Année 1996

MÉMOIRE

Présenté devant

l'**Université Claude Bernard - Lyon 1**

pour l'obtention de

L'HABILITATION À DIRIGER DES RECHERCHES

par

JEAN THIOULOUSE

OUTILS LOGICIELS, MÉTHODES STATISTIQUES ET IMPLICATIONS BIOLOGIQUES: UNE APPROCHE DE LA BIOMÉTRIE

Présenté le 25 juin 1996

Jury :	Daniel CHESSEL Yves ESCOUFIER Jean-Dominique LEBRETON Alain PAVÉ Cajo J.F. TER BRAAK Richard TOMASSONE Paul TREHEN
Rapporteurs :	Yves ESCOUFIER Jean-Dominique LEBRETON Domitien DEBOUZIE

UMR CNRS 5558 "Biométrie, Génétique et Biologie des Populations"
Université Claude Bernard - Lyon 1

Outils logiciels, méthodes statistiques et implications biologiques : une approche de la biométrie

Plan du mémoire

Introduction

Dix années d'expérience professionnelle dans le domaine de la biométrie m'ont permis d'analyser plusieurs aspects de cette discipline. Ce mémoire est destiné à rendre compte d'une part des résultats obtenus au cours de ces dix années, et d'autre part de quelques réflexions sur l'articulation de la biométrie avec les sciences biologiques. Il est organisé autour des trois principaux axes constitutifs de la biométrie: la **méthodologie statistique**, l'**outil informatique**, et les **applications biologiques**. Cette organisation n'est pas originale (c'est par exemple celle de Legay 1986), et elle présente l'inconvénient de découper en trois parties une activité par ailleurs homogène, constituant dans le travail quotidien un tout indissociable (même si ponctuellement l'un des axes peut prendre une place dominante). Elle a cependant le mérite de souligner que des résultats scientifiques originaux peuvent être obtenus dans chacun des trois domaines. Il est d'ailleurs intéressant de noter que, bien que ces trois axes soient effectivement représentés dans la plupart des articles que j'ai publiés depuis la soutenance de ma thèse de doctorat, chacun d'eux a pourtant une dominante à caractère statistique, informatique, ou biologique.

Le premier chapitre de ce mémoire est consacré à la méthodologie statistique. L'analyse multivariée y tient une place importante, avec trois méthodes d'intérêt très général: l'**analyse triadique** (Thioulouse & Chessel 1987), qui doit en fait être appelée analyse triadique **partielle** (Kroonenberg 1989), et qui a connu un succès relativement important, le **reciprocal scaling** (Thioulouse & Chessel 1992), terme dont la traduction en français (mise à l'échelle réciproque) n'a que peu d'intérêt puisqu'il fait référence aux dénominations anglo-saxonnes de deux autres méthodes (reciprocal averaging, Hill 1973 et dual scaling, Nishisato 1980), et enfin les **analyses locales** et **globales** (Thioulouse, Chessel & Champely 1995) qui permettent d'introduire une relation de voisinage entre échantillons dans l'étude des tableaux de données structurés dans l'espace ou le temps. Un autre sujet abordé dans ce chapitre est celui du calcul de la précision d'un échantillon systématique au moyen de la théorie des variables régionalisées.

Le second chapitre traite des outils informatiques, et particulièrement des logiciels statistiques (Thioulouse 1989, Thioulouse 1990, Thioulouse *et al.* 1995), mais aussi des **réseaux informatiques** (Thioulouse & Chevenet, soumis), de la question du développement **d'interfaces utilisateur** conviviales et **portables**, de l'importance des outils graphiques en analyse de données (Thioulouse *et al.* 1991, Devillers *et al.* 1991, Devillers *et al.* 1993), et des graphiques **interactifs**.

Le troisième chapitre aborde les résultats biologiques obtenus, dans le domaine de la biologie des populations (Debouzie & Thioulouse 1986, Thioulouse 1987), de l'agronomie (Cadet & Thioulouse 1989, Cadet *et al.* 1994), de l'écologie (Thioulouse & Chessel 1992), de l'écotoxicologie (Devillers *et al.* 1991, Devillers *et al.* 1993), et de la biologie moléculaire (Thioulouse & Lobry 1995, Perrière & Thioulouse 1996).

Chapitre 1
Méthodes statistiques

Les développements statistiques qui sont présentés dans ce chapitre ont pour la plupart été réalisés en collaboration avec Daniel Chessel. Ils concernent des domaines assez divers : ceux qui se sont traduits par des articles dont nous sommes co-auteurs relèvent de l'analyse multivariée (multitableaux, reciprocal scaling, structures spatiales). J'ai aussi poursuivi de façon un peu annexe un axe de recherche sur les variables régionalisées, dans le cadre du calcul de la précision des échantillons systématiques. Enfin, les réflexions sur les méthodes graphiques proviennent de nombreuses discussions que nous avons eues, en particulier avec Yves Auda.

1. Les analyses multitableaux

Les analyses multitableaux, c'est à dire les analyses portant sur une série de tableaux de données, connaissent depuis quelques années un intérêt croissant (Coppi & Di Ciaccio 1994), ainsi qu'en témoigne le numéro spécial de la revue **Computational Statistics and Data Analysis** qui leur est consacré (**18**, 1). Cet intérêt provient aussi bien de la part des statisticiens pour les développements théoriques auxquels donnent lieu ces méthodes, que de la part des biologistes, en particulier écologues, pour leur bonne adéquation au problème des études intégrant un facteur chronologique.

Dans ce domaine aussi, l'école française se distingue des approches anglo-saxonnes, principalement représentées par les modèles PARAFAC, INDSCAL, IDIOSCAL (Coppi 1994) qui datent des années 70, et par les modèles de Tucker (TUCKALS, Kroonenberg 1994), plus anciens (années 60). Ces modèles dérivent du positionnement multidimensionnel et donc des méthodes aux moindres carrés alternés. Les méthodes françaises sont bien sûr la méthode STATIS (ou ACT, Lavit *et al.* 1994) et l'analyse factorielle multiple (AFMULT, Escofier & Pagès 1994). On peut soupçonner que le succès des méthodes anglo-saxonnes, comparé à la très faible diffusion des méthodes françaises, est largement dû au fait que les aspects logiciels n'ont pas été totalement négligés comme ils l'ont été en France. Il est par exemple significatif de trouver dans la bibliographie classique des références à des modes d'emplois de logiciels sous la forme "Unpublished manuscript (Bell Laboratories, Murray Hill, 1969)", ce qui a longtemps été inimaginable dans la littérature française. Ce mode de fonctionnement (la référence à une procédure informatique, par opposition à l'énoncé d'un théorème), si il peut être critiqué d'un point de vue théorique, a cependant le mérite énorme d'informer l'utilisateur sur la disponibilité effective de la méthode, et donc sur l'utilisation qu'il pourra éventuellement en faire.

De ce point de vue, l'analyse triadique partielle (Thioulouse & Chessel 1987) présente l'avantage de ne nécessiter qu'un programme d'ACP classique pour être mise en oeuvre, ce qui lui a peut-être valu d'être citée par Potvin & Travis (1993) dans le numéro spécial sur les méthodes statistiques de la revue **Ecology** (**74**, 6). Cette méthode dérive de l'analyse triadique de Jaffrenou (1978), et on peut considérer qu'elle correspond à la méthode STATIS portant sur les tableaux de données au lieu des opérateurs. Nous n'en donnerons ici qu'un court descriptif, reprenant l'annexe statistique de l'article original (Thioulouse & Chessel 1987).

Soient $\mathbf{X}_{k\ /\ k=1,\ ...,\ t}$ t tableaux portant sur les mêmes individus et les mêmes variables. Soient n l'effectif d'individus et p celui des variables. Les tableaux sont normés (chaque variable y a une moyenne nulle et une variance unité). On appelle E l'espace euclidien R^p muni de la métrique des poids uniformes, de matrice:

$$\mathbf{D}_p=(1/p)\mathbf{I}_p$$

dans la base canonique (e_i). Soit E^* le dual de E. De même F est l'espace euclidien R^n muni de la métrique des poids, de matrice:

$$\mathbf{D}_n=(1/n)\mathbf{I}_n$$

dans la base canonique (f_j). L'espace des tableaux $L(E^*,F)$, canoniquement isomorphe à E^* F, est muni de la base des vecteurs e_i^* f_j , qui permet de réécrire un tableau comme un vecteur ligne par ligne sur une seule colonne. Appelons **Z** le tableau à np lignes et t colonnes constitué par la juxtaposition des t tableaux $\mathbf{X}_{k\ /\ k=1,\ ...,\ t}$ ainsi réécrits. $L(E^*,F)$ est alors muni du produit scalaire Tr, noté $\circ$, qui vaut pour un couple de tableaux étudiés $\mathbf{X}_k$ et $\mathbf{X}_l$:

$$(\mathbf{X}_k \circ \mathbf{X}_l) = \mathrm{Tr} (\mathbf{D}_p\ {}^t\mathbf{X}_k\ \mathbf{D}_n\ \mathbf{X}_l)$$

où Tr désigne la trace de l'endomorphisme et ${}^t\mathbf{X}_k$ la transposée de $\mathbf{X}_k$. Dans la base considérée la matrice de ce produit scalaire est $\mathbf{D}_p \otimes \mathbf{D}_n$ (produit de Kronecker, voir Glaçon 1981, p.8), soit tout simplement la matrice diagonale:

$$\mathbf{D}_{np} = (1/np)\ \mathbf{I}_{np}$$

Le produit scalaire des tableaux $\mathbf{X}_k$ et $\mathbf{X}_l$ vaut la moyenne des coefficients de corrélation entre les valeurs d'une même variable prises dans chacun des deux tableaux. L'ACP dans l'espace des tableaux (interstructure) muni de la pondération uniforme I_t et de la métrique Tr se trouve donc être par rapport à la base canonique de $G=R^t$ et la base ci-dessus, *exactement* l'ACP de **Z**, qu'il est inutile de recentrer, par rapport aux métriques I_t et D_{np}, soit l'ACP ordinaire de **Z**.

En utilisant les théorèmes généraux du schéma de dualité (cf. par exemple Caillez 1984, chapitre 3) on vérifie que cette ACP donne comme premier facteur un vecteur $A'=(a_1, a_2, ... ,a_t)$ qui, sous la contrainte:

$$\sum a_k^2 = 1$$

maximise:

$$\| ZA \|^2_{D_p \otimes D_n} = ZA \circ ZA = \mathrm{Tr} (D_p(\sum a_k X_k)\ D_n(\sum a_k X_k))$$

soit, au facteur 1/p près, l'inertie associée à l'ACP du tableau $\sum a_k \mathbf{X}_k$ (compromis), ou encore la somme des variances de ce tableau compromis.

Dans ce point de vue l'interstructure appelle d'elle même l'ACP du compromis comme optimale. Ces remarques justifient la procédure employée dont l'accessibilité ne demande qu'un programme d'ACP classique.

2. Le reciprocal scaling

Le principe théorique du reciprocal scaling (Thioulouse & Chessel 1992) repose sur une propriété de l'analyse des correspondances soulignée par Lebart *et al.* 1977. Cette propriété s'énonce simplement de la façon suivante : l'analyse des correspondances d'une table de contingence est équivalente à la double analyse discriminante des deux variables qualitatives qui définissent cette table de contingence. Le figure 1.1 montre que la table de contingence peut être ré-écrite sous la forme de deux tableaux comportant en lignes les correspondances (i.e., les cases non vides) et en colonnes les modalités des variables qualitatives (espèces et relevés).

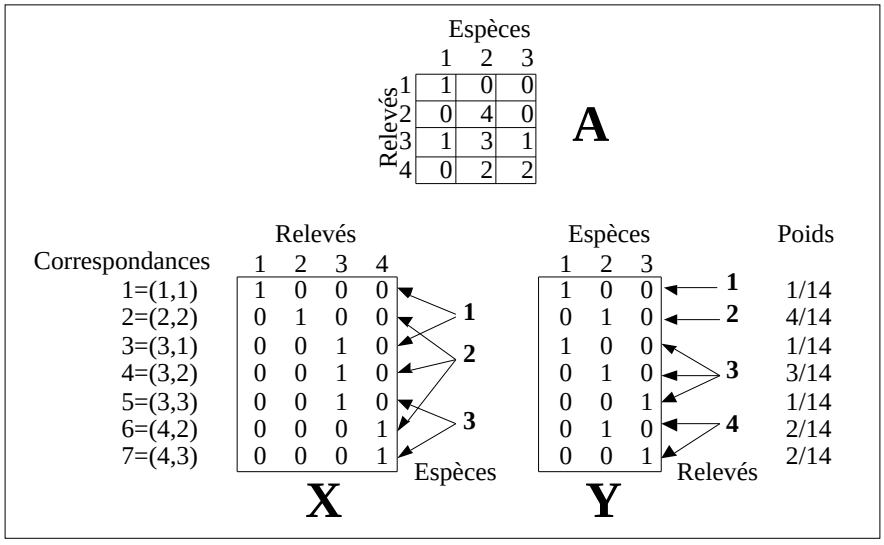


Figure 1.1: Schéma de principe de la double analyse discriminante de Lebart et al. *1984 utilisée dans le reciprocal scaling*

La double analyse discriminante de ces deux tableaux munis d'une pondération lignes définie par les effectifs de la table de contingence (figure 1.1) fournit des codes espèces et des codes relevés qui sont identiques à ceux de l'AFC. Mais elle fournit également des codes des correspondances, qui sont ceux utilisés dans le reciprocal scaling. Ces codes peuvent être calculés simplement à partir des codes espèces $C_k(j)$ et des codes relevés $L_k(i)$ sur l'axe k de la façon suivante :

$$H_k(i,j) = \left(\frac{L_k(i) + C_k(j)}{\sqrt{2\lambda_k \mu_k}} \right)$$

avec : $\mu_k = 1+\sqrt{\lambda_k}$

Il est possible de calculer les moyennes et les variances de ces codes pour chaque relevé (ainsi que pour chaque espèce) :

$$m_k(i) = \frac{1}{a_{i.}} \sum_{j=1}^{t} a_{ij} H_k(i,j)$$

$$s_k^2(i) = \frac{1}{a_{i.}} \sum_{j=1}^{t} a_{ij} H_k^2(i,j) - m_k^2(i)$$

On peut également calculer la covariance entre les deux codes k et l pour le relevé i :

$$c_{kl}(i) = \frac{1}{a_{i.}} \sum_{j=1}^{t} a_{ij} H_k(i,j) H_l(i,j) - m_k(i) m_l(i)$$

Ces quantités sont plus facilement obtenues à partir des codes relevés :

$$m_k(i) = \frac{\sqrt{\mu_k}}{\sqrt{2\lambda_k}} L_k(i)$$

$$s_k^2(i) = \frac{1}{\sqrt{2\lambda_k \mu_k}} \left[\frac{1}{a_{i.}} \sum_{j=1}^{t} a_{ij} C_k^2(j) - \lambda_k L_k^2(i) \right]$$

$$c_{kl}(i) = \frac{1}{2\sqrt{\lambda_k \lambda_l}\ \sqrt{\mu_k \mu_l}} \left[\frac{1}{a_{i.}} \sum_{j=1}^{t} a_{ij} C_k(j) C_l(j) - \sqrt{\lambda_k \lambda_l}\ L_k(i) L_l(i) \right]$$

Elles peuvent être utilisées pour tracer dans le plan des axes k et l une ellipse pour chaque relevé, dont le centre est positionné par les moyennes $m_k(i)$ et $m_l(i)$, les amplitudes horizontales et verticales étant données par les variances $s_k^2(i)$ et $s_l^2(i)$, la covariance $c_{kl}(i)$ donnant la pente de l'axe principal. De la même façon, on peut tracer une ellipse pour chaque espèce, et les graphiques obtenus sont superposables.

Ces représentations permettent de comparer les variabilités intra et inter-relevés (diversité spécifique) et intra et inter-espèces (amplitude d'habitat et séparation des niches). Elles fournissent des figures explicites dont l'interprétation en termes écologiques est souvent édifiante (cf. chapitre 3).

3. Les structures spatiales

L'analyse des structures spatiales et spatio-temporelles était le principal objectif de ma thèse de troisième cycle en 1985, avec comme matériel d'étude un insecte ravageur du colza. J'avais principalement fait appel à l'époque aux tests non paramétriques et à l'analyse des correspondances simple (Debouzie & Thioulouse 1986, Thioulouse 1987), avec de nombreuses représentations graphiques, en particulier cartographiques.

L'ouvrage de synthèse de Upton & Fingleton (1985) donne une bonne revue des méthodes utilisées dans ce domaine. Un certain nombre de progrès ont eu lieu depuis, en particulier grâce aux travaux de Lebart (Banet & Lebart 1984) et de Wartenberg (1985a, 1985b). De plus, l'étude des structures spatiales a connu récemment un succès grandissant dans le domaine écologique, et plusieurs articles méthodologiques leur ont été consacrés (Borcard *et al.* 1992, Legendre 1993, Borcard & Legendre 1994).

Le formalisme du schéma de dualité (Escoufier 1987), joint à l'utilisation des relations de voisinage nous a permis de proposer une approche globale des méthodes d'analyse multivariée prenant en compte les structures spatiales. Cette présentation est basée sur un article publié dans **Ecological and Environmental Statistics** (Thioulouse, Chessel & Champely, 1995).

L'idée la plus simple pour introduire une composante spatiale dans une méthode d'analyse multivariée est d'utiliser les coordonnées en X et en Y des points d'échantillonnage sur une carte géographique. Ce point de vue avait été introduit en écologie par Gittins (1968) et par Lee (1969, 1981) dans le domaine géologique. La "trend surface analysis" consiste à utiliser des polynômes à deux variables (X et Y) de degré fixé a priori. Il est ainsi possible de coupler un tableau écologique avec le tableau des coefficients du polynôme, par l'analyse canonique (Gittins 1968), l'analyse canonique des correspondances ou l'analyse des redondances (Borcard *et al.* 1992). Wartenberg (1985a) se limite à un polynôme de degré deux (cinq paramètres), tandis que Borcard *et al.* emploient un polynôme de degré trois mais dont ils sélectionnent les termes par une régression pas à pas.

Il existe donc un problème intrinsèque dans cette démarche : elle demande d'ajuster un modèle spatial, et les analyses qui en seront déduites intégreront les contraintes de ce modèle. Les modèles polynomiaux étant particulièrement peu efficaces dans ce domaine, des tentatives avec d'autres types de modèles spatiaux (splines) auraient pu être faites. L'utilisation d'une relation de voisinage présente de multiples avantages, mais l'inconvénient de forcer une réduction de l'information spatiale (les distances entre points sont perdues, au profit d'une simple relation binaire). L'importance de cet inconvénient doit être jugée par rapport à ceux des autres méthodes. Les avantages sont multiples, à la fois du point de vue théorique et du point de vue de l'application des méthodes qui en sont tirées.

Si l'espace échantillonné est considéré comme homogène (figure 1.1A), la relation de voisinage est facile à établir : elle est déduite de la triangulation de Delaunay de l'ensemble des points d'échantillonnage (Green & Sibson 1977, Sibson 1980, exemples d'utilisation dans Upton & Fingleton 1985, ou Pigliucci & Barbujani 1991). Par définition, deux points sont voisins si leurs polygones de Voronoï ont au moins un côté en commun. Sur le plan pratique, on peut utiliser le logiciel XYZ GeoBench de P. Schorn (Schorn 1991, ftp://ftp.inf.ethz.ch/pub/xyz/).

Dans le cas des études portant sur un réseau hydrographique (figure 1.1B), utiliser les coordonnées en X et en Y des stations n'aurait pas de sens. C'est la relation amont-aval qui est importante, ainsi que l'appartenance au cours d'eau. Il est donc parfaitement logique d'utiliser la relation de voisinage suivante: chaque point est voisin des deux points situés immédiatement en amont et en aval sur le même cours d'eau.

L'utilisation d'une relation de voisinage peut aussi être une bonne façon de prendre en compte un obstacle séparant des sites géographiquement proches (figure 1.1C). Dans le cas de la figure 1.1C, seuls les sites situés du même côté de l'obstacle seront pris comme étant voisins.

Afin de présenter les méthodes d'analyse locales et globales basées sur une relation de voisinage entre individus, il convient d'abord de définir les notions de variance locale et de variabilité globale.

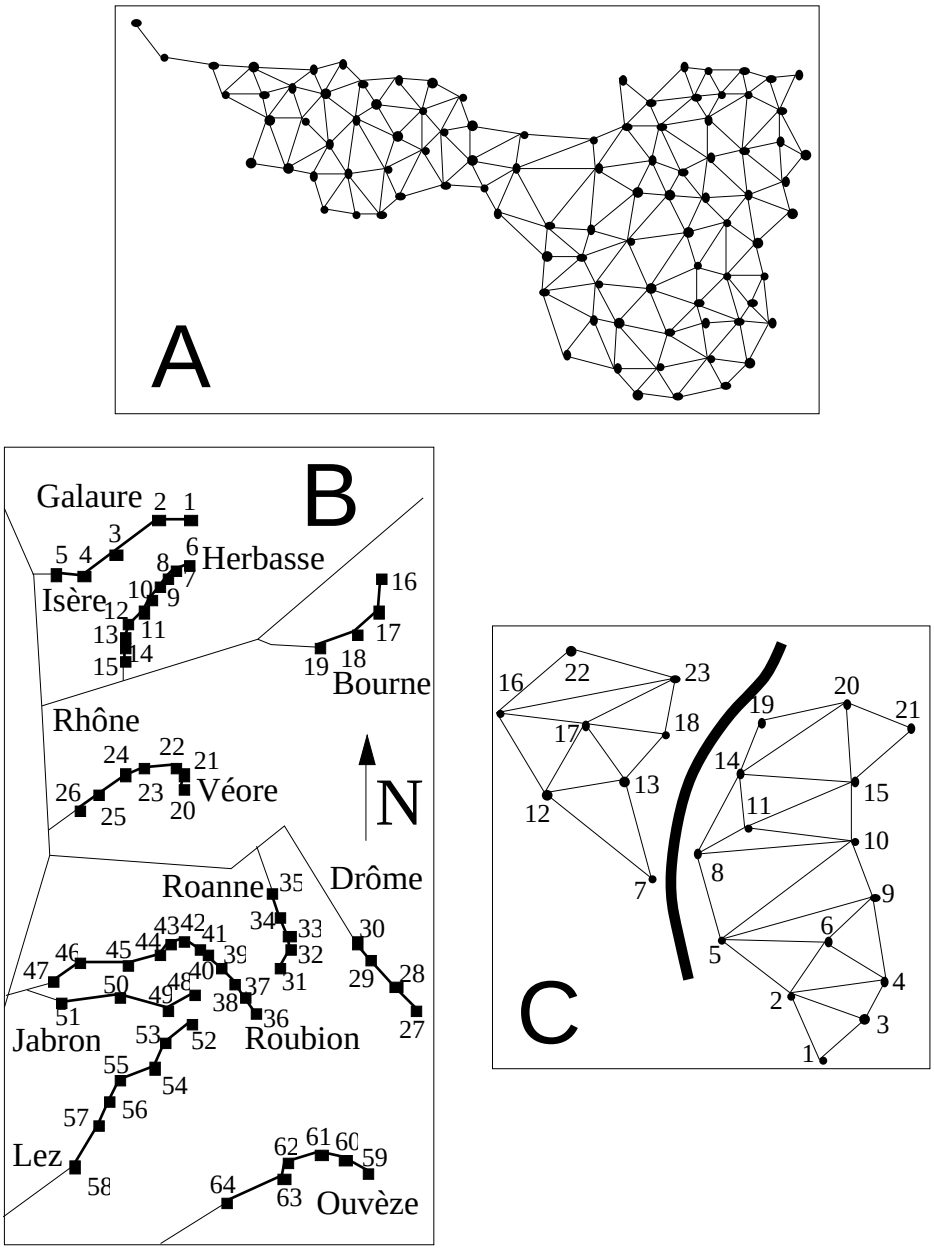


Figure 1.2: Exemples d'utilisation d'une relation de voisinage

3.1. Variance totale, variance locale et variabilité globale

3.1.1. *Notations*

$\mathbf{X} = \left[x_{ij} \right]$ est la matrice à *n* lignes et *p* colonnes contenant le tableau de données (*p* variables mesurées en *n* sites d'échantillonnage). $\mathbf{X}^t$ est le transposé de **X**.

$\mathbf{x} = \left[x_i \right]$ est un vecteur de composantes x_i (un vecteur colonne de **X**).

$\mathbf{M} = \left[m_{ij} \right]$ est la matrice de la relation de voisinage (n lignes et n colonnes) : si le site i est voisin du site j, alors $m_{ij} = 1$, sinon $m_{ij} = 0$. De plus, pour tout i, $m_{ii} = 0$.

La matrice $\mathbf{P} = \left[p_{ij} \right]$ est simplement déduite de **M** par : $p_{ij} = \frac{1}{2m} m_{ij}$, m étant le nombre total de paires de voisins, et donc $\sum_{ij} p_{ij} = 1$.

$\mathbf{D} = \text{Diag}(p_1, p_2, ..., p_n)$ est la matrice diagonale des "poids de voisinage" : $p_i = \frac{1}{2m} \sum_j m_{ij}$.

3.1.2. *Définitions*

La moyenne de la variable **x** étant donnés les poids **D** est égale à :

$$\overline{x_{\mathbf{D}}} = \sum_i p_i x_i = \mathbf{x}^{\mathbf{t}} \mathbf{D} \mathbf{1}_n .$$

Sa variance, que nous appellerons ici variance totale est:

$$Var(\mathbf{x}) = \sum_i p_i \left(x_i - \overline{x_{\mathbf{D}}} \right)^2 .$$

Si **x** est **D**-centrée, cette variance totale peut s'écrire sous forme matricielle :

$$Var(\mathbf{x}) = \mathbf{x}^{\mathrm{t}} \mathbf{D} \mathbf{x} . \qquad (1)$$

La variance locale (Banet & Lebart 1984) est tout simplement la variance entre points voisins:

$$LV(\mathbf{x}) = \sum_i \sum_j p_{ij} \left(x_i - x_j \right)^2 ,$$

et elle peut s'écrire matriciellement :

$$LV(\mathbf{x}) = \mathbf{x}^{\mathrm{t}} (\mathbf{D} - \mathbf{P}) \mathbf{x} = \mathbf{x}^{\mathrm{t}} \mathbf{D} (\mathbf{I}_n - \mathbf{D}^{-1} \mathbf{P}) \mathbf{x} . \qquad (2)$$

La variabilité globale, ou autocovariance spatiale, est définie par :

$$GV(\mathbf{x}) = \sum_i \sum_j p_{ij} \left(x_i - \overline{x_D} \right) \left(x_j - \overline{x_D} \right) ,$$

ce qui, si **x** est **D**-centrée peut s'écrire:

$$GV(\mathbf{x}) = \mathbf{x}^{\mathrm{t}} \mathbf{P} \mathbf{x} = \mathbf{x}^{\mathrm{t}} \mathbf{D} (\mathbf{D}^{-1} \mathbf{P}) \mathbf{x} . \qquad (3)$$

Cette quantité n'étant pas toujours positive, nous ne l'appellerons pas variance globale.

La seconde forme dans l'équation (3) montre que la variabilité globale représente la covariance entre les valeurs observées en un point et la moyenne des valeurs voisines de ce point, tandis que l'équation (2) montre que la variance locale peut être considérée comme la covariance entre les valeurs observées en un point et la différence entre ces valeurs et la moyenne des valeurs voisines. Ces notions correspondent donc tout à fait à l'intuition assimilant variabilité globale élevée = phénomène lisse, variance locale élevée = phénomène agité.

Des équations (1), (2), et (3), nous pouvons déduire une décomposition de la variance sous la forme:

$$Var(\mathbf{x}) = LV(\mathbf{x}) + GV(\mathbf{x}),$$

c'est-à-dire une décomposition de la variance totale en composantes locales et globales vis à vis de la relation de voisinage considérée.

3.1.3. *Relation avec les indices usuels*

Lorsque les poids de voisinage sont uniformes (comme par exemple dans les relations de voisinage circulaire, ou de voisinage linéaire si on excepte les deux extrémités), le rapport de la variance locale à la variance totale $LV(\mathbf{x})/Var(\mathbf{x})$ est égal, à un facteur $(n-1)/n$ près, au coefficient d'autocorrélation de Geary, dont est déduit l'indice de Geary (Geary 1954, Cliff & Ord 1973).

De même, on peut noter que l'indice de Moran (Moran 1948, Cliff & Ord 1973, Ripley 1981) est égal, sous la même hypothèse, au rapport de la variance globale à la variance totale $GV(\mathbf{x})/Var(\mathbf{x})$.

3.2. Analyses totales, locales et globales

Les notions de variance totale, variance locale et de variabilité globale ayant été présentées, il convient maintenant de définir les analyses qui maximisent ces quantités. Le schéma de dualité (Escoufier 1987) est bien sûr le cadre choisi pour ces définitions.

3.2.1. *Analyse générale d'un triplet*

Un triplet statistique (**X**, **C**, **R**) est constitué de trois matrices: la matrice des données **X**, à n lignes et p colonnes, sur laquelle on peut avoir effectué une transformation préalable (centrage, normalisation, etc.), la matrice des poids des lignes (**R**), et la matrice de la métrique utilisée pour mesurer les distances entre les lignes (**C**). De façon duale, **C** peut aussi être considérée comme la matrice des poids des colonnes et **R** comme la matrice de la métrique utilisée pour mesurer les distances entre colonnes. L'analyse d'un triplet est alors basée sur la diagonalisation de la matrice $\mathbf{X}^t\mathbf{RXC}$.

Quand **C** n'est pas proportionnelle à l'identité, cette matrice n'est pas symétrique, mais l'équation aux valeurs propres peut être écrite sous la forme

$$\mathbf{C}^{1/2}\mathbf{X}^t\mathbf{RXC}^{1/2}\mathbf{u}_\alpha = \lambda_\alpha\mathbf{u}_\alpha,$$

ce qui conduit à la diagonalisation d'une matrice qui elle, est symétrique. Les vecteurs propres $\mathbf{u}_\alpha$ et $\mathbf{u}_\beta$, correspondant aux valeurs propres λ_α et λ_β, vérifient

$\mathbf{u}_\alpha^t \mathbf{u}_\beta = \delta_{\alpha\beta}$. Les axes principaux $\mathbf{a}_\alpha$ et les coordonnées lignes $\mathbf{c}_\alpha$ sont respectivement égaux à :

$$\mathbf{a}_\alpha = \mathbf{C}^{-1/2}\mathbf{u}_\alpha \text{ et } \mathbf{c}_\alpha = \frac{1}{\sqrt{\lambda_\alpha}}\mathbf{X}\mathbf{C}^{1/2}\mathbf{u}_\alpha .$$

3.2.2. *Analyses en composantes principales totale, locale et globale*

L'ACP usuelle est donc l'analyse du triplet ($\mathbf{X}_c$, $\mathbf{I}_p$, $\frac{1}{n}\mathbf{I}_n$) où $\mathbf{I}_p$ est la matrice identité d'ordre *p*, $\mathbf{I}_n$ la matrice identité d'ordre *n*, et $\mathbf{X}_c$ est le tableau de données (*n* x *p*) centré ou normé. Les coordonnées lignes de cette analyse maximisent la variance usuelle.

L'analyse totale est simplement l'analyse du triplet ($\mathbf{X_D}$, $\mathbf{I}_p$, $\mathbf{D}$) où $\mathbf{X_D}$ est le tableau de données **D**-centré (or **D**-normé), **D** étant toujours la matrice diagonale des poids de voisinage. Les coordonnées lignes de cette analyse maximisent la variance totale.

L'analyse locale est l'analyse du triplet ($\mathbf{X_D}$, $\mathbf{I}_p$, $\mathbf{D} - \mathbf{P}$), et les coordonnées lignes de cette analyse maximisent la variance locale (Le Foll 1982).

L'analyse globale est l'analyse du triplet ($\mathbf{X_D}$, $\mathbf{I}_p$, $\mathbf{P}$), et les coordonnées lignes de cette analyse maximisent la variance globale. Elle est proche de l'analyse de Wartenberg (Multivariate spatial correlation analysis, Wartenberg 1985b), qui est celle du triplet ($\mathbf{X}$, $\mathbf{I}_p$, $\mathbf{M}$), mais elle en diffère par l'utilisation de $\mathbf{P} = (1/2m)\mathbf{M}$ et du **D**-centrage.

Les valeurs propres de l'analyse globale ne sont pas toujours positives (elles ne représentent en effet pas des variances mais des autocovariances spatiales). Ceci ne pose pas de problème d'interprétation : une autocovariance spatiale négative signifie simplement qu'une valeur positive élevée en un point est entourée de valeurs négatives élevées aux points voisins.

L'introduction du **D**-centrage permet donc d'obtenir trois analyses comparables, dont les opérateurs sont liés, et qui fournissent une décomposition de la variance totale :

$$\mathbf{X_D^t D X_D} = \mathbf{X_D^t (D - P) X_D} + \mathbf{X_D^t P X_D} \qquad (4)$$

De plus, ces analyses peuvent être généralisées facilement, en particulier au cas de l'analyse des correspondances.

3.2.3. *Analyses des correspondances totale, locale et globale*

La généralisation des analyses précédentes à l'analyse des correspondances se heurtent au fait qu'en AFC les pondérations lignes sont imposées. Il convient donc de définir une AFC modifiée, utilisant les poids de voisinage. L'AFC usuelle est définie par Escoufier (1982) de la façon suivante. Soit $x..$ la somme totale de la matrice **X**, qui est donc ici une table de contingences floro-faunistique :

$$x.. = \sum_i \sum_j x_{ij} ,$$

et soit $\mathbf{F} = \left[f_{ij} \right]$ la matrice des fréquences:

$$f_{ij} = \frac{x_{ij}}{x..}$$

$$\sum_i \sum_j f_{ij} = 1.$$

$\mathbf{F}_n = \text{Diag}\left[f_{i.} \right]$ et $\mathbf{F}_p = \text{Diag}\left[f_{.j} \right]$ sont les matrices diagonales contenant les fréquences marginales :

$$f_{i.} = \sum_j f_{ij}$$

$$f_{.j} = \sum_i f_{ij} .$$

Avec ces notations, l'AFC usuelle est l'analyse du triplet ($\mathbf{F}_n^{-1}\mathbf{F}\mathbf{F}_p^{-1} - \mathbf{1}_{np}$, $\mathbf{F}_p$, $\mathbf{F}_n$). $\mathbf{1}_{np}$ est la matrice unité de dimensions (n x p). Le terme général a_{ij} de la matrice $\mathbf{A} = \mathbf{F}_n^{-1}\mathbf{F}\mathbf{F}_p^{-1} - \mathbf{1}_{np}$ est égal à :

$$a_{ij} = \frac{f_{ij}}{f_{i.} f_{.j}} - 1.$$

$\mathbf{F}_w = \text{Diag}\left[w_{.j} \right]$ est la matrice diagonale contenant la pondération calculée à l'aide des poids de voisinage :

$$w_{.j} = \sum_i p_i \frac{f_{ij}}{f_{i.}} .$$

L'AFC totale est l'analyse du triplet ($\mathbf{F}_n^{-1}\mathbf{F}\mathbf{F}_w^{-1} - \mathbf{1}_{np}$, $\mathbf{F}_w$, $\mathbf{D}$), l'AFC locale est l'analyse du triplet ($\mathbf{F}_n^{-1}\mathbf{F}\mathbf{F}_w^{-1} - \mathbf{1}_{np}$, $\mathbf{F}_w$, $\mathbf{D} - \mathbf{P}$), et l'AFC globale est l'analyse du triplet ($\mathbf{F}_n^{-1}\mathbf{F}\mathbf{F}_w^{-1} - \mathbf{1}_{np}$, $\mathbf{F}_w$, $\mathbf{P}$). Ces analyses fournissent des codes espèces $\mathbf{F}_w$-normés, qui, par averaging fournissent des codes relevés qui maximisent les variances totale et locale, et la variabilité globale.

3.3. Discussion

Les résultats que nous avons obtenus avec ces analyses sur des données simulées (structures spatiales imposées) et sur des jeux de données réels montrent qu'elles atteignent efficacement leur objectif.

On peut distinguer deux situations extrêmes. Dans le cas d'un gradient pur, on doit observer l'égalité des résultats des trois types d'analyse (structures totales = structures globales = structures locales). L'interprétation se fait alors en termes d'ordination des relevés. Dans une partition pure, les structures locales doivent être différentes des structures totales et globales. Dans la pratique, toutes les situations intermédiaires

peuvent être rencontrées, avec par exemple des structures locales imbriquées dans des structures globales. L'analyse totale peut alors, pour une raison d'échelles, être incapable de révéler l'un ou l'autre type de structure, alors que les analyses locales et globales en seront capables.

Les analyses locales et globales fournissent des codes qui ont comme propriété de maximiser la variance locale et la variance globale. Pour profiter au mieux de cette propriété il faut utiliser des représentations graphiques adaptées à son expression. Le plan factoriel est loin d'être optimal de ce point de vue : s'agissant d'une propriété spatiale, des représentations spatiales sont nécessaires. Trois types de représentations spatiales peuvent être réalisées dans le logiciel ADE-4 : les courbes de niveaux et les polygones jointifs avec niveaux de gris, et les cartes par cercles et carrés (cf. chapitre 2, § 1.3).

