Généralisation de l'analyse factorielle multiple à l'étude des tableaux de fréquence et comparaison avec l'analyse canonique des correspondances

Lila Abdessemed, Brigitte Escofier

INRIA

UNITÉ DE RECHERCHE
INRIA-RENNES

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P.105
78153 Le Chesnay Cedex
France
Tél. (1) 39 63 55 11

Rapports de Recherche

1 9 9 2

25*ème anniversaire*

N° 1820

Programme 4
Robotique, Image et vision

GENERALISATION DE L'ANALYSE FACTORIELLE MULTIPLE A L'ETUDE DES TABLEAUX DE FREQUENCE ET COMPARAISON AVEC L'ANALYSE CANONIQUE DES CORRESPONDANCES

Lila ABDESSEMED
Brigitte ESCOFIER

Décembre 1992

RR-1820

IRISA

INSTITUT DE RECHERCHE EN INFORMATIQUE
ET SYSTEMES ALEATOIRES

Campus Universitaire de Beaulieu
35042 - RENNES CEDEX FRANCE
Tél. : 99 84 71 00 - Télex : UNIRISA 950 473 F
Télécopie : 99 38 38 32

Généralisation de l'analyse factorielle multiple à l'étude des tableaux de fréquence et comparaison avec l'analyse canonique des correspondances

Publication Interne n°688 - Novembre 1992, 34 pages

Programme 4

Lila Abdessemed, Brigitte Escofier

Résumé

L'analyse factorielle multiple est une technique qui permet seulement d'étudier de façon simultanée des groupes de variables numériques et/ou qualitatives. Nous proposons ici, une méthode qui permet également l'étude des tableaux de fréquence.

Nous présentons cette méthode dans le cas d'un seul tableau de fréquence et d'un tableau de variables numériques. Celle-ci consiste en une transformation du tableau de fréquence et l'introduction de pondérations adéquates sur les lignes et les colonnes.

Nous illustrons cette méthode par un exemple d'application sur des données écologiques. Nous comparons cette méthode avec l'analyse canonique des correspondances, et nous discutons le cas de plusieurs tableaux de fréquence.

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE (U.R.A 227) UNIVERSITÉ DE RENNES I INSA DE RENNES
INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE (UNITÉ DE RECHERCHE DE RENNES)

Multiple Factor Analysis generalization to two-way contingency tables and comparison with Canonical Correspondence Analysis

Abstract

Multiple Factor Analysis is a method which only allows to analyze together groups of numerical variables and/or groups of qualitative variables. We propose here a method wich allows to analyze simultaneously two-way contingency tables, quantitative and qualitative tables.

We present this method in the case of one contingency table only and one group of numerical variables. It consists of the transformation of the frequency table and the introduction of metrics on rows and columns of the two groups.

A comparison is made with the Canonical Correspondence Analysis, and the case of multiple two-way contingency tables is studied.

1 Introduction

L'étude de la liaison entre plusieurs groupes de variables a toujours été et est toujours un sujet d'actualité, qui a eu des éléments de réponse selon la problématique d'une part, et selon la nature des variables.

Nous nous intéressons dans ce cadre, à l'étude de la liaison entre les structures induites par un tableau de fréquence et par un tableau de variables descriptives, ces dernières pouvant être de type numérique, ou qualitatif.

Parmi les méthodes qui permettent d'étudier de façon simultanée des groupes de variables de nature différente, l'Analyse Factorielle Multiple (AFM) [3] peut apporter des réponses à notre problème, mais celle-ci ne s'applique qu'à des groupes de variables numériques ou qualitatives et ne prend pas en compte des informations de type profils qui caractérisent les tableaux de fréquence.

Dans cette étude, nous nous proposons de généraliser l'AFM à l'étude des tableaux de fréquence.

Par ailleurs, l'Analyse Canonique des Correspondances (ACC) étant une technique permettant d'étudier la liaison entre les structures induites par un tableau de fréquence, et par un tableau de variables numériques, nous la comparons avec la méthode que nous proposons.

Nous présentons notre méthode d'abord dans le cas d'un tableau descriptif numérique et d'un tableau de fréquence. Nous envisageons ultérieurement une généralisation pour le cas de plusieurs tableaux de variables numériques et/ou qualitatives et d'un tableau de fréquence. Enfin, nous examinons le cas de plusieurs tableaux de fréquence.

2 Les notations et les données

Notations

Nous considérons un tableau de fréquence F de terme général f_{ij}, ayant I lignes et J colonnes.
Ses marges sur I et sur J sont notées respectivement $f_{i.}$ et $f_{.j}$.
Par ailleurs, nous considérons un tableau de variables descriptives X de terme général x_{ik}, décrivant les I individus à l'aide de K variables numériques.

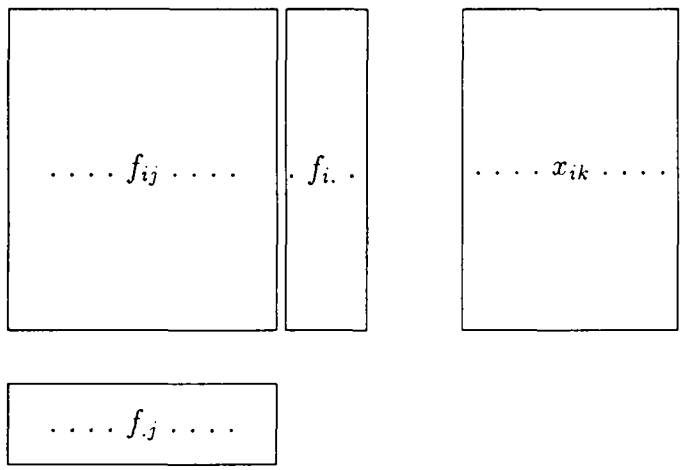

Table 1 : Tableau de fréquence F Tableau descriptif X

Les données

Pour notre étude, nous avons repris les données écologiques illustratives de l'ACC présentées dans [7].

Ces données se présentent sous la forme d'un tableau F=(f_{ij}) indiquant la présence ou l'absence de 22 espèces de plantes sur 12 sites, qui sont par ailleurs décrits par 4 variables d'environnement dans le tableau X=(x_{ik}). Le tableau de présence-absence codé en 0/1 est donc considéré comme un tableau de fréquence.

Ces variables sont :

- l'indice de salinité (g/m^2)
- la distance à la mer (m)

- la profondeur du sol (cm)
- la pente (degré)

Nous les notons respectivement SALI, DIST, SOIL, SLOP.
On trouvera en annexe l'intitulé et les noms des 22 espèces de plantes.

	Espèces	Variables descriptives

Sites

1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
0	1	1	1	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
0	0	1	1	1	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0
0	0	1	1	1	1	0	1	1	0	1	0	1	0	0	0	0	0	0	0	0	0
0	1	1	1	1	0	0	1	1	1	1	0	1	1	0	0	0	0	0	0	0	0
0	1	1	0	1	0	0	1	1	1	1	0	1	1	1	1	1	0	0	0	0	0
0	0	1	0	0	1	0	1	1	1	0	0	1	1	1	1	0	1	1	0	0	0
0	0	1	1	0	0	0	0	0	1	1	0	0	0	1	0	0	0	1	1	0	0
0	0	0	1	0	0	0	1	1	1	1	0	1	0	1	1	0	0	1	0	1	1

sali	dist	soil	slop
1483	23	3	30
777	29	2	20
298	35	11	25
875	57	1	25
509	57	11	20
542	60	4	20
382	62	12	20
272	66	6	27
214	80	10	18
105	85	12	10
148	100	17	5
62	123	16	15

Table 2 : Tableau de fréquence F Tableau descriptif X

3 Les objectifs

A travers cette étude, nous cherchons à mettre en évidence des relations entre la fréquence des espèces sur les sites et les valeurs des variables descriptives et à les évaluer.

Aussi, notre premier objectif consiste à voir si les structures induites sur les sites par les deux tableaux F et X sont liées, en d'autres termes, s'il existe des directions de dispersion communes pour les deux structures. On cherche alors, à mesurer leur importance pour chacun des groupes.

On étudie aussi les relations entre les espèces caractérisées par les fréquences sur les différents sites, ainsi que les corrélations entre les variables descriptives. On cherche surtout à mettre en relief, si elle existe une relation spécifique espèce-variable.

L'objectif suivant est de trouver les sites qui se ressemblent de par leur répartition sur les espèces et par les valeurs des variables descriptives, et on veut détecter ceux qui sont proches seulement au sens de la répartition des espèces ou ceux qui le sont uniquement au sens des variables environnementales. On cherche alors, une représentation simultanée des sites au sens de la répartition "espèces et variables", et au sens des deux autres points de vue.

Dans l'étude de la liaison entre un tableau de fréquence et un groupe de variables, ces objectifs sont donc exactement les mêmes que ceux assignés dans l'AFM, pour l'analyse de la liaison entre plusieurs groupes de variables numériques et/ou qualitatives.

