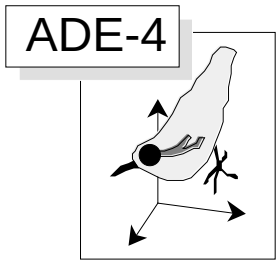


Analyses canoniques et listes d'occurrences d'espèces

Résumé

La fiche précise la perturbation théorique introduite par le traitement des listes d'occurrences dans l'ensemble des méthodes linéaires d'analyse des données écologiques. Les méthodes d'analyses canoniques comme l'AFC, l'analyse canonique des correspondances et l'analyse canonique classique jouent un rôle central dans ce domaine. Les liens entre ces techniques et les méthodes dites à variables instrumentales sont abordés. Une révision globale de la place de l'analyse canonique en écologie semble nécessaire.

Plan

D. Chessel & C. Gimaret

1 — Introduction

Nous avons décrit dans la fiche 2.8 l'origine des tableaux d'occurrences. Il n'y a pas de relevés mais récoltes dans les herbiers et les collections des Musées d'Histoire Naturelle de faits établis, les rencontres en un lieu et un temps en général connus de l'occurrence d'un individu d'une espèce donnée.

Ces listes d'occurrence constituent un type d'informations radicalement original qui mérite qu'on y réfléchisse de manière nouvelle. Nous allons voir que l'analyse de cette information, acceptée comme telle, n'est pas sans surprise.

LatLonAltSai.xls

	A	B	C	D	E	F	G
1	numesp	lat	lon	alt	saison	numqua	nomesp
2	1	9.83333333	76.9166667	850	2.5	16	actiboud
3	1	9.71666667	77.15	1125	3.5	25	actiboud
3948	269	8.5	77.3333333	1300	2.5	125	verntrav
3949	269	8.5	77.3833333	1300	2.5	125	verntrav
3950	269	8.5	77.3833333	1400	2.5	125	verntrav
3951	269	8.51666667	77.3333333	1175	2.5	125	verntrav
3952	269	8.48333333	77.385	1300	2.5	128	verntrav
3953	269	8.48333333	77.4666667	1300	2.5	129	verntrav
3954	269	8.38333333	77.4666667	1100	3.5	132	verntrav

Dans le tableau ci-dessus, 3953 occurrences ont été consignées (extrait de Ramesh, B.R. & Pascal, J.P. (1998) *Atlas of endemics of the Western Ghats (India)*. Institut Français de Pondichery. In press). Sur une ligne, on trouve le numéro du taxon, la latitude, la longitude, l'altitude, la longueur de la saison sèche, le numéro du quadrat et le nom de code du taxon. Cette information se décompose en deux parties. La première porte sur l'individu rencontré, la seconde sur le milieu dans lequel s'est faite cette rencontre. Ces deux types doivent être soigneusement séparés, tant au plan théorique que pratique.

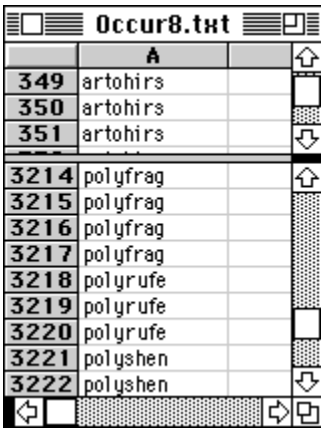
Occur.xls

	A	B	C
1	numesp	nomesp	Nomfami
2	1	actiboud	LAURACEAE
3	1	actiboud	LAURACEAE
3948	269	verntrav	ASTERACEAE
3949	269	verntrav	ASTERACEAE
3950	269	verntrav	ASTERACEAE
3951	269	verntrav	ASTERACEAE
3952	269	verntrav	ASTERACEAE
3953	269	verntrav	ASTERACEAE
3954	269	verntrav	ASTERACEAE

L'individu rencontré porte une information biologique. Il est d'une espèce, d'un genre, d'une famille. Il peut être d'un sexe, d'une taille, d'un type morphologique donné. Il peut présenter toute sorte de traits biologiques. La forme la plus simple de cette information ne contient que le nom de l'espèce. Il faut intégrer cette seule information dans un schéma formel. C'est l'objet du premier paragraphe.

2 — Tableau pour une liste d'occurrences

Un tableau d'occurrences est issu d'une seule variable qualitative dont chaque modalité est une espèce.



Occur8.txt

	A
349	artohirs
350	artohirs
351	artohirs
3214	polyfrag
3215	polyfrag
3216	polyfrag
3217	polyfrag
3218	polyrufe
3219	polyrufe
3220	polyrufe
3221	polyshen
3222	polyshen

On appellera **o** une telle liste. Elle a *o* lignes (occurrences) et 1 colonne. Elle contient *s* noms différents. Les *s* taxa sont numérotés de 1 à *s* et des fichiers à *s* lignes peuvent contenir de l'information qui leur est associée, qu'ils s'agissant de noms complets, de noms de code, de position taxonomique. Dans l'exemple utilisé, il y a 269 taxa :

Dico-esp

	A	B	C	D	E
1	numesp	code	numgenre	famille	numfam
2	1	actiboud	1	LAURACEAE	21
3	2	acticaca	1	LAURACEAE	21
4	3	acticaob	1	LAURACEAE	21
268	267	vateindi	103	DIPTEROCARPACEAE	12
269	268	veprbilo	104	RUTACEAE	35
270	269	verntrav	105	ASTERACEAE	6

Conceptuellement, une telle liste donne **O**, un tableau disjonctif complet comportant *o* lignes et *s* colonnes. Sur chaque ligne toutes les valeurs sont nulles à l'exception d'une seule qui vaut 1 et est associée à la colonne de l'espèce concernée par l'occurrence. On notera $\mathbf{o}_i$ le numéro de l'espèce apparaissant dans l'occurrence *i* ($1 \le i \le o$). Donc ($1 \le j \le s$) :

$$\mathbf{O}_{ij} = 1 \Leftrightarrow \mathbf{o}_i = j$$

$$\mathbf{O}_{ij} = 0 \Leftrightarrow \mathbf{o}_i \neq j$$

Concrètement cette liste est lue par l'option OccurData : Read_Occur_File.

Comme cette information doit être couplée avec l'information mésologique donnant les conditions de la rencontre, il faut donner au tableau **O** un statut de tableau d'analyse des données. Cette opération *pratiquement* inutile (deux lignes sont soit identiques soit différentes !) est *conceptuellement* fondamentale. Elle doit donner à la variabilité totale exprimée dans **O** le statut de mesure de la diversité totale de la liste et permettre de décomposer cette diversité totale en diversité expliquée par le milieu (diversité β) et diversité intrinsèque (diversité α).

Chaque colonne du tableau **O** est une indicatrice (variable prenant la valeur 0 ou 1). Chaque occurrence a le même poids $1/o$, ce qui donne une matrice diagonale des poids du type $\Delta_o = \mathbf{D}_{1/o} = (1/o)\mathbf{I}_o$. Δ_o est la pondération *naturelle* des occurrences. Elle pourrait ne pas être uniforme en toute généralité. Pour le moment, on la conservera uniforme pour simplifier.

Les moyennes des colonnes du tableau **O** sont alors simplement les proportions de chaque espèce. Pour $1 \le j \le s$, on notera $\mathbf{O}_{\bullet j}$ le nombre d'occurrences portant sur l'espèce *j* et f_j la fréquence associée, soit :

$$f_j = \frac{\mathbf{O}_{\bullet j}}{o}$$

Évidemment, $\sum_{j=1}^{s} f_j = 1$. Dans tout ce qui suit on utilisera f_j comme le poids *naturel* du taxon *j*. Ceci donne une matrice diagonale des poids des taxa :

$$\Delta_s = Diag\left(f_1, \quad , f_j, \quad , f_s\right)$$

On aura encore besoin du vecteur de ces fréquences, qui sera noté :

$$\mathbf{f}_s = \begin{bmatrix} f_1 \\ \vdots \\ f_j \\ \vdots \\ f_s \end{bmatrix} \quad \mathbf{f}_s^t = \begin{bmatrix} f_1 & & f_j & & f_s \end{bmatrix}$$

Le symbole ^{t} sera utilisé pour la transposition des matrices. $\mathbf{O}^j$ est la colonne *j* du tableau **O**. Sa moyenne est f_j et sa variance vaut $f_j\left(1 - f_j\right)$.

2.1 — Inertie, diversité et métrique canonique

Ces premières observations conduisent à voir dans les indicatrices des classes de simples variables qui induisent une ACP centrée, fictive évidemment. Si $\mathbf{O}_c$ est le tableau centré associé au tableau **O**, le triplet $\left(\mathbf{O}_c, \mathbf{I}_s, \Delta_o\right)$ a comme inertie totale l'indice de Simpson :

$$Iner\left(\mathbf{O}_c, \mathbf{I}_s, \Delta_o\right) = Trace\left(\mathbf{O}_c^t \Delta_o \mathbf{O}_c \mathbf{I}_s\right) = 1 - \sum_{j=1}^{s} f_j^2$$

Cette remarque conduit à voir que mesurer le diversité d'un ensemble, c'est apprécier quantitativement la différence qui existe entre tous les éléments de cet ensemble. C'est clairement vrai pour une variable quantitative, dont on mesure la variabilité par la variance. Pour une pondération uniforme des éléments de la collection de valeurs on a :

$$Var(x) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - m(x)\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \frac{1}{n}\sum_{j=1}^{n} x_j\right)^2 = \frac{1}{n}\sum_{i=1}^{n} {x_i}^2 - \left(\frac{1}{n}\sum_{j=1}^{n} x_j\right)^2$$

Or,

$$\frac{1}{n^2}\sum_{i,j=1}^{n}\left(x_i - x_j\right)^2 = \frac{1}{n^2}\sum_{i,j=1}^{n}\left({x_i}^2 + {x_j}^2 - 2x_i x_j\right) = \frac{1}{n}\sum_{i=1}^{n}{x_i}^2 + \frac{1}{n}\sum_{j=1}^{n}{x_i}^2 - \frac{2}{n^2}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{j=1}^{n} x_j\right)$$

Donc :

$$2Var(x) = \frac{1}{n^2}\sum_{i,j=1}^{n}\left(x_i - x_j\right)^2$$

Curieusement, si on ne compte que les couples de deux points distincts, on donne le même sens à la variance estimée (dite sans biais) :

$$2\hat{Var}(x) = \frac{1}{n(n-1)}\sum_{\substack{i,j=1 \\ i\neq j}}^{n}\left(x_i - x_j\right)^2$$

Une variabilité est toujours la moyenne des carrés des distances entre objets de la collection. La diversité d'un ensemble dont on sait mesurer la différence entre deux éléments par un indice de distance et dans lequel chaque élément a un poids sera donc définie par :

$$Div(A) = \frac{1}{2}\sum_{x,y}^{n}{}_{A} p_x p_y d^2(x,y)$$

Si A est une partie finie d'un espace euclidien muni d'un produit scalaire Ω, on a alors une mesure euclidienne de la diversité par :

$$Div_{\Omega}(A) = \frac{1}{2}\sum_{x,y}^{n}{}_{A} p_x p_y \|x - y\|_{\Omega}^2$$

Ceci n'est rien d'autre que l'inertie du nuage des points de A au sens de la norme Ω. Nous avons discuté de la concordance des points de vue dans la comparaison entre l'indice CATANOVA de Light & Margolin [1] et le procédé qu'avait utilisé Lebart [2] pour introduire les graphes de voisinages en analyse des données et la concordance entre l'indice de Simpson et l'inertie d'un tableau d'indicatrices [3] (Voir fiche 2.8).

La convergence entre variance, inertie, moyenne des distances entre éléments des couples, indice de diversité est un élément très important. L'inertie est, sans conteste, une mesure algébrique de la diversité.

On notera pour résumer qu'un tableau d'occurrences conduit à un triplet du type $\left(\mathbf{O}_C, \mathbf{I}_S, \Delta_O\right)$ dont l'inertie est l'indice de Simpson associé à l'ensemble des occurrences. Retrouvons directement ce résultat du point de vue de la différence entre deux occurrences. Si deux occurrences portent sur deux espèces différentes *j* et *k*, les deux lignes du tableau $\mathbf{O}_C$ sont :

$$a = \begin{bmatrix} -f_1 & 1-f_j & -f_k & -f_s \end{bmatrix}$$
$$b = \begin{bmatrix} -f_1 & -f_j & 1-f_k & -f_s \end{bmatrix}$$
$$a-b = \begin{bmatrix} 0 & 1 & -1 & 0 \end{bmatrix}$$

La distance entre deux occurrences d'espèces différentes est donc toujours de 2. Il y a, au total c^2 couples et :

$$\sum_{j=1}^{s}\sum_{\substack{k=1\\k\neq j}}^{s} n_j n_k = \sum_{j=1}^{s} n_j \sum_{\substack{k=1\\k\neq j}}^{s} n_k = \sum_{j=1}^{s} n_j\left(c-n_j\right)$$

couples portant sur deux espèces distinctes. D'où :

$$Div_{\mathbf{I}_s}(\mathbf{O}) = \frac{1}{2}\sum_{j=1}^{s}\frac{1}{c}\frac{1}{c}2n_j\left(c-n_j\right) = 1-\sum_{j=1}^{s} f_j^2 = Iner\left(\mathbf{O}_c,\mathbf{I}_s,\Delta_o\right)$$

Ceci souligne que choisir le triplet $\left(\mathbf{O}_c,\mathbf{I}_s,\Delta_o\right)$ revient à dire que deux occurrences ont une distance nulle si elles portent sur la même espèce et 2 si elles portent sur deux espèces distinctes, ou encore que la diversité totale est celle de l'indice de Simpson, ou encore que chaque espèce participe à la diversité d'autant plus fortement qu'elle est plus abondante (tant que f_j n'atteint pas 50% des occurrences, ce qui est normalement le cas). Ce n'est pas le seul point de vue possible.

2.2 — Décorrélation des indicatrices

Dans l'analyse fictive du triplet $\left(\mathbf{O}_c,\mathbf{I}_s,\Delta_o\right)$, la matrice de covariance à diagonaliser est :

$$\mathbf{V_O} = \mathbf{O}_c^t\Delta_o\mathbf{O}_c\mathbf{I}_s = \mathbf{O}_c^t\Delta_o\mathbf{O}_c = \frac{1}{c}\mathbf{O}_c^t\mathbf{O}_c = \Delta_s - \mathbf{f}_s\mathbf{f}_s^t$$

On retrouve encore l'indice de Simpson immédiatement comme trace de cette matrice à *s* lignes et *s* colonnes. Cette matrice est de rang *s*-1. Les indicatrices des classes sont donc artificiellement covariantes. La variance de chaque variable est $f_j\left(1-f_j\right)$ et la covariance de deux variables vaut $-f_j f_k$. C'est évident pour deux espèces abondantes puisque chaque occurrence ne porte que sur une seule espèce.

On peut chercher à décorréler les variables, ce qui se fait par le biais de l'inversion de norme. $\mathbf{V_O}$ a un inverse généralisé $\mathbf{V_O^-}$ (voir fiche 2.8) qui s'écrit ($\mathbf{1}_{ss}$ est la matrice à *s* lignes et *s* colonnes dont toutes les valeurs égalent l'unité) :

$$\mathbf{V_O^-} = \Delta_s^{-1} - \mathbf{1}_{ss}$$

Il suffit de vérifier que :

$$\mathbf{V_O}\mathbf{V_O^-}\mathbf{V_O} = \mathbf{V_O} \quad \mathbf{V_O^-}\mathbf{V_O}\mathbf{V_O^-} = \mathbf{V_O^-}$$

On notera au passage la forme particulière du produit (opérateur de centrage des codes espèces) :

$$\mathbf{V_O^-V_O} = \mathbf{I}_s - \begin{pmatrix} f_1 & \cdots & f_s \\ \vdots & \vdots & \vdots \\ f_1 & \cdots & f_s \end{pmatrix}$$

$\mathbf{V_O^-}$ est une semi-norme dans l'espace orthogonal au vecteur $\mathbf{1}_s$ dans lequel se trouve le nuage des occurrences (lignes de $\mathbf{O}_c$). On peut donc considérer le triplet $\left(\mathbf{O}_c, \mathbf{V_O^-}, \Delta_o\right)$ utilisant la norme dite du Khi2. Quelle son inertie totale ?

$$Iner\left(\mathbf{O}_c, \mathbf{V_O^-}, \Delta_o\right) = Trace\left(\mathbf{O}_c^t \Delta_o \mathbf{O}_c \mathbf{V_O^-}\right) = Trace\left(\mathbf{V_O V_O^-}\right) = s - 1$$

Le nombre d'espèce (moins 1, car une liste à une seule espèce a une diversité nulle !) est donc une inertie totale, un indice de diversité ! Décorréler les indicatrices et mesurer les distance entre deux occurrences par la métrique du Khi2, c'est choisir le nombre d'espèce pour décrire la diversité. Une occurrence d'espèce rare y joue alors un rôle bien plus grand.

