MASTER EN SCIENCES BIOLOGIQUES

Comparaison de l'analyse en composantes principales et de l'analyse factorielle des correspondances, appliquées sur des tableaux de données écologiques du type échantillons-espèces

Frisque, Pascal

Award date:
1987

Awarding institution:
Universite de Namur

FACULTÉS UNIVERSITAIRES N.D.DE LA PAIX
NAMUR
FACULTÉ DES SCIENCES

Comparaison de l'analyse en composantes principales et de l'analyse factorielle des correspondances, appliquées sur des tableaux de données écologiques du type échantillons-espèces.

Mémoire présenté pour l'obtention du grade
de Licencié en Sciences
biologiques
par

Frisque Pascal

1986-1987

Facultés Universitaires Notre-Dame de la Paix
Faculté des Sciences
rue de Bruxelles 61, B-5000 **NAMUR**
Tél. 081-22.90.61 Télex 59222 facnam-b Téléfax 081-23.03.91

Comparaison de l'analyse en composantes principales et de l'analyse factorielle des correspondances, appliquées sur des tableaux de données écologiques du type échantillons-espèces.

FRISQUE Pascal

Errata.

Quelques modifications se sont avérées nécessaires, après publication, de façon à permettre une compréhension optimale.

Les pages suivantes représentent les copies corrigées. Si vous désirez un nouvel exemplaire du mémoire, je vous serais très obligé de me le faire savoir par l'intermédiaire du laboratoire de Biologie Quantitative, de même que vos remarques éventuelles.

Je signale aussi que les pages 102 et 103 ne devraient plus faire partie de la nouvelle édition. Cependant, la numérotation des pages suivantes restent identique, pour éviter des changements trop nombreux.

Je vous présente mes excuses pour les problèmes que ces modifications ont pu vous créés. Merci de votre compréhension.

Mémoire de licence en Sciences Zoologiques

Juin 1987

Laboratoire de Biologie Quantitative

Promoteurs : Prof. E. Feytmans et E. Depiereux

Facultés Universitaires Notre-Dame de la Paix
Faculté des Sciences
rue de Bruxelles 61, B-5000 **NAMUR**
Tél. 081-22.90.61 Télex 59222 facnam-b Téléfax 081-23.03.91

Comparaison de l'analyse en composantes principales et de l'analyse factorielle des correspondances, appliquées sur des tableaux de données écologiques du type échantillons-espèces.

FRISQUE Pascal

Résumé.

L'analyse en composantes principales (ACP) et l'analyse factorielle des correspondances (AFC) sont deux méthodes d'analyse multivariée couramment utilisées pour l'étude des tableaux de données écologiques du type échantillons-espèces, avec présence d'un gradient.

Ces deux techniques fournissent généralement des résultats très proches lorsqu'elles sont appliquées à des données réelles.

L'emploi de données simulées a permis de préciser les possibilités de chacune des deux analyses.

L'AFC se révèle supérieure dans la majorité des cas. Seul avantage de l'ACP : les déformations moindres des structures graphiques obtenues lorsque les colonnes présentent des poids très différents.

L'AFC s'avère donc souvent un meilleur choix, quoique certaines difficultés de méthodologie et d'interprétation requièrent une pratique courante pour tirer le meilleur parti de la richesse indiscutablement plus élevée des résultats fournis.

Abstract.

Principal components analysis (PCA) and correspondence analysis (CA) are two methods of multivariate analysis currently used for the study of tables of ecological data of the samples-species type, with the presence of a coenocline.

These two techniques generally give very similar results when applied to real data.

The use of simulated data lets it possible to specify the possibilities of each analysis.

AFC reveals itself to be superior in most cases. The only advantage of PCA : the slighter deformations of graphic structures obtained when the columns present very different weights.

So, AFC often becomes the best choice to make, though somes methodological and interpretative difficulties require a current practice to take the best of the obviously higher wealth of the given results.

Mémoire de licence en Sciences Zoologiques

Juin 1987

Laboratoire de Biologie Quantitative

Promoteurs : Prof. E. Feytmans et E. Depiereux

J'exprime ma profonde reconnaissance à Monsieur le Professeur Feytmans pour avoir mis à ma disposition tous les moyens nécessaires à la réalisation de ce mémoire.

Je tiens à manifester toute ma gratitude à Monsieur Depiereux qui m'a particulièrement stimulé par son intérêt soutenu et sa critique constructive.

Je remercie les personnes qui m'ont fourni des indications essentielles avec une grande disponibilité, spécialement Madame Dansart-Jacques et Madame Orban.

Je conserverai un souvenir agréable de cette année grâce notamment à Madame Van Vyve-Genette, Madame Beguin, Dominique Derême, André Decat et Etienne Pirotte.

A mes parents.

1. Introduction.

De plus en plus nombreuses sont les études permettant l'obtention de grands tableaux de résultats chiffrés.

Encore faut-il savoir qu'en faire.

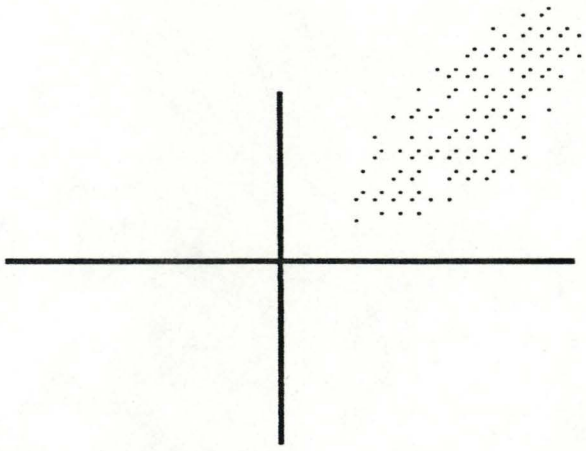
Les techniques d'analyse factorielle peuvent aider le chercheur à synthétiser ses données en y mettant de l'ordre et en révélant les aspects cachés d'une masse de chiffres. Ces analyses sont souvent un préliminaire indispensable au traitement statistique proprement dit, par leur faculté de générer des hypothèses qu'il faudra alors vérifier à l'aide d'une panoplie d'outils mathématiques appropriés.

Un exemple simple permettra de comprendre intuitivement ce que l'on peut attendre d'une telle analyse.

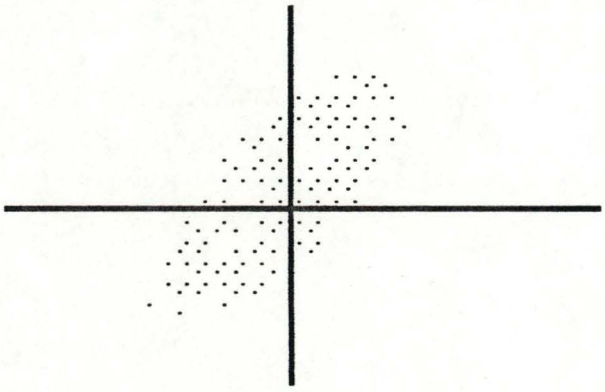
Considérons l'expérience suivante : un biologiste mesure la taille, le poids, le volume pour une série d'individus et il consigne les résultats dans un tableau en associant les individus aux lignes et les variables aux colonnes. Ses résultats acquis, il se demande ce qu'il va pouvoir en tirer. Il fait tourner une analyse en composantes principales sur son tableau et il constate qu'il peut résumer ses trois variables en une seule : la croissance, et que les individus se répartissent le long d'un axe représentant ce phénomène de croissance. Cet axe définit un nouvel espace, appelé espace factoriel, et représente une variable nouvelle, combinaison linéaire des variables originales. Il a donc mis de l'ordre en reclassant les individus et généré une hypothèse : les trois variables sont corrélées entre elles, et l'analyse les réunit pour n'en former qu'une, associée au nouvel axe. Dans ce cas, on parle d'ordination en espace réduit, le nombre de dimensions, ou nombre de variables, ayant diminué.

Ces méthodes ne sont évidemment utiles que si les dimensions du tableau ne permettent pas de l'interpréter à vue.

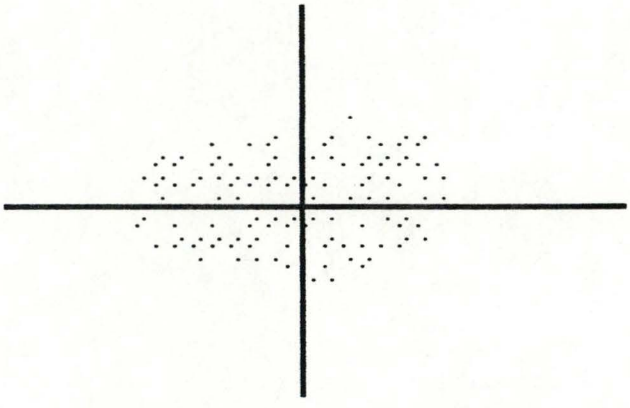
Les méthodes d'analyse factorielle sont très nombreuses et chacune possède ses spécificités propres. Il existe cependant des domaines d'études pour lesquels l'emploi d'une technique d'analyse préférentiellement à une autre n'est pas toujours clairement justifié. C'est une de ces situations de mitoyenneté indécise qui fournit le sujet de ce mémoire.



Représentation des données originales



Centrage des données



Rotation des axes due à l'analyse

Résumé visuel d'une analyse multivariée.

2. Historique des méthodes factorielles.

On peut ranger les techniques d'analyse des données, ou méthodes de statistique descriptive multidimensionnelle, en deux grandes familles : les méthodes factorielles et les méthodes de classification.

Les méthodes factorielles se proposent de fournir des représentations synthétiques de vastes ensembles de valeurs numériques. Les principes dont elles s'inspirent sont anciens, mais le développement et la diversification de ces techniques sont récents, et dûs incontestablement à la diffusion des possibilités du calcul électronique.

Les méthodes factorielles s'apparentent dans leur principe aux techniques d'analyse factorielle proposées et mises au point au début du siècle par les psychologues. Elles utilisent des calculs d'ajustement qui font essentiellement appel à l'algèbre linéaire, et produisent des représentations graphiques où les objets à décrire deviennent des points sur un axe ou dans un plan. L'analyse factorielle est basée sur la considération qu'un certain nombre de variables peuvent être la conséquence ou le résultat de qualités ou de caractères non mesurables (facteurs), les rapports entre eux s'exprimant analytiquement au moyen d'un système d'équations.

Les méthodes de classification sont plus récentes. Elles mettent en jeu une formulation et des calculs algorithmiques et produisent des classes ou des familles de classes, permettant de grouper et de ranger les objets à décrire (Benzécri, 1973 ; Dagnelie, 1975 ; Lebart & al., 1979 ; Legendre & Legendre, 1979).

Ces deux familles de méthodes sont plus complémentaires que concurrentes, et peuvent avec profit être utilisées conjointement sur un même jeu de données. Elles donnent chacune un point de vue sur les matériaux statistiques qui leur sont soumis.

2.1. Les différentes méthodes factorielles.

Les résultats obtenus par ces différentes méthodes d'approche ne peuvent pas être toujours considérés comme des résultats définitifs ; en effet, la décomposition en facteurs, quelque soit le modèle choisi, ne donne pas lieu à une solution unique. Par transformation ou rotation spatiale, on peut passer d'une solution directement trouvée à d'autres solutions plus représentatives ou plus suggestives (Torrens-Isbern, 1972).

L'analyse factorielle classique (ou analyse en facteurs communs et spécifiques), bien que n'étant pas une méthode purement descriptive, doit incontestablement être citée comme antécédent. Fondée et perfectionnée par Spearman (1904) et Thurstone (1947), cette méthode se proposait d'aller au-delà des apparences en permettant d'exhiber les variables qui échappent à l'observation directe. Ainsi les nombreuses notes obtenues à des tests psychologiques pourraient être expliquées par un très petit nombre de facteurs cachés, tels que la mémoire ou l'intelligence par exemple. Cette méthode, la plus ancienne, n'est plus guère utilisée en dehors du domaine de la psychologie, parce qu'elle fait appel à un modèle à priori assez restrictif (Torrens-Isbern, 1972).

Les difficultés qui devaient naître des tentatives de forcer la réalité à s'adapter à une conception préétablie conduisirent à des aménagements plus ou moins déguisés, avec développement de techniques particulières (Burt, Holzinger, Delaporte, etc). La rigueur de la théorie psychologique spearmanienne disparaît donc, et donne lieu à des conceptions plus souples. Thurstone, dans cette direction, arrive à l'extrême, puisqu'il accepte les facteurs en nombre indéfini, déterminés empiriquement par les résultats de l'analyse.

L'analyse en composantes principales (ACP) (Pearson, 1901 ; Hotelling, 1933) permet d'obtenir un résumé descriptif (sous forme de graphiques le plus souvent) d'un ensemble de n observations effectuées sur p variables numériques continues (Lebart & al., 1979).

L'analyse factorielle des correspondances (AFC) fut développée par Benzécri (1964, 1973, 1976, 1984). Les principes sous-jacents aux aspects techniques de cette méthode ont des antécédents assez anciens : on peut les faire remonter à certains travaux de Hirschfeld (1935) et de Fisher (1940). On consultera sur ce point les notes historiques de Hill (1974) et de Benzécri (1976). Ce sont les travaux de ce dernier auteur qui, avec l'avènement des ordinateurs, sont responsables du développement de cette méthode. Dans l'article précité, Benzécri mentionne également les travaux de Guttman (1941) et Hayashi (1950, 1952).

Si l'analyse permet d'obtenir la description de tableaux de contingence, elle peut aussi s'étendre à l'étude de tableaux de nombres positifs assez divers : tableaux d'incidence (codage binaire de présence-absence), ou encore tableaux codés sous forme disjonctive complète. Le choix de l'analyse des correspondances comme algorithme privilégié pour ces types de tableaux est justifié par certaines propriétés de la méthode. Les rôles analogues dévolus à chacune des deux dimensions des tableaux analysés trouvent des applications générales et permettent une lecture aisée des résultats, car les règles d'interprétation des proximités sont les mêmes pour les deux ensembles. Enfin le caractère optimal de la représentation simultanée facilite l'interprétation des résultats.

D'autres méthodes peuvent être considérées comme dérivées des précédentes : l'analyse des covariances partielles, l'analyse des rangs qui est une variante non paramétrique de l'analyse en composantes principales. L'analyse de Hotelling joue un rôle théorique important : elle contient en effet comme cas particulier la régression multiple et l'analyse discriminante. L'analyse des correspondances d'une table de contingence peut d'ailleurs être considérée comme une analyse discriminante appliquée à des codages particuliers (Lebart & al., 1979).

Lors des descriptions de tableaux, les méthodes d'analyse des données jouent le rôle d'instruments d'observations. Comme l'infiniment petit qui rend nécessaire l'usage du microscope, ou l'infiniment éloigné qui nécessite le téléscope ou la lunette, le multidimensionnel doit, pour être déchiffré, être soumis à des "programmes de calcul". On peut comparer ces algorithmes de réduction de données à un appareil radiographique qui fournit des images à partir d'une réalité inobservable (l'opacité des tissus, obstacle à la vision directe du squelette ou des organes étant alors l'analogue du caractère multidimensionnel des données, obstacle à leur assimilation). L'usage de l'appareil en vue du dépistage ou du diagnostic suppose une certaine préparation du sujet qui doit, par exemple, absorber des produits opacifiants ; il s'agira pour nous de procéder à un éventuel recodage des données, ou de transformer celles-ci en tenant compte d'informations exogènes.

L'interprétation des résultats est liée certainement à la *connaissance des principes* de fonctionnement de l'appareil : l'opacité aux rayons X dépend de la densité, du volume, de la composition chimique des organes : pour nous, il s'agira de connaître les principes géométriques des opérations effectuées sur les données. Mais l'obtention de clichés, comme celle des listages de résultats, ne constitue pas la phase la plus délicate ; l'expérience clinique du médecin, difficile à caractériser de façon précise, est irremplaçable. De même *l'expérience pratique*, que le statisticien devra acquérir "sur le tas", est indispensable.

2.2. Les applications.

L'analyse factorielle était, au début, surtout un outil de recherche psychologique, la plus grande majorité des études d'applications se rapportaient donc à ce domaine ; mais depuis une vingtaine d'années des essais d'utilisations sont apparus dans d'autres disciplines, chaque jour plus nombreux et de conceptions plus larges.

Ainsi Harman (1968) constate l'étendue de la gamme des applications autres que psychologiques de l'analyse factorielle dans les articles et livres, au nombre de 549, mentionnés dans la bibliographie de son ouvrage "Modern factor analysis". On y trouve des recherches dans les domaines de la sociologie, météorologie, rapports internationaux, économie, médecine, biologie, urbanisation et développement économique, communications, entomologie, étude des accidents, systèmes homme-machine, géologie, etc ... .

L'étendue des possibilités de l'analyse factorielle dans la recherche scientifique est énorme, presque illimitée, parce que les seules limites aux applications ne sont pas des préalables : on saura si l'hypothèse factorielle convenait au problème, uniquement après réalisation de l'analyse.

3. Utilisations de l'ACP et de l'AFC en écologie.

L'emploi des techniques d'ordination en espace réduit en écologie est pratiqué depuis une quinzaine d'années.

Ainsi Descy (1976) réalise une analyse en composantes principales sur un tableau composé des valeurs de 21 paramètres physico-chimiques et de densités relatives de 35 espèces de diatomées mesurées dans plusieurs stations de la Meuse et de la Sambre. Suivant l'ordination des paramètres physico-chimiques le long de la première composante principale, celle-ci peut être assimilée à un gradient de pollution. Les différentes espèces de diatomées s'ordonnent, par rapport à cet axe, en fonction de leur degré de résistance à la pollution.

De nombreuses études ont été réalisées depuis, notamment dans le cadre de doctorats réalisés en collaboration avec les facultés universitaires de Namur (Depiereux, 1982 ; Cornet, 1986 ; Dansart, 1986 ; Maquet, 1987 ; ...).

La plupart de ces études portaient sur des tableaux de type échantillons-espèces et utilisaient l'ACP comme méthode d'analyse multivariée.

Et pourquoi pas l'AFC ?

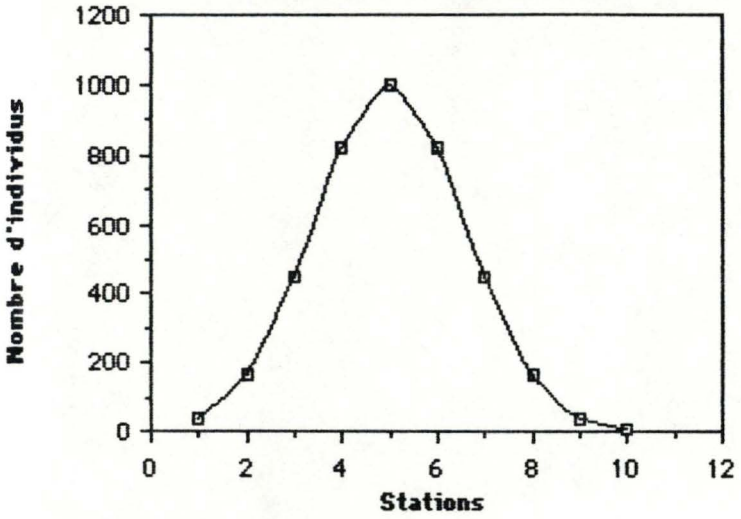
La majorité des travaux de ce genre publiés dans le monde utilisent l'ACP, plus ancienne et d'origine anglo-saxonne, en raison notamment de la puissance et de la simplicité de son modèle qui facilitent sa compréhension et son utilisation dans des domaines variés.

Néanmoins, un doute subsiste quant à l'opportunité de ce choix par rapport à l'AFC, les résultats fournis par les deux analyses étant très proches. Il s'avère donc nécessaire de préciser les capacités et les limites de chacune d'elles, afin de permettre leurs utilisations optimales dans le cadre bien défini des tableaux échantillons-espèces, obtenus en écologie.

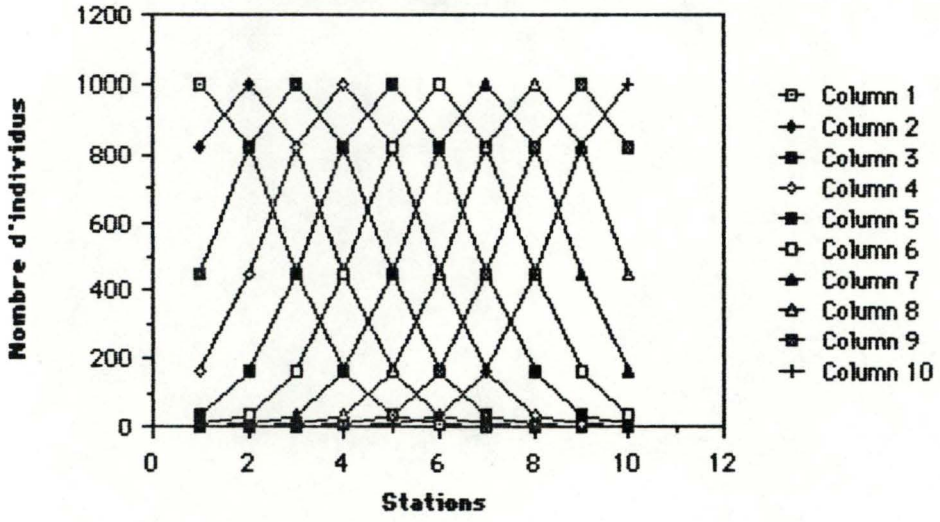
3.1. La notion de gradient.

Ces différentes études ont au moins une caractéristique commune : elles portent toutes sur un gradient, qu'il soit de pollution, climatique, physico-chimique, etc ... . Dans le cas présent, ce gradient est perceptible à travers les variations d'abondances des espèces dans les divers échantillons. La répartition des espèces dans ces stations sera fonction de leurs affinités pour les conditions régnant à ces endroits. C'est ici qu'interviennent les notions de valence écologique et de structure de peuplement. La valence écologique va définir la répartition des individus d'une espèce le long du gradient avec un maximum correspondant aux conditions optimales de développement de l'espèce. La structure de peuplement d'un échantillon décrit la composition faunistique ou floristique de ce prélèvement et les abondances relatives des différents taxons.

Distribution d'une espèce le long d'un gradient



Exemple de gradient



4. Objectifs.

1) Réaliser les présentations parallèles de l'AFC et de l'ACP, en utilisant comme références les principaux ouvrages et travaux publiés dans ce domaine.

2) Ecriture d'un programme original, rendu nécessaire pour répondre aux besoins particuliers du travail.

3) Tester l'information fournie par deux méthodes d'analyse factorielle, l'ACP et l'AFC, appliquées sur des tableaux du type échantillons-espèces, en utilisant des données simulées.

Cette étude devrait permettre de guider le choix de l'utilisateur potentiel de l'une des deux méthodes, en fonction des caractéristiques de ses données et du but qu'il poursuit.

4.1. Remarques préliminaires.

Le but des exposés sur l'AFC et l'ACP n'est pas de fournir des démonstrations détaillées pour ces analyses, mais de rassembler en un texte des éléments théoriques présents dans divers travaux, en unifiant leurs notations.

Le choix de présenter l'AFC avant l'ACP, contrairement à l'usage courant, est justifié par l'introduction d'éléments théoriques nouveaux pour l'ACP, mais d'utilisation habituelle pour l'AFC .

On pourra trouver des compléments d'information et toutes les démonstrations dans les ouvrages et publications cités dans la bibliographie sélective proposée M. Volle (1985), reprise dans la bibliographie générale.

Je me permet de recommander le livre de M. Volle (1985) pour la qualité de son information et son caractère attractif.

5. Théorie de l'AFC.

5.1. Présentation.

L'analyse factorielle des correspondances a d'abord pour but l'analyse des tableaux de contingence qui comparent les différentes descriptions d'un premier descripteur (les lignes du tableau), aux descriptions d'un second descripteur (les colonnes du tableau). Chaque cellule du tableau représente la fréquence d'observation des réalisations conjointes des deux descriptions correspondantes, l'une associée à la ligne et l'autre à la colonne. Ces fréquences d'objets sont donc des entiers positifs ou nuls.

L'analyse peut aussi porter sur des tableaux de contingence qui comparent deux groupes de descripteurs : chaque descripteur comportant autant de lignes ou de colonnes qu'il a de descriptions différentes (ex. : descripteur = genre, descriptions = espèces).

Un tableau de contingence n lignes fois p colonnes met en correspondance deux ensembles finis : l'ensemble I des n vecteurs lignes dans l'espace à p dimensions des colonnes et l'ensemble J des p vecteurs colonnes dans l'espace à n dimensions des lignes.

L'analyse des correspondances est une méthode d'ordination qui préserve, dans l'espace factoriel, la distance euclidienne entre des profils de probabilités conditionnelles pondérées. Ceci équivaut à préserver la distance du χ^2 (distance distributionnelle ou distance du CHI-2) entre les lignes ou entre les colonnes du tableau de contingence. En d'autres mots : on compare des profils.

L'analyse permet l'étude des proximités entre les lignes ou les colonnes, ainsi que des correspondances entre lignes et colonnes du tableau de contingence.

Elle effectue la mesure des écarts entre les données initiales et les valeurs qu'elles auraient en situation d'indépendance des descripteurs. S'il n'y a aucune dépendance, tous les écarts sont nuls et il n'y a rien à étudier. La matrice de données définit le modèle d'indépendance (distance du χ^2.

5.2. Notations.

Le ' indique la transposition d'une matrice.

R = espace des réels.
λ = valeur propre.

SQR = racine carrée.

K = $[k_{ij}]$ (matrice des données, des fréquences absolues.)

$k_{i.} = \Sigma_j k_{ij}$ (fréquences marginales (poids) des lignes.)
$k_{.j} = \Sigma_i k_{ij}$ (fréquences marginales (poids) des colonnes.)

$k_{..} = \Sigma_i \Sigma_j k_{ij}$

$f_{ij} = k_{ij} / k_{..}$ (fréquences relatives.)
$f_{i.} = \Sigma_j f_{ij}$ (fréquences relatives marginales (masses) des lignes.)
$= k_{i.} / k_{..}$
$f_{.j} = \Sigma_i f_{ij}$ (fréquences relatives marginales (masses) des colonnes.)
$= k_{.j} / k_{..}$

H = $[h_{ij}]$ (matrice des données transformées suivant la théorie).
X = $[x_{ij}]$ (matrice des données transformées suivant la voie "calcul").
T = $[t_{jj}]$ (matrice de dispersion (de variance-covariance)).
θ = $[\theta_{i\alpha}]$ (coordonnées des points-lignes dans l'espace factoriel).
Ω = $[\Omega_{j\alpha}]$ (coordonnées des points-colonnes dans l'espace factoriel).
V = $[v_{j\alpha}]$ (matrice des vecteurs propres).
U = $[u_{j\alpha}]$ (matrice des vecteurs propres normés à 1).

i = 1 ... n (nombre de lignes).
j = 1 ... p (nombre de colonnes).
α = 1 ... q (nombre de valeurs propres non nulles).

Le nombre de valeurs propres non nulles est inférieur ou égal à la plus petite dimension de la matrice des données initiales moins 1.

5.3. Notions préliminaires.

5.3.1. *La notion de profil.*

L'AFC ne compare pas des effectifs absolus mais des proportions. La notion de profil est essentielle : on appelle profil d'une ligne ou d'une colonne la distribution des fréquences conditionnelles relatives de ses éléments.

Ainsi chaque ligne a un profil sur les colonnes :

Profil de la ligne i = $\{ k_{ij} / k_{i.} ; j = 1 \ldots p \}$

Et chaque colonne a un profil sur les lignes :

Profil de la colonne j = $\{ k_{ij} / k_{.j} ; i = 1 \ldots n \}$

L'AFC mettra en évidence comment et avec quelle intensité deux profils se ressemblent ou se distinguent entre eux et par rapport à un profil de référence qui est le profil moyen.

Le profil de référence pour les lignes :

$\{ k_{.j} / k_{..} ; j = 1 \ldots p \}$

Le profil de référence pour les colonnes :

$\{ k_{i.} / k_{..} ; i = 1 \ldots n \}$

5.3.2. *Construction du nuage de points.*

a) Dans l'espace R^p, on construit un nuage de n points, chaque point i ayant pour coordonnées les quantités :

$\{ f_{ij} / f_{i.} ; j = 1 \ldots p \}$ (profil de la ligne i)

et étant affecté de la masse $f_{i.}$.

Ces n points seront situés dans un sous-espace à p-1 dimensions puisque leurs p coordonnées vérifient la relation :

$\Sigma_j (f_{ij} / f_{i.}) = 1$ pour tout i = 1 ... n

Les proximités entre points s'interprètent alors en termes de proximités entre profils.

b) Dans l'espace R^n, on construit un nuage de p points, chaque point j ayant pour coordonnées les quantités :

$\{ f_{ij} / f_{.j} ; i = 1 \ldots n \}$ (profil de la colonne j)

et étant affecté de la masse $f_{.j}$

Ces p points seront situés dans un sous-espace à n-1 dimensions puisque leurs n coordonnées vérifient la relation :

$\Sigma_i (f_{ij} / f_{.j}) = 1$ pour tout j = 1 ... p

5.3.3. *Choix des distances.*

Le fait de travailler dans les deux espaces R^p et R^n incite à munir ces espaces d'une distance différente de la distance euclidienne usuelle. La distance entre deux lignes i et i' sera donnée par la formule qui exprime la distance dite du CHI-2 (ou distance distributionnelle) :

$$(1)\quad d^2(i,i') = \Sigma_j \; (1 / f_{.j}) * ((f_{ij} / f_{i.}) - (f_{i'j} / f_{i'.}))^2$$

De la même façon, la distance entre deux colonnes sera donnée par :

$$(2)\quad d^2(j,j') = \Sigma_i \; (1 / f_{i.}) * ((f_{ij} / f_{.j}) - (f_{ij'} / f_{.j'}))^2$$

Cette distance "pondérée" a surtout l'avantage de vérifier le principe d'équivalence distributionnelle :

> Si deux points-lignes i_1 et i_2 sont confondus dans R^p et si on les considère comme un seul point affecté de la somme des masses de i_1 et de i_2 (i_1 et i_2 seront remplacés par i_0) alors les distances entre tous les couples de points dans R^p et R^n restent inchangées. Il en est de même pour deux points-colonnes j_1 et j_2 ayant les mêmes propriétés.

Ce qui peut se dire mathématiquement :

Si les points i_1 et i_2 sont confondus dans R^p, on a, pour tout j :

$$f_{i_1 j} / f_{i_1 .} = f_{i_2 j} / f_{i_2 .} = f_{i_0 j} / f_{i_0 .}$$

On a donc en particulier :

$$(f_{i_1 j} + f_{i_2 j}) / (f_{i_1 .} + f_{i_2 .}) = f_{i_0 j} / f_{i_0 .}$$

D'où, puisque les dénominateurs sont égaux, pour tout j :

$$f_{i_1 j} + f_{i_2 j} = f_{i_0 j}$$

Les calculs des quantités $f_{.j} = \Sigma_i \; f_{ij}$ ne sont donc pas affectés, et donc les distances $d^2(i,i')$ données par la formule (1) sont invariantes.

5.4. Analyse dans R^p.

5.4.1. *Transformation des données initiales.*

On peut vérifier que la distance distributionnelle entre deux lignes ou colonnes du tableau initial est égale à la distance euclidienne entre les deux lignes ou colonnes correspondantes calculée sur les données initiales transformées par la formule suivante :

(3) $f_{ij} / (f_{i.} * SQR(f_{.j}))$

alors la distance euclidienne usuelle de deux points i et i' vaut :

(4) $d^2(i,i') = \Sigma_j \{ [f_{ij} / (f_{i.} * SQR(f_{.j}))] - [f_{i'j} / (f_{i'.} * SQR(f_{.j}))] \}^2$

5.4.2. *Calcul du centre de gravité.*

Le centre de gravité du nuage est défini par ses coordonnées dans R^p :

$\{ SQR(f_{.j}) ; j = 1 \ldots p \}$

Après translation de l'origine au centre de gravité du nuage, les coordonnées du point i s'écrivent :

$h_{ij} = \{ [f_{ij} / (f_{i.} * SQR(f_{.j}))] - SQR(f_{.j}) ; i = 1 \ldots n \ \& \ j = 1 \ldots p \}$

5.4.3. *Calcul de la matrice T de dispersion (de variance-covariance).*

Le terme général s'écrit :

$t_{jj'} = \Sigma_i \{ f_{i.} * h_{ij} * h_{ij'} \}$

Il est possible de donner à cette matrice une forme simple, on pose :

$x_{ij} = (f_{ij} - (f_{i.} * f_{.j})) / SQR(f_{i.} * f_{.j})$

Alors la matrice T s'exprime en fonction de la matrice $X = [x_{ij}]$:

$T = X' * X$

5.4.4. *Calcul des valeurs et vecteurs propres.*

On calcule les valeurs propres λ_α et les vecteurs propres u_α correspondants au moyen de l'équation habituelle :

$(T - (\lambda_\alpha * I)) * u_\alpha = 0$

Puisque T est symétrique, les vecteurs propres sont orthogonaux.

5.4.5. *Normalisation des vecteurs propres.*

Les vecteurs propres sont normés à 1 :

$u_{j\alpha} = v_{j\alpha} / SQR(norme\alpha)$ $\qquad$ $norme\alpha = \Sigma_j (v_{j\alpha})^2$

5.4.6. *Coordonnées des points-lignes dans l'espace factoriel.*

La position des lignes du tableau initial dans l'espace factoriel est donnée par la matrice transformée des composantes initiales :

$$\theta = H * U$$

La projection du point i sur l'axe α aura pour coordonnées :

$$(5)\ \Sigma_j \{ [(f_{ij} / (f_{i.} * SQR(f_{.j}))) - SQR(f_{.j})] * u_{j\alpha} \}$$

Les vecteurs propres étant orthogonaux :

$$\Sigma_j \{ u_{j\alpha} * SQR(f_{.j}) \} = 0$$

En vertu de cette propriété, la relation (5) se simplifie :

$$\Sigma_j \{ [f_{ij} / (f_{i.} * SQR(f_{.j}))] * u_{j\alpha} \}$$

Les coordonnées barycentrées (de telle sorte que l'origine des axes corresponde au centre des masses) seront :

$$(1 / SQR(\lambda_\alpha)) * \Sigma_j \{ [f_{ij} / (f_{i.} * SQR(f_{.j}))] * u_{j\alpha} \}$$

5.5. Analyse dans R^n.

5.5.1. *Calcul du centre de gravité.*

Les coordonnées du centre de gravité du nuage dans R^n sont :

$$\{ SQR(f_{i.}) \; ; \; i = 1 \; \ldots \; n \}$$

5.5.2. *Coordonnées des points-colonnes dans l'espace factoriel.*

Il existe une relation simple entre les coordonnées des points-lignes et les coordonnées des points-colonnes.