Les courbes de niveaux sont calculées par une régression lowess. Elles sont particulièrement bien adaptées à la mise en évidence de structures globales. L'utilisation de cette régression soulève toutefois le problème de l'estimation du nombre de voisins pris en compte dans la régression. Une procédure basée sur la minimisation de l'erreur d'estimation en fonction du nombre de voisins permet de disposer d'un critère objectif et stable. Les structures locales sont par contre mal rendues par les courbes de niveaux. Les cartes par cercles et carrés permettent de bien visualiser les structures locales, tout en montrant clairement les structures globales, et constituent donc un bon compromis. Les cartes par polygones jointifs sont les moins efficaces pour visualiser une autocorrélation spatiale, mais elles sont efficaces pour les structures locales.

La relation de voisinage est introduite dans les analyses grâce à la pondération par le nombre de voisins, et par la matrice du graphe de voisinage. La pondération par le nombre de voisins et le **D**-centrage correspondant permettent d'accorder une importance plus grande aux points ayant de nombreux voisins, et donc d'atténuer celle des points marginaux. Elle permet de plus d'unifier le formalisme mathématique des diverses analyses avec celui de plusieurs travaux antérieurs : les indices de Geary et de Moran, ainsi que les propositions de Lebart 1969, Le Foll 1982, Banet & Lebart 1984, Wartenberg 1985b.

Un autre avantage lié à l'utilisation d'une relation de voisinage est la possibilité de construire des modèles de répartition spatiale. En effet, si on utilise des polynômes classiques, les coefficients ne sont pas indépendants, ce qui pose des problèmes pour leur interprétation écologique. Borcard & Legendre (1994, p. 59) proposent dans ce cas d'utiliser des polynômes orthogonaux : "The terms of the spatial polynomials originally proposed by Legendre (1990) are not independent of one another. If the interpretation of the regression or canonical coefficients relating these terms to the community structure is of special interest, orthogonal polynomials should be used instead of the classical polynomials". Les opérateurs de lissage $\mathbf{D}^{-1}\mathbf{P}$ et $\mathbf{I}_n - \mathbf{D}^{-1}\mathbf{P}$ (équation 3) fournissent une alternative intéressante: leurs vecteurs propres (qui sont identiques) définissent une base **D**-orthonormale sur laquelle une projection permet d'obtenir une décomposition du phénomène spatial.

La figure 1.3 montre la représentation des six premiers vecteurs propres d'un opérateur de lissage d'ordre 90. On constate que leur répartition spatiale est lisse, allant d'un gradient pur (premier vecteur propre) à un structure à six extrema (sixième vecteur propre). Le graphique 1.2D montre que l'erreur de lissage augmente avec l'ordre du vecteur propre (1 à 6) et qu'elle croit très rapidement avec le nombre de voisins utilisés pour la régression locale dans le cas des vecteurs propres d'ordre supérieur ou égal à trois.

Un autre problème soulevé par Borcard & Legendre (1994, p. 60) est celui de la modélisation de la partie non spatiale d'un phénomène. En utilisant les derniers vecteurs

propres de l'opérateur de lissage, qui sont fortement auto-corrélés négativement, on obtient de bons modèles de structures locales, pour lesquels il serait impossible d'utiliser des polynômes.

Enfin, la généralisation de l'introduction d'une relation de voisinage dans d'autres méthodes est possible. Elle est déjà réalisée pour l'analyse des correspondances multiples et l'analyse de co-inertie, et elle devrait l'être bientôt pour les autres méthodes de couplage de tableaux (analyses sur variables instrumentales et régression PLS).

4. Analyse de données et représentations graphiques

La graphique a depuis longtemps représenté pour nous une **interface** entre les **théorèmes d'analyse de données** et **l'interprétation en termes biologiques** des résultats des analyses (donc entre statisticien et biologiste) ce qui lui donne le statut intrinsèque **d'outil biométrique**.

Ce point de vue avait été initié par Yves Auda dans le début des années 80 et nous avons poursuivi dans cette direction. Lors de mon arrivée au laboratoire de biométrie en DEA en 1981, le seul périphérique graphique était une table traçante Tektronix au format A3 et bien peu de personnes savaient comment l'utiliser. C'est sans doute une consultation statistique avec Y. Bouchery (INRA Colmar) en 1982 qui m'a démontré tout l'intérêt des représentations graphiques en analyse de données : il nous avait montré une carte géographique, tracée à la main, représentant une parcelle cultivée où les relevés (systématiques) étaient symbolisés par des cercles de diamètre proportionnel à l'infestation du ravageur. La programmation de la table traçante et l'écriture des premiers programmes graphiques en Basic avec l'aide d'Yves Auda me parurent une tache légère en comparaison de la réalisation de cette carte.

L'arrivée des premières consoles graphiques (consoles Secapa en 1983) nous permit ensuite d'obtenir des vitesses de tracé plus compatibles avec une démarche exploratoire et interactive. L'écriture du programme Graphique par Yves Auda marqua une étape décisive. Il fut utilisé pendant plusieurs années, et les graphiques de l'article de Thioulouse et Chessel (1987) en sont issus, grâce à une configuration qui paraît aujourd'hui anecdotique, mais qui témoigne de la permanence de nos efforts dans ce domaine. Ils furent en effet réalisés en utilisant un Macintosh comme terminal virtuel sur l'Eclipse S/140, avec le logiciel VersaTerm en mode émulation Tektronix, donc compatible avec le programme Graphique.

4.1. Correspondances graphiques - théorèmes

L'évolution des moyens informatiques nous a depuis permis d'améliorer à la fois les conditions d'utilisation et les performances des applications graphiques. L'objectif est par contre resté le même : offrir aux biologistes une traduction dans un langage universel des propriétés mathématiques utilisées en analyse de données. Voici quelques exemples qui illustrent cette position et mettent en parallèle des graphiques et les propriétés mathématiques correspondantes. Cette liste n'est bien sûr pas exhaustive, elle présente simplement quelques cas classiques.

Le **cercle de corrélation** traduit la propriété de **maximisation de l'inertie projetée** sur les axes principaux de l'ACP normée. La représentation triangulaire et les biplots reposent sur la même propriété, ainsi que la projection de variables supplémentaires en ACP normée.

Le **graphe canonique en ACP normée**, qui consiste à représenter les variables initiales (normées) en fonction des axes principaux traduit la **maximisation de la somme des carrés des covariances** variables/facteurs.

Les **cartes factorielles** simples ou munies de cercles et de carrés proportionnels aux données, et les **représentations fonctionnelles**, c'est à dire la représentation des valeurs des coordonnées factorielles en fonction d'une structure externe (par exemple le plan d'échantillonnage dans l'espace ou le temps) donnent en général des figures explicites car elles sont basées sur le fait que l'analyse maximise la **variance** des coordonnées factorielles.

Les représentations par **ellipses**, par **courbes de Gauss**, par **enveloppes convexes**, par **étoiles**, par **collections d'histogrammes** permettent de représenter la maximisation d'une **variance inter-classes**, par exemple en analyse discriminante ou en analyse inter/intra.

Le même genre de graphique (par **ellipses** en particulier) permet de représenter la **maximisation des moyennes conditionnelles** des codes des relevés et des espèces en AFC.

De la même façon, la maximisation d'un **rapport de corrélation** (ou d'une somme ou d'une moyenne de rapports de corrélation) en analyse discriminante ou en ACM peut être représentée par des **ellipses**, des **étoiles**, ou des **enveloppes convexes**.

La maximisation du **coefficient de corrélation canonique** en AFC peut être représentée par un graphique de type tableau où les lignes et les colonnes de la table de contingence sont **réordonnées en fonction de leurs coordonnées factorielles**.

La maximisation de la **somme des variances conditionnelles** en AFC dans la méthode du reciprocal scaling peut être représentée par des **ellipses** ou des **étoiles**.

La minimisation de la **somme des carrés des écarts** en régression (polynomiale ou Lowess par exemple) peut se faire en superposant les valeurs observées (points) et la prédiction par le modèle (courbe) reliées par des traits.

La maximisation de la **variance locale ou globale** peut être représentée en traçant les coordonnées factorielles sur le **graphe de voisinage** avec des cercles et des carrés.

La maximisation de la **covariance locale ou globale** peut être représentée en traçant le **graphe de voisinage sur le plan factoriel** de l'analyse locale.

En plus de ces éléments fondamentaux, certaines considérations pratiques doivent être étudiées (quoi tracer, où et comment). Plusieurs discussions avec Daniel Chessel et Yves Auda ont fait émerger les réflexions suivantes.

4.2. Implantation et composante G, protocole et mesure

Après plusieurs tentatives infructueuses de généralisation ou de systématisation de la typologie courbes / cartes / tableaux d'Auda 1983, nous avons tenté de distinguer dans un graphique les notions **d'implantation** et de **composante G** (la **composante graphique** de Bertin 1967).

L'implantation correspond à la position des éléments graphiques. Elle peut être à une ou plusieurs dimensions, mais le cas n>2 est en général trop complexe pour fournir des graphiques interprétables, sauf lorsque la composante G est très simple (histogrammes et surfaces 3D).

La **composante G** est l'élément graphique utilisé. Elle peut s'exprimer de diverses façons : ponctuelle, linéaire, surfacique, ou complexe, et ceci indépendamment de son implantation. Elle peut de plus faire appel à des variables de séparation (labels, symboles géométriques), de valeur (niveaux de gris, labels numériques, courbes de niveaux), de taille (cercles et carrés, bâtons verticaux), ou être complexe (boîtes à moustaches, intervalles de variation, étoiles). Les variables de séparation doivent traduire des relations de ressemblance/dissemblance, les variables de valeur des relations d'ordre, et les variables de taille des relations de proportionnalité. Il est possible de croiser les deux typologies pour avoir par exemple une variable de taille surfacique (cercles et carrés) ou linéaire (bâtons).

On peut aussi distinguer les notions de **protocole** et de **mesure**. Le **protocole** est un élément fixé a priori et externe, même s'il peut faire partie du jeu de données. Ce peut être le protocole expérimental, ou bien une notion implicite comme la structuration en lignes et en colonnes dans un tableau. La **mesure** est l'information résultant de l'expérience, ou bien un code numérique issu de calculs sur ces informations.

Dans un graphique, **l'implantation** et la **composante G** peuvent chacune provenir de la **mesure** ou du **protocole**. Prenons l'exemple d'un plan factoriel simple en AFC: les points sont positionnés sur le plan en fonction de leurs coordonnées factorielles, et une chaîne de caractères permet d'identifier l'élément correspondant (dans ce cas un relevé ou une espèce du tableau de données). L'implantation en X et en Y est liée à une mesure (coordonnée factorielle), et la composante G (variable de séparation, de type ponctuelle) est liée au protocole (identificateur du relevé ou de l'espèce). Si à la place des labels, les points sont représentés par des cercles et des carrés de taille proportionnelle à l'effectif d'une espèce (variable de taille de type surfacique), la composante G représente alors elle aussi une mesure, et non plus le protocole.

Inversement, l'implantation peut représenter le protocole : c'est souvent le cas en analyse de données spatiales, ou dans les représentations de tableaux. Dans un graphique représentant une carte géographique sur laquelle est symbolisée l'abondance d'une espèce (par des cercles de taille variable), l'implantation en X et en Y provient du protocole, alors que la composante G est une mesure. Dans la représentation d'un tableau, où chaque case d'un tableau est symbolisée par un cercle de taille variable, l'implantation provient du protocole et la composante G représente la mesure.

L'utilisation d'une relation de voisinage peut se traduire par une implantation correspondant à une mesure, et une composante G correspondant au protocole, mais on peut également se trouver dans la situation inverse. On peut par exemple tracer le graphe de voisinage sur le plan factoriel des relevés, en reliant d'un trait les points du plan factoriel reliés par la relation de voisinage (implantation = mesure, composante G = protocole), ou bien tracer, sur le graphe de voisinage représenté dans l'espace géographique, les valeurs des coordonnées factorielles des noeuds du graphe par un cercle ou un carré (implantation = protocole, composante G = mesure).

Les graphiques avec des ellipses, des étoiles, des enveloppes convexes correspondent également à une composante G de type complexe, qui traduit en général un protocole. Une ellipse résume cinq valeurs : deux moyennes, deux variances et une covariance. Elle représente le protocole car il s'agit des moyennes, variances, covariances des groupes qui sont issus du protocole (ellipses de dispersion en analyse inter/intra ou en analyse discriminante). Le fonctionnement est le même pour les enveloppes convexes (polygone entourant un nuage représentant une catégorie d'individus), et les étoiles (traits reliant les points d'un nuage à son centre de gravité).

Le cas des courbes est intéressant car il souligne le fait qu'une courbe peut exprimer des notions totalement différentes, ce qu'une typologie basée sur la forme des éléments graphique ne peut pas expliciter. L'implantation en Y traduit en général une mesure. L'implantation en X peut traduire le protocole (courbe de valeurs régulièrement

espacées dans le temps) ou une autre mesure (courbe de durées de développement d'un stade larvaire en fonction de la température). La composante G peut traduire le protocole (traits reliant les points successifs, labels placés à chaque point de la courbe), une mesure (points de taille variable en fonction d'une autre variable), ou les deux (intervalles de confiance de chaque point).

Les **modes d'assemblage** des graphiques élémentaires dans une collection sont la **superposition** et la **juxtaposition**. Ces modes d'assemblage proviennent généralement du protocole. Par exemple dans les analyses multitableaux, on peut tracer la collection des plans factoriels correspondants à l'analyse simple de chaque tableau. La juxtaposition des graphiques élémentaires est alors dictée par le protocole (répétition des tableaux de mesure au cours du temps). De même dans les analyses simples, on représente souvent les graphiques des différentes variables par une collection de graphiques élémentaires juxtaposés. Il existe aussi des modes d'assemblages complexes, comme par exemple dans le graphique qui consiste à représenter sur le plan factoriel la distribution des espèces présentes dans les relevés par des histogrammes. La composante G est alors complexe (puisqu'il s'agit d'un histogramme), liée à la fois au protocole (liste des espèces) et aux mesures (abondances). Aucun module graphique d'ADE-4 ne permet actuellement de réaliser automatiquement ce type de graphique, à l'exception du module ADEScatters, mais dans une présentation dynamique (l'utilisateur clique sur le relevé dont il veut connaître la distribution spécifique).

On voit donc bien qu'il serait illusoire de vouloir dresser une typologie des représentations graphiques. La graphique est bien un langage, pas un ensemble fini de modèles et en faire une typologie serait équivalent à vouloir faire une typologie de toutes les phrases possibles de la langue française : le résultat n'aurait ni utilité ni sens.

Il existe cependant des structures de données associées aux mesures et au protocole, que les logiciels peuvent mettre à profit. Par exemple, les fichiers de coordonnées factorielles ont une structure fixée a priori par le logiciel d'analyse multivariée utilisé. Si le logiciel graphique tient compte de cette structure, cela pourra faciliter (ou même rendre possible) la réalisation de certains graphiques. Dans SAS, les fichiers de coordonnées factorielles comportent en lignes tous les éléments du tableau de données (lignes et colonnes) et en colonnes les facteurs. Cette structure est différente de celle d'ADE-4 qui sépare dans deux fichiers les coordonnées factorielles des lignes et celles des colonnes. Les modules graphiques d'ADE-4 sont adaptés à cette structure et la mettent à profit pour permettre de réaliser facilement des plans factoriels et un grand nombre d'autres graphiques (mais il est quand même possible d'utiliser les fichiers SAS). Le module ADEScatters utilise simultanément les deux fichiers de coordonnées : celui des espèces pour faire un plan factoriel des espèces et, lorsque l'utilisateur clique sur l'une d'elles, celui des relevés pour tracer, dans une petite fenêtre temporaire, le plan factoriel des relevés avec la distribution de l'espèce sélectionnée (par cercles et carrés).

Enfin, l'introduction des notions de mesure et de protocole dans la structure des graphiques présente l'avantage de correspondre à la tendance actuelle qui va vers une amélioration de la prise en compte de la structure des jeux de données (et donc en particulier des protocoles expérimentaux) dans les méthodes d'analyse multivariée. La conception des modules graphiques d'ADE-4 a été directement influencée par ces considérations.

5. Variables régionalisées et échantillonnage systématique

J'ai commencé à m'intéresser à la théorie des variables régionalisées (VR) dès mon arrivée au laboratoire de biométrie (fin 1981), à la suite des travaux de M. El Bahi et de D. Chessel, et en particulier à l'occasion d'une consultation de statistique qui avait eu lieu avec B. Mathy en Février 1981. L'objectif principal était le calcul de la précision d'un échantillon systématique, appliqué dans ce cas à un problème cytologique : l'estimation du volume du disque alaire de *Bombyx mori*. Après diverses péripéties anecdotiques (dont la perte d'un manuscrit par l'éditeur de la revue **Mikroskopie**, et deux courriers qui n'arrivèrent jamais...), un article fut publié en 1985 (Thioulouse *et al.* 1985a). Il décrivait la façon d'utiliser la théorie des VR pour calculer la précision d'un échantillon systématique obtenu par la technique des coupes sériées, avec la possibilité de prévoir une certaine diminution de l'intensité d'échantillonnage tout en préservant une bonne précision de l'estimation. Ce travail était relativement novateur à l'époque, et il offrait un cadre théorique satisfaisant à des pratiques jusque là empiriques, ainsi qu'en témoigne le succès qu'a ensuite connu cette technique en stéréologie.

Les indications qui suivent résument très brièvement les principaux points de la méthode. Soit *f(x)* la valeur d'une VR définie sur un domaine **d** et mesurée au point x.

$$Q = \int_{\mathbf{d}} f(x)dx,\ f(x) = 0 \text{ si } x \quad \mathbf{d}$$

Un échantillon systématique est défini par sa maille t, par le nombre total de points échantillonnés m, et par l'abscisse du premier point x_1, choisie de façon aléatoire entre 0 et t. Les abscisses des points d'échantillonnage x_1, x_2, ..., x_m sont définies par $x_{j+1} = x_j + t$ pour (j=1, 2, ..., m-1). Pour un échantillon donné, un estimateur de la quantité totale Q est $[Q_t]$:

$$[Q_t] = t\sum_{j=1}^{m} f(x_j)$$

L'estimation de la variance de $[Q_t]$ fait appel au covariogramme g(h) :

$$g(h) = \int_{\mathbf{d}} f(x)f(x+h)dx$$

La variance de $[Q_t]$ est alors donnée par :

$$Var([Q_t]) = t\sum_{k=-\infty}^{+\infty} g(kt) - \int_{-\infty}^{+\infty} g(h)dh$$

Mais g(h) ne peut être calculé que pour les valeurs de h qui sont des multiples de la maille, et il faut donc utiliser un modèle du covariogramme. Les règles de correspondances permettent alors d'obtenir la variance de $[Q_t]$ en fonction des paramètres de ce modèle. Si on utilise un développement limité à droite et à l'ordre s de g(h) :

$$g_1(h) = \sum_{j=1}^{s} c_j h^j$$

la variance vaut :

$$Var([Q_t]) = \sum_{j=0}^{s} c_j t^{j+1} \left[-2\sin\left(j \frac{\pi}{2}\right) \frac{B_{j+1}}{j+1} \right]$$

Les B_j étant les nombres de Bernoulli. Dans l'article de Mikroskopie, deux modèles de covariogramme étaient utilisés : un modèle linéaire sur les deux premiers points du covariogramme g(0) et g(1), et un polynôme de degré trois sur les 5 premiers points en excluant g(0) pour prendre en compte un éventuel effet de pépite. Nous avons ainsi pu montrer qu'il était possible de calculer la précision des estimations du volume à partir de la série de coupes et que, dans le cas de l'exemple utilisé, on pouvait définir une maille de sous-échantillonnage optimale.

L'arrivée de François Houllier au laboratoire de biométrie nous a permis d'élargir le champ d'application de cette méthode à l'agronomie, pour la calcul de la précision des dénombrements d'insectes. François a en particulier décrit dans sa thèse de doctorat comment l'EPR (échantillonnage partiellement renouvelé, une technique classique en foresterie) pouvait être généralisé par la théorie des VR. Nous avons ainsi pu montrer dans une note aux Comptes Rendus de l'Académie des Sciences (Thioulouse *et al.* 1985b), comment calculer la précision de l'effectif d'aleurodes infestant un rang de tomates en serre pour diverses intensités d'échantillonnage. La présence d'une agrégativité vraie très élevée dans les dénombrements d'insectes nous a conduits à proposer un modèle de covariogramme tenant compte de l'effet de pépite (discontinuité à l'origine) induit par ce phénomène. Ce modèle est un modèle linéaire ajusté sur les 50 premiers points du covariogramme en excluant g(0), additionné d'une discontinuité à l'origine C_Δ. La règle de correspondance permet dans ce cas d'obtenir une expression de la variance très simple :

$$Var([Q_t]) = c_\Delta t - \frac{c_1}{6} t^2$$

qui présente l'avantage de fournir une décomposition en deux termes, l'un dû à l'agrégativité des insectes ($c_\Delta t$, variance intrinsèque), et l'autre étant la variance de régionalisation ($\frac{c_1}{6} t^2$), liée à la distribution des insectes le long du rang de tomates. L'amélioration de la précision par rapport à l'échantillonnage aléatoire (de l'ordre de 20%) est due à la prise en compte du second terme.

Une consultation statistique avec Jean-Pierre Royet en 1989 m'a ensuite conduit à écrire un troisième article consacré à la théorie des VR, dans lequel nous insistions sur les dangers d'utiliser des formules toutes prêtes en présence de covariogrammes présentant une régionalisation inconnue a priori. En effet, la présence d'un effet de pépite important invalide complètement la méthode utilisée classiquement dans la littérature (Gundersen & Jensen 1987). Cette méthode consiste à appliquer une formule simplifiée donnant directement la précision (ou plus exactement le coefficient d'erreur, $CE([Q_t])$) en fonction des trois premiers termes du covariogramme :

$$CE([Q_t]) = 1/S \sqrt{\frac{3A - 4B + C}{12}}$$

avec

$$CE([Q_t]) = \frac{\sqrt{Var([Q_t])}}{[Q_t]}$$

$$S = \sum_{j=1}^{m} f(x_j)$$

$$A = [g(0)]/t = \sum_{j=1}^{m} f(x_j)^2$$

$$B = [g(t)]/t = \sum_{j=1}^{m-1} f(x_j) f(x_j + t)$$

$$C = [g(2t)]/t = \sum_{j=1}^{m-2} f(x_j) f(x_j + 2t).$$

La présence d'un effet de pépite peut dans ces conditions passer inaperçue, et conduire à une sur-estimation importante de la précision (l'estimation étant alors bien moins bonne que ne le laisse prévoir le modèle simplifié). En utilisant six modèles différents et des données réelles ainsi que des données simulées, nous avons pu montrer que la méthode simplifiée est en fait généralement fiable pour les données histologiques, qui sont le plus souvent très lisses. En présence d'un effet de pépite, un modèle quadratique avec une discontinuité à l'origine, présenté lui aussi sous la forme d'une formule simplifiée, permet d'obtenir une bonne estimation de la précision.

$$CE([Q_t]) = 1/S\sqrt{A - \frac{31}{12}B + \frac{7}{3}C - \frac{3}{4}D}$$

avec

$$D = [g(3t)]/t = \sum_{j=1}^{m-3} f(x_j) f(x_j + 3t).$$

Chapitre 2
Outils logiciels

L'informatique a constitué, depuis le début de mon activité de recherche scientifique, une part importante de mon travail. Le développement de logiciels en biométrie trouve sa justification dans la simple constatation que l'innovation méthodologique ne passe dans le champ de l'utilisation en biologie qu'à travers la disponibilité de logiciels adaptés à son application dans les champs expérimentaux concernés (écologie, biologie des populations, agronomie, etc.). Par définition, l'innovation méthodologique ne peut être véhiculée par les logiciels standards, et sa diffusion, dont la lenteur a souvent été déplorée, implique donc la mise au point de logiciels spécialisés. Il y a là un véritable goulot d'étranglement dans le processus de fécondation réciproque entre disciplines. Mais la traduction des théorèmes en algorithmes et en logiciels ne peut se faire sans une compréhension (même minimale) des dits théorèmes.

Ce point de vue a toujours été une des caractéristiques de l'activité de l'URA 243, et il nécessite une "rentabilisation" de l'effort de développement des logiciels. En effet, le logiciel en lui-même n'est en général pas considéré comme un résultat scientifique, ce qui est une erreur car il contient l'information méthodologique et il est indispensable à sa mise en oeuvre. La multiplication des revues scientifiques consacrées à ce thème (Statistics and Computing, Computational Statistics, Computational Statistics and Data Analysis, Environmental Software, Journal of Statistical Software), ainsi que la part croissante consacrée à l'évaluation des logiciels dans les revues (par exemple les "Software reviews" dans Applied Statistics) montrent bien que, au moins dans le domaine de la statistique, cet état de fait a déjà bien évolué. Une autre voie de rentabilisation passe bien sûr par des collaborations avec des collègues biologistes, ces collaborations étant d'ailleurs un élément constitutif de l'activité biométrique.

L'évolution rapide du matériel informatique vers une puissance toujours accrue a eu des conséquences diverses sur notre discipline. Certaines sont positives: par exemple, la vitesse de calcul n'est pratiquement plus un facteur limitant sur les machines actuelles dans le domaine de l'analyse multivariée. Quelques minutes (cinq ou six, au lieu de plusieurs heures il y a 10 ans) suffisent pour extraire les valeurs propres et les vecteurs propres d'une matrice 500×500, ce qui de toutes façons représente une limite supérieure à la taille des jeux de données récoltés sur le terrain. De plus cette puissance a ouvert la voie aux tests de permutations, ce qui est une excellente chose, et a aussi permis une démultiplication des capacités graphiques, par exemple dans le domaine des graphiques dynamiques.

Par ailleurs, l'augmentation de la puissance des machines (vitesse, mais aussi capacité mémoire) a permis une évolution des logiciels vers une plus grande facilité d'utilisation, principalement grâce aux interfaces utilisateurs graphiques (IUG). Ceci a été vrai en ce qui concerne les système d'exploitation des ordinateurs, avec l'apparition du Macintosh et de son Finder, de Microsoft Windows, et de X-Windows sur les stations de travail Unix (avec Motif ou OpenLook par exemple). L'ancienne programmathèque d'analyse multivariée du laboratoire de biométrie a suivi cette évolution avec le portage des anciens programmes de l'Eclipse S/140 (Data General) sur Macintosh, l'élaboration de l'interface HyperCard, l'écriture des logiciels MacMul, GraphMu et MacDendro, puis plus récemment avec la réécriture en langage C dans ADE-4. Le principal inconvénient de cette évolution réside dans la non compatibilité entre les différentes IUG disponibles sur les trois principaux types d'ordinateurs actuels: stations de travail Unix, micro-ordinateurs compatibles PC, et Macintosh. Divers systèmes de développement croisés

existent (SUIT : simple user interface toolkit, Vibrant, ainsi que des solutions commerciales), mais en pratique ces systèmes ne fournissent pas les résultats attendus, ou imposent des limites trop contraignantes (pertes de fonctionnalités sur certaines plates-formes).

Une nouvelle voie semble cependant apparaître actuellement, avec la croissance exponentielle des capacités et des fonctionnalités des réseaux informatiques, qui permettra sans doute de résoudre rapidement à la fois le problème de la diffusion de l'innovation méthodologique vers les utilisateurs, et celui de la création d'IUG conviviales et portables.

1. Les logiciels d'analyse multivariée : historique et présentation d'ADE-4

La disparition des mini-ordinateurs dans la deuxième moitié des années 80 nous a contraints à amorcer le virage de la micro-informatique. Il était de plus nécessaire de maintenir une programmathèque d'analyse multivariée qui soit à la fois un **outil de recherche**, apte à faciliter la mise au point rapide de programmes de calcul pour tester les avancées théoriques, et un **outil de diffusion méthodologique,** destiné à être utilisé par un large public. Cette contrainte, additionnée au fait que l'ancienne programmathèque, qui fonctionnait sur un mini-ordinateur Eclipse S/140, était écrite en langage BASIC, nous a conduits, à partir de 1988, à choisir le langage Microsoft QuickBasic sur micro-ordinateurs Macintosh. Les performances exceptionnelles de ces micro-ordinateurs dans le domaine graphique, qui restent d'ailleurs inégalées même actuellement, nous ont de plus permis de développer des fonctionnalités graphiques avancées, en suivant la ligne du travail novateur d'Yves Auda (Auda, 1983). Ce choix de la micro-informatique n'était pas dénué d'implications : à la même époque apparaissaient en effet au laboratoire les stations de travail Sun. Mais notre volonté de développer un outil de diffusion méthodologique à l'usage des biologistes ne pouvait s'accorder avec ce type de matériel, de même qu'elle ne peut actuellement être compatible avec l'usage de logiciels spécialisés en statistique, comme SAS ou S.

Après quelques hésitations, nous avions choisi d'intégrer les différents programmes grâce à une interface basée sur le logiciel HyperCard d'Apple (les premières piles ADECO datent de 1989). Cette stratégie a été poursuivie pendant plusieurs années par Daniel Chessel, pour aboutir en 1994 à la version 3.7 du logiciel ADE. Tout en ayant participé activement à l'initiation de ces travaux, j'avais pour ma part choisi ensuite une voie un peu différente, préférant une intégration plus poussée et une interface utilisateur plus perfectionnée, ainsi que le langage FORTRAN. Ces choix aboutirent à la mise au point dès 1988 des logiciels MacMul et GraphMu (Thioulouse, 1989; Thioulouse 1990), puis de MacDendro, qui furent l'objet d'une large diffusion (plus de 200 exemplaires envoyés sur disquettes, et un nombre inconnu diffusé par le réseau Internet), et au sujet desquels je reçois encore plusieurs messages chaque mois.

Les développements originaux réalisés entre-temps par Daniel Chessel dans le domaine de l'analyse de données avaient cependant été à l'origine d'une inflation très importante du logiciel ADE, qui rendait problématique sa diffusion et son utilisation par des néophytes. De plus, le langage QuickBasic ayant été abandonné par Microsoft, de nombreux problèmes étaient rencontrés sur les nouveaux modèles de Macintosh.

Nous avons alors décidé en 1993 de regrouper à nouveau nos efforts et de repartir à zéro pour une nouvelle version du logiciel ADE, baptisée ADE-4 et entièrement réécrite en langage C (initialement avec le compilateur Think C de Symantec, puis avec le CodeWarrior de MetroWerks). J'avais auparavant envisagé d'autres solutions, en

particulier le langage C++ et la bibliothèque de classes MacApp d'Apple Computer pour la génération de l'IUG, ou bien le générateur d'interfaces SUIT (Conway *et al.*, 1992), basé sur la bibliothèque SRGP (Foley *et al.*, 1990) disponible sur Macintosh, stations de travail Sun Microsystems et compatibles IBM PC sous Windows. Mais la lourdeur de MacApp et le manque de support technique et de suivi à long terme de SUIT m'ont fait préférer une solution ad hoc.

L'objectif global était de restructurer le logiciel ADE en mettant à profit l'expérience acquise grâce à MacMul et GraphMu dans le domaine de l'intégration et de l'interface utilisateur. Ce travail, entrepris en Octobre 1993, a débouché sur la diffusion d'ADE-4, qui a commencé en Mars 1995. Il a commencé par l'écriture de plusieurs couches de sous-programmes (plus de 300 au total) destinés à:

- standardiser les fonctions de lecture-écriture des fichiers,
- assurer les fonctions de calcul simples (moyenne, variance, calcul matriciel, etc.),
- fournir des primitives de tracé graphique multifenêtré en coordonnées logiques,
- gérer complètement et de façon transparente l'interface utilisateur.

Les premiers modules d'analyses à un tableau (ACP, AFC, ACM) furent prêts début 1994, les principaux modules graphiques ont été écrits durant l'été 1994. Les méthodes plus sophistiquées furent introduites progressivement dans le courant de l'année 1994 par Daniel Chessel. Sylvain Dolédec et Jean-Michel Olivier (qui s'occupaient déjà des versions précédentes) ont pris en charge la rédaction et la mise en forme de la documentation (deux volumes en anglais, plus quatre volumes en français) ainsi que la diffusion des disquettes et des manuels.

L'ensemble de ce travail a été soutenu financièrement par le Programme Environnement du CNRS dans le cadre du contrat "Méthodes, modèles et théories" (projet "Méthodes linéaires et graphiques pour l'analyse des données environnementales").

Comme les versions précédentes d'ADE, ADE-4 est constitué de plusieurs modules (au nombre de 41 actuellement). Des différences fondamentales sont cependant à relever, en plus du gain en vitesse d'exécution apporté par le langage C, par l'optimisation des compilateurs, et par les nouveaux modèles de Macintosh (gain d'au moins deux ordres de grandeur).

J'ai toujours accordé une place importante à l'architecture matérielle des ordinateurs : la programmathèque BASIC de l'Eclipse S/140 était basée sur des procédures de calcul matriciel écrites en langage assembleur. La gamme des modèles de Macintosh évoluant très rapidement, il était nécessaire de profiter de cette évolution, qui a été encore accéléré par l'introduction de machines basées sur une nouvelle série de micro-processeur (les PowerPC 601, 603 et 604). La volonté de profiter des performances de ces nouveaux processeurs (gain d'un facteur 10) m'a alors conduit à proposer trois types de modules dans ADE-4 : des modules de base, compilés pour les micro-processeurs de la famille 68000 (68000, 68020, 68030, 68040), des modules permettant d'utiliser les coprocesseurs arithmétiques 68881/68882 et la partie spécialisée dans le calcul en virgule flottante du processeur 68040, et enfin des modules compilés en code PowerPC. Nous offrons ainsi aux utilisateurs un choix qui leur permet d'exploiter au mieux la puissance des machines dont ils disposent.

De plus, afin de permettre aux utilisateurs de MacMul et GraphMu de profiter de ces avantages, j'ai rassemblé dans deux modules spéciaux, appelés MacMul 4 et GraphMu 5 les principales fonctionnalités correspondant aux anciens programmes MacMul et GraphMu.

Un effort important a été consacré à la **portabilité** du code source. En effet, toute la partie calculatoire des analyses est entièrement écrite en langage C ANSI, et elle est donc très facilement portable. J'ai d'ailleurs mis à profit cette portabilité pour réaliser le

système NetMul dont il sera question plus loin. Cette portabilité est accrue par l'existence des couches de sous-programmes de base qui isolent la partie calcul de l'implémentation des entrées-sorties, toujours problématique. Le problème de l'IUG (qui, elle, n'est pas portable) a été minimisé par l'écriture d'un ensemble de sous-programmes qui sont automatiquement liés au programme principal et qui assurent l'homogénéité de l'interface utilisateur tout en épargnant complètement au programmeur le souci de l'implémenter dans les modules qu'il écrit.

Les paramètres entrés par l'utilisateur à travers l'IUG sont rendus disponibles au programmeur par l'intermédiaire de variables globales. Il est ainsi très simple de remplacer l'IUG Macintosh des modules d'ADE-4 par une interface en mode ligne (cette opération a été réalisée sur des stations de travail Sun Microsystems sous Unix). Le portage des modules sur micro-ordinateurs compatibles IBM PC serait envisageable de cette façon. Il avait par ailleurs commencé, à l'aide du générateur d'interfaces SUIT, mais il me paraît plus intéressant actuellement de poursuivre la voie ouverte avec NetMul, qui offre des possibilités beaucoup plus larges, par exemple la compatibilité Mac, Sun, PC, et même Newton, le PDA (personal digital assistant, ou assistant numérique personnel) d'Apple.

Les modules d'ADE-4 sont **indépendants** et possèdent chacun leur propre interface utilisateur graphique. L'interface avec les piles HyperCard a été conservée, mais elle est devenue facultative. L'utilisateur peut ainsi ne posséder et n'utiliser **que les modules qui lui sont nécessaires**, ce qui permet d'éviter le problème de la taille des anciennes versions (plus de 10 Mo pour ADE 3.7). L'interface des programmes est simple et elle est uniforme dans tous les modules, ce qui permet à l'utilisateur de se repérer plus facilement. Elle respecte les recommandations d'Apple (Apple, 1987) et plus généralement les règles de base des interfaces utilisateurs (cf par exemple Coutaz, 1990). Les principaux progrès dans ce domaine ont été :
- l'intégration dans l'environnement du Finder Macintosh,
- les dialogues non modaux (*i.e.*, non bloquants),
- la gestion d'un dossier de travail de façon souple,
- le filtrage des fichiers présentés à l'utilisateur lors d'un choix,
- la possibilité d'effectuer les calculs en arrière-plan avec une visualisation de l'avancement des calculs (particulièrement utile lors des tests de permutations, qui peuvent être relativement longs),
- la présentation du rapport d'exécution des calculs dans une fenêtre texte facilement manipulable,
- la technique du "glisser-déposer" (drag and drop) dans certains modules.

