

# Analyse Non Symétrique des Correspondances

## Résumé

La fiche décrit le principe et l'usage de l'analyse non symétrique des correspondances créée par Lauro, N. & D'Ambra, L. (1984, *in* : Data Analysis and Informatics III. Diday, E. & Coll. (Eds.) Elsevier, North-Holland. 433-446). Le test de Light, R.J. & Margolin, B.H. (1971, An analysis of variance for categorical data. *Journal of the American Statistical Association* : 66, 534-544) lui est associée dans les deux options NSCA du module COA. L'enregistrement des occurrences d'espèces à partir des collections, de la littérature ou des relevés exploratoires est particulièrement en accord avec cette approche. L'asymétrie des relations espèces-relevés vues par l'averaging est abordée.

## Plan

1 — Introduction.....	2
2 — L'analyse des correspondances.....	4
2.1 — Centrage des indicatrices.....	4
2.2 — Discrimination des occurrences espèces entre les relevés .....	6
2.3 — Discrimination des occurrences relevés entre les espèces .....	6
2.4 — Analyse canonique des occurrences .....	7
2.5 — Les fichiers d'occurrences .....	7
3 — Principes .....	10
3.1 — L'indice CATANOVA.....	10
3.2 — Propriétés de l'ANSC .....	12
3.3 — Deux programmes pour une même analyse.....	13
4 — Exemples .....	14
4.1 — Histogrammes et classes de taille.....	14
4.2 — ANSC et ordination .....	18
4.3 — Question de centrage.....	22
4.4 — Vers la mesure de l'asymétrie d'un tableau faunistique.....	26
Références .....	29

D. Chessel & C. Gimaret

# 1 — Introduction

L'analyse des correspondances est réputée comme étant une analyse discriminante ou une analyse canonique. Il doit donc exister une version inter-classe associée. Elle s'appelle l'Analyse Non Symétrique des Correspondances (ANSC) décrite dans <sup>1</sup>. Elle s'appuie sur le critère CATANOVA inventé dans <sup>2</sup>. Une relecture de ces deux articles est proposée dans cette introduction dans le cadre d'une situation écologique très précise qui convient particulièrement bien à l'expression des principes en jeu.

Les écologues connaissent la pratique qui permet de passer du cahier de terrain à un tableau floro-faunistique. Dans cette opération, ils exécutent une opération mathématique fondamentale dans la théorie de l'analyse des correspondances. Reprenons l'exemple donné dans <sup>3</sup>. Le cahier de terrain a la forme suivante :

```
Relevé n° 1
  Taxon 1 Ab = 1
Relevé n° 2
  Taxon 2 Ab = 4
Relevé n° 3
  Taxon 1 Ab = 1
  Taxon 2 Ab = 3
  Taxon 3 Ab = 1
Relevé n° 4
  Taxon 2 Ab = 2
  Taxon 3 Ab = 2
```

Supposons que l'abondance (Ab) soit un effectif d'individus. Si on note individu par individu le taxon auquel il appartient, l'enregistrement primaire prend la forme :

```
Relevé n° 1
  Individu 1 : Taxon = 1
Relevé n° 2
  Individu 2 : Taxon = 2
  Individu 3 : Taxon = 2
  Individu 4 : Taxon = 2
  Individu 5 : Taxon = 2
Relevé n° 3
  Individu 6 : Taxon = 1
  Individu 7 : Taxon = 2
  Individu 8 : Taxon = 2
  Individu 9 : Taxon = 2
  Individu 10 : Taxon = 3
Relevé n° 4
  Individu 11 : Taxon = 2
  Individu 12 : Taxon = 2
  Individu 13 : Taxon = 3
  Individu 14 : Taxon = 3
```

Un individu est alors un événement auquel on associe deux caractères, respectivement l'espèce à laquelle il appartient et le relevé dans lequel on l'a trouvé. On donnera le terme générique de tableau d'occurrences à l'écriture :

```
Individu 1 : Taxon = 1 Relevé = 1
Individu 2 : Taxon = 2 Relevé = 2
Individu 3 : Taxon = 2 Relevé = 2
Individu 4 : Taxon = 2 Relevé = 2
Individu 5 : Taxon = 2 Relevé = 2
Individu 6 : Taxon = 1 Relevé = 3
Individu 7 : Taxon = 2 Relevé = 3
Individu 8 : Taxon = 2 Relevé = 3
Individu 9 : Taxon = 2 Relevé = 3
Individu 10 : Taxon = 3 Relevé = 3
Individu 11 : Taxon = 2 Relevé = 4
Individu 12 : Taxon = 2 Relevé = 4
Individu 13 : Taxon = 3 Relevé = 4
```

Individu 14 : Taxon = 3 Relevé = 4

On peut ranger les occurrences par taxon :

Individu 1 : Taxon = 1 Relevé = 1  
Individu 6 : Taxon = 1 Relevé = 3  
Individu 2 : Taxon = 2 Relevé = 2  
Individu 3 : Taxon = 2 Relevé = 2  
Individu 4 : Taxon = 2 Relevé = 2  
Individu 5 : Taxon = 2 Relevé = 2  
Individu 7 : Taxon = 2 Relevé = 3  
Individu 8 : Taxon = 2 Relevé = 3  
Individu 9 : Taxon = 2 Relevé = 3  
Individu 11 : Taxon = 2 Relevé = 4  
Individu 12 : Taxon = 2 Relevé = 4  
Individu 10 : Taxon = 3 Relevé = 3  
Individu 13 : Taxon = 3 Relevé = 4  
Individu 14 : Taxon = 3 Relevé = 4

Ceci s'écrit comme couple de tableaux contenant une seule variable qualitative sous forme d'indicatrices des classes. On dira qu'on a le couple des tableaux d'occurrences :

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

Les lignes identiques peuvent être regroupées. On donnera alors le terme générique de tableau de correspondances l'écriture :

1 0 0	1 0 0 0	poids = 1
1 0 0	0 0 1 0	poids = 1
0 1 0	0 1 0 0	poids = 4
0 1 0	0 0 1 0	poids = 3
0 1 0	0 0 0 1	poids = 2
0 0 1	0 0 1 0	poids = 1
0 0 1	0 0 0 1	poids = 2

La même information est consignée dans le tableau final (relevés en lignes, taxons en colonnes). On utilise alors le terme générique de table de contingence :

1	0	0
0	4	0
1	3	1
0	2	2

En général, on se donnera les notations suivantes.  $I$  est le nombre de relevés,  $J$  est le nombre de taxon,  $O$  est le nombre d'occurrences et  $C$  est le nombre de correspondances. Notons  $\mathbf{X}$  le tableau croisant  $C$  occurrences et  $J$  taxons et  $\mathbf{Y}$  le tableau croisant  $C$  occurrences et  $I$  relevés. La table de contingence  $\mathbf{T}$  est simplement la matrice  $\mathbf{T} = \mathbf{Y}^t \mathbf{X}$ .

On notera, comme d'habitude :

$$\mathbf{T} = [n_{ij}]_{1 \leq i \leq I, 1 \leq j \leq J} \quad n_{i.} = \sum_{j=1}^J n_{ij} \quad n_{.j} = \sum_{i=1}^I n_{ij} \quad n_{..} = \sum_{j=1}^J n_{.j}$$

Dans ces conditions  $O = n_{..}$ . Les distributions marginales sont :

$$f_{i.} = \frac{n_{i.}}{n_{..}} \quad \sum_{i=1}^I f_{i.} = 1 \quad f_{.j} = \frac{n_{.j}}{n_{..}} \quad \sum_{j=1}^J f_{.j} = 1$$

On note :

$$\mathbf{F} = [f_{ij}]_{1 \leq i \leq I, 1 \leq j \leq J} \quad \mathbf{F}_0 = [f_{ij} - f_{i.}f_{.j}]_{1 \leq i \leq I, 1 \leq j \leq J}$$

Les distributions conditionnelles sont :

$$f_{ij|j} = \frac{n_{ij}}{n_{.j}} \quad \sum_{j=1}^J f_{ij|j} = 1 \quad f_{j|i} = \frac{n_{ij}}{n_{i.}} \quad \sum_{i=1}^I f_{j|i} = 1$$

Le relevé prend un poids  $f_{i.}$ , le taxon prend un poids  $f_{.j}$ . On conserve les poids des espèces dans la matrice diagonale  $\mathbf{D}_J = \text{Diag}(f_{.1}, \dots, f_{.j}, \dots, f_{.J})$  et le poids des relevés dans  $\mathbf{D}_I = \text{Diag}(f_{1.}, \dots, f_{i.}, \dots, f_{I.})$ .

## 2 — L'analyse des correspondances

C'est la présentation sous la forme du couple des tableaux d'occurrences  $\mathbf{X}$  et  $\mathbf{Y}$  qui nous intéresse. Chacun des deux peut être examiné séparément et chacun des deux engendre une partition sur les lignes de l'autre. Enfin, ils sont appariés par les lignes de manière symétrique si on le veut. Toute la question de l'analyse des correspondances se retrouve ici.

### 2.1 — Centrage des indicatrices

$\mathbf{X}$  est le tableau des occurrences des espèces. Il a  $O$  lignes et  $J$  colonnes. Chaque occurrence porte le poids  $1/O$ . A la ligne  $o$  et à la colonne  $j$  on rencontre la valeur 0 ou 1. La somme des valeurs de la colonne  $j$  est le nombre d'apparition du taxon  $j$  et la moyenne de la colonne  $j$  est la fréquence  $f_{.j}$ . La variance associée de la colonne  $j$  est alors simplement la quantité  $f_{.j}(1 - f_{.j})$ .

Le tableau centré  $\mathbf{X}_0$  a  $O$  lignes et  $J$  colonnes. A la ligne  $o$  et à la colonne  $j$  on rencontre la valeur  $-f_{.j}$  ou  $1 - f_{.j}$ . L'inertie totale de ce tableau s'écrit comme somme des variances, soit :

$$I_T = \sum_{j=1}^J f_{.j}(1 - f_{.j}) = 1 - \sum_{j=1}^J f_{.j}^2 = \text{Div}_S(f_{.1}, \dots, f_{.J})$$

On retrouve l'indice de diversité de Simpson (Cf. 4). Le tableau centré définit encore une matrice de covariances entre indicatrices centrées sous la forme :

$$\mathbf{C} = \mathbf{X}_0^t \mathbf{D}_{1/O} \mathbf{X}_0 = [Cov(j,k)]_{j, k} \quad j, k = 1, \dots, J$$

avec  $Cov(j,k) = -f_j f_k \quad Cov(j,j) = f_j(1 - f_j)$

Cette matrice peut encore s'écrire :

$$\mathbf{C} = \mathbf{X}_0^t \mathbf{D}_{1/O} \mathbf{X}_0 = \mathbf{D}_J - \mathbf{V}_J \mathbf{V}_J^t \quad \mathbf{V}_J^t = [f_1, \dots, f_j, \dots, f_J]$$

Un tableau d'occurrences permet de définir le triplet  $(\mathbf{X}_0, \mathbf{Id}_J, \mathbf{D}_{1/O})$  dont l'analyse n'a pas de sens.

$\mathbf{Y}$  est le tableau des occurrences des relevés. Il a  $O$  lignes et  $I$  colonnes. Chaque occurrence porte le poids  $1/O$ . A la ligne  $o$  et à la colonne  $i$  on rencontre la valeur 0 ou 1. La somme des valeurs de la colonne  $i$  est le nombre d'apparition du relevé  $i$  et la moyenne de la colonne  $i$  est la fréquence  $f_i$ . La variance associée de la colonne  $i$  est alors simplement la quantité  $f_i(1 - f_i)$ .

La symétrie est parfaite entre espèces et relevés. Le tableau centré  $\mathbf{Y}_0$  a  $O$  lignes et  $I$  colonnes. A la ligne  $o$  et à la colonne  $i$  on rencontre la valeur  $-f_i$  ou  $1 - f_i$ . L'inertie totale de ce tableau s'écrit comme somme des variances, soit :

$$I_T = \sum_{i=1}^I f_i(1 - f_i) = 1 - \sum_{i=1}^I f_i^2$$

La référence à l'indice de diversité de Simpson n'a plus de sens écologique. Le tableau centré définit encore une matrice de covariances entre indicatrices centrées sous la forme :

$$\mathbf{D} = \mathbf{Y}_0^t \mathbf{D}_{1/O} \mathbf{Y}_0 = [Cov(i,l)]_{i, l} \quad i, l = 1, \dots, I$$

avec  $Cov(i,l) = -f_i f_l \quad Cov(i,i) = f_i(1 - f_i)$

Cette matrice peut encore s'écrire :

$$\mathbf{D} = \mathbf{Y}_0^t \mathbf{D}_{1/O} \mathbf{Y}_0 = \mathbf{D}_I - \mathbf{V}_I \mathbf{V}_I^t \quad \mathbf{V}_I^t = [f_1, \dots, f_i, \dots, f_I]$$

Ce tableau d'occurrences permet de définir le triplet  $(\mathbf{Y}_0, \mathbf{Id}_I, \mathbf{D}_{1/O})$  dont l'analyse n'a pas de sens.