Les points-colonnes sur l'axe α auront les coordonnées barycentrées suivantes :

$$(5) \quad \Omega_{j\alpha} = (1 \; / \; SQR(\lambda_\alpha)) * \Sigma_i \; \{ (f_{ij} \; / \; f_{.j}) * \theta_{i\alpha} \}$$

De même :

$$(6) \quad \theta_{i\alpha} = (1 \; / \; SQR(\lambda_\alpha)) * \Sigma_j \; \{ (f_{ij} \; / \; f_{i.}) * \Omega_{j\alpha} \}$$

Remarques.

1) Les relations (5) et (6) impliquent que pour tout α, on a $\lambda_\alpha \leq 1$: toutes les valeurs propres sont inférieures à 1 (Lebart & al., 1979).

2) L'origine des axes correspond aux profils marginaux, ou profils moyens des lignes et des colonnes. Les proximités entre profils s'interprètent donc toujours par référence aux profils moyens.

5.6. Aides à l'interprétation.

5.6.1. *La trace et les valeurs propres.*

Les valeurs propres sont des éléments très importants dans l'interprétation des graphiques des projections, car elles permettent de relativiser en termes d'écart au profil moyen les valeurs des échelles des axes.

La trace se définit comme la somme des valeurs propres. Les différentes voies de calcul sont :

$$\text{trace} = \Sigma_\alpha \lambda_\alpha$$
$$= \Sigma_{ij} (x_{ij})^2$$

5.6.2. *Les contributions.*

Pour interpréter les axes déterminés lors d'une analyse des correspondances, on calcule trois séries de coefficients pour chacun des éléments des deux ensembles mis en correspondance :

a) les contributions à l'analyse, qui expriment la part prise par un élément donné (ligne ou colonne) dans la variance totale (trace) exprimée par l'ensemble des facteurs ;

b) les contributions absolues, qui expriment la part prise par un élément donné dans la variance exprimée par un facteur ;

c) les contributions relatives, ou corrélations éléments-facteurs, qui expriment la part prise par un facteur dans "l'explication" de la dispersion d'un élément. C'est la mesure de la qualité de la représentation d'un point par sa projection sur l'axe α.

Alors que les contributions absolues permettront de savoir quelles variables sont responsables de la construction d'un facteur, les contributions relatives exhiberont celles qui sont des caractéristiques exclusives de ce facteur. Les contributions absolues et relatives sont estimées à partir des coordonnées des points-lignes et des points-colonnes non barycentrés. L'utilisation de ces coefficients sera précisée à l'occasion de l'exemple numérique.

a) Contributions à l'analyse.

Contribution (C) de la ligne i à l'analyse :

$$C(i) = \Sigma_j x_{ij}^2 \ / \ \text{trace}$$

Contribution (C) de la colonne j à l'analyse :

$$C(j) = \Sigma_i x_{ij}^2 \ / \ \text{trace}$$

On notera :

$$\Sigma i \ C(i) = 1$$

$$\Sigma j \ C(j) = 1$$

Plusieurs lignes ou colonnes ne peuvent donc contribuer ensemble et fortement à l'analyse, c'est-à-dire à la variance totale.

b) Contributions absolues.

Contribution absolue (Ca) de la ligne i à l'axe α :

$$Ca_\alpha(i) = f_{i.} * \theta_{i\alpha}^2 / \lambda_\alpha$$

Contribution absolue (Ca) de la colonne j à l'axe α :

$$Ca_\alpha(j) = f_{.j} * \Omega_{j\alpha}^2 / \lambda_\alpha$$

On notera :

$\Sigma_i\ Ca_\alpha(i) = 1$ pour tout axe α

$\Sigma_j\ Ca_\alpha(j) = 1$ pour tout axe α

Plusieurs lignes ou colonnes ne peuvent donc contribuer ensemble et fortement à la construction d'un facteur.

c) Contributions relatives.

Les axes factoriels de chaque espace constituent des bases orthonormées. Le carré de la distance d'un point au centre de gravité (noté G ou H suivant l'espace) se décompose donc en somme des carrés des coordonnées sur ces axes.

Pour un point i de R^p :

$$d^2(i,G) = \Sigma_j \{ [(f_{ij} / (f_{i.} * SQR(f_{.j}))) - SQR(f_{.j})]^2 \}$$

Pour un point j de R^n :

$$d^2(j,H) = \Sigma_i \{ [(f_{ij} / (f_{.j} * SQR(f_{i.}))) - SQR(f_{i.})]^2 \}$$

On remarque que ces distances s'annulent lorsque le profil d'un point est égal au profil moyen.

Contribution relative (Cr) du facteur α à la position de l'élément i :

$$Cr_\alpha(i) = \theta_{i\alpha}^2 / d^2(i,G)$$
$$= \theta_{i\alpha}^2 / \Sigma_\alpha\ \theta_{i\alpha}^2$$

est donc le carré du cosinus du point i avec l'axe α. Il s'interprète comme le carré d'un coefficient de corrélation.

Contribution relative (Cr) du facteur α à la position de l'élément j :

$$Cr_\alpha(j) = \Omega_{j\alpha}^2 / d^2(j,H)$$
$$= \Omega_{j\alpha}^2 / \Sigma_\alpha\ \Omega_{j\alpha}^2$$

On notera :

$\Sigma_\alpha\ Cr_\alpha(i) = 1$ pour tout i

$\Sigma_\alpha\ Cr_\alpha(j) = 1$ pour tout j

Plusieurs facteurs ne peuvent être associés fortement et en même temps à un élément (ligne ou colonne).

5.6.3. *Test d'ajustement à la loi d'indépendance.*

Si les f_{ij} étaient des probabilités, l'hypothèse d'indépendance des lignes et des colonnes s'écrirait :

$$f_{ij} = f_{i.} * f_{.j}$$

On est donc ramené à comparer la fréquence observée kij à la fréquence théorique :

$$k.. * f_{i.} * f_{.j}$$

La quantité χ^2 s'écrit :

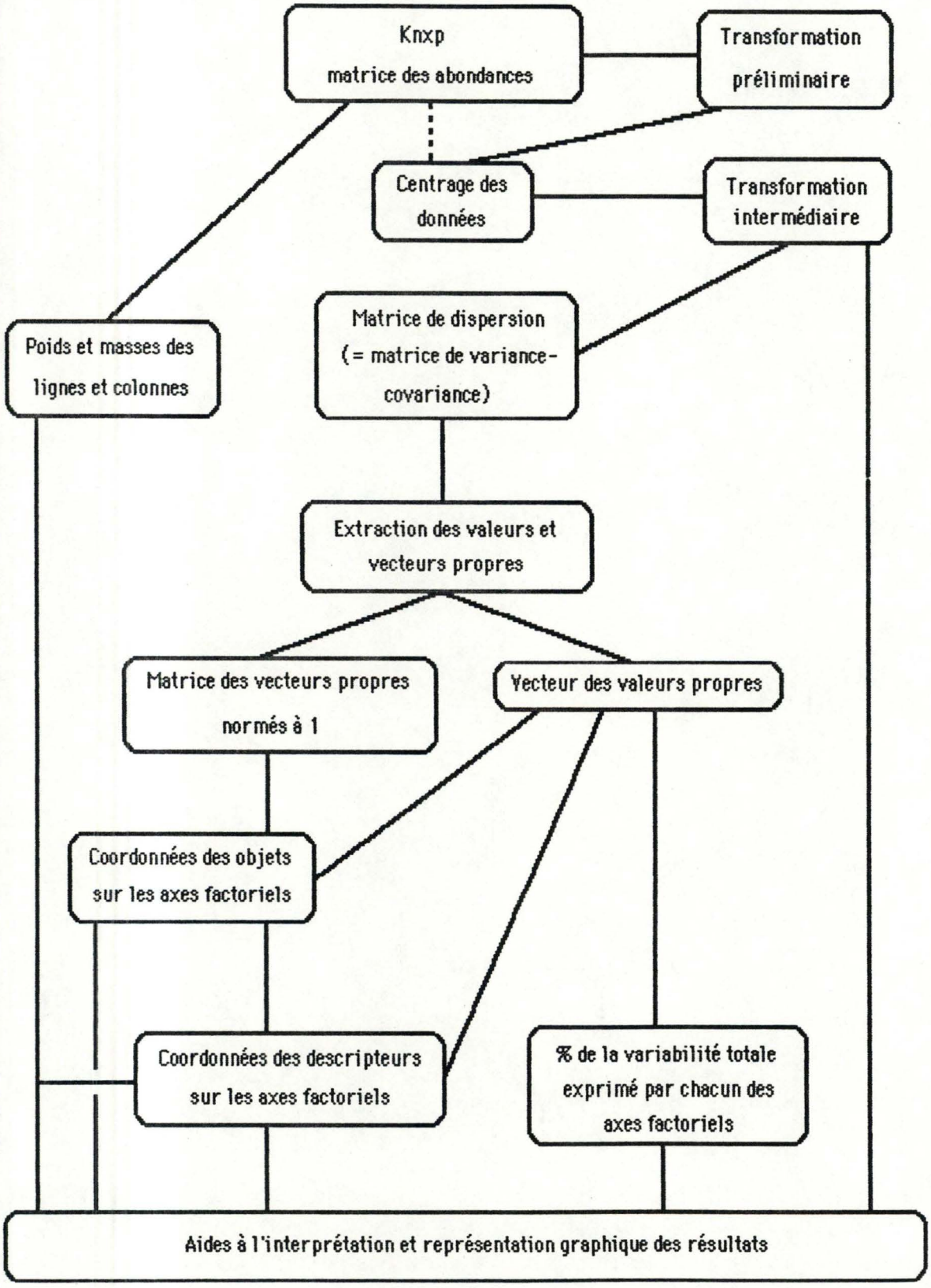
$$\chi^2 = k.. * \Sigma_i\Sigma_j \{ (f_{ij} - (f_{i.} * f_{.j}))^2 / (f_{i.} * f_{.j}) \}$$

avec (n-1) * (p-1) degrés de liberté.

On peut montrer aussi que :

$$\chi^2 = k.. * \text{trace}$$

Cette relation est importante, car en montrant que la trace est une fonction directe du χ^2, elle permet, comme indiqué plus haut, de relativiser en termes d'écarts au profil moyen les valeurs des échelles des axes de l'espace factoriel (par l'intermédiaire de la valeur propre associée à l'axe).



Résumé schématique de l'AFC.

6. Théorie de l'ACP.

6.1. Généralités.

On utilisera l'ACP lorsqu'il s'agira de décrire un tableau K de valeurs numériques continues du type : "variables-individus". L'ACP intervient dans l'interprétation des relations existant entre une série de variables interdépendantes. Son but principal est de condenser l'essentiel des informations données par ces variables interdépendantes, observées directement, en un nombre plus restreint de variables fondamentales indépendantes, qu'on ne peut observer directement. Tel est notamment le cas en biométrie, lorsque l'on a mesuré diverses caractéristiques (poids, taille, etc ...) d'un ensemble d'organismes vivants, et que l'on désire condenser ces observations en un petit nombre de caractéristiques fondamentales (croissance en hauteur, en largeur, etc).

Son utilisation est le plus souvent associée à l'analyse de tableaux autre que les tableaux de contingence.

6.2. Notations.

Le ' indique la transposition d'une matrice.

R = espace des réels.
λ = valeur propre.

SQR = racine carrée.

$K = [\ k_{ij}\ ]$ (matrice des données originales).

$m_j = \Sigma_i\ k_{ij} / n$ (moyennes des variables par colonnes).
$s_j = SQR(\ \Sigma_i\ (k_{ij} - m_j)^2 / n\ )$ (écart-types des variables par colonnes).

$k_{i.} = \Sigma_j\ k_{ij}$ (fréquences marginales (poids) des lignes).
$k_{.j} = \Sigma_i\ k_{ij}$ (fréquences marginales (poids) des colonnes).

$Z = [\ z_{ij}\ ]$ (matrice des données standardisées).
$C = [\ c_{ij}\ ]$ (matrice de dispersion (des corrélations entre variables)).
$Y = [\ y_{i\alpha}\ ]$ (coordonnées des points-lignes dans l'espace factoriel).
$F = [\ f_{j\alpha}\ ]$ (matrice des vecteurs normés aux valeurs propres).
$U = [\ u_{j\alpha}\ ]$ (matrice des vecteurs propres).

i = 1 ... n (nombre de lignes de K).
j = 1 ... p (nombre de lignes de K).
α = 1 ... q (nombre de valeurs propres non nulles).

Le nombre de valeurs propres non nulles est inférieur ou égal à la plus petite dimension de la matrice des données originales.

6.3. Choix de la distance.

On utilisera la métrique euclidienne canonique :

$$d^2(i,i') = \Sigma_j\ ((k_{ij} - k_{i'j})^2 / s_j^2)$$

6.4. Analyse dans R^p.

6.4.1. *Transformation des données initiales.*

Les n points de cet espace sont des individus. On veut obtenir une représentation des proximités entre ces points dans un espace de faible dimension. Or le sous-espace à une dimension qui assure une déformation minimale des proximités entre points n'a aucune raison d'être assujetti à passer par l'origine des axes. En effet, ce n'est pas la position du nuage par rapport à l'origine qui nous intéresse, mais la forme du nuage.

Cela va nous conduire à prendre comme nouvelle origine le centre de gravité du nuage dont les p composantes sont les p moyennes arithmétiques des p variables (colonnes).

La transformation est :

$$k_{ij} - m_j$$

Une modification supplémentaire du tableau de départ peut également être nécessaire si les dispersions des variables sont très différentes ou si les valeurs que prennent ces variables pour chacun des individus sont exprimées dans des unités différentes, auquel cas une comparaison de ces variables n'a plus de sens.

Ces modifications conduisent à l'analyse en composantes principales normée, c'est-à-dire l'analyse du tableau dont les données sont standardisées (centrées et réduites) :

$$z_{ij} = (k_{ij} - m_j) / s_j$$

On notera que les moyennes des variables standardisées sont nulles et les variances égales à un ; ainsi chaque variable aura une contribution analogue dans la détermination des proximités entre les points-lignes.

Ce cas sera le seul considéré dans la suite de ce travail.

6.4.2. *Calcul de la matrice C de dispersion.*

La matrice de variance-covariance s'obtient par :

$$(1/n)\ Z' * Z$$

Du fait de la division de chaque terme de K par l'écart-type de la variable colonne correspondante, la matrice de variance-covariance représente en fait la matrice de corrélation C.

6.4.3. *Calcul des valeurs et vecteurs propres.*

On calcule les valeurs propres λ_α et les vecteurs propres u_α au moyen de l'équation habituelle :

$$(C - (\lambda_\alpha * I)) * u_\alpha = 0$$

Puisque C est symétrique, les vecteurs propres sont orthogonaux.

6.4.4. *Normalisation des vecteurs propres.*

Les vecteurs propres sont normés à 1 :

$u_{j\alpha} = u_{j\alpha} / SQR(norme\alpha)$ $\qquad$ $norme\alpha = \Sigma_j \ (u_{j\alpha})^2$

6.4.5. *Coordonnées des points-lignes dans l'espace factoriel.*

La position des lignes du tableau des zscores dans l'espace factoriel est donnée par la matrice transformée des composantes initiales :

$Y = Z * U$

6.5. Analyse dans Rn.

6.5.1. *Développement théorique.*

Les p points du nuage dans cet espace sont maintenant les points-variables. Les indices i et j ne jouant pas des rôles similaires dans la transformation initiale, les interprétations géométriques associées à cette transformation seront radicalement différentes.

Ainsi la transformation $k_{ij} - m_j$ qui était interprétée comme une translation de l'origine dans R^p (centrage), revêt maintenant une signification différente puisque toutes les composantes de la coordonnée d'un point sont affectées de la même façon, mais chaque point l'est différemment. C'est donc chaque point qui subit une translation particulière qui provoque une première déformation du nuage.

Le changement d'échelle des axes, c'est-à-dire la division de chaque coordonnée de R^p par s_j, devient une seconde déformation du nuage qui ramène chacun des points-variables à la distance 1 de l'origine, si on inclut la division par n dans le calcul de z_{ij} :

$$d^2(j,0) = (1/n) \Sigma_i z_{ij}^2 = 1$$

ce qui équivaut à calculer la variance des données standardisées.

Les p points-variables sont donc sur une hypersphère de rayon 1, centrée à l'origine qui est le point moyen du nuage ainsi transformé.

Les coordonnées de projection d'un point-variable représentent la position de la projection du sommet de cet axe-descripteur dans l'espace réduit.

La distance entre deux points-variables j et j' s'écrit :

$$d^2(j,j') = \Sigma_i (z_{ij} - z_{ij'})^2$$

Après développement du carré et sommation sur l'indice i, on trouve :

$$d^2(j,j') = 2 * (1 - c_{jj'})$$

où $c_{jj'}$ est le coefficient de corrélation entre les variables j et j'.

Ainsi les proximités entre points-variables pourront s'interpréter en termes de corrélations : les points sont très proches si leur corrélation est fortement positive et éloignés si leur corrélation est fortement négative.

De même, on peut montrer que :

$$(1) \quad \cos(\tau_{jj'}) = c_{jj'}$$

où $\tau_{jj'}$ est l'angle formé par les vecteurs (0,j) et (0,j') dans l'espace factoriel et $c_{jj'}$ est le coefficient de corrélation entre les deux variables (descripteurs) j et j'.

Ainsi, la coordonnée de chaque point-variable sur un axe factoriel représente aussi le coefficient de corrélation entre la variable et le facteur α considéré.

Cet aspect de l'analyse dans Rn sera développé à l'aide d'un exemple numérique.

6.5.2. *Voie calcul.*

Il s'agit de normer les vecteurs propres aux valeurs propres :

$f_{j\alpha} = u_{j\alpha} * SQR(\lambda\alpha) / SQR(norme\alpha)$ $\qquad$ $norme\alpha = \Sigma_j (u_{j\alpha})^2$

On notera :

$\Sigma_\alpha f_{j\alpha}^2 = 1$

une même variable ne peut donc être fortement corrélée avec plusieurs composantes, et :

$\Sigma_j f_{j\alpha}^2 = \lambda_\alpha$

une composante associée à une valeur propre faible ne peut donc pas être fortement corrélée avec plusieurs variables.

6.6. Aides à l'interprétation.

6.6.1. *La trace et les valeurs propres.*

La trace est une mesure de la variance totale expliquée par les axes factoriels, la valeur propre est une mesure de cette même variance expliquée par axe.

valeur de la trace :

$$\begin{aligned} \text{trace} &= \Sigma_\alpha \lambda_\alpha \\ &= \Sigma_{ij} (z_{ij})^2 \\ &= p \end{aligned}$$

6.6.2. *Les contributions.*

Bien que l'analyse se suffise déjà à elle-même par l'intermédiaire de ses résultats, on peut, comme pour l'AFC, disposer des aides à l'interprétation que sont les contributions absolues, relatives et à l'analyse.

Les indications que fournissent ces contributions ne sont pas mentionnées dans la plupart des logiciels connus, celles-ci faisant souvent double emploi avec les résultats propres à l'analyse. Cependant, ces valeurs peuvent être intéressantes si l'on désire comparer l'ACP et l'AFC tournant sur un même tableau de données initiales.

Contribution (C) de la ligne i à l'analyse :

$$C(i) = \Sigma_j z_{ij}^2 \; / \; (\text{trace} * n)$$

Contribution absolue (Ca) de la ligne i à l'axe α :

$$Ca_\alpha(i) = y_{i\alpha}^2 \; / \; (\lambda_\alpha * n)$$

Contribution relative (Cr) du facteur α à l'élément i :

$$Cr_\alpha(i) = y_{i\alpha}^2 \; / \; \Sigma_\alpha \, y_{i\alpha}^2$$

Contribution (C) de la colonne j à l'analyse :

$$C(j) = \Sigma_i z_{ij}^2 \; / \; (\text{trace} * n)$$

Contribution absolue (Ca) de la colonne j à l'axe α :

$$\begin{aligned} Ca_\alpha(j) &= f_{j\alpha}^2 \; / \; \lambda_\alpha \\ &= f_{j\alpha}^2 \; / \; \Sigma_j \, f_{j\alpha}^2 \end{aligned} \qquad \lambda_\alpha = \Sigma_j \, f_{j\alpha}^2$$

Contribution relative (Cr) du facteur α à l'élément j :

$$\begin{aligned} Cr_\alpha(j) &= f_{j\alpha}^2 \; / \; \Sigma_\alpha \, f_{j\alpha}^2 \\ &= f_{j\alpha}^2 \end{aligned} \qquad \Sigma_\alpha \, f_{j\alpha}^2 = 1$$

D'après la formule (1), $Cr_\alpha(j)$ s'interprète comme le carré du coefficient de corrélation de la variable j avec l'axe α.

Il faut rapprocher ceci des $\cos^2$ obtenus par l'AFC, qui ont la même signification.

On notera :

$\Sigma_i\ C(i) = 1$ $\qquad\qquad$ $\Sigma_j\ C(j) = 1$

$\Sigma_i\ Ca_\alpha(i) = 1$ $\qquad\qquad$ $\Sigma_j\ Ca_\alpha(j) = 1$

$\Sigma_\alpha\ Cr_\alpha(i) = 1$ $\qquad\qquad$ $\Sigma_\alpha\ Cr_\alpha(j) = 1$

Les enseignements que l'on pourra en tirer seront les mêmes que pour l'AFC.

Ils peuvent donc servir de points de comparaison entre les deux analyses, ce qui sera montré avec l'exemple numérique.

6.6.3. *Superposition des deux espaces R^n et R^p.*

Si l'échelle des coordonnées des points-variables a une interprétation en termes de corrélations, il n'en est pas de même pour les points-individus. On applique à leurs coordonnées un coefficient dont la valeur :

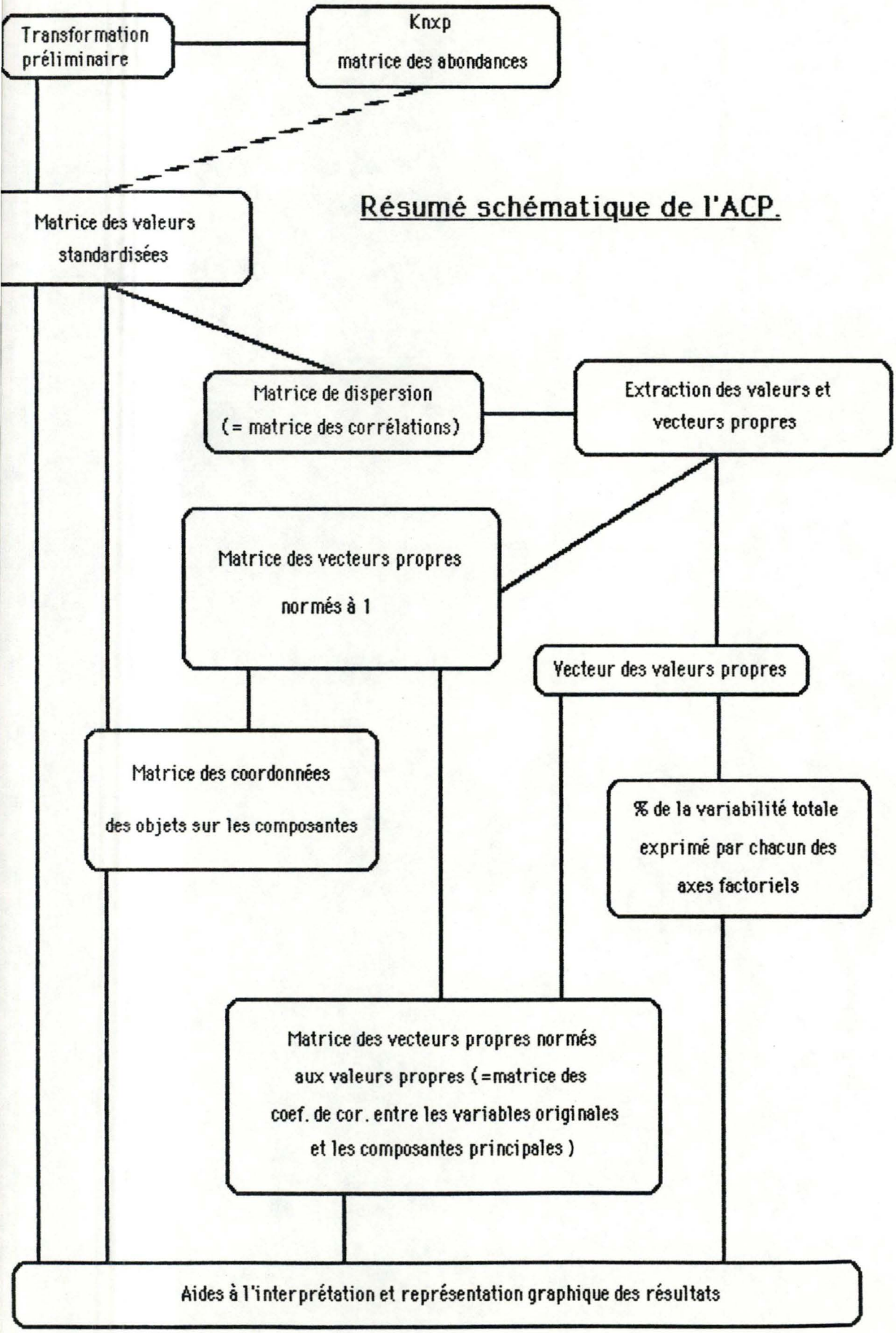
$$c = SQR(a\ /\ b)$$

est déterminée pour assurer un positionnement dans le plan compatible avec la répartition des points-variables, et permettre ainsi une représentation simultanée des deux nuages.

La superposition, avec des précautions d'interprétation, rend plus vivante la visualisation, en suggérant quelles variables sont responsables des proximités.

Ce type de représentation est rarement adopté, il ne le sera pas non plus dans le cadre de ce travail.

Résumé schématique de l'ACP.



Etapes de calcul différentes

valuation de la matrice de dispersion à partir de :

ACP : Z'Z

$$Zij = \frac{Kij - Mj}{Sj * SQR(N)}$$

AFC : X'X

$$Xij = \frac{Fij - (Fi. * F.j)}{SQR(Fi. * F.j)}$$

Extraction des valeurs et vecteurs propres de la matrice de dispersion

ACP

matrice des corrélations entre les descripteurs.

AFC

matrice des variances-covariances des descripteurs.

Calcul des coordonnées des descripteurs dans l'espace factoriel

ACP

Matrice des vecteurs propres normés aux valeurs propres.
Interprétation en termes de corrélations.

AFC

Coordonnées des points-colonnes dans l'espace compatible des points-lignes.

7. Types de données utilisées.

7.1. Remarque préliminaire.

Certaines définitions et termes utilisés sont uniquement valables dans le cadre de ce travail.

7.2. Caractéristiques générales des tableaux.

Les tableaux seront :
- complets,
- non structurés,
- à deux entrées fixes,
- à n lignes (objets),
- à p colonnes (descripteurs).

L'adjectif "complet" indique que chaque cellule du tableau doit contenir une donnée chiffrée.

Dans un tableau non structuré, on ne considère aucun regroupement possible de lignes ou de colonnes. Dans le cas présent, c'est à l'analyse de le proposer.

Un tableau peut être "à une entrée fixe" ou "à deux entrées fixes", selon que l'on fixe à priori une seule ou deux dimensions, une dimension étant le nombre de lignes ou de colonnes. Le choix de deux entrées fixes indique seulement le fait que l'on ne rajoutera pas d'éléments supplémentaires, ce qui est cependant possible pour les deux techniques d'analyse.

Les données utilisées devront constituer un tableau de contingence. On appelle tableau de contingence, un tableau qui donne la ventilation d'une population (ou d'une quantité) selon deux caractères qualitatifs que l'on croise. On reconnaît un tableau de contingence à ceci : en calculant des sommes en ligne ou en colonne, on obtient des quantités qui ont un sens. Sur un tel tableau, on peut aussi calculer des fréquences en divisant la valeur d'une case par la valeur de la somme du tableau (Volle, 1985).

7.3. Caractéristiques des lignes.

Ce seront des échantillons prélevés le long d'un gradient. Ce pourront être par exemple des stations le long d'une rivière, des morceaux d'une carotte de sondage, un relevé phytosociologique, ou tout autre type de données écologiques capables de constituer un tableau de contingence.

7.4. Caractéristiques des colonnes.

Ce seront des unités taxonomiques, de niveau variable dans la réalité. Pour ce travail, on considérera qu'elles ont toutes le même niveau : l'espèce, en tenant compte des variations possibles au sein de cette espèce. Ceci se justifie par le fait que l'étude d'un gradient est réalisée à travers les aptitudes des individus à s'adapter, on doit donc se situer au niveau taxonomique le plus bas possible.

Chaque cellule du tableau contiendra donc un nombre entier positif ou nul.

7.5. Caractéristiques inhérentes aux données écologiques.

1) Les données écologiques sont, comme toutes les données biologiques, entachées d'une grande variabilité. Cette variabilité peut être décomposée comme suit :

   a) une variabilité expliquée qui montre l'évolution du phénomène étudié au cours de mesures séparées dans le temps ou dans l'espace ;

   b) une variabilité inexpliquée ou résiduelle due à une multitude de facteurs expérimentaux et dont l'importance relative empêche parfois l'interprétation des résultats.

Pour limiter l'influence de la variabilité résiduelle, on est souvent amené à répéter la mesure, à condition que chacune n'altère pas la validité des autres. Ainsi, lors de l'étude de la qualité biologique de l'eau d'une rivière, on réalise plusieurs prélèvements d'organismes benthiques dans une même station, de façon à obtenir des abondances moyennes, en principe plus proches de la réalité.

Cette caractéristique pourra être utilisée pour vérifier les capacités de chacune des deux analyses à regrouper les échantillons provenant d'une station. Cependant, ce point de comparaison est surtout intéressant dans le cas de données réelles, caractérisées par une grande variabilité, même entre réplicats.

2) Le nombre de colonnes (descripteurs) est souvent très supérieur au nombre de lignes (objets). Cette caractéristique aura une influence nette pour l'AFC qui travaille en termes d'écarts entre profils.

3) Les poids des lignes ou des colonnes sont très variables. Cette caractéristique est très importante, car la première différence entre les deux analyses réside dans leurs façons de considérer les lignes et les colonnes.

4) Les données écologiques sont généralement constituées de plusieurs espèces pauvrement représentées ou même absentes, alors que seules quelques espèces sont très abondantes : une telle répartition ne correspond pas à une distribution normale, puisqu'elle est fortement dissymétrique. De telles données sont le plus souvent transformées au moyen du logarithme ou de la racine carrée. La racine carrée est la transformation la moins drastique qui permet de normaliser les données distribuées selon la loi de Poisson, alors que la transformation logarithmique est appliquée aux données qui s'écartent fortement de la distribution normale (Legendre & Legendre, 1979).
   Ce type de transformation des données visent à rendre leur distribution conforme à la loi normale et à ramener leur variance indépendante de la moyenne, en vue de stabiliser les variances pour rendre aléatoire leur distribution (Legendre & Legendre, 1979).
   La transformation utilisée sera :

$$\ln (x+1)$$

   le "+1" est justifié par la présence de valeurs nulles.

5) Les descripteurs peuvent être répartis en quatre groupes, en croisant les deux paramètres importants de leur distribution : la valence écologique et le nombre maximum d'individus observés.

a) Les espèces "ubiquistes", caractérisées par une valence écologique large, et par des nombres d'individus plus ou moins constants et importants. Ce sont les plus importantes pour la détection du gradient.

b) Les espèces "rares" ou bioindicateurs, associés à un type de milieu particulier et à valence écologique étroite, dont les niveaux d'abondances varient fortement en fonction d'un gradient. Le nombre maximum d'individus est souvent élevé.

c) Les espèces à valence écologique étroite et niveau d'abondance faible. Elles sont à rapprocher des bioindicateurs.

d) Les espèces "accidentelles", sans signification écologique, présentes en très faibles quantités, de zéro à quelques individus, dans tous les échantillons.

Remarques.

Ces distinctions ne peuvent souvent être faites qu'à partir du tableau lui-même. Ainsi, la valence écologique, estimée à partir du nombre de stations dans lesquelles l'espèce est présente, dépend du nombre de stations situées sur le tronçon prospecté. D'autre part, l'abondance maximum d'une espèce est à comparer avec les autres, de façon à déterminer des abondances relatives.

6) Les objets peuvent être répartis en quatre groupes suivant leurs poids et le nombre d'espèces présentes.

a) Les échantillons riches diversifiés, qui sont alors les témoins de conditions particulièrement favorables à toutes les espèces.

b) Les échantillons riches non diversifiés, qui sont le signe de conditions favorables à quelques espèces adaptées.

c) Les échantillons à profils moyens, sans significations particulières, mais qui peuvent indiquer le point moyen d'un gradient.

d) Les échantillons pauvres à tous points de vue, témoins de conditions défavorables permanentes, d'une forte pollution récente ou de l'inadaptation du mode de prélèvement.

7.6. Commentaire.

Ces caractéristiques sont les éléments de base pour la construction d'un tableau de valeurs simulées.

