ESA CNRS 5023
Écologie des Eaux Douces et des Grands Fleuves
Université Claude Bernard Lyon I
Bâtiment 401C
69622 Villeurbanne CEDEX

DIPLÔME D'ÉTUDES APPROFONDIES :
«Analyse Et Modélisation Des Systèmes Biologiques»

UTILISATION DES LISTES D'OCCURRENCES SPÉCIFIQUES SPATIALISÉES EN ÉCOLOGIE ET EN BIOGÉOGRAPHIE

- Rapport technique -

DRAY Stéphane

Directeur de recherche : Daniel CHESSEL

Parrains : Roland ALLEMAND
Jean-Pierre PASCAL

Janvier 1999

L'ordination des listes d'occurrences :
quand l'analyse canonique des correspondances est une analyse canonique

SOMMAIRE

Introduction

En écologie, de nombreuses études s'attachent à définir les structures des communautés végétales ou animales ainsi que les facteurs environnementaux qui en sont responsables. Ainsi, les aires de distribution des espèces et les variations de la biodiversité à travers le temps et l'espace sont fréquemment analysées lors de travaux à vocation théorique ou pratique. La manipulation de grands jeux de données rend nécessaire l'utilisation d'outils statistiques dans cette discipline. La statistique exploratoire multidimensionnelle a pour objet d'identifier la partie structurée, assimilable à un modèle, des systèmes biologiques ; ce modèle participera ensuite à la formulation des hypothèses sur l'origine des structures rencontrées et le fonctionnement du modèle (Yoccoz 1988). Cette démarche nécessite une connaissance préalable du modèle sous-jacent aux données observées car chaque méthode a des objectifs, des propriétés et des conditions d'applications qui lui sont propres. L'utilisation d'une méthode revient alors, implicitement, à faire des suppositions sur le système étudié. Le choix d'un outil statistique en adéquation avec les données à traiter apparaît donc comme une première étape primordiale. Pour ce faire, il importe de préciser les propriétés inhérentes à chacune des méthodes. Le cadre mathématique du modèle euclidien bâti autour du schéma de dualité (Cailliez & Pagès 1976, Escoufier 1987), enrichi par l'utilisation des projecteurs (Takeuchi *et al.* 1982), permet d'appréhender, à l'aide d'un langage universel, les relations existantes entre les différentes méthodes.

Des études en écologie ou en biogéographie, s'intéressant à l'aire de distribution d'espèces ou à l'étude de la biodiversité, ont conduit leurs auteurs à utiliser des listes d'occurrences. Ces listes résultent de la compilation de données tirées d'atlas, de spécimens préservés dans des herbiers ou muséums… Ces occurrences relèvent donc de rencontres indépendantes entre un organisme et, par exemple, un botaniste en un lieu et à une date donnée. Le développement des ordinateurs et des bases de données a permis de stocker cette information grâce à laquelle il est facile d'appréhender des phénomènes à l'échelle régionale (Hill 1991), voire continentale (Pearson & Ghorpade 1989). Cet engouement pour des études à grande échelle spatiale ou temporelle (Schaffer *et al.* 1998) a permis de revaloriser un grand nombre de données historiques contenues dans les muséums. Bien que ce type de données constitue une observation écologique basique :

"*Recording the occurrence of a species at a given place and time is at once an elementary and an integrative ecological observation. At minimum, noting the existence of a species precedes any other biological knowledge about it.*" (Wright *et al.* 1998), aucune méthode existante ne permet l'analyse de structure spatiale multispécifique à partir de listes d'occurrences.

En effet, l'analyse des listes d'occurrences fait le plus souvent appel à la définition d'unités spatiales élémentaires. La technique la plus courante consiste en un découpage de la zone analysée en quadrats par superposition d'une grille sur les cartes de distribution (Brown & Opler 1990, Mourelle & Ezcurra 1996). Les variations de composition spécifique et les facteurs environnementaux qui en sont à l'origine sont alors étudiés à l'échelle du quadrat par l'intermédiaire de techniques préalablement utilisées sur des tableaux espèces-relevés. Ainsi, au lieu de développer des outils statistiques adaptés au traitement des listes d'occurrences, les écologistes ont modifié leurs données pour les adapter aux méthodes existantes (Kadmon & Danin 1997, Cronk 1989). Pourtant, le traitement des listes d'occurrences introduit des perturbations théoriques dans l'ensemble des méthodes d'analyse de

données. En effet, le statut d'un tableau relevés-espèces présente des différences fondamentales avec un tableau quadrats-espèces construit à partir d'une liste d'occurrences. Par exemple, lorsque la grille de quadrats a été définie *a posteriori*, il se pose le problème du biais d'échantillonnage : les listes d'occurrences ne fournissent aucune information sur l'intensité de prospection de chacun des quadrats et ceci peut entraîner des biais lors de l'étude des relations espèces-milieu. De plus, chaque occurrence possède de l'information qui lui est propre (coordonnées spatiales, nom du collecteur, …) et l'étude des relations espèces-milieu à l'échelle des occurrences apparaît donc théoriquement possible.