On observera, en outre, que les matrices de covariances ont des inverses généralisés définies analytiquement par ( $\mathbf{1}_I$  est la matrice à  $I$  lignes et  $I$  colonnes de valeurs toutes égales à 1) :

$$\mathbf{D}^- = (\mathbf{D}_I - \mathbf{V}_I \mathbf{V}_I^t)^- = \mathbf{D}_I^{-1} - \mathbf{1}_I$$

En effet, on vérifie les deux relations fondamentales :

$$(\mathbf{D}_I^{-1} - \mathbf{1}_I)(\mathbf{D}_I - \mathbf{V}_I \mathbf{V}_I^t)(\mathbf{D}_I^{-1} - \mathbf{1}_I) = \mathbf{D}_I^{-1} - \mathbf{1}_I$$

$$\left(\mathbf{D}_I - \mathbf{V}_I \mathbf{V}_I^t\right) \left(\mathbf{D}_I^{-1} - \mathbf{1}_{II}\right) \left(\mathbf{D}_I - \mathbf{V}_I \mathbf{V}_I^t\right) = \mathbf{D}_I - \mathbf{V}_I \mathbf{V}_I^t$$

De même :

$$\mathbf{C}^- = \left(\mathbf{D}_J - \mathbf{V}_J \mathbf{V}_J^t\right)^- = \mathbf{D}_J^{-1} - \mathbf{1}_{JJ}$$

On retrouve les métriques dites du Khi2.

## 2.2 — Discrimination des occurrences espèces entre les relevés

Le tableau des occurrences n'a, en général pas d'existence réelle, mais sa présence conceptuelle en analyse des correspondances est considérable. Sir R. Fisher<sup>5</sup> est souvent considéré comme le père de l'analyse des correspondances <sup>6</sup> parce qu'il a le premier appliqué l'analyse discriminante à un tableau d'occurrences.

Les deux tableaux d'occurrences définissent donc deux analyses discriminantes. La première fait la discrimination sur  $\mathbf{X}_0$  avec la variable qualitative définie dans  $\mathbf{Y}$ . On calcule, pour chaque classe, la moyenne des lignes du tableau  $\mathbf{X}_0$  associées à cette classe et on affecte au résultat le poids de la classe.

Le calcul est immédiat et on obtient le triplet statistique :

$$\left(\mathbf{L}_0, \mathbf{D}_J^{-1} - \mathbf{1}_{JJ}, \mathbf{D}_I\right) \quad \mathbf{L}_0 = \left[ f_{j|i} - f_j \right]_{i \quad I,1 \quad j \quad J}$$

strictement équivalent au triplet :

$$\left(\mathbf{L}_0, \mathbf{D}_J^{-1}, \mathbf{D}_I\right)$$

On obtient des codes canoniques des espèces centrés pour la pondération marginale des taxons, de variance unité maximisant la variance des positions moyennes par relevés. Le pourcentage de variance inter-classes est la valeur propre associée. Les résultats numériques sont strictement ceux d'une analyse des correspondances sur le tableau  $\mathbf{T}$  (COA : COReSPondence ANalysis), les taxons utilisant les scores de variance 1 (DDUtil : Add normed scores) et les relevés étant positionnés par les coordonnées ordinaires. La signification écologique du principe statistique est claire <sup>7</sup>.

## 2.3 — Discrimination des occurrences relevés entre les espèces

Le raisonnement s'étend à la situation symétrique. La seconde analyse discriminante annoncée fait la discrimination sur  $\mathbf{Y}_0$  avec la variable qualitative définie dans  $\mathbf{X}$ . On calcule, pour chaque classe, la moyenne des lignes du tableau  $\mathbf{Y}_0$  associées à cette classe et on affecte au résultat le poids de la classe. Par souci de conserver la même disposition du tableau ( $I$  lignes et  $J$  colonnes), les classes sont positionnées en colonnes. On obtient le triplet statistique :

$$\left(\mathbf{K}_0, \mathbf{D}_J, \mathbf{D}_I^{-1} - \mathbf{1}_{II}\right) \quad \mathbf{K}_0 = \left[ f_{i|j} - f_i \right]_{i \quad I,1 \quad j \quad J}$$

strictement équivalent au triplet :

$$\left(\mathbf{K}_0, \mathbf{D}_J, \mathbf{D}_I^{-1}\right)$$

On obtient des codes canoniques des relevés centrés pour la pondération marginale des relevés, de variance unité maximisant la variance des positions moyennes par

espèces. Le pourcentage de variance inter-classes est la valeur propre associée. Les résultats numériques sont strictement ceux d'une analyse des correspondances sur le tableau **T** (COA : CO<sup>R</sup>respondence Analysis), les relevés utilisant les scores de variance unité (DDUtil : Add normed scores) et les taxons étant positionnés par les coordonnées ordinaires.

Le fait que deux discriminantes sont assurées par la même procédure souligne la symétrie explicite lignes-colonnes dans une analyse des correspondances. Les doubles représentation par averaging sont explicitées en détail dans <sup>8</sup> et présentées en écologie par <sup>9</sup>.

## 2.4 — Analyse canonique des occurrences

Le fond de la spécificité de l'AFC est obtenue avec le point de vue de l'analyse canonique. Il s'agit simplement de l'analyse du tableau croisé :

$$\mathbf{F}_0 = \mathbf{Y}_0^\dagger \mathbf{D}_{VO} \mathbf{X}_0$$

En effet, quand on calcule les covariances croisées de la colonne  $j$  du tableau  $\mathbf{X}_0$  et de la colonne  $i$  du tableau  $\mathbf{Y}_0$ , on additionne quatre types de valeurs avec les fréquences respectives :

<i>Valeurs</i>	<i>Fréquences</i>
$f_i \cdot f_j$	$1 - f_i - f_j + f_{ij}$
$-f_i (1 - f_j)$	$f_j - f_{ij}$
$-f_j (1 - f_i)$	$f_i - f_{ij}$
$(1 - f_j)(1 - f_i)$	$f_{ij}$

Si on faisait une co-inertie des deux tableaux d'occurrences, on étudierait simplement l'analyse du triplet  $(\mathbf{F}_0, \mathbf{I}_{d_j}, \mathbf{I}_{d_I})$  qui reste à envisager, mais pour une analyse canonique, on introduit les deux normes inverses pour étudier le triplet :

$$(\mathbf{F}_0, \mathbf{D}_J^{-1} - \mathbf{1}_{JJ}, \mathbf{D}_I^{-1} - \mathbf{1}_{II})$$

équivalent pour des raisons simples de centrage à  $(\mathbf{F}_0, \mathbf{D}_J^{-1}, \mathbf{D}_I^{-1})$ .

La solution est encore l'analyse des correspondances dans la version découverte par <sup>10</sup> et popularisée en écologie par <sup>11</sup>. L'équivalence stricte des trois approches est décrite dans <sup>12</sup>. L'AFC, comme analyse canonique des tableaux d'occurrences est sans concurrent. Elle seule a cette propriété qui conduit au *reciprocal scaling* de l'amplitude des espèces et de la diversités des relevés <sup>3</sup>. Les points de vue utilisateur de cette partie théorique sont illustrées dans la documentation du module COA.

## 2.5 — Les fichiers d'occurrences

La présentation d'un tableau floristique ou faunistique par matrice d'occurrences est un exercice théorique dans la plupart des cas. Certes, les données ont été consignées d'abord dans un cahier de terrain, mais dès que le tableau relevés-taxons a été constitué on ne revient pas en arrière et toute l'information consignée dans le tableau final est équivalente strictement aux enregistrements initiaux.

Il peut en être tout autrement avec l'exploration des gisements des observations naturalistes, qui s'apparente au *Data mining*. Les exemplaires conservés dans les collections des musées d'histoire naturelle, en particulier dans les herbiers, indiquent qu'on a rencontré à une date connue, en un lieu précis, un individu d'un taxon donné.

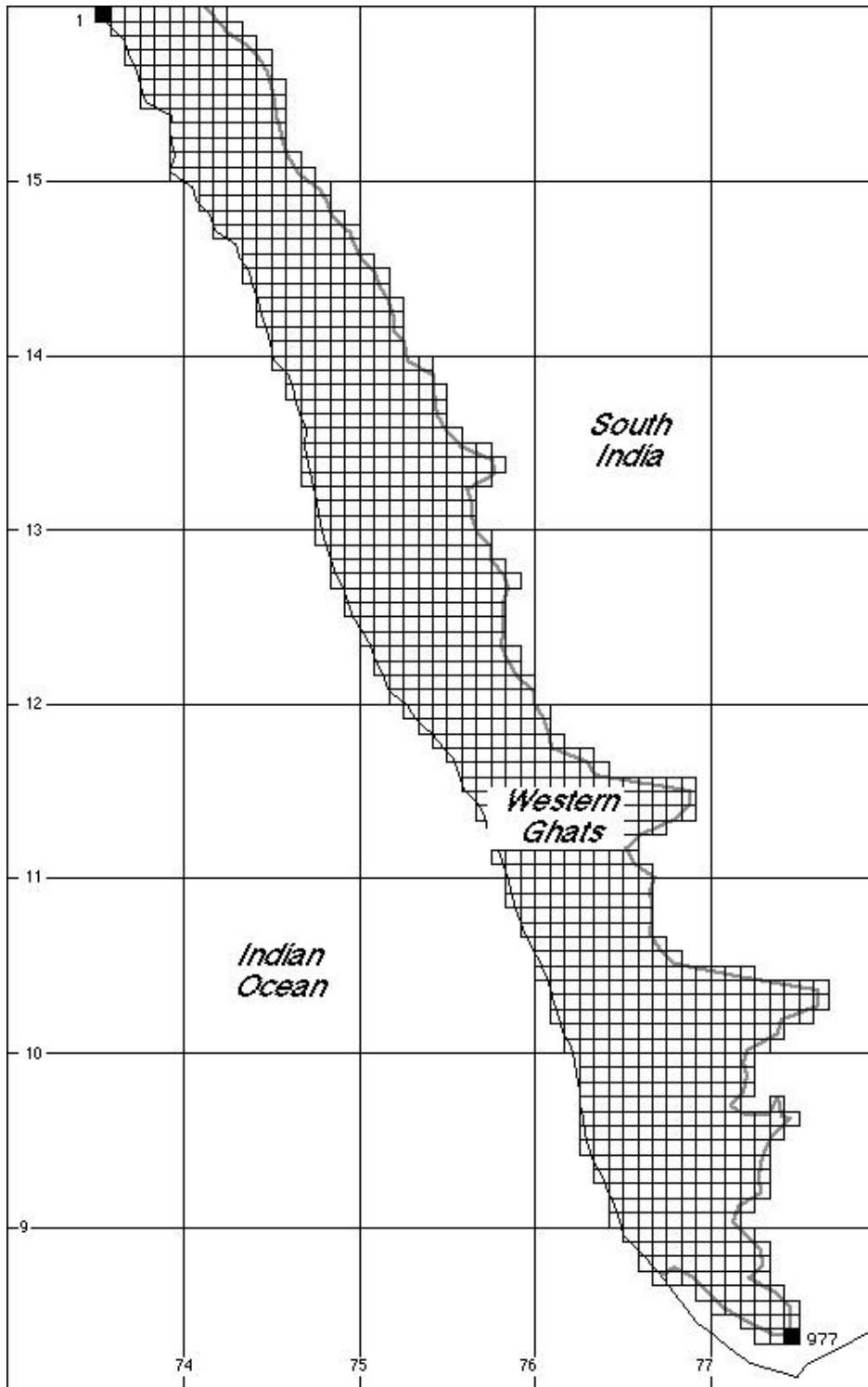


Figure 1 : Définition de 977 quadrats de 5 minutes de côté.

Il n'y a pas eu forcément relevé au sens expérimental du terme, mais simplement rencontre entre — par exemple — un botaniste et une plante. Le fichier d'occurrence existe alors comme information de base incompressible sous la forme :



## Nom d'espèce, date de récolte, auteur, lieu, conditions de milieu,...

En collectant ces occurrences, on peut parvenir à des listes fort longues. Il ne s'agit pas d'un échantillonnage *objectif* mais d'un ensemble de rencontres. C'est ainsi que J.P. Pascal a réuni, lors de la constitution d'un atlas en cours d'édition, plus de 10 000 occurrences d'espèces endémiques des forêts sempervirentes des Ghâts occidentaux (Ouest de l'Inde). Les observations réunissent les exemplaires de la plupart des herbiers connus, des observations publiées dans plusieurs centaines d'articles et des relevés ordinaires d'un institut de recherche. Les dates d'enregistrement débutent en 1815.