4 Technique de généralisation

Un "vrai" tableau de fréquence est un tableau croisant deux variables qualitatives. Il est communément traité par l'Analyse Factorielle des Correspondances (AFC) qui permet d'étudier l'écart à l'hypothèse d'indépendance. En AFC, lignes et colonnes sont traitées avec une complète symétrie ; elles sont représentées par leurs profils avec la distance du χ^2.

Pour étudier conjointement le tableau de fréquence et le tableau de variables par l'AFM, nous voulons conserver le point de vue de l'AFC pour le tableau de fréquence

et donc conserver les distances entre lignes ou colonnes, pour pouvoir en restituer la structure .

D'autre part, l'AFM est une Analyse en Composantes Principales (ACP) particulière. La technique de généralisation sera alors basée sur l'équivalence entre l'AFC du tableau de fréquence F et l'ACP non normée d'un tableau transformé F', ses lignes et ses colonnes étant munies de métriques adéquates.

En effet, il est équivalent de réaliser une AFC du tableau de terme général f_{ij} ou une ACP non normée du tableau transformé de terme général $f_{ij}/f_{i.}f_{.j}$, les individus étant munis des poids $f_{i.}$ et les variables des poids $f_{.j}$.

De ce fait, en considérant au lieu du tableau de fréquence F, le tableau F' centré de terme général $(f_{ij}/f_{i.}f_{.j} - 1)$ comme un groupe de variables numériques, nous sommes ramenés à la situation classique de l'AFM, avec deux groupes de variables numériques (F' et X).

Cependant, l'aspect symétrique de l'AFC ne pourra être restitué dans l'AFM, puisqu'en ACP la notion d'individu est complètement différenciée de celle de variable et la notion de ressemblance entre les variables dans l'AFC est complètement différente de celle définie dans l'ACP : la ressemblance entre espèces est définie dans le tableau de fréquence par la distance entre les profils au sens du χ^2, alors que pour les variables dans l'ACP, elle s'exprime en terme de liaison, à travers leur corrélation.

Partant du résultat précédent, on applique donc l'AFM au tableau (table 3) composé de deux groupes de variables numériques définis par F' et X. Le tableau F' est $f_{i.}$ centré et a pour terme général $(f_{ij}/f_{i.}f_{.j} - 1)$ et le tableau X est centré et réduit pour les poids $f_{i.}$. Nous notons $\bar{x}_k$ et s_k la moyenne et l'écart-type de la variable k calculés avec les poids $f_{i.}$. Le terme général du tableau X est $(x_{ik} - \bar{x}_k)/s_k$.

Les sites i sont munis des poids $f_{i.}$ et les espèces j ont le poids $f_{.j}$. Nous notons D=diag($f_{i.}$) et M=diag($f_{.j}$) les métriques induites par les poids $f_{i.}$ et $f_{.j}$ sur les espaces R^I et R^J dans lesquels sont situés respectivement les variables et les sites .

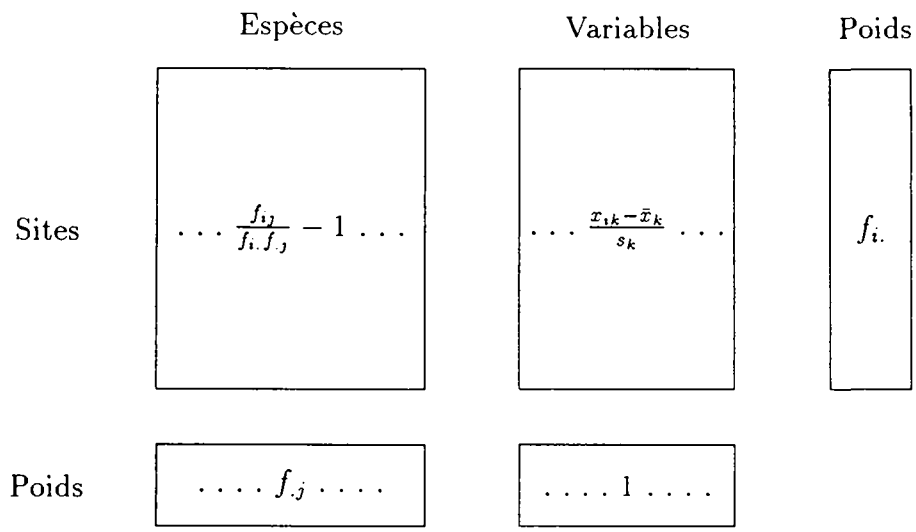

Table 3 : Le tableau de fréquence F' et le tableau descriptif X centrés pour les poids $f_{i.}$

Centrage et réduction

Les deux nuages de sites associés aux tableaux F' et X sont donc centrés, ce qui élimine l'influence de la position des barycentres et ne prend en compte que la forme des nuages. Par ailleurs, cela permet d'interpréter le cosinus entre deux vecteurs représentant les variables de X comme un coefficient de corrélation. Celui-ci est toujours calculé avec la métrique D.

La réduction de X est due à la situation classique des variables exprimées dans des unités différentes, alors que la non réduction de F est nécessaire pour conserver l'équivalence entre l'ACP de F' et l'AFC de F.

Les variables descriptives peuvent avoir un poids quelconque, selon l'importance qu'on veut leur attribuer. Dans notre exemple, nous avons opté pour un poids uniforme égal à 1.

5 Influence des pondérations D et M sur les nuages d'individus et de variables associés à chacun des tableaux F' et X

L'AFM de groupes de variables numériques se fait en deux étapes. La première est une ACP de chacun des groupes de variables pris séparément. De ce fait, nous étudions l'influence des métriques introduites sur les différents nuages associés à chacun des deux tableaux F' et X.

Le tableau F'

Pour le tableau de fréquence transformé F', les métriques M et D n'ont d'autre effet que la conservation de la structure AFC du tableau de fréquence initial F.
Elles permettent la restitution complète de l'information contenue dans le tableau F (distances du χ^2 entre les profils des lignes et des colonnes, inerties, indices ...etc).

Le tableau X

Le tableau de variables descriptives n'est influencé que par la métrique D.

Celle-ci intervient dans le calcul de la moyenne et de la variance ainsi que dans l'inertie du nuage des sites décrits par les variables descriptives. Elle donne à la distance d'un site à l'origine, une importance proportionnelle à son effectif exprimé dans F.

Par ailleurs, dans le nuage des variables, la métrique influe en particulier sur la liaison entre deux variables, puisque celle-ci s'exprime à travers une $f_{i.}$ corrélation.

Conclusion

La pondération D=diag($f_{i.}$) s'avère indispensable pour la conservation de la distance du χ^2 pour les profils des colonnes du tableau de fréquence initial; elle s'interprète facilement pour le tableau numérique, car on travaille souvent en ACP avec des poids d'individus non nécessairement uniformes, et les différents indices définis (moyennes, corrélations ...) sont des notions qui existent quels que soient les poids choisis.

6 Surpondération des groupes

La seconde étape de l'AFM est une ACP de l'ensemble des variables de tous les groupes.

Les variables de chaque groupe sont pondérées en divisant leur poids dans l'analyse séparée par la première valeur propre obtenue dans chacune des ACP séparées. Cette pondération a pour effet d'obtenir dans une direction quelconque pour chacun d'eux, une inertie maximale égale à 1, et ce dans le but d'équilibrer leur influence respective.

Pour un tableau de fréquence, le problème de la surpondération ou de la non surpondération va se poser avec une certaine acuité, compte tenu des particularités de ce type de tableau. En effet, les valeurs propres obtenues dans une AFC sont toutes majorées par 1.

Par ailleurs, l'inertie associée au tableau dépend de la structure du tableau et cette dernière se traduit plus particulièrement sur la première valeur propre. Plus la structure est forte, plus la première valeur propre est grande.

Nous examinons l'influence de cette surpondération dans trois formules différentes pour les deux tableaux F' et X, pour dégager celle qui nous semble la plus adaptée. Celles-ci consistent à surpondérer les deux tableaux F' et X, ou à ne surpondérer que le tableau de variables numériques, ou encore à n' introduire aucune surpondération.

On note λ_1 la première valeur propre obtenue dans l'AFC de F (ou ACP de F') et μ_1 celle obtenue dans l'ACP de X.

6.1 Surpondération du tableau de fréquence F' et du tableau de variables numériques X

Ce premier cas correspond à l'application classique de l'AFM.

Surpondérer les espèces j du tableau de fréquence F' qui sont initialement munies du poids $f_{.j}$, revient à leur attribuer le poids $f_{.j}/\lambda_1$ et surpondérer les variables de X revient à leur attribuer le poids $1/\mu_1$.