En effet,

$$\|a - b\|^2_{\mathbf{V_O^-}} = \frac{1}{f_j} + \frac{1}{f_k}$$

donne :

$$Div_{\mathbf{V_O^-}}(\mathbf{O}) = \frac{1}{2} \sum_{\substack{j,k=1 \\ k \neq j}}^{s} \frac{1}{c}\frac{1}{c} n_j n_k \left(\frac{1}{f_j} + \frac{1}{f_k}\right) = \frac{1}{2} \sum_{\substack{j,k=1 \\ k \neq j}}^{s} \left(f_j + f_k\right) = \frac{1}{2}(2s - 2)$$

$$Iner\left(\mathbf{O}_c, \mathbf{V_O^-}, \Delta_o\right) = s - 1$$

La distance entre deux occurrences de deux espèces rares est donc considérable. On a là le point de vue proprement naturaliste qui donne aux raretés une importance très forte.

De toute manière, l'utilisation dans la suite des deux schémas $\left(\mathbf{O}_c, \mathbf{I}_s, \Delta_o\right)$ et $\left(\mathbf{O}_c, \mathbf{V_O^-}, \Delta_o\right)$ s'impose.

3 — Typologie de classes d'occurrence

En face de la liste des occurrences, apparaît de l'information soit biologique soit mésologique. Le type de cette information va déterminer diverses pratiques. Le plus simple de ces types est une partition des occurrences en classes. Soit **y** un vecteur qui associe à chaque occurrence un numéro de groupe compris entre 1 et *g*. Soit **Y** le tableau disjonctif complet associé, $\mathbf{Y}_c$ le tableau centré associé, Δ_g la matrice diagonale contenant le poids des classes (h_j est le rapport du nombre d'occurrences dans le groupe j sur le nombre d'occurrences total o) :

$$\Delta_g = Diag\left(h_1, \quad , h_j, \quad , h_g\right)$$

$\mathbf{V_Y}$ est la matrice de covariances des indicatrices centrées de s g groupes et $\mathbf{V_Y^-}$ est son inverse généralisée. On pense immédiatement à la table de contingence qui croise les deux partitions des occurrences (celle qui est maintenant introduite et celle des noms d'espèces). On la note $\mathbf{T} = \mathbf{Y}^t\Delta_o\mathbf{O}$. Elle a g lignes (groupes) et s colonnes (taxa). Dans chaque cellule y figure la proportion (total = 1) des occurrences de chaque espèce dans chaque groupe.

On notera, cependant, qu'en passant à la table de contingence des occurrences (nombres d'apparition de chaque espèce dans un quadrat si on a divisé la zone des occurrences en quadrat, par exemple) on se représente un tableau classes-espèces. En croisant avec une variable quantitative de la même manière on aura (Cf. infra) un tableau de moyenne de positions de chaque espèce. On perd donc implicitement l'existence physique des occurrences et par là leur représentation.

Il convient donc de préserver la notion d'occurrences comme lignes d'un tableau ce qui se fait grâce au projecteur associé à la partition instrumentale. Dans R^o, les espèces sont les indicatrices des classes de **O**, tandis que les classes sont les indicatrices des classes de **Y**. Pour moyenner par classe les occurrences, il n'est pas nécessaire de centrer **Y**. On observe simplement que la matrice du projecteur sur les indicatrices des classes s'écrit :

$$\boldsymbol{P}_\mathbf{Y} = \mathbf{Y}\left(\mathbf{Y}^t\Delta_o\mathbf{Y}\right)^{-1}\mathbf{Y}^t\Delta_o = \mathbf{Y}\Delta_g^{-1}\mathbf{Y}^t\Delta_o$$

On aura alors deux analyses définies par les triplets $\left(\boldsymbol{P}_\mathbf{Y}\,\mathbf{O}_c, \mathbf{I}_s, \Delta_o\right)$ et $\left(\boldsymbol{P}_\mathbf{Y}\,\mathbf{O}_c, \mathbf{V_O^-}, \Delta_o\right)$. Le premier correspond à une inter-classes et le second à une analyse discriminante.

3.1 — Analyse inter-classe d'un ensemble d'occurrences

Cette approche décompose l'indice de Simpson total en une partie inter-classes et une partie intra-classes. On peut chercher à savoir comment faire le calcul d'une part, à lui donner sa signification d'autre part. La décomposition des schémas de dualité contenant un projecteur (on peut consulter [4] ou [5]) permet de dire que l'analyse de $\left(\boldsymbol{P}_\mathbf{Y}\,\mathbf{O}_c, \mathbf{I}_s, \Delta_o\right)$ est équivalente à celle de $\left(\mathbf{Y}^t\Delta_o\mathbf{O}_c, \mathbf{I}_s, \Delta_g^{-1}\right)$ ou $\left(\Delta_g^{-1}\mathbf{Y}^t\Delta_o\mathbf{O}_c, \mathbf{I}_s, \Delta_g\right)$ qui ont les mêmes valeurs propres et les mêmes coordonnées des espèces. Dans le tableau de ce dernier triplet on a les moyennes par classes des composantes des colonnes de $\mathbf{O}_c$.

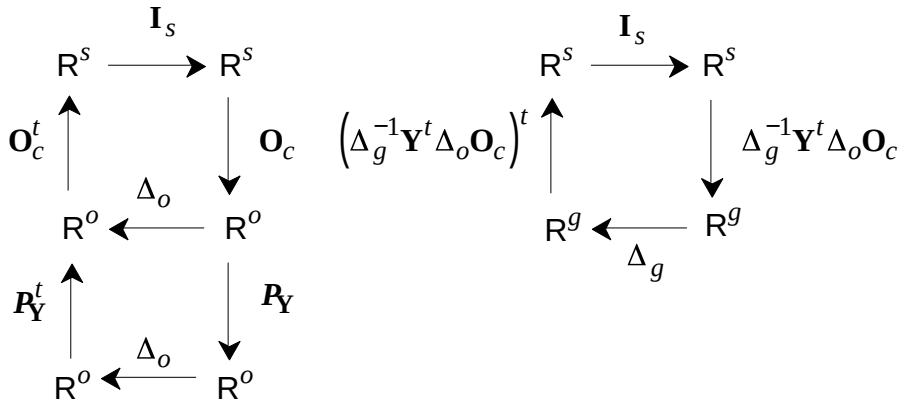
Le tableau $\Delta_g^{-1}\mathbf{Y}^t\Delta_o\mathbf{O}_c$ donne par ligne (classes) les moyennes des lignes de $\mathbf{O}_c$. On utilise **T** avec les notations habituelles. Les lignes de **T** sont indicées par k ($1 \le k \le g$), les colonnes sont indicées par j ($1 \le j \le s$). Dans les cellules de ce tableau on trouve t_{jk}, la proportion des occurrences tombées dans le groupe k et portant sur l'espèce j. Les somme marginales $t_{\bullet k}$ et $t_{j\bullet}$ sont respectivement les fréquences des occurrences par groupes et par espèces. Nécessairement, $t_{\bullet k} = h_k$ et $t_{j\bullet} = f_j$.

On note alors $t_{j/k}$ la fréquence conditionnelle des occurrences de la classe *k* qui porte sur l'espèce *j*. Alors :

$$\Delta_g^{-1}\mathbf{Y}^t\Delta_o\mathbf{O}_c = \left[t_{j/k} - t_{\bullet k}\right]_{1\le k\le g, 1\le j\le s}$$

L'analyse de $\left(\Delta_g^{-1}\mathbf{Y}^t\Delta_o\mathbf{O}_c, \mathbf{I}_s, \Delta_g\right)$ est donc l'analyse des correspondances non symétriques [6] dans la version *profils par classes*.

Les options NSCA du module COA donne l'essentiel. Cela vient de la partie commune des schémas $\left(\boldsymbol{P}_\mathbf{Y}\,\mathbf{O}_c, \mathbf{I}_s, \Delta_o\right)$ et $\left(\Delta_g^{-1}\mathbf{Y}^t\Delta_o\mathbf{O}_c, \mathbf{I}_s, \Delta_g\right)$. Les deux sont liés par le diagramme :



Les opérateurs d'Escoufier associés à R^s sont strictement identiques. A droite cela s'explicite par "code des espèces normés" (axes principaux) donnant par averaging "position des classes de variance maximale". A gauche, le même code des espèces positionne chaque occurrence sur l'espèce qui lui est associée, ce qui positionne chaque classe à la moyenne des occurrences qu'on y trouve, l'optimisation visant à minimiser la distance entre les positions des classes et celle des espèces qui s'y trouvent.

On retiendra donc que l'analyse non symétrique des correspondances (profils classes) est l'analyse inter-classe de la liste des occurrences. Ce point de vue est cependant très formel. Il est purement statistique. Les occurrences sont des observations. Le tableau des indicatrices des espèces est un tableau de données. Le tableau des indicatrices des classes est un tableau de variables instrumentales. Le tableau de données est estimé par le tableau de variables externes par projection. La conséquence abstraite de ce raisonnement conduit à la vision concrète d'un score des taxons optimisant la variance des scores des classes par averaging.

On en arrive donc à une ambiguïté assez forte. On analyse en principe la liste des occurrences des espèces et on en arrive à une information sur la liste des occurrences des quadrats. Cela correspond pourtant exactement à la définition habituelle des variables instrumentales. Si on relit l'article initial de Rao[7], on trouve page 340 :

> *Let* **X** *be the vector of* p *main variables, and* **Z** *the vector of* m *instrumental variables. In theory* **Z** *may include some or all the elements of* **X**. *We wish to replace* **Z** *by a* q *dimensional random variable* $\mathbf{Y} = \mathbf{M}^t\mathbf{Z}$ *in such a way that the predictive efficiency of* **Y** *for* **X** *is a maximum (Rao, 1962[8]).*

Il ne fait pas de doute que les instrumentales sont des explicatives. Cela vient du fait que tout schéma de dualité peut s'interpréter de deux manières cohérentes (lignes-> colonnes et colonnes -> lignes). La vision variables instrumentales par de la notion de

prédicteur, fournit un triplet, analyse ce triplet, et interprète le résultat de l'autre manière.

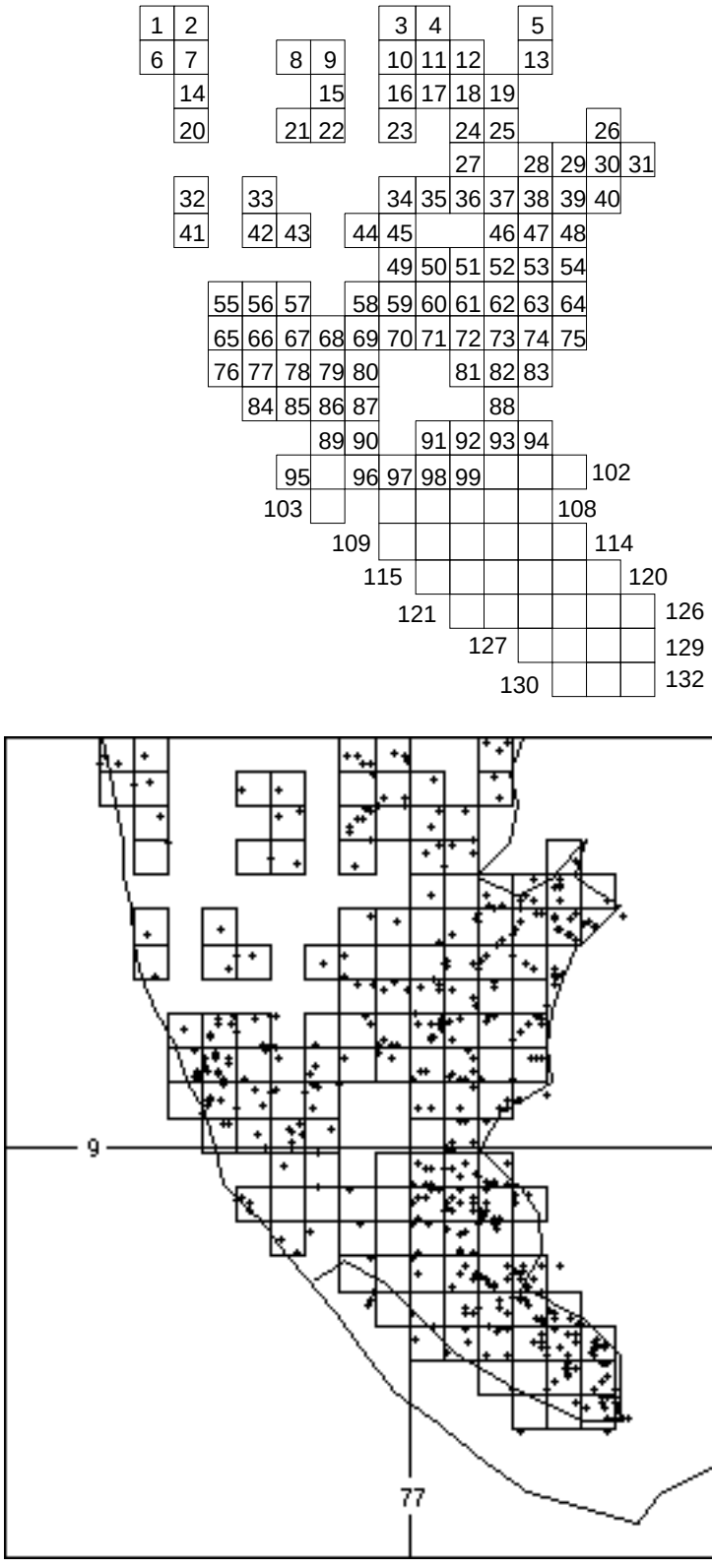


Figure 1 : Position spatiale d'une liste d'occurrences. et définition d'une grille.

On retiendra qu'expliquer les occurrences des espèces par les quadrats revient à discriminer les quadrats par leur contenu en espèces. C'est toujours vrai dès qu'on introduit une partition. Prévoir un groupe de variable par une partition, c'est chercher

une combinaison des indicatrices des classes (un code constant par classe) qui optimise la prédiction des variables. Cela s'interprète systématiquement dans l'autre sens : trouver une combinaison de variables qui sépare les classes. C'est directement lié à la notion de rapport de corrélation (variance inter/variance totale) qui est un carré de corrélation (pourcentage de variance expliquée par l'indicatrice).

Ceci explique aussi la très grande difficulté pour les écologues de relier le modèle d'averaging (ordination directe des espèces par une variable de milieu) et celui de la régression (prédire la liste des espèces par la variable de milieu).

3.2 — Typologie de profils quadrats

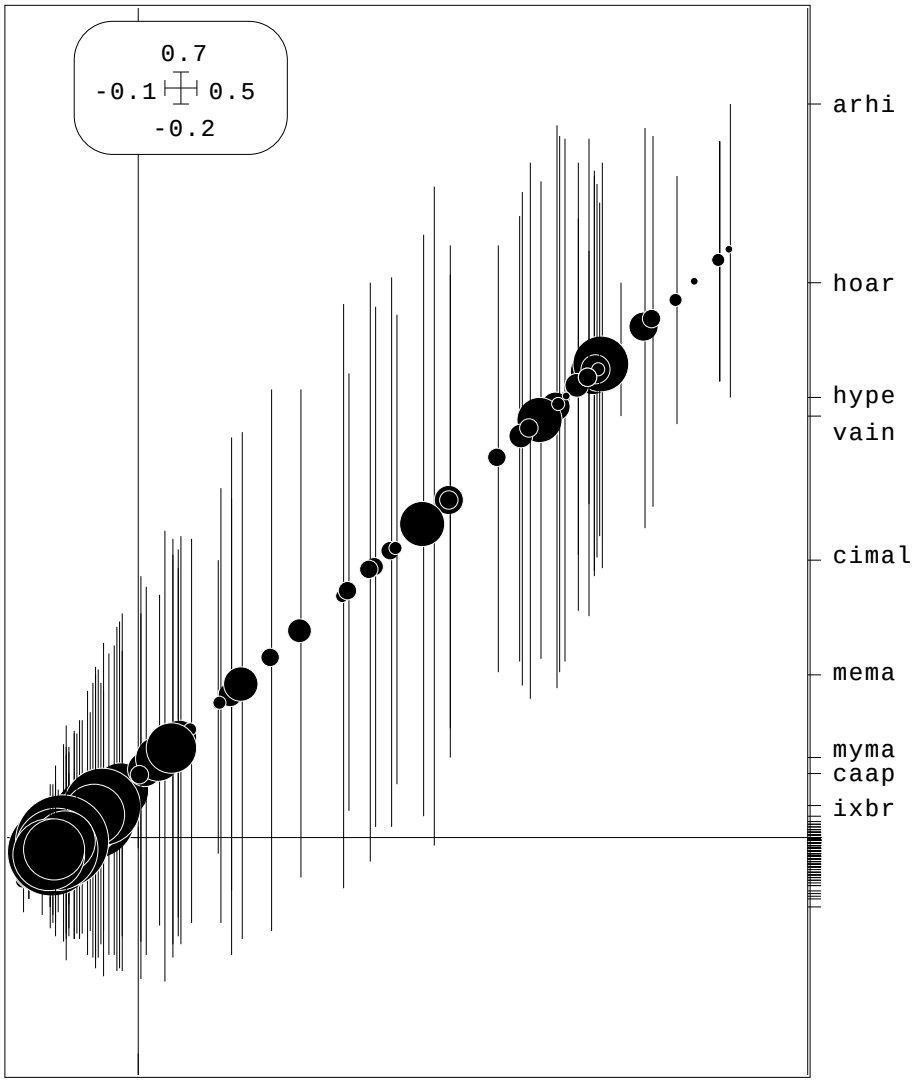
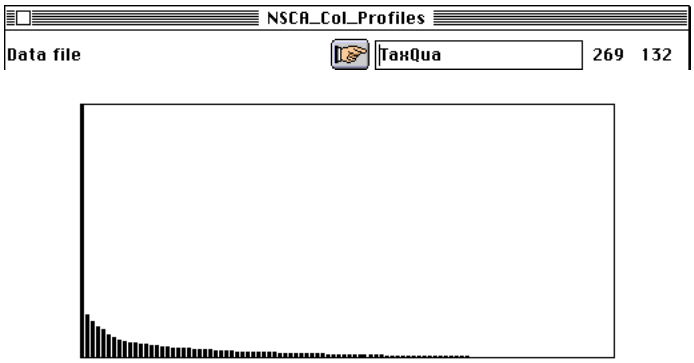
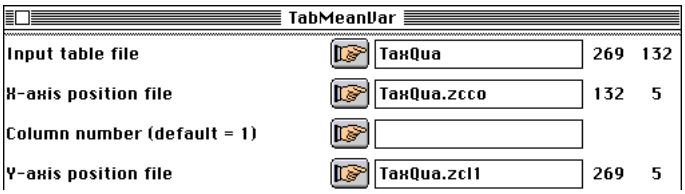


Figure 2 : Séparation des quadrats par un score des taxons et averaging.