<b>8493</b>	289	8.52333333	77.4533333	F1	1988	1070
<b>8494</b>	289	8.53333333	77.1416667	F1	1988	1070
<b>8495</b>	289	8.53333333	77.45	H	1988	1072
<b>8496</b>	289	8.53666667	77.4533333	F1	1988	1070
<b>8497</b>	289	8.55833333	77.2083333	F1	1988	1070
<b>8498</b>	289	8.57333333	77.4083333	F1	1988	1070
<b>8499</b>	289	8.58833333	77.3166667	F1	1988	1070
<b>8500</b>	289	8.59166667	77.305	F1	1988	1070
<b>8501</b>	289	8.63333333	77.25	H	1917	243
<b>8502</b>	289	8.64166667	77.25	F1	1988	1070
<b>8503</b>	289	8.66666667	77.15	H	1971	683
<b>8504</b>	289	8.71666667	77.2333333	H	1956	394
<b>8505</b>	289	8.73333333	77.0833333	H	1955	366
<b>8506</b>	289	8.75166667	77.1166667	F1	1988	1070
<b>8507</b>	289	8.79166667	77.1183333	F1	1988	1070
<b>8508</b>	289	8.805	77.2066667	F1	1988	1070
<b>8509</b>	289	8.82166667	77.2066667	F1	1988	1070
<b>8510</b>	289	8.88333333	77.1	H	1956	392
<b>8511</b>	289	8.9	77.1833333	H	1957	414

On reconnaît sur cet échantillon le numéro de l'occurrence, le numéro de code du taxon, la longitude et la latitude en degrés centésimaux, le type de l'observation, l'année de l'observation, le numéro de campagne (groupe d'observations enregistrées la même année par un même auteur). La position spatiale permet de récupérer les variables fondamentales de l'environnement (altitude, pluviométrie, géomorphologie).

Les points d'occurrence peuvent être regroupés dans l'espace par "quadrats" de 5 minutes d'amplitude (figure 1). Cette division de l'espace introduit une partition dans l'espace des occurrences. On voit dans cette situation que le statut des deux variables qualitatives (numéro du taxon, numéro du quadrat) n'ont pas de statut symétrique. La dissymétrie entre les deux variables qualitatives est la base de l'introduction de l'ANSC.

En faisant une analyse discriminante, on fait automatiquement l'autre et donc une analyse canonique. La symétrie est totale dans la partie statistique de l'AFC. Cette symétrie a un grand intérêt dans le cadre d'un statut symétrique des deux variables. L'archétype est donné dans l'exemple couleur des cheveux - couleur des yeux sur  $O$  individus<sup>5 8</sup>. Dans un ensemble de relevés classiques, la liste des relevés est contrôlée et la liste des taxons rencontrés est une conséquence. Il y a déjà dissymétrie. Mais en partant d'une liste d'occurrences préétablies, puis en construisant une partition, la dissymétrie est totale. On peut en effet changer de partition mais pas de liste. D'un côté on a des observations, de l'autre des contraintes d'analyse de ces observations.

On peut ranger les personnes par couleur des yeux et voir si le profil de la couleur des cheveux évolue : l'inverse est aussi légitime et le plus simple est de faire les deux d'un coup. On peut ranger les occurrences espèces par quadrat et voir si le profil floristique évolue d'un quadrat à l'autre. L'inverse se justifie mal.

Dans l'étude précitée, après répartition des occurrences espèces par quadrat, l'AFC donne des valeurs propres représentées dans la figure 2. On observe immédiatement que les trois premiers facteurs sont des parasites : un quadrat pauvre contient une espèce rare et l'analyse met en avant cette tête d'épingle dans le nuage de points avant toute indication d'une organisation quelconque.

On peut se demander si ce comportement pathologique disparaît avec l'approche non symétrique.

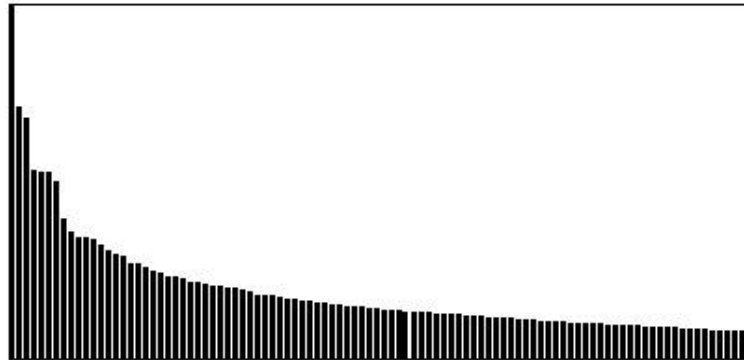


Figure 2 : Valeurs propres de l'AFC d'un tableau distribuant 10080 occurrences d'espèces endémiques entre 583 quadrats de 12 ' de côté. Les trois premières valeurs propres correspondent à des correspondances pointues associant un taxon et un quadrat.

## 3 — Principes

### 3.1 — L'indice CATANOVA

L'approche non symétrique des deux tableaux d'occurrences voit dans l'un des deux un tableau d'observation et dans l'autre l'indicatrice de mise en classes des observations. Comme on voit habituellement les individus en lignes et les classes comme paquets de lignes, nous prenons  $\mathbf{X}_0$  comme tableau des observation et  $\mathbf{Y}$  comme indicateur des classes :

Individu 1	: Taxon = 1	Relevé = 1	1	0	0
Individu 2	: Taxon = 2	Relevé = 2	0	1	0
Individu 3	: Taxon = 2	Relevé = 2	0	1	0
Individu 4	: Taxon = 2	Relevé = 2	0	1	0
Individu 5	: Taxon = 2	Relevé = 2	0	1	0
Individu 6	: Taxon = 1	Relevé = 3	1	0	0
Individu 7	: Taxon = 2	Relevé = 3	0	1	0
Individu 8	: Taxon = 2	Relevé = 3	0	1	0
Individu 9	: Taxon = 2	Relevé = 3	0	1	0
Individu 10	: Taxon = 3	Relevé = 3	0	0	1
Individu 11	: Taxon = 2	Relevé = 4	0	1	0
Individu 12	: Taxon = 2	Relevé = 4	0	1	0
Individu 13	: Taxon = 3	Relevé = 4	0	0	1
Individu 14	: Taxon = 3	Relevé = 4	0	0	1

Un test de signification basé sur l'inertie inter-classe est défini par Light & Margolin<sup>2</sup> sous le nom de CATANOVA pour analyse de variance sur variable qualitative. Nous avons vu que le tableau centré  $\mathbf{X}_0$  a  $O$  lignes et  $J$  colonnes. A la ligne  $o$  et à la colonne  $j$  on rencontre la valeur  $-f_j$  ou  $1 - f_j$ . L'inertie totale de ce tableau s'écrit comme somme des variances, soit :

$$I_T = \sum_{j=1}^J f_j (1 - f_j) = 1 - \sum_{j=1}^J f_j^2 = Div_S(f_1, \dots, f_J)$$

En raisonnant sur les sommes de carrés d'écart (*total sum of squares, to maintain a terminology consistent with the standard analysis of variance*) au lieu des moyennes cela s'écrit (op. cit. p. 535)

$$TSS = \frac{O}{2} \left( 1 - \sum_{j=1}^J f_{.j}^2 \right)$$

Cette variance totale se décompose pour chacune des colonnes de  $\mathbf{X}_0$  en variance inter-classe et variance intra-classe. L'inertie inter-classes s'écrit :

$$B = \sum_{i=1}^I \sum_{j=1}^J f_{ji} \left( f_{ji} - f_{.j} \right)^2 = \sum_{i=1}^I \sum_{j=1}^J f_{ji}^2 - \sum_{j=1}^J f_{.j}^2$$

La somme des carrés inter-classes s'écrit (op.cit. p. 535)

$$BSS = \frac{O}{2} \left( \sum_{i=1}^I \sum_{j=1}^J f_{ji}^2 - \sum_{j=1}^J f_{.j}^2 \right)$$

La proportion de variance expliquée est :

$$R^2 = \frac{BSS}{TSS} = \frac{B}{I_T}$$

Il est tout à fait curieux que le lien entre ANOVA et variable qualitative fait ici par les indicatrices des classes centrées et l'inertie comme somme de variances est fait par les auteurs cités par l'intermédiaire de la distance entre deux occurrences qui vaut 0 si elles concernent la même catégorie et 1 dans le cas contraire. Le passage de la variance vue comme carré de la norme de l'indicatrice centrée à la variance vue comme moyenne des carrés des distances des points deux à deux se fait par (op. cit. p. 534) :

$$(n-1)Var(x) = \sum_{k=1}^n (x_k - moy(x))^2 = \frac{1}{2n} \sum_{i=1}^n \sum_{j=1}^n (x_i - x_j)^2 = \frac{1}{2n} \sum_{i=1}^n \sum_{j=1}^n d_{ij}^2$$

On retrouve exactement le procédé qu'avait utilisé Lebart <sup>13</sup> pour introduire les graphes de voisinages en analyse des données.

L'indice CATANOVA s'écrit alors par (op. cit. p. 538) :

$$C = \frac{(n-1)(J-1)BSS}{TSS} = (n-1)(J-1) \frac{B}{I_T}$$

et se teste comme un Khi2 à  $(I-1)(J-1)$  degrés de liberté.

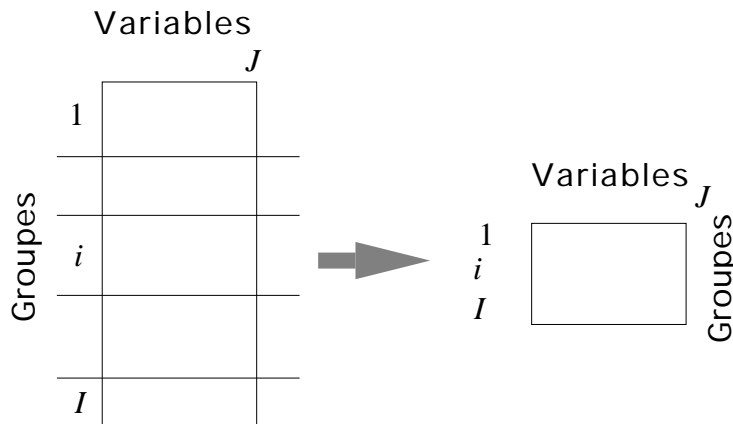
Le coefficient  $R^2$  qui est réécrit par Lauro & D'Ambra <sup>1</sup> (p. 435) sous la forme :

$$R^2 = \frac{\sum_{i=1}^I \sum_{j=1}^J f_{ji}^2 - \sum_{j=1}^J f_{.j}^2}{\sum_{j=1}^J f_{.j}^2} = \frac{\sum_{i=1}^I \sum_{j=1}^J \frac{f_{ij}^2}{f_{.i}} - \sum_{j=1}^J f_{.j}^2}{\sum_{j=1}^J f_{.j}^2}$$

soit encore :

$$R^2 = \frac{\text{Trace}(\mathbf{F}^t \mathbf{D}_I^{-1} \mathbf{F} - \mathbf{V}_J \mathbf{V}_J^t)}{\text{Trace}(\mathbf{D}_J - \mathbf{V}_J \mathbf{V}_J^t)}$$

On reconnaît les mêmes quantités dans le langage du schéma de dualité, soit au dénominateur l'inertie totale de l'analyse de départ et au numérateur l'inertie totale de l'analyse inter-classe. Dans les deux articles cités, les auteurs ont choisi l'indicatrice des colonnes comme variable de contrôle (définition des classes) et l'indice ligne comme variable d'état (définition des observations). Pour conserver la logique :



nous avons fait l'inverse. Les lignes du tableau de contingence définissent les classes du tableau des occurrences. Les relations rapportées sont strictement celles des articles cités à cette permutation près.

### 3.2 — Propriétés de l'ANSC

On appellera donc ANSC l'analyse du triplet :

$$(\mathbf{L}_0, \mathbf{Id}_J, \mathbf{D}_I) \quad \mathbf{L}_0 = [f_{j|i} - f_{.j}]_{i=1, \dots, I, j=1, \dots, J}$$

On reconnaît une simple inter-classe des indicatrices centrées de  $\mathbf{X}_0$  avec la variable  $\mathbf{Y}$ , utilisant les poids des lignes (poids des classes) et la pondération uniforme des colonnes (et non la métrique du Khi2 comme l'analyse des correspondances ordinaires). Cette analyse est disponible dans l'option COA : NSCA\_Row\_Profiles.

Les coordonnées des lignes sont centrées pour la pondération lignes de la table de contingences. En effet :

$$\mathbf{L}_0^t \mathbf{D}_I \mathbf{1}_I = 0 \quad \sum_{i=1}^I (f_{j|i} - f_{.j}) f_{i.} = \sum_{i=1}^I f_{ij} - f_{.j} = 0$$

Les coordonnées des lignes, proportionnelles aux composantes principales, sont donc orthogonales au vecteur  $\mathbf{1}_I$ , donc centrées. Les coordonnées des colonnes sont centrées pour la pondération uniforme. En effet :

$$\mathbf{L}_0^1 \mathbf{1}_J = 0 \quad \sum_{j=1}^J (f_{j|i} - f_{.j}) = 1 - 1 = 0$$

Les coordonnées des colonnes sont de covariances nulles pour la pondération uniforme tandis que les coordonnées des lignes sont de covariances nulles pour la pondération marginale.

Si on applique cette analyse à un tableau floro-faunistique (groupes d'occurrences en lignes, taxons en colonnes), on a les propriétés suivantes.

Les axes principaux (coordonnées colonnes de norme 1) de cette analyse sont des codes colonnes centrés et normés pour la pondération unitaire qui maximise la somme des carrés des écarts entre les moyennes conditionnelles par ligne et la position moyenne des relevés. On a donc une propriété d'averaging espèces --> relevés. La position moyenne des relevés n'est pas à l'origine.