Pour les modules graphiques, nous avons utilisé une approche totalement interactive, dans laquelle l'utilisateur peut, grâce à trois fenêtres non modales, modifier un grand nombre de paramètres graphiques (une trentaine) et visualiser simultanément le résultat à l'écran. Ces modules graphiques représentent un progrès important par rapport aux modules d'ADE 3.7, et aussi par rapport à GraphMu dont l'interface était presque entièrement modale. Un module de graphique dynamique (ADESctatters) a aussi été introduit, dont les fonctionnalités seront détaillées plus loin.

Afin de rendre compte de l'évolution du logiciel ADE-4, il est intéressant de présenter les principales fonctions des modules qui le composent actuellement. Cette présentation reprend en partie et précise certains points d'un article soumis à la revue **Statistics and Computing**.

1.1. L'interface utilisateur

Les modules d'ADE-4 se répartissent en deux catégories: les modules de calcul et les modules graphiques, avec une interface utilisateur un peu différente.

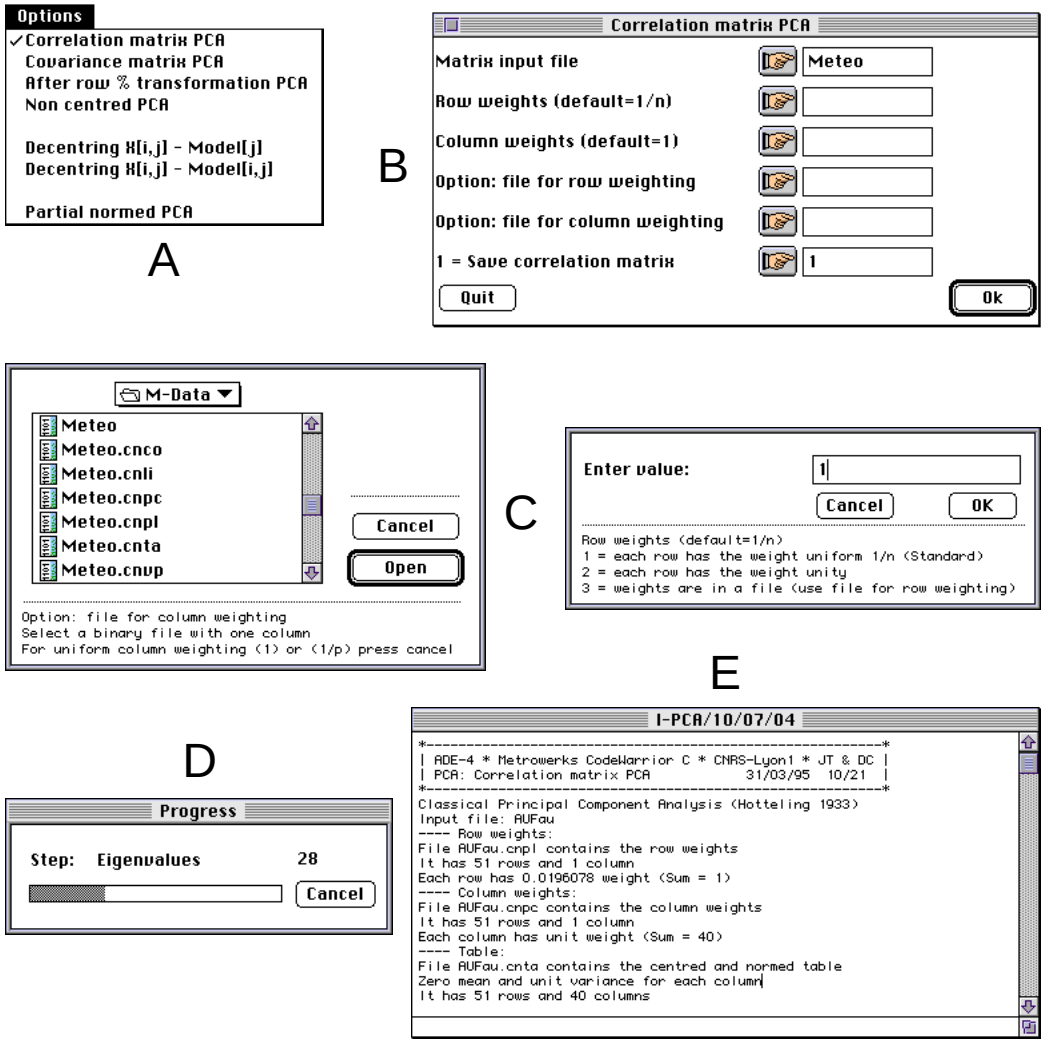


Figure 2.1: Éléments de l'interface utilisateur des modules de calcul d'ADE-4

La figure 2.1 montre les principaux éléments de l'interface utilisateur des modules de calcul. Le menu Options (A) permet de choisir une des méthodes proposées dans le module (par exemple le type d'ACP). La fenêtre de dialogue principale (B) permet de choisir les paramètres de la méthode : fichiers d'entrée, options de pondérations, etc. Son contenu est bien sûr variable en fonction de l'option sélectionnée dans le menu Options. Le bouton Ok permet de lancer l'exécution des calculs, alors que le bouton Quit permet de quitter le module. Lorsque l'utilisateur clique sur un des boutons ornés de l'icone d'une main, des fenêtres de dialogue spéciales apparaissent pour faciliter la sélection d'un fichier, ou l'entrée d'un paramètre numérique (C). Dans ces fenêtres, un texte explicatif guide l'utilisateur sur les choix à effectuer. Lors des opérations longues (diagonalisation, tests de permutations) une fenêtre affiche la progression des calculs (D). A la fin des calculs, un rapport est affiché dans une fenêtre dont le contenu peut être copié dans le presse-papier ou enregistré dans un fichier (E).

Les modules graphiques (figure 2.2) ont aussi un menu Options (A), permettant de choisir entre les divers types de graphiques disponibles. Ils disposent de plus d'un menu

Fenêtres (A), qui donne accès à quatre fenêtres : la fenêtre graphique, dans laquelle sont tracés les graphiques (re-dimensionnables dynamiquement à la souris), et trois fenêtres de paramètres. La fenêtre de dialogue principale (B) est équivalente à celle des modules de calcul : l'utilisateur y entre librement les paramètres de base: fichiers de coordonnées, numéros de colonnes, options particulières à un type de graphique (par exemple le tracé d'un cercle de rayon unité pour faire un cercle de corrélation en ACP normée). Le bouton Draw lance le tracé du graphique avec les paramètres courants, tandis que le bouton Quit permet de quitter le module. Les boutons Copy graph, Save graph, et Print graph permettent respectivement de copier le graphique courant dans le presse-papier du Macintosh, de l'enregistrer dans une fichier (de type PICT), ou de l'imprimer directement. Ces boutons permettent de reproduire en un seul clic de la souris les commandes présentes dans les menus "File" et "Edit" du module.

La fenêtre "Min/Max" (C) présente et permet de modifier les valeurs des paramètres qui sont communs à tous les types de graphiques (ou au moins à plusieurs types) :

- minimum et maximum en abscisses et en ordonnées (en cordonnées utilisateur, avec une option de recalcul automatique ou non de chacune de ces valeurs),
- hauteur et largeur de la fenêtre graphique (en cordonnées physiques),
- nombre de graphiques horizontaux et verticaux dans une collection (avec également une option de recalcul automatique ou non de chacune de ces valeurs),
- nombre de graduations sur les axes des abscisses et des ordonnées,
- valeur du facteur G (facteur de proportionnalité des cercles et des carrés),
- contrainte de l'espace physique à la forme carrée,
- tracé d'un cadre autour de chaque graphique,
- cartouche d'échelles.

La fenêtre de sélection des lignes et des colonnes (D et E) est particulièrement importante. C'est en effet sur elle que repose en grande partie ce qui fait l'originalité et la puissance des modules graphiques. J'ai pour cela repris et généralisé les possibilités de collectionner les graphiques qui existaient déjà dans GraphMu. Il est ainsi toujours possible (sauf lorsque cela n'aurait pas de sens) de sélectionner les colonnes des fichiers de données, chaque colonne conduisant alors au tracé d'un graphique élémentaire dans la collection. Pour les lignes, la fenêtre de sélection possède deux états : l'état "Fichier" (D) et l'état "Clavier" (E). Dans le premier état, l'utilisateur choisit un fichier de sélection des lignes. Ce fichier doit contenir une variable qualitative dont les modalités définissent les numéros des graphiques auxquels seront affectées les lignes des fichiers de données. Dans le second état, l'utilisateur sélectionne lui-même les lignes constituant chaque graphique. Par défaut, toutes les lignes vont dans le même graphique, et chaque colonne correspond à un graphique différent. Ce fonctionnement est à relier au fait que les modules graphiques de base (Graph1D, Curves et Scatters) ont deux versions : une version "normale", dans laquelle les graphiques élémentaires d'une collection sont juxtaposés dans la fenêtre de tracé, et une version, dont le nom se termine par le suffixe "Class", dans laquelle les graphiques élémentaires sont au contraire superposés (Graph1DClass, CurveClass et ScatterClass). La diversité des possibilités ainsi générée est surprenante, et elle est bien illustrée dans les six volumes de documentation actuellement disponibles. Elle permet en particulier d'introduire directement la notion de **protocole expérimental** (ou au moins la structuration des données) dans les illustrations.

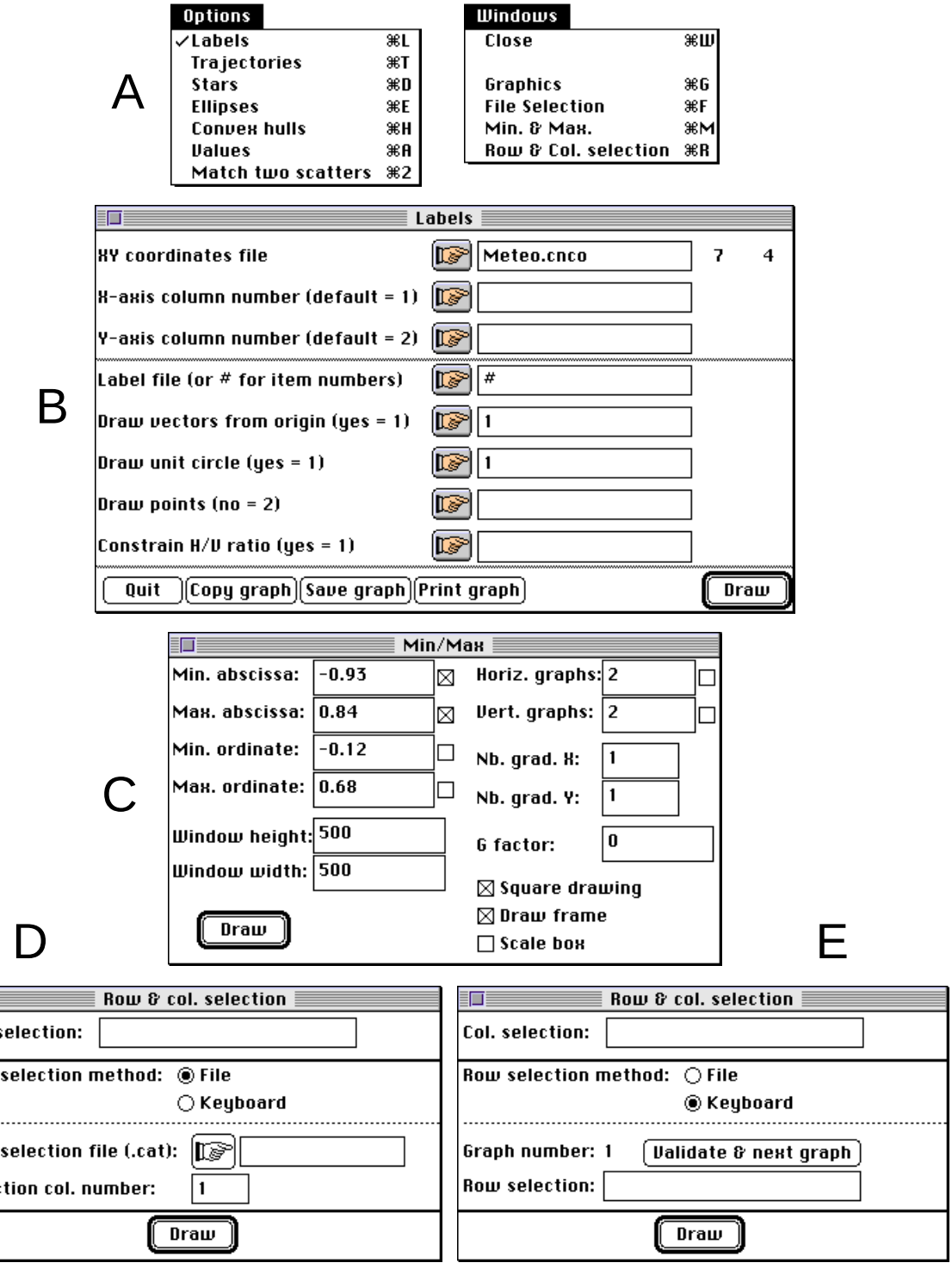


Figure 2.2: Éléments de l'interface utilisateur des modules graphiques d'ADE-4

Que ce soit dans les modules de calcul ou dans les modules graphiques, l'utilisateur peut ainsi choisir facilement l'opération qu'il veut effectuer (menu Option), et il est ensuite guidé dans l'entrée des paramètres par la fenêtre de dialogue principale. Le fait qu'elle soit non modale lui permet par exemple de changer d'avis, de recommencer, de consulter d'autres documents avant de lancer les calculs, etc. Les paramètres d'entrée sont modifiables dynamiquement et il n'est donc pas nécessaire de relancer l'exécution du module pour les modifier. Certains paramètres reçoivent automatiquement des valeurs par défaut (par exemple pondérations uniformes en ACP) mais ils peuvent aussi

être modifiés, ce qui évite d'avoir à se préoccuper de détails triviaux dans les analyses simples, tout en offrant des possibilités sophistiquées. Ces avantages peuvent paraître d'importance secondaire, mais ils permettent de se concentrer sur l'objectif de l'analyse plutôt que sur la façon de l'atteindre, et d'explorer, en quelques minutes, diverses stratégies. Les temps de calcul étant devenus pratiquement négligeables (moins de 5 secondes pour l'ACP d'un tableau 100 x 50, et 10 secondes pour 200 x 100), il était nécessaire d'en faciliter la mise en oeuvre, sous peine de rebuter les utilisateurs qui sont maintenant habitués aux logiciels munis d'une IUG évoluée.

De plus, cette interface utilisateur a permis, sans accroître la complexité apparente des modules, d'augmenter considérablement la granularité du système, et donc de réduire sa taille tout en facilitant sa maintenance : un module d'ADE-4 correspond typiquement à environ 5 à 10 modules d'ADE 3.7. Nous avons longuement discuté cette question de la granularité des modules, et l'état actuel représente un compromis entre complexité de l'interface et facilité de maintenance : la gestion de 45 000 lignes de code (30 000 pour les programmes principaux et 15 000 pour les sous-programmes) ne se fait pas sans quelques problèmes. De plus, elle est aussi soumise à des considérations d'ordre plus stratégique : les grands standards des méthodes d'analyse de données, comme l'analyse canonique des correspondances (ACC) peuvent en effet justifier d'extraire ces méthodes des modules compacts (dans ce cas le module Projectors) pour en faire un module particulier, ce qui en facilitera l'utilisation et favorisera la diffusion des autres modules. Fort heureusement cette opération (ainsi que d'autres types de réorganisation, comme par exemple la construction des modules MacMul 4 et GraphMu 5) est réalisable de façon très simple grâce au fait que l'IUG des modules est complètement séparée de la partie calculatoire : il suffit donc d'extraire le code de calcul et de lui rajouter une partie interface propre.

La notion de dossier de travail a elle aussi été l'objet de nombreuses discussions. En effet, si elle n'apparaît pas directement comme élément de l'interface utilisateur, elle représente quand même un point sous-jacent important dans le fonctionnement des modules. Elle avait été complètement banalisée dans les premières versions des modules que j'avais écrits (le dossier de travail était simplement le dernier dossier ouvert lors d'une sélection de fichier, et tous les fichiers d'entrée et de sortie devaient s'y trouver). Cette solution était certainement la plus souple du point de vue de l'utilisateur, mais elle ne permettait pas de satisfaire certaines contraintes pratiques, issues de l'augmentation de la complexité des méthodes. Les méthodes de couplage et les méthodes multitableaux nécessitent en effet de disposer d'un nombre important de fichiers, à la fois en entrée et en sortie, et la dispersion de ces fichiers dans divers dossiers de la hiérarchie du système Macintosh multipliait les risques d'erreur, auxquels venait s'ajouter le désagrément d'avoir à retrouver ces fichiers pour poursuivre les calculs ou pour réaliser des graphiques. Nous avons finalement conservé cette notion de dossier de travail unique regroupant tous les fichiers, mais en autorisant sa modification à l'intérieur de chaque module (grâce à la commande "Préférences" du menu "Edit"), ainsi qu'à travers l'interface HyperCard. Cette solution minimise les contraintes imposées à l'utilisateur, tout en augmentant la fiabilité générale du système.

L'interface HyperCard a elle aussi été conservée (figure 2.3) mais son emploi, qui nécessite HyperCard version 2.2 (logiciel maintenant commercialisé par Apple Computer), est laissé au choix de l'utilisateur. Tous les modules sont réunis sur une seule carte de la pile ADE•Base, sous forme de menus pop-up permettant de lancer leur exécution, ou d'accéder aux cartes correspondantes de la pile ADE•Doc. Cette dernière offre un résumé des fonctions de chaque module. Dans la partie gauche de la carte de base, un champ texte montre les fichiers présents dans le dossier de travail courant et permet de les manipuler, ainsi que de modifier le dossier de travail. Deux autres piles, ADE•Data et ADE•Biblio ont aussi été conservées et augmentées. Elles contiennent actuellement plus de 1000 références bibliographiques et 150 jeux de données.

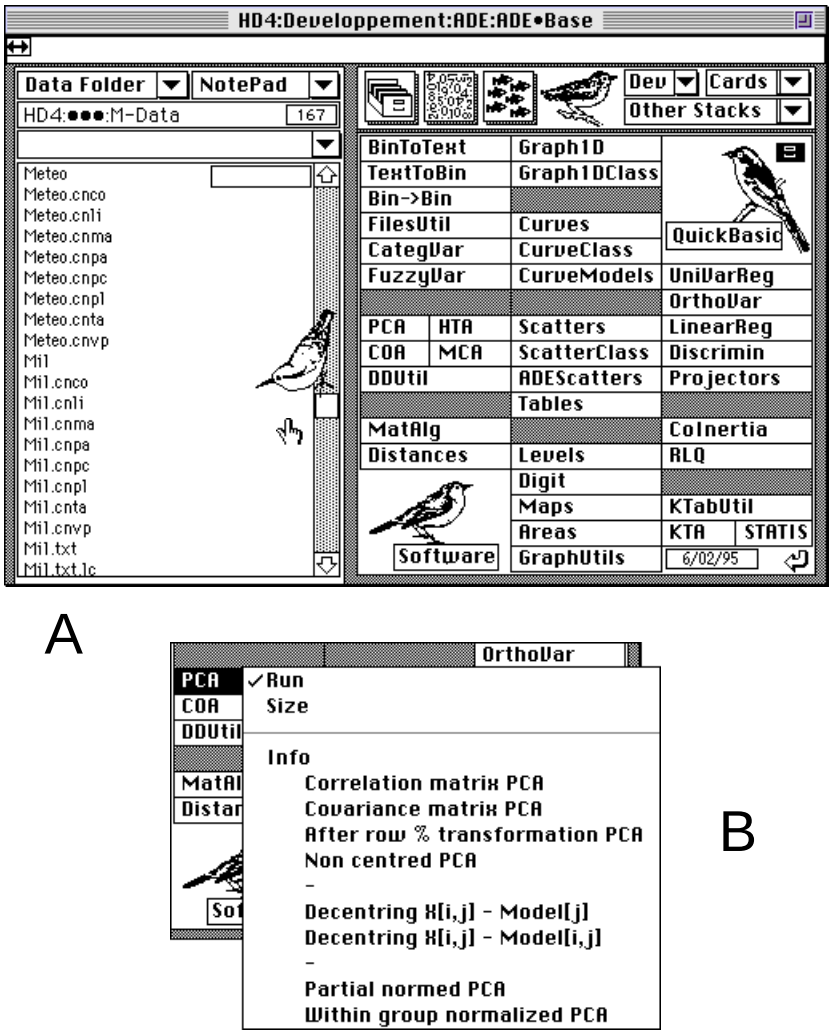


Figure 2.3: Éléments de l'interface utilisateur HyperCard d'ADE-4

1.2. Les modules de calcul

Afin de ne pas surcharger ce mémoire, j'ai résumé les principales fonctionnalités des modules de calcul dans des tableaux présentés en annexe. Les développements théoriques dont sont issues ces fonctionnalités sont le fruit du travail de Daniel Chessel et je ne présente ici qu'un résumé très succinct des différentes méthodes disponibles dans ADE-4.

J'ai choisi une présentation volontairement "mécaniste" de ces méthodes, plutôt qu'une présentation basée sur des objectifs biologiques, car ce point de vue reflète mieux la structure informatique du logiciel et la façon hiérarchique dont il a été construit (le chapitre trois de ce mémoire présentera des applications à caractère biologique). Les différentes méthodes sont donc réparties en : analyses à un tableau, analyses à un tableau avec structures spatiales, analyses à un tableau avec groupes de lignes, analyses à un tableau avec une variable à prédire (méthodes de régression), analyses à deux tableaux, et analyses à k tableaux.

1.2.1. *Les analyses à un seul tableau*

A la base du système se situent trois modules: PCA, COA et MCA (tableaux 2.1, 2.2 et 2.3), qui mettent en oeuvre les trois méthodes simples d'analyse à un tableau : l'analyse en composantes principales (Hotelling 1933) pour traiter les tableaux de variables quantitatives, l'analyse des correspondances (Williams 1952, Greenacre 1984) pour les tables de contingences, et l'analyse des correspondances multiples (Nishisato 1980, Tenenhaus & Young 1985) pour les tableaux de variables qualitatives. Ces modules offrent plusieurs variantes plus originales : ACP non centrée ou décentrée, ACP normée partielle, ACP interne, AFC interne, AFC décentrée, analyse des correspondances non symétrique, reciprocal scaling, analyse des correspondances floue.

Mais le rôle de ces trois modules ne s'arrête pas là : ils constituent aussi la première étape de presque toutes les autres méthodes. L'exemple le plus simple est celui de l'analyse discriminante : pour réaliser une analyse discriminante dans ADE-4, il faut commencer par utiliser l'un des trois modules de base, puis, dans le module Discrimin, choisir la variable qualitative indicatrice des classes et exécuter les calculs de l'analyse discriminante ainsi définie. On voit que cette stratégie permet de mettre en oeuvre de façon tout à fait naturelle trois types d'analyse discriminante: l'analyse discriminante classique (dans le cas où le module de base était le module PCA), l'analyse discriminante sur variables qualitatives (avec le module MCA), et l'analyse discriminante des correspondances (avec le module COA). Ce type de fonctionnement prend tout son intérêt dans les méthodes de couplage de deux tableaux. Par exemple en analyse de co-inertie, chaque tableau est préparé séparément à l'aide d'un des trois modules de base, et le couplage des deux analyses simples s'effectue ensuite avec le module CoInertia. Cette façon de faire réduit la taille des modules, simplifie la mise en oeuvre, et permet de comparer les résultats du couplage à ceux des analyses séparées.

Un quatrième module appartient aussi à la catégorie des analyses à un tableau. Il s'agit du module HTA (homogeneous table analysis). Ce module est dédié à l'analyse des tableaux homogènes, c'est à dire des tableaux constitués de mesures d'une seule variable, répétées selon deux critères. Un exemple est celui des tableaux de toxicité, dans lesquels la même variable (en général la DL50) est mesurée pour divers composés chimiques sur divers types d'organismes. Dans ces conditions, il n'y a aucune raison de privilégier *a priori* un type de centrage par rapport à un autre, et le module HTA permet alors de comparer six méthodes de centrages différentes : aucun centrage, centrage global (*i.e.*, par rapport à la moyenne générale), centrage ligne, centrage colonne, double centrage (ligne et colonnes) additif, et double centrage multiplicatif.

Le module DDUtil apporte de nombreuses aides numériques à l'interprétation, qui peuvent être utilisées à la suite des trois modules de base : représentation biplot (Gabriel 1971), analyse d'inertie des lignes et des colonnes, projection de lignes et de colonnes supplémentaires, reconstitution de données (Lebart et al., 1984).

1.2.2. *Les analyses à un tableau avec une relation de voisinage*

Lorsque les données disponibles incluent une relation de voisinage (par exemple spatial ou temporel) entre les relevés, le module Distances peut être utilisé pour réaliser les analyses locales et globales. Lebart avait utilisé ces termes dès 1969, avec un sens un peu différent de celui proposé par Thioulouse *et al.*, 1995a. Le point fondamental réside dans l'utilisation de la matrice de voisinage entre relevés pour obtenir à la fois une pondération de voisinage (chaque relevé est pondéré par son nombre de voisins) et une décomposition de la variance totale en variance locale et variance globale (qui peuvent être reliées de façon simple aux indices de Geary et de Moran), ce qui conduit à des analyses dont les codes des relevés maximisent ces quantités (cf chapitre 1). Les tableaux 2.4 et 2.5 de l'annexe résument les fonctionnalités du module Distances, qui offre de plus la possibilité de mettre en oeuvre trois autres méthodes classiques

d'analyse de données spatiales : le test de Mantel (comparaison de deux matrices de distances, Mantel 1967), l'analyse en coordonnées principales (analyse d'une matrice de distances, Manly 1972), et l'arbre de longueur minimale (Kevin & Whitney 1972).

1.2.3. *Les analyses à un tableau avec des groupes de lignes*

Le module Discrimin permet d'analyser un tableau dont les lignes sont réparties en plusieurs groupes. Selon les objectifs, on pourra effectuer une analyse discriminante, ou une analyse inter/intra. L'analyse discriminante est vue ici comme l'ACP des centres de gravité des groupes de lignes. L'inverse de la matrice de covariance est calculée par diagonalisation, ce qui fournit directement une inverse généralisée (inverse de la restriction au sous-espace de dimension égal au rang) dans le cas où la matrice n'est pas de rang plein. Le module offre, de plus, un calcul d'analyse de la variance, des tests de permutation pour juger de la significativié de la discrimination, et la projection de lignes supplémentaires. Le tableau 2.6 de l'annexe résume ces possibilités.

1.2.4. *Les analyses à un tableau avec une variable à prédire*

Trois modules introduisent les méthodes de régression linéaire. UniVarReg (tableau 2.7 en annexe) traite de la régression polynomiale et du modèle Lowess (locally weighted regression and smoothing scatterplots), Cleveland 1979, Cleveland & Devlin 1988. OrthoVar (tableau 2.8 en annexe) réalise la régression linéaire multiple dans le cas où les variables explicatives sont orthogonales. C'est le cas par exemple de la régression sur composantes principales (PCR) (Naes 1984), ou de la projection sur le sous-espace engendré par les vecteurs propres d'un opérateur de lissage (Méot *et al.* 1993, Thioulouse *et al.* 1995a). Enfin, le module LinearReg (tableau 2.9 en annexe) permet d'effectuer la régression linéaire multiple classique et la régression PLS (partial least squares) (Geladi & Kowalski 1986, Höskuldsson 1988, Lindgren 1994).

1.2.5. *Les analyses à deux tableaux*

On peut distinguer deux types d'analyses à deux tableaux : les analyses simples et celles à norme inversée. Une autre typologie distingue les analyses symétriques (les deux normes sont inversées ou pas) et les analyses non symétriques (une seule des deux normes est inversée). Le croisement des deux conduit à distinguer l'analyse de co-inertie, qui est une analyse symétrique simple, l'analyse canonique, qui est aussi symétrique mais où les deux normes sont inversées, et les analyses en composantes principales sur variables instrumentales (ACPVI) qui sont des analyses non symétriques car la norme est inversée dans un seul des deux espaces. Dans cette dernière catégorie on trouve l'AFCVI (analyse factorielle des correspondances sur variables instrumentales, Lebreton *et al.* 1988a, 1988b), ou CAIV (correspondence analysis with respect to instrumental variables, Lebreton *et al.* 1991) analogue à l'ACC (analyse canonique des correspondances) de ter Braak (1987a, 1987b).

Le module Projectors est dédié à l'utilisation des projecteurs en analyse multivariée (Takeuchi *et al.* 1982). Il permet de générer des bases orthonormales de diverses façons (en particulier à partir d'un tableau) et de projeter les variables d'un second tableau sur le sous-espace vectoriel engendré. Il peut ainsi être utilisé pour effectuer toutes les ACPVI, et donc en particulier l'ACC, avec des tests de permutation. Le tableau 2.10 en annexe donne la liste de ses options.

Le module CoInertia réalise l'analyse de co-inertie (Chessel & Mercier 1993, Dolédec & Chessel 1994, Thioulouse & Lobry 1995, Cadet *et al.* 1994) et offre la possibilité de faire des tests de permutation. La liste des options figure dans le tableau 2.11 en annexe.

Le module RLQ introduit une généralisation de l'analyse de co-inertie au cas de trois tableaux : un tableau de variables environnementales (le tableau R, avec n relevés en lignes et p variables en colonnes), un tableau d'abondances floro-faunistique (le tableau L, avec les mêmes n relevés en lignes et q espèces en colonnes), et un tableau de traits spécifiques (le tableau Q, avec les mêmes q espèces en colonnes et t traits spécifiques en lignes). Le tableau d'abondances floro-faunistique sert de lien entre les deux autres, et l'analyse de co-inertie est réalisée à travers ce lien (Dolédec & Chessel, soumis). Le tableau 2.12 en annexe donne la liste des options.

1.2.6. *Les analyses à k tableaux*

Les analyses à k tableaux sont un exemple typique de la lenteur de la diffusion méthodologique vers le champ des applications en biologie. La méthode STATIS (Escoufier 1980; Lavit 1988; Lavit *et al.* 1994) est l'archétype de ces analyses en France. La thèse de L'Hermier des Plantes date de 1976, et malgré un intérêt évident de cette méthode dans le domaine écologique (étude de l'évolution des structures des communautés écologiques), les utilisations sont restées très rares (citons cependant l'étude d'Amanieu *et al.* 1981). Nous espérons que la disponibilité de cette méthode dans ADE-4 en facilitera la diffusion.

Le module STATIS permet d'utiliser la méthode STATIS classique (compromis d'opérateurs), et deux méthodes dérivées portant sur des compromis de tableaux : l'analyse triadique partielle (Thioulouse & Chessel 1987), et l'analyse d'une suite de tables de contingences (Foucart 1978). Ces trois options sont répertoriées dans le tableau 2.13 de l'annexe.

Le module MFA met en oeuvre l'analyse factorielle multiple (Escofier & Pages 1994). Il s'agit là aussi d'une méthode "classique" (le premier article date de 1982 : Escofier & Pages 1982) mais dont les utilisations en écologie sont presque inexistantes, ou très récentes (Mateille *et al.* 1995). La seule option du module correspond au cas où les tableaux ont leurs lignes (individus) en commun, ce qui est le cas envisagé par les auteurs de la méthode (tableau 2.14 de l'annexe).

Le module KTA permet d'une part, d'effectuer automatiquement toutes les analyses simples d'un multi-tableau, et d'autre part, de mettre en oeuvre la généralisation de l'analyse de co-inertie à k tableaux (méthode ACOM, analyse de co-inertie multiple, Chessel & Hanafi 1995), tableau 2.15 de l'annexe.

Un problème pratique non négligeable rencontré dans les analyses multi-tableaux est celui des diverses manipulations nécessaires à la préparation des fichiers de données. Il est en effet nécessaire de définir une structure permettant de collectionner et de réorganiser les différents tableaux. Le module KTabUtil remplit cette fonction. Il permet de définir un multi-tableau général dans un seul fichier (les tableaux élémentaires peuvent être collectionnés en lignes, en colonnes, ou simultanément en lignes et en colonnes), à partir d'une ou deux indicatrices des groupes de lignes ou de colonnes. On peut ensuite réaliser diverses opérations : transposition, réorganisation, centrage, normalisation, etc. La structure ainsi définie est bien sûr compatible avec les modules graphiques, ce qui permet de **collectionner automatiquement** les graphiques en fonction de la collection de tableaux, et ceci pour représenter soit les données initiales, soit les coordonnées factorielles.

1.3. Les modules graphiques

Comme cela a été souligné dans le chapitre 1, la graphique représente bien plus qu'une procédure automatique en sortie d'un programme d'analyse multivariée. Elle doit faire partie intégrante du processus d'analyse, et nous avons vu comment l'interface utilisateur des modules graphiques a été conçue dans ce sens.

ADE-4 comprend actuellement 14 modules graphiques, qui se répartissent en cinq catégories : graphiques à une dimension, courbes, nuages de points, tableaux , et cartes géographiques.

1.3.1. *Graphiques à une dimension*

Le module **Graph1D** permet de tracer des graphiques "à une dimension". Cette appellation désigne en fait la structure des données sous-jacente : on considère les données comme une série unidimensionnelle de n valeurs $[x_i]_{i=1,n}$. Cette série de valeurs définit l'implantation horizontale ou verticale. La composante G est constituée de labels (variable de séparation ponctuelle), de barres verticales (variable de taille surfacique, cas des histogrammes) ou de traits (courbes de Gauss).

Deux options sont disponibles : **Labels** et **Histograms**. La première trace des labels équirépartis le long d'un axe, reliés aux points de coordonnées $[x_i]$ sur cet axe. Elle est particulièrement utile dans le dépouillement d'axes factoriels individuels (analyse discriminante à deux classes par exemple). La seconde calcule la distribution des valeurs $[x_i]$ en k classes et trace les histogrammes correspondants, éventuellement superposés aux courbes de Gauss de mêmes paramètres que la distribution obtenue. Elle peut être utilisée sur des données brutes ou sur des coordonnées factorielles, et elle est elle aussi surtout intéressante dans le dépouillement d'un axe factoriel (par opposition à un plan factoriel).

Dans les deux options, si le tableau de données comporte plusieurs colonnes, chaque colonne correspond à une série de valeurs $[x_i]$ et elles sont tracées côte à côte. Cette possibilité autorise le dépouillement automatique d'une série d'axes factoriels. Il est possible de sélectionner les colonnes à tracer, ainsi que les groupes de lignes qui formeront chaque graphique élémentaire de la collection, grâce à la fenêtre de sélection des lignes et des colonnes, soit au clavier soit avec un fichier de sélection qualitatif (figure 2.2D et 2.2E).

Dans l'option Labels, on peut tracer les graphiques horizontalement ou verticalement, fixer les labels ou (par défaut) utiliser le numéro d'ordre, et faire apparaître ou non un label pour les différentes séries de valeurs.

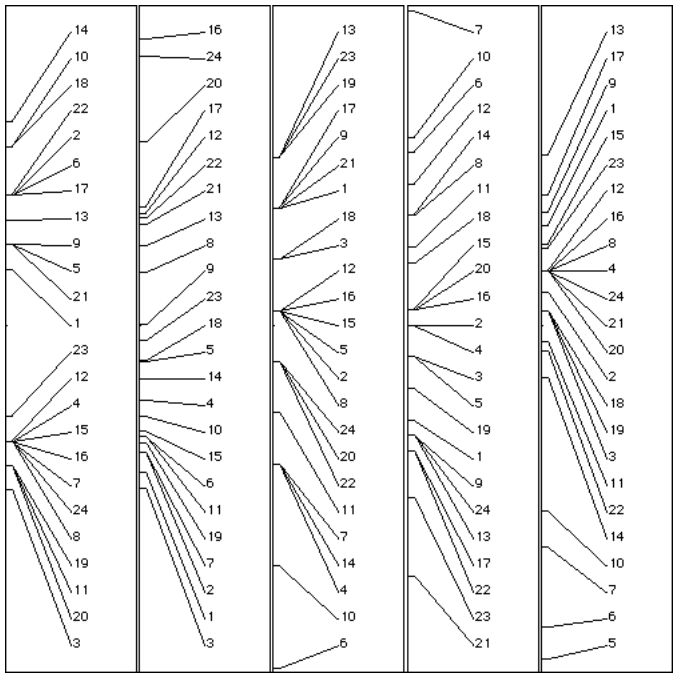


Figure 2.4: Représentation de labels équirépartis le long d'un axe (Module Graph1D, option Labels)

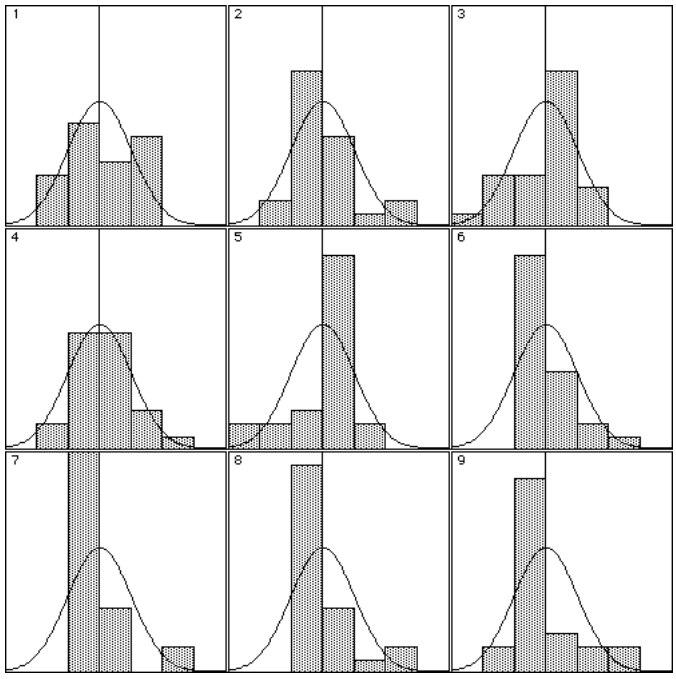


Figure 2.5: Représentation de la distribution de séries de valeurs (Module Graph1D, option Histograms)

Dans l'option Histograms, il est possible de choisir le nombre de classes de l'histogramme, de faire apparaître ou non un label pour les différentes séries de valeurs, de tracer au choix soit l'histogramme, soit la courbe de Gauss, ou les deux superposés, et d'exprimer les histogrammes en pourcentages ou en effectifs. Ces représentations permettent de se faire rapidement une idée des variances inter et intra, par exemple en analyse discriminante ou en analyse inter/intra.