8. Choix méthodologiques particuliers.

a) La variance détectée par les deux analyses est assimilable à de l'information. Celle-ci peut être soit interprétable (variabilité liée au gradient), soit ininterprétable (variabilité résiduelle). Les importances respectives sont elles-mêmes variables dans le cas de données réelles. Cependant, les exemples utilisés au cours de cette étude présenteront toujours une proportion supérieure de variabilité interprétable.
Seul le plan formé par les deux premiers axes sera considéré au cours de ce travail. Ce choix est justifié, d'une part, par la portion habituellement importante de la variance totale (interprétable) exprimée par les facteurs 1 et 2, et, d'autre part, l'examen des axes suivants n'apporte que rarement des informations supplémentaires (Depiereux, 1982 ; Cornet, 1986 ; Dansart, 1986 ; Maquet, 1987).

b) Les coordonnées sont définies au signe près sur chaque axe. L'orientation des axes est arbitraire et dépend de l'algorithme de diagonalisation utilisé.

c) Les deux graphes (colonnes et lignes) sont superposables tels quels pour l'AFC, mais pas pour l'ACP. Ce choix de représentation des projections a été guidé par le fait que la plupart des logiciels considèrent la représentation simultanée pour l'AFC, mais pas pour l'ACP.

d) Le but des analyses est de condenser l'information présente à travers un certain nombre de variables dépendantes en un nombre plus faible de variables linéairement indépendantes.
Cependant, on peut observer dans le plan une structure curvilinéaire indiquant que le deuxième axe est une fonction quadratique du premier, ou vice-versa (effet Guttman ou structure en croissant). Pour pallier cet inconvénient, certains auteurs conseillent de considérer le plan formé des axes 1 et 3 ; dans ce cas, il faut ignorer des taxons influençant fortement le gradient, comme le montre habituellement les contributions absolues au facteur 2.

Cette structure représente le gradient que l'analyse a détecté à partir des variations d'abondance des espèces dans les stations.

Les stations sont reclassées le long de ce gradient en fonction de l'intensité de celui-ci à ces endroits.

Une méthode proposée et développée par Depiereux & al. (1982) pour l'ACP permet d'obtenir le reclassement des stations en utilisant cette structure particulière. On fait tourner un vecteur centré à l'origine des axes et les stations sont reclassées par leur ordre de rencontre avec ce vecteur, en commençant par l'une des extrémités de la courbe. D'autres méthodes ont été proposées pour réaliser un reclassement, en "redressant" cette structure, elles ne seront pas considérées dans ce travail.

9. Exemple numérique illustrant l'AFC.

Le tableau utilisé pour cet exemple provient de données expérimentales récoltées par Genin (1976) sur la Bième, ruisseau confluent avec la Sambre à hauteur de Tamines. Ces données ont l'avantage d'être particulièrement typiques du processus de récupération d'une rivière en aval d'une pollution organique.

9.1. Données initiales.

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15
S1	7	5	2	3	51	6	0	1845	8	0	0	0	172	40	0
	3	7	1	4	103	8	0	1214	7	0	0	1	126	54	0
	2	18	4	3	65	10	0	622	12	0	0	0	260	20	0
S2	0	114	0	1	0	0	1	19	1	0	0	0	0	417	0
	0	145	0	0	0	0	1	34	0	0	0	0	0	482	0
	0	150	0	1	0	0	1	16	1	0	0	0	0	508	0
S3	0	88	2	27	1	0	7	242	14	0	1	0	0	4	0
	0	110	2	10	1	1	4	297	28	0	0	0	0	3	0
	0	111	6	59	4	0	2	338	12	0	0	0	0	7	1
S4	0	45	6	7	1	1	9	559	207	0	7	3	51	2	0
	0	11	4	2	3	2	14	713	160	0	10	5	45	2	0
	0	80	2	3	3	0	11	918	116	0	15	2	39	5	0
S5	8	41	0	4	3	19	0	530	53	22	533	29	584	57	35
	19	70	0	3	7	20	0	744	144	24	606	18	710	75	38
	6	58	0	3	8	17	0	791	103	8	608	26	484	55	41

9.1.1. *Description du tableau.*

Cinq stations (objets) de prélèvements ont été choisies : la station 1 (S1) est en amont d'une source importante de pollution organique ; les stations 2 à 5 (S2 à S5) s'échelonnent en aval des rejets.

Dans chaque station, trois prélèvements ont été réalisés (réplicats) en hiver et au printemps suivant. Tous les organismes récoltés ont été identifiés (descripteurs) et dénombrés.

Le tableau de données analysé reprend les valeurs d'abondance des quinze taxons les mieux représentés dans les quinze prélèvements hivernaux.

9.1.2. *Transformation préliminaire.*

Les donnée sont transformées en ln (x+1), comme indiqué plus haut.

9.2. Résultats de l'analyse.

val. propres	%relatifs	%cumulés
0.2494	46.3928	46.3928
0.1335	24.8215	71.2143
0.1036	19.2712	90.4856
0.0345	6.4252	96.9108
0.0045	0.8420	97.7528
0.0038	0.7149	98.4677
0.0029	0.5346	99.0023
0.0020	0.3777	99.3801
0.0013	0.2481	99.6282
0.0010	0.1887	99.8169
0.0005	0.0930	99.9099
0.0004	0.0813	99.9912
0.0000	0.0087	100.0000
0.0000	0.0000	100.0000
0.0000	0.0000	100.0000
0.5376		

ure du chi2 = 244.6649 196 degrés de liberté

me des éléments = 455.0677

ls des lignes en %

e (1) =	6.7766
e (2) =	6.8861
e (3) =	6.8638
e (4) =	3.4842
e (5) =	3.3868
e (6) =	3.5516
e (7) =	4.8775
e (8) =	4.7581
e (9) =	5.4125
e (10) =	6.9713
e (11) =	6.8463
e (12) =	6.9633
e (13) =	10.7726
e (14) =	11.5349
e (15) =	10.9145

s des colonnes en %

nne (1) = 2.5718
nne (2) = 12.8034
nne (3) = 2.6805
nne (4) = 5.3924
nne (5) = 5.4738
nne (6) = 3.9459
nne (7) = 3.1561
nne (8) = 19.0616
nne (9) = 10.0335
nne (10) = 1.8792
nne (11) = 5.9428
nne (12) = 3.2108
nne (13) = 10.1421
nne (14) = 11.1399
nne (15) = 2.5662

ributions des lignes l'analyse, en %

e (1) = 5.7595
e (2) = 5.3102
e (3) = 5.3517
e (4) = 8.4408
e (5) = 10.0995
e (6) = 8.9802
e (7) = 5.9692
e (8) = 4.9007
e (9) = 6.7391
e (10) = 3.9923
e (11) = 4.3374
e (12) = 3.3321
e (13) = 9.7148
e (14) = 9.2687
e (15) = 7.8039

ributions absolues des lignes

e (1) =	0.0000	0.0205	0.2591
e (2) =	0.0000	0.0143	0.2439
e (3) =	0.0003	0.0384	0.2119
e (4) =	0.0924	0.1639	0.0004
e (5) =	0.0956	0.2119	0.0034
e (6) =	0.0961	0.1779	0.0004
e (7) =	0.0755	0.0234	0.0570
e (8) =	0.0660	0.0230	0.0248
e (9) =	0.0692	0.0262	0.0105
e (10) =	0.0064	0.0786	0.0559
e (11) =	0.0006	0.0844	0.0406
e (12) =	0.0077	0.0347	0.0494
e (13) =	0.1750	0.0408	0.0212
e (14) =	0.1711	0.0368	0.0079
e (15) =	0.1442	0.0252	0.0137

ributions relatives (cos2) des lignes

e (1) =	0.0001	0.0881	0.8669
e (2) =	0.0001	0.0668	0.8854
e (3) =	0.0023	0.1782	0.7631
e (4) =	0.5077	0.4821	0.0010
e (5) =	0.4391	0.5207	0.0065
e (6) =	0.4964	0.4919	0.0009
e (7) =	0.5866	0.0974	0.1840
e (8) =	0.6246	0.1165	0.0974
e (9) =	0.4763	0.0966	0.0299
e (10) =	0.0743	0.4890	0.2700
e (11) =	0.0069	0.4828	0.1802
e (12) =	0.1078	0.2582	0.2858
e (13) =	0.8358	0.1043	0.0420
e (14) =	0.8566	0.0986	0.0163
e (15) =	0.8571	0.0801	0.0339

ribution des colonnes l'analyse, en %

nne (1) = 4.5278
nne (2) = 8.6845
nne (3) = 4.8010
nne (4) = 5.7115
nne (5) = 6.7076
nne (6) = 3.6907
nne (7) = 7.7536
nne (8) = 2.1547
nne (9) = 3.1706
nne (10) = 7.2739
nne (11) = 9.5780
nne (12) = 4.6813
nne (13) = 6.9804
nne (14) = 16.0223
nne (15) = 8.2621

ributions absolues des colonnes

nne (1) =	0.0566	0.0027	0.0688
nne (2) =	0.1286	0.0810	0.0253
nne (3) =	0.0264	0.1252	0.0004
nne (4) =	0.0343	0.0459	0.0156
nne (5) =	0.0020	0.0619	0.2484
nne (6) =	0.0455	0.0017	0.0585
nne (7) =	0.0735	0.0378	0.1406
nne (8) =	0.0291	0.0201	0.0066
nne (9) =	0.0000	0.0727	0.0565
nne (10) =	0.1123	0.0444	0.0225
nne (11) =	0.1373	0.0065	0.1434
nne (12) =	0.0698	0.0007	0.0469
nne (13) =	0.0781	0.0332	0.0735
nne (14) =	0.0879	0.4223	0.0614
nne (15) =	0.1187	0.0440	0.0317

tributions relatives (cos2) des colonnes

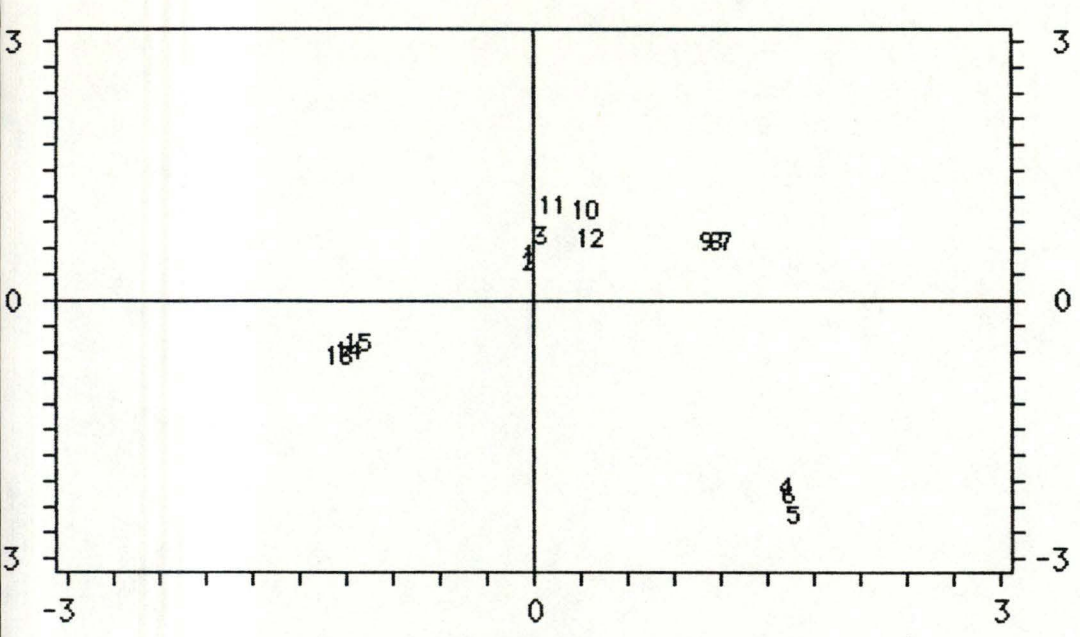
nne (1) =	0.5797	0.0148	0.2927
nne (2) =	0.6873	0.2315	0.0562
nne (3) =	0.2551	0.6474	0.0015
nne (4) =	0.2787	0.1994	0.0526
nne (5) =	0.0139	0.2289	0.7136
nne (6) =	0.5714	0.0112	0.3056
nne (7) =	0.4397	0.1210	0.3495
nne (8) =	0.6279	0.2321	0.0590
nne (9) =	0.0001	0.5687	0.3435
nne (10) =	0.7165	0.1516	0.0595
nne (11) =	0.6650	0.0168	0.2885
nne (12) =	0.6918	0.0037	0.1933
nne (13) =	0.5190	0.1181	0.2029
nne (14) =	0.2544	0.6542	0.0738
nne (15) =	0.6663	0.1323	0.0738

données des lignes sur les axes

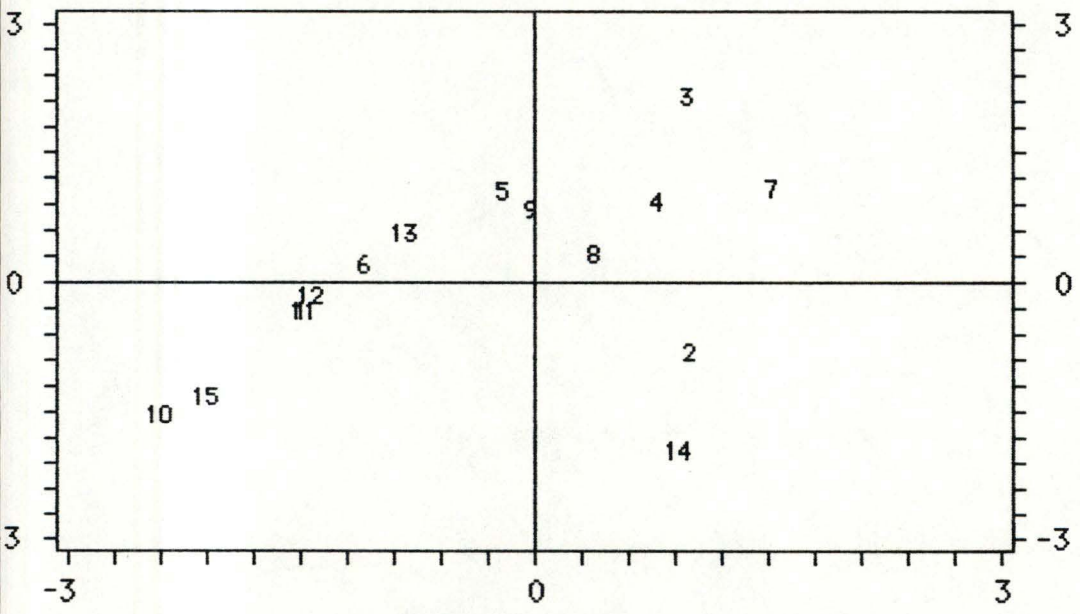
t 1 :	-0.0154	0.5494	-1.9552
t 2 :	-0.0128	0.4556	-1.8821
t 3 :	0.0619	0.7483	-1.7571
t 4 :	1.6281	-2.1690	-0.1120
t 5 :	1.6800	-2.5011	-0.3164
t 6 :	1.6448	-2.2384	-0.1089
t 7 :	1.2439	0.6929	1.0809
t 8 :	1.1775	0.6953	0.7214
t 9 :	1.1306	0.6960	0.4398
t 10 :	0.3029	1.0621	0.8956
t 11 :	0.0970	1.1101	0.7697
t 12 :	0.3334	0.7055	0.8423
t 13 :	-1.2745	-0.6156	0.4435
t 14 :	-1.2180	-0.5648	0.2609
t 15 :	-1.1493	-0.4804	0.3544

données des colonnes sur les axes

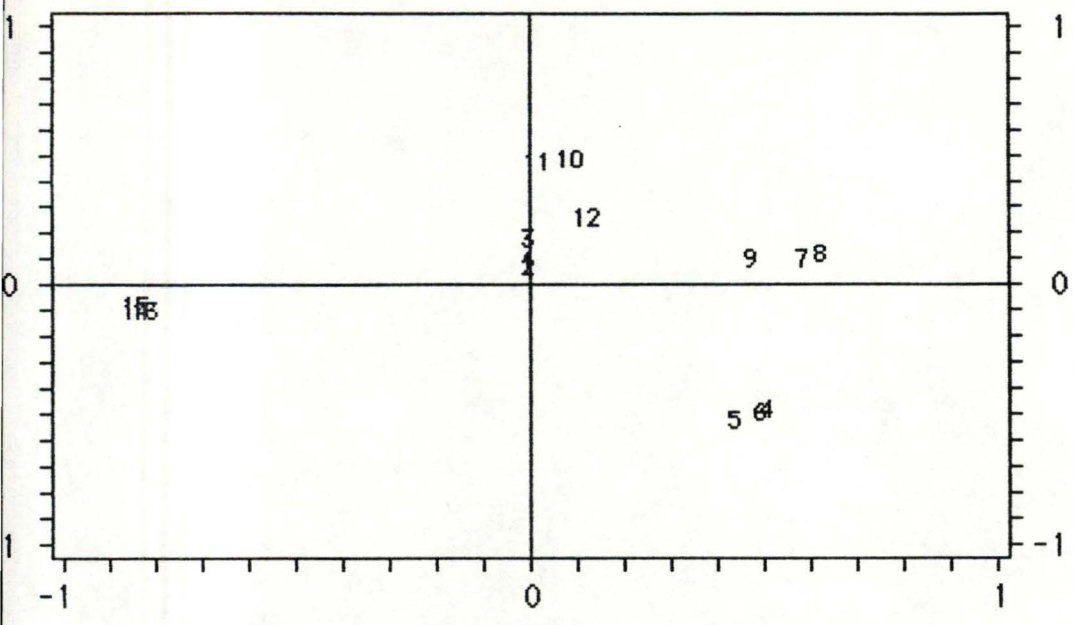
r. 1 :	-1.4829	-0.3236	-1.6351
r. 2 :	1.0023	-0.7952	0.4445
r. 3 :	0.9924	2.1614	0.1177
r. 4 :	0.7977	0.9225	0.5375
r. 5 :	-0.1914	1.0631	-2.1301
r. 6 :	-1.0733	0.2058	-1.2179
r. 7 :	1.5260	1.0942	2.1109
r. 8 :	0.3909	0.3250	-0.1859
r. 9 :	-0.0081	0.8509	0.7506
r. 10 :	-2.4449	-1.5377	1.0932
r. 11 :	-1.5200	-0.3299	1.5534
r. 12 :	-1.4742	-0.1476	1.2091
r. 13 :	-0.8774	0.5723	-0.8513
r. 14 :	0.8881	-1.9470	-0.7423
r. 15 :	-2.1503	-1.3099	1.1106



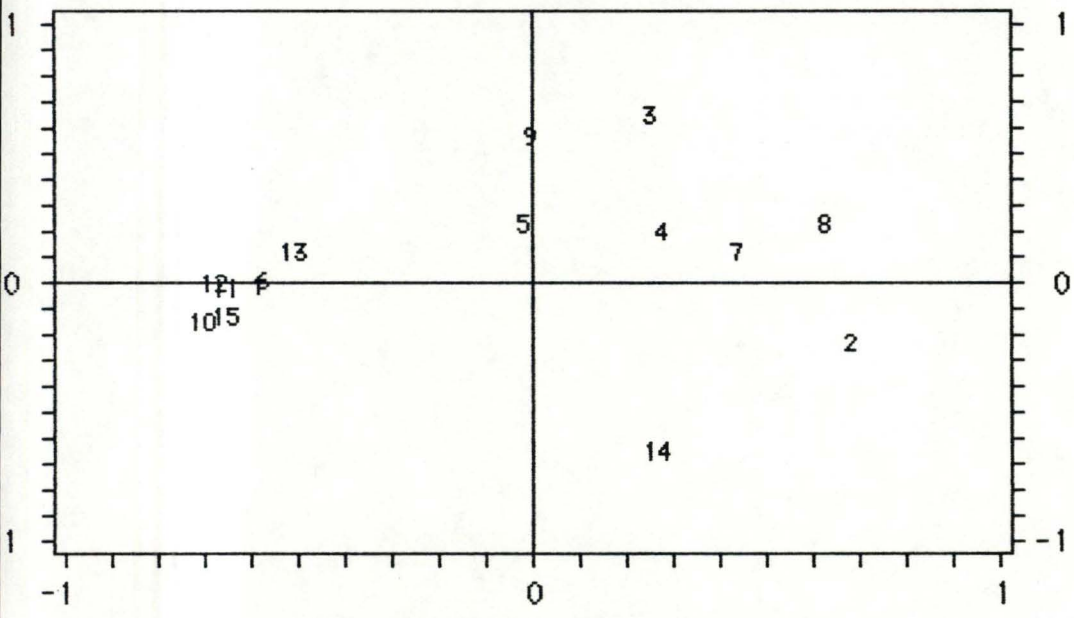
lignes dans le plan des axes factoriels 1 & 2



colonnes dans le plan des axes factoriels 1 & 2



cos2 des lignes dans le plan des axes factoriels 1 & 2



cos2 des colonnes dans le plan des axes factoriels 1 & 2

9.3. Interprétation des résultats chiffrés.

9.3.1. *Les axes factoriels et leurs valeurs propres associées.*

La trace (= somme des valeurs propres, 0 ≤ trace ≤ 1) est une indication importante, car elle permet, comme la mesure du χ^2, de relativiser les échelles des axes en termes d'écarts au profil moyen. En d'autres mots, elle est un signe de l'intensité de la dépendance existant entre les lignes ou les colonnes du tableau, dépendance sans laquelle l'analyse ne servirait à rien.

Dans le cas présent, les deux mesures montrent une forte dépendance.

Si on applique la règle empirique selon laquelle on ne puisse interpréter que les axes factoriels dont les valeurs propres sont supérieures à la moyenne de toutes les valeurs propres :

moyenne = trace / la plus petite dimension de la matrice originale moins 1
= 0.5376 / 14
= 0.0384,

seuls les trois premiers axes sont interprétables en termes écologiques.

Ibanez (1973) a fourni un cadre théorique à ce test empirique.

Les trois premiers axes représentent 90 % de l'information exprimée par l'ensemble des facteurs. Dans la pratique, on considère rarement plus de trois axes, ceux-ci fournissant habituellement toute l'information interprétable, alors que les suivants sont associés à la variabilité résiduelle ininterprétable. Cependant, il existe des cas où le contraire se produit, la variabilité résiduelle étant la plus importante : il faut alors pousser plus loin ses investigations.

9.3.2. *Les contributions des lignes.*

a) Contributions à l'analyse.

La contribution la plus importante est apportée par la ligne 5 qui possède aussi le poids le plus faible. Il est aussi intéressant de constater que les lignes à poids proportionellement plus élevés (13, 14, 15) ou faibles (4, 5, 6) ont des contributions plus élevées que les autres stations. On est donc à un point où les extrêmes se rejoignent dans une même action. Ceci est explicable par le fait que les lignes à poids faibles et élevés sont souvent celles qui s'écartent le plus du modèle d'indépendance.

b) Contributions absolues.

Les lignes 13, 14 et 15 contribuent le plus à la formation de l'axe 1. Ces lignes font partie de la même station de prélèvements.

La construction de l'axe 2 est surtout influencée par les lignes 4, 5 et 6. Ce groupe de lignes constitue, lui aussi, une même série de réplicats.

c) Contributions relatives.

Si on tient compte des erreurs d'approximations, des observations analogues aux contributions absolues peuvent être effectuées.

Commentaires.

L'AFC fait très bien la différence entre les différents groupes de réplicats en associant chacun d'eux à un axe factoriel selon les écarts au modèle d'indépendance exprimés par chaque groupe.

La concordance entre les contributions relatives et absolues indiquent une bonne conservation des distances dans le plan de projection formé des facteurs 1 et 2.

9.3.3. *Contributions des colonnes.*

a) Contributions à l'analyse.

Les colonnes 14, 11 et 2 apportent les contributions les plus importantes. Il s'agit de taxons montrant le plus fort taux de variations du nombre d'individus en fonction du gradient, ce sont donc les taxons les plus intéressants à considérer pour la détection de ce gradient. Leur courbe de valence écologique est caractérisée par un rapport hauteur sur base plus important que pour les autres taxons.

b) Contributions absolues.

Les colonnes 11, 2 et 15 détiennent les contributions les plus élevées à l'axe 1. On retrouve deux des trois taxons cités au paragraphe précédent, le gradient devrait donc être bien représenté par le premier axe. Les taxons 14, 3 et 2 influençant le plus le facteur 2 sont le signe du phénomène de dépendance existant entre les deux premiers facteurs. Il faudra donc tenir compte du deuxième axe pour effectuer le reclassement des taxons.

c) Contributions relatives.

Les variables 10, 12, 2 et 14, 3, 9 sont les plus proches, respectivement, de l'axe 1 et de l'axe 2.

Commentaires.

Si on poussait plus loin l'analyse des concordances entre les classements des variables par les deux séries de coefficients, on pourrait en déduire que la qualité de la représentation des projections des variables dans le premier plan factoriel est bonne. Ce qui doit se traduire par la présence d'une structure curvilinéaire, le long de laquelle le reclassement des points-variables pourra être fait.

9.4. Représentation graphique des projections.

9.4.1. *Projections des lignes.*

On peut déjà observer une déformation de la courbe par rapport à une structure quadratique théorique. Ces déformations qui sont propres à l'AFC constitueront le principal obstacle au reclassement et un point de comparaison avec l'ACP .

Remarque.

Le reclassement des réplicats le long du premier axe fournit un résultat identique.

9.4.2. *Projections des colonnes.*

La structure curvilinéaire est très nette sauf pour les espèces 3, 7, 9 et 8 qui présentent des distributions quelque peu inattendues en fonction du gradient. Ce qui a été déduit des valeurs des contributions se retrouve donc, un reclassement fiable des taxons peut être réalisé.

Commentaires.

Le graphe des lignes montre que l'AFC fait peu de différence entre les stations 1 et 4. De plus, les lignes 10, 11, 12 forment un groupe moins serré que les autres. Cela se vérifie à l'examen du graphe des colonnes où les taxons 3, 7, 9, ayant leurs maxima à la station 4, sont fortement éloignés les uns des autres. Le taxon 8, associé à la station 1, fausse aussi l'analyse. Il conviendrait donc de retirer une ou plusieurs de ces quatres colonnes du tableau et de recommencer l'analyse.

9.5. Conclusions.

L'AFC se révèle donc parfaitement apte à la détection d'un gradient. Cependant, la présence d'une variabilité résiduelle importante est néfaste, surtout pour le reclassement des taxons, ceux-ci étant souvent plus nombreux que les stations et plus difficiles à distinguer les uns des autres.

10. **Exemple numérique illustrant l'ACP.**

10.1. Description des données.

Le tableau utilisé pour cet exemple provient de données expérimentales récoltées par Genin (1976) : ce sont les mêmes que celles proposées pour l'AFC .

10.1.1. *Transformation préliminaire.*

Les donnée sont transformées en ln (x+1), comme indiqué plus haut.

10.2. Résultats de l'analyse.

val. propres	%relatifs	%cumulés
6.9791	46.5273	46.5273
3.8934	25.9557	72.4830
2.5860	17.2397	89.7228
0.9857	6.5715	96.2943
0.1611	1.0738	97.3681
0.1409	0.9391	98.3072
0.0929	0.6195	98.9267
0.0736	0.4907	99.4175
0.0347	0.2315	99.6489
0.0237	0.1578	99.8067
0.0163	0.1088	99.9156
0.0101	0.0672	99.9827
0.0026	0.0173	100.0000
0.0000	0.0000	100.0000
0.0000	0.0000	100.0000
15.0000		

me des éléments = 455.0677

s des lignes en %

e (1) =	6.7766
e (2) =	6.8861
e (3) =	6.8638
e (4) =	3.4842
e (5) =	3.3868
e (6) =	3.5516
e (7) =	4.8775
e (8) =	4.7581
e (9) =	5.4125
e (10) =	6.9713
e (11) =	6.8463
e (12) =	6.9633
e (13) =	10.7726
e (14) =	11.5349
e (15) =	10.9145

s des colonnes en %

nne (1) = 2.5718
nne (2) = 12.8034
nne (3) = 2.6805
nne (4) = 5.3924
nne (5) = 5.4738
nne (6) = 3.9459
nne (7) = 3.1561
nne (8) = 19.0616
nne (9) = 10.0335
nne (10) = 1.8792
nne (11) = 5.9428
nne (12) = 3.2108
nne (13) = 10.1421
nne (14) = 11.1399
nne (15) = 2.5662

ributions des lignes l'analyse, en %

e (1) = 6.1141
e (2) = 5.3711
e (3) = 4.4203
e (4) = 7.4648
e (5) = 8.5560
e (6) = 8.0498
e (7) = 4.6573
e (8) = 3.2566
e (9) = 6.2709
e (10) = 4.4611
e (11) = 5.1532
e (12) = 3.3721
e (13) = 10.7342
e (14) = 12.5574
e (15) = 9.5612

tributions absolues des lignes

e (1) =	0.0105	0.0221	0.2776
e (2) =	0.0105	0.0107	0.2519
e (3) =	0.0029	0.0265	0.1776
e (4) =	0.0688	0.1601	0.0010
e (5) =	0.0687	0.1913	0.0042
e (6) =	0.0721	0.1777	0.0003
e (7) =	0.0558	0.0220	0.0374
e (8) =	0.0414	0.0147	0.0223
e (9) =	0.0461	0.0426	0.0120
e (10) =	0.0049	0.1041	0.0591
e (11) =	0.0003	0.1076	0.0155
e (12) =	0.0060	0.0411	0.0510
e (13) =	0.1950	0.0314	0.0344
e (14) =	0.2364	0.0296	0.0287
e (15) =	0.1807	0.0185	0.0268

ributions relatives (cos2) des lignes

e (1) =	0.0795	0.0938	0.7827
e (2) =	0.0908	0.0517	0.8086
e (3) =	0.0301	0.1558	0.6926
e (4) =	0.4287	0.5568	0.0024
e (5) =	0.3738	0.5803	0.0085
e (6) =	0.4167	0.5729	0.0006
e (7) =	0.5572	0.1224	0.1386
e (8) =	0.5919	0.1169	0.1181
e (9) =	0.3419	0.1765	0.0330
e (10) =	0.0515	0.6054	0.2285
e (11) =	0.0023	0.5421	0.0520
e (12) =	0.0829	0.3166	0.2609
e (13) =	0.8453	0.0759	0.0553
e (14) =	0.8758	0.0613	0.0394
e (15) =	0.8795	0.0501	0.0483

tribution des colonnes l'analyse, en %

nne (1) = 6.6667
nne (2) = 6.6667
nne (3) = 6.6667
nne (4) = 6.6667
nne (5) = 6.6667
nne (6) = 6.6667
nne (7) = 6.6667
nne (8) = 6.6667
nne (9) = 6.6667
nne (10) = 6.6667
nne (11) = 6.6667
nne (12) = 6.6667
nne (13) = 6.6667
nne (14) = 6.6667
nne (15) = 6.6667

tributions absolues des colonnes

nne (1) =	0.1211	0.0047	0.0253
nne (2) =	0.0222	0.0665	0.1659
nne (3) =	0.0136	0.2089	0.0004
nne (4) =	0.0019	0.0888	0.0268
nne (5) =	0.0356	0.0430	0.2052
nne (6) =	0.1269	0.0002	0.0251
nne (7) =	0.0393	0.0697	0.1214
nne (8) =	0.0544	0.1407	0.0079
nne (9) =	0.0455	0.0921	0.1023
nne (10) =	0.1073	0.0253	0.0424
nne (11) =	0.1020	0.0026	0.1005
nne (12) =	0.1064	0.0006	0.0741
nne (13) =	0.1190	0.0136	0.0135
nne (14) =	0.0004	0.2217	0.0417
nne (15) =	0.1046	0.0216	0.0476

ributions relatives (cos2) des colonnes

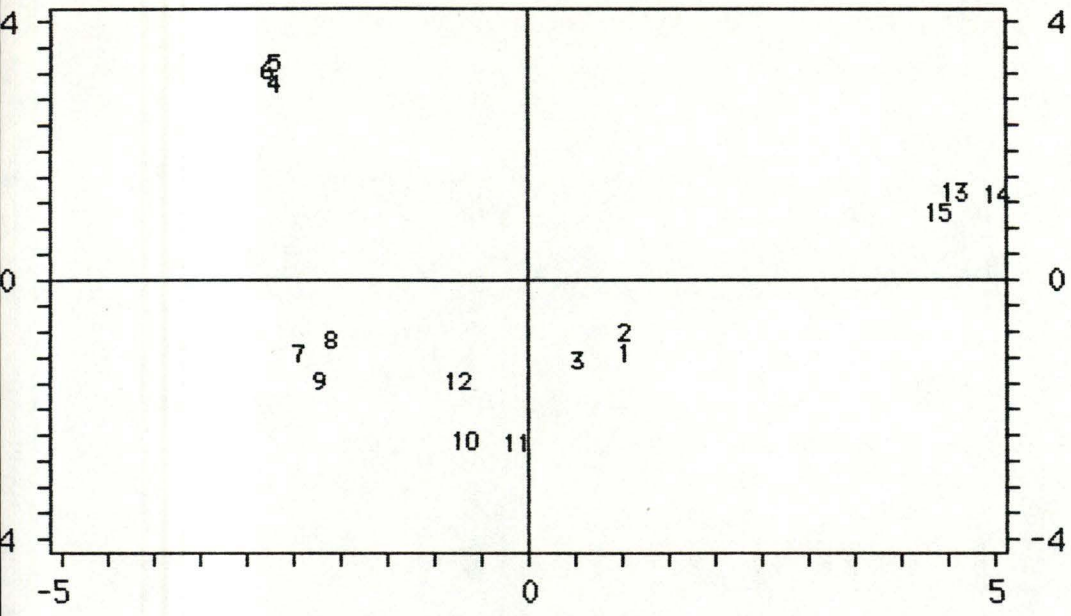
nne (1) =	0.8453	0.0182	0.0655
nne (2) =	0.1546	0.2591	0.4289
nne (3) =	0.0949	0.8133	0.0011
nne (4) =	0.0130	0.3458	0.0692
nne (5) =	0.2482	0.1675	0.5306
nne (6) =	0.8855	0.0008	0.0649
nne (7) =	0.2741	0.2715	0.3140
nne (8) =	0.3797	0.5479	0.0204
nne (9) =	0.3178	0.3584	0.2646
nne (10) =	0.7487	0.0984	0.1097
nne (11) =	0.7116	0.0102	0.2598
nne (12) =	0.7427	0.0022	0.1915
nne (13) =	0.8307	0.0528	0.0349
nne (14) =	0.0025	0.8632	0.1077
nne (15) =	0.7298	0.0840	0.1231