Les composantes principales (coordonnées lignes de norme 1) de cette analyse sont des codes lignes centrés et normés pour la pondération marginale. La coordonnée d'une colonne ne représente pas la position moyenne d'une espèce mais cette position moyenne multipliée par le poids marginal de l'espèce. La somme des carrés des produits poids\*moyenne est maximisée. Donc une forte prime est accordée aux espèces abondantes et l'effet des espèces rares est fortement limitée contrairement à l'AFC. Dans un tableau d'occurrences, un grand nombre de taxons rares ne sera donc pas un facteur limitant. Un poids très faible ramène automatiquement un taxon vers l'origine.

La dissymétrie du statut lignes-colonnes est donc très forte. Une carte factorielle dans laquelle les poids modifient directement les coordonnées obtenues par averaging ne semble pas être d'un usage facile. Le programme retiendra donc la relation colonnes --> lignes (variables --> classes) qui correspond au graphe en étoile individus-classes ordinaire dans une inter-classe.

### 3.3 — Deux programmes pour une même analyse

Pour explorer la dissymétrie à la base de l'analyse nous proposons donc deux options. Ce n'est pas normal en soi. Il y a deux ANSC sur une même table de contingence : soit les occurrences-lignes sont triées par colonnes, soit les occurrences-colonnes sont triées par lignes. L'ANSC dans un cas n'est rien d'autre que l'ANSC du tableau transposé de l'autre. Il ne devrait donc y avoir qu'un programme.

Cependant, dans l'architecture de l'ensemble, les choses doivent être plus souples. Quand on aborde les questions de traits biologiques et de dispersion des espèces, les taxons sont les lignes des tableaux et les modules de couplage sont disponibles. Quand on aborde les questions de caractères de l'environnement et de contenu des communautés, les relevés sont en lignes et les modules de couplage sont disponibles. Pour couvrir tous les cas possibles (espèces ou relevés en lignes, approche inter-espèces ou inter-sites) deux options sont indispensables.

Les deux options sont dans le module COA, respectivement COA : NSCA\_Row\_Profiles et COA : NSCA\_Col\_Profiles. La première analyse le schéma :

$$(\mathbf{L}_0, \mathbf{Id}_J, \mathbf{D}_I) \quad \mathbf{L}_0 = [f_{j/i} - f_{.j}]_{i \ I, 1 \ j \ J}$$

Elle correspond à l'analyse inter-classe des occurrences colonnes entre paquets-lignes. La seconde analyse le schéma :

$$(\mathbf{K}_0, \mathbf{D}_J, \mathbf{Id}_I) \quad \mathbf{K}_0 = [f_{i/j} - f_{.i}]_{i \ I, 1 \ j \ J}$$

Elle correspond à l'analyse inter-classe des occurrences lignes entre paquets-colonnes. Quand on n'a aucune raison claire de faire l'une plutôt que l'autre, on est tout simplement en face d'une AFC ordinaire !



It has 12 rows and 1 column  
 File Hepta.zrma contains:  
   the number of rows: 24  
   the number of columns: 12  
   the total number: 2045

Le listing rappelle la constitution du triplet étudié. Il donne le test CATANOVA :

Total Simpson Diversity = 8.7853e-01  
 Inertia = 3.4944e-02  
 Light & Margolin 1971 R2 = 3.9776e-02  
 CATANOVA Index = C = 8.9432e+02  
 Approximated as Khi2 253 DF  
 Proba Khi2 > Obs = 0.0000e+00

Le triplet est diagonalisé :

-----  
 DiagORC: General program for two diagonal inner product analysis  
 Input file: Hepta.zrta  
 --- Number of rows: 24, columns: 12  
 -----

Total inertia: 0.0349442  
 -----

Num.	Eigenval.	R.Iner.	R.Sum	Num.	Eigenval.	R.Iner.	R.Sum
01	+1.8420E-02	+0.5271	+0.5271	02	+4.5225E-03	+0.1294	+0.6565
03	+3.3702E-03	+0.0964	+0.7530	04	+2.3824E-03	+0.0682	+0.8212
05	+2.0896E-03	+0.0598	+0.8810	06	+1.6250E-03	+0.0465	+0.9275
07	+1.5254E-03	+0.0437	+0.9711	08	+7.0281E-04	+0.0201	+0.9912
09	+2.0287E-04	+0.0058	+0.9970	10	+8.7147E-05	+0.0025	+0.9995
11	+1.6225E-05	+0.0005	+1.0000	12	+0.0000E+00	+0.0000	+1.0000

File Hepta.zrvp contains the eigenvalues and relative inertia for each axis  
 --- It has 12 rows and 2 columns  
 File Hepta.zrco contains the column scores  
 --- It has 12 rows and 1 columns  
 File :Hepta.zrco

Col.	Mini	Maxi
1	-6.849e-02	6.165e-02

Le fichier ---.zrco ne sera pas utilisé. On préfère l'association ---.zrc1 (norme unité)  
 ---.zrli (lignes à la moyenne) :

File Hepta.zrli contains the row scores  
 --- It has 24 rows and 1 columns  
 File :Hepta.zrli

Col.	Mini	Maxi
1	-4.029e-01	2.581e-01

File Hepta.zrc1 contains the column scores with unit norm  
 It has 12 rows and 1 columns

File :Hepta.zrc1

Col.	Mini	Maxi
1	-5.046e-01	4.543e-01

La construction de la figure associée au principe de l'analyse se fait en plusieurs temps. Construisons d'abord une image de référence simple des données avec le module Tables :

**Values**

Input table file

H-axis position file

Column number (default = 1)

Y-axis position file

Column number (default = 1)

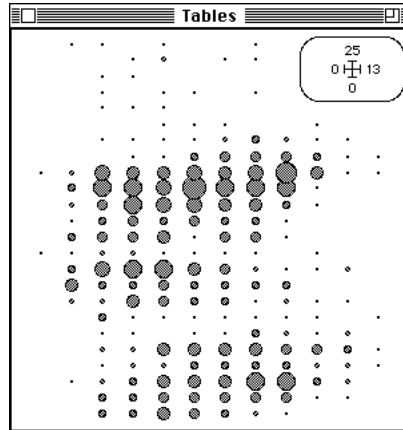
---

H-axis: Ordination (1) or Ranking (2)

Y-axis: Ordination (1) or Ranking (2)

Grid (yes = 1)

Upside down (yes = 1)



**TabMeanVar**

Input table file

H-axis position file

Column number (default = 1)

Y-axis position file

Column number (default = 1)

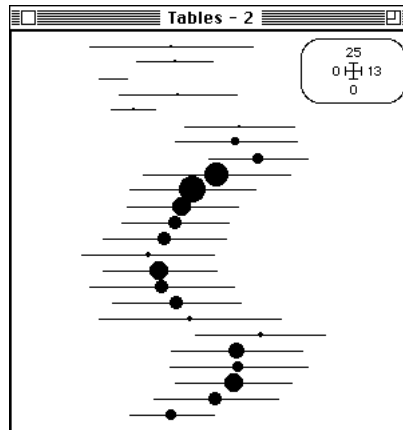
---

H-axis: Ordination (1) or Ranking (2)

Y-axis: Ordination (1) or Ranking (2)

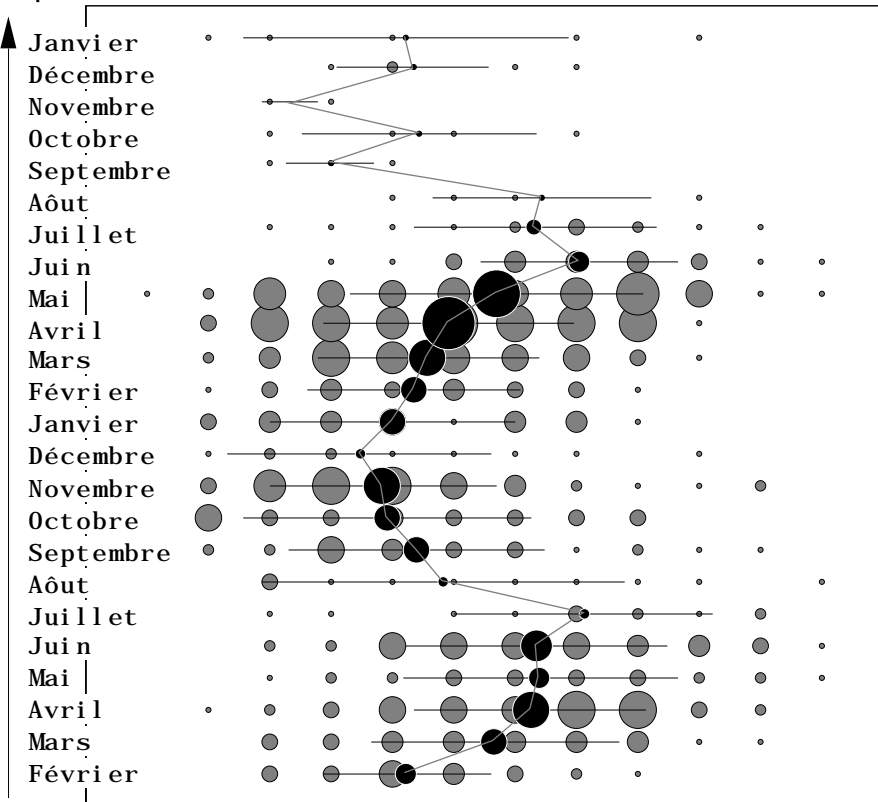
1 = Row distribution

Upside down (yes = 1)



Taille →

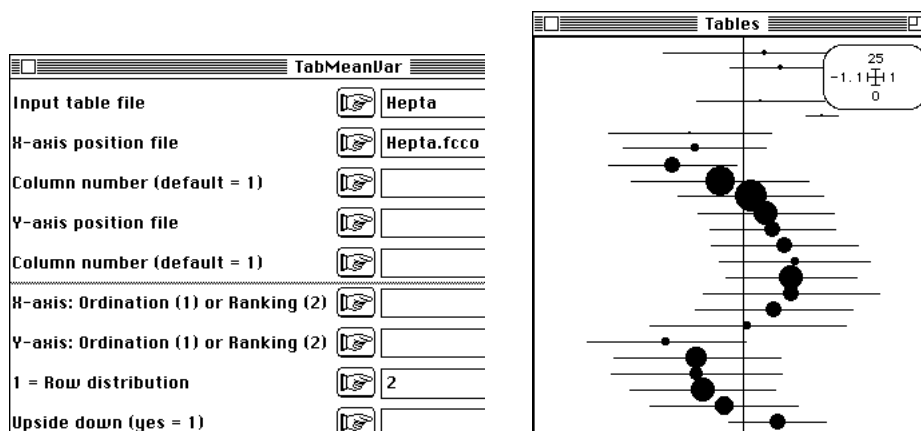
↑ Temps



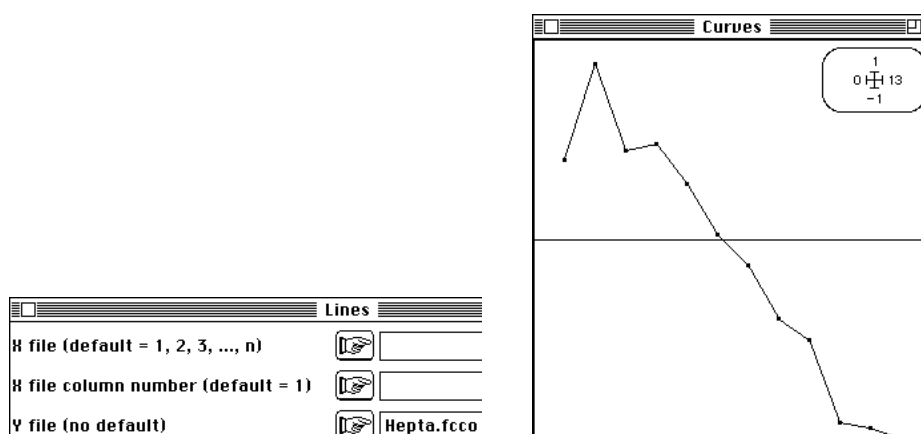


L'opération n'est rien d'autre que le tracé de la courbe de régression, c'est-à-dire de la moyenne de la taille des individus par classes.