La distance entre deux sites i et i' devient donc dans l'ACP du tableau global F'UX :

$$d^2(i,i') = \sum_j \frac{f_{.j}}{\lambda_1} \left(\frac{f_{ij}}{f_{i.}f_{.j}} - \frac{f_{i'j}}{f_{i'.}f_{.j}} \right)^2 + \sum_k \frac{1}{\mu_1} \left(\frac{x_{ik} - x_{i'k}}{s_k} \right)^2$$

soit encore :

$$d^2(i,i') = \frac{d_1^2(i,i')}{\lambda_1} + \frac{d_2^2(i,i')}{\mu_1}$$

en notant d_1 et d_2 les distances entre sites décrits par le tableau F' et par le tableau descriptif X.

Le carré de la distance entre deux sites i et i' décrits par F et X s'écrit donc comme somme des carrés des distances normalisées pour ces deux sites décrits d'une part par la répartition des espèces et d'autre part, par les variables descriptives.

L'inertie du nuage des sites N(I) décrits par les espèces et par les variables descriptives devient donc :

$$Inertie(N(I)) = \frac{Inertie(N(I_1))}{\lambda_1} + \frac{Inertie(N(I_2))}{\mu_1}$$

où $N(I_1)$ est le nuage des sites décrits par les espèces (AFC du tableau F) et $N(I_2)$ le nuage des sites décrits par les variables descriptives (ACP du tableau X avec les poids $f_{i.}$).

Pour notre exemple, dans les ACP séparées, nous obtenons les inerties suivantes :

$$\text{Inertie}(N(I_1)) = \text{Inertie}(N(J)) = 1{,}79 \qquad \text{avec} \qquad \lambda_1 = 0{,}43$$

$$\text{Inertie}(N(I_2)) = \text{Inertie}(N(K)) = 4 \qquad \text{avec} \qquad \mu_1 = 3{,}05$$

L'inertie de $N(I_1)$ après surpondération, donc divisée par 0,43, passe de 1,79 à 4,16, tandis que celle de $N(I_2)$ divisée par 3,05 tombe de 4 à 1,31. L'inertie globale du nuage n'a pas tellement varié : elle passe de 5,79 à 5,47.

L'inertie et la première valeur propre du tableau de fréquence, sont beaucoup plus faibles que celles du tableau X. Ceci est normal puisque dans l'ACP de X, l'inertie est égale au nombre de variables et la première valeur propre est toujours supérieure à 1, tandis que dans l'AFC de F, la première valeur propre est inférieure à 1.

La surpondération augmente donc systématiquement l'importance du tableau de fréquence.

Plus la première valeur propre λ_1 est faible, plus l'importance du tableau de fréquence est amplifiée. Ceci peut conduire dans le cas extrême où λ_1 est nulle ou quasi nulle, à donner une place trop importante à un tableau de fréquence qui aura donc une inertie nulle ou quasi-nulle.

La surpondération ne change pas les distances entre espèces (tableau F), ni les corrélations entre variables (tableau X). Elle équilibre l'influence des deux sous nuages en donnant à chacun d'eux une inertie égale à 1, dans sa direction d'inertie maximum.

D'autre part, en AFM, en plus des résultats de l'ACP pondérée, on obtient des indices de liaison entre chaque facteur et chaque groupe de variables. Cet indice qui n'est autre que l'inertie du groupe projeté sur le facteur est du fait de la pondération, toujours compris entre 0 et 1, et un intervalle de variation fixe est nécessaire à l'interprétation d'un indice.

6.2 Surpondération du tableau descriptif X

Dans le cas où on ne surpondère que le tableau de variables X, la distance entre deux sites i et i' s'exprime comme suit :

$$d^2(i,i') = d_1^2(i,i') + \frac{d_2^2(i,i')}{\mu_1}$$

et l'inertie totale du nuage change dans le même rapport.

$$Inertie(N(I)) = Inertie(N(I_1)) + \frac{Inertie(N(I_2))}{\mu_1}$$

Dans ce cas, nous rendons plus proches les inerties associées aux deux groupes: 1,79 pour le tableau F et 1,31 pour le tableau X.

Pour chacun des nuages associés à chacun des deux tableaux, l'inertie dans une direction quelconque est encore inférieure ou égale à 1. Pour F, c'est une propriété de l'AFC ($\lambda_1 \leq 1$) et pour X, cela découle de la surpondération.

Si λ_1 est proche de 1, l'influence des deux tableaux sera équilibrée. Si au contraire, λ_1 est très faible, ce qui traduit le fait que le tableau de fréquence a une structure proche de l'indépendance, il sera presque éliminé de l'analyse. En quelque sorte, il a une importance proportionnelle à sa structure.

Du fait de la non pondération du tableau de fréquence, la liaison entre ce dernier et un facteur de l'analyse globale est au maximum égale à la première valeur propre λ_1, alors que la liaison entre le groupe de variables et le facteur sera au plus égale à 1. Ceci ne permet pas de conserver le même intervalle de variation pour la mesure de liaison entre une variable générale et un groupe, à savoir, l'intervalle [0,1] et complique l'interprétation.

6.3 Pas de surpondération pour les deux tableaux

Si on ne surpondère aucun des deux tableaux, ceux-ci ne seraient plus comparables et l'influence du tableau descriptif l'emporterait toujours sur le tableau de fréquence et les structures communes ne pourront pas être détectées. Ceci du fait que l'influence de X est fortement tributaire du nombre de ses variables et de sa structure, et celle de F n'est tributaire que de sa structure.

6.4 Choix du critère de surpondération

Nous retenons comme solution la première possibilité qui consiste à pondérer les deux groupes de façon classique. La première de nos motivations est en faveur de la conservation des propriétés de l'AFM pour la comparaison des groupes, à savoir un même intervalle de variation pour la mesure de liaison entre groupes et facteurs. Dans les autres cas, nous avons un intervalle de variation qui est variable, ce qui rend l'indice mesurant la liaison entre groupes et facteurs, difficilement interprétable.

D'autre part, dans l'étude de la structure d'un tableau, on prend en compte celle-ci, qu'elle soit forte ou faible. A fortiori, dans l'étude conjointe des structures de deux tableaux, il ne parait logique pas d'éliminer l'une au profit de l'autre.

Or, cette solution permet justement de donner la même importance dans l'analyse globale, aux distances entre sites définis par les espèces et par les variables, ceci que la structure définie par la répartition des espèces soit forte ($\lambda_1 \simeq 1$) ou faible ($\lambda_1 \simeq 0$).

Le cas extrême d'un tableau de fréquence de structure très faible, pourra justement être pris en compte, car dans le cas contraire, l'influence du tableau descriptif serait prépondérante, masquant ainsi tous les points communs qu'on voudrait mettre en relief.

7 Individus, Variables et Groupes

En AFM, on analyse trois types d'objets : les individus, les variables et les groupes. Ainsi, nous sommes amenés à étudier les caractéristiques des différents nuages les représentant.

7.1 Les individus

On cherche comme dans l'AFM classique, une représentation graphique du nuage des individus N(I) caractérisés par l'ensemble des variables, et une représentation superposée des deux nuages de sites $N(I_1)$ et $N(I_2)$ caractérisés par la répartition des espèces et par les variables d'environnement.

L'espace $R^{J\oplus K}$ de dimension J+K contient les représentations par le nuage N(I) des individus définis par l'ensemble des colonnes des tableaux X et F.

La projection de N(I) sur le sous espace de $R^{J\oplus K}$ engendré par les J premières composantes, se déduit de $N(I_1)$ par une simple homothétie de rapport $1/\lambda_1$. De même, la projection de N(I) sur le sous espace de $R^{J\oplus K}$ engendré par les K dernières composantes, se déduit de $N(I_2)$ par une homothétie de rapport $1/\mu_1$.
Ces homothéties sont dues à la surpondération.

Caractéristiques des différents nuages de sites

Trois nuages de sites sont donc en présence.

- **Le nuage des sites décrits par les espèces : $N(I_1)$**

  Il est $f_{i.}$ centré et la distance entre deux sites s'exprime par la distance du χ^2 et correspond au nuage des sites obtenu dans l'AFC du tableau de fréquence initial F.

- **Le nuage des sites décrits par les variables descriptives: $N(I_2)$**

  Ce nuage qui correspond aux lignes du tableau X est encore $f_{i.}$ centré et la distance entre deux sites est la distance euclidienne que l'on a dans une ACP, et les proximités sont celles d'une ACP.

- **Le nuage des sites décrits par les espèces et par les variables descriptives : N(I)**

  Le nuage des sites N(I) qui est la "réunion" de deux sous nuages $f_{i.}$ centrés est aussi $f_{i.}$ centré, et correspond aux lignes du tableau F'UX.

  Le carré de la distance entre deux sites s'écrit comme la somme des carrés pondérés des distances entre les sites dans chacun des sous nuages $N(I_1)$ et $N(I_2)$ (cf 6.1).

7.2 Les variables

Les variables (espèces et variables descriptives) associées aux deux groupes sont situées dans R^I, espace dans lequel nous les représentons ainsi que les composantes principales des ACP séparées.

Comme dans l'AFM classique, nous cherchons s'il existe un ou plusieurs facteurs communs à ces deux groupes.