Le module **Graph1DClass** fonctionne en superposant les graphiques élémentaires d'une collection au lieu de les juxtaposer comme le fait Graph1D. La superposition de graphiques de type labels équirépartis ou de type histogrammes serait illisible, et la seule option disponible dans ce module concerne donc les courbes de Gauss. Les graphiques élémentaires ne sont cependant pas superposés tous ensembles : seuls ceux correspondants à des groupes de lignes le sont. Un fichier supplémentaire (".cat") contenant une (ou plusieurs) variable qualitative définissant ces groupes de lignes est donc nécessaire.

Si ce fichier qualitatif de sélection des lignes possède plusieurs colonnes (correspondant donc à plusieurs partitions des lignes), les graphiques correspondants à ces colonnes sont juxtaposés. Si le fichier de données possède également plusieurs colonnes, les graphiques correspondants à ces colonnes sont aussi juxtaposés, et la juxtaposition se fait en priorité par rapport aux colonnes du fichier de sélection (*i.e.*, les différents graphiques correspondants aux colonnes du fichier de sélection des lignes sont juxtaposés pour la première colonne du fichier de données, puis pour la seconde, la troisième, etc.).

Comme dans Graph1D, il est aussi possible de choisir le nombre de classes pour le calcul de la distribution, et de faire apparaître ou non un label pour les différentes séries de valeurs. Il est de plus possible de faire apparaître le label de la classe sur chaque courbe de Gauss. Ce label provient automatiquement du fichier ".123" créé lors de l'exécution du module CategVar sur le fichier qualitatif.

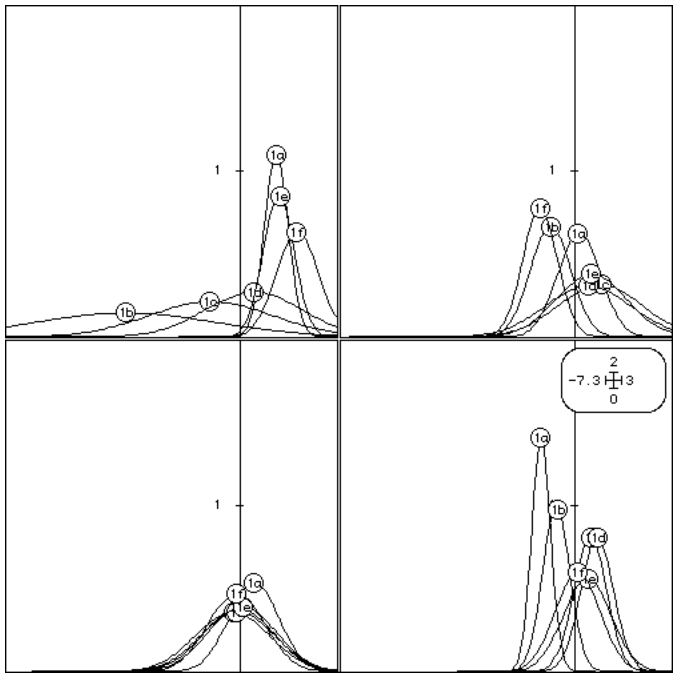


Figure 2.6: Représentation de la distribution de quatre séries de valeurs,
chaque série étant répartie en six classes (Module Graph1DClass, option Gauss curves).
Le fichier de données possède quatre colonnes.
Le fichier qualitatif de sélection des lignes en a une seule,
correspondant à une variable qualitative à six modalités.
Ces six modalités définissent six groupes de lignes
et chaque courbe de Gauss correspond à un de ces groupes.

Ce type de représentation est très utile dans le dépouillement des ACM ou plus généralement des analyses portant sur des variables qualitatives. Il est en effet possible de représenter de cette façon la variabilité des coordonnées factorielles des individus

pour chaque modalité des variables qualitatives du tableau de données, ce qui facilite beaucoup leur interprétation. La représentation bi-dimensionnelle correspondante est bien sûr celle des ellipses d'inertie tracées sur les plans factoriels.

1.3.2. *Courbes*

Trois modules permettent de tracer des courbes : Curves, CurveClass et CurveModels. La structure de données utilisée ici correspond à deux séries de valeurs $[x_i]_{i=1,n}$ et $[y_i]_{i=1,n}$. La série $[x_i]$ correspond à l'implantation horizontale (positions des abscisses des points de la courbe), et la série $[y_i]$ à l'implantation verticale (ordonnées des points). La composante G est variable : traits, barres verticales, escaliers, ou complexe (option Boxes).

Le module **Curves** trace des courbes simples (option **Lines**), des diagrammes en bâtons (option **Bars**), des courbes en escaliers (option **Steps**), et des courbes où chaque sommet est une "boîte à moustaches" (option **Boxes**). Les abscisses des points de chaque courbe peuvent provenir d'un fichier ou bien être équiréparties, de 1 jusqu'au nombre total de points. Les ordonnées proviennent obligatoirement d'un fichier de données. Comme dans la plupart des modules, si ce fichier possède plusieurs colonnes, les courbes correspondant à chaque colonne sont juxtaposées, avec la possibilité de sélectionner ces colonnes, ainsi que des groupes de lignes. Il est possible de faire apparaître ou non un label pour les différentes séries de valeurs, et de tracer des courbes cumulées (uniquement pour les trois premières options).

Dans l'option Lines, on peut tracer séparément les points et/ou la courbe, affecter un label à chaque point (qui peut être simplement son numéro d'ordre ou bien provenir d'un fichier), et superposer un nombre arbitraire de courbes dans chaque graphique élémentaire d'une collection (chaque courbe est alors numérotée).

Dans l'option Bars on peut fixer la largeur des barres verticales (en pixels). Par défaut elle est telle que les barres soient jointives.

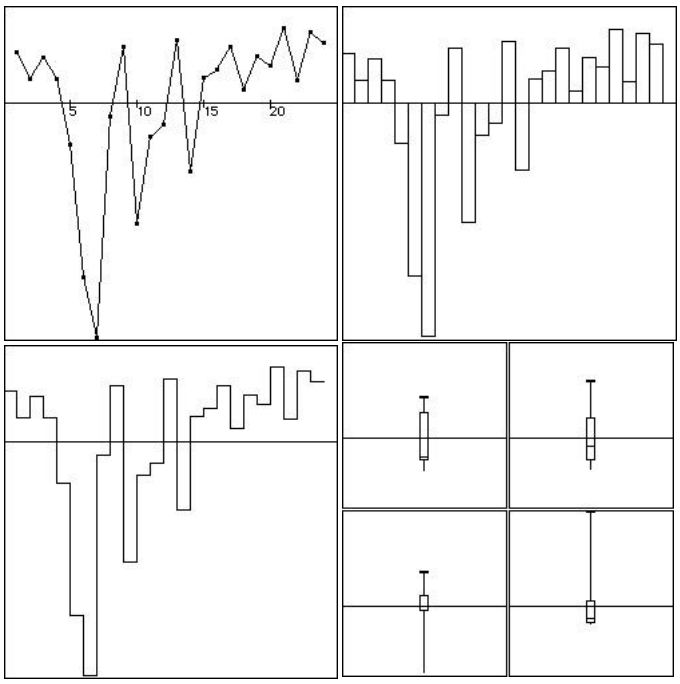


Figure 2.7: Exemple de graphiques tracés avec le module Curves (les quatre options disponibles sont représentées)

Le module **CurveClass** fonctionne comme le module Curves, mais, comme le module Graph1DClass, il superpose les graphiques élémentaires dans les collections provenant de la définition de groupes de lignes (il faut donc définir un fichier qualitatif pour la sélection des lignes). Pour cette raison, seules les options **Lines** et **Boxes** sont disponibles (la superposition de barres ou de courbes en escaliers a une très mauvaise lisibilité).

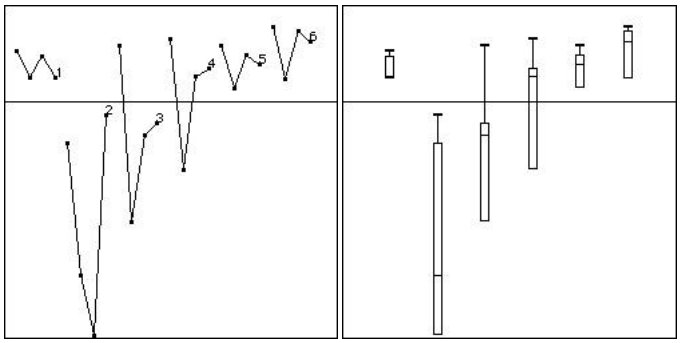


Figure 2.8: Exemple de graphiques tracés avec le module CurveClass (les deux options disponibles sont représentées)

Le module **CurveModels** permet de tracer des modèles de courbes en superposant données observées (points) et courbes ajustées (traits). Trois options sont disponibles : régression polynomiale, régression Lowess, et modèle numérique (qui permet simplement de superposer deux séries de valeurs numériques).

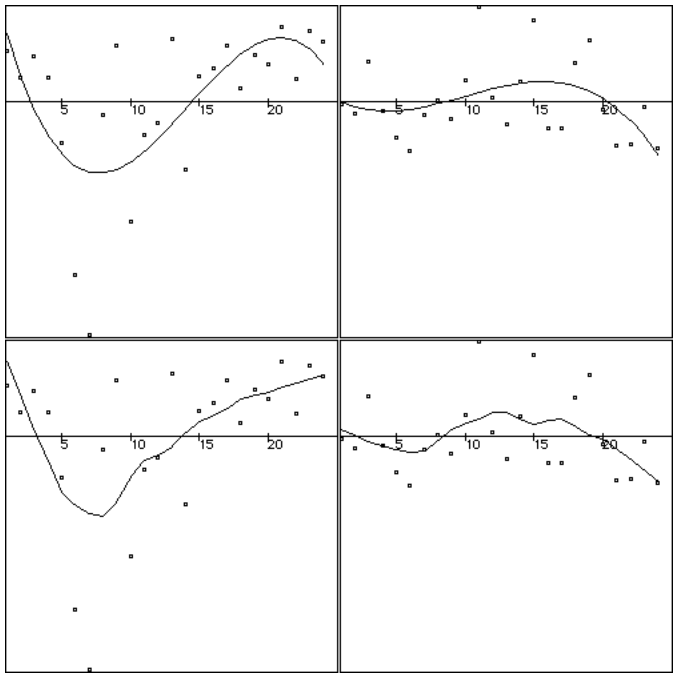


Figure 2.8: Exemple de graphiques tracés avec le module CurveModels : En haut deux ajustements avec des polynômes de degré 3, en bas avec une régression lowess sur 10 points.

Dans les options **Lowess** et **Polynomials**, les abscisses des points de chaque courbe peuvent provenir d'un fichier, ou bien être équiréparties entre 1 et le nombre total de points. Les ordonnées proviennent obligatoirement d'un fichier de données dont il est possible de sélectionner les colonnes, mais pas les lignes. Un fichier de pondération des points peut être utilisé (sinon tous les points ont le même poids). Il est possible de faire

apparaître ou non un label pour les différentes courbes, de tracer séparément les points (données) et/ou les courbes (modèles), et de tracer des traits verticaux reliant chaque point à la courbe (résidus de régression).

Dans l'option Polynomials, l'utilisateur fixe le degré du polynôme, alors que dans l'option Lowess il fixe le nombre de points utilisés pour la régression locale.

1.3.3. *Nuages de points*

Le module **Scatters** permet de tracer divers types de représentation graphiques basées sur des nuages de points. La structure de données correspondante est une série de couples de valeurs $[x_i, y_i]_{i=1,n}$ et éventuellement une seconde série $[g_i]_{i=1,n}$. L'implantation provient de la série de couples valeurs $[x_i, y_i]$. La composante G provient de la série de valeurs $[g_i]$ (option Values), ou bien elle est une caractéristique de l'option choisie. L'objectif premier de ce module est bien sûr la réalisation de plans factoriels, avec de nombreuses variantes possibles. Pour les cartes géographiques, qui utilisent aussi des séries de couples de valeurs $[x_i, y_i]$, on utilisera de préférence le module Maps, permettant de superposer le graphique avec un fond de carte.

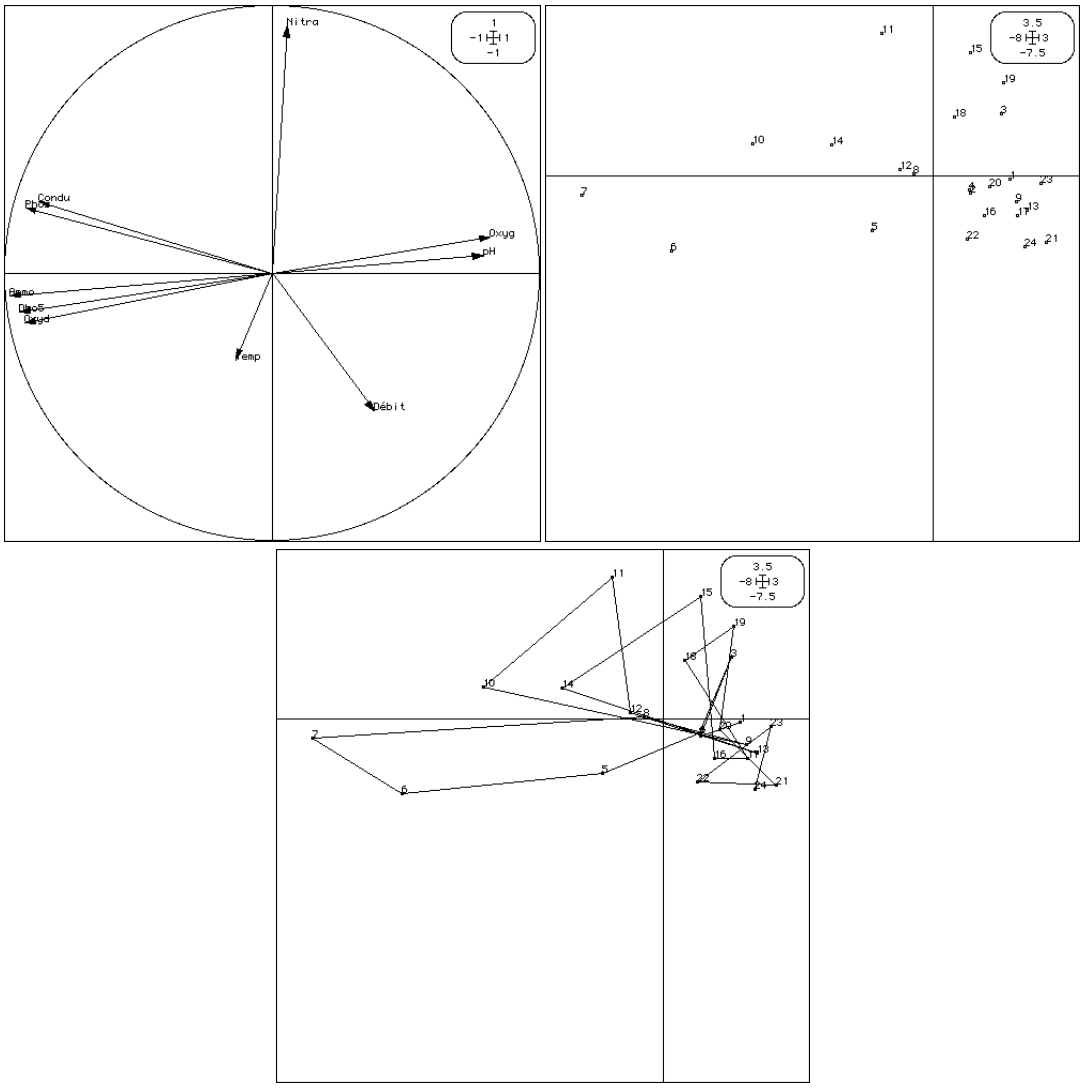


Figure 2.9: Exemple de graphiques tracés avec le module Scatters (option Labels et Trajectories): cercle de corrélation, plan factoriel simple, trajectoires.

L'option de base est intitulée **Labels**. Elle trace une carte factorielle classique, c'est à dire un nuage de points muni d'étiquettes. Les coordonnées des points proviennent d'un fichier XY, dont on peut choisir deux colonnes (une pour les abscisses et une pour les ordonnées) et dont les lignes correspondent aux points affichés sur le plan. Il n'y a pas de possibilité de collection automatique sur les colonnes de ce fichier. Il est par contre possible de sélectionner les lignes et donc de faire des collections de plans factoriels par catégories de lignes. Les labels des points sont optionnels, peuvent êtres lus dans un fichier, ou être égaux au numéro d'ordre des points. On peut tracer un vecteur reliant l'origine à chaque point, un cercle de rayon unité (cercle de corrélation), et afficher ou non les points. On peut aussi contraindre le rapport hauteur/largeur de la fenêtre physique à être égal au rapport hauteur/largeur de la fenêtre utilisateur, ce qui permet de ne pas déformer le plan factoriel. L'autre solution consiste à modifier manuellement les valeurs du minimum et du maximum des abscisses et des ordonnées, ainsi que la hauteur et la largeur de la fenêtre de façon à ce que les rapports hauteur/largeur soient égaux.

L'option **Trajectories** relie les points par un trait dans l'ordre où ils se présentent dans le fichier XY. Ceci permet surtout de souligner l'existence d'une évolution, par exemple une relation d'ordre chronologique entre les points.

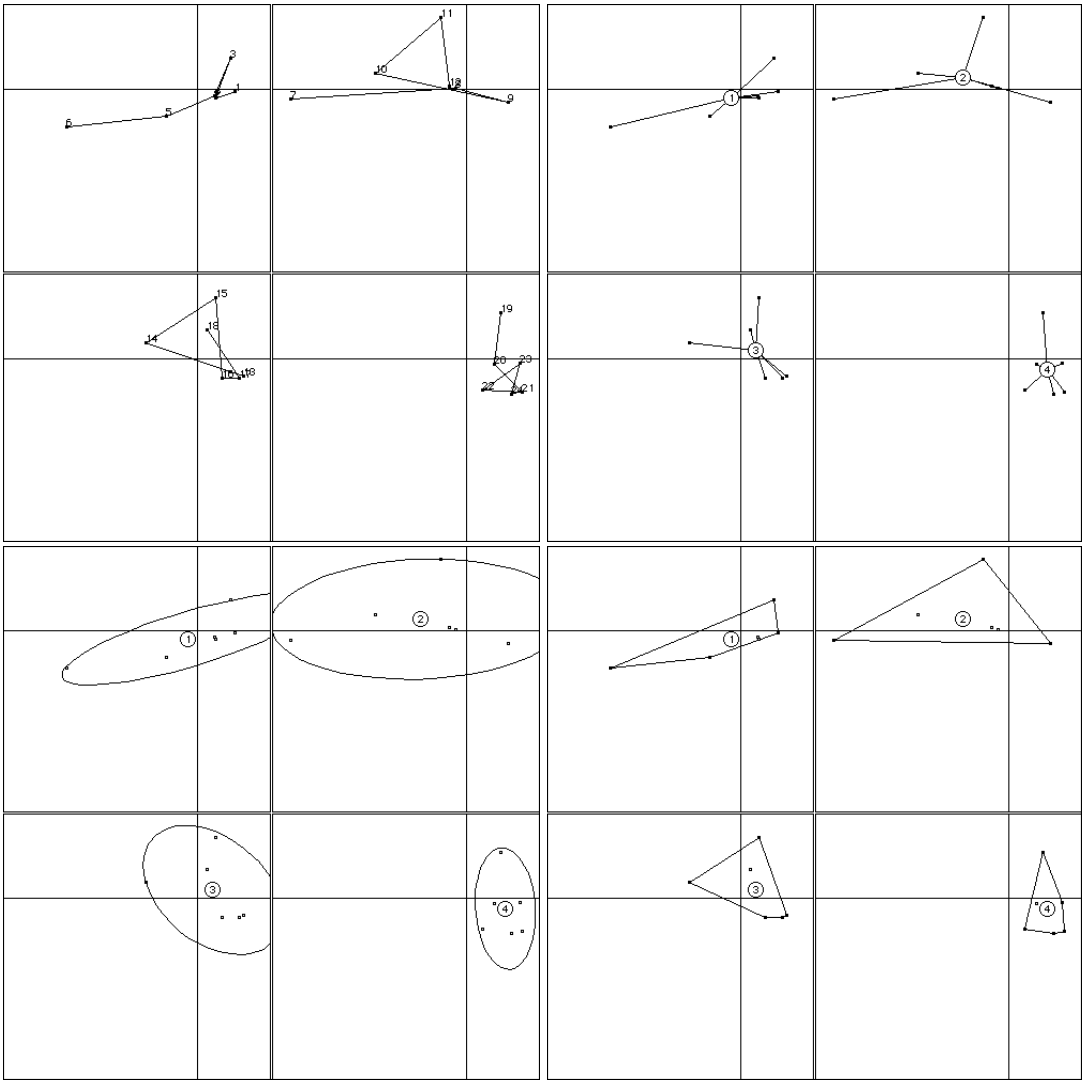


Figure 2.10: Exemple de graphiques tracés avec le module Scatters (option Trajectories, Stars, Ellipses, Convex hulls).

L'option **Stars** relie chacun des points du nuage à son centre de gravité. Dans le cas d'une collection, chaque point d'un sous-nuage de la collection est relié au centre de gravité du sous-nuage.

L'option **Ellipses** trace une ellipse dont les paramètres (coordonnées du centre, longueur des deux axes, pente de l'axe principal) sont fixés par les moyennes, les variances, et la covariance des coordonnées en X et en Y des points du nuage (ou du sous-nuage dans une collection). Elle offre de plus la possibilité de tracer les ellipses de confiance à 50, 70, 90, et 99%.

L'option **Convex hulls** trace l'enveloppe convexe d'un nuage, c'est à dire le polygone (convexe) passant par les points les plus extérieurs du nuage. Elle permet en plus d'utiliser la technique dite de "l'épluchage" (peeling) d'un nuage, en éliminant les points les plus périphériques et en re-traçant ensuite l'enveloppe. Dans cette version, l'élimination des points périphériques est basée sur une classification des points en fonction de leur distance au centre de gravité du nuage.

Ces trois types de représentation sont assez comparables du point de vue fonctionnel. Elles permettent principalement de comparer la variabilité inter- et intra-groupes. L'option Stars souligne plus particulièrement la relation avec le centre de gravité du nuage, et donc la notion de dispersion (variance intra), alors que les options Ellipses et Convex hulls font surtout apparaître les classes (variance inter). Les ellipses peuvent rappeler la distribution multi-normale, ce qui n'est pas toujours intéressant, et on pourra donc leur préférer les enveloppes convexes dans un contexte non inférentiel. L'enveloppe convexe est plus complexe que l'ellipse : une ellipse est définie par cinq paramètres, alors qu'une enveloppe convexe n'a pas de définition paramétrique et peut être formée d'un nombre de sommets variable. L'ellipse est donc meilleure pour un point de vue synthétique.

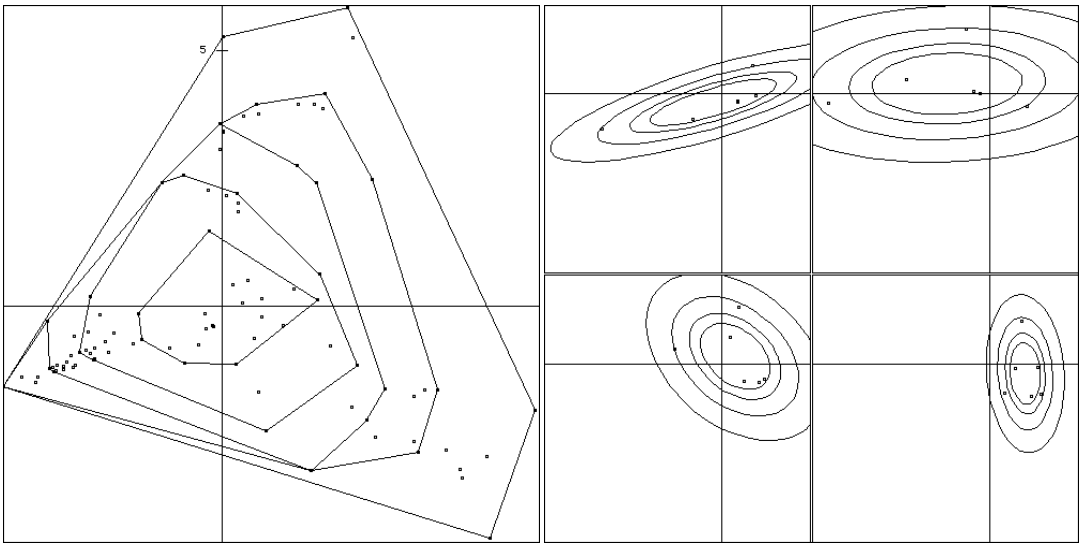


Figure 2.11: Exemple de graphiques tracés avec le module Scatters: option Convex hulls avec peeling en 5 classes et option Ellipses avec représentation des ellipses de confiance à 50, 70, 90 et 99%.

L'option **Values** utilise, en plus de la série de couples $[x_i, y_i]$, une série de valeurs $[g_i]$ (stockées dans le "fichier G") pour tracer des cercles ($g_i>0$) ou des carrés ($g_i<0$) sur le plan factoriel. La relation de proportionnalité entre les valeurs $[g_i]$ et la taille des cercles peut être modifiée par l'intermédiaire du "Facteur G" dans la fenêtre Min/Max (figure 2.2C). Le module autorise simultanément les collections sur les colonnes du fichier G et sur les lignes. La priorité est donnée, comme dans les autres modules, à la sélection des lignes.

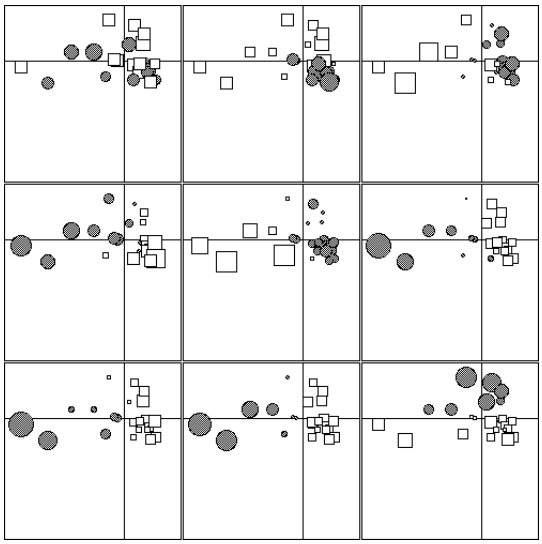


Figure 2.12: Exemple de graphiques tracés avec le module Scatters: option Values avec un fichier G à 9 colonnes.

Ce type de représentation est souvent utilisé pour représenter sur le plan factoriel les valeurs du tableau de données : on réalise une collection de plans factoriels sur lesquels les cercles et les carrés sont proportionnels aux variables initiales (brutes ou de préférence transformées). L'objectif est l'interprétation directe des axes factoriels.

L'option **Match two scatters** permet de comparer deux nuages ayant le même nombre de points. La structure de données employée est celle de deux séries de couples de valeurs [x_i, y_i]. Une flèche est tracée reliant le point de coordonnées (x_i, y_i) de la première série au point correspondant dans la seconde. Cette représentation est particulièrement intéressante dans les analyses à deux tableaux (analyse de co-inertie, analyse canonique des correspondances, etc.) car elle permet de comparer les codes obtenus pour chaque tableau.

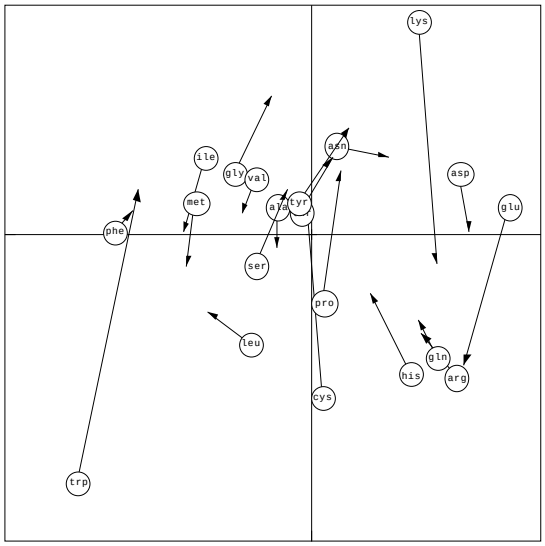


Figure 2.13: Exemple de graphiques tracés avec le module Scatters: option Match two scatters.

L'option **Neighborhood graph** utilise en plus un graphe de voisinage (sous la forme de la matrice du graphe $M = [m_{ij}]$; $m_{ij} = 1$ si i et j sont voisins, 0 sinon). Elle permet simplement de représenter sur le plan factoriel des traits reliant les points voisins. Cette représentation permet d'apprécier la qualité de la prise en compte d'une relation de voisinage dans une analyse.

Le module **ScatterClass** fonctionne de la même façon que les autres modules de type Class : les graphiques correspondant à des groupes de lignes sont superposés. Seules les options **Labels**, **Trajectories**, **Stars**, **Ellipses** et **Convex hulls** sont présentes. Pour ces deux dernières, les ellipses de confiance concentriques et le peeling ne sont pas disponibles. L'avantage par rapport à la version simple (module Scatters) réside dans la diminution du nombre de graphiques élémentaires, et dans une plus grande facilité de comparaison des éléments graphiques qui sont superposés.

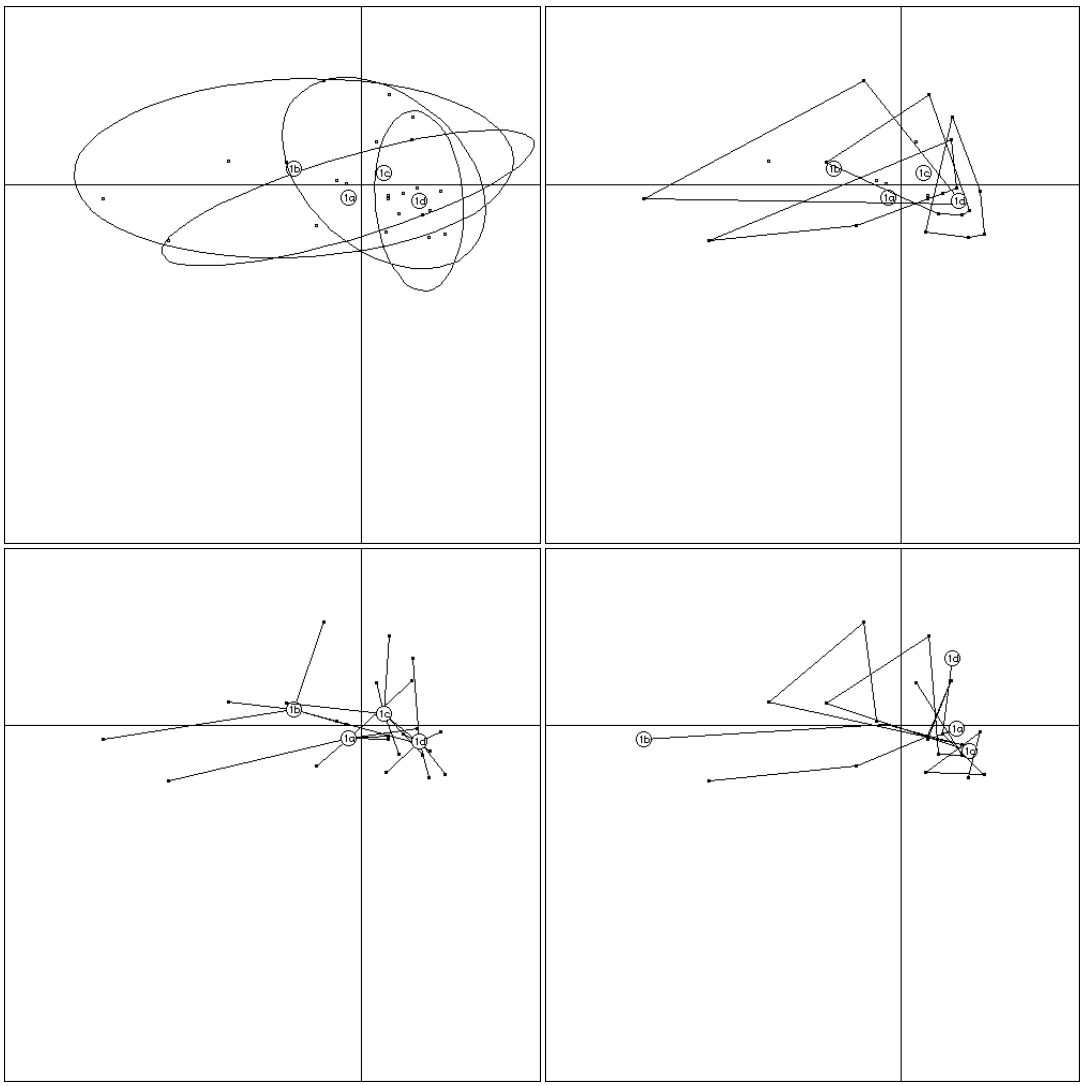


Figure 2.14: Exemple de graphiques tracés avec le module ScatterClass (options Ellipses, Convex hulls, Stars, Trajectories).

1.3.4. *Tableaux*

La structure de données utilisée dans le module **Tables** est une matrice de valeurs $M = [m_{ij}]$. Elle peut s'accompagner de deux séries de valeurs $[x_i]$ et $[y_i]$ qui définissent l'implantation en X et en Y (dans ce cas, l'implantation provient d'une mesure). Dans

l'option **Values**, la composante G est constituée de cercles et de carrés de taille variable (composante G = mesure). Dans l'option **Paint** la composante G est un carré de niveau de gris variable (mesure).

L'option **Values** trace des cercles ($m_{ij} > 1$) et des carrés ($m_{ij} < 1$) de taille proportionnelle aux valeurs m_{ij}. Par défaut la position des cercles et des carrés est fixée par le numéro d'ordre de la ligne et de la colonne de l'élément m_{ij}. Il est possible de rendre ces positions (en lignes et en colonnes) égales aux valeurs lues dans des fichiers, ou de seulement **ré-ordonner** les positions des lignes et des colonnes en fonction de ces valeurs, tout en les laissant équiréparties. On peut aussi inverser verticalement le tableau, et afficher un quadrillage.

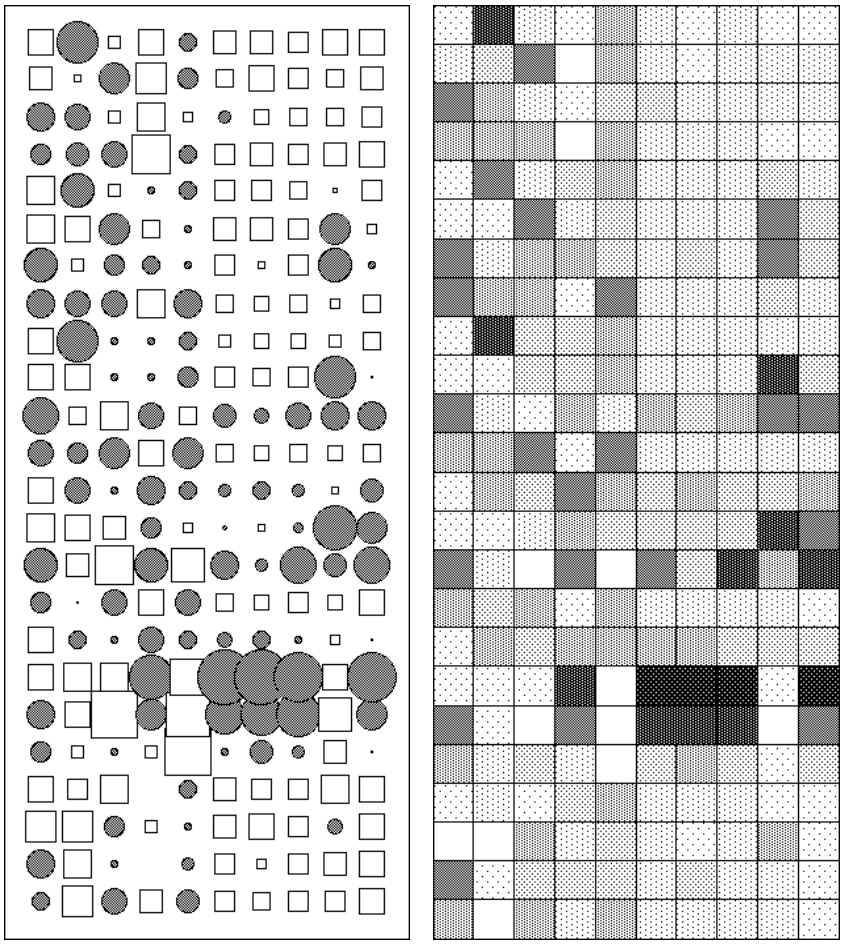


Figure 2.15: Exemple de graphiques tracés avec le module Tables (options Values et Paint).