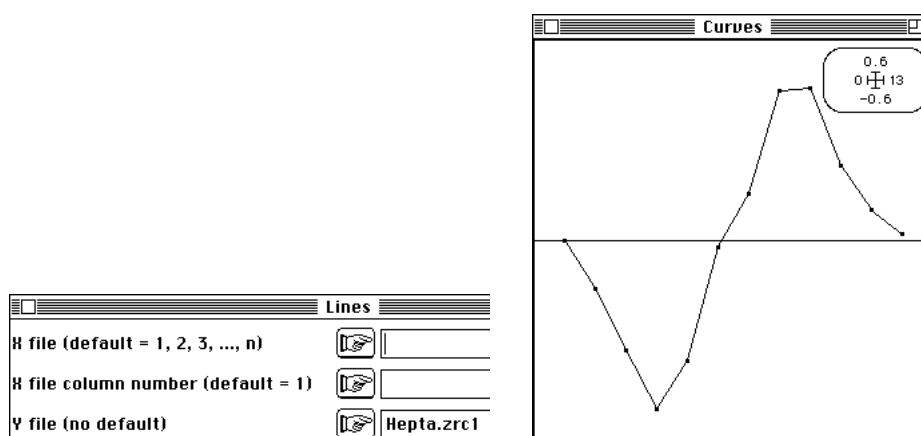
L'analyse des correspondances retrouve automatiquement cette figure (à l'orientation aléatoire près):



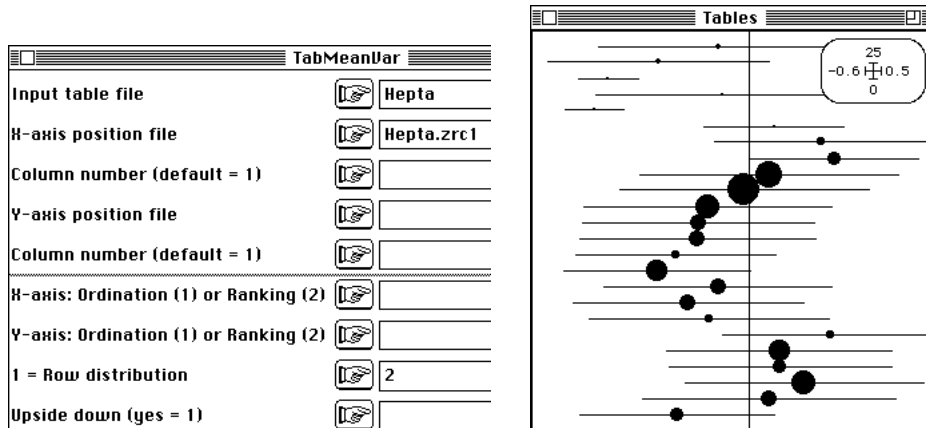
En effet la première coordonnée des colonnes de l'AFC est une fonction simple du numéro de la classe de taille :



Il est assez inattendu que l'ANSC ne fasse pas la même chose :



Ce qui donne un graphe canonique pourtant peu éloigné, en première lecture :



L'analyse non symétrique voit deux groupes d'individus (les deux cohortes) et une rupture là où l'AFC semble souligner un processus continu. Il faut bien convenir que la taille moyenne des individus ne peut diminuer que par l'émergence des adultes et le recrutement d'une nouvelle génération. La symétrie et la double régression linéaire sous-jacente <sup>17</sup> à l'AFC n'est donc pas sans conséquence. Ici l'ANSC est cohérente dans son principe avec la situation biologique et s'impose.

## 4.2 — ANSC et ordination

Pour un tableau faunistique classique l'usage de l'AFC est basé sur la notion de double averaging ou dual scaling<sup>18</sup>. Espèces à la moyenne des relevés occupés et relevés à la moyenne des espèces présentes : un des avantages décisifs est qu'il n'est nullement besoin de choisir. On a les deux modèles d'un coup.

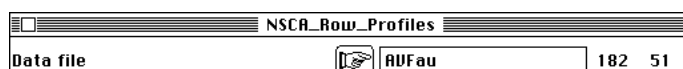
Dés qu'on s'écarte de ce point central, il faut décider. Sur un plan purement conceptuel, si les relevés sont sous la direction de l'expérimentateur (c'est lui qui décide où et quand l'abondance des taxons est mesurée) l'occurrence des espèces est un facteur aléatoire (en principe, il ne décide pas de ce qu'il convient d'y trouver). Il y a dissymétrie et le facteur contrôlé (variable instrumentale) est l'indicatrice relevé. Si les relevés sont en lignes, l'option COA : NSCA\_Row\_Profiles est autorisée et on en reste à l'averaging relevés sur espèces.

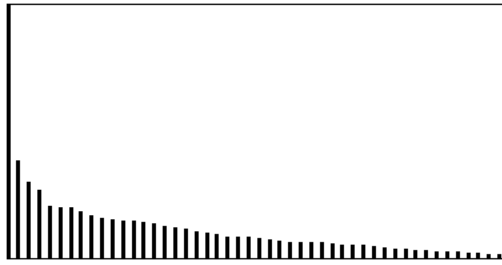
Dans ce cas, les associations ---.fcc1 (colonnes-espèces) et ---.fcli (relevés à la moyenne) comme ---.zrc1 et ---.zrli ont le même principe sous des contraintes de fonctionnement différentes. Ces contraintes peuvent-elles conduire à des interprétations différentes ? La statistique ne dira rien de plus sur ce fait et seule la pratique expérimentale donnera des indications. On peut essayer divers jeux de données bien étudiés.

En ornithologie, les données de Prodon et Lebreton<sup>19</sup> permettent un essai (carte Avi-Veg de la pile de données). En AFC :

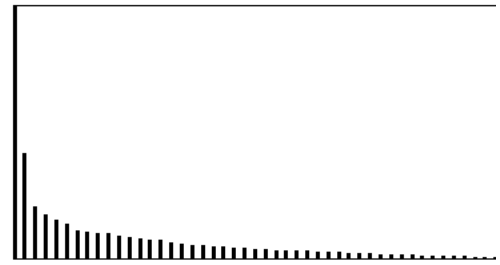


En ANSC :



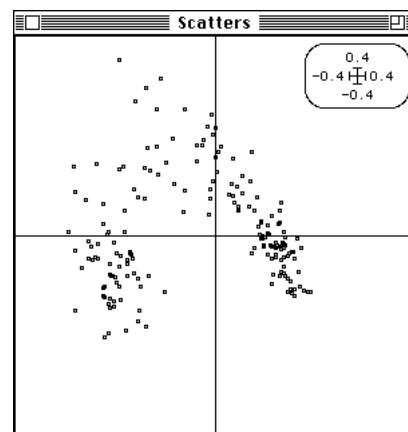
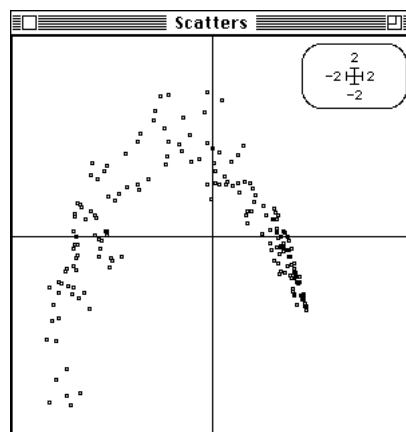
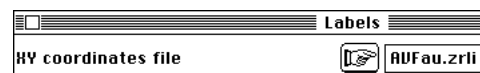
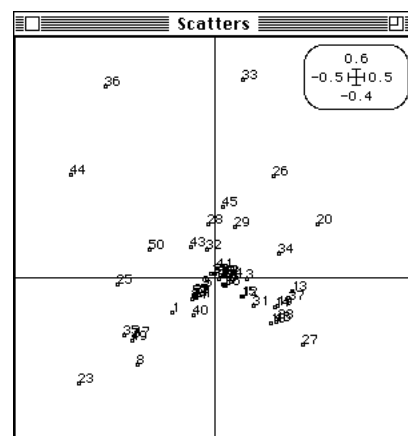
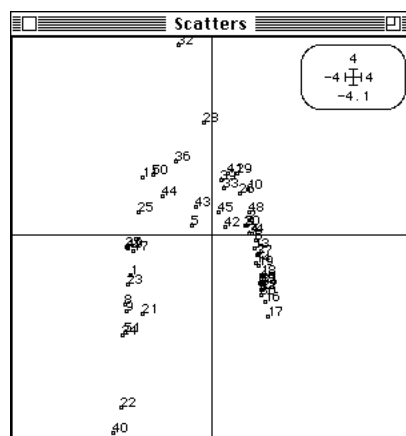
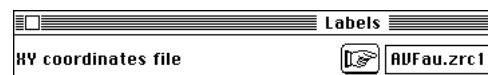


Number of axes ?



Number of axes ?

Le spectre des valeurs propres de la seconde est un peu plus clair (à gauche AFC, à droite ANSC).



Il ne suffit donc pas de préciser qu'on ordonne les relevés par averaging sur les espèces pour avoir un résultat unique. L'AFC apparaît alors comme la partie commune des deux ANSC possibles.

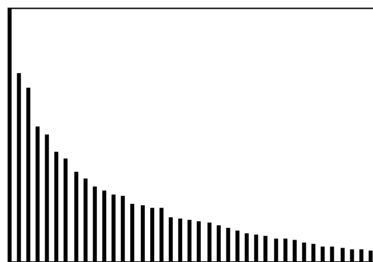
Pour le tableau de relevés phytosociologique de G. de Belair<sup>20</sup> (Carte Mafragh de la pile de données) on a :

Correspondence Analysis

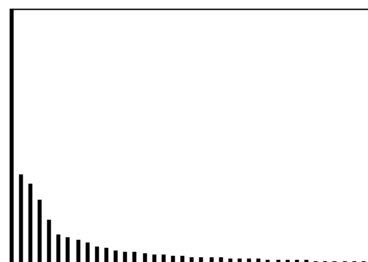
Mafragh\_Flo

NSCR\_Row\_Profiles

Mafragh\_Flo 97 56



Number of axes ? 3



Number of axes ? 5

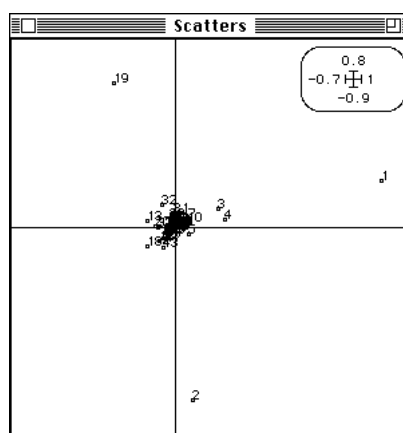
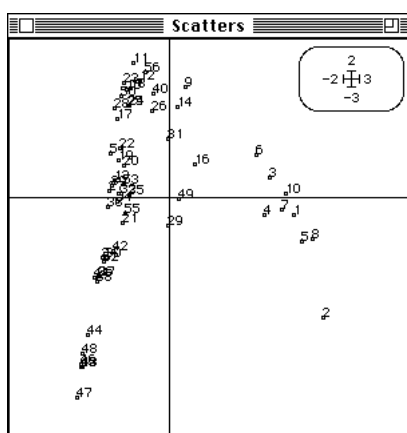
Labels

Mafragh\_Flo.fcc1

Labels

HY coordinates file

Mafragh\_Flo.zrc1



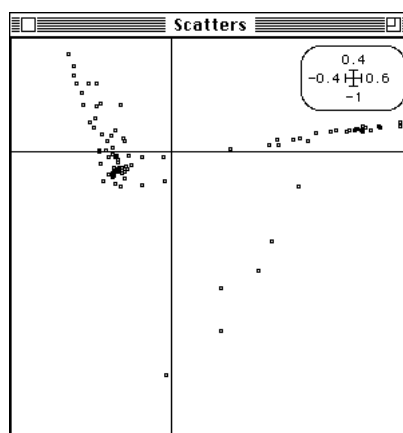
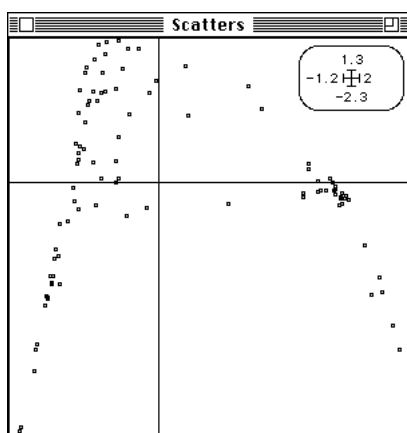
Labels

Mafragh\_Flo.fcl1

Labels

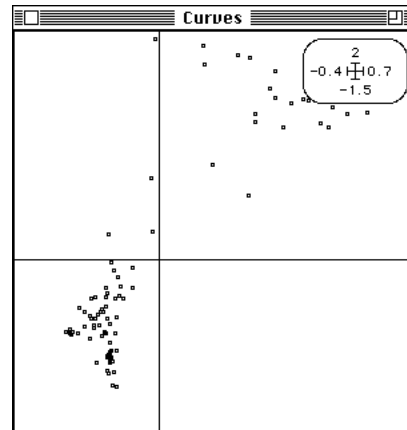
HY coordinates file

Mafragh\_Flo.zrli



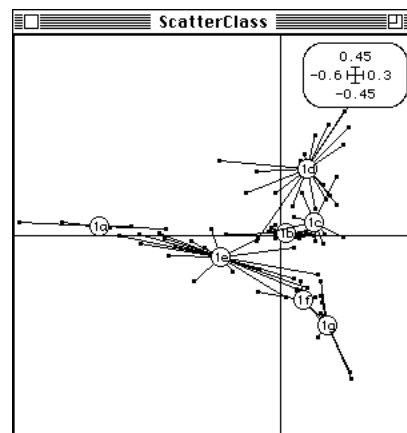
La divergence des deux appréciations ne fait guère de doutes. Et pourtant, les deux axes 1 sortent les relevés de la même association :

Lines	
X file (default = 1, 2, 3, ..., n)	Mafragh_Flo.zrli
X file column number (default = 1)	
Y file (no default)	Mafragh_Flo.fcli
Cumulated data (1=yes, 2=no)	
Variable label file (or #)	
Draw curves (1=yes, 2=no)	2



Il ne faut pas s'y tromper, l'ANSC restitue la totalité de l'information décrite dans l'interprétation des auteurs :

Stars	
HV coordinates file	Mafragh_Flo.zrli
X-axis column number (default = 1)	3
Y-axis column number (default = 2)	4
Categories file (.cat)	Mafragh_Cla.cat

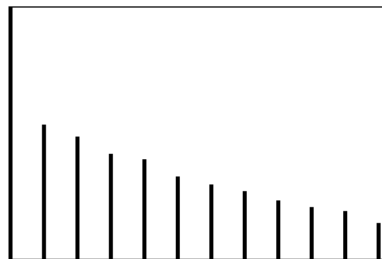


La dissymétrisation espèces-relevés semble donc limiter l'effet Guttman qui serait alors une conséquence de l'ordination réciproque. De toute manière, on peut encore affirmer que la dissymétrisation a un impact non négligeable sur les résultats.