Les nuages de variables

Nous étudions également trois nuages, dont l'un est la réunion des deux autres.

- **Le nuage des espèces : N(J)**

  Le nuage des espèces N(J) est $f_{.j}$ centré et l'interprétation pour ce nuage de variables est analogue à celle d'une AFC, puisque la distance entre deux espèces est la distance du χ^2 et leur liaison s'exprime sur un axe par leur proximité au sens euclidien.

- **Le nuage des variables descriptives : N(K)**

  Le sous nuage des variables descriptives N(K), n'est pas centré et on retrouve pour celui-ci les interprétations de l'ACP pour les variables, avec entre autre, une interprétation des liaisons sur un plan, à l'aide d'un cercle de corrélation qui est en fait un cercle des $f_{i.}$ corrélations.

- **Le nuage des espèces et des variables descriptives : N(JUK)**

  Le nuage N(JUK) qui est situé dans R^I est également non centré, puisqu'il est la réunion de deux sous nuages dont l'un n'est pas centré, soit N(K). Il correspond aux colonnes du tableau F'UX.

  La liaison entre une espèce et une variable peut être abordée comme en ACP. Elle est alors mesurée par la covariance entre le profil de l'espèce et la variable.

Cette covariance est délicate à interpréter, mais nous verrons que sur un facteur, grâce aux formules de transition, l'interprétation des positions relatives espèces-variables est simple.

7.3 Les groupes

Les deux groupes sont représentés dans R^{I^2} par les opérateurs W_1D et W_2D avec $W_1 = {}^t F'M_1F'$ et $W_2 = {}^t XM_2X$, muni du produit scalaire $\langle A, B\rangle$ = Trace A ^{t}B. M_1 et M_2 sont les matrices diagonales des poids des variables pour les groupes 1 et 2.

Dans cet espace, un groupe est représenté par sa coordonnée sur l'axe factoriel $z_s{}^t z_s D$ qui est l'inertie de la projection du nuage défini par le groupe sur la composante principale z_s de l'analyse globale de F'UX.

Deux groupes sont proches si la distance $d^2(G_1, G_2)$ est faible et ceci correspond au cas où les distances entre individus sont semblables dans $N(I_1)$ et $N(I_2)$.

8 Représentation des sites et des variables sur les facteurs et interprétation des résultats

Notations

Soient:

- γ_s la $s^{ième}$ valeur propre obtenue dans l'AFM de F'UX (i.e. dans l'ACP pondérée)
- $F_s(i)$ le facteur d'ordre s pour le site i obtenu dans l'AFM de F'UX
- $G_s(j)$ le facteur d'ordre s pour l'espèce j obtenu dans l'AFM de F'UX
- $H_s(k)$ le facteur d'ordre s pour la variable descriptive k obtenu dans l'AFM de F'UX

Rappelons que dans l'analyse globale, les sites sont munis des poids $f_{i.}$, les espèces des poids $f_{.j}/\lambda_1$ et les variables descriptives des poids $1/\mu_1$.

8.1 Les groupes et les facteurs

1. **Premier facteur**

   L'inertie du premier facteur vaut 1,95. Ceci indique déjà l'existence d'une structure commune entre les deux groupes, puisque la valeur maximum de cette inertie qui est 2, est atteinte lorsque les directions principales d'inertie des deux nuages se confondent, et que le lien facteur-groupe vaut 1. Ici, ce lien vaut respectivement 0,97 et 0,98 pour le premier et le deuxième groupe.

   La corrélation entre ce facteur et la projection associée du nuage $N(I_1)$ vaut 0,991 et vaut également 0,991 pour le nuage $N(I_2)$. Donc, on peut conclure que le facteur commun est confirmé : il existe dans les deux nuages $N(I_1)$ et $N(I_2)$ une direction de dispersion commune. Ces directions sont très proches des premiers facteurs de F et de X. La $f_{i.}$ corrélation vaut 0,966 pour le premier et 0,990 pour le second.

2. **Deuxième facteur**

   L'inertie du deuxième facteur vaut 0,810 et le lien avec le premier groupe est de 0,792 et avec le second groupe, il est de 0,020. Il s'agit donc, d'un facteur de répartition des espèces non lié aux variables descriptives. Il se confond presque avec le deuxième facteur de F (corrélation = 0,983).

3. **Facteurs suivants**

   Les facteurs suivants sont également des facteurs spécifiques à la répartition des espèces sur les sites. En fait, cette situation était attendue, puisque dans les ACP séparées, le tableau X est quasi-monodimensionnel et le tableau F est multidimensionnel.

Nous nous limitons dans la suite, à l'interprétation des deux premiers facteurs.

8.2 Représentation des variables

1. **Les variables descriptives**

   Formules de transition

   Pour une variable descriptive k, nous avons la formule de transition suivante :

$$H_s(k) = \frac{1}{\sqrt{\gamma_s}} \sum_i f_{i.} \left(\frac{x_{ik} - \bar{x}_k}{s_k} \right) F_s(i) = \frac{1}{\sqrt{\gamma_s}} \sum_i f_{i.} \left(\frac{x_{ik}}{s_k} \right) F_s(i) \qquad k \in K$$

   Interprétation (Fig 1)

   L'interprétation de la projection du nuage des variables descriptives se fait de manière analogue à celle de l'ACP. En effet, la coordonnée de la projection d'une variable k sur un facteur représente la $f_{i.}$ corrélation entre cette variable et le facteur, de la même manière qu'en ACP. On trace pour cela le cercle des corrélations pour l'interprétation des projections des variables descriptives.

   On note sur le premier facteur, une opposition très marquée entre l'indice de salinité (SALI) et la distance à la mer (DIST). La profondeur du substrat (SOIL) est très corrélée avec la distance à la mer et la pente du site (SLOP) l'est avec la salinité (SALI).

   Comme ces variables sont très corrélées avec le premier facteur, on s'attend à retrouver les projections des sites salés et proches de la mer du coté de la variable SALI et ceux qui sont loin de la mer et donc moins salés du coté de la variable DIST.

   Pour le deuxième facteur, les variables descriptives n'interviennent quasiment pas.

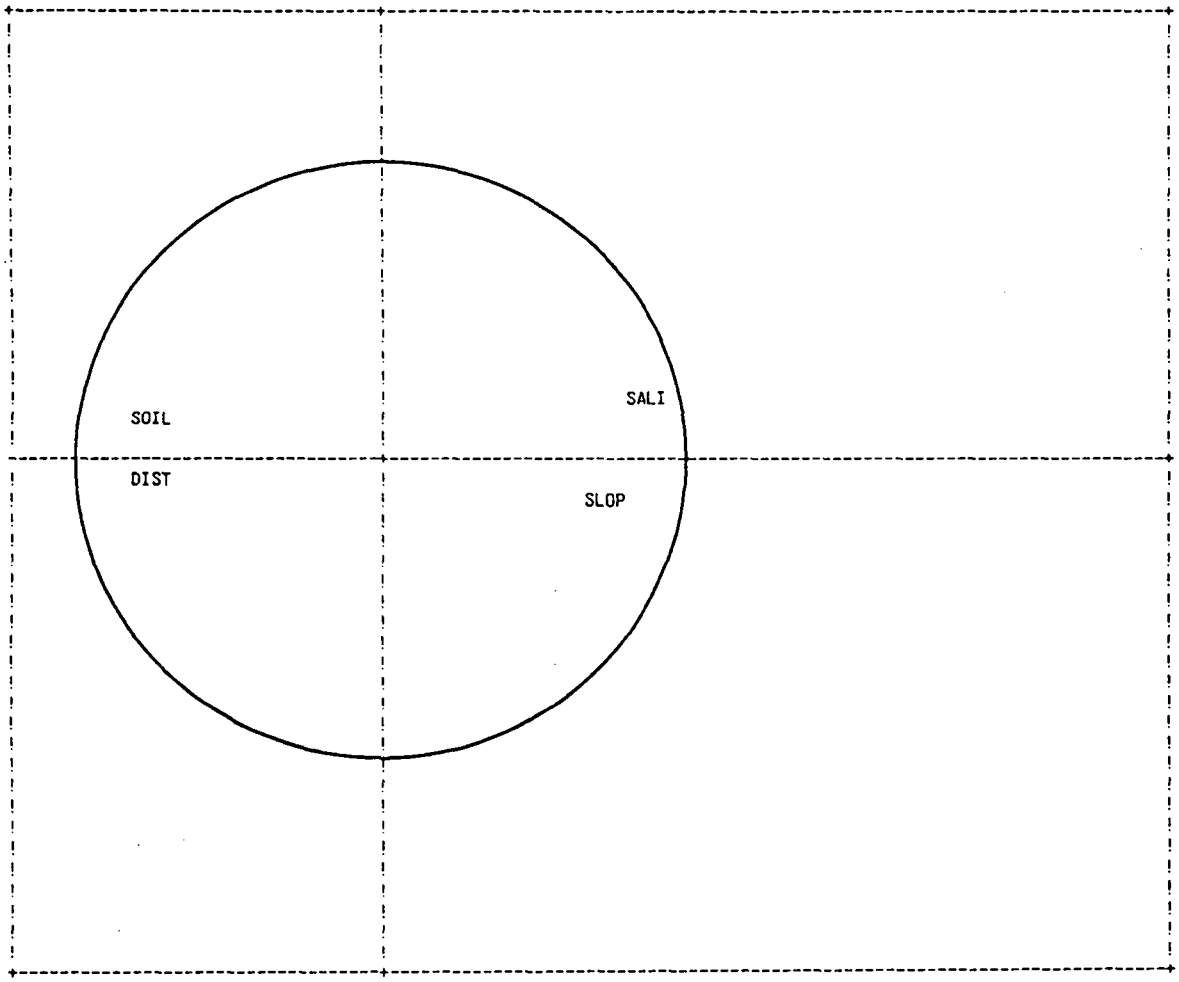

Fig 1:

Représentation des variables descriptives sur le plan 1x2

2. **Les espèces**

Formules de transition

Pour une espèce j du tableau de fréquence:

$$G_s(j) = \frac{1}{\sqrt{\gamma_s}} \sum_i f_{i.} \left(\frac{f_{ij}}{f_{i.} f_{.j}} - 1 \right) F_s(i) = \frac{1}{\sqrt{\gamma_s}} \sum_i \frac{f_{ij}}{f_{.j}} F_s(i) \qquad j \in J$$

Interprétation (Fig 2)

La pondération des sites par $f_{i.}$ fait d'une espèce j le barycentre des sites qui la contiennent (car $\sum_i f_{i.} F_s(i) = 0$). De ce fait, nous retrouvons la propriété barycentrique de l'AFC.

La liaison entre deux espèces se traduit en terme de distance ; celle-ci est en fait la distance du χ^2. On aura donc une interprétation du nuage projeté des espèces N(J) qui sera analogue à celle de l'AFC.

De même, pour les espèces, il y a une opposition entre les espèces qui se projettent du coté de la distance à la mer (DIST) et celles qui se projettent du coté de l'indice de salinité (SALI).

On peut en conclure qu'une espèce qui se projette du même coté que SALI, et à l'opposé de DIST, est plus fréquente dans les sites salins.

A l'opposé, une espèce qui est du même coté que DIST est plus fréquente dans les sites éloignés de la mer et donc moins salins.

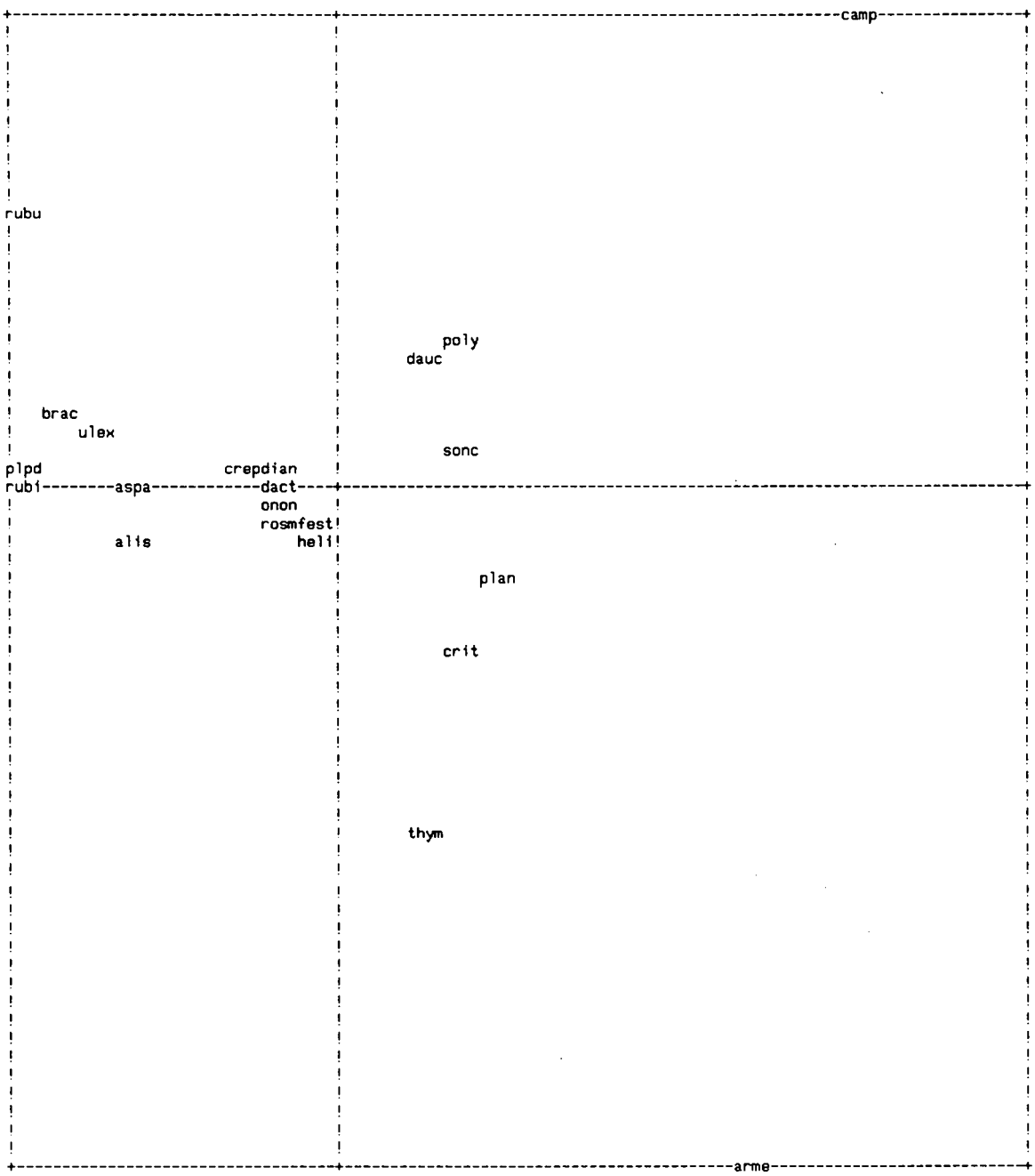

Fig 2:

Représentation des espèces sur le plan 1x2

3. **Les variables descriptives et les espèces**

   Interprétation de la proximité espèces-variables descriptives

   L'espèce j et la variable k se projetteront du même coté, si des sites qui ont une fréquence de j supérieure à la fréquence moyenne des sites, ont aussi une valeur supérieure à la valeur moyenne pour k.

   L'espèce "camp" et la variable SALI se projettent du même coté. On peut en conclure que cette espèce est beaucoup plus fréquente que la moyenne dans les sites qui ont un indice de salinité supérieur à la valeur moyenne.

8.3 Représentation des sites

Divers points de vue pour les sites seront examinés : le premier concerne les sites vus au travers des deux groupes de variables et les suivants décrivent les sites vus au travers de chacun des deux groupes.

1. **Représentation du nuage des sites vus au travers des 2 groupes**

   La représentation du nuage des sites N(I) s'obtient en faisant l'ACP pondérée du tableau conjoint F'UX.

   Formules de transition

$$F_s(i) = \frac{1}{\sqrt{\gamma_s}} \left[\sum_j \frac{f_{.j}}{\lambda_1} \left(\frac{f_{ij}}{f_{.j}} - 1 \right) G_s(j) + \sum_k \frac{1}{\mu_1} \left(\frac{x_{ik} - \bar{x}_k}{s_k} \right) H_s(k) \right] \qquad i \in I$$

   Interprétation (Fig 3)

   Deux sites sont proches, s'ils sont en même temps proches au sens de la répartition sur les espèces et au sens des variables descriptives.

   Sur le facteur 1, il y a une opposition très nette entre les sites éloignés de la mer : ce sont ceux qui se projettent du coté de distance à la mer et ceux qui sont les plus salins qui se projettent à l'autre extrémité de l'axe. Ils se projettent aussi du coté où les espèces qu'ils contiennent sont les plus fréquentes.