Donnons un exemple. Les 3953 occurrences sont positionnées dans l'espace (figure 1). L'espace peut être divisée en unités de 5' de côté, ce qui définit une partition. Le lien entre la liste des espèces et la liste des quadrats s'exprime entièrement dans une table de contingence. La constitution de cette table se fait avec l'option OccurData : Occurrence classes (voir annexe 1). Si on veut faire une typologie de quadrats (classes) par la liste de leur contenu floristique, on utilisera donc COA : NSCA_Col_Profiles puisque les classes sont en colonnes :

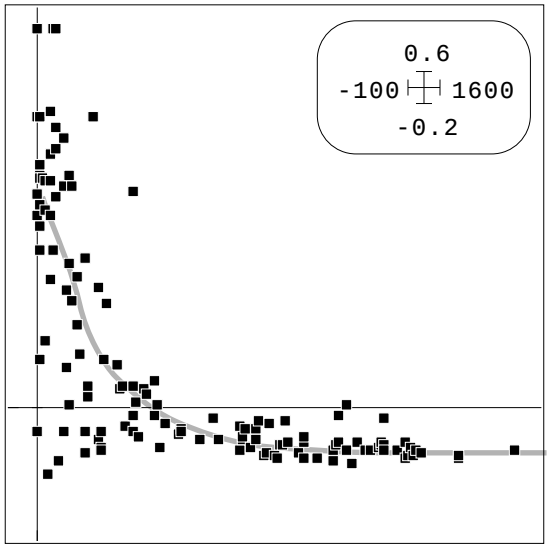


La présence d'un facteur dominant s'impose. On obtient la figure 2 par Tables : TabMeanVar :



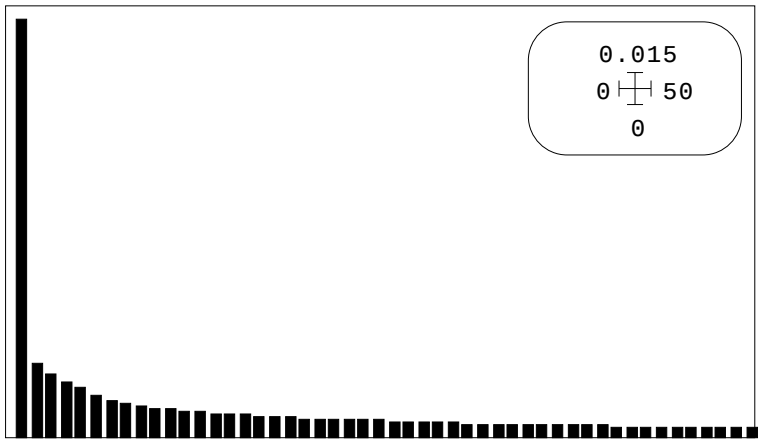
Dans la figure 2, on trouve l'essentiel du principe de cette analyse. En abscisse, sont positionnées les espèces par un code centré et réduit pour la pondération unitaire. Chaque quadrat est à la moyenne des espèces qu'on y trouve. Le recentrage de la coordonnée des colonnes est un effet technique sans signification.

Les scores des quadrats sont très liés à l'altitude moyenne des occurrences qui sont dedans (abscisse : altitude moyenne des occurrences dans un quadrat, ordonnée : coordonnée factorielle du quadrat, modèle loess) :



Donc, en voulant prédire la présence des espèces par la variable qualitative quadrat, on en arrive à exprimer que la séparation des quadrats par leur contenu en espèces se fait essentiellement par l'altitude.

Quand à savoir si les facteurs suivants sont interprétables, le graphe des valeurs propres semble répondre négativement :



Nous y reviendrons. On retiendra pour l'instant que séparer les quadrats par leur profil espèces ou séparer les espèces par leur profils quadrat sont deux opérations distinctes correspondant respectivement à prédire les indicatrices espèces par les indicatrices quadrats et inversement. Prédire dans un sens, par exemple prédire espèce par quadrat, c'est séparer par averaging dans l'autre, par exemple séparer quadrat per espèce! On peut s'y tromper.

3.3 — Analyse discriminante d'un ensemble d'occurrences

On peut aussi faire les deux à la fois. Un argument décisif va pousser à ce choix. Le paragraphe 2.2 va d'abord dans ce sens. Nous sommes partis *a priori* avec deux choix possibles, travailler sur les indicatrices centrées, ce qui vient d'être fait, ou sur les indicatrices décorélées, ce qui semble légitime vue que la covariance des indicatrices centrées est totalement artéfactuelle.

La même notion de variables instrumentales, la même manipulation sur les schémas de dualité donne directement la solution par le schéma $\left(\Delta_g^{-1}\mathbf{Y}^t\Delta_o\mathbf{O}_c, \Delta_s^{-1} - \mathbf{1}_{ss}, \Delta_g\right)$ dans lequel on reconnaît immédiatement le schéma $\left(\Delta_g^{-1}\mathbf{Y}_c^t\Delta_o\mathbf{O}_c\Delta_s^{-1}, \Delta_s, \Delta_g\right)$ qui est celui de l'analyse des correspondances du tableau croisé espèces-classes, vue comme analyse canonique et reconnue comme telle par Estève [9].

On voit alors dans l'AFC une double analyse discriminante qui met au centre de l'interprétation la notion de correspondances (cases non vides du tableau), une double prédiction réciproque, une mesure symétrique de l'amplitude des espèces et de la diversité des relevés [10].

Ce choix se justifie en outre par une remarque fort simple. Nous avons mis la liste des occurrences en face d'un numéro de quadrat. Si on y met qu'une seule variable quantitative (l'altitude s'impose dans l'exemple traité) que doit-on faire ?

L'analyse directe du gradient s'impose, bien que posée à l'origine par le biais des relevés (synthèse et plus de 250 références dans Whittaker 1967 [11]). Sur le gradient, chaque occurrence de chaque espèce est un fait certain qui définit une densité de probabilité qui représente la niche de l'espèce sur ce gradient, niche qu'on résumera — si on n'a pas d'autres moyens — par la moyenne et la variance. On trouvera dans la figure 3 un exemple de la modélisation (par estimation non paramétrique de densité de probabilité, base dans [12], introduction en écologie par [13]) des courbes de réponse à l'altitude par S-PLUS

```
nomesp<-"ixbr"
z<-sort(ALTI[ESP==nomesp])
wz<-ksmooth(z,kernel="box",bandwidth=500,
x.points=seq(0,max(ALTI),by=100))
plot (wz$x,wz$y,xlab="Altitude",ylab="Densite")
lines (wz$x,wz$y)
title(nomesp)
```

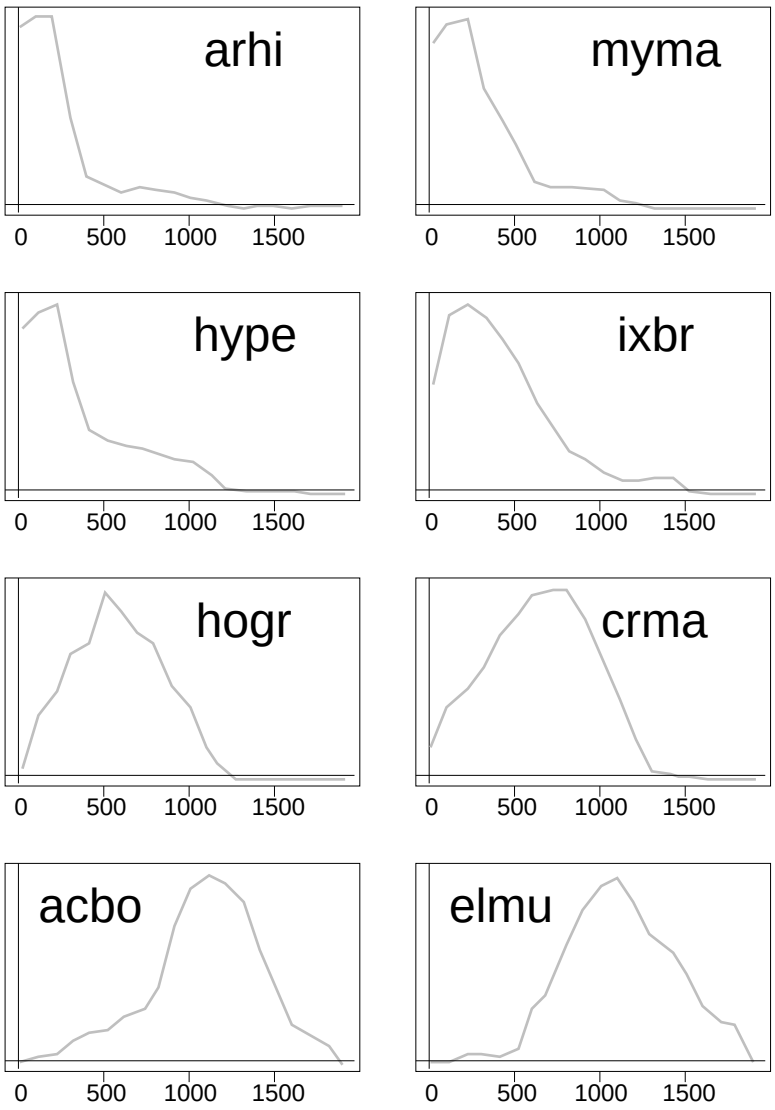


Figure 3 : Exemples d'estimation non paramétrique de densité de probabilité pour l'analyse directe des niches sur le gradient d'altitude. Les altitudes moyennes des occurrences de chaque espèce sont respectivement de 126, 191, 234, 352, 568, 635, 1090 et 1125 m. En haut la procédure en langage S-PLUS [14]

L'analyse du gradient pose simplement le calcul de la position moyenne des occurrences de chaque taxon sur le gradient et le calcul de la variance des positions moyennes comme indice de séparation des niches. Si on considère la variable de milieu quantitative comme une seule variable instrumentale, si on suppose que cette variable est centrée (moyenne nulle) pour l'ensemble des occurrences et normalisée (variance unité) ce qui n'enlève strictement rien à la généralité du raisonnement, la prédiction de l'indicatrice de l'espèce *k* par cette variable **y** prend la valeur $f_k moy(\mathbf{y}/k)$. Il faut utiliser la métrique Δ_s^{-1} pour récupérer comme inertie projetée la variance des positions moyennes des taxa.

La variance des positions moyennes des espèces (comprise entre 0 et 1, puisque rapportée à la variance initiale) est donc le carré du cosinus de l'angle entre la variable **y** et le sous-espace engendré par les indicatrices centrées. Il s'agit *implicitement* d'une analyse canonique.

Si donc le lien avec une variable quantitative est celui d'une analyse canonique, le lien avec une variable qualitative devrait être de même nature et l'analyse d'un tableau espèces-occurrences devrait être une analyse des correspondances.

Quand la mesure écologique de base est l'enregistrement d'une occurrence, l'analyse canonique joue un rôle central. L'AFC et ses dérivées sera la méthode privilégiée de l'information biologique quand on ne veut tenir compte que des présences. On démontre mathématiquement [15] que l'analyse canonique des correspondances est l'analyse discriminante du tableau de milieu dont les lignes sont dupliquées pour chaque occurrence d'une espèce, discrimination sur la variable nom d'espèce. Cette pratique a été explicitement introduite par Green [16] avec l'argument fondamental suivant :

> *If the species is absent, there are three possible interpretations :(i) The species cannot live there; that is, its niche does not include that point. (ii) The species can live there, but never had the opportunity for zoogeographic reasons. (iii) The species can and does live there, but the sample failed, by chance, to include a representative of that species.*

Comment mieux dire qu'un zéro dans un tableau écologique n'a pas de sens ! Seule l'observation du taxon est une information. Ceci a des conséquences importantes. Dans certaines situations expérimentales, ce peut être totalement faux car on a pris toutes les précautions pour qu'une valeur nulle soit l'indice d'une absence (études sur la pollution). Dans d'autres situations ce peut être totalement vrai (les conditions météorologiques déterminent la quantité de chants des oiseaux et l'AFC est optimale[17]. On est peut être souvent entre les deux (une proportion d'absence est un alea, une autre une information).

Dans une seule situation, il n'y a pas d'ambiguïté, c'est celle que nous étudions. Seule des occurrences sont comptabilisées et il n'y a pas de relevés. Cette situation particulière montre qu'utiliser l'AFC et ses dérivées, c'est ne voir dans un tableau qu'une liste d'occurrence, ce qui a été explicité dans [10].

Le principal défaut de l'analyse canonique porte sur l'exigences du nombre d'individus par rapport aux nombres de variables des deux tableaux. Dès qu'il y a plusieurs occurrences par espèces, aucune difficulté n'est à craindre. En multipliant par contre les espèces rares, donc les variables surnuméraires, le danger de sortir du domaine de validité apparaît (ce qui est bien connu en AFC).

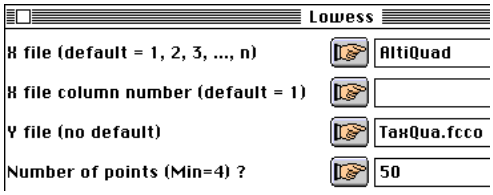
On est maintenant sûr que l'ordination doit comporter plusieurs méthodes et des choix. La première question sera "dans le tableau tout zéro est-il une absence d'information ? "

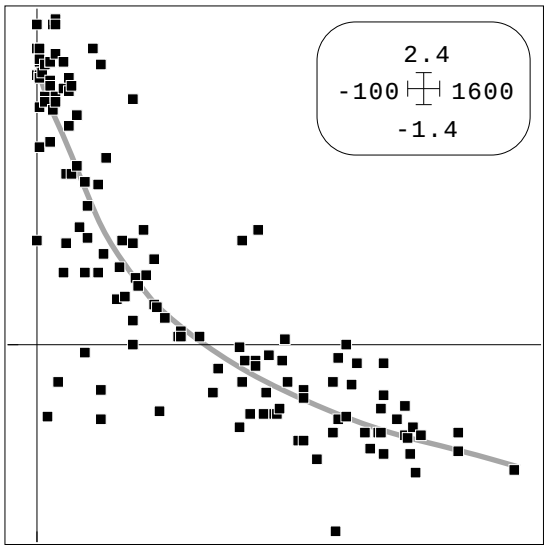
Pour un tableau issu d'une liste d'occurrence, la réponse est certainement oui. On ne sait rien de l'intensité de la prospection et on ne saura jamais si un quadrat pauvre n'est pas le résultat pur et simple du manque de pistes, du caractère peu amène des habitants ou de la mauvaise réputation du potentat local.

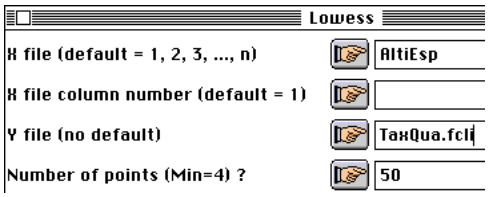
Pour un relevé en milieu semi-aride, la réponse est certainement non. On peut commettre des oublis, faire des erreurs de détermination mais l'essentiel se voit. Pour un relevé ichtyologique, cela peut dépendre. Si l'échantillon est un coup de sonde dans une eau boueuse un jour de crue ce sera plutôt oui. Si l'échantillon est le bilan de plusieurs visites, par plusieurs méthodes, pendant plusieurs années, ce sera non.