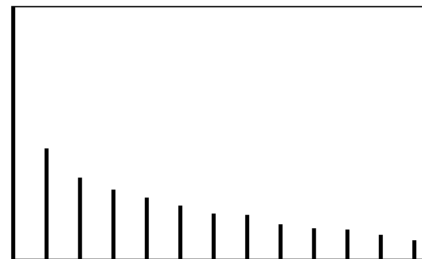
En hydrobiologie, dans le tableau de L. Friday <sup>21</sup>, les relevés sont en colonnes. Le même raisonnement conduit à utiliser l'autre option de l'ANSC :

Correspondence Analysis	
	Fau

NSCA_Col_Profiles	
	Fau

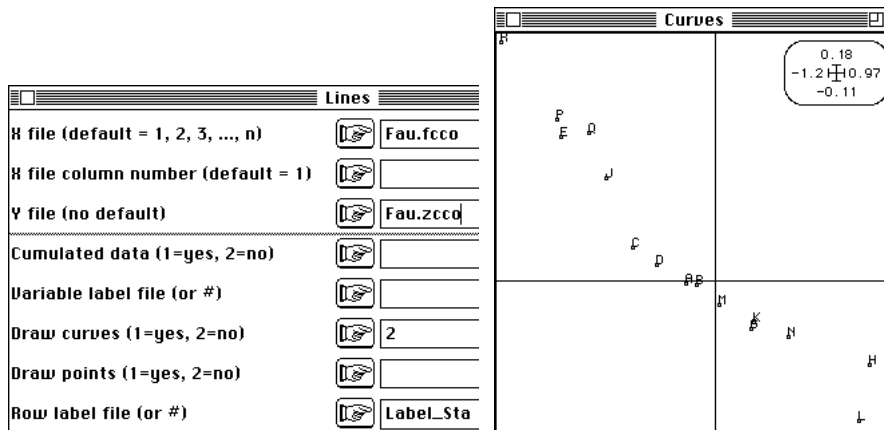


Number of axes ? 1

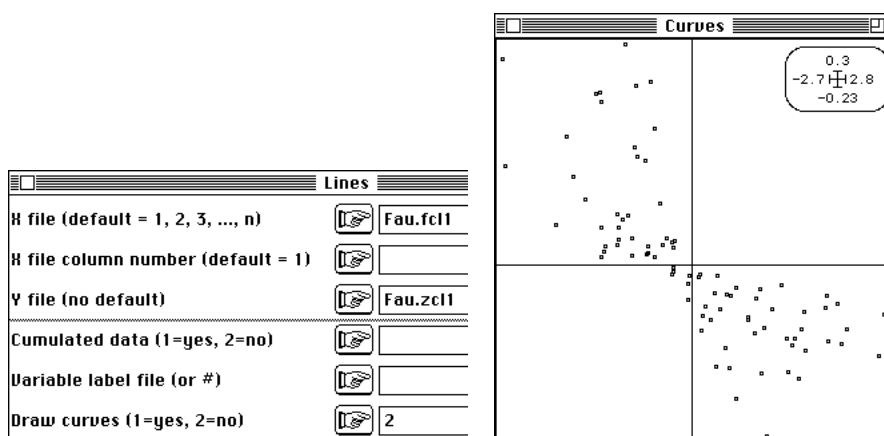


Number of axes ? 1

Il y a une même indication d'une structure unidimensionnelle. L'ordination des stations est très comparable :

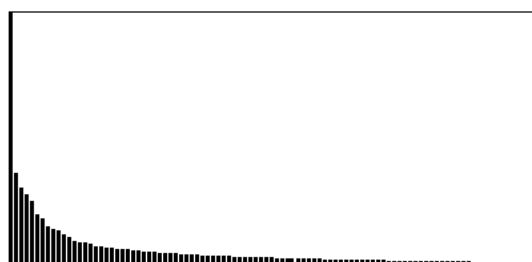


Elle est obtenue par averaging sur une ordination des espèces qui l'est beaucoup moins :



En commun, on trouve par contre une indifférence totale à la question des variations de richesse spécifique (centrage de type AFC).

Une dernière observation concerne l'exemple abordé au paragraphe 2.4 et qui a provoqué cette étude. Les valeurs propres de l'AFC sont dans la figure 2. Les valeurs propres de l'analyse non symétrique sont :



On a donc sensiblement modifié la situation et on retiendra que l'ANSC n'est pas une variante de l'AFC mais bien une nouvelle méthode d'ordination à prendre en considération. Pour des tables de contingence comptant des individus sur un facteur contrôlé et un facteur mesuré, comme sur des tableaux d'occurrences partitionnés elle doit pouvoir être employée avec profit.

### 4.3 — Question de centrage

Un détail sépare encore l'AFC et l'analyse non symétrique. Il porte sur les centrages. Pour illustrer cet élément, prenons l'exemple proposé par les auteurs <sup>1</sup>.

Les données sont dans la carte Medica de la pile de données :

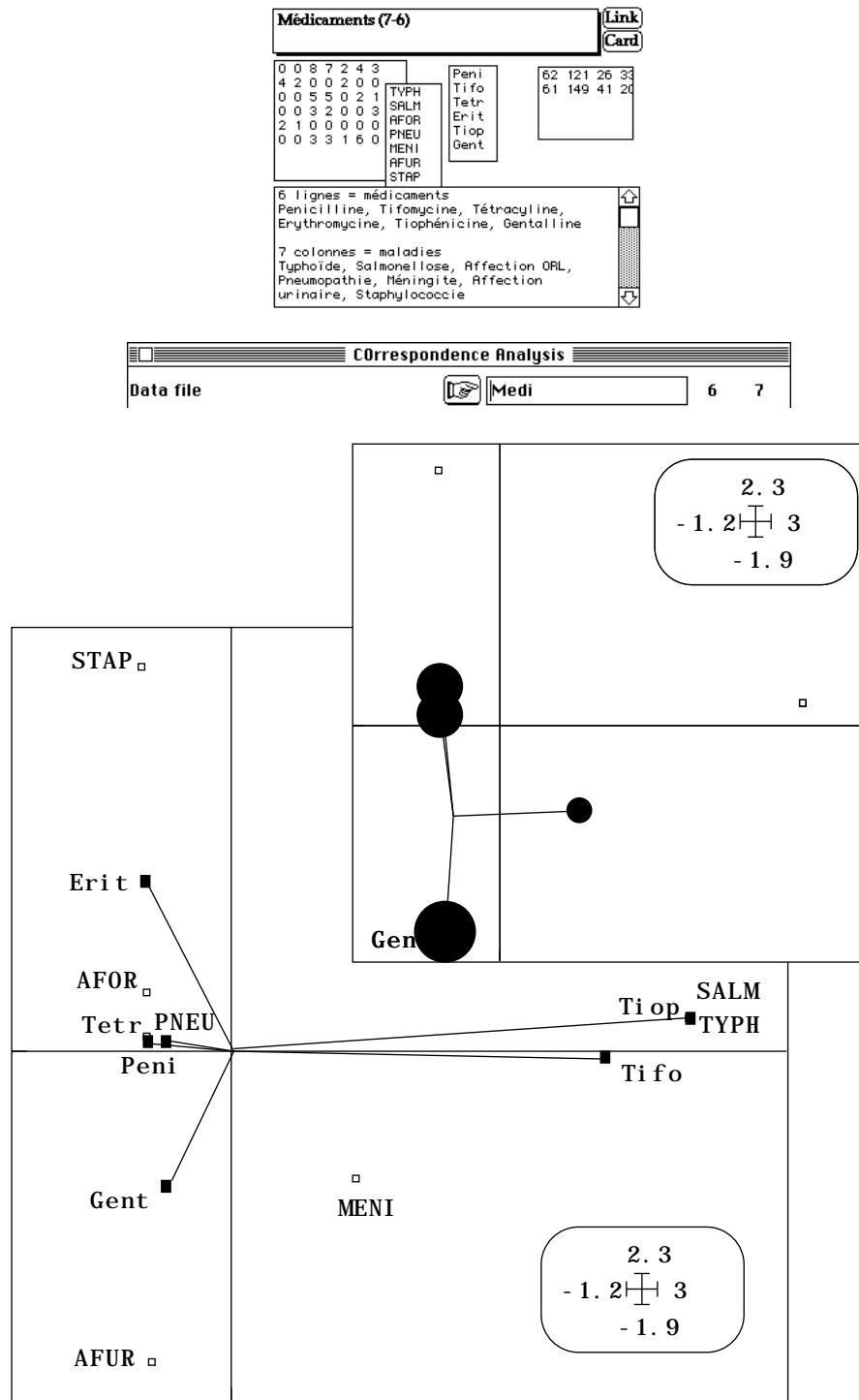
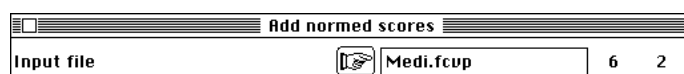
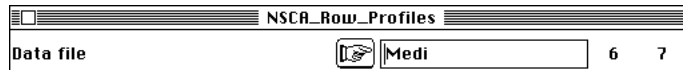


Figure 3 : AFC comme analyse discriminante. Les lignes (médicaments) sont à la moyenne des colonnes (maladies) pour leur distributions. La somme des carrés des distances à l'origine) est maximisée (fichiers ---.fcc1 et .fcli). En haut pour illustrer le calcul de la moyenne, utilisation du module ScatterDistri.



Les colonnes (maladies) précèdent les lignes (prescriptions), le diagnostic est antécédent au médicament (op.cit. p. 442) et on compare AFC et ANSC en pensant la variable maladie comme explicative et la variable prescription comme prédite. La

variable observée est la prescription et la variable de contrôle est la maladie. On fait de la discrimination inter groupes de malades sur la variable prescription ou encore on étudie les profils de prescription en fonction des groupes de malades (profils colonnes) :

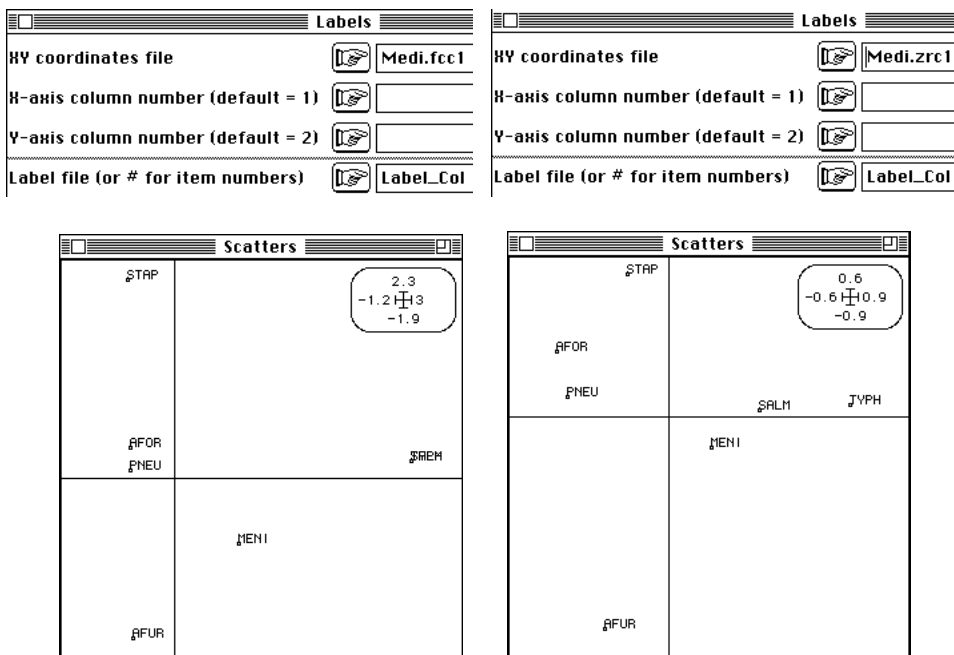


La variable contrôlée définit les modalités et la variable observée définit les profils. En écologie, le raisonnement est très explicite pour Hill <sup>22</sup>:

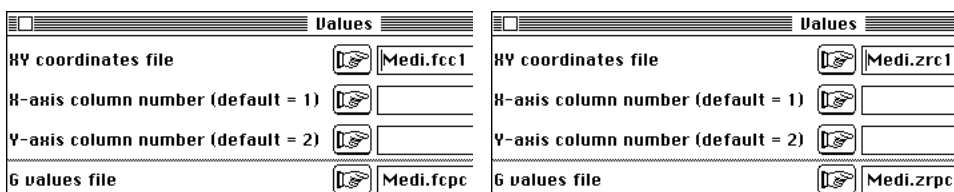
*“A remarkable feature of correspondance analysis is that it turns out to be the solution of a wide range of seemingly different problems. In particular, the axes of correspondances analysis can be shown to minimize the ratio of the within-species variance of the sample scores to the between-species variance (Hill, 1977)”* <sup>23</sup>.