L'option **Paint** trace simplement un carré de niveau de gris proportionnel aux valeurs m_{ij}. Il est possible d'étiqueter les lignes et les colonnes (avec leur numéro d'ordre ou des labels lus dans un fichier), d'inverser verticalement le tableau, et de fixer le nombre de niveaux de gris (huit au maximum).

Ces pratiques sont particulièrement bien adaptées à l'expression des théorèmes de ré-écriture de tableaux et de reconstitution de données (en ACP et en AFC).

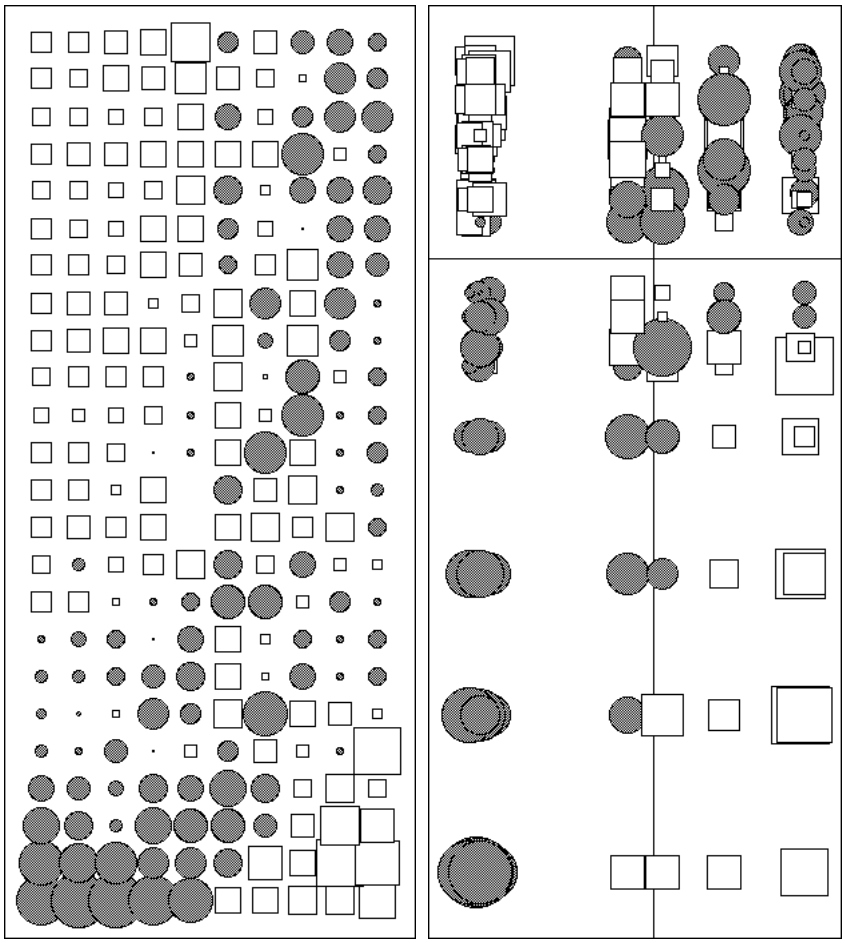


Figure 2.16: Exemple de graphiques tracés avec le module Tables (options Values avec re-ordination et re-positionnement des lignes et des colonnes en fonction des coordonnées factorielles lignes et colonnes de l'analyse du tableau).

1.3.5. *Cartes géographiques*

L'importance de la prise en compte de l'espace dans l'étude des relations espèces-milieu nous a conduits à privilégier cet aspect dans les développements graphiques d'ADE-4. Cette position se traduit par l'existence de quatre modules graphiques dédiés à la cartographie : Digit, Maps, Levels, et Areas.

Le module **Digit** est un utilitaire de digitalisation. La première option (**Digitize**) permet, en partant d'une carte géographique, d'obtenir les coordonnées en X et en Y des points d'échantillonnage repérés sur la carte. La digitalisation se fait "à la souris" en cliquant sur chaque point. Comme le montre la figure 2.17, le module Digit affiche la carte géographique dans une fenêtre, et une petite fenêtre auxiliaire affiche les coordonnées et le nombre de points digitalisés. Ces coordonnées sont enregistrées dans un fichier de sortie.

La seconde option (**Grid preparation**) a pour objectif de définir un maillage régulier sur une zone géographique. Ce maillage est ensuite utilisé dans le module Levels pour interpoler les valeurs observées (qui sont souvent réparties dans l'espace de façon non régulière), et permettre ainsi de tracer des courbes de niveaux. La définition du maillage se fait de façon interactive : la fenêtre affiche la carte géographique, et l'utilisateur fixe le nombre de noeuds horizontaux et verticaux. Une grille est alors superposée à la carte, et l'utilisateur peut sélectionner à la souris les carrés dont les sommets constitueront les

noeuds du maillage définitif (figure 2.18). Il est ainsi possible de définir une zone de forme non rectangulaire.

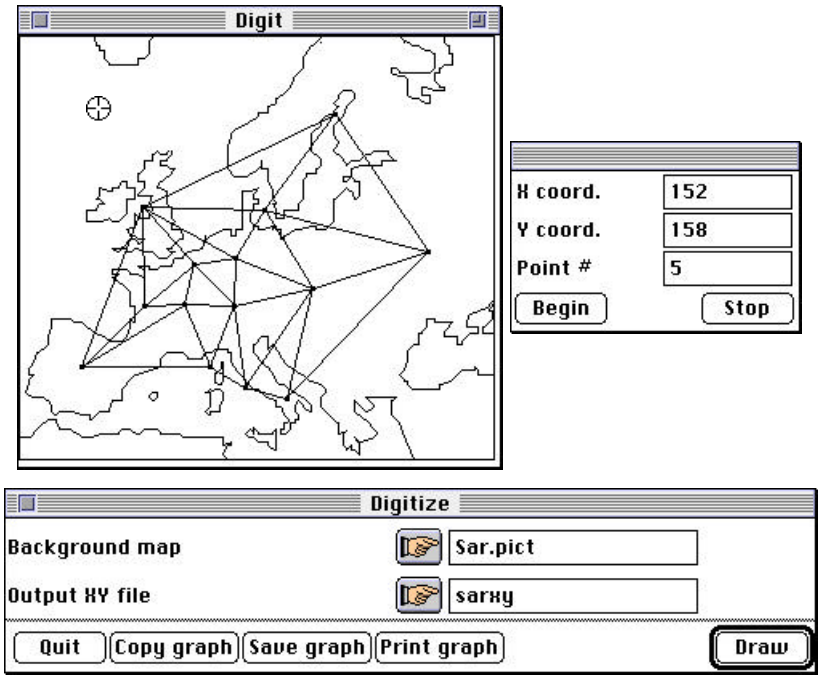


Figure 2.17: Interface utilisateur du module Digit (option Digitize).

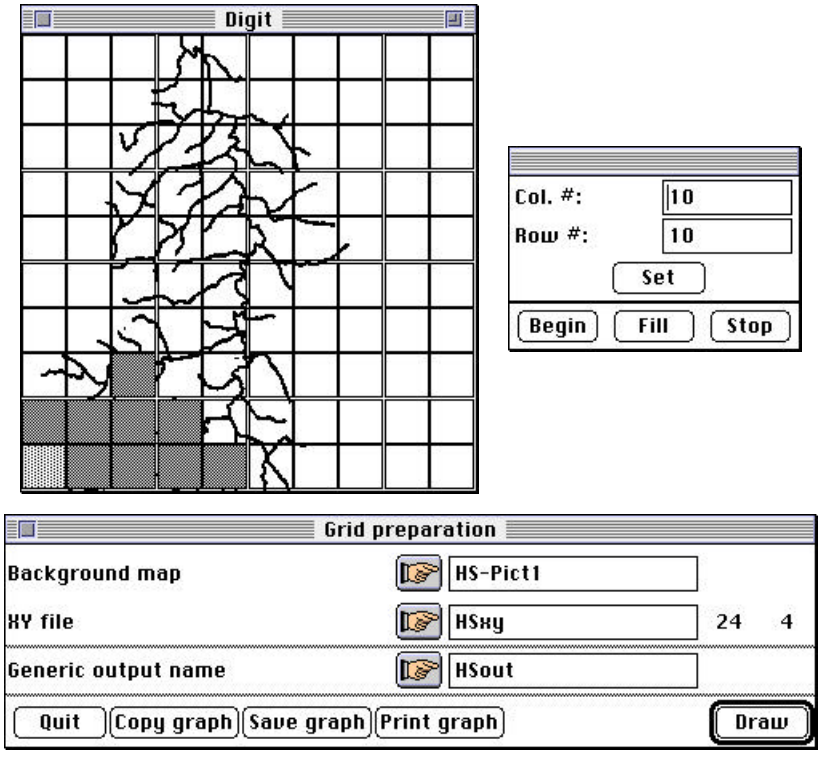


Figure 2.18: Interface utilisateur du module Digit (option Grid preparation).

L'option **Labels** du module **Maps** permet de placer des labels sur une carte aux points de coordonnées obtenus par digitalisation avec le module Digit. Il faut pour cela

disposer d'un fichier de type PICT contenant la carte géographique (scannée à partir d'un document sur papier), du fichier de coordonnées des points, et d'un fichier de labels (ce fichier est facultatif, les labels pouvant être automatiquement pris comme le numéro d'ordre des relevés).

L'option **Neighborhood graph** permet de tracer sur la carte le graphe d'une relation de voisinage (présentée sous la forme de la matrice du graphe) entre les relevés.

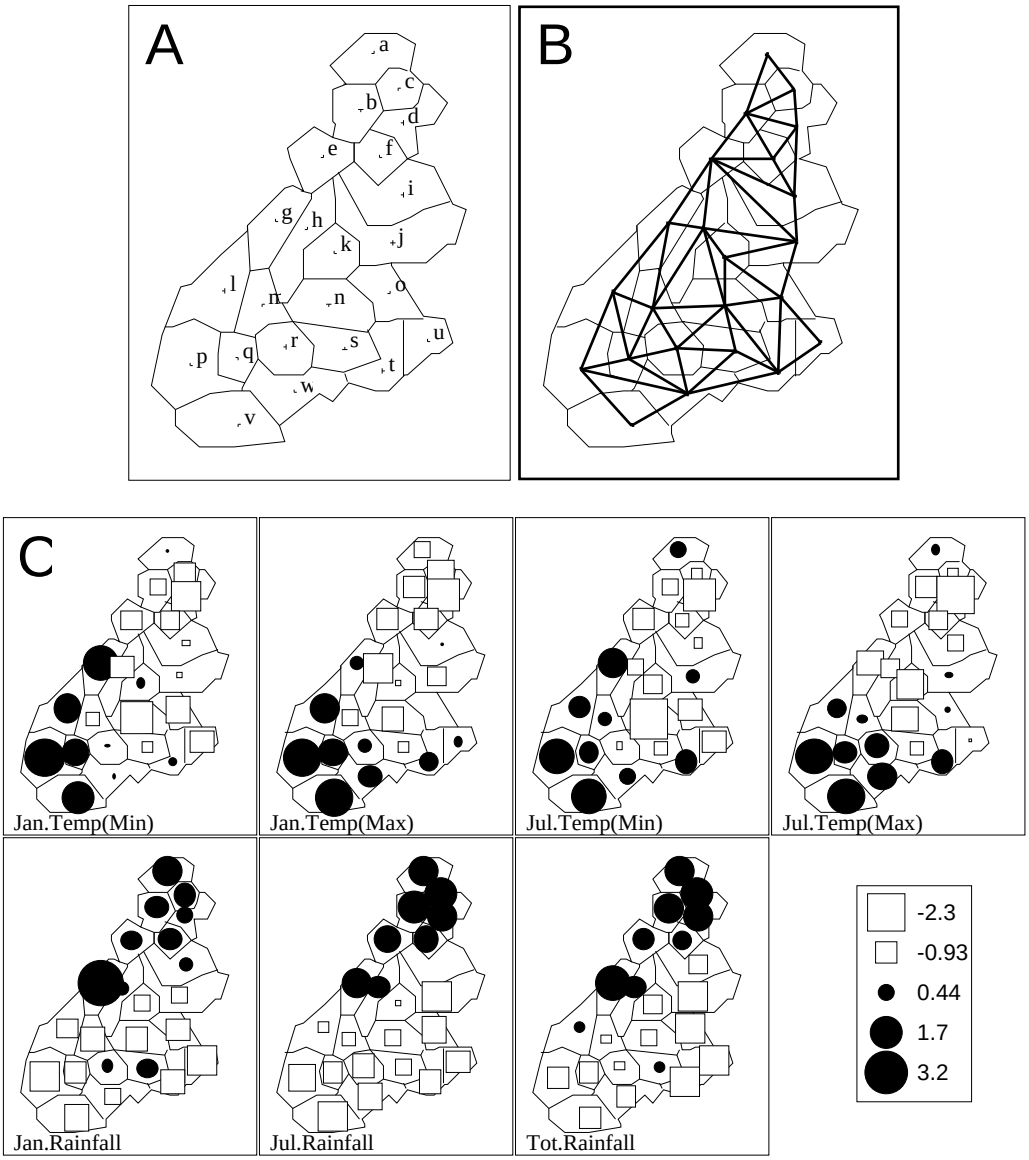


Figure 2.19: Exemple de graphiques obtenus avec le module Maps : options Labels (A), Neighboring graph (B), et Values (C)

Des cercles et des carrés peuvent être tracés sur une carte géographique avec l'option **Values**. Ils seront de taille proportionnelle aux valeurs contenues dans un fichier de données, et il est possible de réaliser des collections de cartes en fonction des colonnes de ce fichier (mais pas en fonction de groupes de lignes). Il est possible de tracer toutes les cartes d'une collection à la même échelle, ou de définir une échelle pour chaque carte. Les données peuvent, de plus, être remises à l'échelle entre 0 et 1 afin d'obtenir des cartes plus comparables.

Le module **Areas** est un module de cartographie surfacique. Il permet de tracer des polygones jointifs possédant un motif dont le niveau de gris est proportionnel aux valeurs mesurées dans chaque zone. Comme dans le module Maps, il est possible de réaliser des collections de cartes en fonction des colonnes du fichier de données et de choisir une échelle identique pour toutes les cartes ou non. Le nombre de niveaux de gris est paramétrable (huit au maximum), et il est possible d'affecter un label à chaque carte.

Les classes correspondant aux niveaux de gris sont calculées par une procédure de classification automatique qui maximise l'inertie inter-classe.

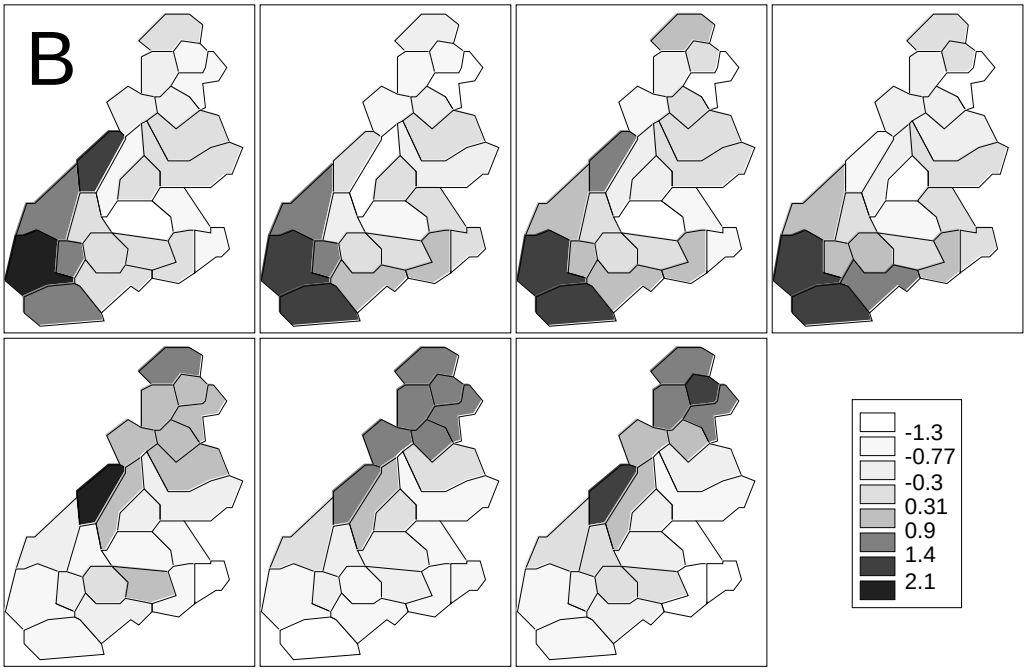


Figure 2.20: Exemple de graphiques obtenus avec le module Areas

Le module **Levels** permet de tracer des courbes de niveaux. Chaque courbe est, de plus, remplie d'un motif dont le niveau de gris correspond à la valeur de la courbe. Les courbes de niveaux sont tracées par interpolation aux noeuds d'un maillage régulier (le maillage est défini dans le module Digit). Une procédure de régression Lowess est utilisée pour réaliser l'interpolation. L'utilisateur doit donc indiquer le nombre de voisins utilisés dans cette régression. Le module GraphUtils permet d'orienter ce choix en calculant la somme des carrés des erreurs et en fournissant une carte des résidus de régression. Plus le nombre de voisins utilisés est grand et plus les courbes de niveau seront lisses, le problème étant de trouver un minimum dans la relation entre la somme des carrés des erreurs et le nombre de voisins utilisés.

La comparaison des trois techniques de cartographie (cercles et carrés, polygones jointifs, et courbes de niveaux) montre que, de façon assez surprenante, la cartographie surfacique est la moins apte à mettre en évidence une autocorrélation spatiale. Les courbes de niveau y sont au contraire particulièrement bien adaptées, et les cartes par cercles et carrés représentent un bon compromis en permettant de mettre relativement bien en évidence à la fois des gradients et des partitions. L'utilisation conjointe des deux (courbes de niveau plus cercles et carrés) peut se révéler nécessaire (Thioulouse *et al.* 1995a).

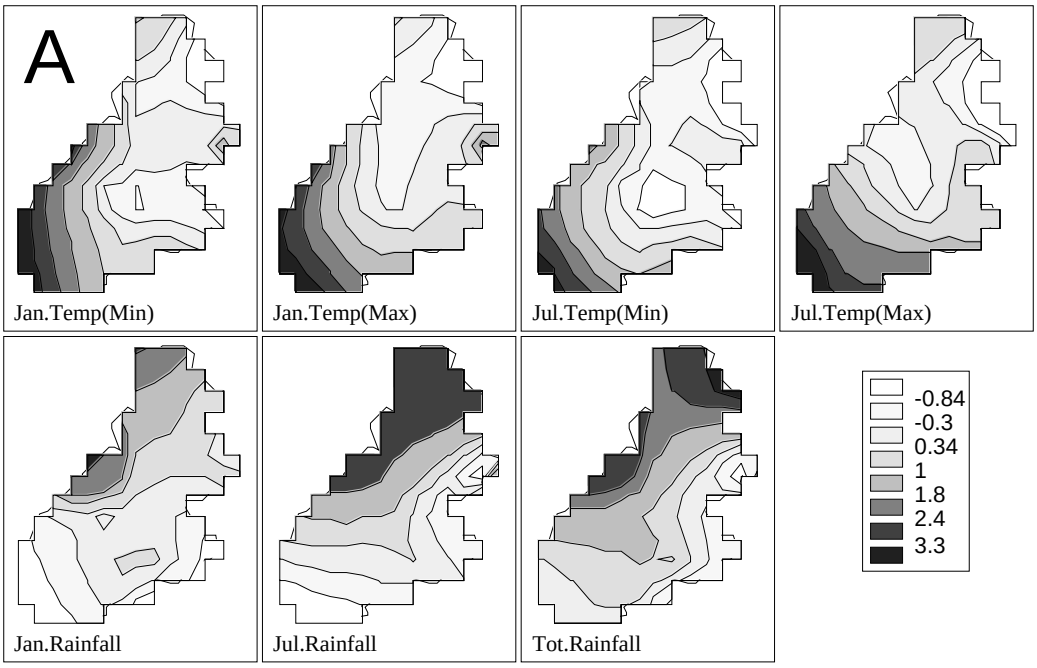


Figure 2.21: Exemple de graphiques obtenus avec le module Levels

1.3.6. *Conclusion*

Plusieurs modules restent encore à écrire, en particulier pour la représentation de données de type distributionnel et de tableaux. Mais la diversité des représentations graphiques déjà disponibles permet de choisir les illustrations en tenant compte à la fois de la diversité des propriétés mathématiques des méthodes d'analyse de données et de la multiplicité des situations biologiques rencontrées. De plus, les 14 modules déjà existants ne sont pas des blocs monolithiques figés. Il est facile d'y rajouter ou d'en retirer des options grâce à la structure de programmation qui a été décrite précédemment (§1.1).

Deux modules n'ont pas été décrits. Le premier est ADEPict, un utilitaire d'affichage. Il permet, lorsque l'utilisateur clique sur un fichier PICT créé par un module graphique d'ADE-4 d'afficher le graphique, ainsi que les paramètres qui avaient été utilisés pour le tracer (minimum et maximum des abscisses et des ordonnées, facteur G, etc.). Ce module, ainsi que les modules ADEBin (pour l'affichage du contenu d'un fichier de données) et ADETrans (pour l'importation/exportation des fichiers textes) supportent le "Glisser-Déposer" (Drag and Drop), c'est à dire que l'utilisateur peut faire glisser l'icône d'un fichier sur celle du module pour l'ouvrir.

Le second module est ADEScatters, dont les fonctionnalités seront décrites plus loin (§3)

1.4. La politique de diffusion

Nous avons choisi de distribuer ADE-4 gratuitement. Ce choix a fait l'objet de discussions répétées dans le groupe de développeurs. Il s'est avéré finalement qu'aucun d'entre nous n'avait le temps (ou l'envie) de s'occuper d'un autre mode de diffusion. Une solution alternative serait en effet d'établir un contrat entre le CNRS et une société qui prendrait en charge la diffusion commerciale. Cette solution implique un certain investissement dans la recherche de cette société et dans l'établissement du contrat. Le

principal bénéfice de cette solution serait probablement un accroissement de la diffusion, en particulier dans le domaine privé et/ou industriel.

L'ensemble du logiciel est donc disponible gratuitement sur le réseau internet par FTP (file transfer protocol) anonyme sur biom3.univ-lyon1.fr dans le répertoire /pub/mac/ADE/ADE4. De plus, une page WWW (world-wide web) existe à l'URL (uniform resource locator) suivant: http://biomserv.univ-lyon1.fr/ADE-4.html. Elle propose une présentation détaillée du logiciel et offre la possibilité de récupérer les modules, en groupe ou individuellement.

Une version sur disquette est également disponible, avec un exemplaire de la documentation complète sur papier (six volumes, dont deux en anglais) pour une somme forfaitaire destinée à couvrir les frais de duplication des disquettes, de photocopie de la documentation et d'affranchissement postal.

2. Les réseaux informatiques

2.1. Historique

L'équipement informatique de l'URA 243 a subi une évolution constante et la part de l'équipement réseau dans le budget global a augmenté rapidement ces dernières années. Ce développement a d'ailleurs constitué une partie importante de mon activité. Les premières étapes dans ce domaine ont commencé par l'installation de liaisons série sur l'Eclipse S/140 ainsi que sur le Vax 730, ce qui nous a permis de répartir des terminaux dans les différents bureaux (consoles vt100, Secapa, et Macintosh avec un logiciel d'émulation de terminal).

L'apparition des stations de travail au laboratoire en 1987 (Sun 3) et 1988 (Sun 4), qui se substituèrent aux mini-ordinateurs, entraîna une progression dans ce sens, avec par exemple l'achat d'un multiplexeur 16 voies. Cette solution rencontra rapidement ses limites et les débuts du véritable réseau Ethernet se situent à la même époque (1988). Celui-ci se généralisa rapidement : le raccordement des Macintosh (qui étaient déjà connectés en réseau LocalTalk) au réseau Ethernet grâce à un boîtier FastPath fut effectué en 1988. Quatre autres boîtiers furent achetés en 1990 pour relier les réseaux de Macintosh des laboratoires de Biologie moléculaire et cellulaire (UMR 106), d'Écologie microbienne (URA 1977), d'Écologie des eaux douces et des grands fleuves (URA 1974) et le secrétariat de l'institut IASBSE.

C'est aussi de cette année là que date le raccordement de notre réseau interne à ROCAD et donc à l'Internet. Le passage d'un réseau en câble coaxial fin à un réseau en paire torsadée (10baseT) avec étoiles actives a eu lieu en 1994. Nous disposons actuellement, pour le seul laboratoire de Biométrie, d'environ 80 ports Ethernet 10baseT, et la saturation de la bande passante nous conduit à envisager le passage à un réseau Ethernet commuté.

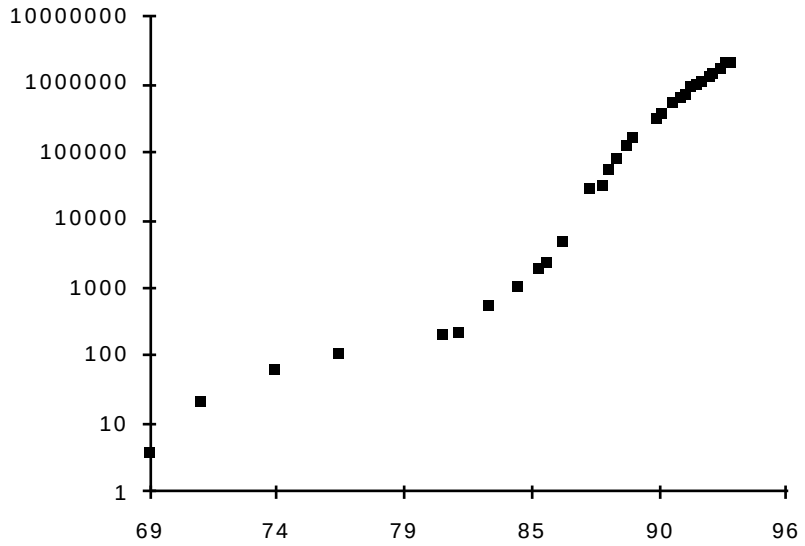


Figure 2.22: Croissance du nombre total d'ordinateurs connectés au réseau Internet (1969-1994, échelle logarithmique). Source : Internet Business Center (http://tig.com/IBC/)

L'accès au réseau Internet est aujourd'hui devenu une nécessité pour un grand nombre de chercheurs, ne serait-ce que pour l'utilisation du courrier électronique. Les services offerts, en particulier dans le domaine de la documentation et de la diffusion de logiciels sont souvent irremplaçables. Toutes les disciplines scientifiques ne sont pas affectées au même titre. La biologie moléculaire, du fait de ses besoins très importants en matière d'informatique (stockage des banques de séquences) se situe en tête des disciplines biologiques, mais on assiste à une augmentation importante des services offerts aussi en écologie et dans les sciences de l'environnement, ainsi que dans le domaine des mathématiques et des statistiques.

L'augmentation exponentielle du nombre d'ordinateurs connectés au réseau Internet (figure 2.22) est rendue possible par la standardisation des protocoles de communication utilisés. Le protocole de base est TCP/IP (transmission control protocol et internet protocol, correspondant aux couches 3 et 4 du modèle ISO), et il est disponible sur pratiquement tous les types d'ordinateurs actuels, y compris les micro-ordinateurs PC compatible et Macintosh. Les protocoles de niveaux supérieurs (FTP : file transfer protocol, SMTP : simple mail transfer protocol, NNTP network news transfer protocol) sont eux aussi standardisés. Les applications qui les mettent en oeuvre sont ainsi disponibles sur toutes les plates-formes ce qui permet de les employer indépendamment du modèle d'ordinateur utilisé.

En 1990, l'apparition du protocole HTTP (hypertext transfer protocol), sur lequel sont basés les services WWW (world-wide web), a permis un saut qualitatif dans la présentation des informations disponibles, en autorisant la consultation de texte, d'images, de sons, etc., à travers des liens hypertexte (Berners-Lee *et al.* 1992). Des logiciels clients (capable d'interroger les serveurs WWW) sont apparus rapidement sur stations unix, PC et Macintosh.

Les documents WWW sont écrits en langage HTML (hypertext markup language). Ce langage permet de définir des éléments d'interface utilisateur simples comme des menus, des boutons et des champs texte éditables, qui sont suffisants pour créer des IUG fonctionnelles. J'ai utilisé ces possibilités pour créer un service WWW appelé NetMul. La présentation qui en est faite ici est basée sur un article soumis actuellement à la revue **Computational Statistics and Data Analysis**.

2.2. NetMul

NetMul permet d'utiliser un sous-ensemble d'ADE-4, limité actuellement aux trois analyses de base (ACP, AFC, ACM) avec les aides à l'interprétation numériques correspondantes (analyse d'inertie, éléments supplémentaires, reconstitution de données), auxquelles j'ai rajouté récemment l'analyse en coordonnées principales (PCO, pour l'analyse de matrices de distances), l'analyse de co-inertie, les analyses inter/intra et discriminantes, et les analyses locales et globales. L'URL de NetMul est la suivante : http://biomserv.univ-lyon1.fr/NetMul.html. Le serveur est actuellement une station de travail Sun (Sparcstation 20).

Le principe d'utilisation est le suivant : l'utilisateur commence par transférer son fichier de données sur le serveur dans le répertoire /pub/NetMul/data. Il peut effectuer cette opération soit par FTP anonyme sur biom3.univ-lyon1.fr avec n'importe quel logiciel de transfert FTP, soit par un simple copier/coller dans le client Web. Afin de limiter l'encombrement, ce répertoire est vidé automatiquement tous les matins, et un quota maximum de 10 mega-octets est imposé.

On peut ensuite se connecter au service NetMul en utilisant un client Web. La page d'accueil (figure 2.23) présente les fonctionnalités du serveur et son mode d'utilisation. Un menu déroulant permet de choisir l'option de calcul (figure 2.24), et le bouton "Let's go !" lance l'exécution d'un programme (NetMul-query) sur le serveur.

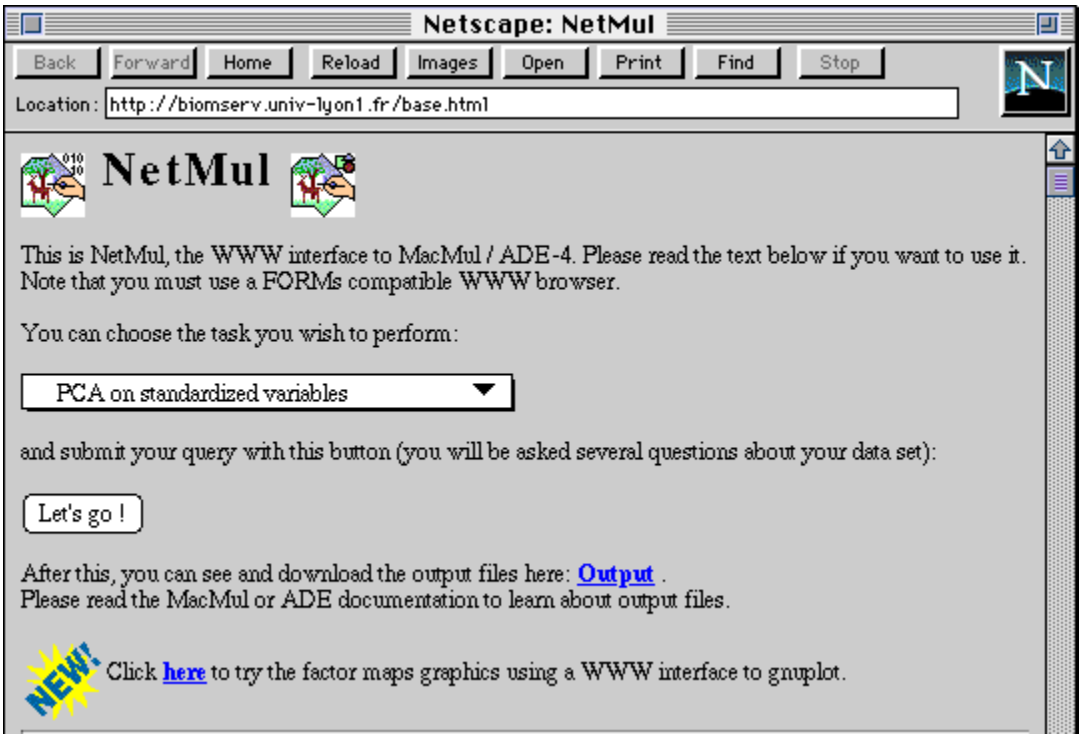


Figure 2.23: Page d'accueil de NetMul (début)

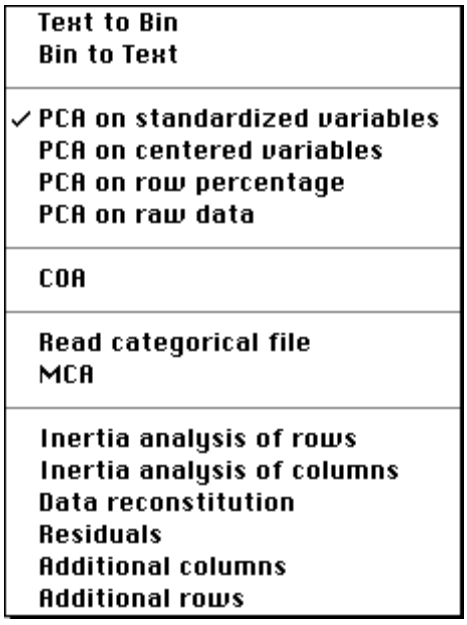


Figure 2.24: Menu déroulant de NetMul

NetMul-query génère un document HTML contenant les éléments d'interface correspondant à l'option choisie dans le menu, et envoie le formulaire résultant au client qui l'affiche (figure 2.25). L'utilisateur remplit ce formulaire avec les informations relatives à ses données et aux options d'analyse qu'il désire. Le bouton "Submit Query" lui permet ensuite de lancer l'exécution des calculs. Le programme NetMul-query récupère les informations données par l'utilisateur, lance l'exécution du programme de calcul et renvoie au client le listing d'exécution.

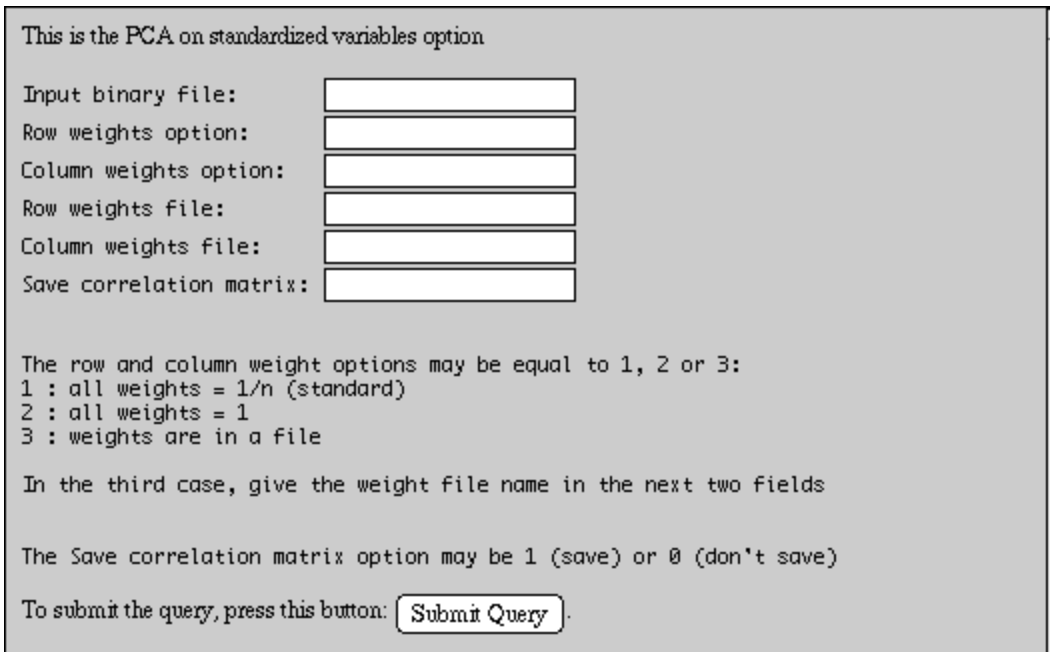


Figure 2.25: Dialogue de préparation d'une ACP normée. Ce dialogue est généré automatiquement par le programme NetMul-Query

Tous les fichiers de sortie sont créés dans le répertoire initial (/pub/NetMul/data) et l'utilisateur peut donc les récupérer s'il le désire, pour effectuer des représentations graphiques localement. Il peut également réaliser un plan factoriel simple à l'aide de l'interface avec le programme gnuplot (figure 2.26). Ce service de représentation graphique a été intégré dans un système plus complexe, appelé WWW-Query, qui a été développé par Guy Perrière (chercheur à l'URA 2055). WWW-Query permet l'accès à un service d'interrogation de banques de séquences de biologie moléculaire, couplé aux méthodes d'analyse multivariée disponibles dans NetMul. Ce système est présenté en détail dans un article publié dans la revue **Computer Applications in the Biosciences** (Perrière & Thioulouse 1996).