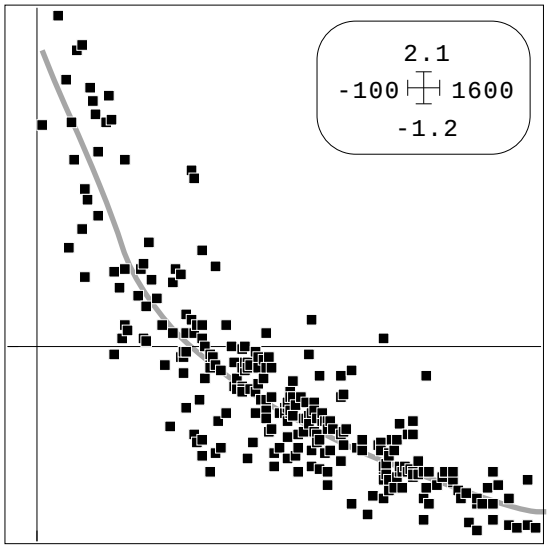
Si c'est oui, seules importent les cases non nulles et l'AFC comme analyse canonique est valide. Si c'est non, l'AFC rapportant tout aux occurrences peut être une calamité : une part importante d'information est perdue d'entrée de jeu.

L'analyse des correspondances du tableau de l'exemple donne un premier axe lié à l'altitude. On peut calculer l'altitude moyenne des occurrences arrivés dans chaque quadrat (altitude du quadrat) et l'altitude moyenne des occurrences de chaque espèce (position de niche). La coordonnée retrouve globalement le lien :

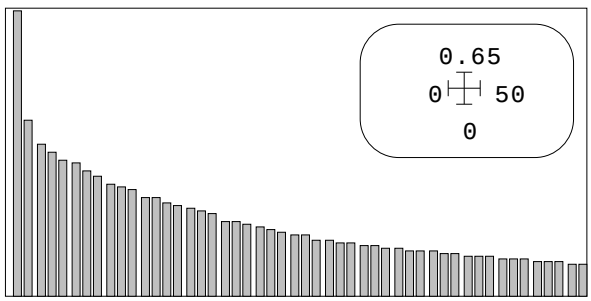


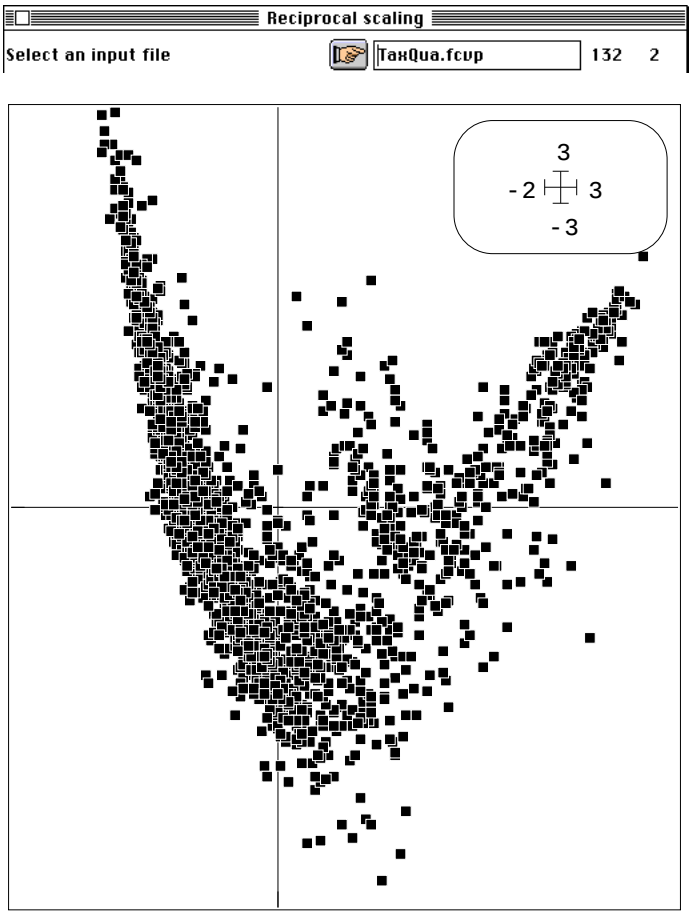






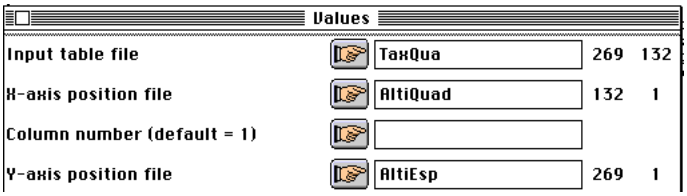
On notera simplement la différence : il y a plus de quadrat en bas et plus d'espèces en haut. Or, on ne peut faire l'impasse sur l'axe 2 :

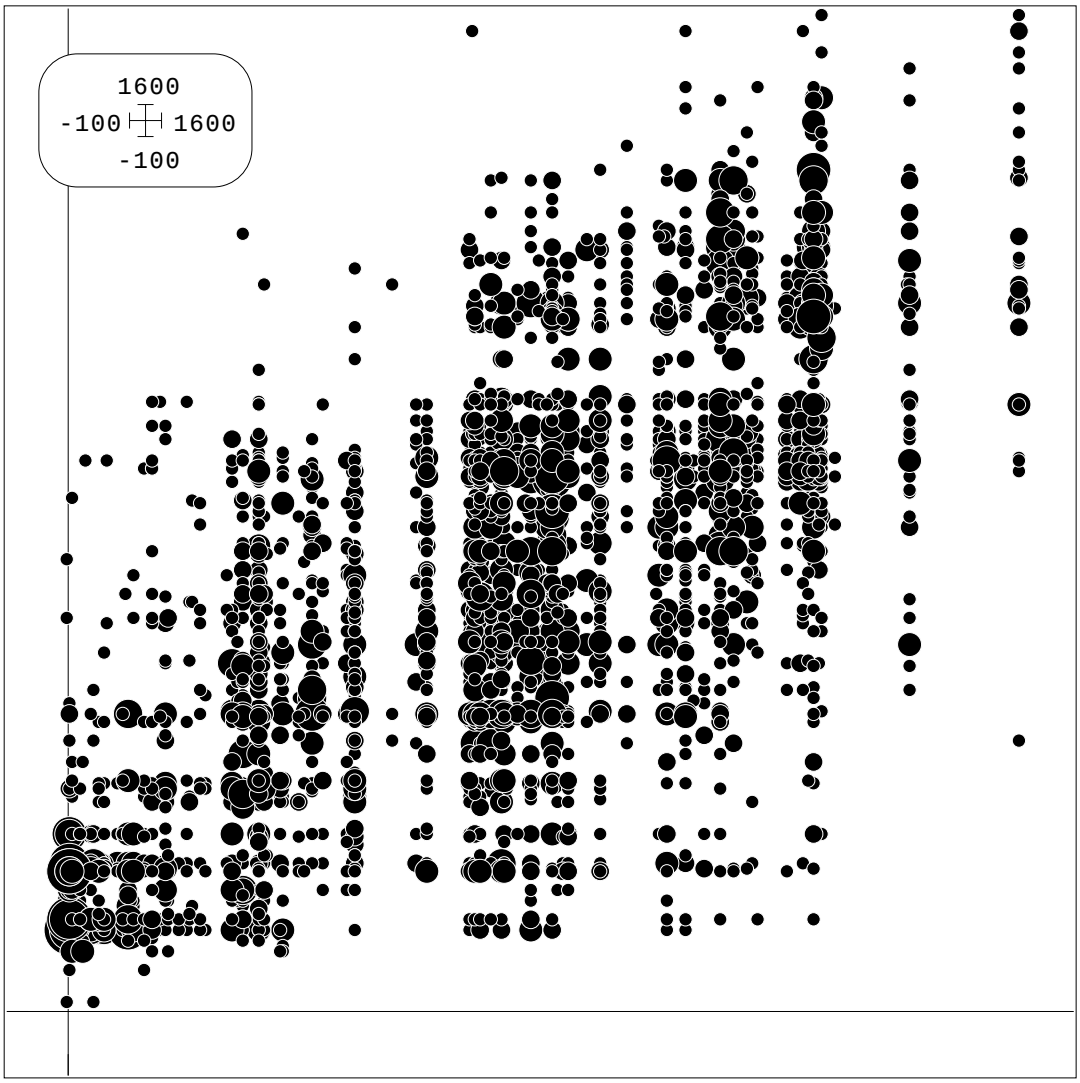




Les occurrences sont positionnés par leur score sur deux axes pour optimiser d'un seul coup et à égalité les classes espèces et les classes quadrats. On parlera sans doute d'effet Guttman. On peut y voir *a contrario* un fait différent. L'ordination altitudinale se fait sur un double schéma par "plaine contre montagne" d'une part, par "étagement continu des niches" en partie haute d'autre part. Ces deux faits sont liés au centre *artefactuellement*.

Ce qui frappe, de quelque manière qu'on s'y prenne reste l'omniprésence des grandes amplitudes de niche qui entretiennent une diversité intra-quadrat très forte. Dans la figure qui suit, on trouvera la représentation de l'altitude par quadrat et par espèce pour souligner la symétrie des deux informations dans une liste d'occurrences. Chaque occurrence est un numéro d'espèce et un numéro de quadrat. Une variable mesurée sur les occurrences peut alors attribuer une valeur par averaging sur une partition comme sur l'autre. On fait la même chose implicitement en réécrivant un tableau avec ordination des lignes et des colonnes sur les scores d'AFC. Avec l'analyse, on soulignera que les taxa de plaine peuvent monter fort haut, tandis que les taxons de montagne ne descendent pas. Par contre en montagne, la richesse est plus forte et l'ordination par l'étagement beaucoup plus continu. Apparaît peut-être un autre facteur écologique. Il n'y a pas un gradient unique mais une dissymétrie qui exprimée par deux dimensions. La figure simple sur l'altitude est obtenue par :





Pour conclure, on dira que coupler une liste d'occurrences avec une variable par averaging c'est faire l'analyse canonique entre **O** et **y**. Pour cette raison coupler une liste d'occurrences avec une variable qualitative sera totalement cohérente avec l'analyse du gradient si on utilise l'analyse canonique entre les deux paquets d'indicatrices, c'est-à-dire si on utilise l'AFC. Ceci va nous conduire à comprendre en quoi l'analyse canonique des correspondances [18] est vraiment une analyse canonique. Le titre même de l'article initial de Ter Braak la relie à la pratique de l'ordination directe et en ce sens nous allons voir que le traitement des listes d'occurrence la valide totalement.

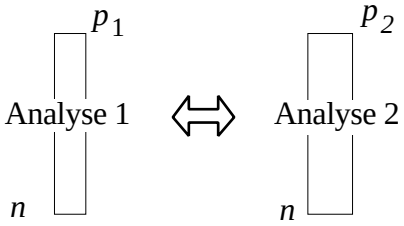
4 — L'ordination directe des occurrences

En face de **O** on trouve maintenant un tableau **Y**, c'est-à-dire plusieurs variables. On pourrait mélanger dans **Y** des variables quantitatives (altitude) et des variables qualitatives (numéro du quadrat), des variables positives (dont on cherche à définir l'effet) comme des variables négatives (dont on cherche à éliminer l'effet).

Il s'en suit une série de problèmes tous attachés au couple (**O**, **Y**). La première observation porte sur les conditions numériques. Elles se caractérise par la grandeur de *o*, le nombre d'observations (ici 3953) et *s* le nombre de colonnes de **O** (ici 269) ou *p* le nombre de colonnes de **Y** (au plus une dizaine).

Cette situation est très favorable à l'analyse canonique qui *a priori* est un cadre approprié. Cette situation est rare en écologie, le couplage habituel flore-milieu excluant cette stratégie dès qu'on parle de relevés. C'est pourquoi l'analyse canonique des correspondances est d'abord conçue comme une analyse sur variables instrumentales explicitement dans [19] ou implicitement dans l'assemblage du programme CANOCO [20]. Il suffit de savoir que l'analyse des redondances est un autre nom utilisé pour l'ACP sur variables instrumentales, laquelle est d'abord une alternative à l'analyse canonique [21].

On doit donc revenir sur la dichotomie ACPVI et AC (analyse canonique). Considérons deux tableaux d'ACP (le raisonnement s'étend à tout type) partageant la même pondération des individus.



On a deux triplets $\left(\mathbf{X}, \mathbf{I}_{p_1}, \mathbf{D}_n\right)$ et $\left(\mathbf{Y}, \mathbf{I}_{p_2}, \mathbf{D}_n\right)$. L'analyse inter-batterie [22] ou plus généralement l'analyse de co-inertie du couple [23] est basée sur le triplet :

$$\left(\mathbf{Y}^t\mathbf{D}_n\mathbf{X}, \mathbf{I}_{p_1}, \mathbf{I}_{p_2}\right)$$

Les deux ACPVI sont celles des triplets :

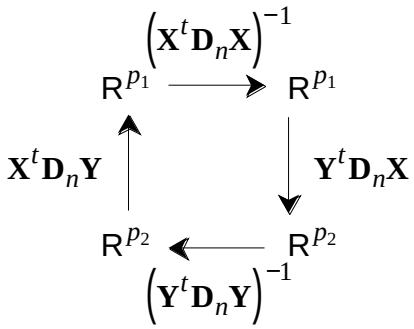
$$\left(\mathbf{Y}^t\mathbf{D}_n\mathbf{X}, \mathbf{I}_{p_1}, \left(\mathbf{Y}^t\mathbf{D}_n\mathbf{Y}\right)^{-1}\right) \text{ et } \left(\mathbf{Y}^t\mathbf{D}_n\mathbf{X}, \left(\mathbf{X}^t\mathbf{D}_n\mathbf{X}\right)^{-1}, \mathbf{I}_{p_2}\right)$$

L'analyse canonique (synthèse dans [24]) utilise :

$$\left(\mathbf{Y}^t\mathbf{D}_n\mathbf{X}, \left(\mathbf{X}^t\mathbf{D}_n\mathbf{X}\right)^{-1}, \left(\mathbf{Y}^t\mathbf{D}_n\mathbf{Y}\right)^{-1}\right)$$

Seules les métriques (identité ou inverse des matrices de covariance) sont en jeu. Ceci est le repère principal car les aides à l'interprétation basées sur ce calcul central varient à l'infini. Fondamentalement, en AC on trouve des combinaisons linéaires de variables de chacun des deux tableaux de variance unité maximisant leur corrélation.

Le schéma de l'AC contient implicitement deux projecteurs, ceux des ACPVI en contiennent un seul et celui de la co-inertie n'en contient pas. La principale propriété de l'AC s'écrit par le schéma :



Il suffit de noter que :

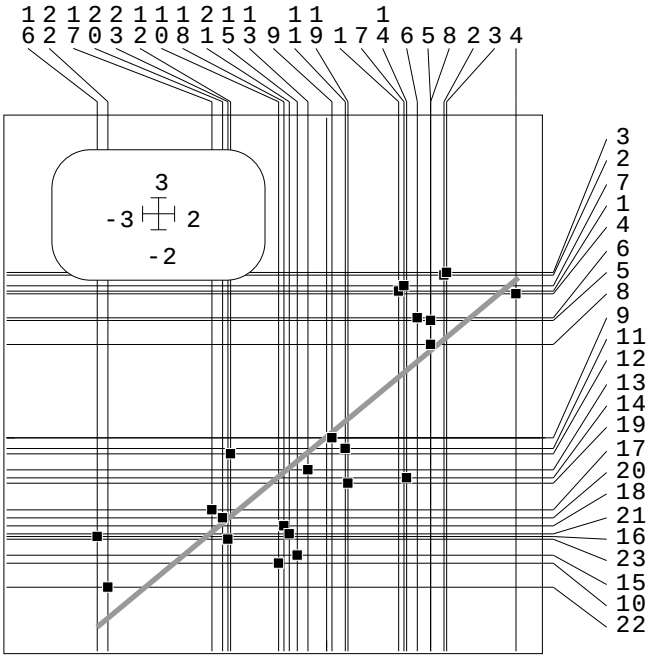
$$\|\mathbf{u}\|^2_{\left(\mathbf{X}^t\mathbf{D}_n\mathbf{X}\right)^{-1}} = 1 \qquad \mathbf{u}^t\left(\mathbf{X}^t\mathbf{D}_n\mathbf{X}\right)^{-1}\mathbf{u} = 1 \qquad \left\|\mathbf{X}\left(\mathbf{X}^t\mathbf{D}_n\mathbf{X}\right)^{-1}\mathbf{u}\right\|^2_{\mathbf{D}_n} = 1$$

La constitution de scores orthonormés (moyenne nulle, variance unité, covariance nulle quand il y en a plusieurs) est caractéristique des analyses à inversion de norme. On maximise la corrélation en maximisant la covariance sous la contrainte de variance unité. ceci est fondamental pour bien comprendre le lien avec l'averaging sous-jacent aux méthodes d'ordination sur gradient.