rdonnées des lignes sur les axes

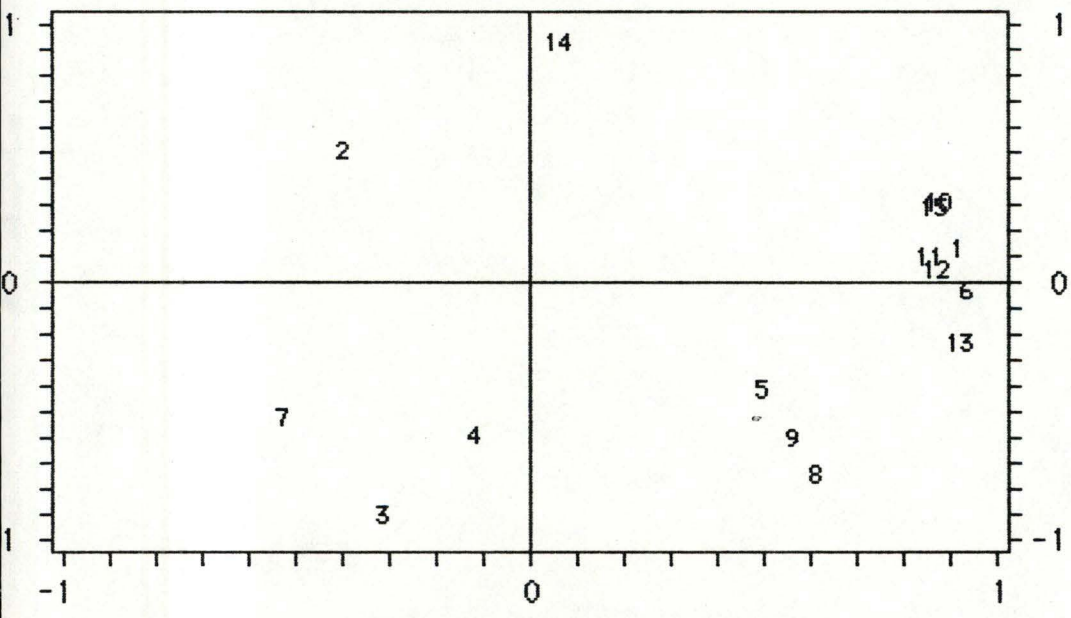
et 1 :	1.0461	-1.1362	-3.2813
et 2 :	1.0476	-0.7903	-3.1260
et 3 :	0.5474	-1.2447	-2.6246
et 4 :	-2.6832	3.0580	-0.2015
et 5 :	-2.6825	3.3424	-0.4052
et 6 :	-2.7471	3.2213	-0.1064
et 7 :	-2.4164	-1.1326	1.2049
et 8 :	-2.0826	-0.9255	0.9304
et 9 :	-2.1964	-1.5781	0.6822
et 10 :	-0.7189	-2.4651	1.5144
et 11 :	-0.1631	-2.5072	0.7765
et 12 :	-0.7932	-1.5499	1.4069
et 13 :	4.5183	1.3537	1.1560
et 14 :	4.9744	1.3156	1.0550
et 15 :	4.3497	1.0385	1.0189

rdonnées des colonnes sur les axes

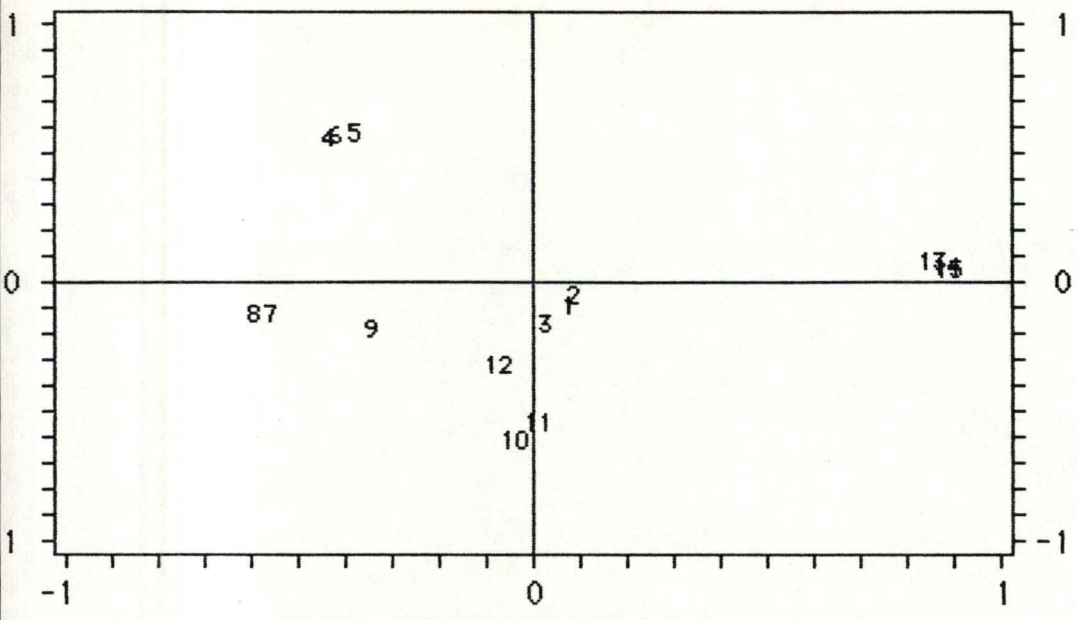
cr. 1 :	0.9194	0.1349	-0.2559
cr. 2 :	-0.3932	0.5090	0.6549
cr. 3 :	-0.3080	-0.9019	-0.0327
cr. 4 :	-0.1141	-0.5880	0.2631
cr. 5 :	0.4982	-0.4093	-0.7284
cr. 6 :	0.9410	-0.0274	-0.2547
cr. 7 :	-0.5235	-0.5211	0.5603
cr. 8 :	0.6162	-0.7402	-0.1428
cr. 9 :	0.5637	-0.5987	0.5144
cr. 10 :	0.8653	0.3137	0.3312
cr. 11 :	0.8436	0.1010	0.5097
cr. 12 :	0.8618	0.0468	0.4376
cr. 13 :	0.9114	-0.2298	-0.1869
cr. 14 :	0.0501	0.9291	-0.3282
cr. 15 :	0.8543	0.2899	0.3509



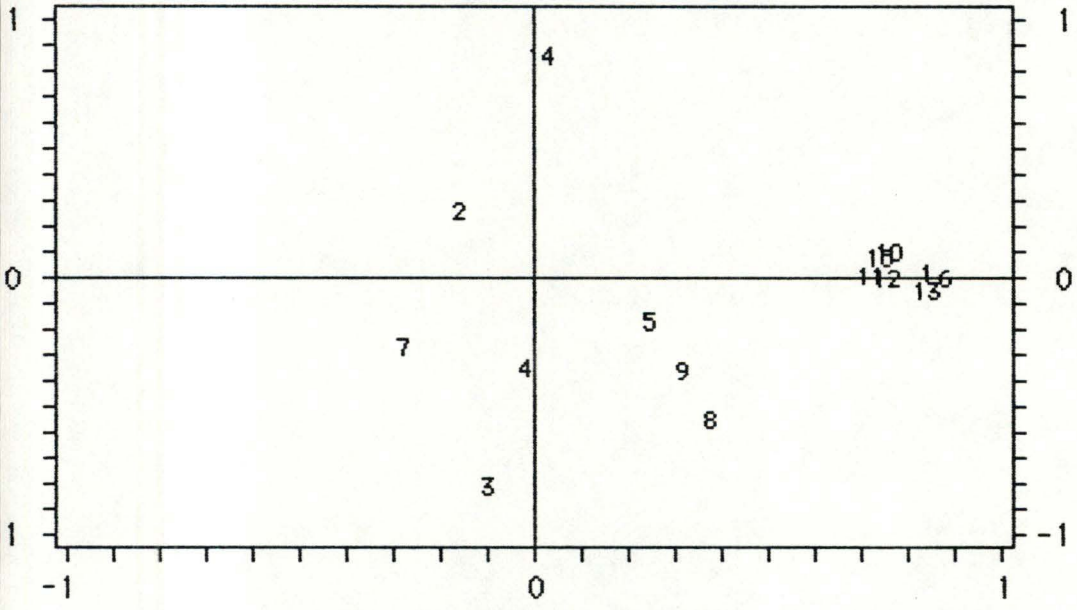
lignes dans le plan des composantes 1 & 2



colonnes dans le plan des composantes 1 & 2



cos2 des lignes dans le plan des axes factoriels 1 & 2



cos2 des colonnes dans le plan des axes factoriels 1 & 2

10.3. Interprétation des résultats chiffrés.

10.3.1. *Les composantes principales et leurs valeurs propres asssociées.*

Les composantes principales successives correspondent à des fractions de plus en plus faibles de la variance totale : le problème se pose donc de déterminer combien de composantes seraient éventuellement interprétables en termes écologiques.

Un test est fourni par une règle empirique, selon laquelle on ne puisse interpréter que les composantes principales correspondant aux valeurs propres λ plus grandes que la moyenne des λ. Dans le cas particulier de données centrées-réduites, la moyenne des λ est 1 et on ne pourrait interpréter les composantes dont les λ sont plus petits ou égaux à 1. Ibanez (1973) a d'ailleurs fourni un cadre théorique à ce test empirique.

Dans le cas présent, l'espace réduit est constitué de 13 dimensions (axes) dont trois sont associés à une valeur propre supérieure à 1. Ces trois premiers axes représentent 90 % de la variance totale. Ce chiffre nous donne une première indication quant à la qualité de la représentation des points dans l'espace factoriel. Au plus ce pourcentage est élevé, meilleure sera la représentation des projections des points dans les plans formés par ces axes, car les distances entre ces points seront mieux conservées. En pratique, on considère trois axes au maximum, les suivants n'exprimant, en général, qu'une variabilité résiduelle. Le contraire peut cependant se produire. Auquel cas, l'expérimentateur devra pousser plus loin ses investigations.

10.3.2. *Les contributions des lignes.*

a) Contributions à l'analyse.

Les lignes 8, 12 et 3 possèdent les contributions les plus faibles, elles se situent au milieu du gradient, et les lignes 13, 14 et 15 les plus élevées, elles se situent à l'une des extrémités de ce gradient.

b) Contributions absolues.

Elles montrent la part prise par une ligne dans la variance exprimée par un axe.

Pour l'axe 1 : ce sont les lignes 13, 14 et 15 qui prennent le plus part à sa construction et les coordonnées de ces lignes sur l'axe 1 sont aussi les plus importantes en valeurs absolues. Les valeurs les plus faibles sont détenues par les lignes 11, 3 et 10 qui possèdent les coordonnées les plus proches de zéro sur cet axe.

Pour l'axe 2 : les observations sont identiques avec respectivement les lignes 4, 5 et 6 pour les contributions les plus élevées et les lignes 2, 8 et 15 pour les contributions les plus faibles.

c) Contributions relatives.

Elles montrent la part prise par chaque axe dans la position d'un élément.

Pour l'axe 1 : les ligne 13, 14, 15 et les lignes 11, 3, 10 sont respectivement plus et moins influencées.

Pour l'axe 2 : il s'agit des lignes 10, 5, 4 et des lignes 13, 14, 15.

Commentaires.

Hormis les inévitables erreurs d'approximation à partir de la troisième décimale, on peut dire, pour les contributions absolues :

a) les contributions absolues ne sont utiles que pour fournir un résultat chiffré, sinon l'examen du graphe des projections est plus rapide,
b) les points situés aux extrémités du gradient contribuent le plus à la construction des axes impliqués dans la représentation de cette structure, chaque extrémité étant préférentiellement associée à l'un des deux premiers axes (ce n'est pas une règle générale).

pour les contributions relatives :

a) elles permettent de relativiser les proximités entre points observées dans un plan par rapport à leurs proximités réelles dans l'espace factoriel.
b) les points situés aux extrémités du gradient sont le plus pris en compte par les nouveaux axes (ils sont plus proches).

10.3.3. *Contributions des colonnes.*

a) Contributions à l'analyse.

Toutes les contributions sont identiques. C'est un effet de la standardisation préliminaire à l'analyse qui a pour conséquence d'égaliser toutes les variances des colonnes. L'ACP se rapprochant d'une analyse de la variance, toutes les colonnes contribuent de la même façon à l'analyse.

Cela pourrait constituer un élément défavorable : les espèces ayant des caractéristiques différentes, elles ne devraient pas apporter des contributions identiques à l'analyse.

b) Contributions absolues.

Les observations sont parallèles au graphe des projections et ne donnent qu'un résultat chiffré, comme pour les lignes. Pour chaque axe, elles sont proportionnelles aux coordonnées.

Pour le facteur 1, ce sont les colonnes 6, 1, 13 et pour le facteur 2, les colonnes 14, 3 et 8 qui apportent les contributions les plus élevées.

c) Contributions relatives.

Elles font surtout double emploi avec les coordonnées des colonnes puisqu'elles ne sont que les carrés de ces dernières. Mais elles sont intéressantes car elles permettent une interprétation plus fine du graphe des projections en accentuant les écarts relatifs entre celles-ci.

Les cosinus carrés les plus élevés pour l'axe 1 sont réalisés avec les points-variables 6, 1 et 13 ; et pour l'axe 2, il s'agit des points 14, 3 et 8. Ce classement est identique à celui obtenu pour les contributions absolues.

Commentaires.

Les contributions des colonnes font encore plus double emploi avec le graphe des projections que les contributions des lignes, ce qui était prévisible au vu des algorithmes de calcul.

Leur intérêt réside dans les nuances qu'elles peuvent apporter pour l'interprétation des graphes.

10.4. Représentation graphique des projections.

10.4.1. *Projections des lignes.*

On peut déjà observer une déformation de la courbe par rapport à une structure quadratique théorique. Ces déformations qui sont propres à l'ACP constitueront le principal obstacle au reclassement et un point de comparaison avec l'AFC .

Remarque.

Le reclassement des réplicats le long du premier axe ne fournit pas un résultat identique.

10.4.2. *Projections des colonnes.*

Les coordonnées des points-variables sont inférieures à 1 en valeurs absolues ; en effet, ces points sont à la distance 1 de l'origine dans R^n, et l'opération de projection est "contractante" : elle ne peut que diminuer les distances.

Un point de la sphère de rayon unité dans R^n représente une variable (dont les n composantes sont normalisées). La valeur du produit scalaire des vecteurs joignant l'origine à deux points de cette sphère n'est autre que le coefficient de corrélation des variables correspondantes (c'est également le cosinus de l'angle des deux vecteurs). Ainsi, les coordonnées des points-variables sur le premier axe ne sont autres que les coefficients de corrélation entre les diverses variables et le premier facteur, considéré lui-même comme une variable artificielle (c'est une combinaison linéaire des variables initiales). On pourra chercher à interpréter chaque axe en fonction de l'association des variables les plus corrélées au facteur correspondant.

Les coefficients de corrélation les plus élevés en valeurs absolues avec le premier axe sont observés par ordre décroissant pour les variables 6, 1, 13, 10, 12, 15 et 11. Les espèces concernées sont les plus sensibles au gradient.

L'association du deuxième facteur avec les espèces 14, 3 et 8 à caractéristiques très différentes, montre que le gradient est moins bien représenté par la deuxième composante : en effet, le reclassement des stations obtenu le long de cet axe revêt peu de significations, alors que le reclassement effectué le long du premier est très proche de la réalité.

Le premier axe devrait donc représenter ce gradient, mais le phénomène de dépendance entre les facteurs 1 et 2 oblige à tenir compte du deuxième axe pour le reclassement.

La structure curvilinéaire est assez lâche, cela peut être dû à une variabilité résiduelle importante. Le reclassement est perturbé par les taxons 3, 7 et 9, dont il faudrait reconsidérer les présences dans le tableau initial.

Commentaires.

Les stations 2 et 5 représentant les extrémités du gradient sont bien différenciées, de même que les taxons ayant leurs maxima à ces endroits. Le graphe des colonnes montre une sensibilité plus grande de l'analyse vis-à-vis de la variabilité résiduelle tant pour le reclassement des taxons que pour celui des stations.

10.5. Conclusion.

La puissance du modèle des composantes principales a permis d'appliquer cette technique d'analyse à une grande variété de tableaux. Son emploi en écologie est très répandu. Cependant, l'ACP peut perdre de sa puissance pour des utilisations particulières, le choix d'une technique plus appropriée est alors justifié.

11. Comparaison des résultats concernant l'exemple.

11.1. Résultats comparables.

1) Les valeurs propres et la trace,

2) les contributions à l'analyse,

3) les contributions absolues,

4) les contributions relatives et les graphiques associés (cos^2),

5) les graphiques des projections et les reclassements effectués.

11.2. Les valeurs propres et la trace.

Les nombres de valeurs propres non nulles obtenus par les deux analyses sont identiques, les deux espaces possèdent donc les mêmes dimensions.

La corrélation entre les pourcentages relatifs des valeurs propres est remarquable (0.92), mais ne signifie pas encore que les que les deux méthodes prennent en compte la même information.

La trace n'a pas de signification particulière pour l'ACP, tandis qu'elle représente une mesure des écarts au modèle d'indépendance pour l'AFC. Une valeur de 0.5376 montre une dépendance réelle, cependant la variabilité résiduelle intervient aussi dans le calcul.

11.3. Les contributions.

Les contributions les plus importantes sont données pour les deux analyses par ordre décroissant.

11.3.1. *Contributions à l'analyse.*

	ACP	AFC
lignes :	14, 13, 15, 5, 6, 4	5, 13, 14, 6, 4, 15
colonnes :	-	14, 11, 2, 15, 7, 10

11.3.2. *Contributions absolues.*

	ACP		AFC	
	F 1	F 2	F 1	F 2
lignes :	14, 13, 15	5, 6, 4	13, 14, 15	5, 6, 4
colonnes :	6, 1, 13	14, 3, 8	11, 2, 15	14, 3, 2

11.3.3. *Contributions relatives.*

	ACP		AFC	
	F 1	F 2	F 1	F 2
lignes :	15, 14, 13	10, 5, 6	15, 14, 13	5, 6, 10
colonnes :	6, 1, 13	14, 3, 8	10, 12, 2	14, 3, 9

Commentaire.

Comme on pouvait s'y attendre, les principales différences sont trouvées pour les contributions des colonnes. Ce sont donc ces dernières qui sont les plus importantes dans la comparaison.

Les contributions à l'analyse indiquent quels taxons doivent être associés aux deux premiers axes. Pour l'ACP aucune indication n'est fournie puisque toutes les contributions sont identiques. L'AFC, par contre, respecte bien cette logique, à l'exception du taxon 3 qui constitue un élément perturbateur comme indiqué lors des commentaires concernant l'AFC.

Il a été montré que le premier axe est le plus important pour réaliser un reclassement. Or les trois taxons pris en compte pour le premier facteur sont différents d'une analyse à l'autre : le deuxième axe ayant, ou devant avoir, un intérêt moindre. A priori, l'AFC semble donc plus indiquée pour cet exemple.

Cependant, il ne faut pas oublier qu'il s'agit d'un cas type, dans lequel presque tous les taxons sont intéressants à considérer.

11.4. Les graphiques des projections.

Le tableau initial indique à quelle station, chaque taxon présente son maximum d'abondance :

station		taxon(s)
1	--	5, 8
2	--	2, 14
3	--	4
4	--	3, 7, 9
5	--	1, 6, 10, 11, 12, 13, 15

11.4.1. *Les projections des lignes.*

le reclassement par la méthode du vecteur tournant donne des résultats identiques, qui sont aussi ceux obtenus le long du premier axe pour l'AFC, l'ACP présentant quant à elle de légères différences. La dépendance entre les deux axes serait donc plus marquée pour l'ACP.
L'ordre est le suivant : S2, S3, S4, S1, S5, la pollution diminuant.

11.4.2. *Les projections des colonnes.*

La structure curvilinéaire apparait plus nettement avec l'AFC.

Les taxons associés aux stations situées aux extrémités du gradient (S2 et S5) sont bien différenciés par les deux méthodes. Les deux analyses reconnaissent les taxons 3, 7, 9 comme éléments perturbants, l'AFC y ajoute le 8, mais ce n'est pas le plus important. Dans les deux cas, il conviendrait donc de reconsidérer les présences dans le tableau initial des taxons ayant leurs maxima à la station 4, ou la station elle-même.

11.5. Conclusion.

L'AFC gagne au points cette fois-ci, mais tout n'est pas perdu pour l'ACP.

12. Utilisation de données simulées.

L'utilisation de données simulées permet de contrôler les paramètres des distributions des espèces, et, ainsi, d'obtenir facilement les effets désirés. Par contre, cela présente le désavantage "d'oublier" la variabilité aléatoire, caractéristique fondamentale des données biologiques, qui ne peut pas être imitée pour de très nombreuses raisons.

Cependant, l'emploi de nombres artificiels est rendu nécessaire pour multiplier les données et pour permettre le contrôle complet des analyses.

Les données utilisées seront synthétisées en utilisant le modèle mathématique proposé par Austin & Noy-Meir (1971) et adapté. Ce modèle permet d'imiter la courbe de la distribution d'une espèce le long d'un gradient et d'en contrôler les paramètres.

12.1. Méthode.

La formule utilisée est la suivante :

$$x = x_{max} * \exp((-d^2 * 2) / k)$$

- x = nombre d'individus à une distance d du point exprimant le x_{max} ; x sera entier, positif ou nul.
- x_{max} = nombre maximum d'individus observés ; x_{max} sera entier, positif ou nul.
- d = distance à l'optimum, en nombre de stations, d sera entier, positif ou nul.
- k = coefficient d'étalement de la courbe en nombre de stations, k sera entier positif ou nul. Si k est nul, il définira, par convention, une abondance constante égale à x_{max}, dans toutes les stations. k indique le nombre de stations, à une ou deux unités près, dans lesquelles l'espèce sera présente.

Les paramètres définis par l'expérimentateur sont donc la station où on observe x_{max}, x_{max} et k.

Une légère variabilité peut être introduite pour éviter des problèmes de calcul pouvant se présenter avec des distribution trop régulières.

13. **Apports préliminaires extérieurs.**

Il s'agit d'un recueil d'extraits ou de conclusions tirés d'articles ou d'ouvrages dont on pourra trouver les références dans la bibliographie.

Pour le groupement des espèces en associations, l'AFC ne semble pas échapper aux problèmes que rencontre l'ACP (Reyssac & Roux, 1972 ; Ibanez & Séguin, 1972 ; Binet & al., 1972).

Dans le cas de gradients prononcés, l'AFC serait préférables aux composantes principales (Hill, 1973), ce qui a aussi été démontré par Gauch & al. (1977) à l'aide de données floristiques simulées et expérimentales.

In most practical applications, reciprocal averaging gives stand ordinations which are similar to those derived by principal component analysis of standardized data. As a general method for use in phytosociological contexts it is preferable because it generates good simultaneous species ordinations. The rationale of the method is close to that of gradient analysis, so that it is more suitable than principal component analysis for displaying strong floristic gradients (Hill, 1973).

Chardy & al. (1976) conseillent pour leur part d'effectuer simultanément plusieurs types d'ordination en espace réduit en vue de dégager des structures d'intérêts complémentaires.

The use of nonmetric multidimensional scaling (une autre technique) as an ordination method has been studied by the use of simulated coenoclines and coenoplanes. It was found that the method always produced better ordinations than principal components analysis and in most cases better than reciprocal averaging (Fasham, 1977). Ce pourrait être une solution au dilemne posé par le choix à faire entre l'ACP et l'AFC.

Reciprocal averaging (R.A.) and P.O. (polar or Bray-Curtis method, une autre méthode) are little subject to the involution of axis extremes that affects non-standardised principal component analysis (P.C.A.). Despite the arch effect, R.A. is superior to P.C.A. at high beta diversities (variabilité entre les échantillons) and on the whole preferable to P.C.A. at low beta diversities. Second and higher axes of P.C.A. and R.A. may express ecologically meaningless, curvilinear functions of lower axes. When curvilinear displacements are combined with samples errors, axis interpretation is difficult. None of the techniques solve all the problems for ordination that result from the curvilinear relationships characteristic of community data. For applied organisation research consideration of sample set properties, careful use of supporting information to evaluate axes, , and comparison of results of R.A or P.C.A. with P.O. and direct ordinations (exemple : le cluster analysis) are suggested (Gauch & al, 1977).

L'AFC produit des ordinations des échantillons très semblables à celles que l'on peut obtenir par l'ACP de matrices de corrélations entre espèces (données centrées réduites), ce qui est normal puisque la première étape du calcul consiste justement à pondérer chaque donnée par la somme (ou la probabilité) de la ligne et de la colonne, ce qui élimine les effets dûs à la plus grande variance de certaines lignes ou colonnes (Legendre & Legendre, 1979).

L'AFC est une ACP qui préserve dans l'espace factoriel la distance distributionnelle entre les lignes et les colonnes du tableau de contingence (Lebart & al., 1979).

... it will be demonstrated that correspondence analysis has large advantages over principal component analysis, especially for highly skewed distributed variables ... correspondence analysis is superior, especially for variables which have distributions which exhibit large deviations from normality in which case principal component analysis can only be used after a non linear transformation of the variables to a normal distribution, which is not always possible (Eppink, 1984).

Lorsqu'on a affaire à un tableau de contingence, on préférera en général utiliser l'AFC : cette méthode convient particulièrement aux tableaux de contingence, en raison notamment du rôle logiquement symétrique qu'elle fait jouer aux lignes et aux colonnes du tableau (Volle, 1985).

14. Comparaisons de l'ACP et de l'AFC.

14.1. Remarque préliminaire.

Les tableaux utilisés ont presque tous les mêmes dimensions, c'est-à-dire dix lignes et dix colonnes. Ce choix est justifié par les raisons suivantes :

1) le hardware et le software employés ne permettent pas une vitesse d'exécution élevée, celle-ci étant inversement proportionnelle au nombre de colonnes ;

2) l'augmentation du nombre des dimensions de la matrice originale possède un effet "noyant" pour les particularités que le tableau peut avoir ;

3) la clarté des représentations graphiques des résultats nécessite un nombre limité de points ;

4) ces dimensions sont suffisantes pour observer les conséquences de diverses manipulations du tableau de départ.

Il s'agit donc d'un compromis entre les dimensions souvent grandes des tableaux de données réelles et les contraintes de temps et de clarté.

5) Les résultats consignés dans ce chapitre ont été choisis pour leur caractère illustratif. Cependant, il ne faut pas oublier qu'il s'agit de données simulées qui, comme telles, s'éloignent plus ou moins de la réalité. De plus, leur nombre théorique étant infini, cela implique que tous les cas ne peuvent être considérés. Elles restent, néanmoins, le meilleur choix pour ce type d'étude, comme cela a déjà été expliqué.

14.2. Références.

La matrice, dix lignes et dix colonnes, servira de point de départ pour étudier les effets de diverses transformations appliquées à ce tableau.

Les dix colonnes représentent les distributions de dix espèces dans dix stations. Les abondances sont simulées sans variabilité et les abréviations utilisées pour décrir les différents paramètres de chacune des distributions sont :

espèces : Esp.,
abondance maximum de l'espèce : A. M.,
station présentant l'abondance maximum : St. max.,
coefficient d'étalement en nombre de stations : Coef.

14.2.1. *Natures des distributions.*

Les paramètres servant à simuler les distributions sont :

Esp.	A. M.	St. max.	Coef.
1 à 10	1000	1 à 10	10

Ces caractéristiques ont été choisies pour la grande analogie existant entre les résultats fournis par les deux analyses.

Une caractéristique très importante de ce tableau est sa symétrie parfaite. Cela indique que les lignes et les colonnes peuvent jouer des rôles identiques dans la détection du gradient. Il s'agit donc bien d'un tableau de contingence.

Le croissant, témoin du gradient, est parfaitement visible tant pour le graphe des lignes que pour celui des colonnes.

14.2.2. *Evolutions des données.*

Divers graphiques sont fournis pour montrer l'évolution des données transformées selon les deux analyses. Ce qui donne l'occasion de découvrir un magnifique "papillon" (Frisque, 1987), qui est d'autant plus beau, qu'il ne sert pas à grand chose... . De quoi faire mentir ceux qui prétendent que les statistiques manquent de poésie !

14.2.3. *Contributions.*

On peut constater que, pour chaque analyse, ce sont les extrémités du gradient (lignes ou colonnes) qui ont les contributions les plus élevées à l'analyse, sauf pour les colonnes de l'ACP qui ont des contributions identiques. Le graphique rendant compte de l'évolution des variances-covariances pour l'AFC indique la raison de ce phénomène : le "diabolo" (Frisque, 1987) montre les colonnes les plus importantes, à variances élevées, et les colonnes à variances faibles et influences réduites.

Ceci laisse présager que si des difficultés doivent survenir, elles interviendront avec les lignes et colonnes associées au milieu du gradient, ce qui a déjà été constaté lors de l'exemple numérique.

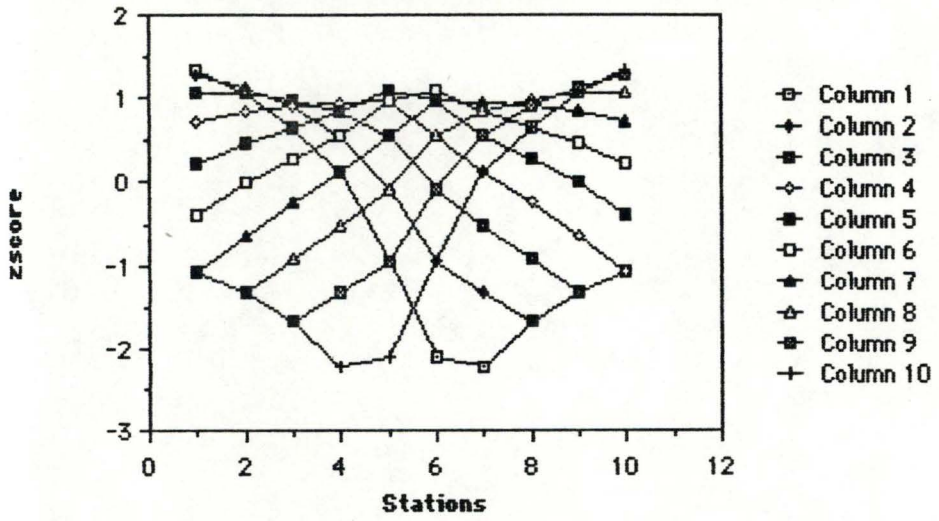
Les extrémités du gradient (variances élevées) sont donc surtout associées au premier axe et le centre (variances faibles) au deuxième.

Les valeurs des contributions relatives des lignes et des colonnes au premier axe sont, en règle générale, plus élevées avec l'AFC, surtout pour le centre du gradient. C'est un résultat très intéressant car il indique que le gradient est mieux représenté sur le premier axe factoriel proposé par l'AFC que sur la première composante de l'ACP. En effet, le reclassement des lignes et des colonnes s'avère plus facile le long du premier axe de l'AFC, les coordonnées des extrémités du gradient étant plus différenciées.

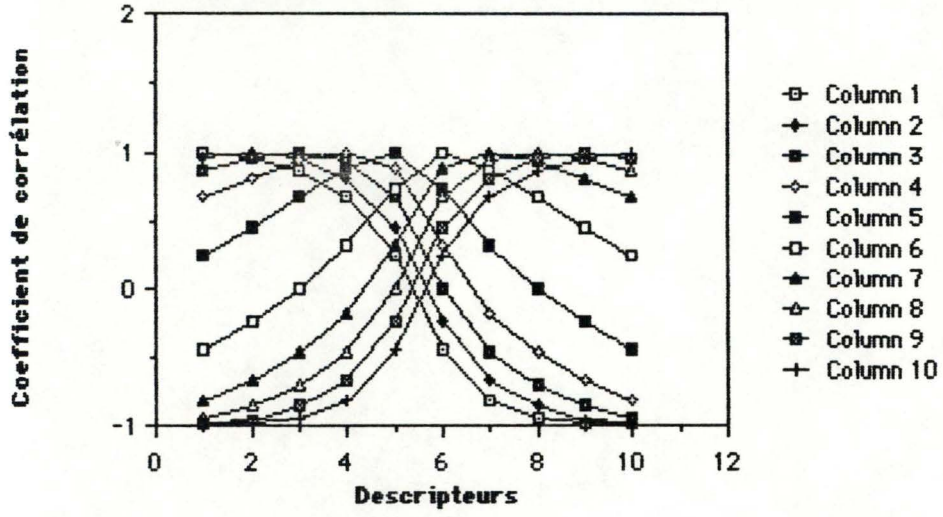
14.2.4. *Conclusion.*

L'AFC se révèle donc quelque peu supérieure à l'ACP, en produisant une structure en croissant légèrement redressée le long du premier axe. Cependant, il sera montré plus loin un exemple beaucoup plus parlant en ce qui concerne les déformations engendrées par les deux analyses.