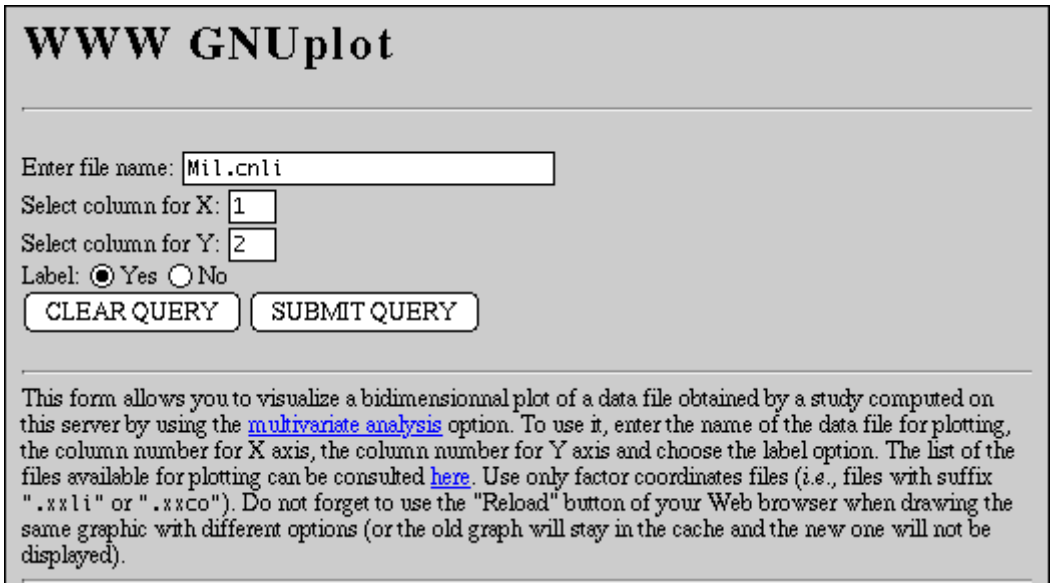


Figure 2.26: Dialogue de préparation à l'utilisation de gnuplot

Le graphique obtenu (figure 2.27) est une image GIF qui peut être copiée et utilisée ensuite dans d'autres logiciels.

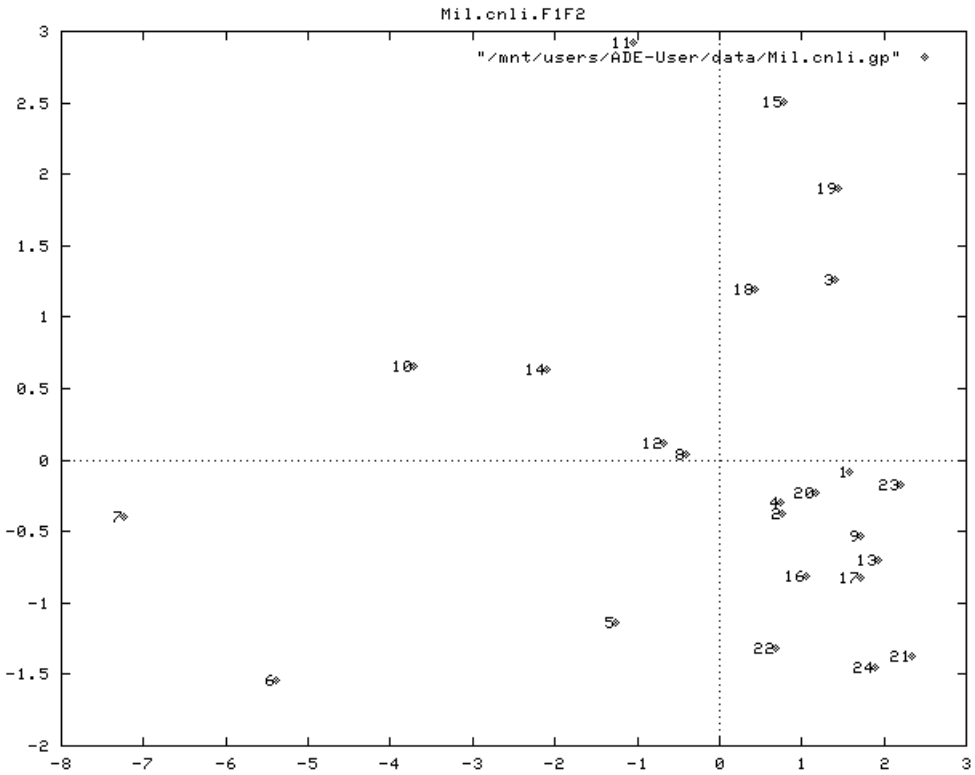


Figure 2.27: Plan factoriel F1 x F2 généré par le programme gnuplot

2.3. Discussion - Perspectives

Le principal intérêt de NetMul (et de WWW-Query) est la portabilité. Il peut en effet être utilisé sur n'importe quel ordinateur disposant d'un logiciel client Web, ce qui représente la majorité des machines existantes. De plus, les logiciels clients sont gratuits (sauf dans certains cas : par exemple, l'utilisation du logiciel Netscape n'est gratuite que pour les étudiants et le personnel des établissements d'enseignement ou des organisations caritatives).

Les méthodes qui sont disponibles actuellement dans NetMul sont encore restreintes, mais il est très facile d'en ajouter d'autres. En effet, il suffit d'extraire des modules d'ADE-4 le code de calcul d'une méthode, et de le recompiler sur le serveur, en rajoutant au programme NetMul-query la partie destinée à générer le formulaire de l'interface utilisateur.

Un autre point important est le fait que les calculs sont réalisés sur l'ordinateur serveur. Il n'est donc pas nécessaire de posséder un ordinateur puissant pour utiliser NetMul, même pour des jeux de données de grande taille. NetMul est actuellement disponible librement depuis n'importe quel ordinateur de l'Internet, mais il est aisé de restreindre son utilisation de diverses façons, par exemple en fixant un nombre maximum d'utilisateurs simultanés, ou bien le domaine de l'ordinateur client (pour créer un serveur à l'échelle d'un laboratoire ou d'un campus).

La limitation la plus importante de NetMul est la faiblesse des possibilités de représentations graphiques. L'évolution rapide des fonctionnalités offertes tant par les logiciels serveurs que par les clients donne à penser que cette limitation sera très vite dépassée. La voie la plus prometteuse actuellement semble être le langage Java, un langage orienté objet dérivé des langages C et C++. Son originalité repose sur la notion d' "Applets", petites applications compilées par le serveur et transmises sous forme de code objet intermédiaire au client. Celui-ci exécute ces applications sur l'ordinateur

local, ce qui lui permet de disposer de fonctionnalités originales et évolutives (graphiques, animations, interactivité, etc.) Le code objet intermédiaire est indépendant de l'architecture, et il peut donc être exécuté sur n'importe quel ordinateur, mais il nécessite un interpréteur, qui est intégré dans le logiciel client.

La puissance de ces outils nous permet d'envisager une version d'ADE–4 dans laquelle l'interface utilisateur sera totalement implémentée à travers le réseau, ce qui lui permettra d'être utilisée à partir de n'importe quel modèle d'ordinateur.

3. Graphiques interactifs

3.1. Introduction

Les progrès réalisés en matière d'IUG sont importants et ils apportent une amélioration significative de la facilité d'utilisation de tous les logiciels. Dans le domaine des logiciels graphiques, le bénéfice est encore plus net car le nombre de paramètres que l'utilisateur doit fournir est plus élevé, et les sorties du programme sont elles aussi de type graphique. L'utilisation des graphiques en analyse de données n'est pas une démarche formalisable : une grande liberté d'action doit être laissée à l'utilisateur. **L'interactivité** est donc une nécessité, et nous avons vu comment l'IUG des modules graphiques d'ADE-4 permettait d'aborder ce problème.

Afin de développer un peu plus les possibilités d'interactivité dans l'exploration graphique des analyses multivariées, j'ai mis au point un module de "plan factoriel interactif", appelé ADEScatters. Ce paragraphe est consacré à la description des possibilités offertes par ce module. Il reprend en partie et développe un article publié dans la revue **Computational Statistics**. Il est intéressant de noter que le CISIA (Centre International de Statistique et d'Informatique Appliquées) a également développé récemment un programme d'exploration interactive des plan factoriels, qui s'intègre dans la famille du logiciel SPAD. Ce logiciel s'appelle SPAD•GF, et il permet d'analyser les plans factoriels issus de SPAD•N et de SPAD•T. Dans le même domaine, notons l'existence du logiciel AMADO, également commercialisé par le CISIA, qui permet la ré-écriture graphique de tableaux à la façon du module Tables d'ADE-4 (§1.3.4, figure 2.16). Par ailleurs, le groupe "Logiciels" de l'ASU (Association pour la statistique et ses utilisations) a vu la création en 1994 du groupe "Cercle Factoriel", qui doit s'intéresser "aux fonctionnalités nécessaires à une bonne exploitation statistique de résultats d'analyses factorielles". L'utilisation du logiciel SAS, si elle est compréhensible de la part de statisticiens professionnels, ne facilitera sans doute pas la diffusion de leurs recommandations auprès des utilisateurs biologistes.

3.2. ADEScatters : principales fonctionnalités

Il existe plusieurs difficultés dans l'interprétation des plans factoriels classiques. La plus simple concerne la superposition des points lorsque le nombre d'éléments (lignes ou colonnes) est grand. La solution classique a consisté à éditer une liste des points superposés, ce qui n'est qu'un pis-aller, introduisant une étape supplémentaire dans la démarche d'interprétation : l'utilisateur est obligé de rechercher dans une (ou plusieurs) liste(s) les points qui l'intéressent. De plus, cette solution interdit une approche globale dans l'interprétation du plan factoriel : il est difficile de se faire une idée de la densité des points dans une région donnée du plan si on n'a aucune idée du nombre de points superposés.

Une autre difficulté souvent rencontrée est la localisation d'un élément particulier dans le plan : même si le nombre d'éléments est raisonnable (inférieur à 100), il est souvent difficile de déterminer la position d'un élément choisi a priori. Ceci est pourtant souvent nécessaire lors de l'étape d'interprétation des axes : il faut pouvoir localiser rapidement sur le plan factoriel un point dont on sait qu'il a des caractéristiques particulières, par exemple du point de vue du protocole expérimental.

La troisième difficulté est le retour au données. Cette étape de l'interprétation consiste à rechercher, dans le tableau de données, quelles sont les caractéristiques prises en compte par les facteurs. On recherche par exemple souvent les points possédant des valeurs extrêmes pour certaines variables, en cherchant à corréler cette caractéristique avec une position extrême dans le plan factoriel. Cette opération faite à la main devient rapidement pénible, voire impossible pour les grands tableaux. Elle peut être partiellement automatisée en utilisant des graphiques, par exemple le tracé, sur le plan factoriel, de cercles et de carrés de taille proportionnelle aux données. Mais cette stratégie est limitée elle aussi par la taille des tableaux de données.

ADEScatters a été écrit pour essayer de résoudre ces problèmes. Comme les autres modules graphiques d'ADE-4, il peut être utilisé pour réaliser des collections de graphiques, et les trois fenêtres de paramètres sont utilisables. En fait, il fonctionne de la même façon que le module Scatters, mais il ne possède que l'option Labels. La première possibilité caractéristique d'ADEScatters est la recherche d'un élément sur le plan factoriel : l'options **Find...** du menu **Edit** permet à l'utilisateur d'indiquer le numéro de l'élément qu'il recherche. Cet élément apparaît alors sur le plan factoriel, entouré d'un cadre, en caractères gras, et clignote cinq fois (figure 2.28). L'option **Find label...** réalise la même opération, mais l'utilisateur indique le label de l'élément recherché au lieu de son numéro, ce qui évite d'avoir à chercher ce numéro quand on connaît le label.

Une autre possibilité est le zoom, qui peut être réalisé de plusieurs façons. La plus simple consiste à cliquer en un point du plan factoriel en maintenant les touches commande et majuscule appuyées. La zone entourant le point choisi est alors agrandie d'un facteur égal à 1,5. Si par contre ce sont les touches commande et option qui sont maintenues appuyées, le zoom est inversé, et la zone est réduite au lieu d'être agrandie. Il est ainsi facile de zoomer rapidement sur une zone du plan factoriel où de nombreux points sont superposés afin de déterminer quels sont ces points. L'autre façon consiste à sélectionner une zone du plan avec la souris, tout en maintenant la touche commande appuyée (figure 2.28). La zone sélectionnée est alors retracée en occupant toute la fenêtre. Le grossissement du zoom est donc dans ce cas proportionnel à la surface de la zone sélectionnée. L'utilisateur peut par exemple sélectionner une zone du plan particulièrement dense et zoomer uniquement sur cette zone. Un simple clic n'importe où sur le graphique en maintenant la touche commande appuyée recadre le plan factoriel en reprenant comme limites le minimum et le maximum des abscisses et des ordonnées. Le zoom sur une zone sélectionnée à la souris peut être combiné avec la recherche des éléments présents dans cette zone : il suffit de maintenir les touches commande et option appuyées pendant la sélection de la zone pour afficher à l'écran la liste des points (triés par ordre croissant) dans une fenêtre spéciale (figure 2.28). Toutes ces opérations sont aussi réalisables en mode collection : tous les graphiques élémentaires de la collection sont alors remis à l'échelle simultanément.

La troisième catégorie de possibilités offertes par ADEScatters est la représentation des données sur les plans factoriels. Deux types de représentations sont possibles, et dans les deux cas l'utilisateur doit cliquer sur l'élément pour lequel il veut avoir des renseignements. Prenons l'exemple de l'AFC d'un tableau d'abondances faunistiques : en lignes figurent n relevés et en colonnes p espèces. ADEScatters est utilisé de la même façon que le module Scatters pour tracer le plan factoriel des relevés après une AFC.

Afin d'interpréter la signification du premier axe, l'utilisateur peut cliquer sur chaque point du plan factoriel représentant un relevé tout en maintenant la touche contrôle

appuyée. Une petite fenêtre temporaire apparaît alors, contenant un diagramme en bâtons qui montre l'abondance des différentes espèces dans ce relevé (figure 2.29, en bas). La hauteur de chaque bâton est proportionnelle aux valeurs contenues dans le tableau effectivement traité par l'AFC, c'est à dire les abondances relatives centrées ($p_{ij}/p_{i.}p_{.j}$-1). L'ordre des espèces dans cette représentation graphique est simplement l'ordre dans lequel elles apparaissent dans le tableau de données.

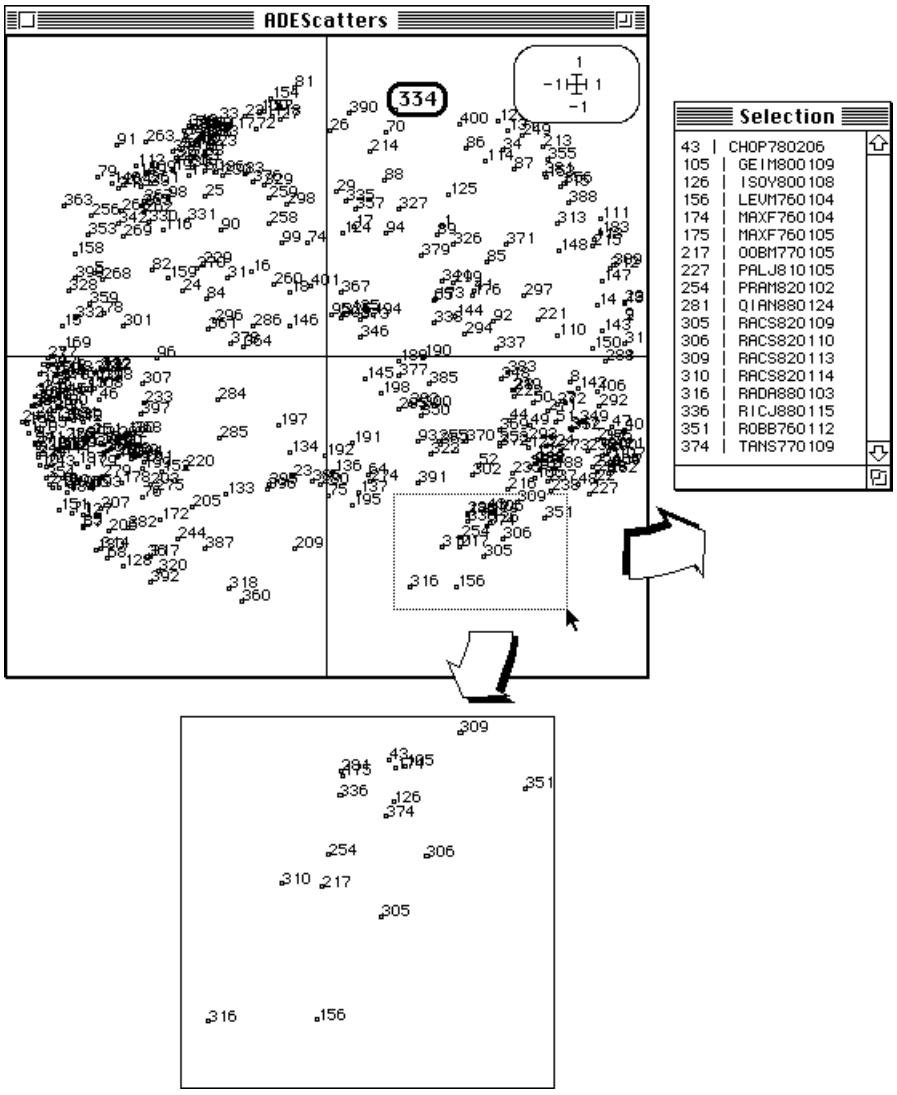


Figure 2.28: Interface du module ADEScatters,
montrant comment est affiché un élément après sa recherche (élément numéro 334, en haut de la fenêtre), ainsi que la possibilité de zoomer sur une zone (en bas)
et d'obtenir la liste des éléments présents dans une zone sélectionnée (à droite).

Afin de mieux faire ressortir la signification des axes factoriels, il est aussi possible d'ordonner les espèces en fonction de leurs coordonnées factorielles le long d'un des axes de l'AFC. Pour cela, il suffit de maintenir à la fois la touche contrôle et la touche option appuyées en cliquant sur un relevé. Une fenêtre de dialogue permet alors d'indiquer l'axe à utiliser pour ré-ordonner les espèces dans le diagramme en bâtons (figure 2.29, en bas).

Le deuxième type de représentation graphique interactive est la carte duale, sur laquelle sont tracés des cercles et des carrés de taille proportionnelle aux données (figure 2.29, à droite). Cette technique s'inspire du biplot (Gabriel 1971) : le graphique tracé dans la

petite fenêtre temporaire n'est plus un diagramme en bâtons, mais le plan factoriel des espèces. Pour chaque espèce un cercle (valeurs positives) ou un carré (valeurs négatives) représente l'abondance relative centrée dans le relevé sur lequel l'utilisateur a cliqué.

De plus, en amenant le pointeur de la souris sur les cercles et les carrés, l'utilisateur peut connaître le numéro de chaque espèce : ce numéro apparaît et reste affiché tant que la souris n'est pas déplacée (figure 2.29, à droite : élément numéro 12)

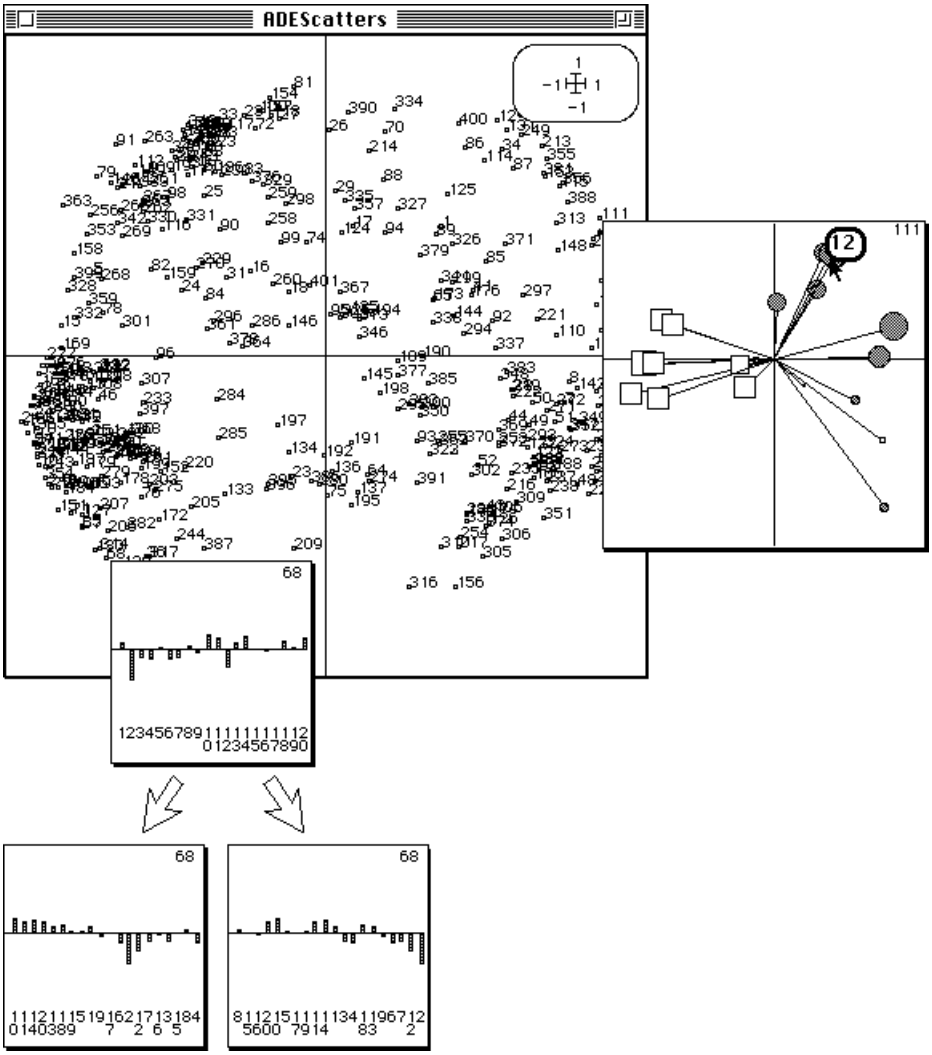


Figure 2.29: Représentation graphique interactive des données sur le plan factoriel dans le module ADEScatters. En bas, représentation par des diagrammes en bâtons : les deux flèches indiquent la possibilité de ré-ordonner les valeurs en fonction de leur coordonnée sur l'axe 1 (à gauche) ou sur l'axe 2 (à droite). Dans la partie droite de la figure, la petite fenêtre temporaire montre le plan factoriel dual, sur lequel sont tracés des cercles et des carrés de taille proportionnelle aux données. Le numéro d'un élément peut être connu en amenant le pointeur sur le cercle ou le carré (élément numéro 12 sur la figure).

En cliquant sur plusieurs relevés successivement, il est ainsi facile de voir quelles sont les espèces abondantes dans ces relevés, et donc de découvrir la signification des axes.

Symétriquement, il est possible de partir du plan factoriel des espèces : l'utilisateur clique alors sur une espèce, et c'est l'abondance de cette espèce dans les différents relevés qui est utilisée pour tracer les graphiques.

Les mêmes possibilités sont offertes en ACP et en ACM, et la relation de proportionnalité entre les données et la taille des cercles et des carrés peut, comme dans le module Scatters, être réglée par l'intermédiaire du facteur G (fenêtre Min/Max).

3.3. Conclusion

Les avantages apportés par l'interactivité dans l'interprétation des analyses multivariées sont importants. Ils sont manifestes dans le domaine des représentations graphiques, et le module ADEScatters tente d'en tirer avantage de façon directe. L'interactivité dans les modules de calcul d'ADE-4 est aussi un facteur à prendre en compte : elle autorise une démarche réellement exploratoire. Il y a dix ou quinze ans, la mise en oeuvre d'une analyse simple demandait plusieurs heures. Quelques minutes suffisent maintenant. Ces progrès sont dus bien sûr à l'augmentation des performances des ordinateurs, mais aussi à la place importante accordée à l'interface utilisateur des logiciels. L'arrivée de nouvelles techniques, comme la réalité virtuelle, permettra peut-être d'accélérer encore cette évolution. Le module QuickTime VR, distribué gratuitement par Apple, permet déjà de se déplacer librement dans un environnement virtuel affiché dans une fenêtre à l'écran, et de manipuler "à la souris" des objets tridimensionnels. L'exploration dynamique d'un nuage de points dans un espace vectoriel (donc évidemment virtuel !) est tout à fait réalisable.

Chapitre 3
Implications biologiques

L'utilisation des méthodes statistiques et des outils logiciels évoqués dans les chapitres précédents m'a conduit à l'obtention de résultats biologiques dans des domaines très divers, généralement à la suite de collaborations avec des collègues spécialistes de ces domaines. Il s'agit principalement des disciplines suivantes: biologie moléculaire, biologie des populations, agronomie, écotoxicologie, et écologie.

1. Biologie moléculaire

Je n'ai commencé à m'intéresser directement à l'analyse de données en biologie moléculaire que récemment, à travers deux collaborations avec Jean Lobry (Thioulouse & Lobry, 1995) et Guy Perrière (Perrière & Thioulouse, 1995), tous deux membres du laboratoire de Biométrie, Génétique et Biologie des Populations. La biologie moléculaire, avec l'énorme masse de données qu'elle accumule à un rythme de plus en plus rapide, devrait de toute évidence être un domaine privilégié d'utilisation des techniques d'analyse multivariée. On n'assiste cependant pas à une généralisation importante de leur usage, qui reste par exemple beaucoup moins développé qu'en écologie (cette discipline bénéficiant il est vrai d'une plus longue tradition dans ce domaine).

Le premier travail (Thioulouse & Lobry, 1995) concerne l'étude des relations entre la composition des protéines en acides aminés et les propriétés physico-chimiques de ces acides aminés. Il s'agit là d'un problème qui avait déjà été abordé par divers auteurs (Sneath 1966, Sjöström & Wold 1985, Kidera *et al.* 1985, Nakai *et al.* 1988), qui avaient tous trouvé un certain nombre de relations en utilisant des méthodes et des jeux de données variés. Afin d'aborder le problème de façon efficace, nous avons utilisé un jeu de données de grande dimension : 999 protéines d'*Escherichia coli* et 402 propriétés physico-chimiques mesurées sur les 20 acides aminés naturels, ainsi qu'une méthode de couplage bien adaptée, l'analyse de co-inertie (Chessel & Mercier 1993, Dolédec & Chessel 1994).

Les résultats que nous avons obtenus montrent qu'il existe effectivement des relations fortes entre les propriétés physico-chimiques des acides aminés et la composition des protéines. La plus importante de ces relations est l'hydrophobicité. Elle est due aux contraintes physico-chimiques liées à la localisation des protéines, qui peuvent se trouver dans un milieu polaire (dans le cytoplasme ou dans le milieu extra-cellulaire) ou apolaire (la bi-couche de phospholipides membranaires). Ces contraintes font que les protéines membranaires sont plus riches en acides aminés ayant des propriétés hydrophobes (figure 3.1: Ile, Leu, Met, Phe, Trp).

Une autre relation importante est liée au niveau d'expressivité des gènes protéiques. Les concentrations des ARN de transfert majeurs varient selon les espèces, et il n'existe donc pas de relation directe entre le niveau d'expressivité d'une protéine et les propriétés physico-chimiques des acides aminés. L'analyse de co-inertie met cependant en relation un faible niveau d'expressivité avec la tendance des acides aminés à former des hélices alpha, et inversement un haut niveau d'expressivité avec la tendance à former des

feuillets béta. Cette relation peut être interprétée de façon indirecte : les gènes faiblement exprimés sont souvent des gènes de régulation, qui interagissent avec l'ADN par le biais de structures en hélice alpha, alors que les gènes hautement exprimés comprennent les gènes codant pour les protéines membranaires, qui sont riches en feuillets béta.

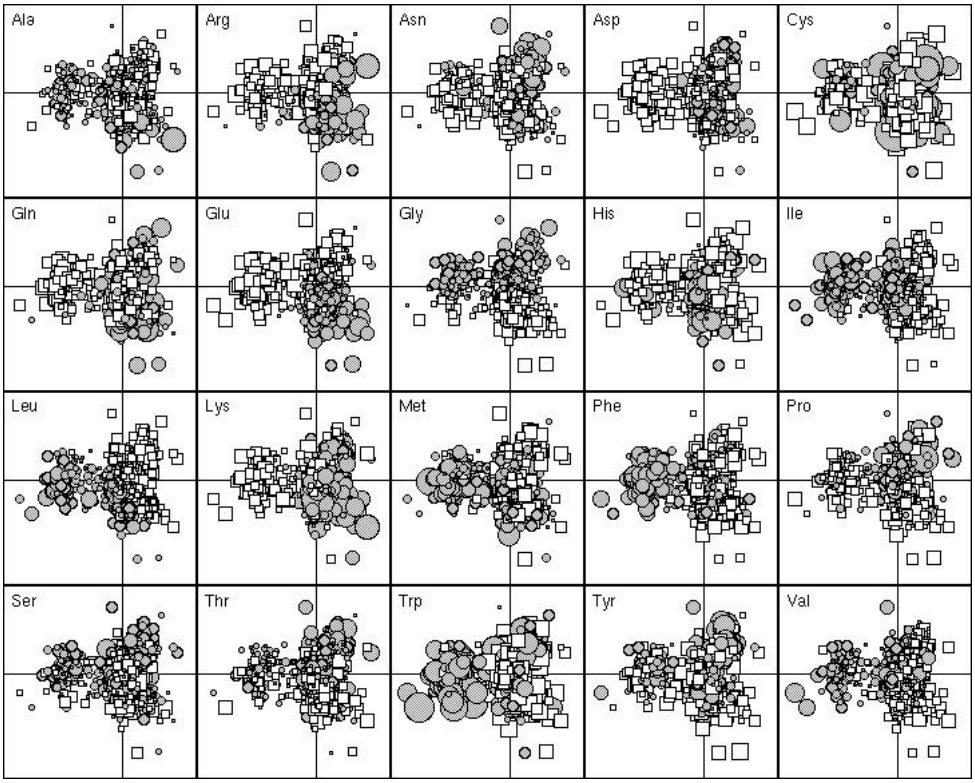


Figure 3.1: Composition en acide aminés des 999 protéines de E. coli représentée sur le plan F1× F2. Les protéines situées sur la gauche du plan sont des protéines membranaires intégrales et on peut vérifier qu'elles contiennent une proportion importante d'acides aminés hydrophobes.

Enfin, la troisième relation observée relie l'aromaticité et le poids moléculaire des acides aminés avec la composition des protéines. Les contraintes sélectives imposées aux protéines en termes de coût énergique pour la biosynthèse font que les acides aminés aromatiques tendent à être évités, sauf lorsqu'ils sont indispensables (participation au site actif des protéines enzymatiques).

La composition des protéines en acides aminés est de toute évidence soumise à une pression de sélection élevée, en particulier pour les acides aminés participant au site actif des protéines enzymatiques. Mais ces derniers sont peu nombreux, et leur contribution à la composition globale de la protéine est très faible. Les relations détectées par l'analyse de co-inertie se situent bien au niveau de la composition globale, et à ce niveau aussi une pression de sélection est observée, dont la principale matérialisation est l'adéquation de la composition de la protéine à son environnement physico-chimique.

Le second article (Perrière & Thioulouse, 1995) concerne le couplage entre un logiciel d'interrogation des banques de séquences d'acides nucléiques (WWW-Query) et le service d'analyse multivariée en ligne NetMul, tous deux utilisables à travers le réseau Internet (Cf. chapitre 2, § 2.2). Nous avons utilisé ce système pour réaliser deux études, l'une portant sur l'usage du code (i.e., le fait que les codons synonymes ne sont pas utilisés de façon équivalente) chez *Haemophilus influenzae* et la seconde sur la phylogénie du gène de l'hormone de croissance de 70 espèces.

La première analyse a été réalisée sur le génome complet d'*H. influenzae* (il s'agit de la première étude de l'usage du code sur un génome complet). La fréquence absolue des 61 codons dans les 1680 gènes a été calculée, et une AFC a été effectuée sur ce tableau. Le premier facteur isole principalement les protéines ribosomales (gènes hautement exprimés) et le second facteur isole les protéines membranaires (figure 3.2).

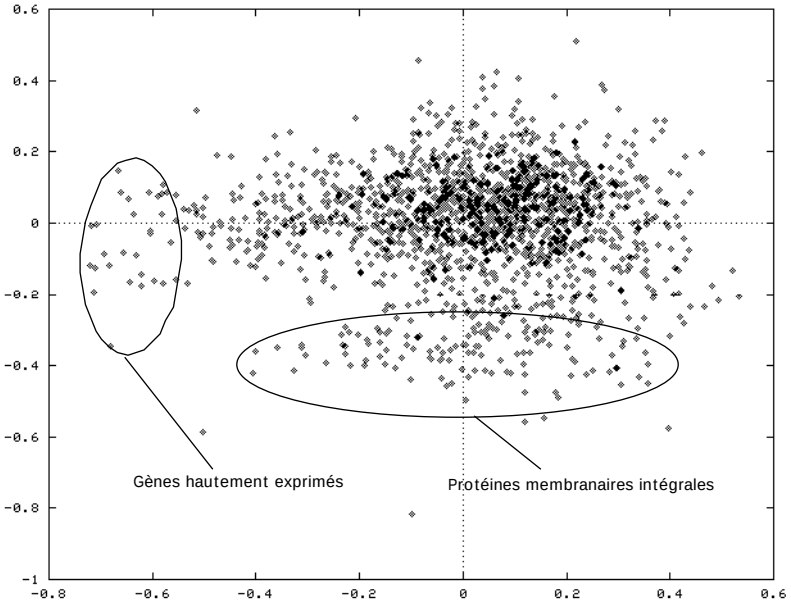


Figure 3.2: Plan F1× F2 de l'AFC de la composition en codons des gènes d'H. influenzae

En utilisant les 30 gènes ayant les coordonnées les plus négatives sur le F1, nous avons calculé l'indice moyen d'usage du code (Sharp & Li 1987). Il est ainsi possible de caractériser les gènes ayant le plus fort biais d'usage du code.

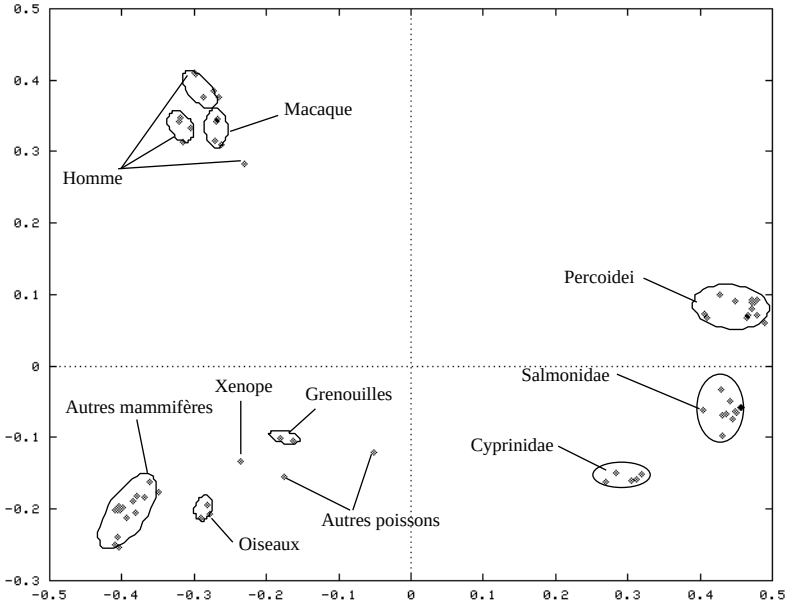


Figure 3.3: Plan F1× F2 de l'ACO (analyse en coordonnées principales) des 70 gènes d'hormone de croissance.