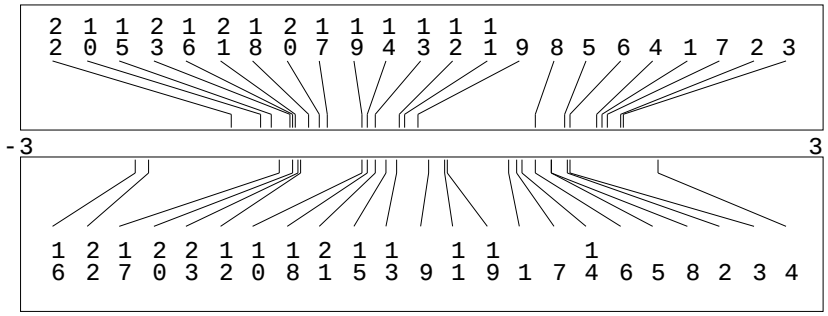
Comme exprimer une corrélation maximale. Cela dérive de la simple observation que pour des variables normées :

$$\|\mathbf{x}-\mathbf{y}\|^2_{\mathbf{D}_n} = \|\mathbf{x}\|^2_{\mathbf{D}_n} + \|\mathbf{y}\|^2_{\mathbf{D}_n} - 2\langle\mathbf{x}|\mathbf{y}\rangle_{\mathbf{D}_n} = 2 - cor(\mathbf{x},\mathbf{y})$$

Il s'en suit qu'exprimer la corrélation, c'est mesurer la somme des carrés des écarts dans chaque couple de valeur. Maximiser la corrélation, c'est minimiser la somme des carrés des écarts. On pense habituellement la corrélation dans un graphe bivarié :



On peut la représenter par deux graphes univariés :

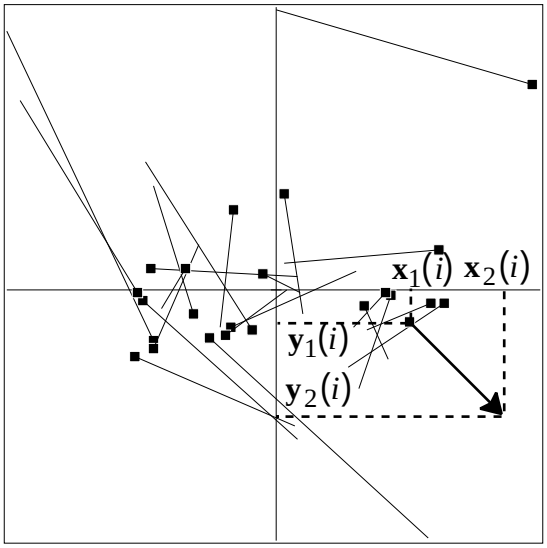


La corrélation est soit l'adéquation de la droite de régression, soit l'erreur commise entre les deux systèmes de position (en négatif).

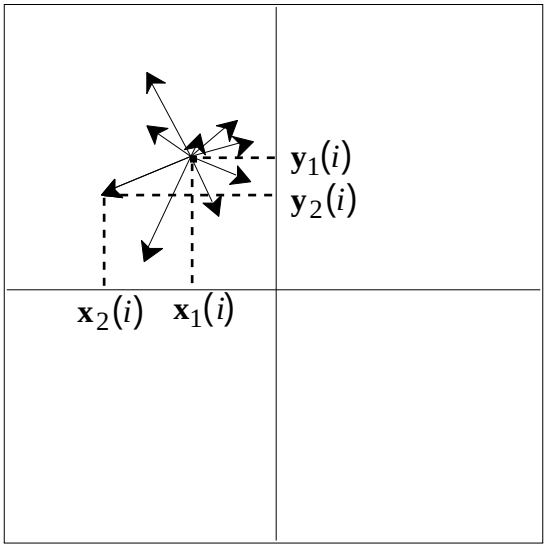
Si on étend ce raisonnement à deux variables $\mathbf{x}_1, \mathbf{x}_2$ et à deux variables $\mathbf{y}_1, \mathbf{y}_2$ non corrélées par couples ($cor(\mathbf{x}_1, \mathbf{x}_2) = 0$ et $cor(\mathbf{y}_1, \mathbf{y}_2) = 0$), alors :

$$\sum_{i=1}^{n} p_i \left[(\mathbf{x}_1(i) - \mathbf{x}_2(i))^2 + (\mathbf{y}_1(i) - \mathbf{y}_2(i))^2 \right] = 4 - 2cor(\mathbf{x}_1, \mathbf{x}_2) - 2cor(\mathbf{y}_1, \mathbf{y}_2)$$

Le graphe qui exprime deux corrélations simultanément est donc :

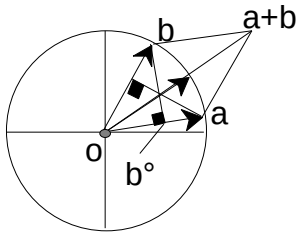


Maximiser la corrélation entre deux systèmes de représentation, c'est minimiser la somme des carrées des écarts entre ces deux systèmes, c'est optimiser la prédiction de l'un par l'autre et réciproquement. Que va t'il se passer alors avec un tableau d'occurrences **O** et un tableau de variables sur ces mêmes occurrences **Y** ?



Toutes les occurrences associées à la même espèce tombent au même endroit puisque la valeur des indicatrices y prennent les mêmes valeurs (toute combinaison d'indicatrice est constante par classe). Par contre chaque occurrence vue par le milieu a sa position propre. Naturellement on se dit que l'optimum sera atteint si la position de l'espèce (occurrence) est à la moyenne des positions de l'espèce (milieu).

Certes, mais dans ce cas la variance des positions des occurrences ne peut plus être unitaire, elle est forcément plus petite puisqu'il y a eu projection. Le paradoxe tient à la notion de bissectrices (dans R^n et entre sous-espaces mais c'est déjà vrai dans R^2) :



a est une combinaison des variables de **X**. Il est dans le sous-espace engendré [**X**]. **b** est une combinaison des variables de **Y**. Il est dans le sous-espace engendré [**Y**]. Comme produit de l'AC, **a** et **b** maximise la corrélation, donc le cosinus de l'angle, donc minimise l'angle. **a** est le vecteur de [**X**] le plus prédictible par [**Y**] et **b** est le vecteur de [**Y**] le plus prédictible par [**X**]. **a** et **b** sont le plus près possibles de leur demi-somme qui est à la même distance angulaire et réelle de chacun d'eux (ils sont normés). On dira la même chose si on cherche une position des espèces (position des occurrences communes de chaque espèce) les plus près possibles en moyenne des positions des occurrences-milieu (de variance 1) ou si on cherche une position des espèces de variance 1 (position des occurrences communes de chaque espèce) la plus corrélée à la des occurrences-milieu de variance 1.

Dans un cas on utilise **a** et **b**, dans l'autre **a** et **b°**. On pourrait de même utiliser **a** et **a°** ou encore la bissectrice portée par **a**+**b** et les projections associée de **a** et **b**. Le très étonnant de cette affaire est simplement que cette observation géométrique explique pourquoi on peut, dans une AFC, indifféremment :

1) mettre les relevés à la moyenne des espèces ;
2) mettre les espèces à la moyenne des relevés ;
3) mettre les relevés et les espèces à la moyenne des occurrences.

La symétrie et la multiplicité des points de vue n'est vraiment utile que dans l'analyse canonique de deux paquets d'indicatrices. Quand il n'y a qu'un des deux ensembles qui a cette propriété le plus simple est de s'en tenir à l'un des trois possibles. Le plus clair, c'est placer les occurrences par les variables de milieu avec des scores non corrélés de variance unité et les espèces (ensemble d'occurrences portant le même nom) à la moyenne des positions qui leur appartiennent.

Or ceci est strictement l'analyse canonique des correspondances qui se trouve être une véritable analyse canonique que dans le cas ici étudié. En se demandant ce qu'il convient de faire de particulier sur les occurrences, on en vient à affirmer qu'il faut faire de l'AFC et de l'ACC, ce qu'il vaut mieux ne pas faire dans le cas des relevés.

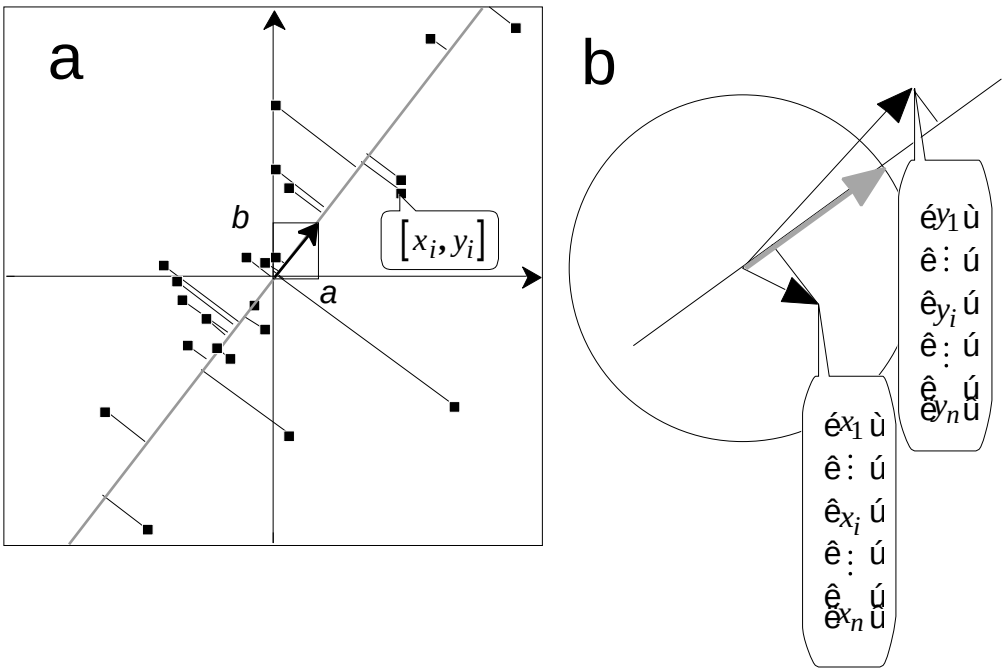
Tous les arguments convergent vers ce point de vue. L'indépendance des lignes du tableau **Y** est totalement invalidée dès qu'on duplique les lignes relevés (milieu) pour chaque espèce qui s'y trouve. Cela n'intervient pas dans les tableaux d'occurrences. De plus, une des propriétés les plus discutables en ACC sur tableaux est le centrage des variables de milieu par les poids des relevés en espèces (comme si un relevé sans espèce ne devait pas participer à la mesure de la variabilité de l'environnement). Par contre, dans le cas des occurrences, la valeur moyenne des variables de milieu est la valeur moyenne pour les occurrences qu'on a effectivement observées. Les biais d'intensité de prospection sont ainsi automatiquement éliminés. Seul compte ce qui est arrivé.

En bref, l'ACC est la méthode de couplage milieu-occurrences. Elle s'étend à son cas habituel d'utilisation que si on ne voit dans le tableau que cet aspect. C'est possible mais pas toujours, tant s'en faut, souhaitable. Reste à réécrire le programme !

5 — Dépouillement d'une analyse canonique

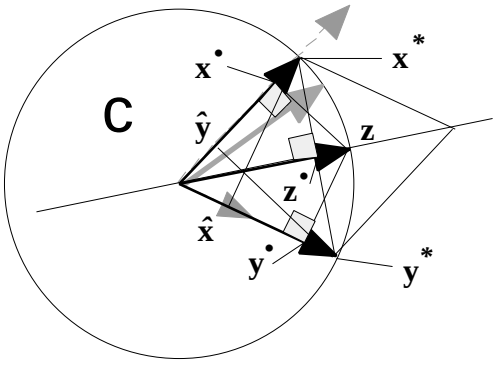
5.1 — Deux variables quantitatives

Faire l'ACP de deux variables quantitatives, c'est trouver l'axe principal du nuage centré :



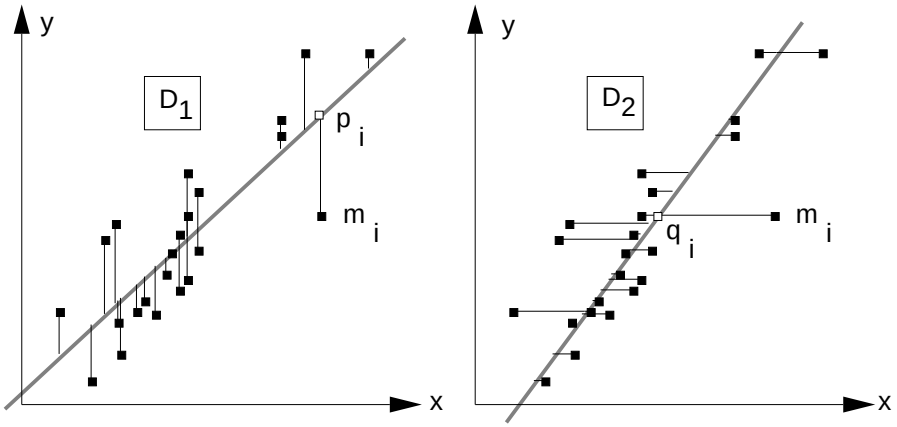
(a) Vue dans $\mathbb{R}^2$, cette ACP donne l'axe principal de composantes *a* et *b* avec $a^2 + b^2 = 1$ qui minimise la somme des carrés des distances des points à la droite portée par l'axe[25]. Quand les variables sont normées, on a toujours $a = b = \sqrt{2}/2$. (b) Vue dans $\mathbb{R}^n$, s'il y a *n* points de mesure, cette ACP donne le vecteur de $\mathbb{R}^n$ (composante principale) qui a la même propriété pour le nuage des deux variables.

Faire l'analyse canonique du couple de variables **x** et **y** est fort différente. Toute notion de variabilité inégale des variables est éliminée au profit de la seule considération apportée à la corrélation. On cherche la bissectrice des deux variables :



On considère donc uniquement les variables normalisées $\mathbf{x}^*$ et $\mathbf{y}^*$. La bissectrice est un score normalisé qui optimise $cor^2(\mathbf{x}, z) + cor^2(\mathbf{y}, z)$. A un coefficient près, mesurer la corrélation entre **x** et **y** est évidemment équivalent à mesurer les corrélations de **x** et

de **y** avec **z**. C'est aussi équivalent de prédire **z** respectivement par **x** et **y** ou de prédire directement **x** par **y** ou **y** par **x**. C'est numériquement équivalent mais pratiquement il n'en n'est rien. On sait en effet que le carré de corrélation mesure la valeur de chacune des deux droites de régression, mais que ces deux droites donnent des résultats et des points de vue fort différents :

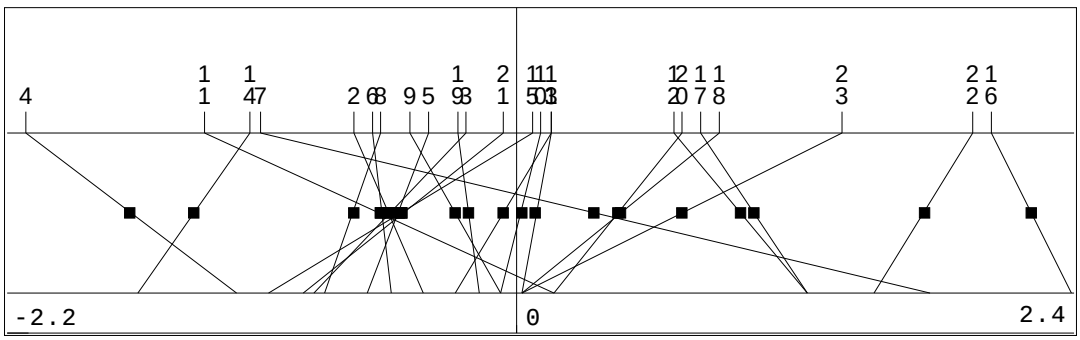


On a immédiatement toute la difficulté du problème dans son cas le plus simple. Dire qu'on a une corrélation de 0.8, c'est dire que 64% de la variabilité d'une variable est prédite par l'autre et réciproquement, ou que chacune des deux est prédite par une même troisième à 81% ou que les deux sont prédites simultanément à 90%. C'est pourquoi l'ACP normée est une analyse canonique, que l'AFC est une analyse canonique, mais que les deux sont aussi des ACP. L'analyse canonique est attribuée à Hotteling (1936)[26] comme l'ACP normée (1933)[27].