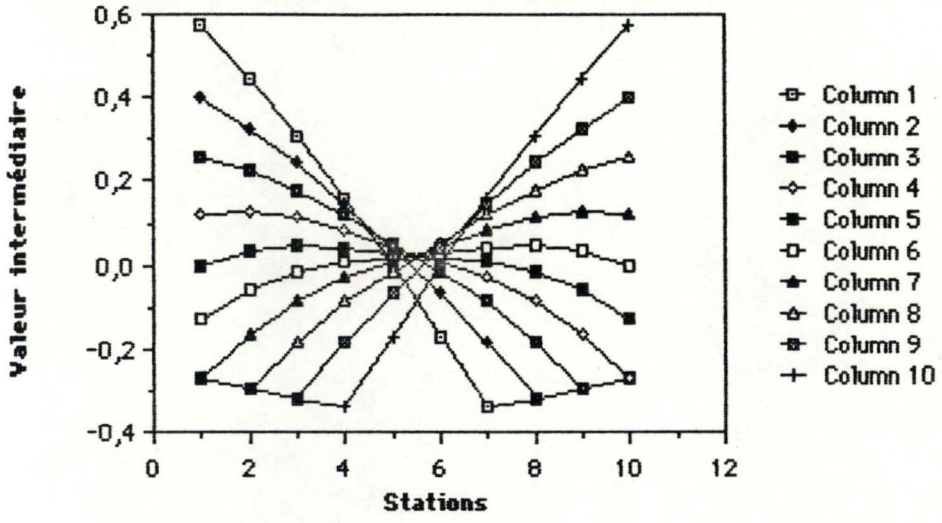
ACP : zscores des espèces dans les stations



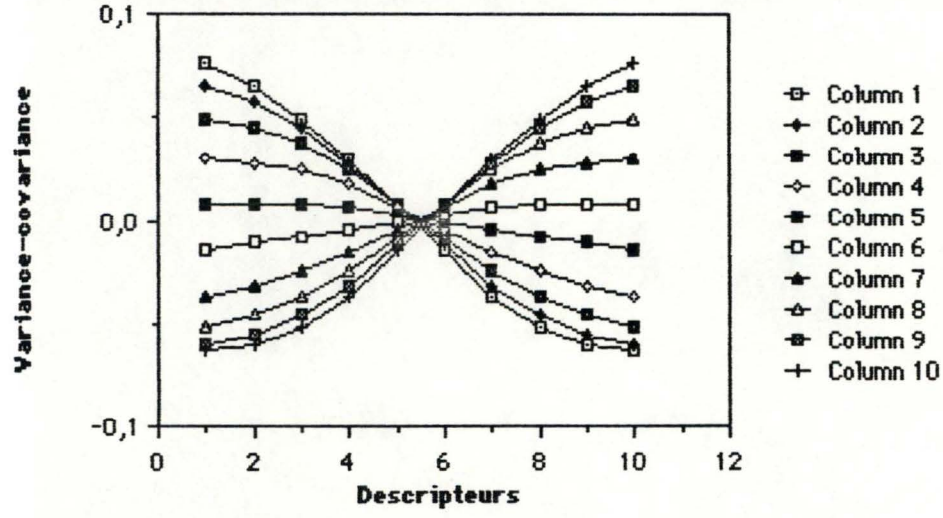
ACP : coefficients des corrélations entre descripteurs



AFC : valeurs intermédiaires des descripteurs ("papillon")



AFC : variances et covariances des descripteurs ("diabolo")



ACP

val. propres	%relatifs	%cumulés
7.0005	70.0051	70.0051
2.8816	28.8161	98.8212
0.0583	0.5826	99.4038
0.0301	0.3012	99.7050
0.0144	0.1442	99.8492
0.0075	0.0754	99.9246
0.0044	0.0438	99.9684
0.0032	0.0316	100.0000
0.0000	0.0000	100.0000
0.0000	0.0000	100.0000
10.0000		

me des éléments = 423.1708

s des lignes en %

e (1) = 7.2069
e (2) = 8.7921
e (3) = 10.2358
e (4) = 11.4438
e (5) = 12.3214
e (6) = 12.3214
e (7) = 11.4438
e (8) = 10.2358
e (9) = 8.7921
e (10) = 7.2069

s des colonnes en %

nne (1) = 7.2069
nne (2) = 8.7921
nne (3) = 10.2358
nne (4) = 11.4438
nne (5) = 12.3214
nne (6) = 12.3214
nne (7) = 11.4438
nne (8) = 10.2358
nne (9) = 8.7921
nne (10) = 7.2069

ACP

ributions des lignes l'analyse, en %

e (1) = 19.3202
e (2) = 12.3952
e (3) = 8.2795
e (4) = 5.8281
e (5) = 4.1770
e (6) = 4.1770
e (7) = 5.8281
e (8) = 8.2795
e (9) = 12.3952
e (10) = 19.3202

ributions absolues des lignes

e (1) =	0.1678	0.2573	0.1707
e (2) =	0.1604	0.0374	0.0146
e (3) =	0.1140	0.0057	0.0981
e (4) =	0.0519	0.0711	0.1918
e (5) =	0.0059	0.1286	0.0248
e (6) =	0.0059	0.1286	0.0248
e (7) =	0.0519	0.0711	0.1918
e (8) =	0.1140	0.0057	0.0981
e (9) =	0.1604	0.0374	0.0146
e (10) =	0.1678	0.2573	0.1707

ributions relatives (cos2) des lignes

e (1) =	0.6081	0.3838	0.0051
e (2) =	0.9057	0.0868	0.0007
e (3) =	0.9639	0.0198	0.0069
e (4) =	0.6231	0.3513	0.0192
e (5) =	0.0996	0.8871	0.0035
e (6) =	0.0996	0.8871	0.0035
e (7) =	0.6231	0.3513	0.0192
e (8) =	0.9639	0.0198	0.0069
e (9) =	0.9057	0.0868	0.0007
e (10) =	0.6081	0.3838	0.0051

ribution des colonnes l'analyse, en %

ACP

nne (1) = 10.0000
nne (2) = 10.0000
nne (3) = 10.0000
nne (4) = 10.0000
nne (5) = 10.0000
nne (6) = 10.0000
nne (7) = 10.0000
nne (8) = 10.0000
nne (9) = 10.0000
nne (10) = 10.0000

ributions absolues des colonnes

nne (1) =	0.1395	0.0039	0.1247
nne (2) =	0.1388	0.0050	0.0620
nne (3) =	0.1213	0.0479	0.0243
nne (4) =	0.0818	0.1432	0.2280
nne (5) =	0.0186	0.3000	0.0611
nne (6) =	0.0186	0.3000	0.0611
nne (7) =	0.0818	0.1432	0.2280
nne (8) =	0.1213	0.0479	0.0243
nne (9) =	0.1388	0.0050	0.0620
nne (10) =	0.1395	0.0039	0.1247

ributions relatives (cos2) des colonnes

nne (1) =	0.9765	0.0111	0.0073
nne (2) =	0.9718	0.0145	0.0036
nne (3) =	0.8489	0.1380	0.0014
nne (4) =	0.5730	0.4126	0.0133
nne (5) =	0.1302	0.8646	0.0036
nne (6) =	0.1302	0.8646	0.0036
nne (7) =	0.5730	0.4126	0.0133
nne (8) =	0.8489	0.1380	0.0014
nne (9) =	0.9718	0.0145	0.0036
nne (10) =	0.9765	0.0111	0.0073

ACP

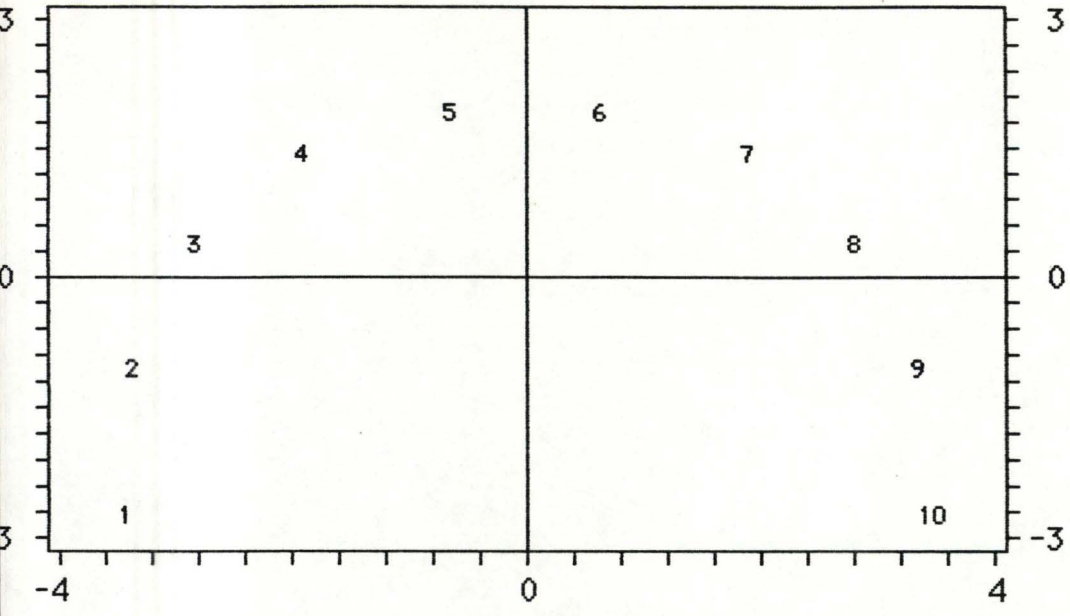
rdonnées des lignes sur les axes

t 1 :	-3.4276	-2.7230	0.3153
t 2 :	-3.3506	-1.0375	0.0922
t 3 :	-2.8250	0.4046	-0.2391
t 4 :	-1.9056	1.4309	-0.3343
t 5 :	-0.6449	1.9250	-0.1202
t 6 :	0.6449	1.9250	0.1202
t 7 :	1.9056	1.4309	0.3343
t 8 :	2.8250	0.4046	0.2391
t 9 :	3.3506	-1.0375	-0.0922
t 10 :	3.4276	-2.7230	-0.3153

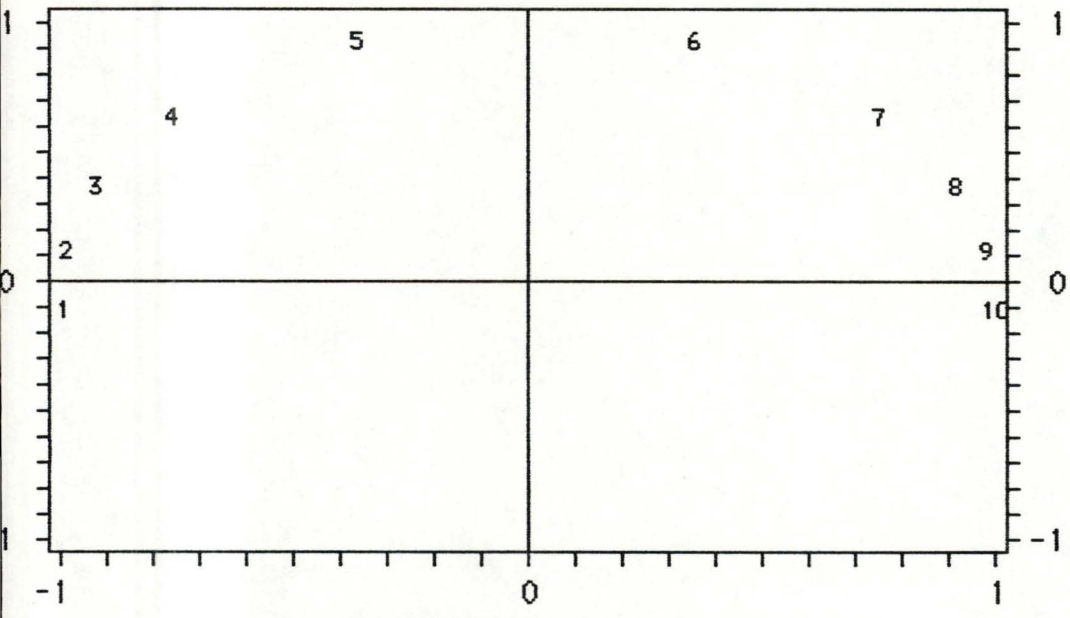
rdonnées des colonnes sur les axes

r. 1 :	-0.9882	-0.1055	-0.0852
r. 2 :	-0.9858	0.1205	-0.0601
r. 3 :	-0.9213	0.3715	0.0377
r. 4 :	-0.7570	0.6423	0.1152
r. 5 :	-0.3608	0.9298	0.0596
r. 6 :	0.3608	0.9298	-0.0596
r. 7 :	0.7570	0.6423	-0.1152
r. 8 :	0.9213	0.3715	-0.0377
r. 9 :	0.9858	0.1205	0.0601
r. 10 :	0.9882	-0.1055	0.0852

ACP

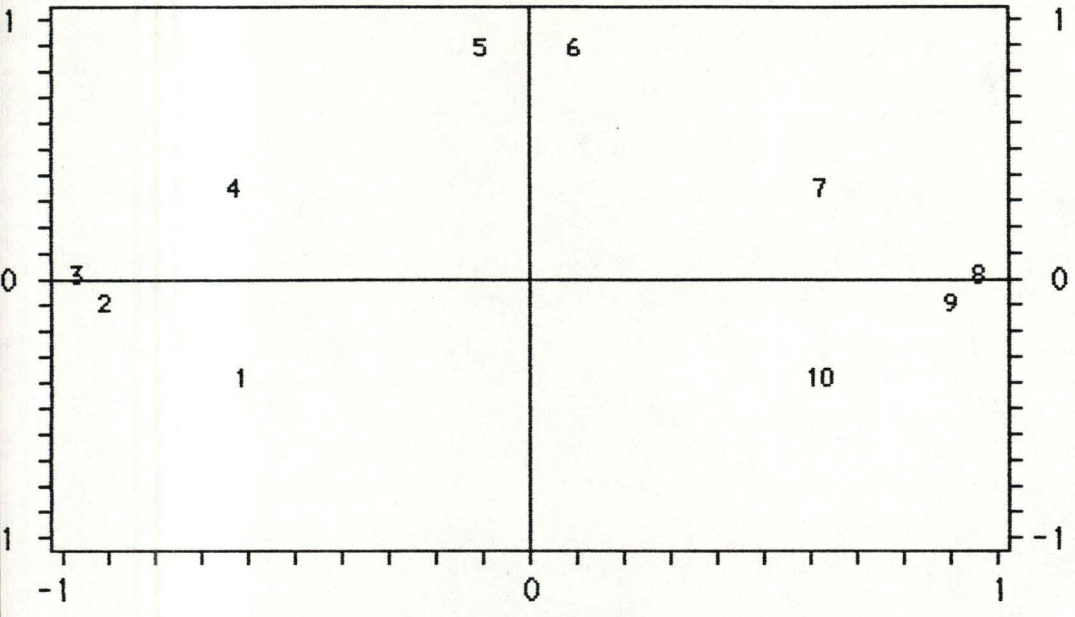


lignes dans le plan des composantes 1 & 2

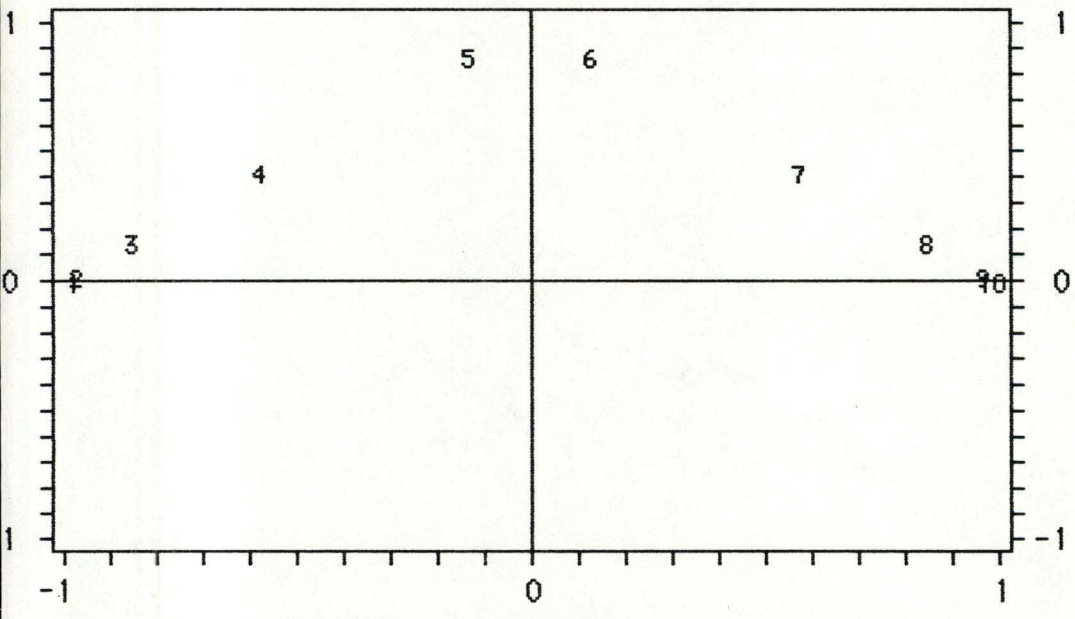


colonnes dans le plan des composantes 1 & 2

ACP



cos2 des lignes dans le plan des axes factoriels 1 & 2



cos2 des colonnes dans le plan des axes factoriels 1 & 2

AFC

val. propres	%relatifs	%cumulés
0.3596	93.5074	93.5074
0.0209	5.4361	98.9435
0.0019	0.4852	99.4287
0.0011	0.2915	99.7202
0.0006	0.1665	99.8867
0.0002	0.0621	99.9489
0.0002	0.0511	99.9999
0.0000	0.0001	100.0000
0.0000	0.0000	100.0000
0.0000	0.0000	100.0000

0.3845

ure du chi2 = 162.7252 81 degrés de liberté

me des éléments = 423.1708

s des lignes en %

e (1) = 7.2069
e (2) = 8.7921
e (3) = 10.2358
e (4) = 11.4438
e (5) = 12.3214
e (6) = 12.3214
e (7) = 11.4438
e (8) = 10.2358
e (9) = 8.7921
e (10) = 7.2069

s des colonnes en %

nne (1) = 7.2069
nne (2) = 8.7921
nne (3) = 10.2358
nne (4) = 11.4438
nne (5) = 12.3214
nne (6) = 12.3214
nne (7) = 11.4438
nne (8) = 10.2358
nne (9) = 8.7921
nne (10) = 7.2069

ributions des lignes l'analyse, en %

e(1) = 19.7399
e(2) = 14.9623
e(3) = 9.7260
e(4) = 4.7323
e(5) = 0.8394
e(6) = 0.8394
e(7) = 4.7323
e(8) = 9.7260
e(9) = 14.9623
e(10) = 19.7399

AFC

ributions absolues des lignes

e(1) =	0.1922	0.3038	0.1162
e(2) =	0.1562	0.0464	0.0257
e(3) =	0.1021	0.0080	0.0965
e(4) =	0.0446	0.0764	0.2331
e(5) =	0.0048	0.0654	0.0284
e(6) =	0.0048	0.0654	0.0284
e(7) =	0.0446	0.0764	0.2331
e(8) =	0.1021	0.0080	0.0965
e(9) =	0.1562	0.0464	0.0257
e(10) =	0.1922	0.3038	0.1162

ributions relatives (cos2) des lignes

e(1) =	0.9106	0.0837	0.0029
e(2) =	0.9765	0.0169	0.0008
e(3) =	0.9818	0.0044	0.0048
e(4) =	0.8814	0.0878	0.0239
e(5) =	0.5329	0.4238	0.0164
e(6) =	0.5329	0.4238	0.0164
e(7) =	0.8814	0.0878	0.0239
e(8) =	0.9818	0.0044	0.0048
e(9) =	0.9764	0.0169	0.0008
e(10) =	0.9106	0.0837	0.0029

ribution des colonnes l'analyse, en %

AFC

nne (1) = 19.7399
nne (2) = 14.9623
nne (3) = 9.7260
nne (4) = 4.7323
nne (5) = 0.8394
nne (6) = 0.8394
nne (7) = 4.7323
nne (8) = 9.7260
nne (9) = 14.9623
nne (10) = 19.7399

ributions absolues des colonnes

nne (1) =	0.1922	0.3038	0.1163
nne (2) =	0.1562	0.0464	0.0258
nne (3) =	0.1021	0.0080	0.0964
nne (4) =	0.0446	0.0764	0.2330
nne (5) =	0.0048	0.0654	0.0284
nne (6) =	0.0048	0.0654	0.0285
nne (7) =	0.0446	0.0764	0.2332
nne (8) =	0.1021	0.0080	0.0966
nne (9) =	0.1562	0.0464	0.0257
nne (10) =	0.1922	0.3038	0.1162

ributions relatives (cos2) des colonnes

nne (1) =	0.9107	0.0837	0.0029
nne (2) =	0.9764	0.0169	0.0008
nne (3) =	0.9819	0.0044	0.0048
nne (4) =	0.8813	0.0878	0.0239
nne (5) =	0.5327	0.4236	0.0164
nne (6) =	0.5332	0.4240	0.0165
nne (7) =	0.8816	0.0878	0.0239
nne (8) =	0.9816	0.0044	0.0048
nne (9) =	0.9765	0.0169	0.0008
nne (10) =	0.9106	0.0837	0.0029

AFC

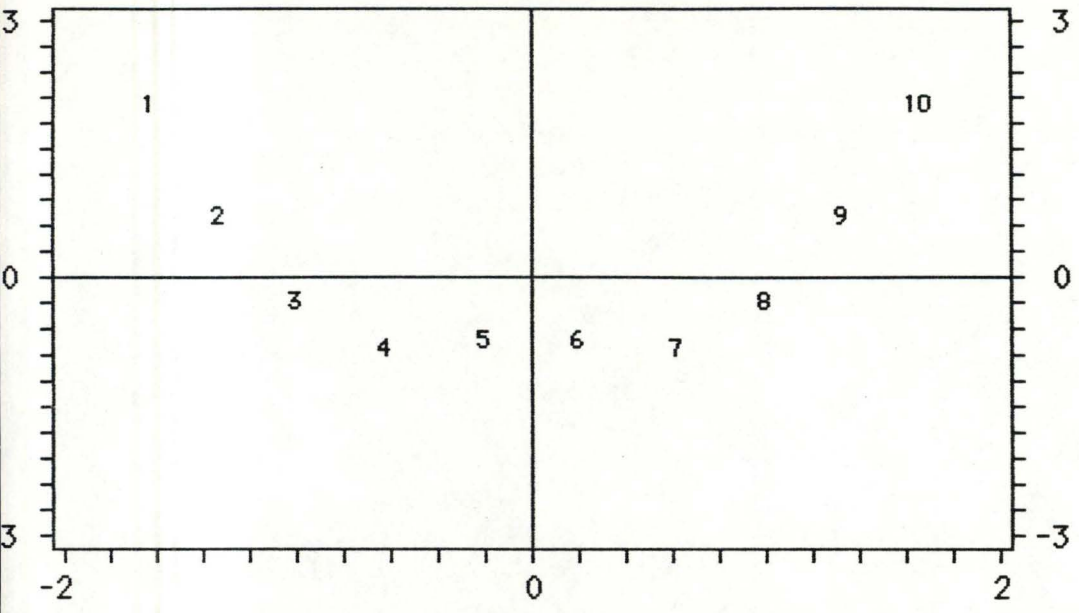
données des lignes sur les axes

t 1 :	-1.6332	2.0531	1.2700
t 2 :	-1.3331	0.7264	0.5410
t 3 :	-0.9988	-0.2787	-0.9710
t 4 :	-0.6243	-0.8171	-1.4271
t 5 :	-0.1970	-0.7288	-0.4805
t 6 :	0.1970	-0.7288	0.4805
t 7 :	0.6243	-0.8171	1.4271
t 8 :	0.9988	-0.2787	0.9710
t 9 :	1.3331	0.7264	-0.5411
t 10 :	1.6332	2.0531	-1.2700

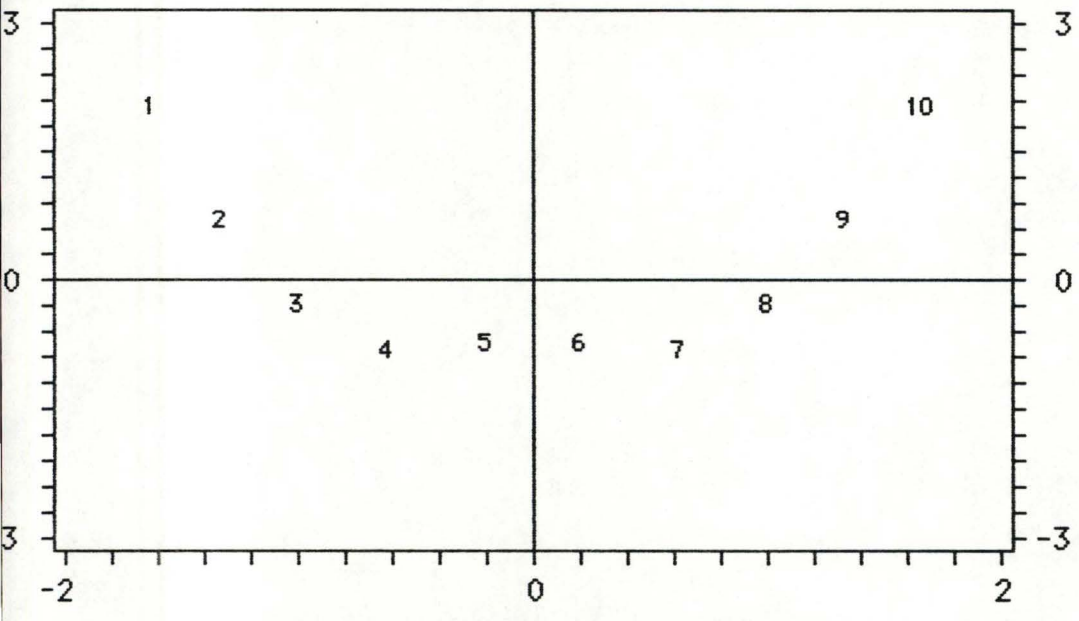
données des colonnes sur les axes

r. 1 :	-1.6332	2.0532	-1.2703
r. 2 :	-1.3331	0.7265	-0.5414
r. 3 :	-0.9988	-0.2787	0.9706
r. 4 :	-0.6243	-0.8171	1.4268
r. 5 :	-0.1970	-0.7288	0.4802
r. 6 :	0.1970	-0.7288	-0.4808
r. 7 :	0.6243	-0.8171	-1.4274
r. 8 :	0.9988	-0.2787	-0.9713
r. 9 :	1.3331	0.7265	0.5408
r. 10 :	1.6332	2.0532	1.2697

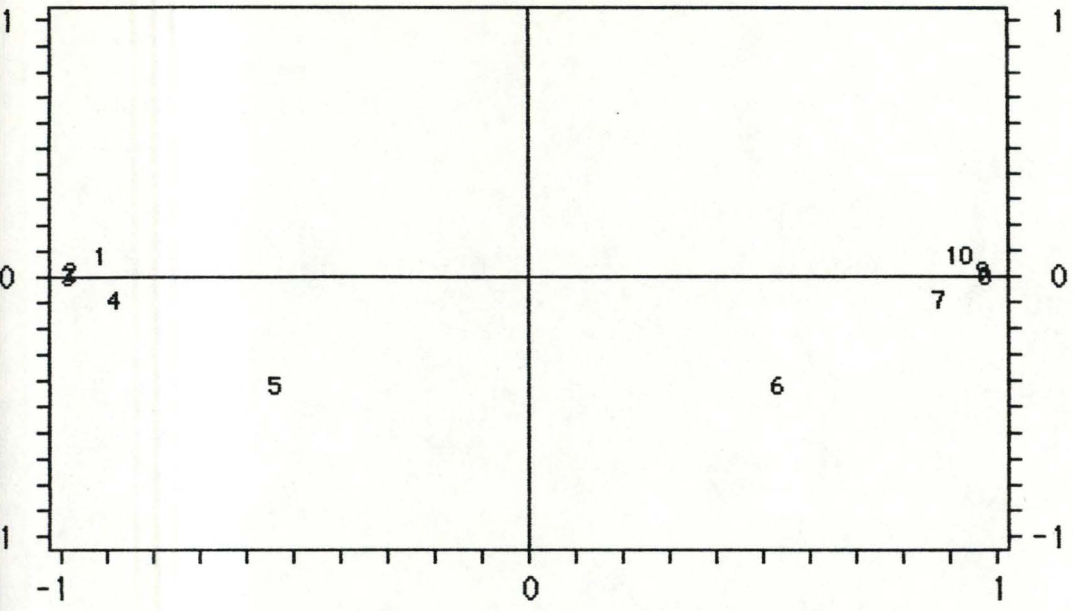
AFC



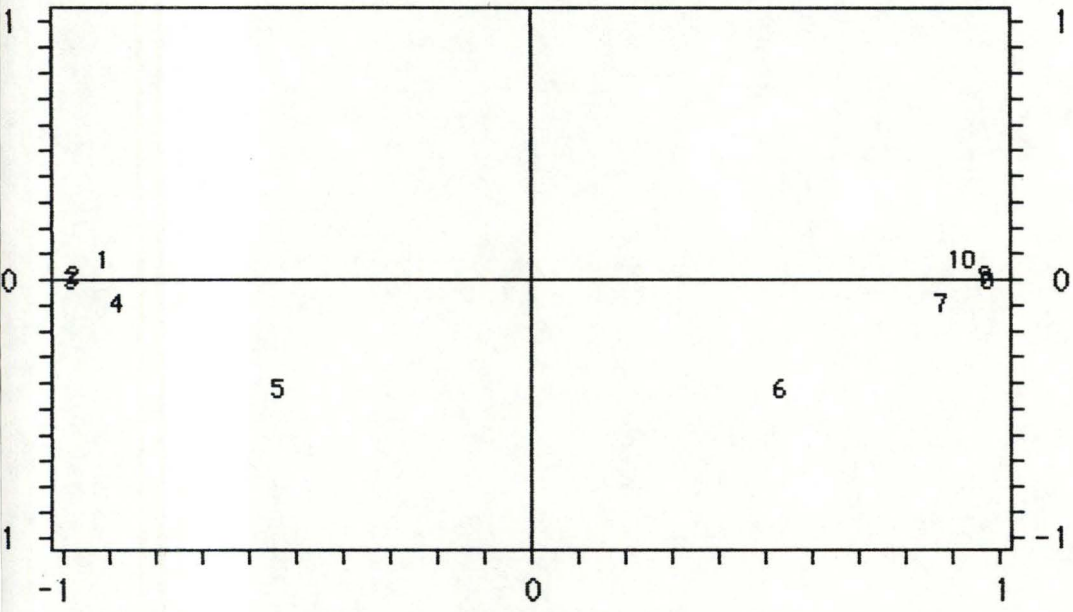
lignes dans le plan des axes factoriels 1 & 2



colonnes dans le plan des axes factoriels 1 & 2



cos2 des lignes dans le plan des axes factoriels 1 & 2



cos2 des colonnes dans le plan des axes factoriels 1 & 2

14.3. Influences des poids des lignes et des colonnes.

14.3.1. *Modifications des références.*

Les modifications apportées au tableau de référence sont :

ligne 1 divisée par 100,
ligne 5 divisée par 100,
colonne 1 divisée par 100,
colonne 5 divisée par 100.

14.3.2. *Projections selon l'ACP.*

Un reclassement correct n'est plus possible pour les lignes, que ce soit par la méthode du vecteur tournant, ou par projections sur le premier axe.

Le reclassement des colonnes est toujours possible par la méthode du vecteur, mais pas par projections.

14.3.3. *Projections selon l'AFC.*

Le reclassement des lignes et des colonnes peut être fait avec l'une ou l'autre des deux méthodes.

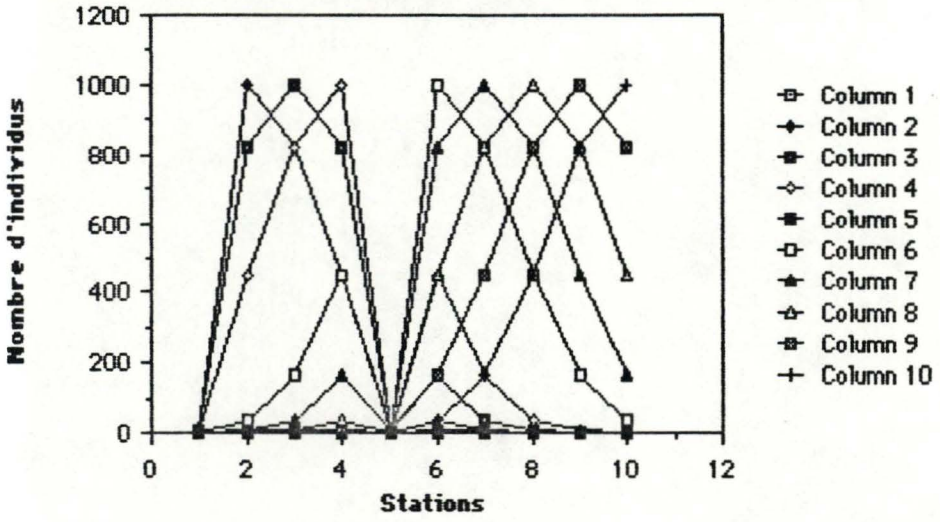
14.3.4. *Déformations dues aux modifications.*

Les déformations sont plus importantes au milieu du gradient (ligne et colonne 5) comme prévu plus haut, sauf pour l'ACP, qui, grâce à l'opération de standardisation selon les colonnes, est beaucoup moins sensible que l'AFC aux modifications concernant les colonnes.

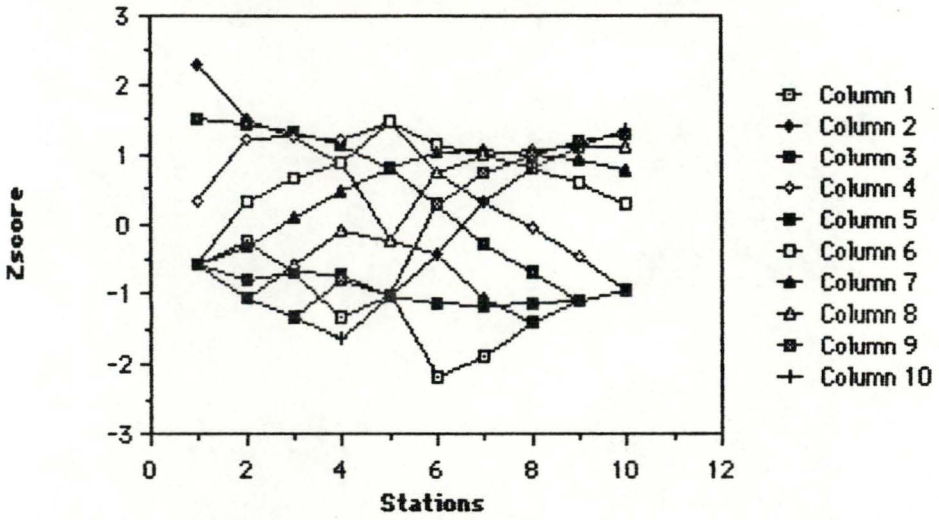
14.3.5. *Conclusion.*

Si l'ACP offre une plus grande résistance aux modifications des colonnes, cela ne constitue pas un avantage décisif, puisque cette dernière est moins fiable au niveau du reclassement tant des lignes que des colonnes. De plus, on peut vouloir utiliser cette propriété de l'AFC pour distinguer les espèces dont les distributions s'écartent le plus du profil moyen.