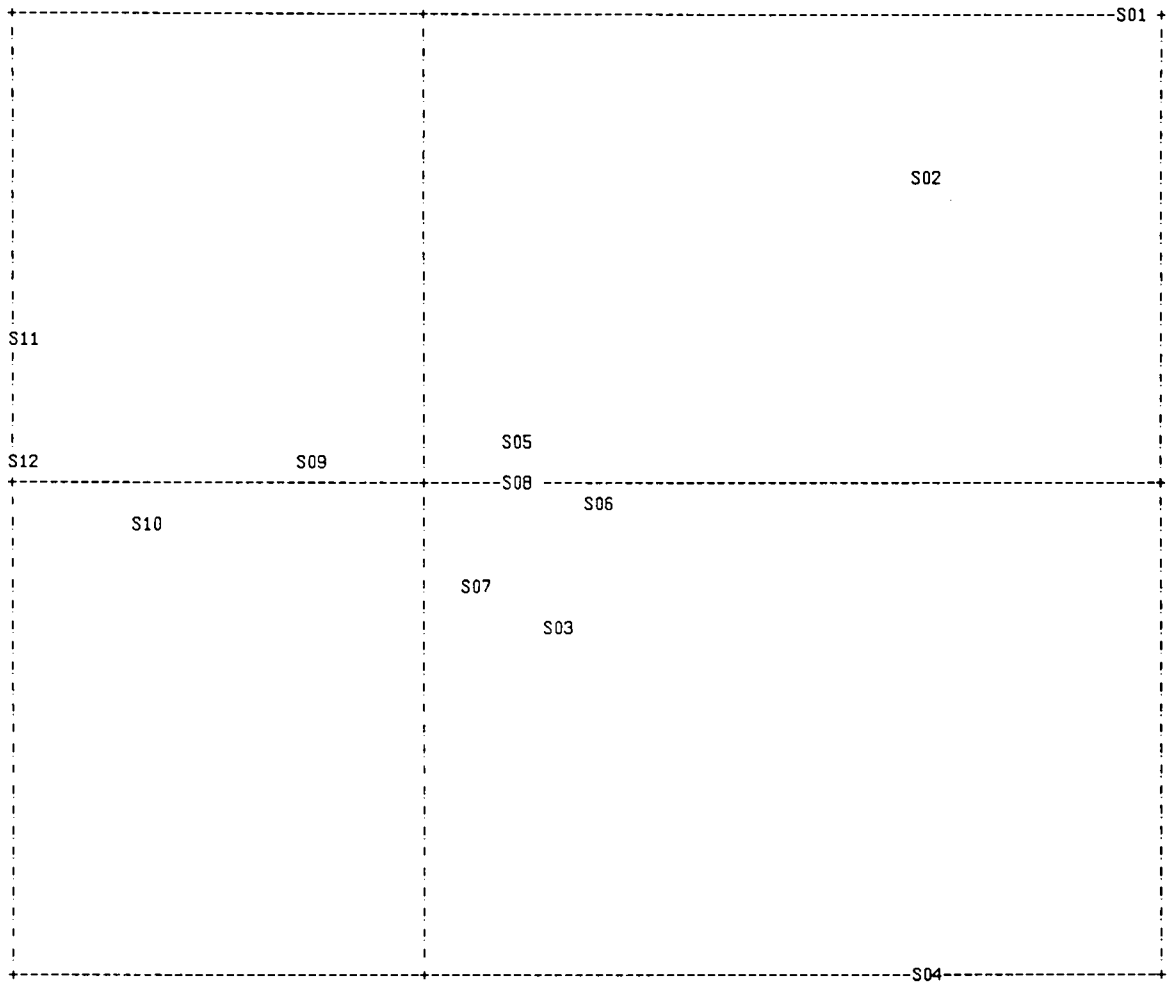

Fig 3:

Représentation des sites au travers des espèces et des variables descriptives sur le plan 1x2

2. **Représentation des sites vus au travers de leur répartition sur les espèces**

Pour obtenir la représentation d'un site i_1 du nuage $N(I_1)$, il suffit de considérer le tableau $\tilde{F}' = [F', 0]$ en supplémentaire dans l'analyse globale.

Un site i_1 aura donc comme coordonnées :

$$i_1 = \left[\left(\frac{f_{ij}}{f_{i.} f_{.j}} - 1 \right)_{j \in J}, (0)_{k \in K} \right]$$

Sa coordonnée sur l'axe s obtenu dans l'ACP globale sera :

$$F_s(i_1) = \frac{1}{\sqrt{\gamma_s}} \sum_j \frac{f_{.j}}{\lambda_1} \left(\frac{f_{ij}}{f_{.j}} - 1 \right) G_s(j) = \frac{1}{\lambda_1 \sqrt{\gamma_s}} \sum_j \frac{f_{ij}}{f_{i.}} G_s(j) - \frac{1}{\lambda_1 \sqrt{\gamma_s}} \sum_j f_{.j} G_s(j)$$

Or le dernier terme est nul, le nuage des espèces étant centré et le site i_1 sera au barycentre des espèces qu'il contient, d'où :

$$F_s(i_1) = \frac{1}{\lambda_1 \sqrt{\gamma_s}} \sum_j \frac{f_{ij}}{f_{i.}} G_s(j)$$

3. **Représentation des sites vus au travers des variables descriptives**

On procède de la même manière que pour le nuage $N(I_1)$.

On projette le tableau $\tilde{X} = [0, X]$ en supplémentaire dans l'ACP globale.

Un point i_2 du nuage $N(I_2)$ a pour coordonnées.

$$i_2 = \left[(0)_{j \in J}, \left(\frac{x_{ik} - \bar{x}_k}{s_k} \right)_{k \in K} \right]$$

Sa projection sur l'axe s'écrit:

$$F_s(i_2) = \frac{1}{\sqrt{\gamma_s}} \sum_k \frac{1}{\mu_1} \left(\frac{x_{ik} - \bar{x}_k}{s_k} \right) H_s(k)$$

Interprétation (Fig 4)

Représentation simultanée des sites pour les trois points de vue

Sur le facteur 1, les sites décrits par les espèces et ceux décrits par les variables descriptives sont quasi confondus, puisque c'est un facteur commun. Notons que le site 6 est très salin, mais a une répartition sur les espèces un peu moins marquée, que ne le laisse attendre les variables descriptives.

En particulier, les sites 2 et 4 sont tous deux très salins, mais ont des répartitions sur les espèces qui sont très différentes. Cette différence s'explique par la présence de l'espèce "camp" pour le site 1 qui y est fréquente, alors que pour le site 4, elle est due à la présence de l'espèce "arme".

On note S01 le site 1 vu au travers des deux groupes, S011 le site 1 vu au travers des espèces (groupe 1), et S012 le site 1 vu au travers des variables descriptives (groupe 2)

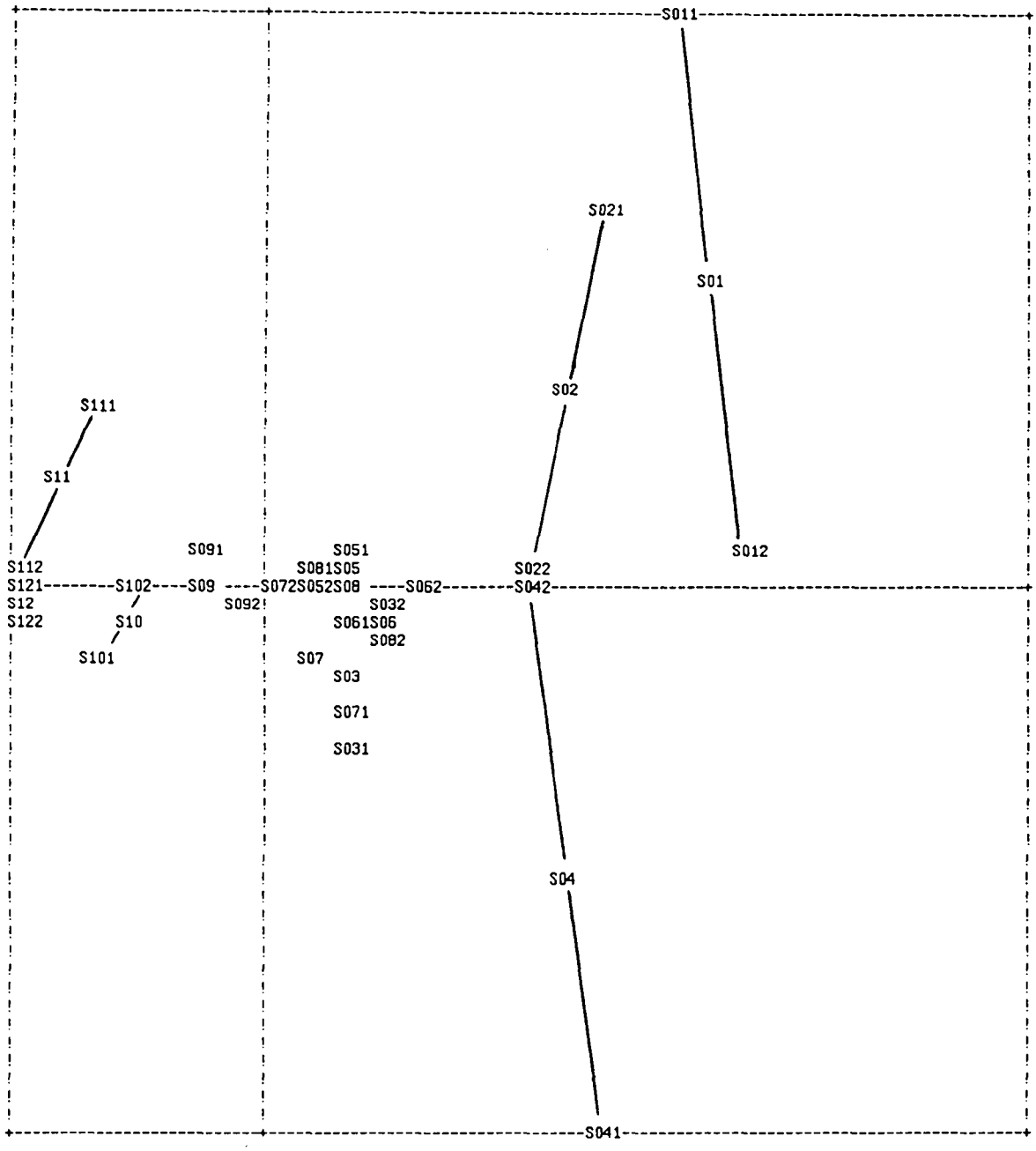

Fig 4:
Représentation des sites vus au travers des trois points de vue sur le plan 1x2

9 Comparaison avec l'Analyse Canonique des Correspondances

Parmi les méthodes permettant d'étudier un tableau de fréqences et un tableau de variables numériques, nous relevons en particulier le cas de l'Analyse Canonique des Correspondances (ACC) [8].