On a ici la source très élémentaire d'un grand nombre de difficultés. En analyse des correspondances prendre **z** avec $\mathbf{x}^{\bullet}$ et $\mathbf{y}^{\bullet}$, c'est placer les correspondances puis les espèces et les relevés par averaging [10] alors que utiliser $\mathbf{x}^{*}$ et $\mathbf{y}^{*}$ avec $\mathbf{z}^{\bullet}$ c'est placer les espèces et les relevés avec des variances unité et les correspondances au milieu (graphe de Heiser[28]) alors qu'on peut encore mettre les espèces à la moyenne des relevés ou les relevés à la moyenne des espèces ($\mathbf{x}^{*}$ et $\hat{\mathbf{y}}$ ou $\mathbf{y}^{*}$ et $\hat{\mathbf{x}}$) comme souligné par Oksanen [29].

Faire l'analyse canonique de deux variables quantitatives, c'est mesurer leur corrélation. Exprimer cette corrélation graphiquement n'est pas aussi simple qu'on pourrait le penser si on veut une vision globale cohérente. On peut privilégier une solution symétrique entre **x** et **y**. Les objets les plus connus de l'analyse canonique étant les variables canoniques (voir encore Rencher 1988 [30]), il semble logique de garder $\mathbf{x}^{*}$ et $\mathbf{y}^{*}$ avec $\mathbf{z}^{\bullet}$ comme indicateur du lien, donc choisir la voie de Heiser.

Dans ce cas, on obtient (Graph1D: Match_two_var) :



On appellera graphe de Heiser à deux variables la figure ci-dessus. En bas, *n* points sont positionnés par leur valeur sur **x**. En haut les mêmes points sont positionnés par leur valeur sur **y**. Les variables sont normalisées (moyennes nulles et variances unitaires). Au milieu, on a la variable $(\mathbf{x}+\mathbf{y})/2$ de variance $1/2+cor(\mathbf{x},\mathbf{y})$. La moyenne des carrés des différences entre les deux séries de positions vaut :

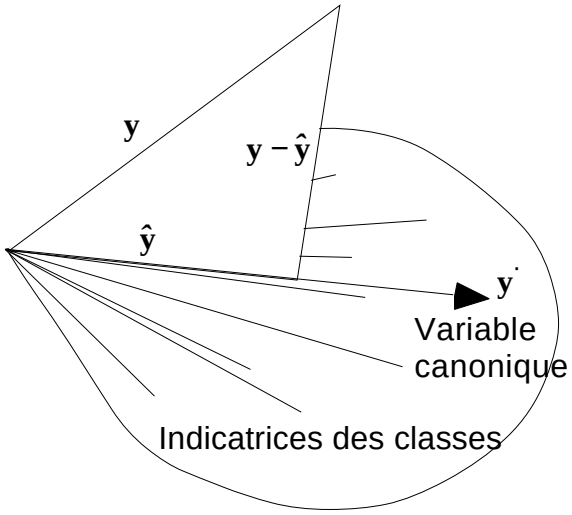
$$E(\mathbf{x},\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2 = 2(1 - cor(\mathbf{x},\mathbf{y}))$$

On lit la valeur de la corrélation tant par la variance des positions médianes (grande avec la corrélation) que par la moyenne des erreurs (petite avec cette corrélation). Cette figure est la représentation de base de deux scores d'analyse canonique. A deux dimensions, on retrouve les graphes obtenus par Scatters: Match two scatters que nous avons proposé pour représenter la partie analyse canonique de l'analyse de co-inertie.

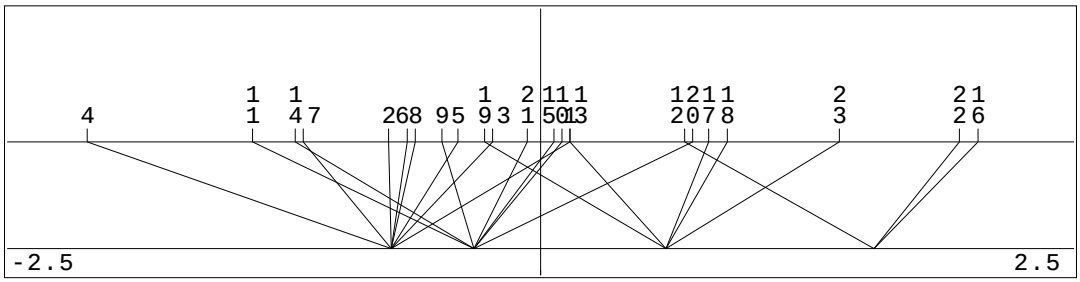
5.2 — Une variable qualitative et une variable quantitative

Quand on met face à face une variable qualitative et une variable quantitative, il est d'usage de faire la moyenne de l'une sur les classes définies par l'autre, ce qui correspond à une projection qui est encore une régression. L'équation de l'analyse de variance qui définit le rapport de corrélation, apparenté au carré de corrélation est un point de vue de variables instrumentales (totalement décrit dans le chapitre 4 de l'ouvrage de Takeuchi et Coll. [31]).

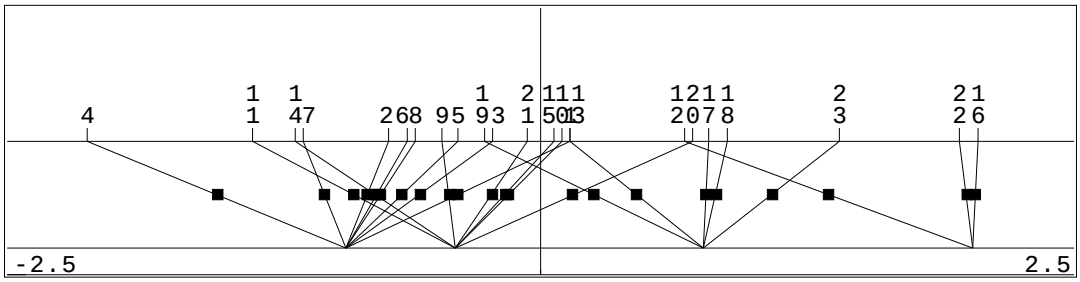
Les indicatrices des classes forment une base orthogonale du sous-espace qu'elles engendrent, ce qu'on représente par :



La variable **y** est normée. La variable $\hat{\mathbf{y}}$ est celle dont les composantes sont obtenues à partir de celles de **y** en remplaçant une composante par la moyenne des valeurs de la même classe. La variable $\mathbf{y}^{\cdot}$ est proportionnelle à la précédente mais sa variance égale 1. Dans le point de vue régression $\hat{\mathbf{y}}$ est le vecteur du sous-espace des indicatrices le plus proche de **y**. Sa variance vaut le rapport de corrélation (théorème de Pythagore). La corrélation entre **y** et $\mathbf{y}^{\cdot}$ est aussi le rapport de corrélation (point de vue analyse canonique). Représenter **y** et $\hat{\mathbf{y}}$ donne :



Les classes sont à la moyenne des points qui sont dedans. Le rapport de corrélation se lit par la variance des positions des classes. Le point de vue analyse canonique s'exprime par :

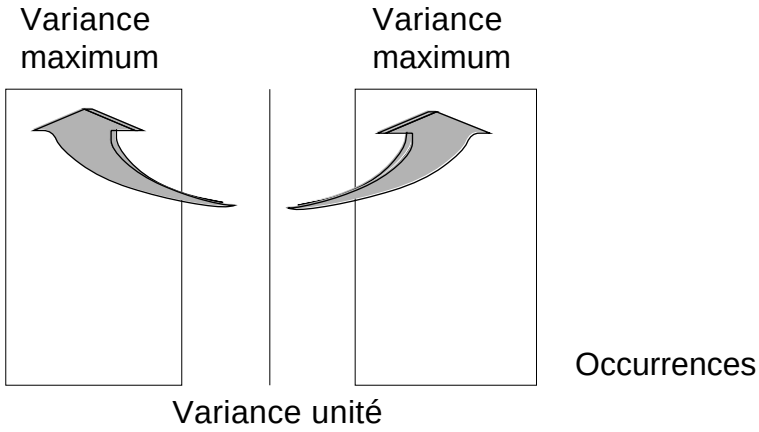


La variance des deux positions est unitaire. Le rapport de corrélation se lit par la corrélation entre les deux scores. La différence est très faible mais elle est explicitée pour des raisons de cohérence de tous les points de vue.

5.3 — Deux variables qualitatives

La corrélation de deux variables qualitatives est exprimée par l'analyse des correspondances de la table de contingence qui les croise. Il est très rare qu'on veuille expliciter cette corrélation au niveau des individus porteurs des numéros de modalité. Il est vrai que si 1000 individus répondent à 2 questions à 4 modalités, il y a 1000 points répartis sur 16 positions possibles. Dans un tableau d'occurrences à 200 espèces répartis en 100 classes, il y a 20 000 positions possibles et les voir peut ne pas être privé d'intérêt.

Si les correspondances sont positionnées avec un score de variance unité espèces et sites se positionnent par averaging avec optimisation simultanée des variances (figure 4). L'intérêt porté aux variances tant inter que intra se retrouvent dans la figure 5 qui montre le long du gradient amont-aval la forte dissymétrie des structures lignes et colonnes. La contrainte et le critère de ce type d'approche s'inscrit dans le schéma :



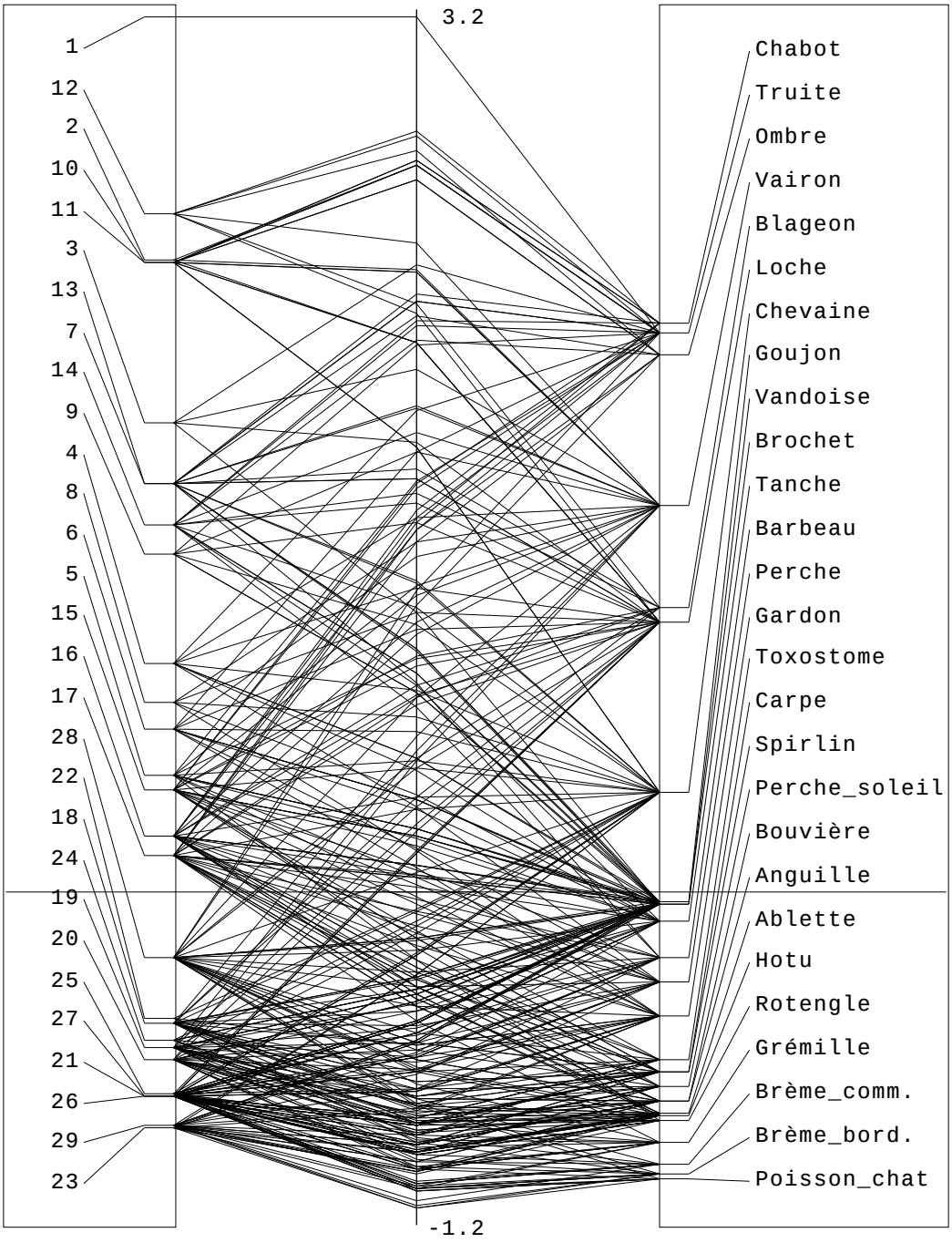


Figure 4 : Analyse des correspondances comme double analyse discriminante. Les occurrences ont un score de variance 1 optimisant simultanément la variance des positions des lignes et des colonnes du tableau. Chaque correspondance est un point raccordé d'une part à la ligne, d'autre part à la colonne qui définissent cette correspondance.

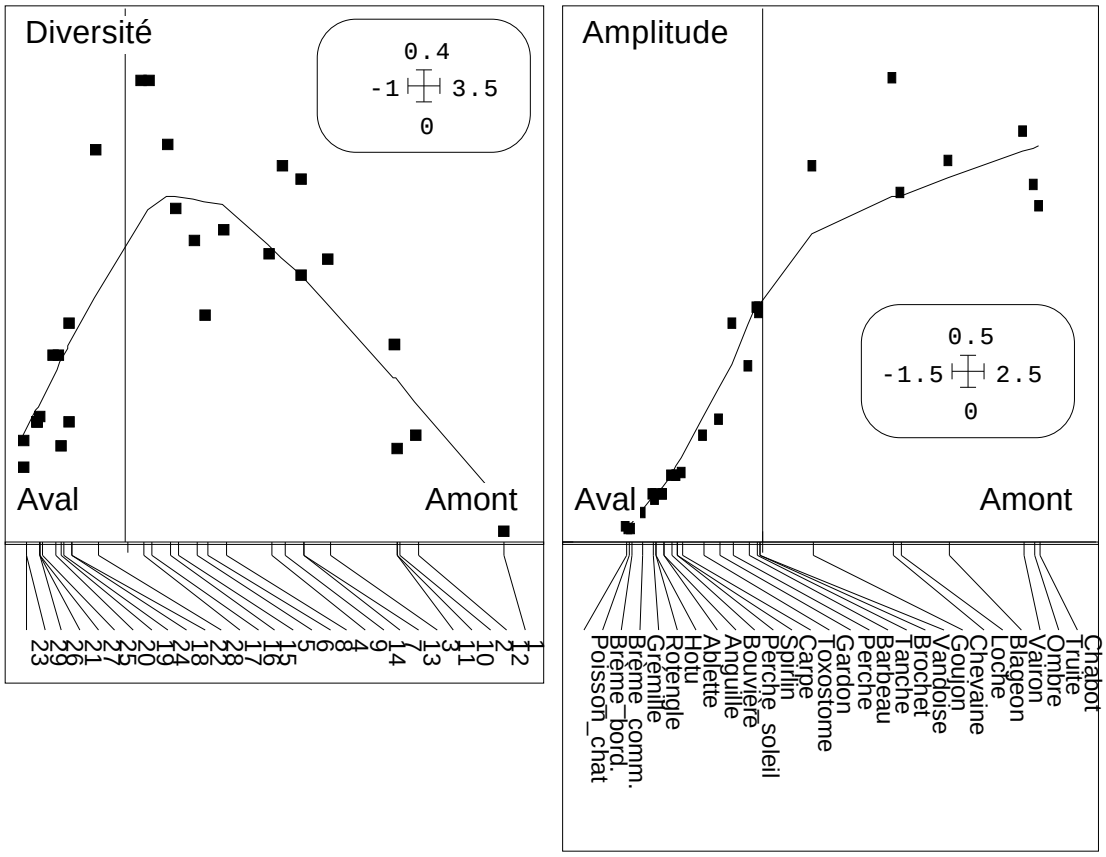
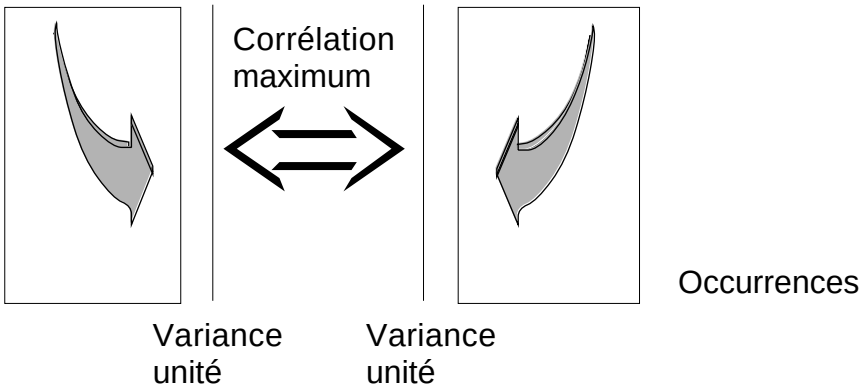


Figure 5 : Symétrie amplitude-diversité dans un tableau faunistique (données de J. Verneaux [32]sur la rivière Doubs.