La définition minimiser la variance intra sur la variance inter, ou encore maximiser la variance inter sur la variance intra correspond à celle de l’analyse discriminante classique. Le lecteur qui penserait que *“this derivation may not immediately mean anything to a biologist”* peut se reporter aux explications très claires de cet article (op. cit. p. 137). La discrimination se fait sur les occurrences observées (espèces) par les classe contrôlées (relevés).

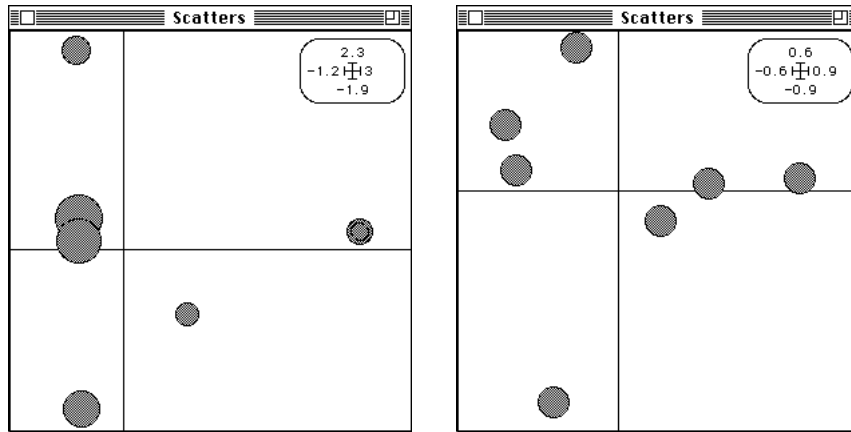
Dans l’exemple en cours, on positionne donc les maladies avec des scores pour maximiser la variance des moyennes par diagnostic :



La figure de gauche (AFC) est centrée pour la distribution marginale (poids des colonnes) alors que celle de droite l’est pour la distribution uniforme :







Les deux analyses maximisent la moyenne des carrés des écarts entre la moyenne du profil ligne et la moyenne du profil global. Quand le profil global tombe à l'origine, il n'y a pas de problème (Figure 3). Quand ce n'est pas le cas, la coordonnée est un écart entre deux moyennes (Figure 4).

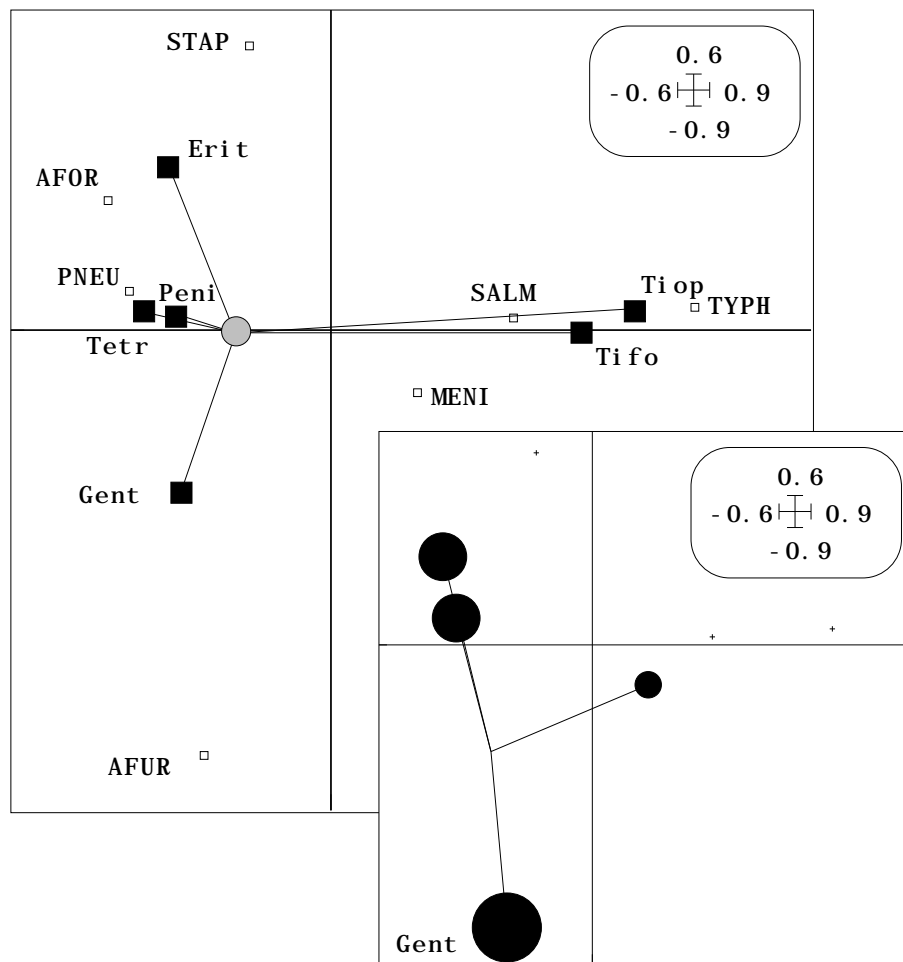


Figure 4 : Critère maximisé en ANSC. Petits carrés blancs : position des catégories de références (---.zrc1). Carrés noirs : position des classes par averaging (---.zrls). Cercle gris : position du centre de gravité des classes (---.zrlg). La somme des carrés des distances est optimisée.

Comme en AFC, la variance des positions des profils est optimisée par l'ANSC mais la moyenne n'est pas à l'origine. Il convient d'en tenir compte (Figure 5) et les options donnent les fichiers utiles pour améliorer la lecture des cartes factorielles.

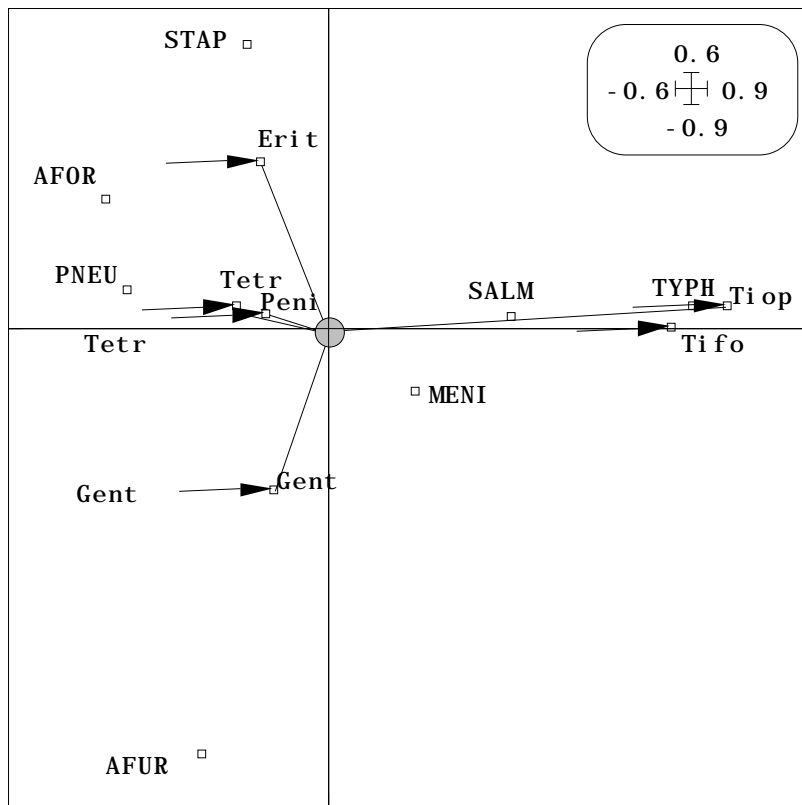
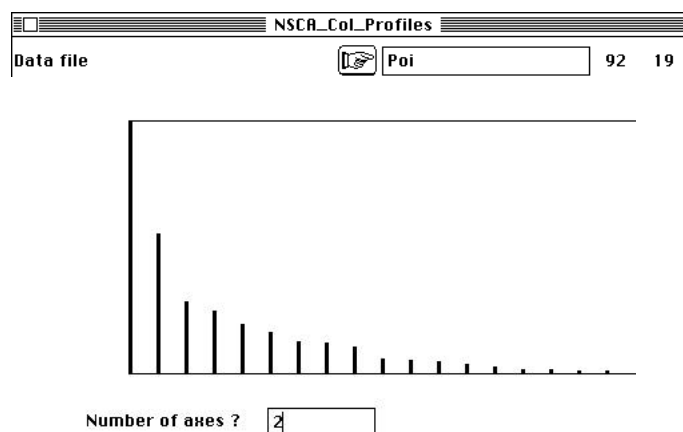


Figure 5 : Cartes factorielles (---.zrc1 et ---.zrli) en ANSC. Le décalage introduit par la différence des contraintes de centrages est souligné par les flèches. Si le décentrage est important, il faut en tenir compte.

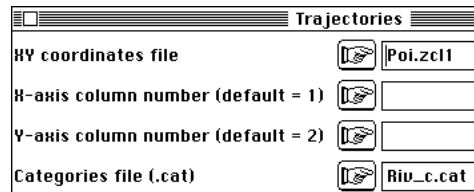
#### 4.4 — Vers la mesure de l'asymétrie d'un tableau faunistique

Un dernier exemple valide complètement les conclusions des auteurs de l'ANSC. L'analyse des correspondances n'indique pas le coût des contraintes imposées pour obtenir le double averaging. Cette question n'apparaît qu'à l'usage d'une méthode très proche sur le fond sans imposer ces contraintes. Une inter-classe peut-elle être profondément différente d'une discriminante, une ACPVI peut-elle être fort éloignée d'une analyse canonique ? A ceux qui pensent que toute méthode linéaire se vaut, les données de J. Verneaux<sup>24</sup> apporte un démenti formel.

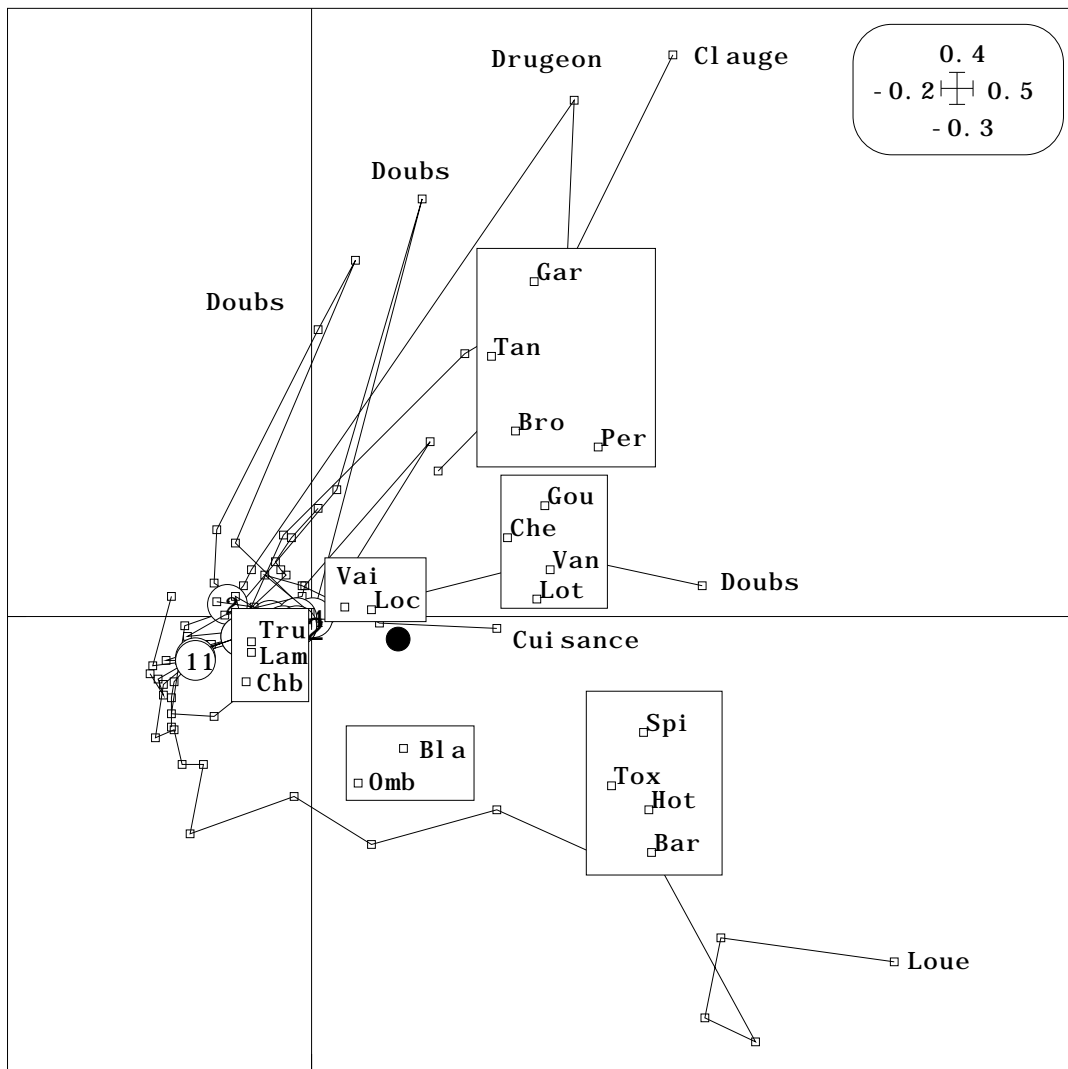
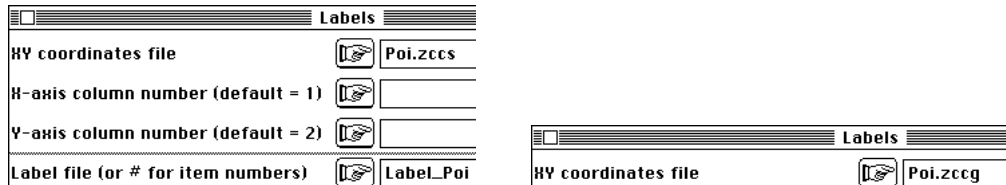
Le tableau faunistique 92 stations sur 12 rivières et 19 espèces de Poissons est soumis aux deux ANSC possibles. La première est la plus naturelle :