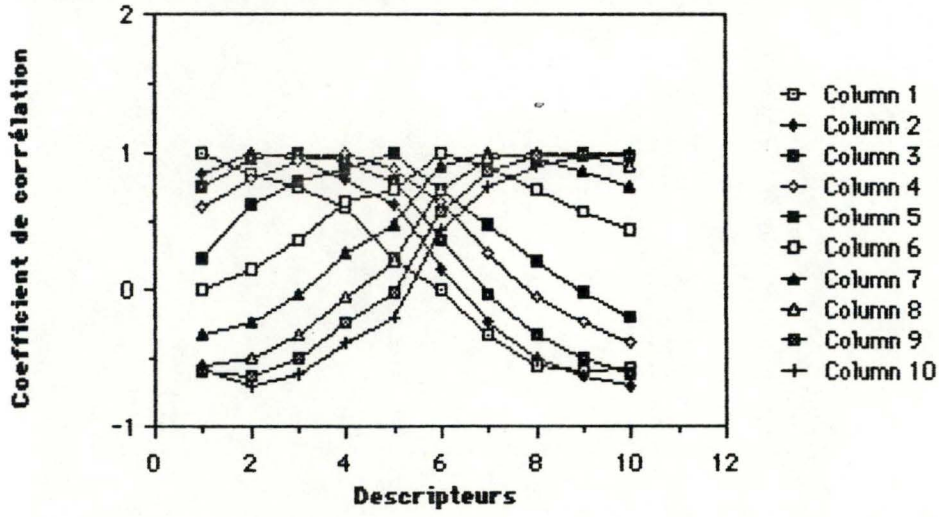
Influences des poids des lignes et des colonnes



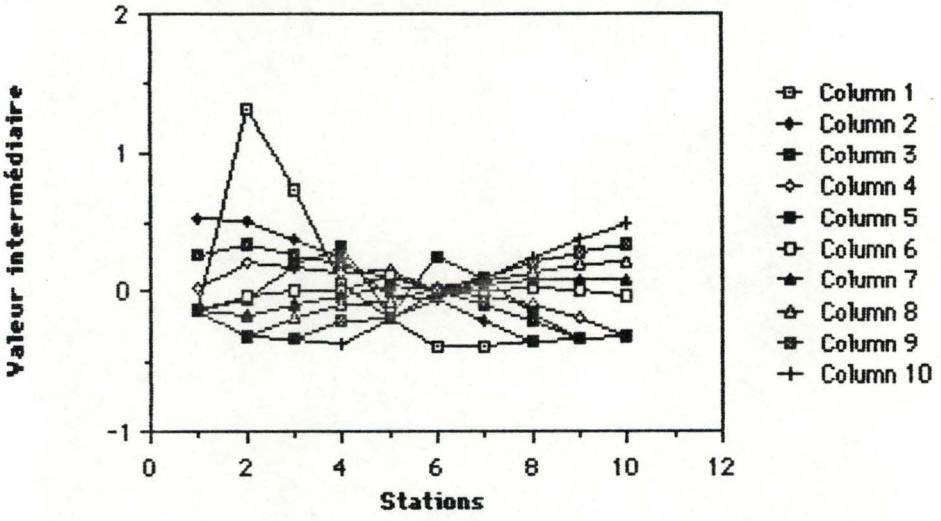
ACP : zscores des espèces dans les stations



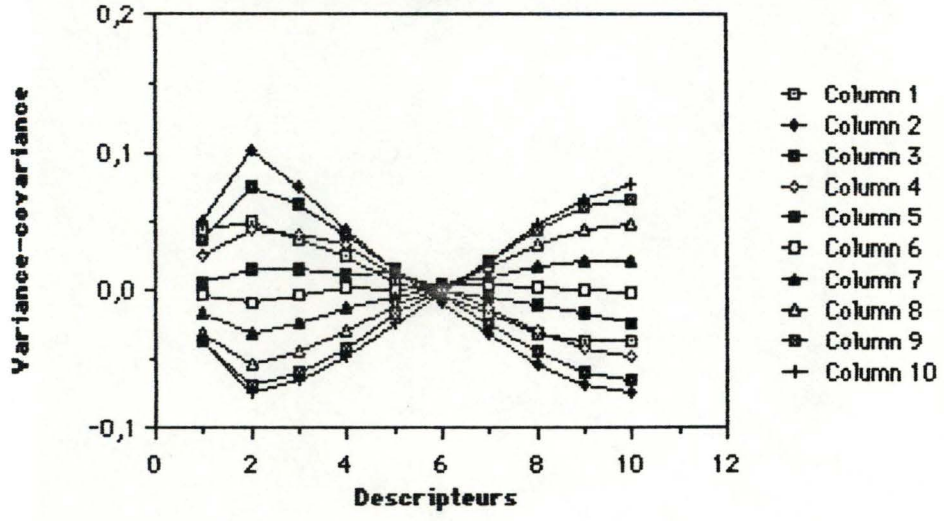
ACP : coefficients des corrélations entre descripteurs



AFC : valeurs intermédiaires des descripteurs



AFC : variances et covariances des descripteurs



ACP

val. propres	%relatifs	%cumulés
5.3502	53.5017	53.5017
3.9762	39.7624	93.2640
0.5323	5.3231	98.5871
0.1058	1.0577	99.6448
0.0200	0.2003	99.8451
0.0101	0.1005	99.9457
0.0035	0.0349	99.9806
0.0019	0.0187	99.9993
0.0001	0.0007	100.0000
0.0000	0.0000	100.0000
10.0000		

me des éléments = 306.1382

s des lignes en %

e (1) = 1.4699
e (2) = 9.2363
e (3) = 11.2091
e (4) = 12.9017
e (5) = 2.9397
e (6) = 14.9226
e (7) = 14.3488
e (8) = 12.7054
e (9) = 10.9402
e (10) = 9.3264

s des colonnes en %

nne (1) = 1.4699
nne (2) = 9.2363
nne (3) = 11.2091
nne (4) = 12.9017
nne (5) = 2.9397
nne (6) = 14.9226
nne (7) = 14.3488
nne (8) = 12.7054
nne (9) = 10.9402
nne (10) = 9.3264

ributions des lignes l'analyse, en %

ACP

(1) = 16.0849
(2) = 15.1750
(3) = 10.7820
(4) = 8.5516
(5) = 9.1509
(6) = 7.1083
(7) = 5.9580
(8) = 6.9211
(9) = 9.5231
(10) = 10.7449

ributions absolues des lignes

(1) =	0.0196	0.3601	0.0976
(2) =	0.2463	0.0009	0.3615
(3) =	0.1744	0.0314	0.0264
(4) =	0.0718	0.1039	0.0947
(5) =	0.0063	0.1975	0.1030
(6) =	0.0034	0.1564	0.1216
(7) =	0.0404	0.0920	0.0207
(8) =	0.1131	0.0165	0.0090
(9) =	0.1656	0.0032	0.0840
(10) =	0.1592	0.0381	0.0816

ributions relatives (cos2) des lignes

(1) =	0.0651	0.8903	0.0323
(2) =	0.8682	0.0024	0.1268
(3) =	0.8655	0.1159	0.0130
(4) =	0.4494	0.4829	0.0589
(5) =	0.0371	0.8580	0.0599
(6) =	0.0256	0.8747	0.0911
(7) =	0.3625	0.6142	0.0185
(8) =	0.8744	0.0947	0.0069
(9) =	0.9301	0.0135	0.0470
(10) =	0.7925	0.1410	0.0404

ribution des colonnes l'analyse, en %

ACP

nne (1) = 10.0000
nne (2) = 10.0000
nne (3) = 10.0000
nne (4) = 10.0000
nne (5) = 10.0000
nne (6) = 10.0000
nne (7) = 10.0000
nne (8) = 10.0000
nne (9) = 10.0000
nne (10) = 10.0000

ributions absolues des colonnes

nne (1) =	0.1208	0.0084	0.5995
nne (2) =	0.1488	0.0394	0.0199
nne (3) =	0.1206	0.0877	0.0001
nne (4) =	0.0636	0.1586	0.0017
nne (5) =	0.0206	0.1911	0.2112
nne (6) =	0.0107	0.2319	0.0022
nne (7) =	0.0708	0.1549	0.0016
nne (8) =	0.1269	0.0793	0.0009
nne (9) =	0.1541	0.0363	0.0338
nne (10) =	0.1630	0.0122	0.1290

ributions relatives (cos2) des colonnes

nne (1) =	0.6462	0.0333	0.3191
nne (2) =	0.7960	0.1567	0.0106
nne (3) =	0.6454	0.3489	0.0001
nne (4) =	0.3402	0.6308	0.0009
nne (5) =	0.1100	0.7599	0.1124
nne (6) =	0.0574	0.9220	0.0012
nne (7) =	0.3790	0.6161	0.0009
nne (8) =	0.6791	0.3155	0.0005
nne (9) =	0.8246	0.1444	0.0180
nne (10) =	0.8723	0.0486	0.0687

données des lignes sur les axes

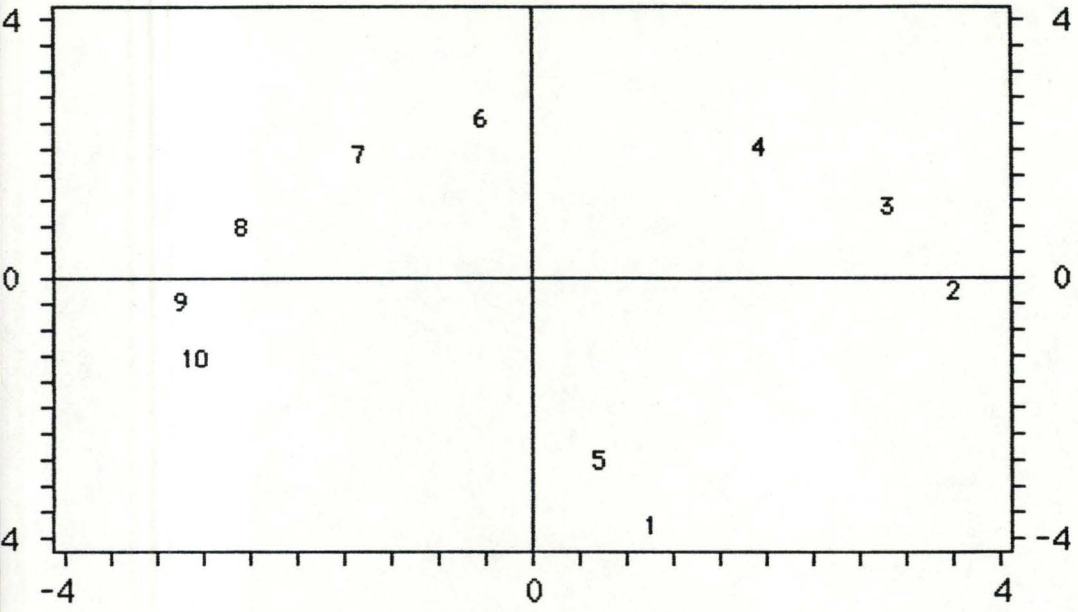
t 1 :	1.0232	-3.7842	-0.7208
t 2 :	3.6297	-0.1897	1.3871
t 3 :	3.0548	1.1178	0.3747
t 4 :	1.9603	2.0322	-0.7098
t 5 :	0.5823	-2.8020	-0.7405
t 6 :	-0.4266	2.4935	-0.8046
t 7 :	-1.4696	1.9129	-0.3322
t 8 :	-2.4600	0.8095	0.2186
t 9 :	-2.9761	-0.3589	0.6687
t 10 :	-2.9181	-1.2309	0.6589

ACP

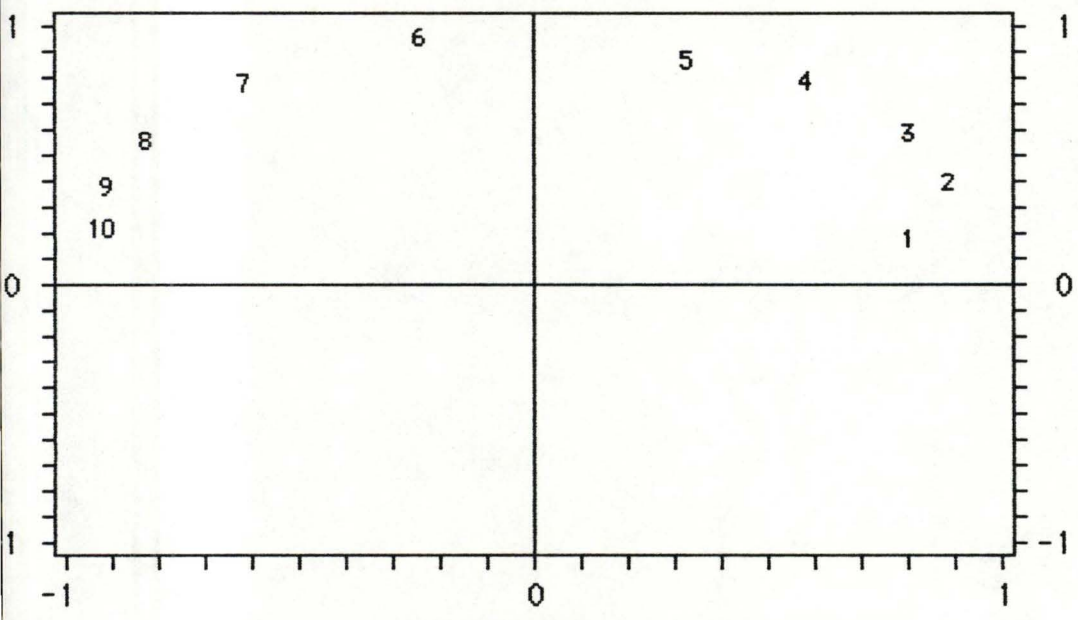
données des colonnes sur les axes

r. 1 :	0.8039	0.1826	0.5649
r. 2 :	0.8922	0.3959	0.1029
r. 3 :	0.8033	0.5906	-0.0074
r. 4 :	0.5833	0.7942	-0.0303
r. 5 :	0.3317	0.8717	-0.3353
r. 6 :	-0.2395	0.9602	0.0340
r. 7 :	-0.6156	0.7849	0.0292
r. 8 :	-0.8241	0.5617	0.0223
r. 9 :	-0.9080	0.3800	0.1342
r. 10 :	-0.9340	0.2204	0.2621

ACP

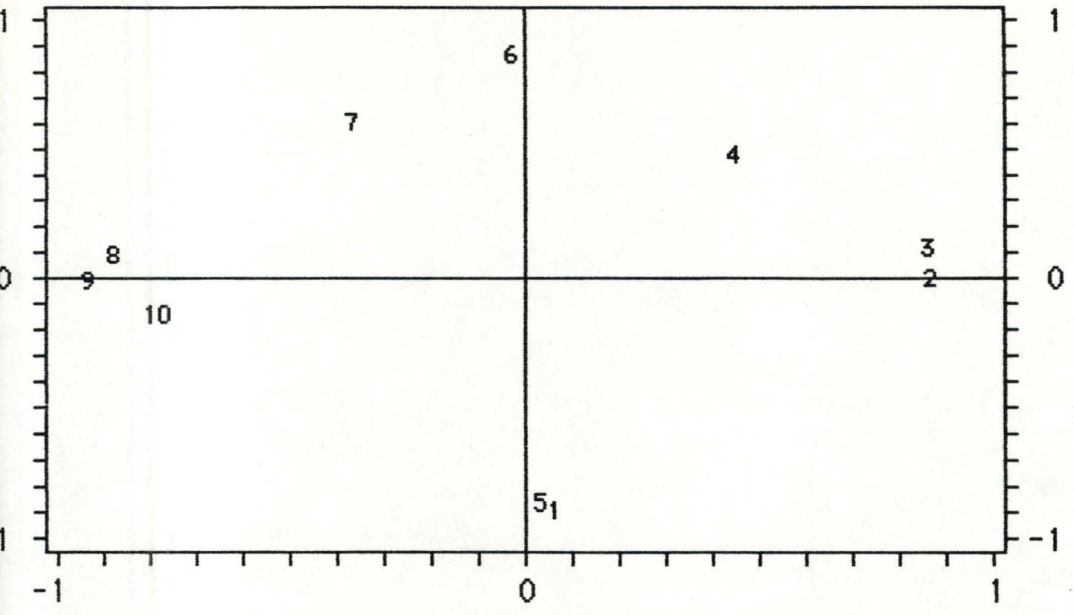


lignes dans le plan des composantes 1 & 2

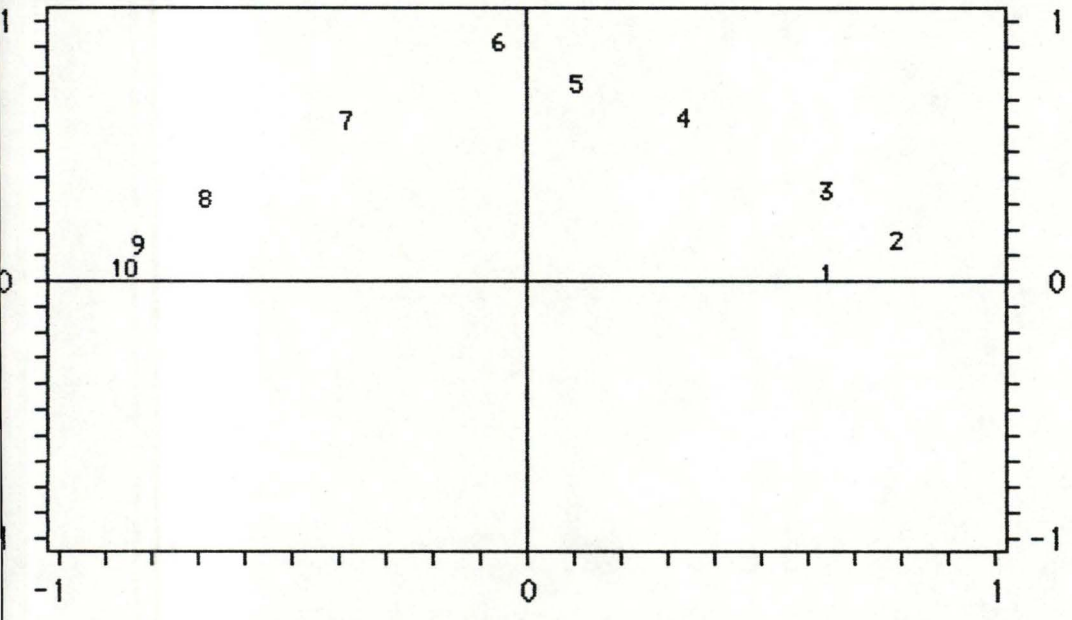


colonnes dans le plan des composantes 1 & 2

ACP



cos2 des lignes dans le plan des axes factoriels 1 & 2



cos2 des colonnes dans le plan des axes factoriels 1 & 2

AFC

val. propres	%relatifs	%cumulés
0.3784	86.6824	86.6824
0.0309	7.0842	93.7666
0.0178	4.0814	97.8481
0.0074	1.6946	99.5427
0.0011	0.2624	99.8051
0.0005	0.1177	99.9227
0.0002	0.0545	99.9772
0.0001	0.0228	100.0000
0.0000	0.0000	100.0000
0.0000	0.0000	100.0000
0.4366		

ure du chi2 = 133.6482 81 degrés de liberté

me des éléments = 306.1382

s des lignes en %

e (1) = 1.4699
e (2) = 9.2363
e (3) = 11.2091
e (4) = 12.9017
e (5) = 2.9397
e (6) = 14.9226
e (7) = 14.3488
e (8) = 12.7054
e (9) = 10.9402
e (10) = 9.3264

s des colonnes en %

nne (1) = 1.4699
nne (2) = 9.2363
nne (3) = 11.2091
nne (4) = 12.9017
nne (5) = 2.9397
nne (6) = 14.9226
nne (7) = 14.3488
nne (8) = 12.7054
nne (9) = 10.9402
nne (10) = 9.3264

AFC

ributions des lignes l'analyse, en %

e(1) = 9.7863
e(2) = 23.0635
e(3) = 14.1649
e(4) = 7.3458
e(5) = 3.2361
e(6) = 1.0304
e(7) = 2.2964
e(8) = 7.5985
e(9) = 13.7575
e(10) = 17.7206

ributions absolues des lignes

e(1) =	0.0685	0.1777	0.5888
e(2) =	0.2385	0.2099	0.2115
e(3) =	0.1604	0.0001	0.0328
e(4) =	0.0727	0.0990	0.0417
e(5) =	0.0110	0.1447	0.0181
e(6) =	0.0003	0.0865	0.0684
e(7) =	0.0221	0.0403	0.0076
e(8) =	0.0856	0.0061	0.0156
e(9) =	0.1530	0.0419	0.0149
e(10) =	0.1879	0.1938	0.0007

ributions relatives (cos2) des lignes

e(1) =	0.6069	0.1286	0.2456
e(2) =	0.8963	0.0645	0.0374
e(3) =	0.9818	0.0000	0.0095
e(4) =	0.8578	0.0955	0.0231
e(5) =	0.2951	0.3169	0.0229
e(6) =	0.0255	0.5984	0.2724
e(7) =	0.8348	0.1244	0.0135
e(8) =	0.9783	0.0057	0.0084
e(9) =	0.9640	0.0216	0.0044
e(10) =	0.9190	0.0775	0.0002

ribution des colonnes l'analyse, en %

AFC

nne (1) = 9.7863
nne (2) = 23.0635
nne (3) = 14.1649
nne (4) = 7.3458
nne (5) = 3.2361
nne (6) = 1.0304
nne (7) = 2.2964
nne (8) = 7.5985
nne (9) = 13.7575
nne (10) = 17.7206

ributions absolues des colonnes

nne (1) =	0.0685	0.1777	0.5888
nne (2) =	0.2385	0.2099	0.2115
nne (3) =	0.1604	0.0001	0.0328
nne (4) =	0.0727	0.0990	0.0417
nne (5) =	0.0110	0.1447	0.0181
nne (6) =	0.0003	0.0865	0.0684
nne (7) =	0.0221	0.0403	0.0076
nne (8) =	0.0856	0.0061	0.0156
nne (9) =	0.1530	0.0419	0.0149
nne (10) =	0.1879	0.1938	0.0007

ributions relatives (cos2) des colonnes

nne (1) =	0.6069	0.1286	0.2456
nne (2) =	0.8963	0.0645	0.0374
nne (3) =	0.9818	0.0000	0.0094
nne (4) =	0.8579	0.0955	0.0232
nne (5) =	0.2951	0.3169	0.0228
nne (6) =	0.0255	0.5983	0.2724
nne (7) =	0.8348	0.1244	0.0135
nne (8) =	0.9783	0.0057	0.0084
nne (9) =	0.9640	0.0216	0.0044
nne (10) =	0.9189	0.0775	0.0002

AFC

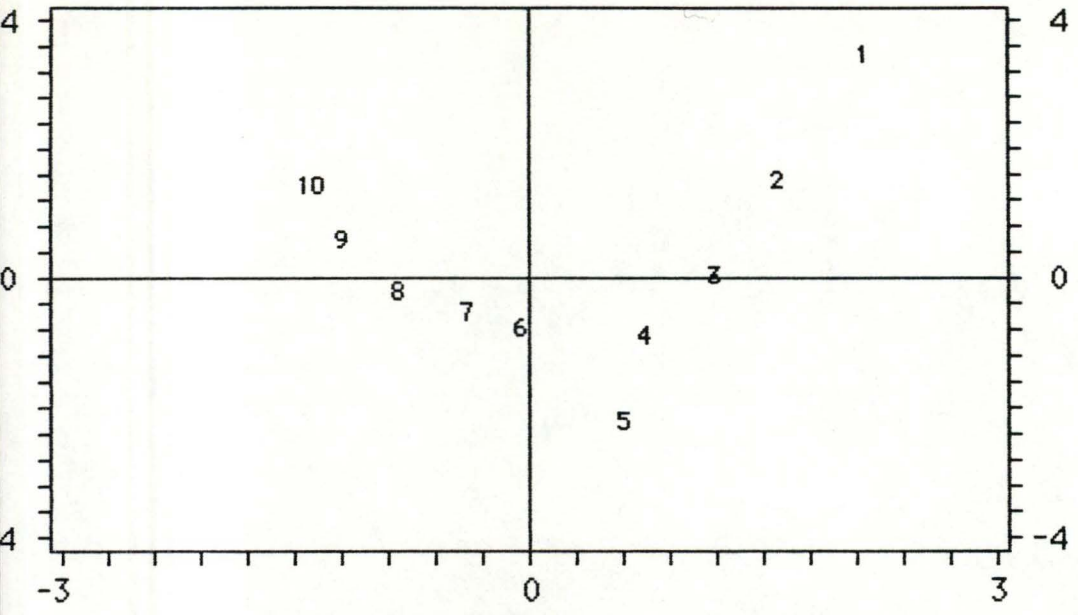
rdonnées des lignes sur les axes

t 1 :	2.1589	3.4765	6.3290
t 2 :	1.6068	1.5076	-1.5132
t 3 :	1.1963	0.0285	-0.5409
t 4 :	0.7506	-0.8760	0.5682
t 5 :	0.6121	-2.2187	-0.7849
t 6 :	-0.0449	-0.7614	0.6768
t 7 :	-0.3926	-0.5301	0.2298
t 8 :	-0.8210	-0.2185	-0.3508
t 9 :	-1.1826	0.6188	-0.3689
t 10 :	-1.4192	1.4416	0.0867

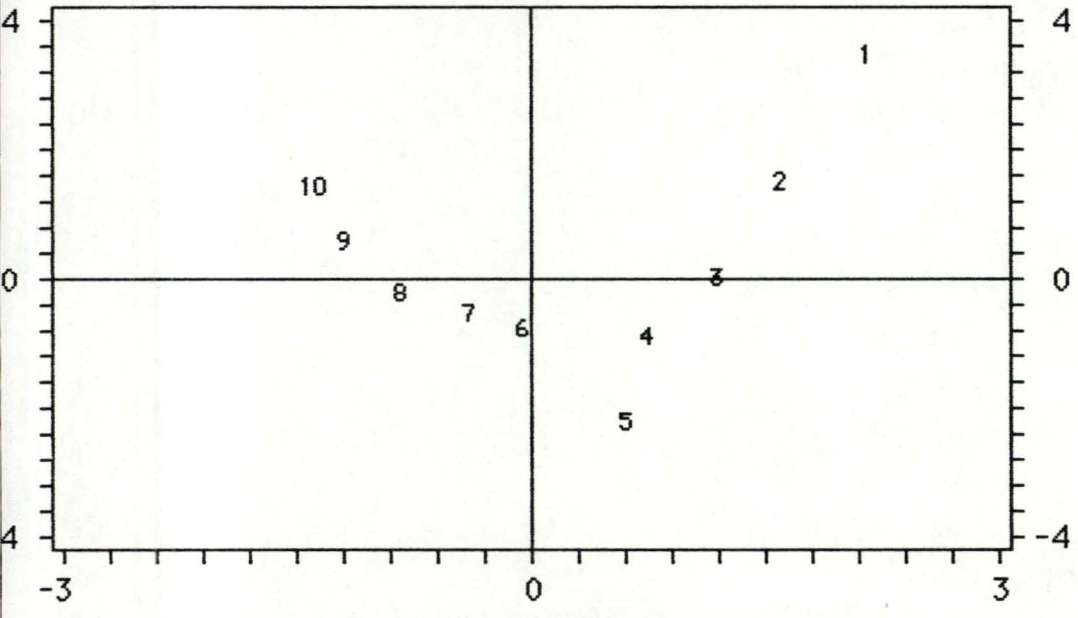
rdonnées des colonnes sur les axes

r. 1 :	2.1589	3.4765	-6.3290
r. 2 :	1.6068	1.5076	1.5132
r. 3 :	1.1963	0.0285	0.5409
r. 4 :	0.7506	-0.8760	-0.5683
r. 5 :	0.6121	-2.2187	0.7849
r. 6 :	-0.0449	-0.7614	-0.6768
r. 7 :	-0.3926	-0.5301	-0.2298
r. 8 :	-0.8210	-0.2185	0.3508
r. 9 :	-1.1826	0.6188	0.3689
r. 10 :	-1.4192	1.4416	-0.0867

AFC

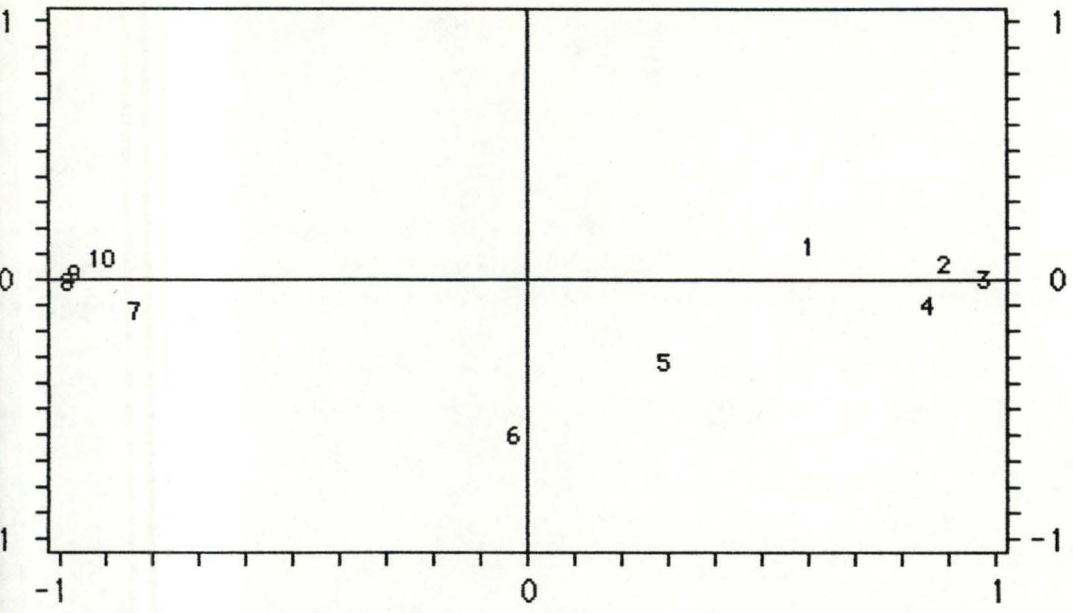


lignes dans le plan des axes factoriels 1 & 2

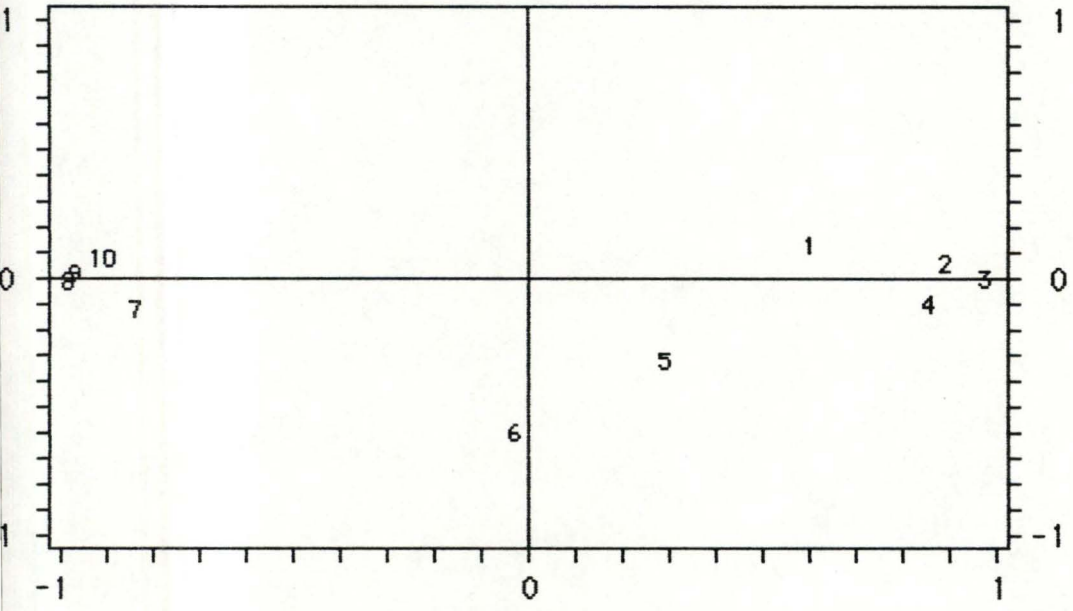


colonnes dans le plan des axes factoriels 1 & 2

AFC



cos2 des lignes dans le plan des axes factoriels 1 & 2



cos2 des colonnes dans le plan des axes factoriels 1 & 2

14.4. Aptitude à la détection d'un gradient.

Cette aptitude pourra être estimée soit pour un gradient de faible intensité, soit dans le cas où seules quelques espèces répondent a ce gradient. C'est ce dernier cas qui sera exploité.

14.4.1. *Natures des distributions.*

Les paramètres servant à simuler les distributions sont :

Esp.	A. M.	St. max.	Coef.
1 à 4	100	-	0
5 à 6	1000	5 à 6	10
7 à 10	100	-	0

Les données sont simulées avec une légère variabilité de façon à éviter les problèmes liés aux variances nulles pour l'ACP. Les stations des maxima sont choisies de façon à assurer une asymétrie, et les coefficients d'étalements assurent le recouvrement de toute les stations.

14.4.2. *Projections des lignes.*

L'ACP est beaucoup moins sensible, et ne détecte pas la légère asymétrie, alors que l'AFC permet le reclassement des stations le long du gradient par la méthode du vecteur tournant.

14.4.3. *Projections des colonnes.*

L'AFC montre bien quelles sont les espèces importantes pour la détection du gradient, l'ACP ne montre rien du tout.

