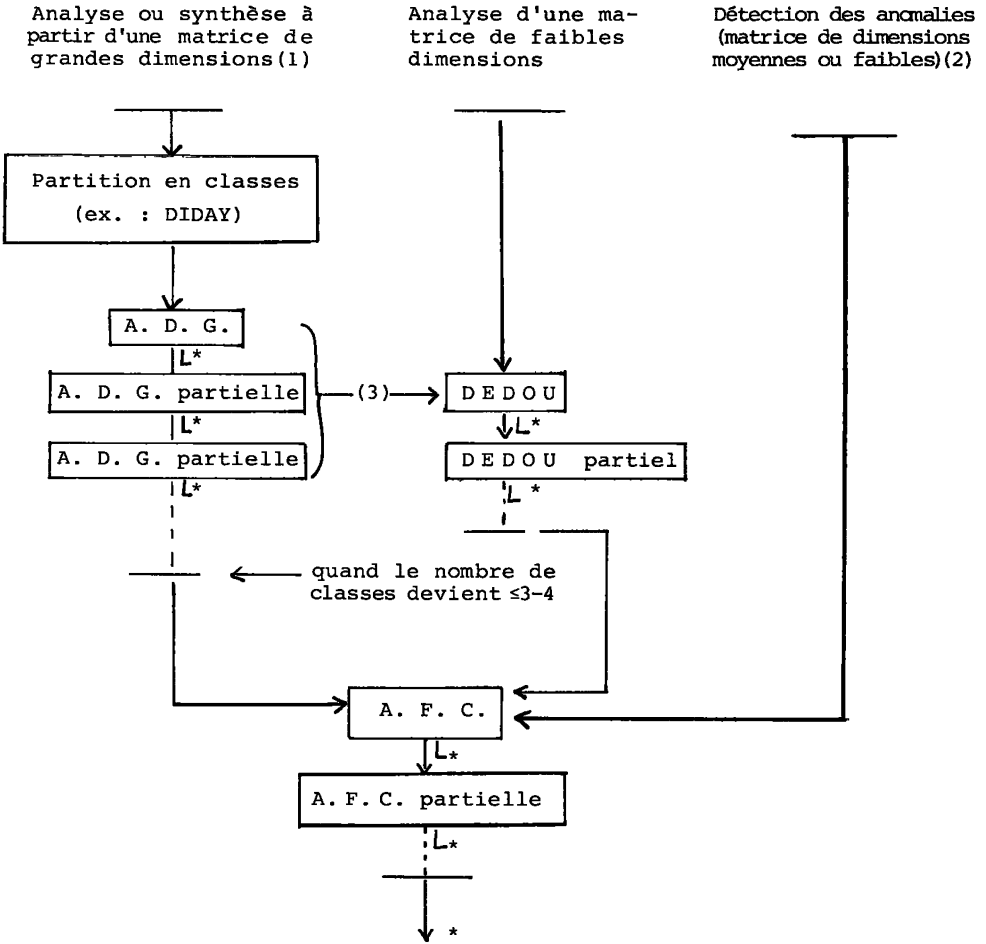


Fig. 2 - Méthodologie proposée pour le traitement des données phytosociologiques

\* Interprétation des cartes factorielles obtenues  
 (1) Cf. exemple pineraies  
 (2) Cf. note (1), p. 171  
 (3) Facultatif (sur une fraction des données)