Les stations sont réparties sur un plan :

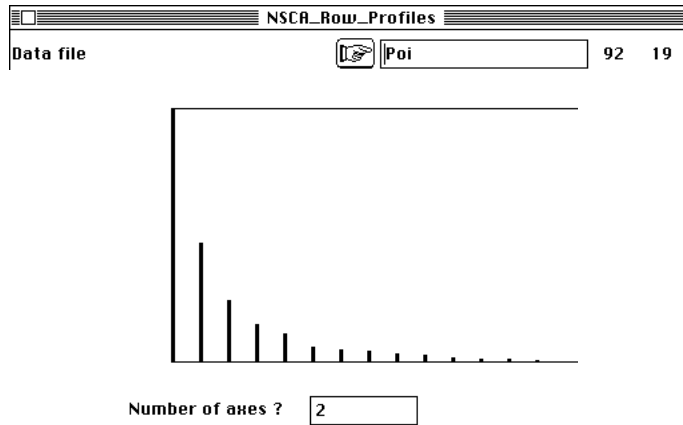


Les espèces sont positionnées à leur moyenne et on représente la position moyenne :



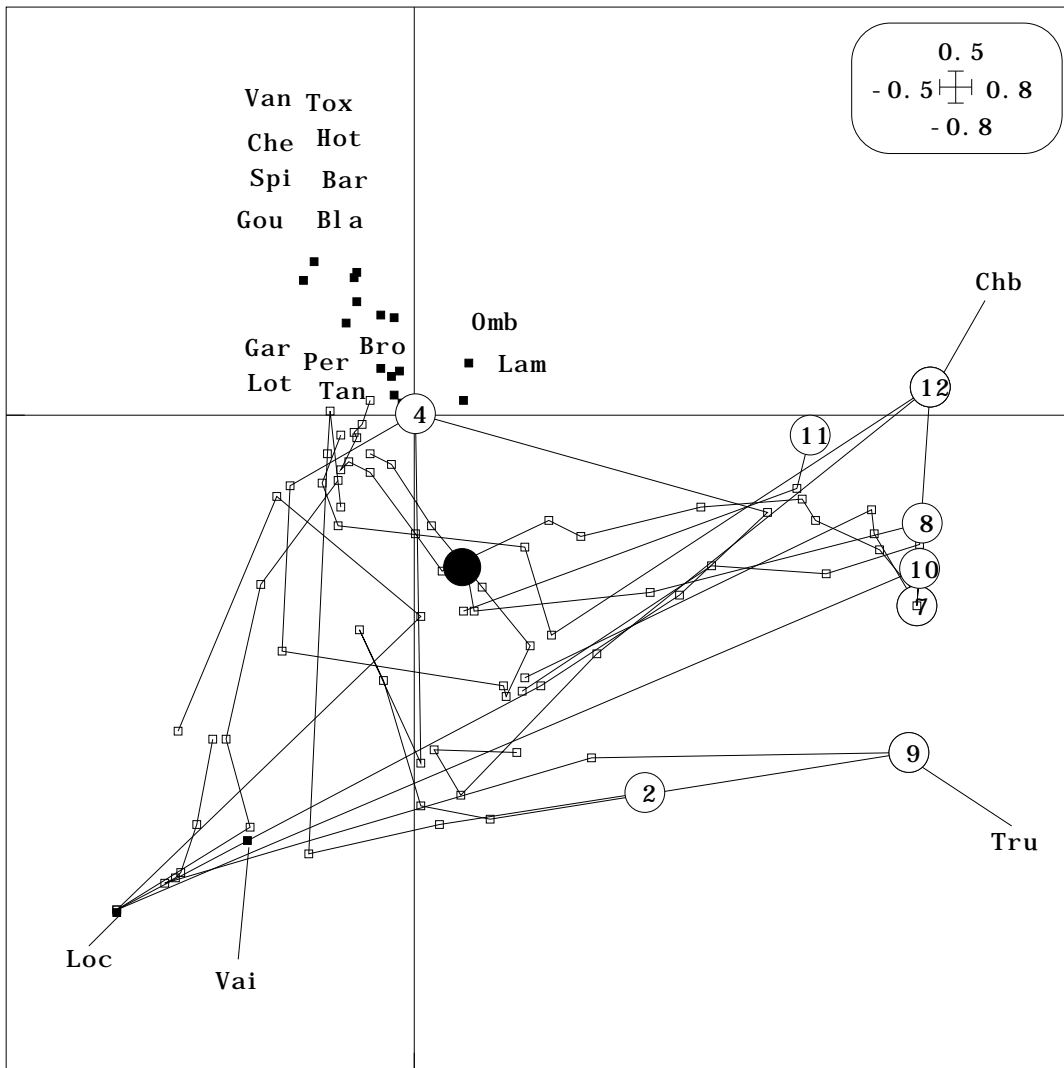
Le tout est très clair et on retrouve à très peu de chose près l'interprétation de l'analyse basée sur la théorie de la niche (fiche thématique 4-7 p. 22). Des rivières passent de la zone à Truite à la zone à Brème tandis que la succession Truite-Ombre-Barbeau est en évidence dans la Loue, rivière de référence.

La seconde est exécutée pour mesurer l'asymétrie du tableau :



Le même principe de dépouillement est utilisé. Les espèces sont distribuées sur un plan, les relevés sont placés à la position moyenne et le centre de gravité est représenté :

Labels		Trajectories	
XY coordinates file	<input type="text" value="Poi.zrc1"/>		<input type="text" value="Poi.zrls"/>
X-axis column number (default = 1)	<input type="text"/>	t = 1)	<input type="text"/>
Y-axis column number (default = 2)	<input type="text"/>	t = 2)	<input type="text"/>
Label file (or # for item numbers)	<input type="text" value="Label_Poi"/>	<input type="text" value="Poi.zrlg"/>	<input type="text" value="Riv_c.cat"/>



On retrouve l'interprétation de la méthode STATIS (fiche thématique 6.2 p. 20) et la distribution triangulaire (Truite-Chabot) (Vairon-Loche) et les autres (Aval).

La conclusion s'impose : le choix de la direction relevés --> espèces ou espèces --> relevés pour l'opération d'averaging est un choix décisif. Ne pas choisir en pratiquant l'AFC est un leurre. En conclusion, les auteurs de l'ANSC (op. cit. p. 444) notent :

*Les considérations méthodologiques qui précèdent, ... nous font penser que l'introduction d'une approche non symétrique dans l'analyse des relations entre variables qualitatives serait pleinement justifiée.*

Du point de vue de l'ordination écologique, c'est à dire de la relations entre espace et espèces, c'est fort vraisemblable.

## Références

<sup>1</sup> Lauro, N. & D'Ambra, L. (1984) L'analyse non symétrique des correspondances. In : *Data Analysis and Informatics III*. Diday, E. & Coll. (Ed.) Elsevier, North-Holland. 433-446.

<sup>2</sup> Light, R.J. & Margolin, B.H. (1971) An analysis of variance for categorical data. *Journal of the American Statistical Association* : 66, 534-544.

<sup>3</sup> Thioulouse, J. & Chessel, D. (1992) A method for reciprocal scaling of species tolerance and sample diversity. *Ecology* : 73, 670-680.

<sup>4</sup> Lande, R. (1996) Statistics and partitioning of species diversity, and similarity among multiple communities. *Oikos* : 76, 5-13.

<sup>5</sup> Fisher, R.A. (1940) The precision of discriminant functions. *Annals of Eugenics* : 10, 422-438.

<sup>6</sup> Gower, J.C. (1990) Fisher's optimal scores and multiple correspondence analysis. *Biometrics* : 46, 947-961.

<sup>7</sup> Hill, M.O. (1977) Use of simple discriminant functions to classify quantitative phytosociological data. In : *Proceedings of the First International Symposium on Data Analysis and Informatics*. Diday, E. (Ed.) INRIA Rocquencourt, France. 181-199.

<sup>8</sup> Lebart, L., Morineau, A. & Piron, M. (1995) *Statistique exploratoire multidimensionnelle*. Dunod, Paris. 1-439.

<sup>9</sup> Oksanen, J. (1987) Problems of joint display of species and site scores in correspondence analysis. *Vegetatio* : 72, 51-57.

<sup>10</sup> Williams, E.J. (1952) Use of scores for the analysis of association in contingency tables. *Biometrika* : 39, 274-289.

<sup>11</sup> Hill, M.O. (1973) Reciprocal averaging : an eigenvector method of ordination. *Journal of Ecology* : 61, 237-249.

Hill, M.O. (1974) Correspondence analysis : A neglected multivariate method. *Journal of the Royal Statistical Society, C* : 23, 340-354.

<sup>12</sup> Cazes, P., Chessel, D. & Dolédec, S. (1988) L'analyse des correspondances internes d'un tableau partitionné : son usage en hydrobiologie. *Revue de Statistique Appliquée* : 36, 39-54.

<sup>13</sup> Lebart, L. (1969) Analyse statistique de la contiguïté. *Publication de l'Institut de Statistiques de l'Université de Paris* : 28, 81-112.

<sup>14</sup> Badia, J. & Do Chi, T. (1976) Etude cinétique de la structure des populations de *Squilla mantis* (Crustacea : Stomatopoda) par l'analyse factorielle des correspondances. *Journal of Experimental Marine Biology and Ecology* : 36, 159-168.

<sup>15</sup> Persat, H. & Chessel, D. (1989) Typologie de distributions en classes de taille : intérêt dans l'étude des populations de poissons et d'invertébrés. *Acta Œcologica, Œcologia Generalis* : 10, 2, 175-195.

<sup>16</sup> El Kallab-Wakim, A. (1978) *Etude du cycle biologique d'Heptagenia sulphurea* (Müller, 1776) (Ephemeroptera : Heptageniidae) dans le Rhône à Lyon. Thèse 3ème Cycle, Université Lyon 1. 1-65.

<sup>17</sup> Hirschfeld, H.O. (1935) A connection between correlation and contingency. *Proceedings of the Cambridge Philosophical Society, Mathematical and Physical Sciences* : 31, 520-524.

<sup>18</sup> Nishisato, S. (1980) *Analysis of categorical data : dual scaling and its applications*. University of Toronto Press, London. 1-276.

<sup>19</sup> Prodon, R. & Lebreton, J.D. (1981) Breeding avifauna of a Mediterranean succession : the holm oak and cork oak series in the eastern Pyrénées. 1 : Analysis and modelling of the structure gradient. *Oikos* : 37, 21-38.

<sup>20</sup> Belair, G. de. (1981) *Biogéographie et aménagement : la plaine de La Mafragh* (Annaba, Algérie). Thèse de 3<sup>e</sup> cycle. Université Paul Valéry, Montpellier. 1-150.

Belair, G. de & Bencheikh-Lehocine, M. (1987) Composition et déterminisme de la végétation d'une plaine côtière marécageuse : La Mafragh (Annaba, Algérie). *Bulletin d'Ecologie* : 18, 4, 393-407.

<sup>21</sup> Friday, L.E. (1987) The diversity of macroinvertebrate and macrophyte communities in ponds. *Freshwater Biology* : 18, 87-104.

<sup>22</sup> Hill, M.O. (1988) How effective is ordination as a means of relating vegetation to ecological factors?. *Bulletin de la Société Royale de Botanique de Belgique* : 121, 134-141.

<sup>23</sup> Hill, M.O. (1977) Use of simple discriminant functions to classify quantitative phytosociological data. In : *Proceedings of the First International Symposium on Data Analysis and Informatics*. Diday, E. (Ed.) INRIA Rocquencourt, France. 181-199.

<sup>24</sup> Verneaux, J. (1973) *Cours d'eau de Franche-Comté (Massif du Jura). Recherches écologiques sur le réseau hydrographique du Doubs. Essai de biotypologie*. Thèse d'état, Besançon. 1-257.