Dans l'ACC, on ne prend en compte que la part de la structure associée au tableau de fréquence, qui est expliquée par les variables de X. Nous rappelons que l'ACC consiste à faire l'ACP des projections des colonnes du tableau F' présenté dans notre méthode, sur le sous espace engendré par les variables de X, avec les métriques D et M définies précédemment. L'ACC propose donc une approche non symétrique.

L'AFM, quant à elle, permet une étude simultanée et non privilégiée des structures associées au tableau de fréquence et au tableau numérique. En plus, elle permet de détecter l'existence de structures communes et de structures spécifiques.

L'AFM doit être un préalable à une ACC qui pourra être faite pour une analyse plus fine surtout si des structures communes ont été décelées.

Interprétation des résultats obtenus par les deux méthodes (Fig 5 et 6)

Nous avons effectué une AFM et une ACC sur l'exemple présenté.

Nous trouvons un premier facteur qui est le même dans les deux analyses. Ceci s'explique puisque le premier facteur de l'AFM est un facteur absolument commun aux deux groupes.

Par contre les deuxièmes facteurs de l'AFM et de l'ACC sont très différents. Ceci est du au fait que le deuxième facteur de l'AFM est un facteur de répartition des espèces qui est très peu lié aux variables descriptives, et qui ne peut apparaitre dans l'ACC.

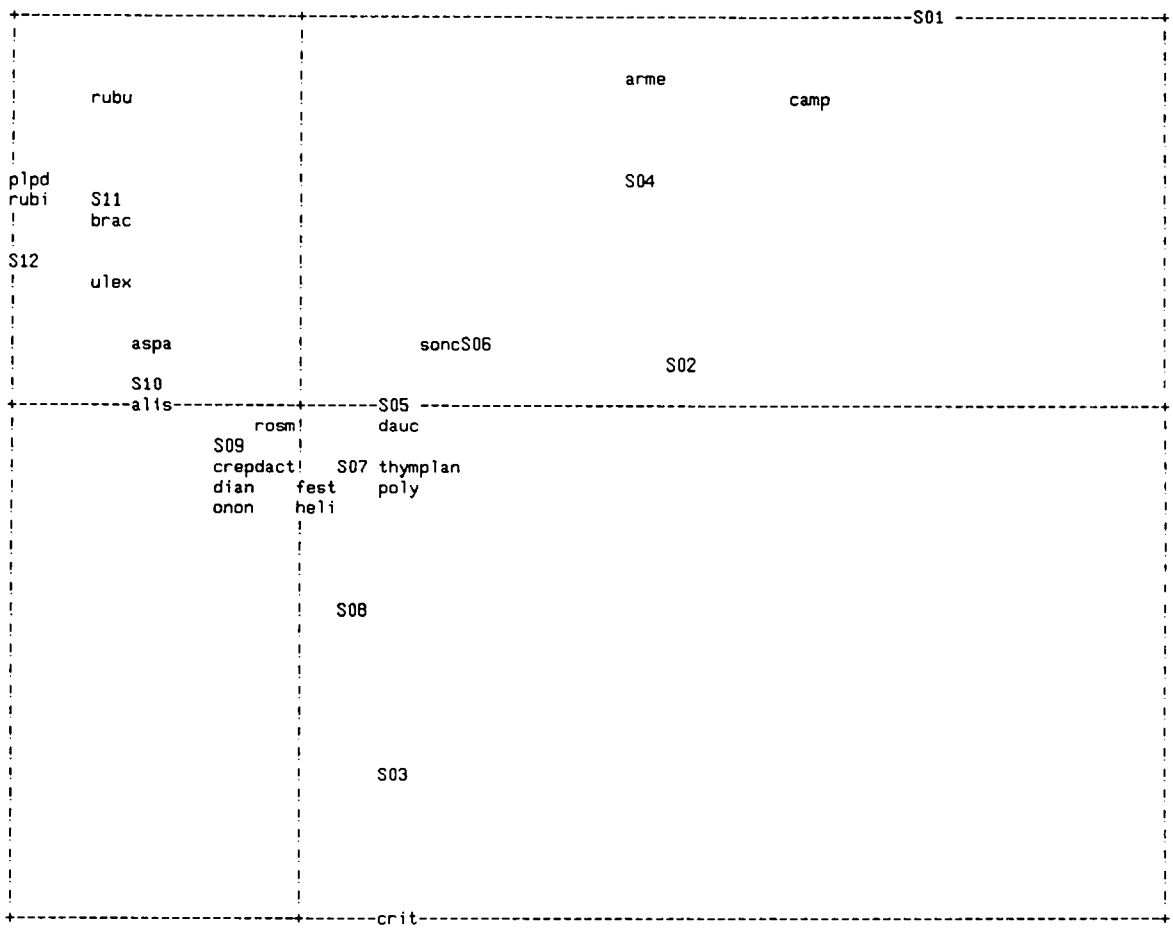

Fig 5:
Représentation des espèces et des sites sur le plan 1x2 de l'ACC

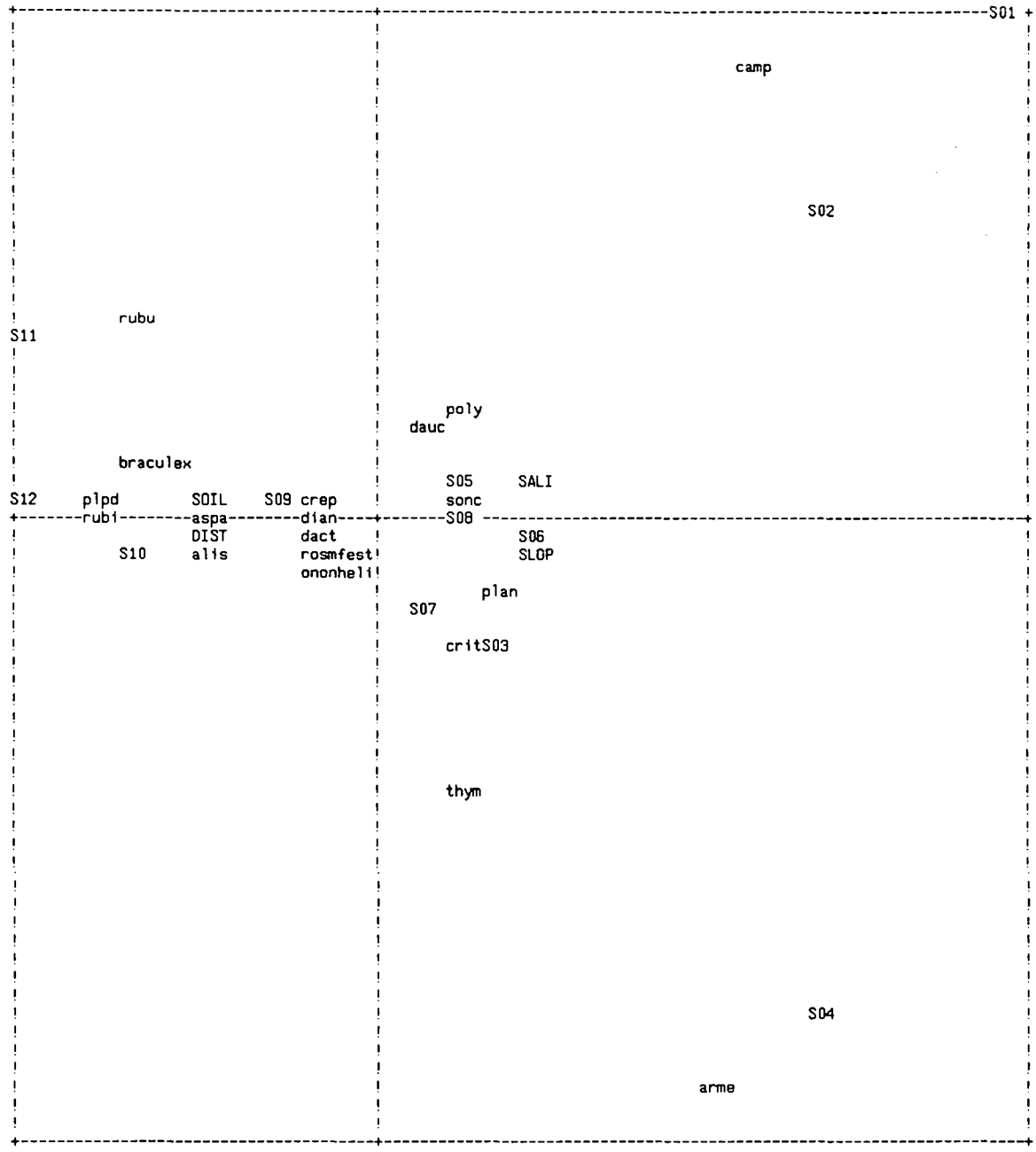

Fig 6:
Représentation des espèces, des variables descriptives et des sites sur le plan 1x2 de l'AFM