Les listes d'occurrences présentent un intérêt pratique indéniable pour des études d'écologie ou de biogéographie à grande échelle avec un coût raisonnable, mais leur exploitation nécessite le développement de méthodes plus adaptées aux caractéristiques atypiques de ce type de données.

Ce travail s'inscrit dans le prolongement de l'étude de l'endémisme dans les Ghâts occidentaux réalisé par Gimaret-Carpentier (1999) à partir de données d'atlas (Ramesh & Pascal 1997). Cette thèse a conduit à certains développements méthodologiques concernant l'usage de l'analyse canonique (AC), de l'analyse canonique des correspondances (ACC) et de l'analyse factorielle des correspondances (AFC) en écologie. Ce rapport permet de développer les aspects mathématiques de l'approche proposée par Gimaret -Carpentier (1999) et de l'appliquer sur des listes d'occurrences de nature différente. Ainsi, dans un premier temps, on s'attachera à présenter les méthodes classiquement utilisées en écologie pour l'étude des relations espèces-milieu. Puis, nous présenterons une approche qui apparaît particulièrement adaptée à l'analyse des listes d'occurrences en évoquant, à l'aide de l'écriture de schémas de dualité et l'utilisation de projecteurs, les relations que différentes méthodes peuvent entretenir entre elles. Enfin, nous nous intéresserons à l'étude des variations de la composition spécifique à partir de deux types de listes d'occurrences. La première illustration permettra d'analyser la structure spatiale multispécifique d'une parcelle de forêt tropicale à l'aide d'une liste exhaustive d'occurrences. La seconde illustration fournira un aperçu de l'analyse d'une liste d'occurrences échantillonnée à une échelle régionale. Ce deuxième exemple est basé sur l'exploitation d'une base de données qui m'a été fournie tardivement (fin Juin 1999) et, dans ce rapport, ne figure donc que les résultats préliminaires à l'analyse de cette base.

I. Un bref aperçu historique

En général, l'information floro-faunistique est présentée sous la forme d'un tableau F croisant les espèces en colonnes et les relevés en lignes. Chaque cellule f_{ij} de ce tableau contient l'abondance ou un indice de présence/absence de l'espèce j dans le relevé i. Le traitement de ce type de tableaux par les méthodes d'ordination permet d'en synthétiser la structure en établissant une typologie des relevés selon leur composition spécifique et une typologie d'espèces en fonction de leur répartition dans les relevés. L'objectif de ces approches consiste à révéler l'existence d'une structure mais également à déterminer quels facteurs environnementaux principaux en sont la cause : "*A major purpose is interpretation of community relationships to environment, and not simply the representation of numerical relationships among samples or species in a hyperspace with a limited number of axes.*" (Gauch & Wentworth 1976).

En face du tableau F, on peut donc retrouver de l'information mésologique contenue dans le tableau M et deux stratégies sont alors possibles (Figure 1) :

- une ordination directe des relevés et des espèces le long du (ou des) gradient(s) de variables environnementales.

- une ordination des relevés selon leur composition spécifique, suivie d'une interprétation du diagramme ainsi obtenu à la lumière de connaissances extérieures sur les espèces ou les sites.

Whittaker (1967) qualifie ces deux stratégies, respectivement, *direct gradient analysis* et *indirect gradient analysis*. Les développements méthodologiques de ces différentes approches ont conduit aux procédures d'analyse euclidienne exploratoire de données du tableau F ou des tableaux F et M conjointement. Par l'intermédiaire d'une synthèse non exhaustive, nous allons présenter quelques caractéristiques principales des méthodes classiquement utilisées afin de comprendre pourquoi certaines se sont imposées comme des standards pour l'analyse des communautés en écologie.