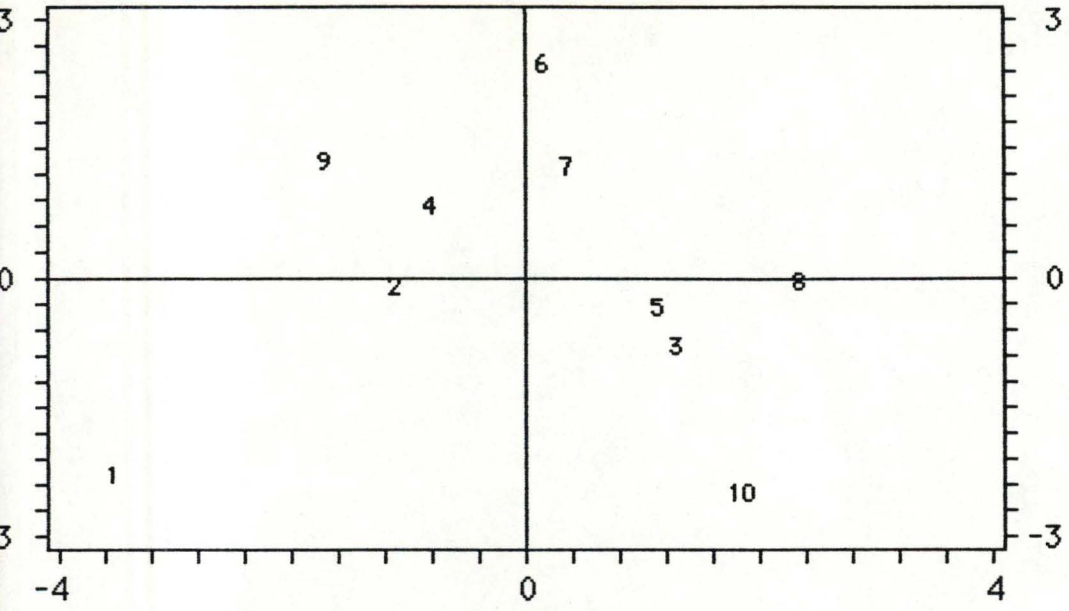
14.4.4. *Influence d'une légère variabilité résiduelle.*

L'AFC montre une résistance beaucoup plus grande que l'ACP, en produisant des graphiques très peu perturbés.

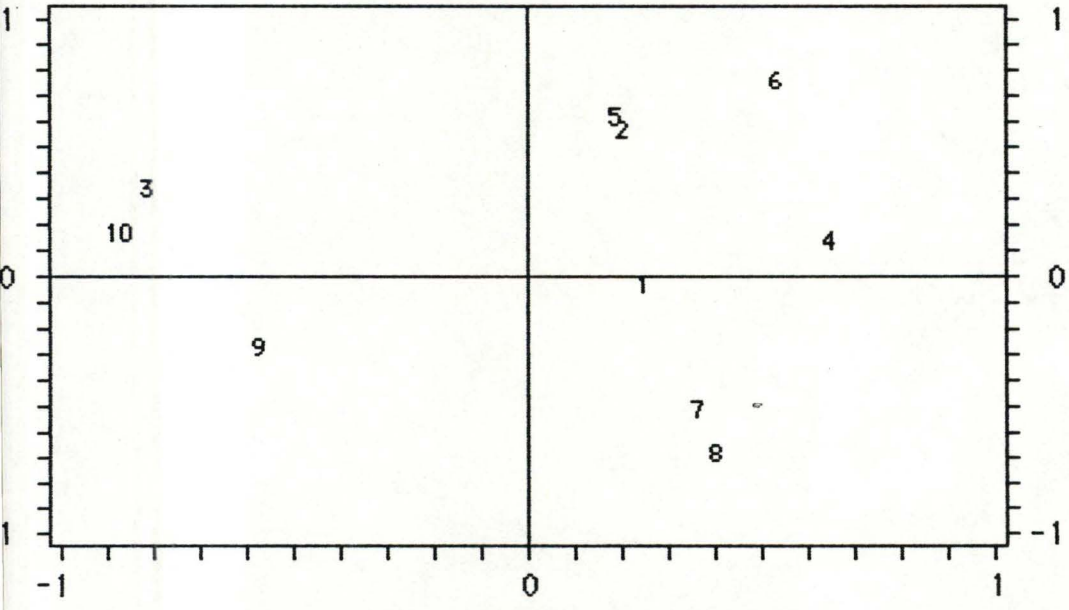
14.4.5. *Conclusion.*

L'AFC semble plus apte à la découverte d'un gradient de faible intensité, ou dans le cas où peu d'espèces présentent une réponse à ce gradient. De plus, l'introduction d'une variabilité résiduelle ne limite pas ses capacités.

ACP



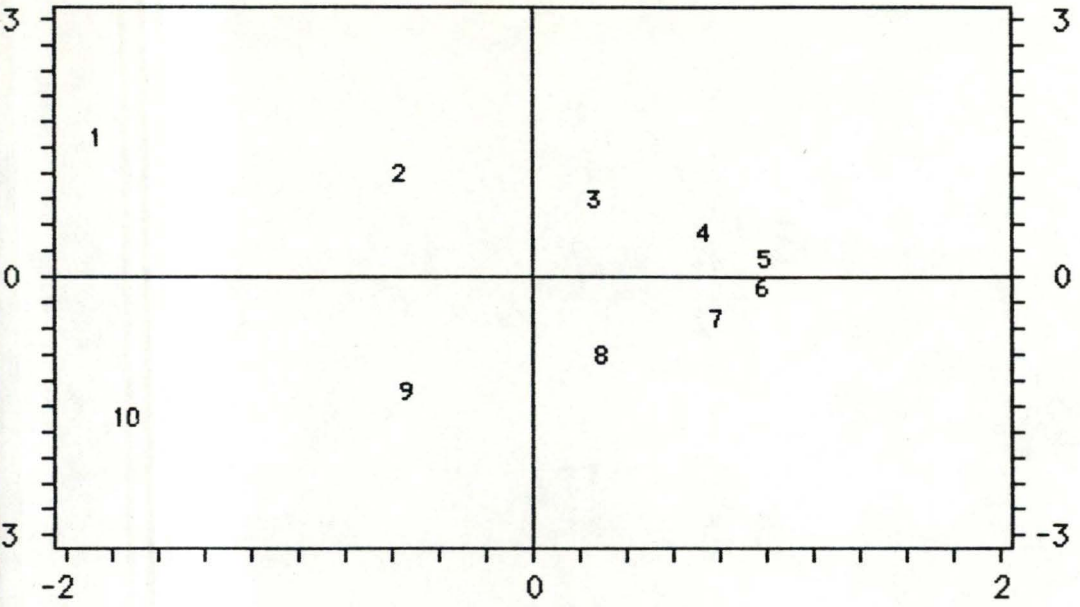
lignes dans le plan des composantes 1 & 2



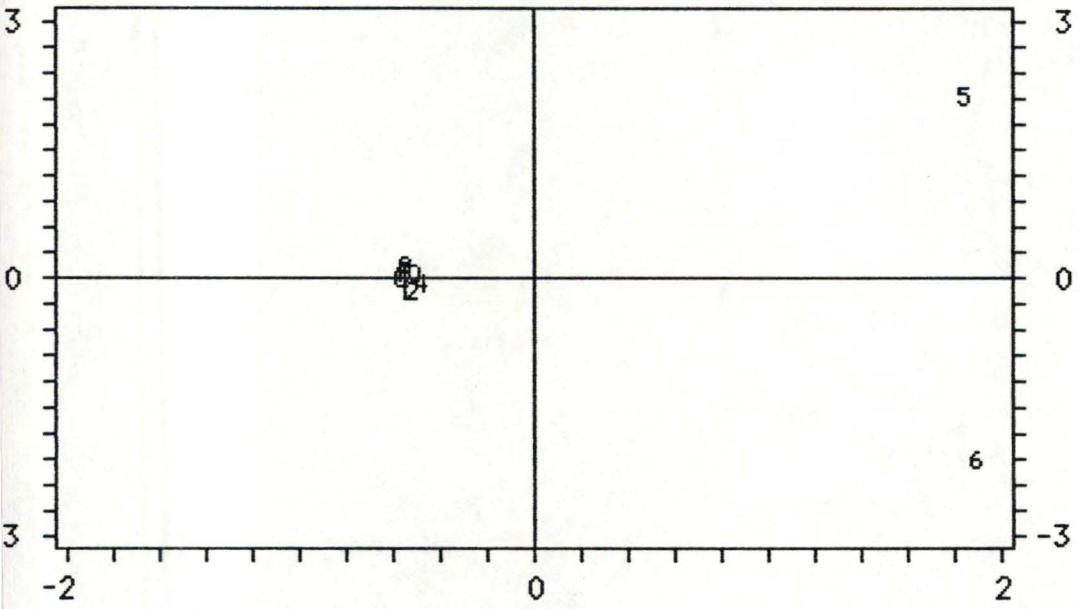
colonnes dans le plan des composantes 1 & 2

AVEC VARIABILITÉ

AFC



lignes dans le plan des axes factoriels 1 & 2



colonnes dans le plan des axes factoriels 1 & 2

AVEC VARIABILITÉ

14.5. Indépendance des facteurs.

On sait déjà que l'AFC produit des structures "en croissant" plus redressées le long du premier axe que l'ACP. L'exemple suivant permet de montrer que l'AFC atteint effectivement mieux l'un des buts communs aux deux analyses : rendre les facteurs indépendants.

14.5.1. *Natures des distributions.*

Les paramètres servant à simuler les distributions sont :

Esp.	A. M.	St. max.	Coef.
1 à 10	100	1 à 10	100

On remarquera qu'il s'agit d'un exemple tout à fait classique. Le coefficient d'étalement est élevé, ce qui limite la dépendance curvilinéaire. Les données sont simulées avec et sans variabilité.

14.5.2. *Projections des lignes et des colonnes.*

Le croissant est bien présent avec l'ACP, mais il devient une ligne on ne peut plus droite avec l'AFC. Si l'on tient compte des valeurs propres et des pourcentages de la variabilité totale exprimés par chacun des axes, on constate que l'ACP répartit l'information interprétable sur au moins deux axes et l'AFC sur un seul axe (simulation sans variabilité). La valeur propre associée au second axe de l'AFC se situe dans les limites de précision de calcul, on ne peut donc pas dire si elle est significativement différente de zéro (simulation sans variabilité).

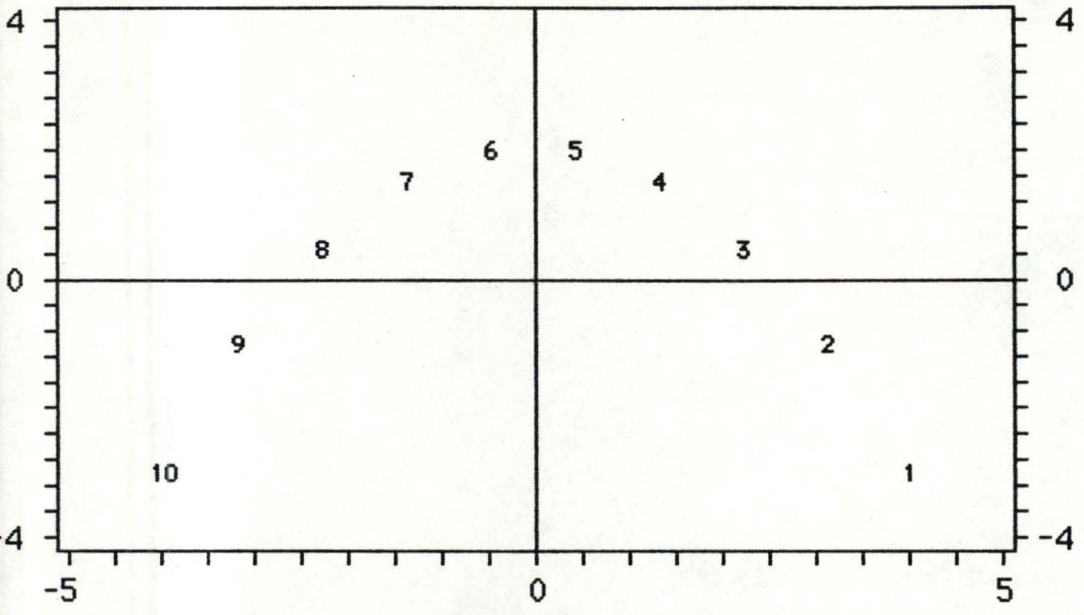
La variabilité détruit cette disposition dans le cas de l'AFC, mais le reclassement est toujours possible par projection sur le premier axe, cela montre que la variabilité introduite est rejetée sur le second axe.

14.5.3. *Conclusion.*

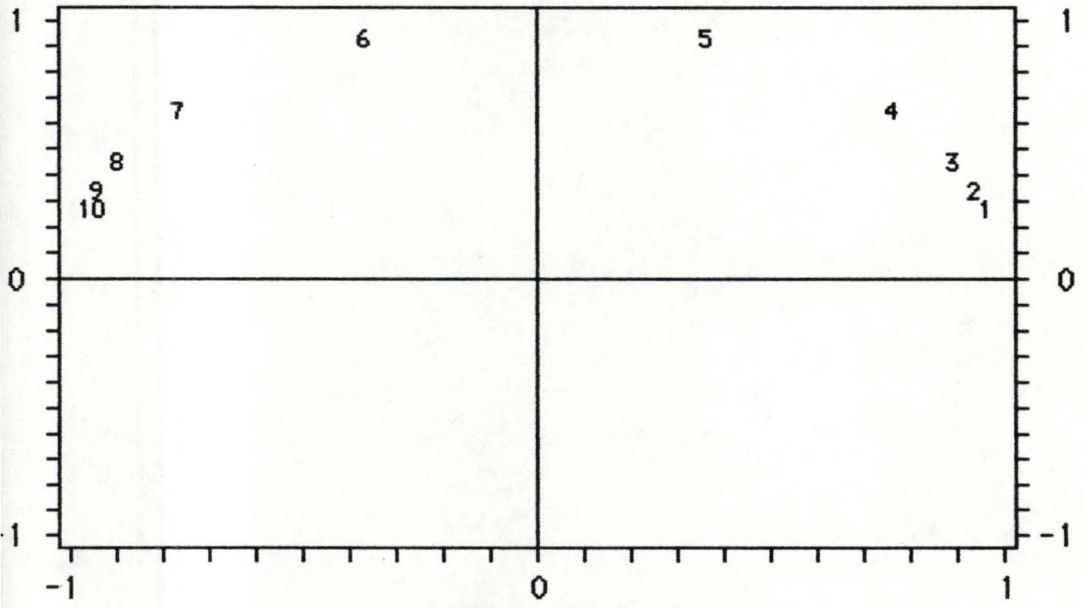
L'AFC est moins sensible au phénomène de dépendance curvilinéaire induit par le type de distribution des espèces le long d'un gradient.

Cela montre que l'AFC est plus fiable, car elle produit moins de déformations.

ACP



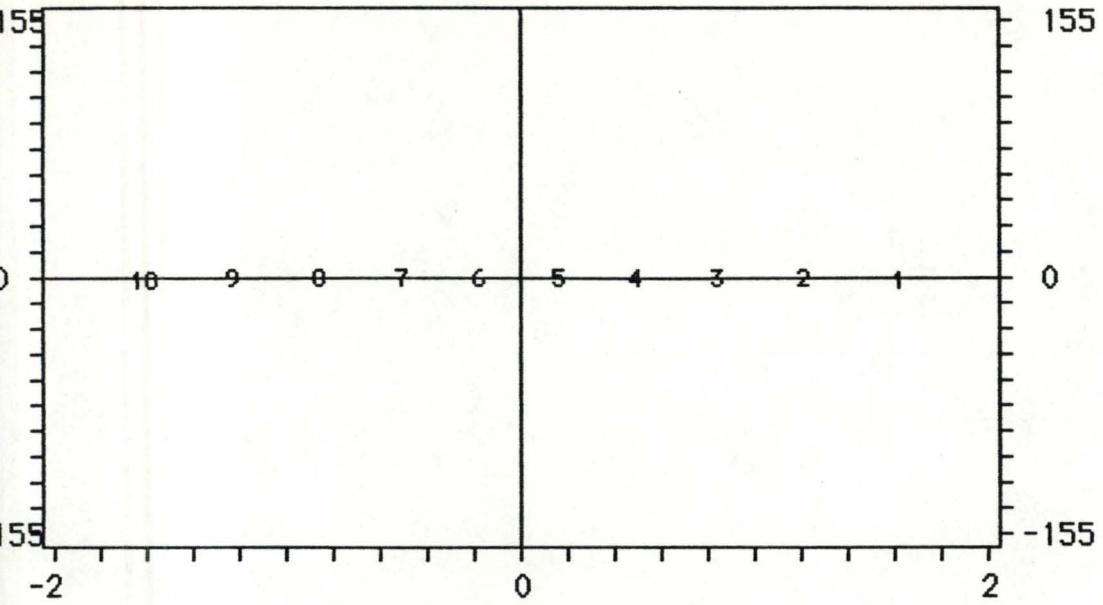
lignes dans le plan des composantes 1 & 2



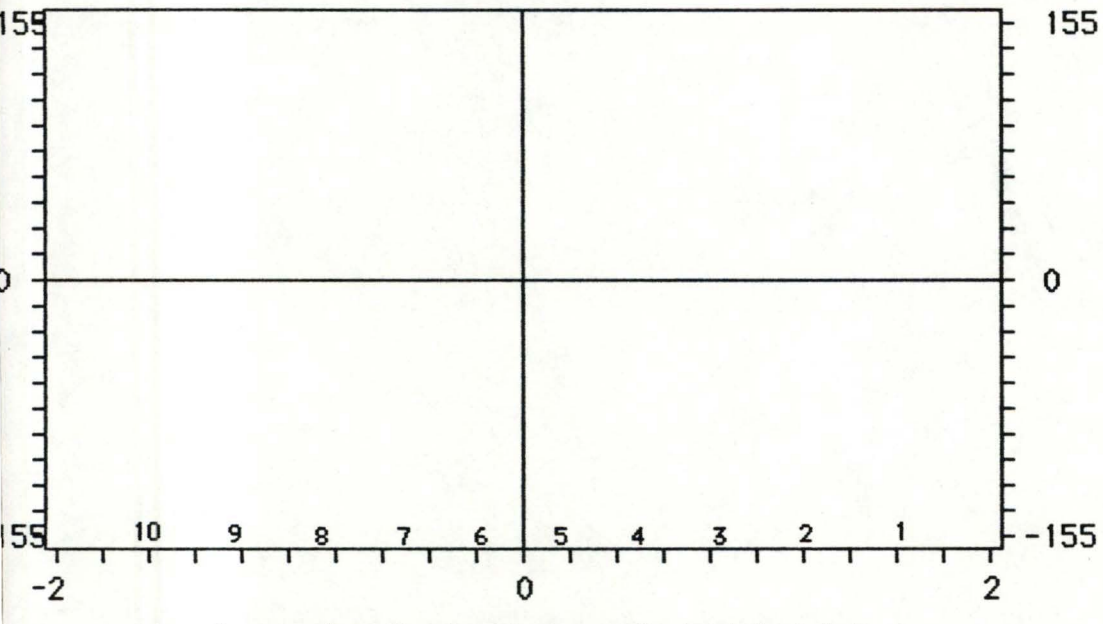
colonnes dans le plan des composantes 1 & 2

SANS VARIABILITÉ

AFC



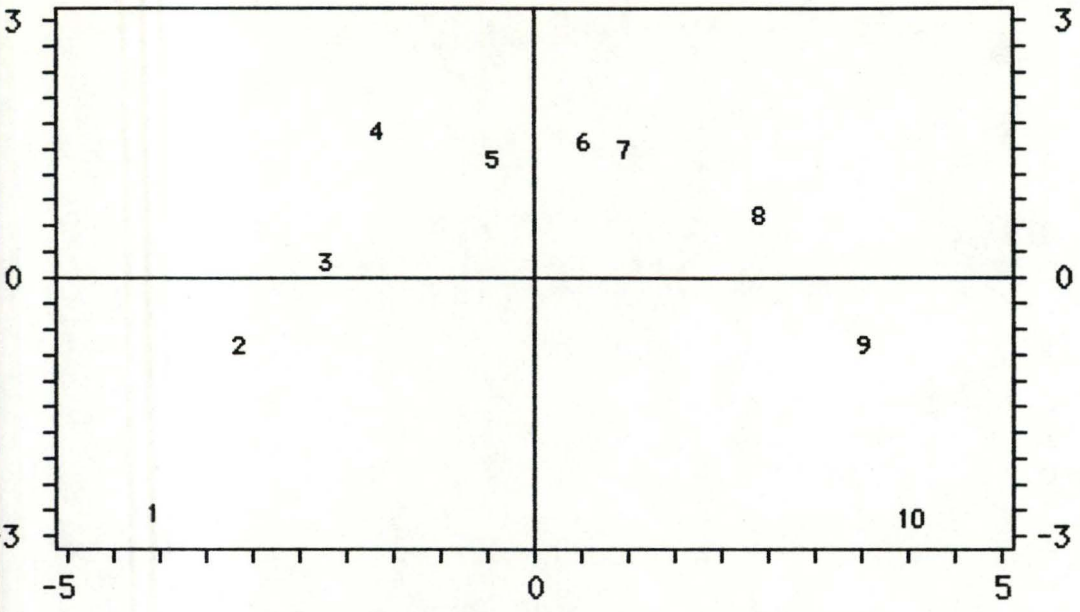
lignes dans le plan des axes factoriels 1 & 2



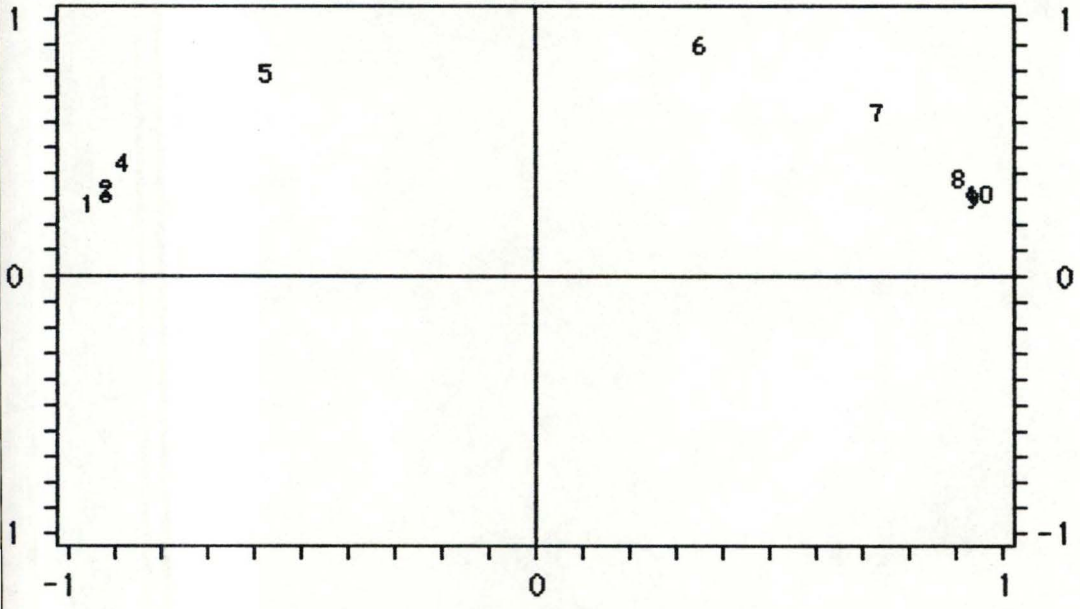
colonnes dans le plan des axes factoriels 1 & 2

SANS VARIABILITÉ

ACP

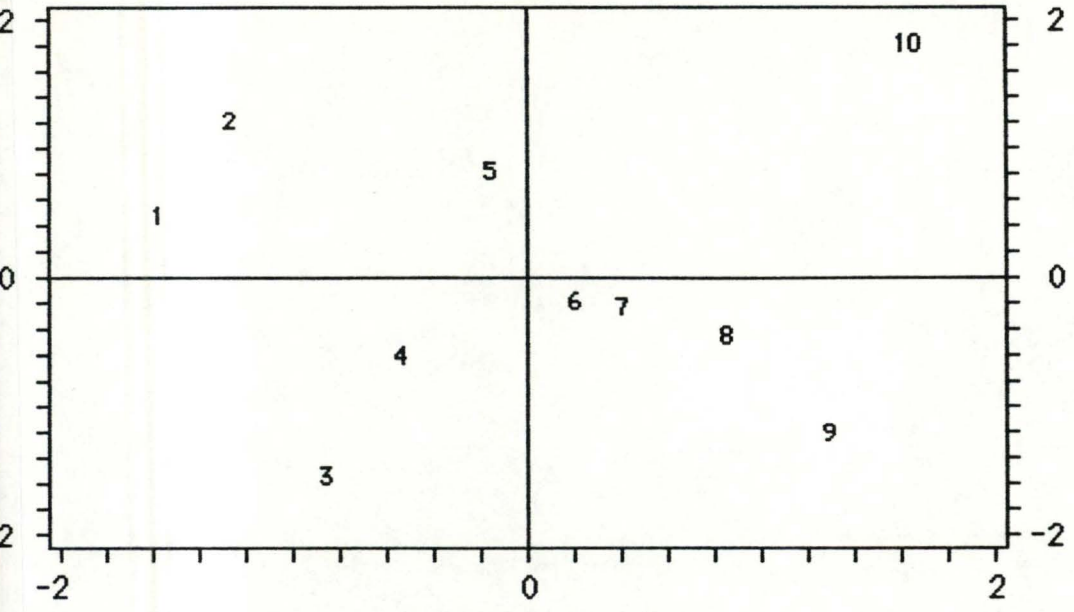


lignes dans le plan des composantes 1 & 2



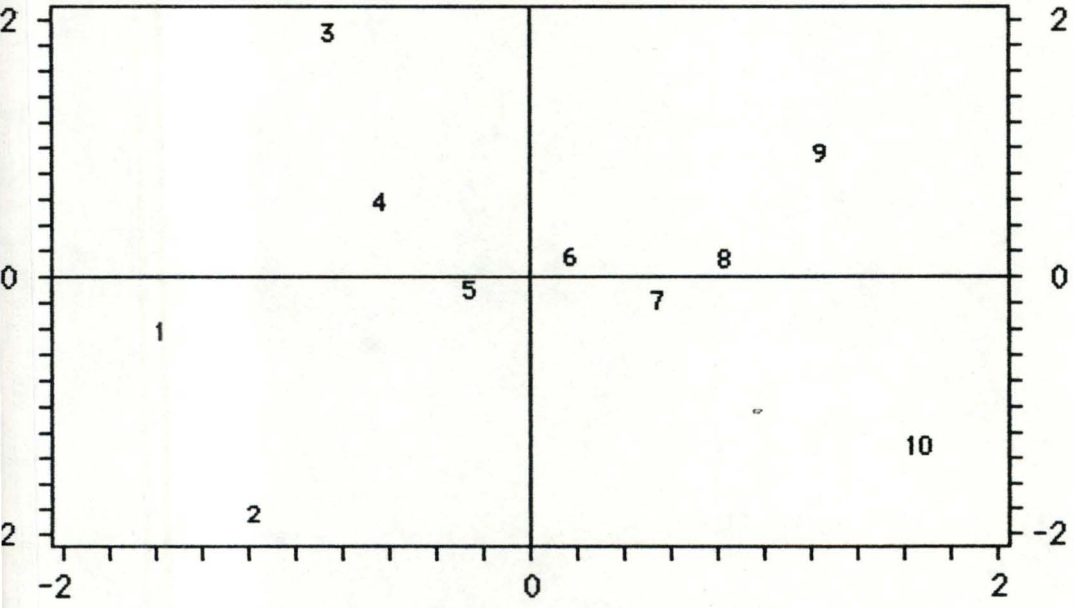
colonnes dans le plan des composantes 1 & 2

AVEC VARIABILITÉ



AFC

lignes dans le plan des axes factoriels 1 & 2



colonnes dans le plan des axes factoriels 1 & 2

AVEC VARIABILITÉ

14.6. Qui se ressemble s'assemble.

Voici un cas très particulier où l'AFC montre qu'elle est beaucoup plus nuancée que l'ACP, et que son interprétation nécessite une certaine pratique pour en dégager toute l'information.

14.6.1. *Natures des distributions.*

Les paramètres servant à simuler les distributions sont :

Esp.	A. M.	St. max.	Coef.
1	1000	1	20
2	100	1	200
3	1000	2	20
4	100	2	200
5	1000	3	20
6	100	3	200
7	1000	4	20
8	100	4	200
9	1000	5	20
10	100	5	200

14.6.2. *Projections des lignes.*

La répartition est identique pour les deux analyses.

14.6.3. *Projections des colonnes.*

L'ACP va reproduire le gradient en confondant les espèces ayant leur maxima aux mêmes stations.

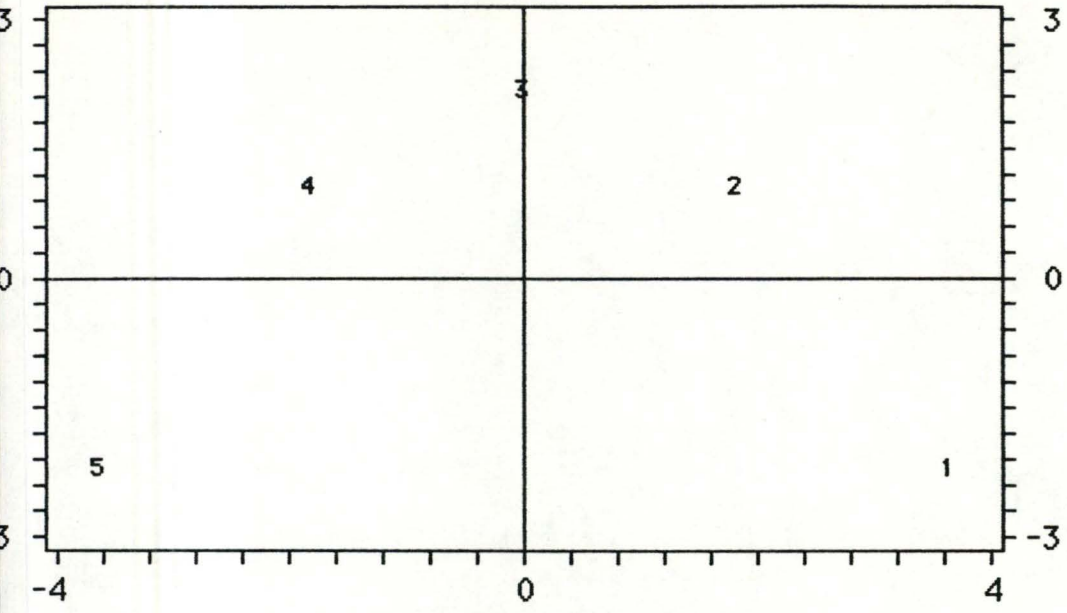
L'AFC sépare les espèces en deux groupes : d'une part, les espèces à variations importantes (nombres impairs) et, d'autre part, les espèces à variations faibles (nombres pairs). Le gradient est détectable, mais dans les deux groupes séparément. Puisque l'on connaît la répartition des stations le long du gradient, il est facile de déterminer la station du maximum.

L'intérêt, c'est que l'AFC montre qu'il y a plus de différences entre les espèces à coefficients d'étalement faibles, qu'il n'y en a entre les espèces à coefficient d'étalement élevés.

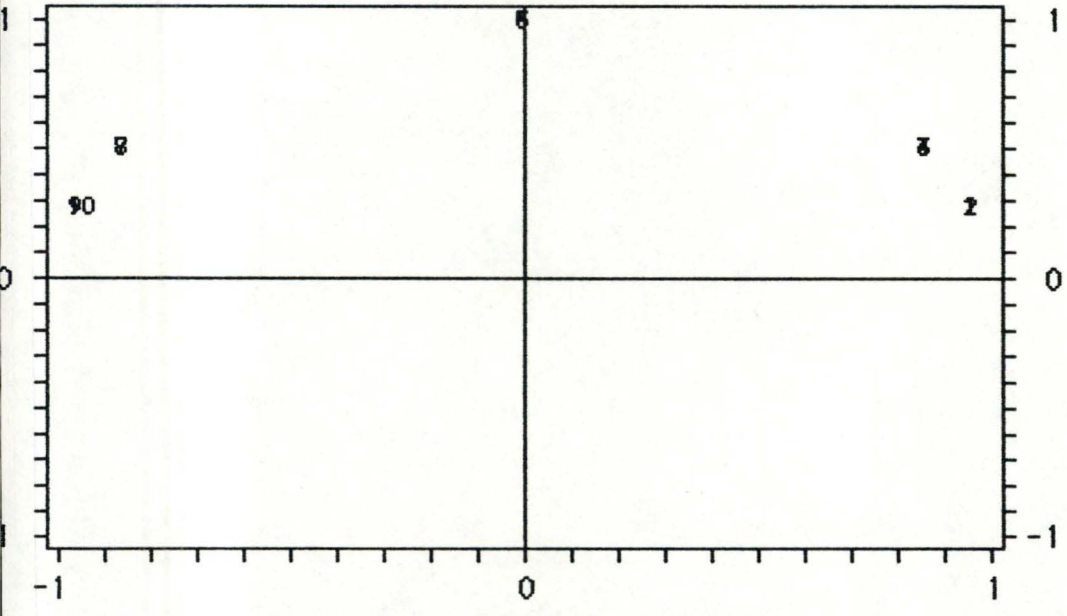
14.6.4. *Conclusion.*

On reconnait ici la finesse des Français face à l'efficacité des Anglo-Saxons. Tout est dans la nuance !