La seconde analyse concerne l'étude de la phylogénie du gène de l'hormone de croissance. L'analyse en coordonnées principales (ACO) peut être utilisée de façon complémentaire aux techniques usuelles de reconstitution de phylogénies (Higgins 1992). Le jeu de données provient de la banque Hovergen, qui contient 70 séquences alignées. L'ACO a été réalisée grâce à NetMul sur une matrice de similarité entre les 70 séquences. La figure 3.3 montre les résultats obtenus dans le plan F1xF2. Ce graphique permet de retrouver l'organisation phylogénétique classique, et souligne des caractéristiques particulières qui n'apparaissent pas sur les arbres, en particulier la proximité de *Anguilla japonica* et *Amia alva* ("Autres poissons" sur la figure 3.3) avec *Xenopus* et le groupe des grenouilles.

2. Biologie des populations d'insectes

Après avoir soutenu ma thèse de 3ème cycle en 1985, j'ai continué à travailler sur les populations d'insectes ravageurs du colza, en collaboration avec le CETIOM (Centre Technique Interprofessionnel des Oléagineux Métropolitains) pendant un an. J'ai ainsi publié un article dans le Journal of Applied Ecology (Thioulouse 1987) qui faisait une synthèse des résultats obtenus sur l'altise du colza, *Psylliodes chrysocephala*.

Je ne présente ici qu'un bref résumé de ces résultats, qui concernent principalement les structures spatio-temporelles, et qui ont été obtenus en conjuguant trois types de méthodes: analyse multivariée, tests non paramétriques, et représentations graphiques, en particulier cartographiques.

Trois échelles de structurations spatio-temporelles ont été mises en évidence. La première se situe à l'échelle de la plante: le comportement de ponte des femelles induit une agrégativité vraie des effectifs de larves par plante, et un mécanisme de régulation densité-dépendant prend place à cette échelle, soit par mortalité différentielle des larves présentes dans les plantes sur-infestées, soit par migration des larves.

A l'échelle de la parcelle de colza, des structures fortes apparaissent dans l'infestation par les adultes. Ces structures évoluent au cours du temps et peuvent être expliquées par l'hétérogénéité du phénomène d'infestation des parcelles lors des vols à partir des sites de l'année précédente (parcelles cultivées en colza ou abris divers), par des différences de survie locales ou des mouvements des insectes dans la parcelle, et par des interactions avec la plante hôte (qualité et densité).

Les variations inter-parcellaires sont aussi significatives, et ont pu être reliées d'une part à la position des parcelles cultivées en colza d'une année à l'autre, et d'autre part à l'existence de sites "relais", susceptibles de fournir un abri aux adultes en dehors des parcelles cultivées. Un modèle d'infestation intégrant la localisation des parcelles et les niveaux d'infestation permet de rendre compte de façon satisfaisante de cette échelle de structuration dans la zone étudiée.

3. Agronomie

Au cours d'une collaboration suivie avec Patrice Cadet (Centre ORSTOM de Dakar), nous avons pu mener à bien plusieurs études concernant la lutte contre les nématodes phyto-parasites.

La première étude date de 1989 (Cadet & Thioulouse 1989); elle portait sur l'analyse de l'influence des traitements chimiques nématicides sur les peuplements de nématodes parasites de la canne à sucre au Burkina Fasso. Les données provenaient d'un plan d'expérimentation complexe, faisant intervenir plusieurs séquences de traitements nématicides, 14 espèces de nématodes, et cinq années d'étude (plantation et années de repousse). Afin de pouvoir aborder simultanément ces divers paramètres, nous avons utilisé l'analyse triadique (Thioulouse & Chessel 1987), en constituant un tableau (dates x séquences) pour chaque espèce.

Les résultats obtenus montrent que seule la fumigation effectuée avant la plantation provoque une baisse importante du peuplement de nématodes. Les traitements appliqués les années suivantes en repousses n'affectent que modérément la multiplication des nématodes. Dans toutes les séquences de traitement comparées, le peuplement a tendance à s'accroître au fil des repousses. Les nématicides ont tendance à modifier la distribution des espèces de nématodes, et conduisent à un peuplement différent des parcelles témoins, mais dont la pathogénicité est supérieure. Pour expliquer ce phénomène, on peut évoquer l'hypothèse suivante: l'effet pathogène des nématodes se traduit par une limitation initiale du développement racinaire de la canne. Les nématicides éliminent momentanément l'incidence des parasites et permettent à la plante de s'enraciner. Cependant, l'action du nématicide étant imparfaite ou réversible, les nouvelles racines peuvent servir de support à la multiplication des nématodes. Certains genres, plus pathogènes, semblent alors plus aptes à coloniser rapidement ces racines, peut-être parce qu'ils résistent mieux aux nématicides, ce qui expliquerait leur prédominance en canne de repousse.

La seconde étude est plus récente (Cadet *et al.* 1994); elle concerne les relations entre les propriétés physico-chimiques du sol et les peuplements de nématodes parasites de la canne à sucre à la Martinique. L'objectif de ce type d'étude est le contrôle mésologique des populations de nématodes, c'est à dire une modification de la composition des peuplements vers un équilibre favorisant les espèces moins pathogènes par l'intermédiaire d'une action sur les paramètres du sol. Cette stratégie offre de nombreux avantages, et en particulier celui d'éviter le phénomène de "vide écologique" créé par l'élimination indifférenciée de toutes les espèces (cf paragraphe précédent).

Les données sur lesquelles ont porté l'étude sont constituées par les dénombrements de six espèces de nématodes le long de trois transects mis en place dans une parcelle de canne à sucre (26 points d'échantillonnage au total). Dans cette parcelle, un nivellement mécanique a amené à l'affleurement trois horizons distincts dont les caractéristiques sont très différentes, et qui sont traversés par les transects. En chaque point de prélèvement, 14 variables physico-chimiques sont mesurées. Une analyse de co-inertie nous a permis de mettre en relation composition du peuplement de nématodes et physico-chimie du sol.

Les résultats montrent que la variation progressive des teneurs de certains éléments physiques et chimiques du sol (matière organique, phosphore, pH) s'accompagne de variations progressives de l'abondance de certaines espèces de nématodes (*Hemicriconemoides* et *Pratylenchus*). Cette analyse révèle également des relations qui n'évoluent pas selon un gradient le long du transect : l'abondance d'*Helicotylenchus* peut, par exemple, être mise en parallèle avec l'existence de fortes teneurs en calcium. Ces résultats ne donnent cependant aucune indication sur le mode d'action des paramètres du sol sur l'équilibre spécifique du peuplement de nématodes. Ces mécanismes devront donc être analysés grâce à des expériences en conditions contrôlées avant de pouvoir aboutir à un contrôle mésologique.

Une troisième étude portait sur l'identification des facteurs pédologiques influençant les peuplements de nématodes ravageurs de la tomate et de l'igname à la Martinique. Ces résultats ont été en partie présentés au 4ème congrès du réseau d'Afrique de l'Est, du

Centre et du Sud de la Société Internationale de Biométrie (Thioulouse, Cadet & Albrecht 1995), et sont actuellement soumis à la revue **Applied Soil Ecology**.

Comme dans le cas de la canne à sucre, nous avons utilisé l'analyse de coinertie pour traiter les données nématologiques et pédologiques, en utilisant de plus des tests de Monte-Carlo pour étudier les différences existantes dans deux situations pour l'igname. Dans le cas de la tomate, nous avons retrouvé une forte relation entre les facteurs pédologiques et les peuplements de nématodes. Dans le cas de l'igname, les résultats varient en fonction de l'espèce d'igname. Dans les cultures de *Dioscorea cayenensis rotundata*, qui est attaquée par le nématode endémique *Pratylenchus coffeae*, les relations nematode-sol sont fortes. Pour *D. alata* par contre, la relation nématode-sol est plus faible. Cette différence, mise en évidence par les tests de Monte-Carlo, peut être attribuée au fait que *D. alata* est attaquée non seulement par *Pratylenchus coffeae*, mais aussi par *Scutellonema bradys*, un ravageur allochtone introduit dans les cultures par le tubercule planté, et ne survivant pas dans le sol après la récolte. On comprend donc bien que l'influence des facteurs du sol sur ce parasite est moindre.

4. Écotoxicologie

L'écotoxicologie s'intéresse à l'impact et au devenir des substances chimiques dans l'environnement. L'analyse de données est largement utilisée dans cette discipline (domaine de la chimiométrie), et j'ai été amené à m'y intéresser à la suite d'une collaboration avec J. Devillers du CTIS (Centre de Traitement de l'Information Scientifique, Lyon, France) et W. Karcher de l'Environment Institute (Ispra, Italie). Nous avons publié en commun plusieurs articles, dont deux à la suite d'un cours Européen que j'ai donné au Joint Research Center (Ispra, Italie) sur les techniques graphiques en analyse de données (Thioulouse *et al.* 1991, Devillers *et al.* 1991), et un dans la revue Ecotoxicology and Environmental Safety (Devillers *et al.* 1993) sur l'analyse des tableaux de données de toxicité multi-spécifiques et multi-composants (tableaux espèces x composés chimiques).

L'analyse des relations structure-activité, qui est un objectif central en écotoxicologie, fait appel depuis de nombreuses années à l'analyse multivariée. Limitées initialement à la régression multiple, les techniques employées se sont diversifiées rapidement avec les méthodes classiques (ACP et AFC), puis des méthodes plus élaborées (régression PLS). Dans un article écrit en collaboration avec J. Devillers (Devillers *et al.* 1991), nous avons pu montrer comment l'utilisation des techniques de projection d'individus supplémentaires et de représentations graphiques systématiques (collections) pouvait permettre de mieux comprendre les prédictions du modèle de Mackay. Ce modèle vise à prédire le devenir de composés chimiques dans l'environnement, et en particulier dans les six compartiments suivants: air, eau, sol, sédiments, sédiments suspendus, et compartiment biologique, à partir de leurs propriétés physico-chimiques.

Dans un autre domaine, l'analyse des données de toxicité fait intervenir un type de tableau un peu particulier dans lesquels figure la toxicité d'une série de composés chimiques vis à vis de différentes espèces (ou de différents stades de développement d'un organisme). Ces tableaux sont des tableaux homogènes, car la quantité mesurée est la même dans toutes les cases du tableau (il s'agit d'une concentration). La méthode que nous avons proposée (Devillers *et al.* 1993) pour analyser ces tableaux est basée sur l'ACP, avec divers types de transformations (centrage colonnes, centrage lignes, double centrage, résidus de la prédiction par les facteurs successifs de l'ACP). Une méthode de représentation graphique particulière (figure 3.4, représentation du tableau par cercles et carrés avec réarrangement des lignes et des colonnes en fonction des coordonnées factorielles) permet de dégager pas à pas les différents effets induisant des structures

dans le tableau (effet espèce, effet produit, interactions). Il nous a ainsi été possible de souligner la sensibilité de certaines espèces à certains produits, ainsi que l'importance du choix du stade de développement de l'organisme.

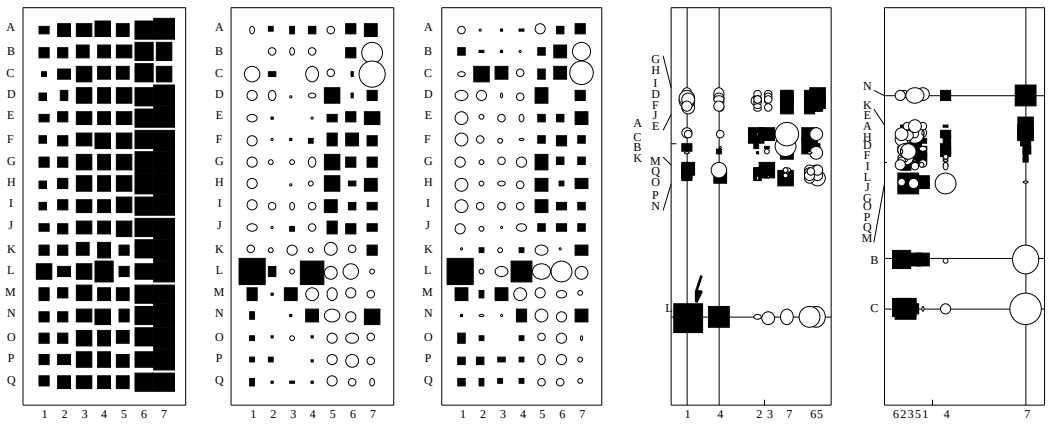


Figure 3.4: Représentations graphiques de tableaux de toxicité.
De gauche à droite: données brutes, données centrées par produits (colonnes), données centrées par produits et par espèces, données doublement centrées avec réarrangement selon les coordonnées sur le premier facteur de l'ACP, résidus par rapport à la reconstitution par le premier facteur de l'ACP avec réarrangement selon les coordonnées sur le second facteur de l'ACP.

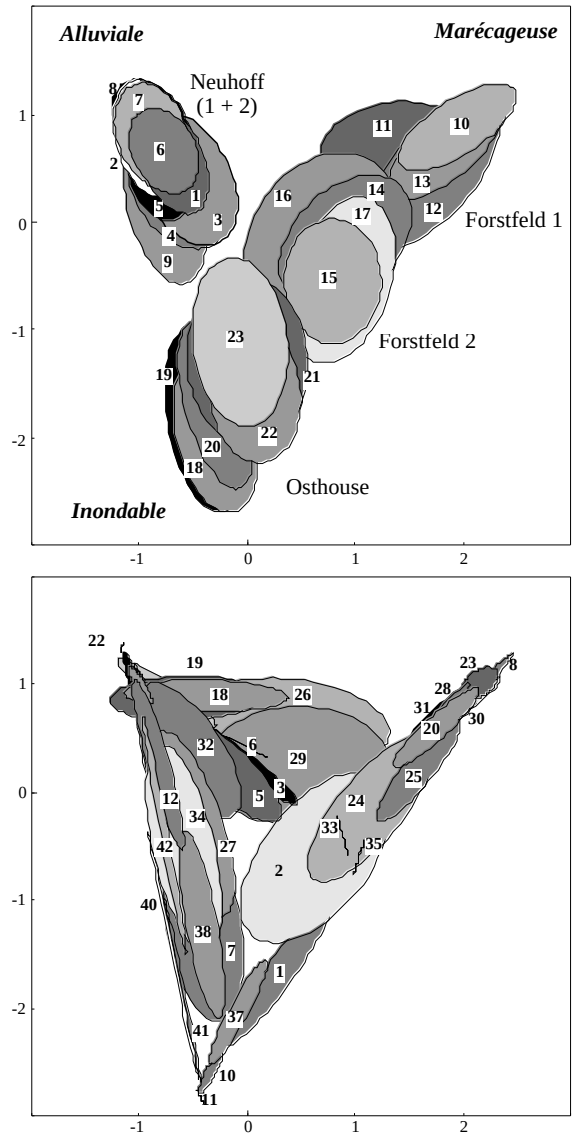
5. Écologie

Le reciprocal scaling, qui a été présenté au chapitre 1, est une méthode bien adaptée à la description de la variabilité intra et inter-relevés (diversité spécifique) et intra et inter-espèces (amplitude d'habitat et séparation des niches) dans les tableaux écologiques (tableaux espèces x relevés).

Les représentations graphiques qu'elle permet d'obtenir sont particulièrement intéressantes, dans la mesure où elles fournissent directement un sens biologique aux théorèmes sur lesquels elle repose. Nous l'avons utilisée sur deux jeux de données, l'un provenant de la thèse d'Hasnaoui (1979), et l'autre d'un rapport de Tatibouet et Broyet (1980).

Dans le premier cas, 23 relevés pratiqués dans des forêts alsaciennes fournissent l'abondance de 43 espèces d'arbres. Le problème est de déterminer si la liste des espèces dans les clairières de régénération est caractéristique de la forêt environnante, ou si elle fait intervenir des espèces allochtones caractéristiques des stades pionniers. Les 23 relevés proviennent de trois types de forêts: alluviale (Neuhoff), marécageuse (Forstfeld), et inondable (Osthouse). A Neuhoff et Forstfeld deux sites d'échantillonnage sont étudiés, avec quatre ou cinq relevés à chaque site.

L'interprétation de la figure 3.5 est très claire: la composition des relevés et leur diversité est entièrement dictée par la forêt d'où ils proviennent (figure 3.5, en haut). Les ellipses sont regroupées en fonction de la forêt où ont été effectués les relevés correspondants (les deux sites de Forstfeld sont bien séparés, alors que ceux de Neuhoff ne le sont pas). Pour des forêts différentes, les ellipses ne se superposent pratiquement pas. Il s'agit là d'un modèle de type classification: il est possible de discriminer les relevés d'après l'abondance des espèces qu'ils contiennent car la diversité inter-relevés est supérieure à la variabilité intra.



1	*Carpinus betulus*
2	*Fraxinus excelsior*
3	*Evonymus europæus*
4	*Ulmus campestris*
5	*Quercus pedunculata*
6	*Ulmus effusa*
7	*Cratægus oxyacantha*
8	*Ribes nigrum*
9	*Lonicera xylosteum*
10	*Acer pseudoplatanus*
11	*Tilia cordata*
12	*Clematis vitalba*
13	*Cornus sanguinea*
14	*Corylus avellana*
15	*Ligustrum vulgare*
16	*Hedera helix*
17	*Cratægus monogyna*
18	*Prunus padus*
19	*Viburnum lantana*
20	*Rubus saxatelis*
21	*Populus alba*
22	*Acer campestre*
23	*Alnus glutinosa*
24	*Urtica dioïca*
25	*Carex brizoïdes*
26	*Carex maxima*
27	*Viola sylvestris*
28	*Solidago serotina*
29	*Glechoma hedecera*
30	*Impatiens nolitangere*
31	*Cirsium palustre*
32	*Sanicula europæa*
33	*Angelica sylvestris*
34	*Anemone nemorosa*
35	*Mentha aquatica*
36	*Paris quadrifolia*
37	*Carex sylvatica*
38	*Fragaria vesca*
39	*Allium ursinum*
40	*Stachys sylvaticus*
41	*Arum maculatum*
42	*Polygonatum multiflorum*
43	*Sonchus asper*

Figure 3.5: Plan F1× F2 du reciprocal scaling. En haut: les ellipses correspondent aux 23 relevés. En bas: les ellipses correspondent aux 43 espèces d'arbres.

Inversement, les ellipses des espèces ne forment pas de groupes (figure 3.5, en bas). Elles sont au contraire très souvent superposées, et elles sont étirées entre deux des sommets du triangle défini par les trois types de forêts (alluviale, marécageuse, inondable). Ceci signifie que seul un petit nombre d'espèces est caractéristique de chaque type de forêt, et que la plupart s'ordonne le long des trois types de gradients. La discrimination entre les forêts n'est donc pas due à la présence ou l'absence d'espèces particulières, mais à des combinaisons d'espèces. La liste des espèces dans les clairières de régénération est caractéristique de la forêt d'où elle provient, plutôt que d'un stade pionnier.

Dans le second cas, l'abondance de 40 espèces d'oiseaux est estimée dans 51 points d'écoute répartis le long d'un gradient d'urbanisation. La figure 3.6 montre le résultat du reciprocal scaling. Ici aussi la différence de structure entre le graphique des espèces (figure 3.6, en haut) et celui des relevés (figure 3.6, en bas) est nette. Les ellipses des relevés montrent une discontinuité entre milieu urbain (à gauche) et rural (à droite), soulignée par le trait vertical pointillé. En milieu urbain les relevés montrent une

diversité plus faible (ellipses plus petites et allongées) qu'en milieu rural (ellipses plus grandes).

Le plan des espèces montre que le maximum de diversité correspond au maximum du changement des conditions environnementales (limite rural-urbain), caractéristique d'un écotone. Les deux graphiques sont superposables, et on peut vérifier qu'un grand nombre d'ellipses se situent au niveau du trait pointillé vertical. Ces ellipses sont plus grandes, traduisant une plus grande amplitude d'habitat pour ces espèces.

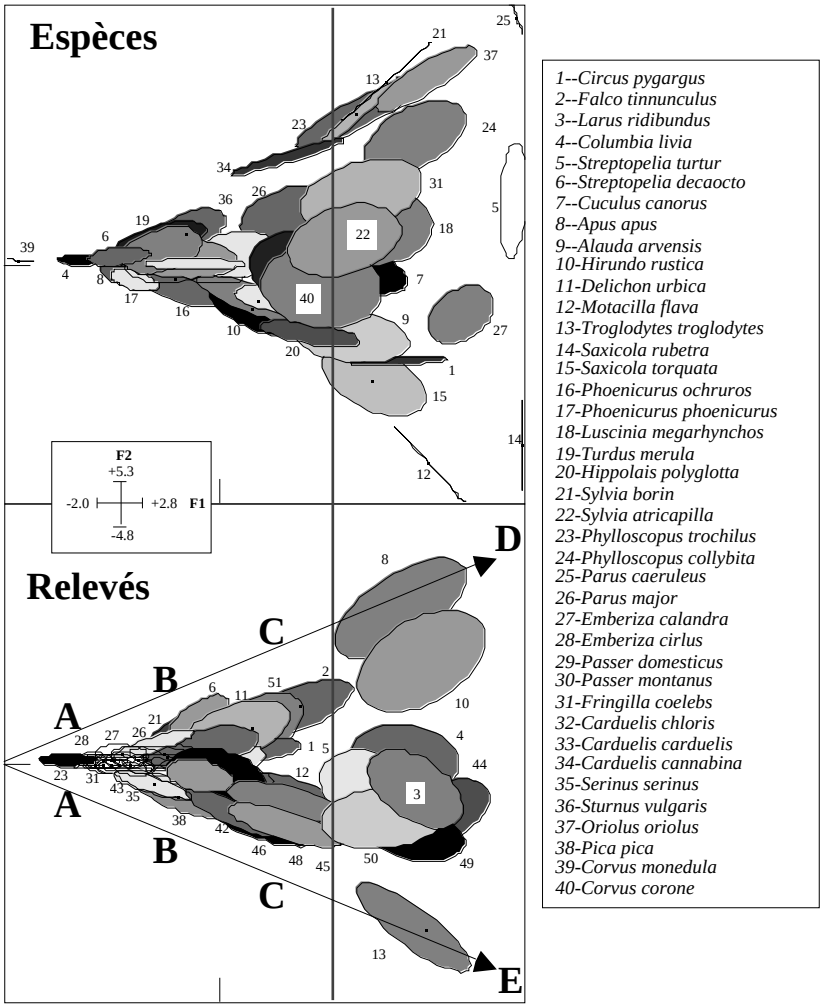


Figure 3.6: Plan F1× F2 du reciprocal scaling. En haut: les ellipses correspondent aux 40 espèces d'oiseaux. En bas: les ellipses correspondent aux 51 relevés.

L'analyse de données spatialisées est particulièrement importante en écologie, et les analyses locales et globales présentées dans le chapitre 1 sont intéressantes de ce point de vue. Nous les avons utilisées sur un jeu de données constitué par les relevés d'abondance de 64 espèces d'oiseaux en 94 points (Bournaud 1990). La figure 3.7 montre les cartes des abondances, et il est aisé de constater que divers types de distribution spatiale sont identifiables (opposition entre les espèces 19 et 28 par exemple).

Le premier facteur des AFC totale, globale et locale est représenté dans la figure 3.8. La figure 3.8B montre que le facteur 1 de l'analyse globale est relativement lisse, et met nettement en évidence un gradient nord-est sud-ouest. Ceci est dû aux contraintes

imposées à ce facteur, et au fait qu'il faut utiliser 9 voisins pour obtenir un lissage satisfaisant : l'erreur de lissage atteint effectivement un minimum pour 9 voisins, au lieu de 6 pour les deux autres analyses. De plus, même pour 5 voisins, l'erreur de lissage est plus faible pour l'analyse globale (0.72) que pour les analyses totale (1.3) et locale (1.7). Le gradient mis en évidence par le facteur 1 correspond à un gradient de végétation sur la zone d'échantillonnage de type ouvert/fermé, qui affecte donc ici clairement la distribution de certaines espèces d'oiseaux.

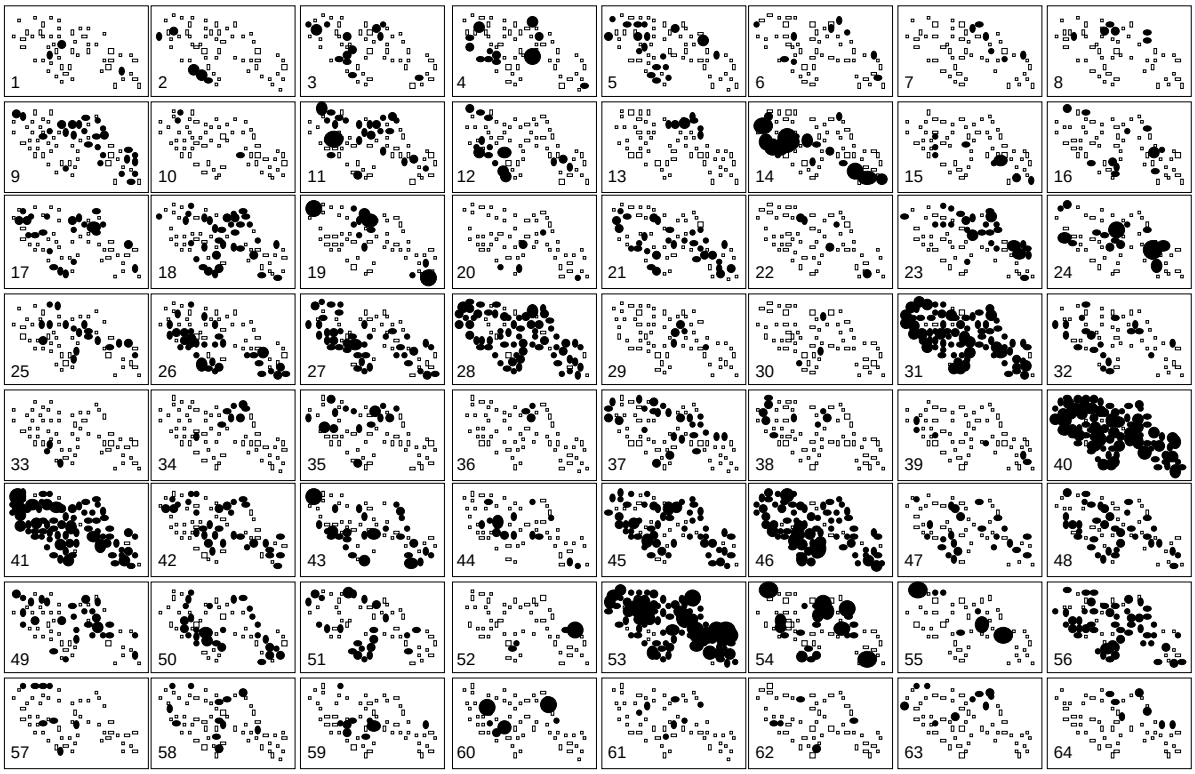


Figure 3.7: Cartes de l'abondance des 64 espèces d'oiseaux dans les 91 relevés.
La taille des cercles est proportionnelle à l'abondance. Un point correspond à une absence.

Les courbes de niveaux de la figure 3.8C (analyse locale) sont beaucoup moins lisses. La représentation par cercles et carrés souligne la présence de zones de faible surface où le facteur 1 est négatif, entourées de zones à valeurs positives. Ces structures locales (puisque le F1 maximise la variance locale) correspondent aussi à des structures de la végétation (clairières et lisières environnantes) qui affectent elles aussi la distribution des oiseaux.

Le F1 de l'AFC totale, qui est simplement une AFC munie de la pondération de voisinage, possède une structure intermédiaire entre les analyses globales et locales. Les caractéristiques de ces deux analyses se retrouvent sur la figure 3.8A, mais de façon beaucoup moins nette. Le gradient ouvert/fermé est moins bien visible, et l'opposition entre les cercles et les carrés est moins évidente. Dans l'analyse totale, les structures globales sont cachées par les structures locales car elles sont représentées à la même échelle.

Cet exemple souligne bien l'importance des représentations graphiques (et donc des logiciels permettant de les réaliser) dans l'étude des structures spatiales en écologie. Les techniques d'interpolation dans le plan et de courbes de niveaux, utilisées conjointement avec les méthodes d'ordination fournissent des représentations qui facilitent considérablement l'interprétation.

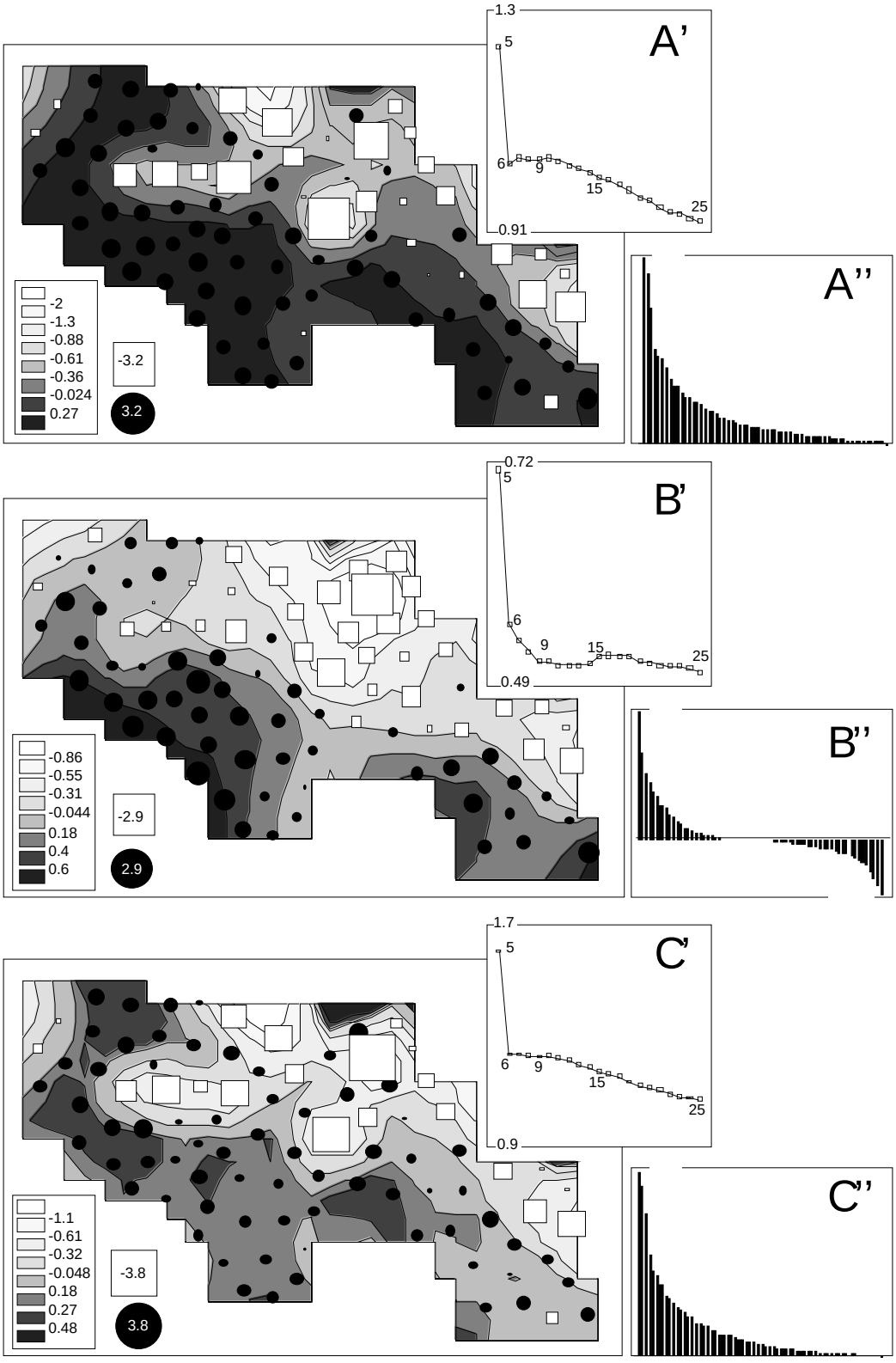


Figure 3.8: Représentation cartographique du premier facteur des AFC totale (A), globale (B) et locale (C). Les valeurs du facteur sont représentés à la fois par des courbes de niveaux avec niveaux de gris et par des cercles (valeurs positives) et des carrés (valeurs négatives).
Les courbes de niveaux sont calculées par régression Lowess sur un nombre de voisins choisi grâce à la représentation de l'erreur de régression en fonction du nombre de voisins (A', B', C').
Les graphiques A'', B'' et C'' représentent les valeurs propres des analyses totale, globale et locale.

Conclusions - Perspectives

Dans le domaine de l'analyse de données, les trois méthodes décrites au chapitre 1 (analyse triadique partielle, reciprocal scaling, analyses locales et globales) sont des exemples de développements méthodologiques directement utilisables par les biologistes. Du point de vue mathématique, les objets manipulés ne sont pas très originaux, mais nous essayons de les rendre utilisables par des non statisticiens. Cet objectif est difficile à atteindre. Il passe par des impératifs scientifiques (publication dans des revues internationales) et techniques (disponibilité des logiciels, facilité de mise en oeuvre) parfois difficiles à concilier. La graphique scientifique en est pour nous l'outil privilégié.

Sur le plan des méthodes statistiques, deux axes particuliers me paraissent devoir être poursuivis et je compte m'y employer, en collaboration avec Daniel Chessel (Laboratoire d'Écologie des Eaux Douces et des Grands Fleuves). Il s'agit d'une part des méthodes d'étude des structures spatiales, en particulier pour ce qui est de l'extension de l'utilisation des relations de voisinage aux méthodes dites de couplages de tableaux. Un article récent (Thioulouse, Chessel & Champely 1995) nous a en effet permis de montrer ce qui était possible dans le cadre des relations de voisinage pour les méthodes à un tableau (ACP et AFC). L'extension aux méthodes de couplage de deux tableaux en analyse de co-inertie a déjà été réalisée, et nous avons l'intention de publier ces travaux prochainement, après les avoir étendus aux autres méthodes de couplage (analyses sur variables instrumentales et régression PLS par exemple).

D'autre part, les méthodes multi-tableaux restent un domaine de recherche active, par exemple pour l'extension de l'analyse de co-inertie à k tableaux (Chessel & Hanafi 1996), et j'espère pouvoir aussi y contribuer. En effet, la répétition au cours du temps d'une campagne d'échantillonnage conduit naturellement à des jeux de données structurés en multitableaux. Ce point de vue débouche sur un objectif fondamental en analyse de données : la typologie de structures. Les méthodes simples permettent de faire de la typologie d'état; les méthodes multitableaux comme STATIS permettent d'étudier des structures moyennes et les écarts à ces moyennes. L'étude des typologies de structures est elle indispensable à l'analyse du fonctionnement des écosystèmes.

Les outils informatiques demandent un investissement important, ne serait-ce que pour maintenir à jour une programmathèque diffusable. En effet, les progrès très rapides de l'industrie informatique se traduisent par une évolution constante des machines et des systèmes d'exploitation, qui apporte toujours plus de puissance et de facilité d'utilisation, mais qui en contrepartie exige des efforts de plus en plus grands pour l'écriture et la maintenance des logiciels. La mise au point d'ADE-4 a représenté un progrès certain dans ce domaine, grâce à l'utilisation d'un langage de programmation standard et relativement portable (le langage C), et d'un système de développement qui dissocie l'interface utilisateur du code de calcul. Un gros travail reste à fournir afin de terminer le portage d'un certain nombre d'anciens modules d'ADE 3.7, de développer des modules graphiques nouveaux, et de compléter la documentation.

L'originalité d'ADE-4 tient autant dans l'originalité et la diversité des méthodes statistiques (co-inertie, variables floues, inter/intra, variables instrumentales, multitableaux, etc.) que dans celles des méthodes graphiques : multifenêtrage systématique sur les lignes et sur les colonnes des fichiers, structuration en graphiques à une dimension / courbes / nuages de points / cartographie, gestion des collection / superposition, graphiques interactifs, etc. Ces deux aspects (statistique et graphique) font plus que se compléter : par exemple, le multifenêtrage décuple les possibilités des

méthodes multitableaux. L'origine de la diversité des méthodes disponibles dans ADE réside dans la diversité des situations biologiques, des questions qu'elles soulèvent, et finalement des jeux de données récoltés par les biologistes. Cette diversité est bien sûr loin d'être maîtrisée, et elle continuera à générer des problèmes méthodologiques nouveaux. Une autre caractéristique d'ADE étant son ouverture, l'intégration de ces nouvelles méthodes se fera sans problème.

L'importance croissante des réseaux informatiques se traduira, à échéance sans doute assez brève, par un autre challenge : celui de la mise à la disposition des biologistes de la connaissance scientifique et des développements méthodologiques sur le réseau Internet. Nous sommes en bonne position pour y participer activement, avec en particulier le serveur Web pour la diffusion d'ADE-4 et le système NetMul. L'axe de recherche principal dans ce domaine sera sans doute l'implémentation de notre méthodologie graphique dans les clients Web grâce au langage Java. En effet, ce langage permet de s'affranchir des contraintes de portabilité du code, tout en disposant de primitives graphiques puissantes, aussi bien du point de vue de la gestion de l'interface utilisateur que pour la programmation graphique en analyse de données. A terme, une des évolutions possibles de la micro-informatique actuelle se traduira peut-être par un changement radical du micro-ordinateur lui-même, transformé en terminal relié au réseau et ne servant qu'à exécuter les logiciels disponibles sur ce réseau. Le serveur Web d'ADE-4 rentre dans cette logique, avec la disponibilité de la totalité du logiciel et de sa documentation (presque 1000 pages).