On peut lui préférer le schéma propre à l'analyse canonique :



Ceci s'exprime dans la figure 6. Sans sectarisme, on peut dire que la figure 4 est plus explicite que la figure 6 dans la mesure où les variances ont une signification écologique incontournable (amplitude-diversité). C'est moins la corrélation qui fait sens que la variation. Placer une espèce à la moyenne de ses occurrences pour parler de son amplitude est un avantage décisif. Si on préfère ce point de vue, il faut le préférer tout le temps. C'est dans l'affirmation de principes généraux mis en œuvre de manières diverses que l'apport du point de vue statistique semble le plus utile. Il introduit une cohérence générale du raisonnement. C'est pourquoi nous explorons les cas possibles.

Il advient alors que si on préfère à la stratégie $\mathbf{x}^*$, $\mathbf{y}^*$, $\mathbf{z}^\bullet$ le triplet $\mathbf{z}$, $\mathbf{x}^\bullet$, $\mathbf{y}^\bullet$ il faut le faire dans toutes les circonstances, donc revenir aux cas précédents.

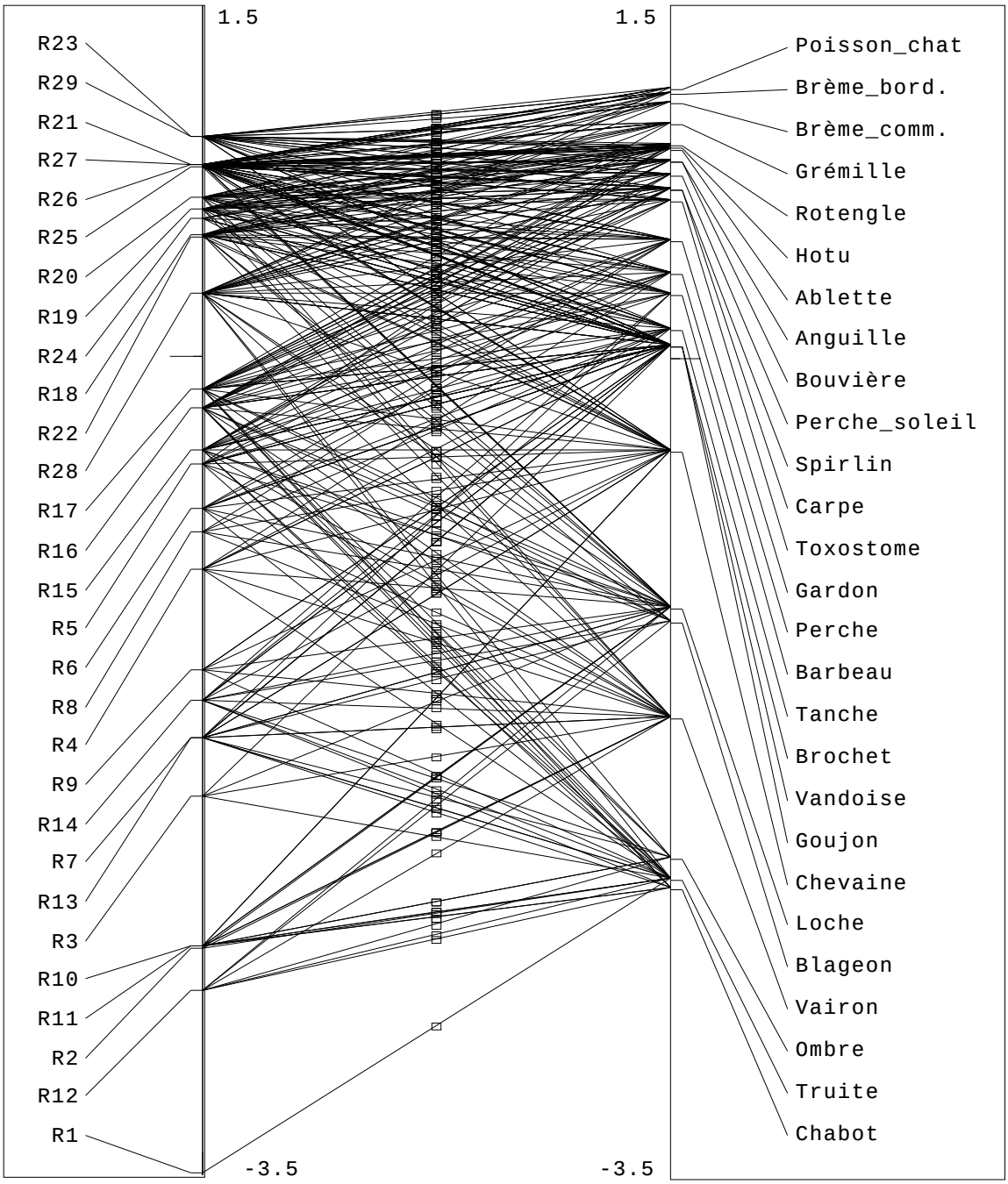
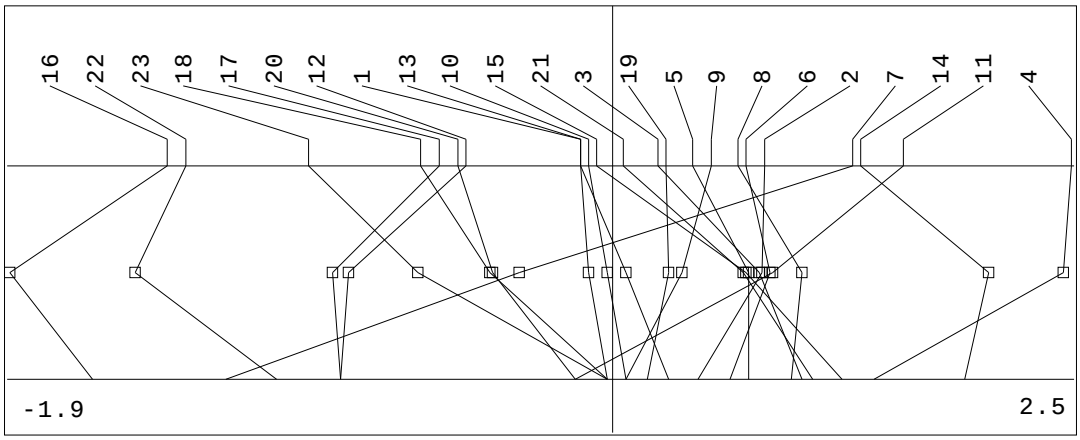


Figure 6 : Analyse des correspondances comme analyse canonique. Les occurrences ont un score de variance 1 constant par espèce et un score de variance 1 constant par relevé. Chaque correspondance est un segment raccordé d'une part à la l'espèce, d'autre part au relevé qui définissent cette correspondance. La corrélation est optimale. On notera l'équivalence des figures 4 et 6. La correspondance est figurée dans la première par un point, dans la seconde par un trait. L'optimisation dans la première porte sur les variances, dans la seconde elle porte sur la corrélation. Dans toute analyse canonique, cette dualité de point de vue est sous-jacente.

Revenons donc au couple de variables quantitatives, le cas le plus simple. Dans la figure ci-dessous est exprimée la corrélation de deux variables par le principe de l'analyse canonique, qui n'est alors rien d'autre que l'ACP normée entièrement interprétée dans l'espace des variables et utilisant la représentation d'un vecteur à *n* composantes comme une série de n positions sur un axe.

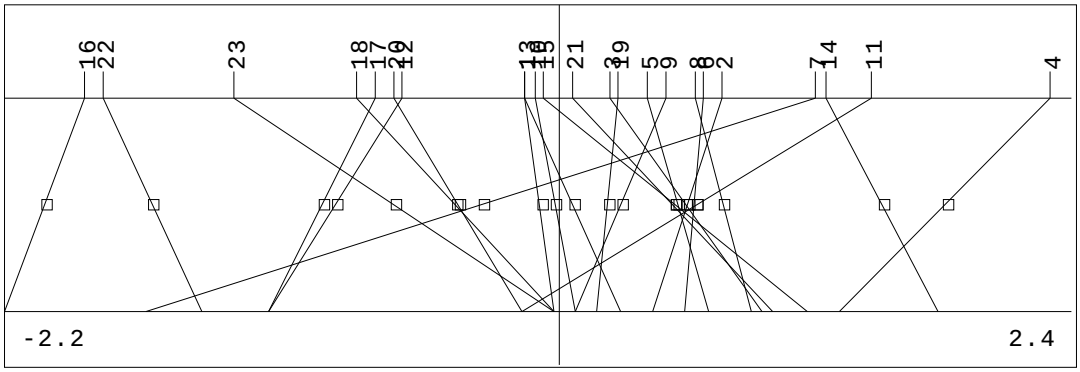


Les variables **x** et **y** sont normalisées (moyenne nulle et variance unité). On construit la variable canonique comme étant définie par $\mathbf{z} = (\mathbf{x} + \mathbf{y}) / \sqrt{2 + 2cor(\mathbf{x}, \mathbf{y})}$. La variable **z** est de moyenne nulle et de variance 1. C'est la première composante principale de l'ACP normée de ces deux variables et la bissectrice de ces deux variables. Les points sont positionnés (carrés) par les composantes de **z** (au milieu de la figure). On exécute alors la prédiction de **z** par **x** et **y** :

$$\hat{\mathbf{z}}_{\mathbf{x}} = \alpha_{\mathbf{x}}\mathbf{x} = \frac{\sqrt{2 + 2cor(\mathbf{x}, \mathbf{y})}}{2}\mathbf{x}$$

$$\hat{\mathbf{z}}_{\mathbf{y}} = \alpha_{\mathbf{y}}\mathbf{y} = \frac{\sqrt{2 + 2cor(\mathbf{x}, \mathbf{y})}}{2}\mathbf{y}$$

Ces deux prévisions, proportionnelles aux observations reconstituent aux mieux et à égalité de performance la variable centrale.



La correspondance entre deux valeurs est en haut un point et en bas (graphe ordinaire de deux variables normées) elle est un trait. On pourrait peut-être croire qu'il s'agit d'un amusement stérile. Pour les tenants du *convenient rescaling* et de bricolages pragmatiques en tout genre, c'est plus que vraisemblable. Il s'agit surtout d'expliciter que les mécanismes mathématiques, s'ils se maîtrisent difficilement au plan formel, peuvent s'expliciter au plan pratique. Le recadrage arbitraire des figures génère en grande partie la nécessité du *detrending*. Si une méthode est définie par sa propriété d'optimalité l'expression de cet optimisation ne peut faire l'objet d'une approximation sans invalider la méthode. Ce dont nous avons besoin, c'est de principes établis (Cf. [33]) et de la possibilité de les exprimer. Si on préfère la figure 4 à la figure 6, il faut préférer la figure du haut à celle du bas. C'est une question d'unité théorique.

Cela revient dans les deux cas à préciser qu'il n'y a qu'un seul type de variables canoniques. L'article de Rencher est exactement centré sur cette difficulté :

Two major types of canonical functions are considered, canonical discriminant functions that separate groups of observation vectors, and canonical variates associated with canonical correlations. The coefficients in both types of canonical functions reflect the joint contribution of the variables to the canonical functions. If the coefficients are converted to correlations, however, they merely reproduce univariate staistice values and become useless in gauging the importance of each variable in the context of the others.

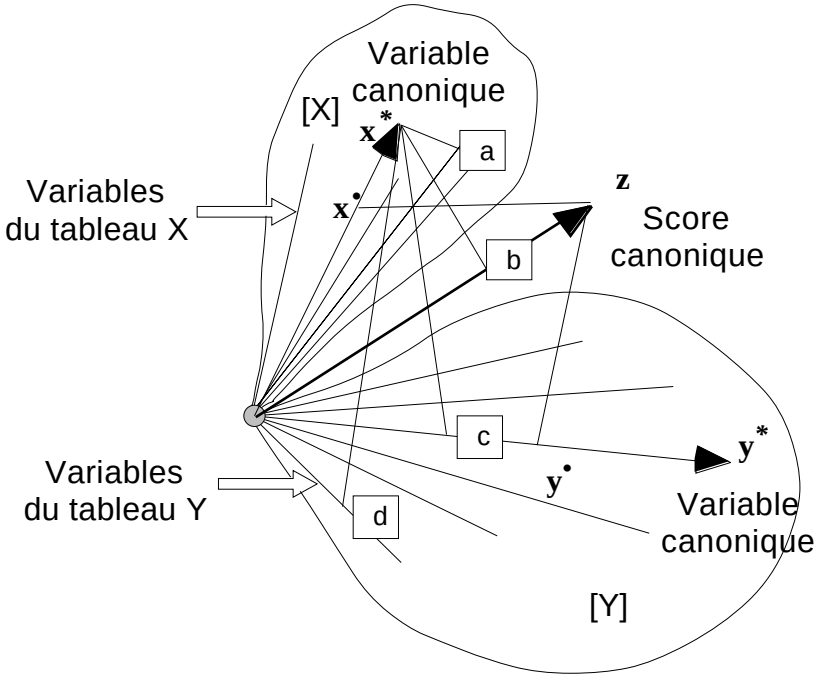
Tel est le sommaire de l'article cité. La difficulté vient de l'unicité du modèle théorique et de la séparation qu'on fait arbitrairement entre le cas (i) une qualitative et *p* quantitatives et le cas (ii) *q* quantitatives et *p* quantitatives. Nous ajoutons dans la discussion le cas (iii) une qualitative et une qualitative. Si plusieurs types de variables canoniques sont possibles, plusieurs types d'interprétation s'en suivront et la confusion ne fera qu'augmenter.

De cette discussion nous retiendrons que dans le cas (iii) le présence d'un score des occurrences qui fait une double discrimination (*canonical discriminant function*) vaut mieux que deux scores des occurrences qui maximise leur corrélations (*canonical variates*) ou que deux scores des occurrences qui optimise chacun un rapport de corrélation (*canonical discriminant functions*). Dans ce cas pour conserver l'unité d'interprétation dans le cas (ii) il vaut mieux partir du centre (score unique des lignes des tableaux) pour remplacer les deux discriminations symétriques par deux régressions multiples symétriques.

Ce point de vue est validé par un dernier argument : il s'étend naturellement aux *K*-tableaux sans rupture. L'analyse canonique vue de cette manière est alors directement le cas particulier de l'analyse canonique généralisée (ACG, [34]) dans le cas $K = 2$.

5.4 — Deux tableaux quantitatifs

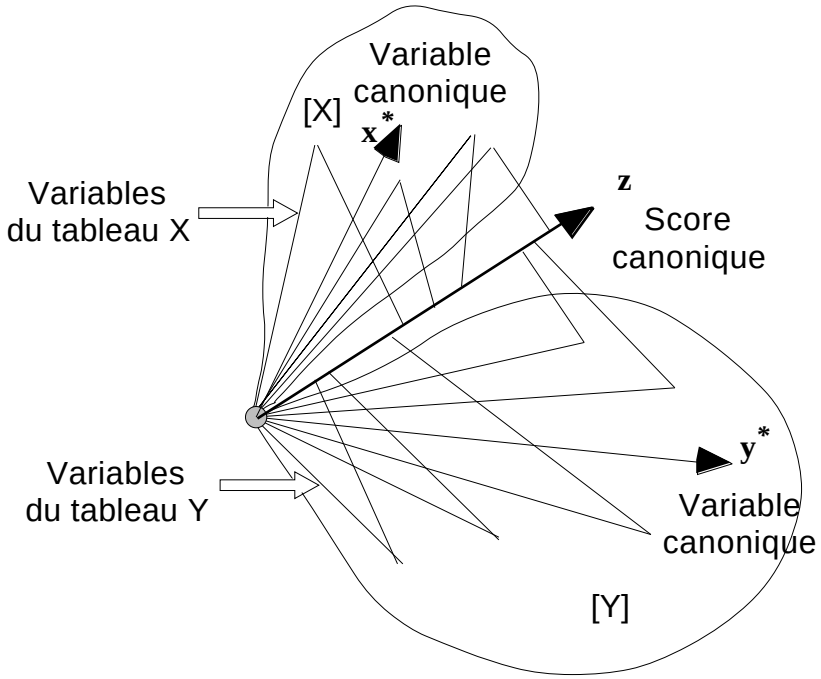
L'analyse canonique (AC) classique a été étudiée à fond, d'un point de vue écologique, par Gittins [35] dont l'ouvrage de synthèse [36] est achevé. L'auteur reconnaît pleinement l'AFC comme analyse canonique (Chapitre 5, *Dual scaling*) Si le principe de l'AC est transparent, les conditions de son utilisation le sont moins de par la multiplicité des approches possibles :



La variable canonique du tableau **X** a en effet une corrélation avec (a) les variables du tableau **X,** (b) le score canonique (bissectrice), (c) la variable canonique du tableau **Y**, (d) les variables du tableau **Y**. Les corrélations réciproques sont évidemment en jeu, de mêmes que l'expression des variables canoniques comme combinaisons des variables de départ (*loadings*), ou encore les corrélations entre variables d'un même groupe (*intraset corrélation*) et entre variables des deux groupes (*interset corrélation*).