10 Généralisation à l'étude de plusieurs tableaux

Dans le cas où l'on étudie plus de deux tableaux, deux principaux cas sont envisagés.

10.1 Cas d'un tableau de fréquence et de plusieurs tableaux de variables

Dans le cas d'un tableau de fréquence et de plusieurs tableaux descriptifs, la méthode présentée peut encore s'appliquer.

Dans le cas où le tableau descriptif est constitué de variables qualitatives, les résultats utilisés dans l'AFM classique sont encore utilisables, à savoir l'équivalence entre l'ACM et l'ACP pondérée des indicatrices. Par ailleurs, pondérer les individus par $f_{i.}$ pour des variables qualitatives, ne pose aucun problème.

10.2 Cas de plusieurs tableaux de fréquence

On peut aussi envisager l'étude des liaisons entre plusieurs tableaux de fréquence et éventuellement d'autres tableaux descriptifs. Ce cas de figure se présente pour des tableaux de fréquence donnant des répartitions d'une population à des époques différentes pour la même variable, ou pour des tableaux de fréquence associés à des variables différentes...etc.
Dans ce cas, la méthode que nous proposons ne peut s'appliquer que si les tableaux de fréquence ont les mêmes marges sur les lignes.

Dans le cas contraire, il faudra se départir de la structure AFC de chacun des tableaux de fréquence, puisque les marges sur les lignes sont différentes ; on ne pourra pas restituer simultanément ces structures AFC, moyennant la transformation des tableaux de fréquence et l'introduction d'une même métrique sur les lignes.

Nous proposons alors d'autres approches ne faisant justement pas intervenir les marges sur les lignes. Parmi celles-ci, nous pourrons faire une analyse par sous tableaux ou nous ramener à des tableaux de pourcentages.

11 Conclusion

Avec l'extension de l'AFM aux tableaux de fréquence, on peut traiter de façon simultanée et symétrique plusieurs tableaux de données de nature différente : ceux-ci peuvent être de type numérique, qualitatif ou être de type fréquence.

Par ailleurs, l'AFM et l'ACC sont deux méthodes complémentaires.
On appliquera l'une ou l'autre des deux méthodes, selon que l'on privilégie une approche symétrique et dans ce cas, on effectuera une AFM, ou non symétrique et là, on utlisera l'ACC qui est plus adaptée.

D'autre part, si avec l'AFM, nous avons la possibilité de détecter des structures communes ou spécifiques, l'ACC viendra utilement affiner les résultats trouvés, essentiellement dans le cas où il existe des structures communes à au moins deux groupes.

Références

[1]. BENZÉCRI J.P. (1973)
L'analyse des données: II, L'analyse des correspondances - Dunod

[2]. DROUET D., ESCOFIER B. (1983)
Comparaison de plusieurs tableaux de fréquence - Cahiers de l'analyse des données Vol. VIII n^0 4

[3]. ESCOFIER B., PAGES J. (1990)
Analyses factorielles simples et multiples - Dunod

[4]. GREENACRE M.J. (1984)
Theory and applications of Correspondence Analysis - Academic Press

[5]. LEBRETON J.D., CHESSEL D., PRODON R., YOCCOZ N. (1988)
L'analyse des relations espèces-milieu par l'analyse canonique des correspondances. Variables de milieu quantitatives - Acta oecologica Oecol. Gen. Vol.9 n^0 1

[6]. LEBRETON J.D., CHESSEL D., RICHARDOT-COULET M., YOCCOZ N. (1988)
L'analyse des relations espèces-milieu par l'analyse canonique des correspondances. Variables de milieu quantitatives - Acta oecologica Oecol. Gen. Vol.9 n^0 2

[7]. LEBRETON J.D., SABATIER R.,BANCO G., BACOU A.M. (1991)
Principal component and correspondence analyses with respect to instrumental variables : an overview of their role in studies of structure (activity and species) environment relationships - Devillers and Karcher

[8]. TER BRAAK C.J.F. (1986)
Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis - Ecology

[9]. TER BRAAK C.J.F. (1987)
The analysis of vegetation-environment relationships by canonical correspondence analysis - Ecology

[10]. TER BRAAK C.J.F. (1988 A)
Classification and related methods of data analysis - Bock H.H.

Annexe

Abréviations et noms des espèces

camp	: Camphorosma monspeliaca
poly	: Polycarpon polycarpoides
dauc	: Daucus carota
sonc	: Sonchus tenerrimus
plan	: Plantago subulata
thym	: Thymelea hirsuta
crit	: Crithmum maritimum
fest	: Festuca arvernensis
heli	: Helichrysum stoechas
crep	: Crepis bulbosa
dact	: Dactylis glomerata
arme	: Armeria ruscinonensis
rosm	: Rosmarinus officinalis
onon	: Ononis spinosa
ulex	: Ulex parviflorus
aspa	: Asparagus acutifolius
dian	: Dianthus pyrenaicus
alis	: Alyssum maritimum
brac	: Brachypodium retusum
rubu	: Rubus sp.
plpd	: Polypodium australe
rubi	: Rubia peregrina

LISTE DES DERNIERES PUBLICATIONS INTERNES PARUES A L'IRISA

PI 678 ETUDE DE QUELQUES ORGANISATIONS D'ANTEMEMOIRES
Nathalie DRACH, André SEZNEC
Octobre 1992, 44 pages.

PI 679 AN ADAPTIVE SPARSE UNSYMMETRIC LINEAR SYSTEM SOLVER
Miloud SADKANE, Roger B. SIDJE
Octobre 1992, 28 pages.

PI 680 BRANCHING BISIMULATION FOR CONTEXT-FREE PROCESSES
Didier CAUCAL, Dung HUYNH, Lu TIAN
Octobre 1992, 36 pages.

PI 681 DEADLOCK MODELS AND GENERAL ALGORITHM FOR DISTRIBUTED DEADLOCK DETECTION
Jerzy BRZEZINSKI, Jean-Michel HELARY, Michel RAYNAL
Octobre 1992, 26 pages.

PI 683 TARGET TRACKING BY VISUAL SERVOING
Aristide S. SANTOS, François CHAUMETTE
Octobre 1992, 50 pages.

PI 684 UNE DESCRIPTION LINEAIRE COMPLETE ET IRREDONDANTE DU POLYTOPE ASSOCIE AU PROBLEME DU VOYAGEUR DE COMMERCE ASYMETRIQUE A 6 SOMMETS
Reinhardt EULER, Hervé LE VERGE
Octobre 1992, 30 pages.

PI 685 MISE EN CORRESPONDANCE DE SEGMENTS DANS UNE SEQUENCE D'IMAGES PAR UNE APPROCHE LOCALE
Samia BOUKIR, Patrick BOUTHEMY, François CHAUMETTE, Didier JUVIN
Octobre 1992, 30 pages.

PI 686 FROM EQUATIONS TO HARDWARE. TOWARDS THE SYSTEMATIC MAPPING OF ALGORITHMS ONTO PARALLEL ARCHITECTURES
François CHAROT, Patrice FRISON, Eric GAUTRIN, Dominique LAVENIER, Patrice QUINTON, Charles WAGNER
Octobre 1992, 18 pages.

PI 687 THE COMPILATION OF PROLOG and its Execution with MALI
Pascal BRISSET, Olivier RIDOUX
Novembre 1992, 90 pages.

PI 688 GENERALISATION DE L'ANALYSE FACTORIELLE MULTIPLE A L'ETUDE DES TABLEAUX DE FREQUENCE ET COMPARAISON AVEC L'ANALYSE CANONIQUE DES CORRESPONDANCES
Lila ABDESSEMED, Brigitte ESCOFIER
Novembre 1992, 34 pages.

ISSN 0249-6399