ACP



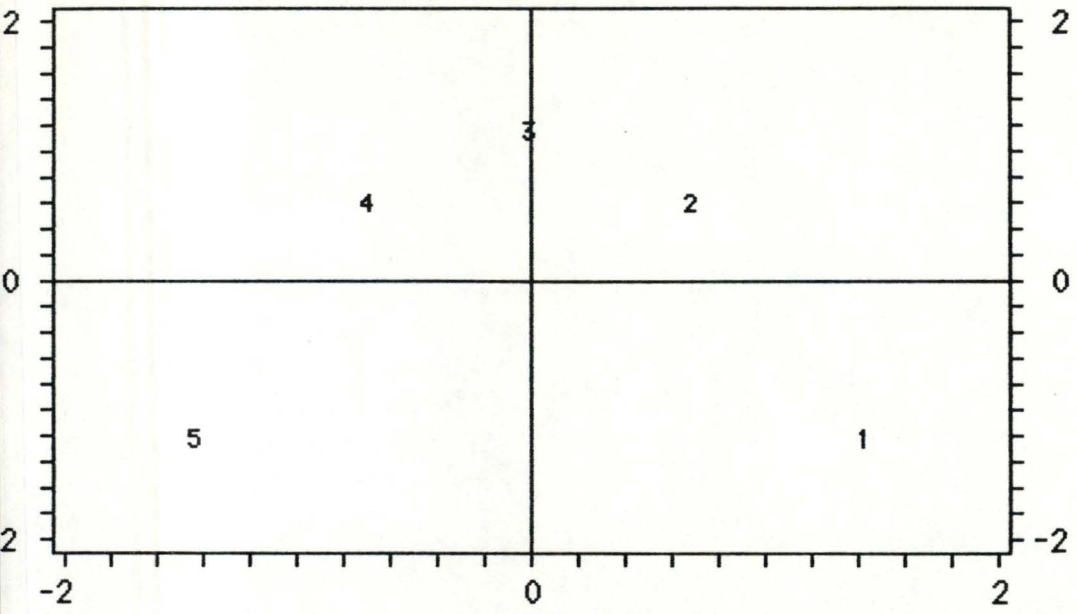
lignes dans le plan des composantes 1 & 2



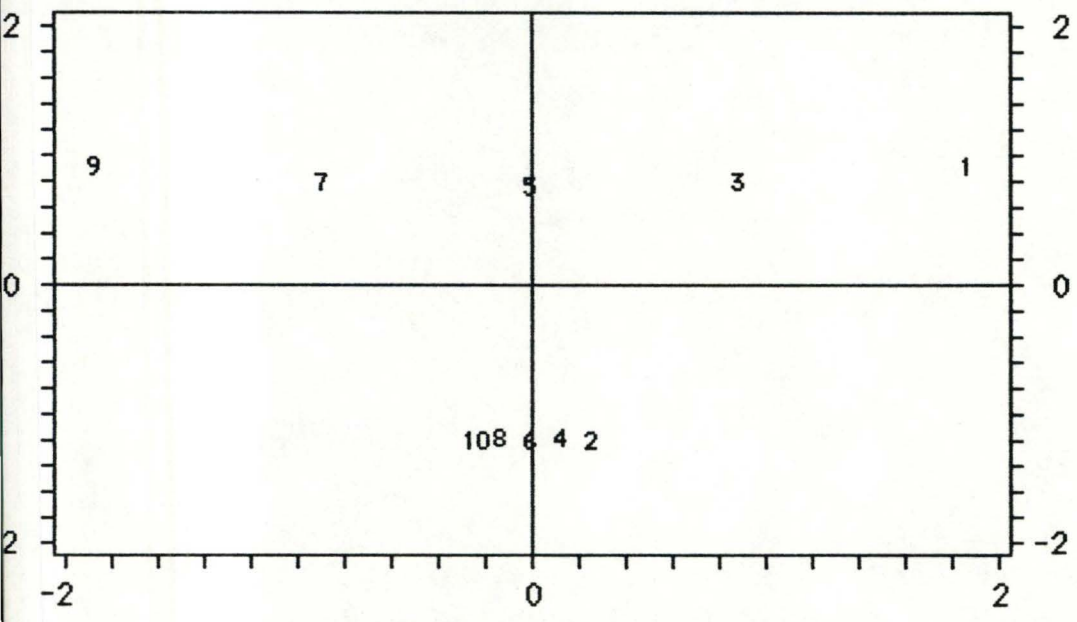
colonnes dans le plan des composantes 1 & 2

SANS VARIABILITÉ

AFC



lignes dans le plan des axes factoriels 1 & 2



colonnes dans le plan des axes factoriels 1 & 2

SANS VARIABILITÉ

14.7. Intérêt de la représentation simultanée.

C'est évidemment de l'AFC dont on va parler.

Ce type de représentation permet d'associer plus facilement une espèce à une station, c'est-à-dire à un type de milieu particulier. Dès lors, une question se pose : comment savoir quelle espèce est la plus caractéristique d'un milieu déterminé, dans le but évident de découvrir des bioindicateurs.

Il ne s'agit pas pour moi de développer une méthodologie nouvelle, mon expérience dans ce domaine étant trop limitée, mais de faire quelques suggestions.

14.7.1. *Exemple N°1 : natures des distributions.*

Les paramètres servant à simuler les distributions sont :

Esp.	A. M.	St. max.	Coef.
1 à 3	1000	2 à 4	10
4	1000	5	50
5	1000	5	10
6	1000	5	5
7 à 10	1000	6 à 9	10

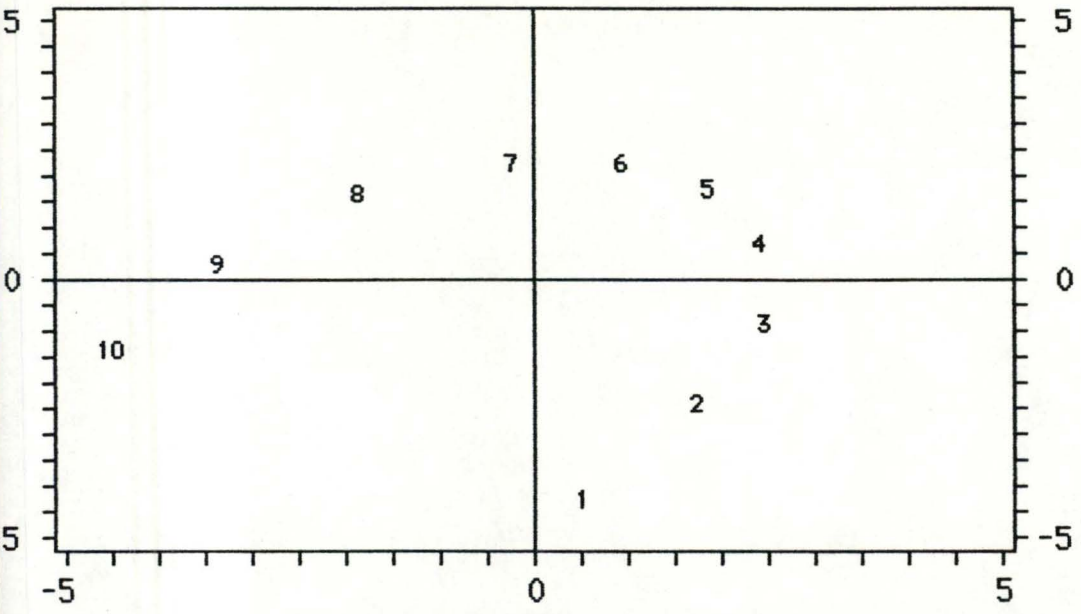
Les trois espèces concernées sont associées à une station située au milieu du gradient, endroit stratégique pour les raisons décrites plus haut, et seul le coefficient d'étalement permet de les distinguer.

14.7.2. *Association espèce-station.*

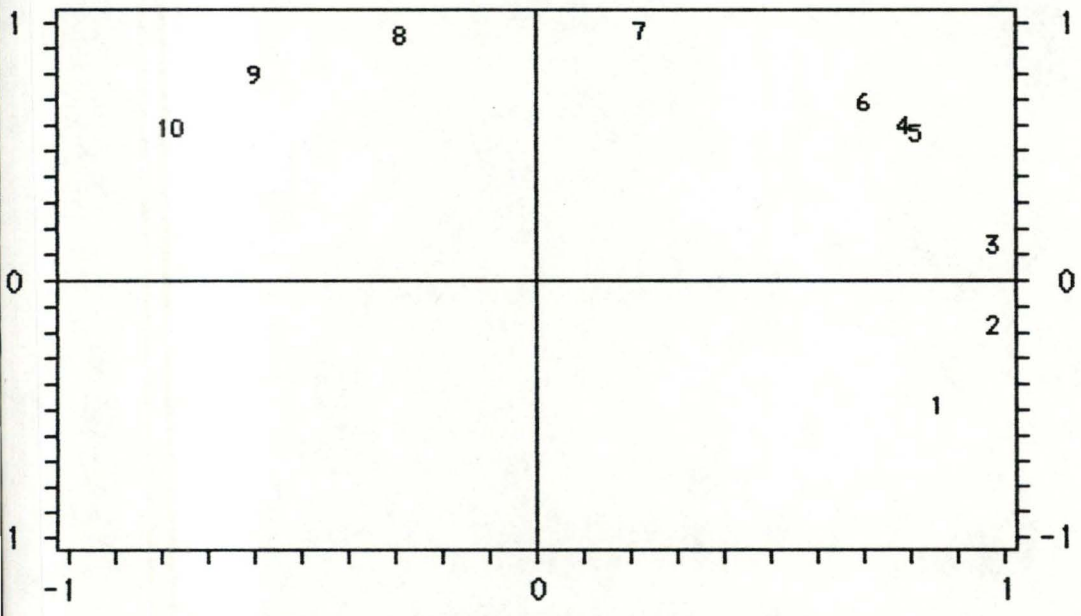
Quelle est l'espèce la plus caractéristique de la station 5 ?

A priori, ce doit être l'espèce s'éloignant le plus de l'origine des axes dans la direction déterminée par le vecteur joignant le centre et le point représentant la station 5. Si l'on suit ce raisonnement, il s'agit de l'espèce 6, qui possède le plus faible coefficient d'étalement. C'est aussi celle qui répond le mieux à la définition d'un bioindicateur.

ACP

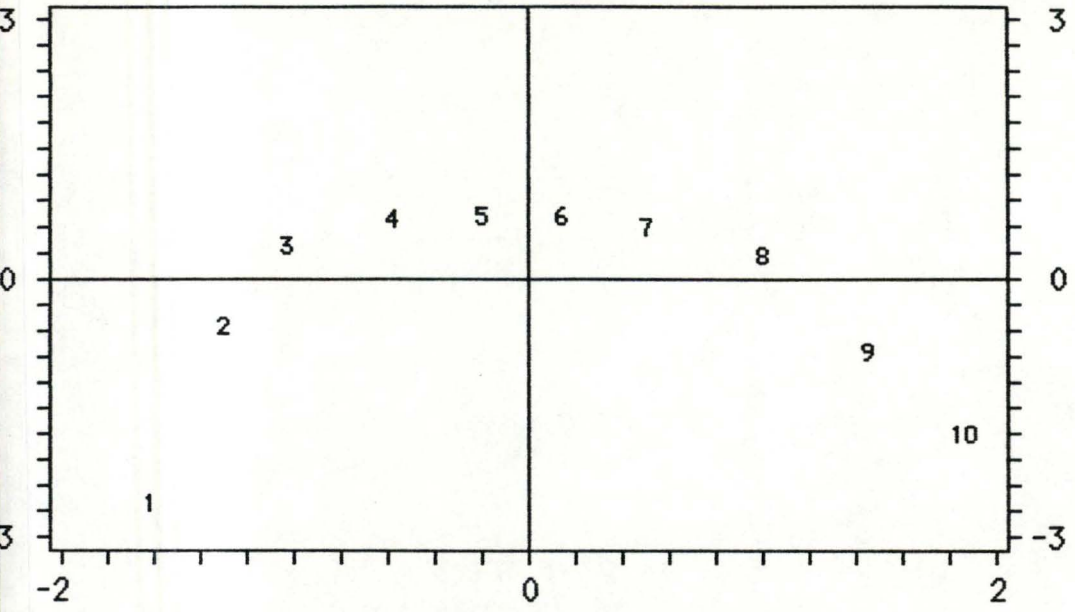


lignes dans le plan des composantes 1 & 2

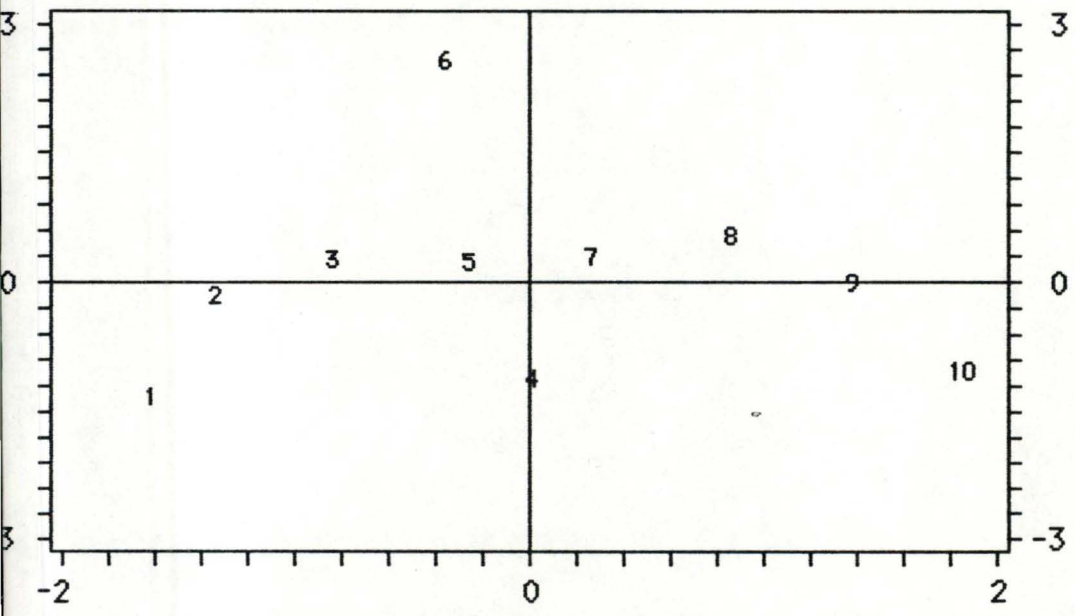


colonnes dans le plan des composantes 1 & 2

AFC



lignes dans le plan des axes factoriels 1 & 2



colonnes dans le plan des axes factoriels 1 & 2

14.7.3. *Exemple N°2 : natures des distributions.*

Considérons maintenant le cas où seule l'abondance maximum différencie les trois espèces.

Les paramètres servant à simuler les distributions sont :

Esp.	A. M.	St. max.	Coef.
1 à 3	1000	2 à 4	10
4	1000	5	10
5	100	5	10
6	10	5	10
7 à 10	1000	6 à 9	10

14.7.4. *Association espèce-station.*

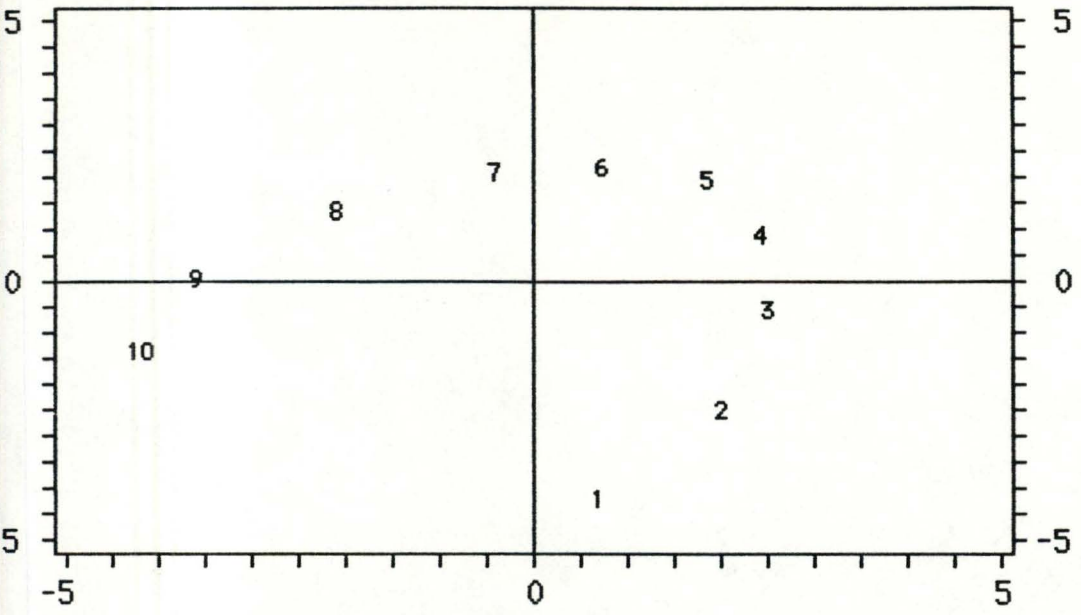
L'espèce 6 se distingue toujours du lot. L'algorithme utilisé pour générer les données a fait que l'espèce 6 possède en réalité un coefficient d'étalement inférieur à 10. Cependant, les caractéristiques de sa distribution indiquent une espèce proche de la définition d'un bioindicateur.

D'autres exemples réalisés en modifiant la station concernée, montrent que l'espèce la plus caractéristique d'une station est habituellement celle qui possède le coefficient d'étalement le plus faible.

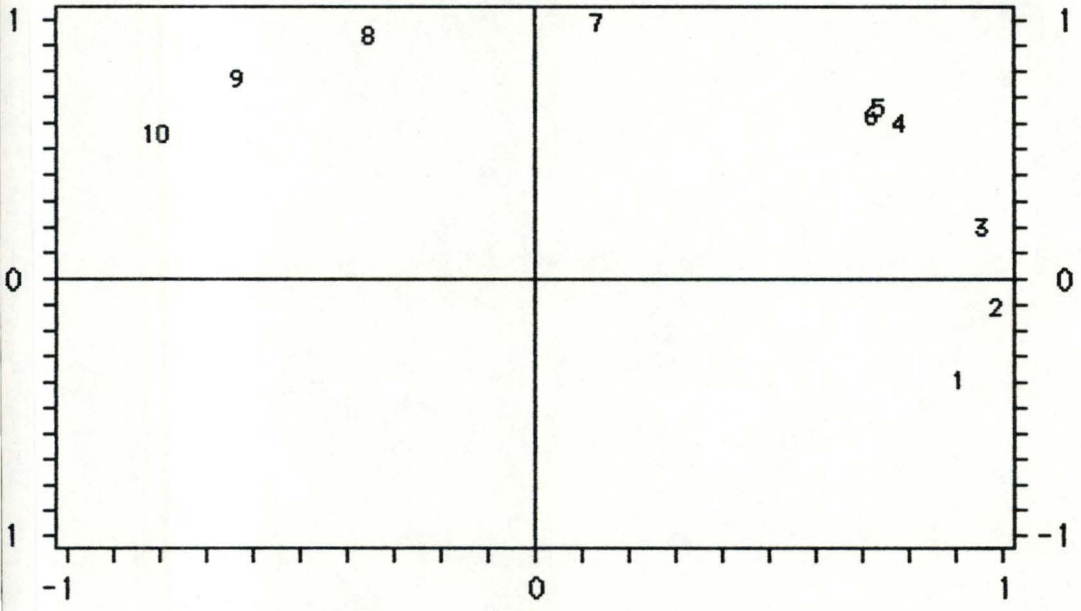
14.7.5. *Conclusion.*

Il s'agit de l'une des nombreuses facettes de l'AFC qui montre que cette dernière n'est pas avare de renseignements au sujet des données qu'on lui propose.

ACP

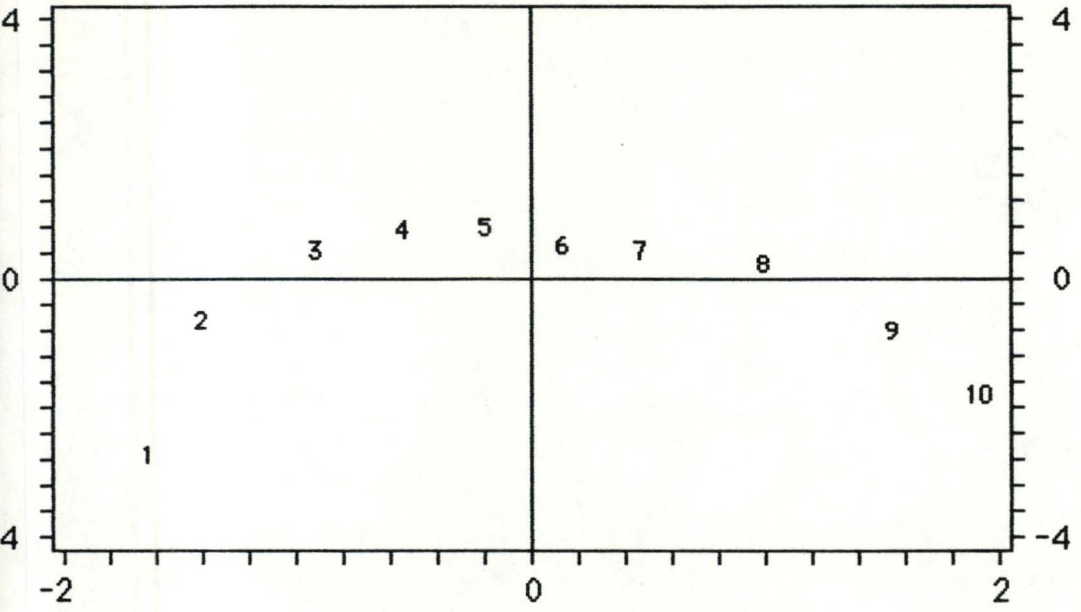


lignes dans le plan des composantes 1 & 2

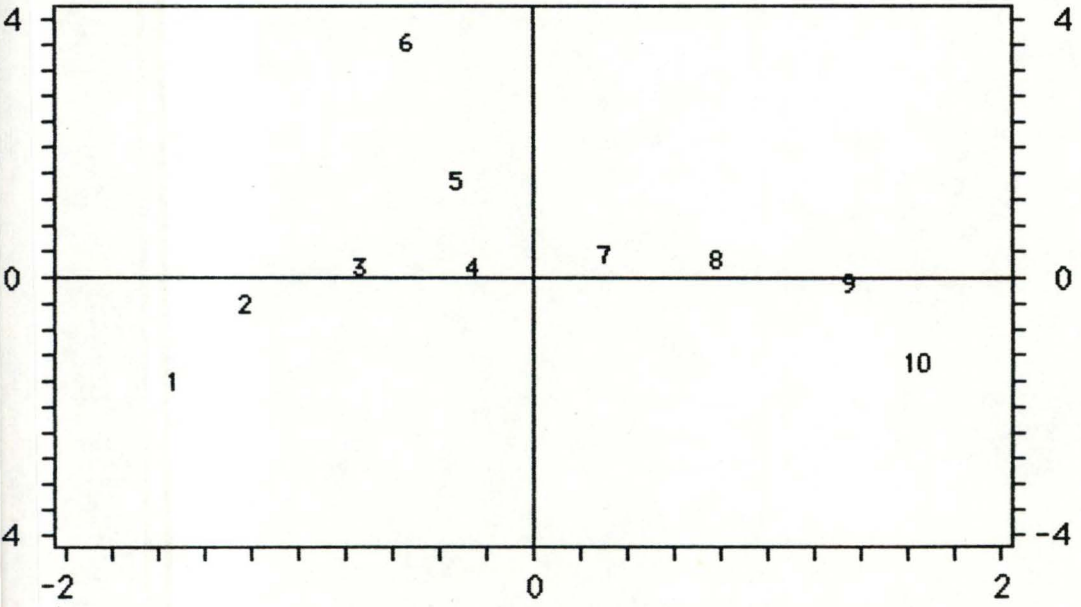


colonnes dans le plan des composantes 1 & 2

AFC



lignes dans le plan des axes factoriels 1 & 2



colonnes dans le plan des axes factoriels 1 & 2

15. Conclusions et commentaires.

Si l'AFC se révèle souvent un meilleur choix que l'ACP lorsque l'on utilise des données simulées, il est difficile de définir à priori laquelle des deux méthodes est préférable dans le cas de données réelles. Les grandes dimensions et la présence d'une forte variabilité aléatoire contribuent ensemble au rapprochement des résultats produits par les deux analyses.

Une façon de résoudre le dilemne est de réaliser les deux analyses conjointement, et avec une méthode d'ordination directe telle que le cluster analysis afin d'obtenir des renseignements complémentaires. De plus le développement du calcul électronique et la présence sur le marché de nombreux logiciels, font que les obstacles du temps et des moyens à mettre en oeuvre se réduisent rapidement.

Une solution définitive est, peut-être, de rechercher une troisième voie pour l'analyse de ce type de tableaux. Hill (1980) propose une variante de l'analyse des correspondances : la DCA, detrented correspondence analysis, qui permet, notamment, de supprimer la dépendance curvilinéaire et qui s'avère supérieure à son aînée dans certaines conditions.

Quoiqu'il en soit, les résultats obtenus me permettent d'établir un tableau de comparaisons qui, je l'espère, et malgré sa valeur relative, pourrait guider le chercheur dans ses choix méthodologiques.

Tableau comparatif en dix points, ACP-AFC.

	ACP	AFC
Difficultés méthodologiques :	+	-
Détection d'un gradient :	-	+
Indépendance des facteurs :	-	+
Déformations graphiques :	-	+
Poids des colonnes :	+	-
Poids des lignes :	-	+
Variabilité résiduelle :	-	+
Différenciation des taxons :	-	+
Association taxon-station :	absence	+
Richesse des résultats :	-	+
Tendance générale :	-	+

+ : meilleur comportement ; - : moins bon comportement.

16. **Petit dictionnaire des synonymes.**

Ce petit dictionnaire reprend quelques-uns des mots utilisés dans les pages précédentes et donnent leurs synonymes les plus fréquents.

Analyse des composantes principales : principal components analysis (PCA).

Analyse factorielle des correspondances : - reciprocal averaging (RA),
- correspondence analysis (CA).

Colonne : - descripteur,
- taxon,
- variable.

Ecart-type : déviation standard.

Ligne : - objet,
- prélèvement,
- station s'il n'y a qu'un seul réplicat,
- échantillon ou réplicat si on étudie plusieurs échantillons équivalents.

Matrice des composantes initiales transformées : - matrice des coordonnées des points dans l'espace factoriel.

Point moyen : barycentre,

Profil moyen : profil marginal,

Trace : - inertie totale,
- variance totale.

17. **A propos d'un logiciel.**

Tous les résultats d'analyses présentés au cours de ces pages ont été obtenus grâce au logiciel ANAMUL.

Il s'agit d'un programme original d'analyse multivariée, adaptés aux besoins propres de cette étude.

17.1. Spécifications.

Il tourne sur Macintosh®, écrit en MS BASIC 2.0®, et sur AMSTRAD "JOYCE" PCW8512®, écrit en CBASIC Compiler®.

Ses principales qualités sont :

1) la portabilité,
2) le nombre réduit de manipulations,
3) le caractère complet des résultats fournis.

Ses principaux défauts sont :

1) la grande place occupée en mémoire par les variables,
2) la faible vitesse d'exécution,
3) la lenteur de l'impression des résultats.

17.2. Caractéristiques.

17.2.1. *Nature des données.*

On peut utiliser des données originales, rentrées "à la main", avec possibilité de contrôle des valeurs.

Il peut simuler les distributions de diverses espèces le long d'un gradient, avec ou sans variabilité aléatoire.

On peut récupérer un tableau de données créé à l'aide d'un tableur.

17.2.2. *Traitement des données.*

Il permet le choix entre :
- une ACP seule,
- une AFC seule,
- une ACP et une AFC.

17.2.3. *Résultats.*

Les résultats proposés sont :

a) les valeurs propres, avec les pourcentages relatifs et cumulés,
b) la trace,
c) la mesure du CHI-2, pour l'AFC,
d) les poids des lignes et des colonnes, en %,
e) les contributions à l'analyses,
f) les contributions absolues,
g) les contributions relatives,
h) les coordonnées des points dans l'espace factoriel,
i) les graphiques des projections,
j) les graphiques des cosinus carrés,

Tout est acheminé automatiquement vers l'imprimante.

De plus, il est possible de récupérer certains des résultats intermédiaires (zscores, corrélations, valeurs intermédiaires de l'AFC, variances-coavariances), en les "collant" dans un tableur.

18. Bibliographies.

18.1. Bibliographie sélective, proposée par M. Volle (1985).

Bastin C. & al., 1980. Analyse des données : abrégé théorique ; étude de cas modèles. Dunod.

Benzécri J.-P. & al., 1973. Analyse des données. T2. Analyse des correspondances. Dunod.

Benzécri J.-P., 1976. Histoire et préhistoire de l'analyse des données. Cahiers de l'analyse des données. Dunod.

Benzécri J.-P. et F., 1980. Analyse des correspondances : exposé élémentaire. Dunod

Bertier P. & Bouroche J.-M., 1977. Analyse des données multi-dimensionelles. PUF, 2ème édition.

Caillez F. & Pages J.-P., 1976. Introduction à l'analyse des données. SMASH.

Fénélon J.-P., 1981. Qu'est-ce que l'analyse des données. Lefonen.

Lebart L., Morineau A. & Fénélon J.-P., 1979. Traitement des données statistiques. Dunod.

Lebart L., Morineau A. & Tabard N., 1977. Techniques de la description statistique. Dunod.

Saporta G., 1978. Théories et méthodes de la statistique. Technip.

18.2. Bibliographie personnelle : livres.

Cornet C., 1986. Contribution a l'étude de l'évolution paléoécologique de la fin du quaternaire dans les Vosges et l'Eifel, d'après les diatomées d'eau douce. Thèse de doctorat en sciences géologiques, U.C.L., Louvain-La-Neuve, 176 p.

Dagnélie P., 1975. Analyse statistique a plusieurs variables. Presses Agronomiques de Gembloux, 362 p.

Dansart-Jacques B., 1986. Etude de la faune benthique de la haute-Semois et de l'impact d'une station d'épuration sur sa restauration. Thèse de doctorat en sciences de l'environnement, F.N.D.P., Namur, 191 p.

Depiereux E., 1982. Utilisation critique de l'analyse en composantes principales et du cluster analysis pour la description d'échantillons d'invertébrés benthiques en eau douce. Etude de la répartition des invertébrés benthiques sur l'Ourthe et la Lesse moyenne et leurs affluents en relation avec la qualité de l'eau. Thèse de doctorat en sciences biologiques, F.N.D.P., Namur, 218 p.

Dwight H.B., 1961 (4ème édition). Tables of integrals and other mathematical data., Macmillan Publishing Co., Inc., New-York, 336 p.

Lebart L., Morineau A. & Fénelon J.-P., 1979. Traitement des données statistiques. Méthodes et programmes. Bordas, Paris, 510 p.

Legendre L. & Legendre P., 1979. Ecologie numérique. 1. Le traitement multiple des données écologiques. Masson, Paris et les Presses de l'Université du Québec, 197 p.

Legendre L. & Legendre P., 1979. Ecologie numérique. 2. La structure des données écologiques. Masson, Paris et les Presses de l'Université du Québec, 247 p.

Legras J., 1971. Méthodes et techniques de l'analyse numérique. Dunod, Paris, 321 p.

Lortal R., 1978. Programmation conversationnelle. Basic. Analyse numérique. Masson, Paris, 120 p.

Nowakowski Cl., 1984. Méthodes de calcul numérique. Programmes en basic et en Pascal. Guide pratique. P.S.I., Lagny, T1 157 p., T2 183 p.

Orban F., 1987. Contribution à l'étude du secteur tertiaire dans les principales villes belges. Structure des effectifs, évolution de 1972 à 1982. Thèse de doctorat en sciences géographiques, Université de Liège, 361 p. .

Reverchon A. & Ducamp M., 1985. Mathématiques sur micro-ordinateur. 2. algèbre. Eyrolles, Paris, 249 p.

Torrens-Isbern J., 1972. Modèles et méthodes de l'analyse factorielle. Dunod, Paris, 202 p.

Volle M., 1985. Analyse des données, 3e édition. Collection "Economie et statistiques avancées". Série : Ecole Nationale de la Statistique et de l'Administration économique et Centre d'études des Programmes Economiques. Economica. Paris. 323 p.

Whittaker R.H., 1973. Handbook of vegetation science. Part V : ordination and classification of communities. Ed by R.H. Whittaker. Dr. W. JUNK b.v.-Publishers-The Hague. Article 11, p. 287 à 321.

18.3. Bibliographie personnelle : articles.

Austin M. P. & Greig-Smith P., 1968. The application of quantitative methods to vegetation survey. II. Some methodological problems of data from rain forest. Journal of ecology, N°56, p. 827-844

Austin M. P. & Noy-Meir I., 1971. The problem of non-linearity in ordination : experiments with two-gradient models. Journal of ecology, N°59, p. 763-773.

Austin M. P. & Orloci L., 1966. Geometric models in ecology. II. An evaluation of some ordination techniques. Journal of ecology, N°54, p. 217-227.

Bannister P., 1968. An evaluation of some procedures used in simple ordinations. Journal of ecology, N°56, p. 27-34.

Depiereux E. & Feytmans E., 1985. Modification progressive de la structure des peuplements d'invertébrés benthiques en fonction de la qualité de l'eau de l'Ourthe et de la Lesse (Meuse belge). Acta oecologica/oecol. appl., N°6, p.81-98.

Depiereux E., Feytmans E. & Micha J.C., 1983. Utilisation critique de l'analyse en composantes principales et du cluster analysis pour la description d'échantillons d'invertébrés benthiques en eau douce. OIKOS, N°40, p. 81-94.

Eppink Th. W. A., 1984. Correspondence analysis versus component analysis for highly skewed distributed variables. Computational statistics quarterly, Vol. 1, N°1, p. 61-76.

Fasham M. J. R., 1977. A comparison of nonmetric multidimensional scaling, principal components and reciprocal averaging for the ordination of simulated coenoclines and coenoplanes. Ecology, N°58, p. 551-561.

Gauch H. G. & al., 1977. A comparative study of reciprocal averaging and other ordination techniques. Journal of ecology, N°65, p. 157-174.

Gauch H. G. & Whittaker R. H., 1972. Comparison of ordination techniques. Ecology, N°53, p. 868-875.

Gauch H. G., Whittaker R. H. & Singer S. B., 1981. A comparative study of nonmetric ordinations. Journal of ecology, N°69, p. 135-152.

Hill M. O., 1973. Reciprocal averaging : an eigenvector method of ordination. Journal of ecology, N°61, p. 237-249.

Hill M. O., 1974. Correspondence analysis, a neglected multivariate method. Applied statistics, Vol. 23, N°3, p. 340-354.

Hill M. O. & Gauch H. G., 1980. Detrended correspondance analysis : an improved ordination technique. Vegetatio, N°42, p. 47-58.

Ibanez F. & Seguin G., 1972. Etude du cycle annuel du zooplancton d'Abidjan. Comparaison de plusieurs méthodes d'analyse multivariable : composantes principales, correspondances, coordonnées principales. Investigacion pesquera, Vol. 36, N°1, p. 81-108.

Kessel S. R. & Whittaker R. H., 1976. Comparisons of three ordination techniques. Vegetatio, Vol. 32, N°1, p. 21-29.

Noy-Meir & Austin M. P., 1970. Principal component ordination and simulated vegetational data. Ecology, N°51, p. 551-552.

Whittaker R. H., 1967. Gradient analysis vegetation. Biological reviews of the Cambridge philos, N°42, p. 207-264.