La multiplication des aides à l'interprétation est un handicap considérable. Nous avons vu que dans le cas des indicatrices de classe, c'est l'expression du vecteur projeté ($\mathbf{x}^{\bullet}$) contient plus d'information explicite que le calcul de ce vecteur. Il vaut mieux voir le vecteur que la façon dont il est calculé. C'est encore vrai dans une analyse des correspondances multiples vue comme analyse canonique généralisée : le plus simple est d'exprimer en quoi le score des lignes rend compte de la partition introduite par chacune des variables (introduction graphique dans [37]). Le même schéma est en jeu dans l'ACP normée vue comme analyse canonique généralisée, quand on représente la composante principale comme prédicteur de chacune des variables (introduction graphique dans [38]. Maximiser une somme de rapport de corrélations (ACM voir le paragraphe 2.5 de la synthèse de Tenenhaus et Young [39]), maximiser une somme de carrés de corrélation (ACP *ibidem* et [40]) conduit à mélanger les deux sans difficultés dans l'analyse de Hill & Smith [41] (module MCA: Hill & Smith Analysis).

Certains auteurs donnent dans une ACP normée les poids (*loadings*) des variables pour indique comment on calcule les coordonnées des lignes à partir des variables. D'autres donnent les corrélations des variables avec les composantes principales. Gittins (p. 132-135) donnent les corrélations des deux ensembles de variables canoniques avec les deux variables de départ mais d'autres donnent les équations des variables canoniques. Ter Braak [42] a fait une analyse très précise de cette difficulté. En introduisant un compromis entre Biplot et reconstitution de matrices, il en arrive à une représentation simultanée des deux ensembles de variables. Le biplot de Ter Braak est une projection des deux ensembles de variables sur les variables canoniques d'un des deux tableaux alors que l'analyse canonique généralisée invite plutôt à la projection simultanée du type :



Projeter sur les variables canoniques d'un des deux espaces impliquent deux graphes possibles mélangeant l'interprétation d'une bloc interset et d'un bloc intraset de

corrélations. Projeter sur les scores canoniques impliquent un principe d'optimalité du type :

$$\frac{1}{p}\sum_{j=1}^{n} cor^2\left(\mathbf{x}^j,\mathbf{z}\right)+\frac{1}{q}\sum_{k=1}^{n} cor^2\left(\mathbf{y}^k,\mathbf{z}\right)$$

On est alors dans le domaine de l'analyse factorielle multiple (pondération par les inerties) et on ne peut admettre qu'une figure optimale pour une analyse soit utilisée pour une autre.

Faut-il alors abandonner l'idée d'avoir un mode de lecture unifié de l'AFC comme analyse canonique, de l'analyse canonique des correspondances et de l'analyse canonique classique ? Il est curieux que Ter Braak passant de l'analyse des correspondances [43] à l'analyse canonique des correspondances [18] puis revenant à l'analyse canonique classique [42] ne fasse pas la connexion alors que Gittins [35] la fait entre AC et AFC. Le biplot de Ter Braak transposé à l'AFC donne les espèces à la moyenne des relevés et le biplot dual donne les relevés à la moyenne des espèces. Ce qui prouve indiscutablement qu'on s'est éloigné de l'analyse canonique. La double représentation classique n'est pas référencée à une analyse discriminante mais à une double analyse d'inertie [44]. L'ACC sera alors plus pensée en termes d'ACPVI que d'analyse canonique en dépit de son nom d'origine.

Il faut donc revenir sur le caractère spécifique des variables qualitatives et en particulier de la variable nom d'espèce dans cette discussion. En effet, pourquoi en croisant deux variables qualitatives la situation semble plus favorable qu'en croisant deux paquets de variables quantitatives ? La différence essentielle porte sur les *intrasets correlations*. Purement artefactuelle pour une variable qualitative après centrage, rien n'est à perdre en enlevant la covariance artificiellement introduite. Pendant de nombreuses années, un programme d'AFC ne centrait pas les variables et éliminait la valeur propre unité artefactuelle qui en résultait. Comme dérivée de la régression, l'analyse canonique ne s'interprète aisément que pour des variables orthogonales.

On ne peut en même temps se débarrasser des corrélations intra *par principe* et en même temps s'en servir dans l'interprétation. Ceci s'exprime remarquablement en disant que l'analyse canonique donne un résultat équivalent par une transformation linéaire des données arbitraire inversible. Prendre les deux variables **x** et **y** ou les deux variables **x** + **y** et **x** - **y** revient strictement au même. Les indicatrices des classes avant centrages sont orthogonales. Toute projection sur l'espace engendré est interprétable. Le centrage et la décorrélation sont des avatars techniques qui ne concernent que le programmeur, l'utilisateur n'étant pas concerné.

6 — Conclusions

Une liste d'occurrences d'espèces introduit une variable qualitative. Le tableau d'indicatrices centrées qui en découle ne présente qu'une seule bonne propriété en termes d'analyse d'inertie : l'inertie totale est l'indice de diversité de Simpson. Son analyse n'a pas de sens, les covariances qui y résident étant purement artefactuelles. Une liste d'occurrences relève de la statistique multivariée dès qu'elle est face à une information externe.

Mise en face d'une autre variable qualitative (numéro de quadrat dans une partition de l'espace par exemple), trois procédures ont un sens. Deux sont des analyses sur variables instrumentales , la troisième est une analyse canonique des deux paquets d'indicatrices. Les deux premières sont des analyses des correspondances non symétriques (ANSC) explicitée dans [45]. La dernière est l'AFC du tableau croisé optimale comme double discriminante ou comme analyse canonique. Dans ce cas, le

dépouillement s'appuie sur une double projection du score canonique sur les deux espaces engendré et donne le *reciprocal scaling* [10].

Cette situation souligne combien l'AFC d'un tableau faunistique est particulière dans l'ensemble des méthodes d'ordination et combien sont diverses, pour un même calcul de base, les possibilités d'usage. Si on veut discriminer les quadrats par leur contenu floristique, l'ANSC sur profils quadrats s'impose : on aura alors une idée de la diversité des relevés. Si on veut discriminer les espèces par leur répartition on fera une ANSC sur profils espèces : on aura alors une idée de l'amplitude des espèces. Si on veut faire les deux d'un coup, avec ce que cela suppose de perte dans le compromis, on fera l'AFC. Ceci indique que l'AFC ne tient compte que des cases non vides d'un tableau et en ce sens ignore les absences.

La complexité de la situation créée par l'analyse des correspondances comme moyen d'ordination, sa spécificité comme double discriminante implicite ou explicite, ses alternatives unilatérales ont été très sous-estimées. L'introduction des données sous forme de listes d'occurrences ont donc d'abord un effet considérable dans le champ théorique. Une analyse des correspondances symétriques ou non introduit sans conteste les données comme des listes d'événements réalisés. Le reste est éliminé (facteurs limitants en particulier).

Mise en face d'un tableau de variables quantitatives (milieu) trois stratégies sont à nouveau concevables, deux analyses non symétriques et une analyse canonique. Les trois analyses sont différentes et recouvrent trois objectifs distincts.

Le premier cherche à discriminer les niches des espèces. Pour faire cela, on peut décorréler les variables ou non. On a donc deux méthodes, l'une de type analyse discriminante des variables de milieu sur la variable nom d'espèce, la seconde est une analyse inter-classe (espèces) sans décorrélation. Dans les deux cas, la liste d'occurrences est une *variable instrumentale* et l'objectif est une typologie de taxa. Les deux se rejoignent si les variables de milieu sont sans corrélation.

Le second cherche l'ordination des relevés par les espèces qui est la plus prédictible par les variables de milieu. Si on veut faire les deux d'un coup, avec ce que cela suppose de perte dans le compromis, on fera une analyse canonique (troisième objectif) qui est exactement pour la partie centrale du calcul une analyse canonique des correspondances. Cette situation souligne à nouveau combien l'ACC d'un tableau floro-faunistique est particulière dans l'ensemble des méthodes d'ordination croisée. L'ACC a donc une extension nouvelle et sa définition actuelle doit être revisitée. Standard logiciel par CANOCO, l'ACC n'est qu'un cas possible, particulier comme analyse canonique.

Le caractère particulier des propriétés du sous-espace engendré par une variable qualitative oblige à conserver la dichotomie traditionnelle entre analyse canonique et analyse discriminante. La diversité des aides à l'interprétation d'une AFC est étendue à des cas encore jamais étudiés. L'analyse canonique comme compromis entre deux analyses sur variables instrumentales doit introduire une séparation plus nette entre les deux stratégies.

L'introduction d'une liste d'occurrences est donc un élément de perturbation dans l'ensemble des méthodes d'ordination. Elle replace AFC et ACC à une place inattendue, génère des variantes inconnues et relance le débat sur l'analyse canonique. Les modules OccurData et Canonical vont expliciter ces remarques.

Références

[1] Light, R.J. & Margolin, B.H. (1971) An analysis of variance for categorical data. *Journal of the American Statistical Association* : 66, 534-544.

[2] Lebart, L. (1969) Analyse statistique de la contiguïté. *Publication de l'Institut de Statistiques de l'Université de Paris* : 28, 81-112.

[3] Lande, R. (1996) Statistics and partioning of species diversity, and similarity among multiple communities. *Oikos* : 76, 5-13.

[4] Cazes, P., Chessel, D. & Doledec, S. (1988) L'analyse des correspondances internes d'un tableau partitionné : son usage en hydrobiologie. *Revue de Statistique Appliquée* : 36, 39-54.

[5] Chessel, D., Lebreton, J.D. & Yoccoz, N. (1987) Propriétés de l'analyse canonique des correspondances. Une utilisation en hydrobiologie. *Revue de Statistique Appliquée* : 35, 55-72.

[6] 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.

[7] Rao, C.R. (1964) The use and interpretation of principal component analysis in applied research. *Sankhya*, A : 26, 329-359.

[8] Rao, C.R. (1962) User of discriminant and applied function in multivariate analysis. *Sankhya* : 22, 317-338.

[9] Estève, J. (1978) Les méthodes d'ordination : éléments pour une discussion. In : *Biométrie et Ecologie*. Legay, J.M. & Tomassone, R. (Eds.) Société Française de Biométrie, Paris. 223-250.

[10] Thioulouse, J. & Chessel, D. (1992) A method for reciprocal scaling of species tolerance and sample diversity. *Ecology* : 73, 670-680.

[11] Whittaker, R.H. (1967) Gradient analysis of vegetation. *Biological Reviews* : 42, 207-264.

[12] Silverman, B.W. (1986) *Density estimation for statistics and data analysis*. Chapman and Hall, London. 1-175.

[13] Rice, J.C. (1993) Forecasting abundance from habitat measures using nonparametric density estimation methods. Canadian J*ournal of Fisheries and Aquatic Sciences* : 50, 1690-1698.

[14] Statistical Sciences. (1995a) S-*PLUS, User's manual*, Version 3.3 for Windows. StatSci, a division of MathSoft, Seattle. 1-470.

Statistical Sciences. (1995b) *S-PLUS, Programmer's manual*, Version 3.2. StatSci, a division of MathSoft, Seattle. 1-425.

Statistical Sciences. (1995c) *S-PLUS Guide to Statistical and Mathematical analysis*, Version 3.3. StatSci, a division of MathSoft, Seattle. 1-650.

[15] Lebreton, J.D., Chessel, D., Prodon, R. & Yoccoz, N. (1988) L'analyse des relations espèces-milieu par l'analyse canonique des correspondances. I. Variables de milieu quantitatives. *Acta Œcologica, Œcologia Generalis* : 9, 1, 53-67.

[16] Green, R.H. (1971) A multivariate statistical approach to the Hutchinsonian niche: bivalve Molluscs of Central Canada. *Ecology* : 52, 543-556.

[17] Lebreton, J.D. & Yoccoz, N. (1987) Multivariate analysis of bird count data. *Acta Œcologica, Œcologia Generalis* : 8, 2, 125-144.

[18] Ter Braak, C.J.F. (1986) Canonical correspondence analysis : a new eigenvector technique for multivariate direct gradient analysis. Ecology : 69, 69-77.

[19] 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. In : *Applied Multivariate Analysis in SAR and Environmental Studies*. Devillers, J. & Karcher, W. (Eds.) Kluwer Academic Publishers. 85-114.

[20] Ter Braak, C.J.F. (1987c) CANOCO - a FORTRAN program for Canonical commnity ordination by [partial][detrended][canonical] correspondence analysis and redundancy analysis. Software documentation. Version 2.1, TNO Institute of Applied Computer Science, Wageningen.

[21] Wollenberg, A.L. (1977) Redundancy analysis, an alternative for canonical analysis. *Psychometrika* : 42, 2, 207-219.

[22] Tucker, L.R. . (1958) An inter-battery method of factor analysis. *Psychometrika* : 23, 2, 111-136.

[23] Dolédec, S. & Chessel, D. (1994) Co-inertia analysis: an alternative method for studying species-environment relationships. *Freshwater Biology* : 31, 277-294.

[24] Gittins, R. (1985) *Canonical analysis, a review with applications in ecology*. Springer-Verlag, Berlin. 1-351.

[25] Pearson, K. (1901) On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* : 2, 559-572.

[26] Hotelling, H. (1936) Relations between two sets of variates. *Biometrika* : 28, 321-377.

[27] Hotelling, H. (1933) Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* : 24, 417-441 , 498-520.

[28] Heiser, W.J. (1987) Joint ordination of species ans sites: the unfolding technique. In : *Developments in numerical ecology*. Legendre, L. & Legendre, P. (Eds.) Springer-Verlag, Berlin, Ecological Sciences, Vol. 14. 189-221.

[29] Oksanen, J. (1987) Problems of joint display of species and site scores in correspondence analysis. *Vegetatio* : 72, 51-57.

[30] Rencher, A.C. (1988) On the use of correlations to interpret canonical functions. *Biometrika* : 75, 363-365.

[31] Takeuchi, K., Yanai, H. & Mukherjee, B.N. (1982) *The foundations of multivariate analysis. A unified approach by means of projection onto linear subspaces.* John Wiley and Sons, New York. 1-458.

[32] 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.

[33] Kenkel, N.C. & Orloci, L. (1986) Applying metric and nonmetric multidimensionnal scaling to ecological studies : some new results. *Ecology* : 67, 919-928.

[34] Carrol, J.D. (1968) A generalization of canonical correlation analysis to three or more sets of variables. *Proceeding of the 76th Convention of the American Psychological Association* : 3, 227-228.

[35] Gittins, R. (1979) Ecological applications of canonical analysis. In : *Multivariate methods in ecological work*. Orloci, L., Rao, C.R. & Stiteler, W.M. (Eds.) Statistical Ecology Series. Vol. 7, International co-operative publishing house, Burtonsville. 309-335.

[36] Gittins, R. (1985) *Canonical analysis, a review with applications in ecology.* Springer-Verlag, Berlin. 1-351.

[37] Pialot, D., Chessel, D. & Auda, Y. (1984) Description de milieu et analyse factorielle des correspondances multiples. *Compte rendu hebdomadaire des séances de l'Académie des sciences. Paris, D* : 298, Série III, 11, 309-314.

[38] Carrel, G., Barthelemy, D., Auda, Y. & Chessel, D. (1986) Approche graphique de l'analyse en composantes principales normée : utilisation en hydrobiologie. *Acta Œcologica, Œcologia Generalis* : 7, 2, 189-203.

[39] Tenenhaus, M. & Young, F.W. (1985) An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis ans other methods for quantifying categorical multivariate data. *Psychometrika* : 50, 1, 91-119.

[40] Saporta, G. (1975) *Liaisons entre plusieurs ensembles de variables et codage de données qualitatives*. Thèse de 3° cycle, Université Pierre et Marie Curie, Paris VI. 1-102.

[41] Hill, M.O. & Smith, A.J.E. (1976) Principal component analysis of taxonomic data with multi-state discrete characters. *Taxon* : 25, 249-255.

[42] Ter Braak, C.J.F. (1990) Interpreting canonical correlation analysis through biplots of structure correlations and weights. *Psychometrika* : 55, 519-531.

[43] Ter Braak, C.J.F. (1985) Correspondence analysis of incidence and abundance data : properties in terms of a unimodal reponse model. *Biometrics* : 41, 859-873.

[44] Benzecri, J.P. & Coll. (1973) *L'analyse des données. II L'analyse des correspondances*. Bordas, Paris. 1-620.

[45] Gimaret-Carpentier, C., Chessel, D. & Pascal, J.P. (1998) Non-symmetric correspondence analysis: an alternative for community analysis with species occurrences data. *Plant Ecology* : in press.