Dans le champ des applications biologiques, la biologie moléculaire est un domaine où l'analyse multivariée devrait permettre des avancées significatives. Les exemples évoqués dans le troisième chapitre de ce mémoire nous ont en effet permis de constater que les structures de données sous-jacentes aux grandes banques (fréquences des codons dans les gènes, composition des protéines, propriétés physico-chimiques ou biologiques des acides aminés et des protéines, structures phylogéniques) justifient l'utilisation de diverses techniques d'analyse de données avancées (couplage de tableaux, multitableaux, etc.), alors que seules les méthodes de base (ACP, AFC, classification) sont employées actuellement. Paradoxalement l'écologie, où les développements informatiques sont beaucoup moins avancés, a connu depuis longtemps une utilisation nettement plus importante de l'analyse multivariée. Nous nous efforcerons donc de développer ce nouveau champ d'applications.

D'autre part, ma participation au contrat "*Nouvelles approches de la bioévaluation à partir des invertébrés benthiques: bases théoriques, outils biométriques et validation à l'échelle des grands bassins*" (J.G. Wasson, CEMAGREF Lyon) me permettra de poursuivre des travaux à orientation écologique.

Enfin, j'ai fait une demande de mise à disposition auprès du CNRS pour travailler pendant une année au laboratoire de Nématologie de l'ORSTOM à Dakar (Sénégal). Ce séjour me permettra de poursuivre une collaboration avec Patrice Cadet (centre ORSTOM de Dakar) sur l'étude des relations entre les nématodes phyto-parasites, le sol, et l'environnement végétal en situation de jachère. Ce programme a été entrepris récemment au Sénégal, et il est intégré au grand programme "*Jachère en Afrique de l'Ouest*", dont la finalité est la restauration de la fertilité des sols par une jachère naturelle "améliorée". C'est un programme multidisciplinaire dans lequel les biologistes (en particulier les nématologistes), interviennent de manière coordonnée avec des agronomes, des pédologues, et des hydrologues. Il y a deux niveaux d'intégration :

- le premier niveau consiste à comprendre le fonctionnement du peuplement de nématodes en fonction de l'évolution de la flore (plantes hôtes) et des transformations induites au niveau des caractéristiques biologiques et physico-chimiques du sol par la jachère.

- le second niveau consiste à intégrer les informations obtenues par les autres disciplines pour développer une technique de gestion et de régulation des peuplements de nématodes phytoparasites (pendant la jachère puis pendant la culture) qui s'appuie sur une sélection, à partir de leur incidence secondaire sur le peuplement de nématodes, des interventions destinées à améliorer les différentes composantes biotiques et abiotiques de la fertilité.

Deux thèses seront en principe achevées en nématologie à cette époque : l'une porte sur les relations plantes de jachère - peuplements de nématodes (Emmanuelle Pate, Université Lyon 1, UMR 5558), l'autre a pour thème les relations entre les facteurs abiotiques telluriques et les peuplements de nématodes (N'Deye N'Diaye, Université de Dakar, Biologie Animale).

De plus, dans le cadre du contrat de programme de l'UR 35 ORSTOM "*Bases Biophysiques de la Gestion Durable des Agrosystèmes Tropicaux*", intitulé provisoirement "*Fertilité Biologique des Sols Tropicaux*", des résultats seront disponibles sur les mécanismes d'action des facteurs abiotiques du sol, sur les stratégies de survie des nématodes et sur leur dispersion passive par les eaux de ruissellement. Mon travail consistera à intégrer ces informations pour déterminer, dans l'évolution des peuplements, la part qui revient aux relations hôte-parasite et la part qui revient aux relations mésologiques.

Références bibliographiques

Amanieu M., Guelorget O. & Nougier-Soule J. (1981). Analyse de la diversité de la macrofaune benthique d'une lagune littorale méditerranéenne. *Vie Milieu*, **31**, 303-312.

Apple Computer, Inc. (1987). *Human Interface Guidelines: The Apple Desktop Interface*. Addison-Wesley, Reading, MA, USA.

Auda Y. (1983). Rôle des méthodes graphiques en analyse des données: application au dépouillement des enquètes écologiques. Thèse de 3ème cycle, Université Lyon 1.

Banet T.A. & Lebart L. (1984). Local and partial principal components analysis (PCA) and correspondence analysis (CA). In *COMPSTAT 84*. International Association for Statistical Computing (eds), Physica-Verlag, Vienna, Austria. pp. 113-123.

Bertin J. (1967). Les diagrammes, les réseaux, les cartes. Mouton et Gautier-Villars, Paris.

Borcard D., Legendre P., & Drapeau P. (1992). Partialling out the spatial component of ecological variation. *Ecology,* **73**, 1045-1055.

Borcard D. & Legendre P. (1994). Environmental control and spatial structure in ecological communities: an example using oribatid mites (Acari, Oribatei). *Environmental and Ecological Statistics*, **1**, 37-61.

Bournaud M. (1990). Peuplement d'oiseaux et propriétés des écocomplexes de la plaine du Rhône: descripteurs de fonctionnement global et gestion des berges. Rapport sur programme SRETIE, Ministère de l'Environnement, pp. 1-135.

Cadet P. & Thioulouse J. (1989). Traitements nématicides et peuplements de nématodes parasites de la canne à sucre au Burkina Faso. 1 - Repousses. *Revue de Nématologie*, **12**, 1, 35-44.

Cadet P., Thioulouse J. & Albrecht A. (1994). Relationships between ferrisol properties and the structure of plant parasitic nematode communities on sugarcane in Martinique (French West Indies). *Acta Œcologica*, **15**, 6, 767-780.

Cailliez F. (1984). *Analyse des données*. Les presses de l'Universités de Montréal, Montréal, Canada.

Chessel D. & Mercier P. (1993). Couplage de triplets statistiques et liaisons espèces-environnement. In *Biométrie et Environnement*., Lebreton J.D. and Asselain B. (eds), 15-44. Paris, Masson.

Cliff A.D. & Ord J.K. (1973). *Spatial autocorrelation*. Pion, London, England.

Cleveland W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* **74**, 829-836.

Cleveland W.S. & Devlin S.J. (1988). Locally weighted repression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association* **83**, 596-610.

Conway M., Pausch R. & Passarelle K. (1992). A tutorial for SUIT, the simple user interface toolkit. Computer science department, University of Virginia.

Coppi R. (1994). An introduction to multiway data and their analysis. *Computational Statistics and Data Analysis*, **18**, 3-13.

Coppi R. & Di Ciaccio (1994). Multiway data analysis - Software and applications. *Computational Statistics and Data Analysis*, **18**, 1-2.

Coutaz J. (1990). *Interfaces homme-ordinateur*. Bordas, Paris.

Debouzie D., Thioulouse J. & Ballanger Y. (1984). Structures spatiales et temporelles des populations d'un ravageur du colza (*Psylliodes chrysocephala* L. (Col. Chrysomelidae)) dans plusieurs parcelles de culture. *Acta Oecologica, Oecologia Applicata*, **5**, 4, 335-353.

Debouzie D. & Thioulouse J. (1986). Statistics to find spatial and temporal structures in populations. In: Mangel M. (Ed.), *Pest control: operations and system analysis in fruit fly managment*, NATO ASI Series, Vol. G11, Springer Verlag, p. 263-282.

Devillers J., Thioulouse J., Domine D., Chastrette M., & Karcher W. (1991). Multivariate analysis of the input and output data in the fugacity model level 1. In: Devillers J. & Karcher W. (Eds) *Applied Multivariate Analysis in SAR and Environmental Studies*, Kluwer Academic Publishers, p. 281-345.

Devillers J., Thioulouse J. & Karcher W. (1993). Chemometrical Evaluation of Multispecies-Multichemical Data by Means of Graphical Techniques Combined to Multivariate Analyses. *Ecotoxicology and Environmental Safety*, **26**, 333-345.

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

Escofier B. & Pagès J. (1982). Comparaison de groupes de variables définies sur le même ensemble d'individus. Rapport de recherche n°149, ISSN 0249-6399. INRIA, Domaine de Voluceau-Rocquencourt, BP 105, 78153 Le chesnay cedex, France.

Escofier B. & Pagès J. (1994). Multiple factor analysis (AFMULT package). *Computational Statistics and Data Analysis*, **18**, 121-140.

Escoufier Y. (1980). L'analyse conjointe de plusieurs matrices de données. In *Biométrie et Temps*. Jolivet, M. (eds), 59-76. Paris, Société Française de Biométrie.

Escoufier Y. (1982). L'analyse des tableaux de contingence simples et multiples. *Metron*, **40**, 53-77.

Escoufier Y. (1987). The duality diagram: a means for better practical applications. In *Developments in numerical ecology*, P. Legendre and L. Legendre (eds), NATO Advanced Study Institute, Series G (Ecological Sciences). Springer Verlag, Berlin, Germany. pp. 139-156.

Foley, van Dam, Feiner, & Hughes (1990). Computer graphics - Principles and practices. Addison-Wesley.

Foucart T. (1978). Sur les suites de tableaux de contingence indexés par le temps. *Statistique et Analyse des données*, **2**, 67-84.

Gabriel K.R. (1971). The biplot graphical display of matrices with application to principal component analysis. *Biometrika* **58**, 453-467.

Geary R.C. (1954). The contiguity ratio and statistical mapping. *The Incorporated Statistician*, **5**, 115-145.

Geladi P. & Kowalski B.R. (1986). Partial least-squares regression: a tutorial. *Anal. Chim. Acta*, **1**, 185, 19-32.

Gittins R. (1968). Trend surface analysis of ecological data. *J. Ecol.*, **56**, 845-869.

Glaçon F. (1981). Analyse conjointe de plusieurs matrices de données. Comparaison de différentes méthodes. Thèse de 3° cycle, Université de Grenoble.

Green P.J. & Sibson R. (1977). Computing Dirichlet tesselations in the plane. *The Computer Journal*, **21**, 168-173.

Greenacre M. (1984). *Theory and applications of correspondence analysis*. London: Academic Press.

Gundersen H.J.G. & Jensen E.B. (1987). The efficiency of systematic sampling in stereology and its prediction. *Journal of Microscopy*, **147**, 229-263.

Hasnaoui B. (1979). Etude structurale des trouées de régénération naturelle en forêts humides de la plaine d'Alsace. Thèse, Université de Strasbourg, Strasbourg, France.

Higgins D.G. (1992). Sequence ordinations: a multivariate analysis approach to analysing large sequence data sets. *Comput. Applic. Biosci.*, **8**, 15-22.

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

Höskuldsson A. (1988). PLS regression methods. *Journal of Chemometrics* **2**, 211-228.

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

Jaffrenou P.A. (1978). Sur l'analyse des familles finies de variables vectorielles. Bases algébriques et application à la description stastistique. Thèse de 3ème cycle, Université Claude Bernard Lyon I.

Kevin V. & Whitney M. (1972). Algorithm 422. Minimal Spanning Tree [H]. *Communications of the Association for Computing Machinery* **15**, 273-274.

Kidera A., Konishi Y., Oka M., Ooi T. & Scheraga H.A. (1985). Statictical analysis of the physical properties of the 20 naturally occuring amino-acids. *J. Prot. Chem.*, **4**, 23-55.

Kroonenberg P.M. (1989). The analysis of multiple tables in factorial ecology. III Three-mode principal component analysis:"analyse triadique complète". *Acta Œcologica, Œcologia Generalis*, **10**, 3, 245-256.

Kroonenberg P.M. (1994). The TUCKALS line. A suite of programs for three-way data analysis. *Computational Statistics and Data Analysis*, **18**, 73-96.

L´Hermier des Plantes H. (1976). Structuration des tableaux à trois indices de la statistique. Théorie et applications d'une méthode d'analyse conjointe. Thèse de 3° cycle, USTL, Montpellier.

Lavit Ch. (1988). *Analyse conjointe de tableaux quantitatifs*. Masson, Paris.

Lavit Ch., Escoufier Y., Sabatier R. & Traissac P. (1994). The ACT (Statis method). *Computational Statistics and Data Analysis*, **18**, 97-119.

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

Lee P.J. (1969). Theory and application of canonical trend surface analysis. *Journal of Geology*, **77**, 303-318.

Lee P.J. (1981). The most predictable surface (MPS) mapping method in Petroleum exploration. *Bulletin of Canadian Petroleum Geology*, **29**, 224-240.

Lebart L., Morineau A. & Tabart, N. (1977). *Techniques de la description statistique, méthodes et logiciels pour la description des grands tableaux*. Dunod, Paris.

Lebart L., Morineau L. & Warwick K.M. (1984). *Multivariate descriptive analysis: correspondence analysis and related techniques for large matrices*. New York, John Wiley and Sons.

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**, 53-67.

Lebreton J.D., Richardot-Coulet M., Chessel D. & Yoccoz N. (1988). L'analyse des relations espèces-milieu par l'analyse canonique des correspondances . II Variables de milieu qualitatives. *Acta Œcologica, Œcologia Generalis* **9**, 137-151.

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. and Karcher, W. (eds), 85-114. Dordrecht: Kluwer Academic Publishers.

Le Foll Y. (1982). Pondération des distances en analyse factorielle. *Statistiques et Analyse des Données*, **7**, 13-31.

Legay J.M. (1986). Qu'est ce que la Biométrie ?. *Le courrier du CNRS*, 64, 56-61.

Legendre P. (1993). Spatial autocorrelation: trouble or new paradigm ? *Ecology*, **74**, 1659-1673.

Lindgren F. (1994). *Third generation PLS. Some elements and applications*. Research Group for Chemometrics. Department of Organic Chemistry, 1-57. Umeå: Umeå University.

Manly B.F. (1994). *Multivariate Statistical Methods. A primer*. London: Chapman and Hall.

Mantel M. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research* **27**, 209-220.

Meot A., Chessel D. & Sabatier R. (1993). Opérateurs de voisinage et analyse des données spatio-temporelles. In : *Biométrie et Environment*. Lebreton, J.D. and Asselain, B. (eds), 45-72. Paris: Masson.

Moran P.A.P. (1948). The interpretation of statistical maps. *Journal of the Royal Statistical Society B*, **10**, 243-251.

Næs T. (1984) Leverage and influence measures for principal component regression. *Chemometrics and Intelligent Laboratory Systems*, **5**, 155-168.

Nakai K., Kidera A. & Kanehisa M. (1988). Cluster analysis of amino acid indices for prediction of protein structure and function. *Prot. Eng.*, **2**, 93-100.

Nishisato S. (1980). *Analysis of categorical data: dual scaling and its applications*. University of Toronto Press, Toronto, Canada.

Pigliucci M. & Barbujani G. (1991). Geographical patterns of gene frequencies in Italian populations of *Ornithogalum montanun* (Liliaceae). *Genetical Research, Cambridge*, **58**, 95-104.

Potvin C. & Travis J. (1993). Concluding remarks: a drop in the ocean. *Ecology*, **74**, 1674-1676.

Ripley B.D. (1981). *Spatial statistics.* John Wiley, New York, USA.

Schorn P. (1991). Implementing the XYZ GeoBench: A programming environment for geometric algorithms. In *Computational Geometry: Methods, Algorithms and Applications*, H. Bieri and H. Noltemeier (eds), Proc. CG'91, International Workshop on Computational Geometry, Bern, March 1991, Springer LNCS 553, pp. 187-202.

Sharp P.M. & Li W.-H. (1987). The codon adaptation index – a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.*, **15**, 1281-1295.

Sibson R. (1980). The Dirichlet tesselation as an aid in data analysis. *Scandinavian Journal of Statistics*, **7**, 14-20.

Sjöström M. & Wold S. (1985). A multivariate study of the relationship between the genetic code and the physical-chemical properties of amino-acids.*J. Mol. Evol.*, **22**, 272-277.

Sneath P.H.A. (1966). Relation between chemical structure and biological activity in peptides. *J. Theoret. Biol.*, **12**, 157-195.

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

Tatibouet F., & Broyer J. (1980). Etude des peuplements d'oiseaux nicheurs de la zone urbaine de Lyon. Pages 106-156 *in* Rapport final du Contrat 237-01-78-00314, Ministère de l'Environnement (Ecologie Urbaine), France.

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**, 91-119.

Ter_Braak C.J.F. (1987a). The analysis of vegetation-environment relationships by canonical correspondence analysis. *Vegetatio* **69**, 69-77.

Ter_Braak C.J.F. (1987b). *Unimodal models to relate species to environment.* Wageningen: Agricultural Mathematics Group.

Thioulouse J., Mathy B. & Ploye H. (1985a). Estimation d'un volume : sous-échantillonnage de coupes sériées, covariogramme transitif et calcul de précision. *Mikroskopie (Wien)*, **42**, 215-224

Thioulouse J., Houllier F. & Onillon J.C. (1985b). Variables régionalisées et dénombrement d'insectes : cas unidimensionnel. *Comptes Rendus de l'Académie des Sciences de Paris*, **301**, 9, 423-428.

Thioulouse J. (1987). Space-time structures in a winter rape pest (*Psylliodes chrysocephala* L.) population : methodological proposals and biological interpretations. *Journal of Applied Ecology*, **24**, 435-450.

Thioulouse J. & Chessel D. (1987). Les analyses multitableaux en écologie factorielle. I: De la typologie d'état à la typologie de fonctionnement par l'analyse triadique. *Acta Oecologica, Oecologia Generalis*, **8**, 4, 463-480.

Thioulouse J. (1989). Statistical analysis and graphical display of multivariate data on the Macintosh. *Computer Applications in the Biosciences*, **5**, 4, 287-292.

Thioulouse J. (1990). MacMul and GraphMu: two Macintosh programs for the display and analysis of multivariate data. *Computers and Geosciences*, **16**, 8, 1235-1240.

Thioulouse J., Devillers J., Chessel D., & Auda Y. (1991). Graphical techniques for multidimensional data analysis. In: Devillers J. & Karcher W. (Eds) *Applied Multivariate Analysis in SAR and Environmental Studies,* Kluwer Academic Publishers, p. 153-205.

Thioulouse J. & Chessel D. (1992). A method for reciprocal scaling of species tolerance and sample diversity. *Ecology*, **73**, 2, 670-680.

Thioulouse J., Chessel D. & Champely S. (1995a). Multivariate analysis of spatial patterns: a unified approach to local and global structures. *Environmental and Ecological Statistics*, **2**, (in press).

Thioulouse J., Dolédec S., Chessel D., & Olivier J.M. (1995b). ADE software: multivariate analysis and graphical display of environmental data. In: Guariso G. & Rizzoli A. (Eds) *Software per l'ambiente*, Pàtron editore, Bologne, pp. 57-62.

Thioulouse J. & Lobry J.R. (1995). Co-inertia analysis of amino-acid physico-chemical properties and protein composition with the ADE package. *Computer Applications in the Biosciences,* **11**, 3, 321-329.

Perrière G. & Thioulouse J. (1996). On-line tools for sequence retrieval and multivariate statistics in molecular biology. *Computer Applications in the Biosciences* (sous presse).

Upton G.J.G., & Fingleton B. (1985). *Spatial data analysis by example.* Volume 1: *Point pattern and quantitative data*. Wiley, Chichester, England.

Wartenberg D. (1985a). Canonical trend surface analysis: a method for describing geographic patterns. *Systematic Zoology*, **34**, 259-279.

Wartenberg D. (1985b). Multivariate spatial correlation: a method for exploratory geographical analysis. *Geographical Analysis*, **17**, 263-283.

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

Annexes

Tableau 1

Options du module PCA.

Options	Triplet	Remarques
Covariance matrix PCA	$(\mathbf{X}_1, \mathbf{D}_p, \mathbf{D}_n)$	$\mathbf{X}_1$ est le tableau des variables centrées: $\mathbf{X}_1 = \left[x_{ij} - \overline{x_j}\right]$ ($\overline{x_j}$ est la moyenne de la variable j)
Correlation matrix PCA	$(\mathbf{X}_2, \mathbf{D}_p, \mathbf{D}_n)$	$\mathbf{X}_2$ est le tableau des variables normées: $\mathbf{X}_2 = \left[(x_{ij} - \overline{x_j})/\sigma_j\right]$ (σ_j est l'écart-type de la variable j)
After row % transformation PCA	$(\mathbf{X}_3, \mathbf{D}_p, \mathbf{D}_n)$	$\mathbf{X}_3$ est le tableau des variables centrées et transformées en pourcentages par ligne (pour une représentation par biplot, Gabriel 1981)
Non centered PCA	$(\mathbf{X}, \mathbf{D}_p, \mathbf{D}_n)$	(Noy-Meir, 1973)
Row decentered PCA (or PCA with reference to one row)	$(\mathbf{X}_4, \mathbf{D}_p, \mathbf{D}_n)$	$\mathbf{X}_4 = \left[x_{ij} - r_j\right]$, r_j étant les valeurs d'un modèle lignes (*i.e.*, un individu de référence)
Decentered PCA (or PCA with reference to another table)	$(\mathbf{X}_5, \mathbf{D}_p, \mathbf{D}_n)$	$\mathbf{X}_5 = \left[x_{ij} - r_{ij}\right]$, r_{ij} étant les valeurs d'un modèle lignes×colonnes (*i.e.*, un tableau de référence)
Partial normed PCA	$(\mathbf{X}_6, \mathbf{D}_p, \mathbf{D}_n)$	Les lignes de $\mathbf{X}$ sont supposées appartenir à plusieurs groupes. $\mathbf{X}_6$ est le tableau $\mathbf{X}$ centré pour ces groupes et standardisé par variable (Bouroche 1975)
Within groups normalized PCA	$(\mathbf{X}_7, \mathbf{D}_p, \mathbf{D}_n)$	Les lignes de $\mathbf{X}$ aussi sont supposées appartenir à plusieurs groupes. $\mathbf{X}_7$ est le tableau $\mathbf{X}$ centré et standardisé par groupe (Dolédec and Chessel 1987)

Tableau 2

Options du module COA.

Options	Triplet	Remarques
Correspondence analysis	$(\mathbf{A}, \mathbf{F}_p, \mathbf{F}_n)$	(Escoufier 1982)
Total inertia test		Cette option réalise un test de Monte-Carlo (Good, 1993, chapitre 13) basé sur le critère de l'inertie totale de la table de contingence. Les permutations conservent les distributions marginales.
Dual scaling COA		Cette option calcule les coordonnées factorielles des cases de la table de contingence et en déduit les moyennes, variances, et covariances de chaque relevé et de chaque espèce (Thioulouse and Chessel, 1992)
Row weighted COA	$(\mathbf{A}, \mathbf{F}_p, \mathbf{W}_n)$	$\mathbf{W}_n$ est la matrice diagonale contenant une pondération des lignes a priori. C'est aussi une généralisation de l'analyse des correspondances floues (Chevenet *et al.* 1994)
Internal COA	$(\mathbf{A}_1, \mathbf{F}_p, \mathbf{F}_n)$	Les relevés (et éventuellement les espèces) sont supposés appartenir à plusieurs groupes. $\mathbf{A}_1$ est le tableau $\mathbf{A}$ après un centrage pondéré pour ces groupes (Benzecri, 1973, 1983; Cazes *et al.*, 1988)
Decentered COA	$(\mathbf{A}_2, \mathbf{F}_p, \mathbf{F}_n)$	$\mathbf{A}_1$ est le tableau $\mathbf{A}$ après centrage par rapport à une distribution a priori (Dolédec *et al.*, 1995)

Tableau 3

Options du module MCA.

Options	Remarques
Multiple correspondence analysis	**D** est le tableau disjonctif complet associé aux variables qualitatives. $\mathbf{D}_m$ est la matrice diagonale des poids des classes et $\mathbf{W}_n$ est la matrice diagonale d'une pondération a priori des lignes (Tenenhaus & Young 1985)
Fuzzy correspondence analysis	Il s'agit d'une généralisation de l'analyse des correspondances multiples : le tableau de données contient le degré d'association de chaque individu à chaque modalité des variables qualitatives, au lieu de donner exactement la modalité portée (Chevenet *et al.* 1994)

Tableau 4

Options générales du module Distances.

Options	Remarques
Mantel test	Calcule le test de Mantel pour la comparaison de deux matrices de distances (Mantel 1967)
Table to distance matrix	Cette option calcule la matrice des distances entre les lignes (ou les colonnes) du tableau de données. Les distances peuvent être calculées avec la métrique Euclidienne usuelle (variables quantitatives), ou avec l'indice de Jaccard (variables qualitatives)
Triplet to distance matrix	Calcule aussi la matrice des distances entre les lignes (ou les colonnes) du tableau de données, mais la métrique est fournie par la définition du triplet (matrices $\mathbf{D}_n$ ou $\mathbf{D}_p$). Ceci permet d'utiliser n'importe quelle métrique (par exemple la métrique du χ^2 dans le cas des tables de contingences)
Principal coordinates analysis	Réalise l'analyse en coordonnées principales d'une matrice de distances (Manly 1994)
Minimum spanning tree	Calcule l'arbre de longueur minimale classique (Kevin & Whitney 1972), ainsi que les arbres orthogonaux obtenus en interdisant les arêtes déjà utilisées

Tableau 5

Les six options du module Distances dédiées à la décomposition spatiale de la variance.

Options	Triplet	Remarques
Moran index analysis	$(\mathbf{X_D}, \mathbf{I}_p, \mathbf{P})$	Analyse de la variabilité globale (covariance entre les valeurs observées et la moyenne de leurs voisins). $\mathbf{X_D}$ est le tableau de données **D**-centré (ou **D**-normé) et $\mathbf{I}_p$ est la matrice identité d'ordre p (Thioulouse *et al.* 1995)
Spatial covariance		Analyse de la covariance spatiale entre deux tableaux : $\mathbf{X}^t(\mathbf{P}-\mathbf{DUD})\mathbf{Y}$, ou $\mathbf{X}^t$ est le tableau transposé, **Y** est le second tableau, et **U** la matrice unité d'ordre n (tous les éléments égaux à 1)
Moran eigenvectors		Calcule les vecteurs propres des opérateurs de lissage $\mathbf{D}^{-1}\mathbf{P}$ et $\mathbf{I}_n - \mathbf{D}^{-1}\mathbf{P}$ décrits par Méot *et al.* (1993) and Thioulouse *et al.* (1995)
Geary index analysis	$(\mathbf{X_D}, \mathbf{I}_p, \mathbf{D}-\mathbf{P})$	Analyse de la variabilité locale (covariance entre les valeurs observées et la différence entre ces valeurs et la moyenne de leurs voisins) (Thioulouse *et al.* 1995)
Local covariance		Analyse de la covariance locale de deux tableaux : $\mathbf{X}^t(\mathbf{D}-\mathbf{P})\mathbf{Y}$
Geary test		Calcule l'indice de Geary pour une relation de voisinage quelconque (Geary, 1954; Cliff and Ord, 1973)

Tableau 6

Options du module Discrimin.

Options	Triplet	Remarques
Initialize		Préparation : relie une analyse de type ACP, AFC ou ACM à une variable qualitative décrivant les groupes d'individus (lignes)
Analysis of variance		Analyse de variance de toutes les variables d'un tableau. Les sorties donnent la décomposition de la variance (inter, intra et totale), la valeur du F et la probabilité correspondante
Discriminant analysis: Run	$(\mathbf{G},(\mathbf{X}^t\mathbf{D}_n\mathbf{X})^-,\mathbf{D}_k)$	Si l'analyse initiale porte sur des variables quantitatives (module PCA), on obtient l'analyse discriminante classique (Mahalanobis 1936; Tomassone *et al.* 1988), avec toutes les variantes offertes par le module PCA. Si il s'agit d'un tableau de variables qualitatives (module MCA), on obtient l'analyse discriminante sur variables qualitatives (Saporta 1975; Persat *et al.* 1985). Si il s'agit d'une table de contingence (module COA) on obtient l'analyse discriminante des correspondances (Chessel, comm. pers.)
Discriminant analysis: Test		Exécute un test de Monte-Carlo basé sur le critère de la trace totale : $Tr(\mathbf{G}^t\mathbf{D}_k\mathbf{G}(\mathbf{X}^t\mathbf{D}_n\mathbf{X})^-)$ (Pilai 1955; Tomassone 1988)
Supplementary rows		Calcule le code discriminant de lignes supplémentaires
Between analysis: Run	$(\mathbf{G},\mathbf{D}_p,\mathbf{D}_k)$	Analyse inter-groupes (Dolédec & Chessel 1987, 1989). Les codes des lignes du tableau **X** initial sont calculés automatiquement, ce qui permet de les comparer aux codes des centres de gravité.
Between analysis: Test		Test de Monte-Carlo basé sur le critère de la trace totale : $Tr(\mathbf{G}^t\mathbf{D}_k\mathbf{G}\mathbf{D}_p)$
Within analysis: Run	$(\mathbf{X}_c,\mathbf{D}_p,\mathbf{D}_k)$	Analyse intra-groupes (Dolédec & Chessel 1987, 1989). $\mathbf{X}_c$ est la matrice **X** après centrage par groupe

Tableau 7

Options du module UniVarReg.

Options	Remarques
Initialize	Choix de la variable prédictrice et des variables à prédire
Polynomial -> Error	Pour chaque variable à prédire, et pour des polynômes de degré compris entre 1 et 6, calcule les coefficients des polynômes et l'erreur d'estimation
Polynomial -> Model	Pour un polynôme de degré fixé, calcule l'estimation des valeurs de la variable à prédire
Polynomial -> Curves	Calcule les valeurs estimées par un polynôme pour un nombre de points quelconque régulièrement répartis sur l'intervalle de prédiction.
Polynomial -> New data	Calcule les valeurs estimées pour des valeurs particulières de la variable prédictrice
Lowess -> Error	Pour chaque variable à prédire, et pour un nombre de voisins compris entre 5 et le nombre total de points, calcule l'erreur d'estimation par régression Lowess
Lowess -> Model	Pour un nombre fixé de voisins, calcule l'estimation de toutes les valeurs des variables à prédire par régression Lowes sur la variable prédictrice
Lowess -> Curves	Calcule les estimations de la régression Lowess pour un nombre de points quelconque régulièrement répartis sur l'intervalle de prédiction
Lowess -> New data	Calcule les estimations de la régression Lowess pour des valeurs particulières de la variable prédictrice

Tableau 8

Options du module OrthoReg.

Options	Remarques
Initialize	Choix des variables prédictrices et des variables à prédire. Un centrage ou une normalisation peut être effectué sur les variables à prédire
Variable test	Test de Monte-Carlo sur le coefficient de régression multiple (R^2) entre chaque variable à prédire et les variables prédictrices
Subspace test	Test de Monte-Carlo sur le R^2 entre chaque variable à prédire et un sous-ensemble des variables prédictrices
Partial test	Test de Monte-Carlo sur le R^2 entre chaque variable à prédire et deux sous-ensemble de variables prédictrices. Ces sous-ensembles définissent les sous-espaces vectoriels **A** et **B**, et le test est réalisé pour les régressions sur **B**/**A** et sur **A**/**B**
Modelling	Régression multiple entre chaque variable à prédire et les variables prédictrices. Calcule des estimations et des résidus

Tableau 9

Options du module LinearReg.

Options	Remarques
Initialize	Choix des variables à prédire et prédictrices
MLR -> MultCorCoeff	Calcule le coefficient de corrélation multiple entre les variables à prédire et toutes les combinaisons de variables prédictrices (il est possible de limiter le nombre de variables prédictrices à utiliser)
MLR -> Modelling	Calcule le coefficient de corrélation multiple, les estimations, les résidus et les coefficients de la régression définie dans l'option Initialize
MLR -> New Data	Calcule les estimations de la régression pour des valeurs arbitraires des variables à prédire
PLS -> Randomization test	Test de Monte-Carlo de la régression PLS
PLS -> Modelling	Régression PLS : calcul du rapport des variances expliquée et totale, les estimations et les résidus
PLS -> New Data	Calcule les estimations par régression PLS pour des valeurs arbitraires des variables prédictrices

Tableau 10

Options du module Projectors.

Options	Remarques
Table->Orthonormal basis	Calcule une base orthonormale de l'espace engendré par les colonnes d'un tableau
Triplet->Orthonormal basis	Calcule une base orthonormale de l'espace engendré par les colonnes d'un tableau muni des deux pondérations
One categ. var.->Orthonormal basis	Calcule une base orthonormale de l'espace engendré par les indicatrices d'une variable qualitative
Two categ. var.->Orthonormal basis	Les indicatrices de deux variables qualitatives engendrent les espaces **A** et **B**. Cette option calcule des bases orthonormales des sous-espaces suivants: $\mathbf{A}+\mathbf{B}$, $\mathbf{A}\times\mathbf{B}$, $\mathbf{A}\cdot\mathbf{B}$, $\mathbf{A/B}$, et $\mathbf{B/A}$. Dans ces notations, $\mathbf{A}\cdot\mathbf{B}$ est défini par $\mathbf{A}\times\mathbf{B}=(\mathbf{A}+\mathbf{B})\ (\mathbf{A}\cdot\mathbf{B})$ (Yoccoz & Chessel 1988)
Combine two orthonormal bases	A partir des bases orthonormales des espaces **A** et **B**, calcule des bases orthonormales des sous-espaces $\mathbf{A}+\mathbf{B}$, $\mathbf{A}\times\mathbf{B}$, $\mathbf{A}\cdot\mathbf{B}$, $\mathbf{A/B}$, and $\mathbf{B/A}$
Intersection of two subspaces	A partir des bases orthonormales des espaces **A** et **B**, calcule une base orthonormale de $\mathbf{A}\cap\mathbf{B}$ (Afriat 1957)
Subspace test	Test de Monte-Carlo sur l'inertie totale projetée sur un sous-espace
Triplet inertia decomposition	Décomposition de l'inertie projetée sur un sous-espace et son orthogonal
PCA on Instrumental Variables	ACPVI des triplets issus de deux analyses simples (PCA, COA, MCA)
Orthogonal PCAIV	ACPVI sur l'orthogonal du sous-espace engendré par les variables instrumentales
Table projection	Projection d'un tableau sur un sous-espace : calcul des projections, des résidus et des coefficients de régression

Tableau 11

Options du module Coinertia.

Options	Remarques
Matching two triplets	Couplage des triplets issus de deux analyses simples (PCA, COA, MCA)
Coinertia test - Fixed D	Test de Monte-Carlo sur l'inertie totale. Les lignes des deux tableaux sont permutées
Coinertia test - Fixed Tab 1	Test de Monte-Carlo sur l'inertie totale. Seules les lignes du tableau 2 sont permutées
Coinertia test - Fixed Tab 2	Test de Monte-Carlo sur l'inertie totale. Seules les lignes du tableau 1 sont permutées
Coinertia analysis	Analyse de coinertie. Selon la nature des triplets initiaux, le résultat de cette option correspond à des analyses déjà décrites. Dans le cas de deux ACP, il s'agit par exemple de l'analyse inter-batteries de Tucker (1958). Dans le cas d'une AFC et d'une ACM, on retrouve l'analyse des profils écologiques de Romane (1972). Dans le cas de deux ACM, on obtient l'analyse canonique sur variables qualitatives (Cazes 1980)

Tableau 12

Options du module RLQ.

Options	Remarques
Prepare RLQ analysis	Préparation de l'analyse et choix des trois tableaux
RLQ test - Fixed L	Test de Monte-Carlo sur le lien entre les tableaux R et Q. Les lignes du tableau R et les colonnes du tableau Q sont permutées
RLQ test - Fixed R-Q	Test de Monte-Carlo sur le lien entre les tableaux R et Q. Les lignes et les colonnes du tableau L sont permutées
Diagonalize	Analyse du triplet $(\mathbf{R}_0^t\mathbf{P}_0\mathbf{Q}_0, \mathbf{D}_p, \mathbf{D}_q)$: $\mathbf{R}_0$ est le tableau $\mathbf{R}$ après centrage. $\mathbf{Q}_0$ est le tableau $\mathbf{Q}$ après centrage. $\mathbf{P}_0 = \mathbf{P} - \mathbf{D}_n \mathbf{1}_{np} \mathbf{D}_p$, où $\mathbf{P}$ est le tableau des fréquences relatives centrées ($p_{ij} = \frac{f_{ij}}{f_{i.} f_{.j}} - 1$)
Coinertia analysis	Calcul des coordonnées factorielles
xPy test - Fixed P	Test de Monte-Carlo sur la corrélation entre les coordonnées factorielles (le tableau des fréquences est fixé)
xPy test - Fixed x-y	Test de Monte-Carlo sur la corrélation entre les coordonnées factorielles (les fréquences marginales sont fixées)

Tableau 13

Options du module STATIS.

Options	Remarques
Operator averaging	Méthode STATIS classique (compromis d'opérateurs) Escoufier 1980, Lavit 1988, Lavit *et al.* 1994
Table averaging	Analyse triadique partielle (Thioulouse & Chessel 1987)
Foucart's COA	Analyse d'une suite de tables de contingences (Foucart 1978)

Tableau 14

Options du module MFA.

Options	Remarques
Variable groups	Analyse factorielle multiple (Escofier & Pages 1994) classique (individus en commun)

Tableau 15

Options du module KTA.

Options	Remarques
Separate analyses	Analyses simples de tous les tableaux du multi-tableau
ACOM	Analyse de co-inertie multiple (Chessel & Hanafi 1995)