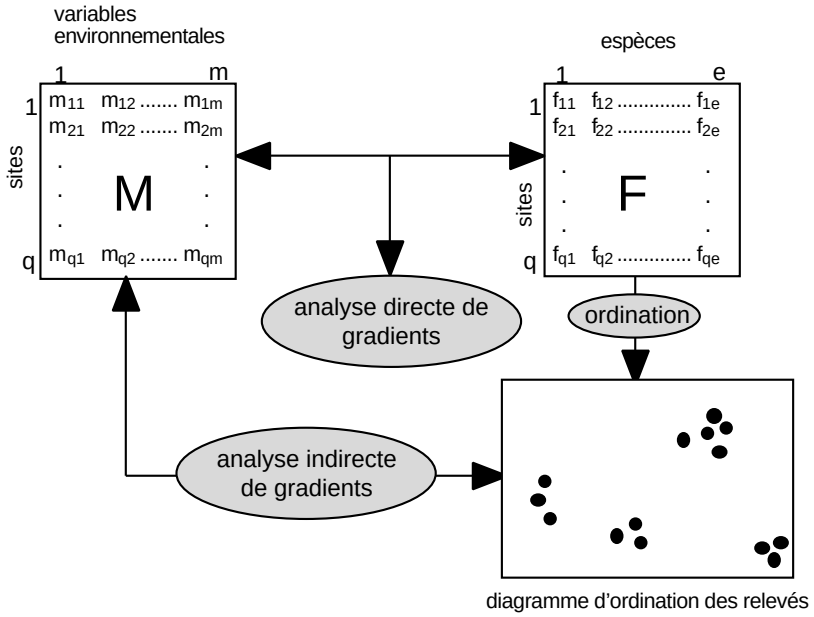


Figure 1 : Représentation schématique du type de données et des procédures d'analyse directe et indirecte de gradients utilisées pour l'analyse des communautés en écologie. (d'après Jongman *et al.* 1987)

I.1. L'ordination indirecte

Malgré le développement de méthodes d'ordination et de classification par les phytoécologistes, certains ont introduit en écologie des techniques ayant fait leurs preuves dans d'autres disciplines. Parmi ces méthodes, on trouve la famille des "analyses à vecteurs propres" (*eigenvectors analysis*), basées sur l'optimalisation d'un critère à l'aide de diagonalisation de matrices, et dans lesquelles la rigueur mathématique séduit : "*These methods were characterized by a higher level of formality and mathematical sophistication than the ad hoc ordination techniques developed directly by ecologists for ecological purposes; and their use was facilitated by the development of computers.*" (Noy-Meir & Whittaker 1977). Les méthodes aux valeurs propres sont étroitement liées au concept d'espace euclidien. Ainsi, pour le tableau F, deux points de vue sont possibles : représenter les relevés (respectivement espèces) comme des objets dans l'espace des espèces (respectivement relevés). Chaque ligne du tableau F est un point F_i de R^e et la similarité entre deux relevés F_i et F_j est alors mesuré par leur produit scalaire (associé à une pondération des espèces D_e) dans R^e :

$$\langle F_i / F_j \rangle_{D_e} = \sum_{k=1}^{e} D_k F_{ik} F_{jk}$$

On a le même raisonnement pour les espèces avec une pondération D_q des relevés dans R^q. L'inertie du nuage des relevés autour d'un point o de R^e est alors :

$$Iner = \sum_{i=1}^{e} D_{q_i} \| F_i - o \|^2_{D_e}$$

Cette quantité mesure les variations de composition spécifique des relevés ; la contribution d'un relevé dépend alors de sa distance à o et de son poids. Les méthodes aux vecteurs propres permettent de décomposer cette inertie au moyen d'axes orthogonaux d'inertie projetée décroissante. Ces axes correspondent aux vecteurs propres associés aux valeurs propres non nulles de la matrice de variances-covariances (matrice des similarités spécifiques). Par conséquent, si les variations de composition spécifique sont principalement dues à un facteur environnemental, le premier axe permettra une ordination des espèces et des relevés le long de ce facteur. On obtient alors un résumé, ayant une signification écologique, dans un sous-espace de dimension réduite, de l'information contenue en F.

Les premiers résultats obtenus par l'application de l'analyse en composantes principales (ACP, Hotelling 1933) sur les tableaux floro-faunistiques apparaissent très prometteurs : "*At some stage it may have seemed to be the definitive solution to the problem of community.*" (Noy-Meir & Whittaker 1977). Cependant, le modèle inhérent à l'ACP ne semble pas permettre une ordination optimale des espèces et des relevés. En effet, les facteurs synthétiques issus de cette procédure maximisent la variance des positions des scores des relevés et maximisent la somme des covariances des espèces. L'ordination des espèces et des relevés ne se fait donc pas selon les mêmes critères. De plus, la possibilité fournie à l'utilisateur de choisir les pondérations ainsi que les transformations éventuelles du tableau F (centrage par espèce, normalisation par espèce, …) n'apparaît pas comme un critère de souplesse de la méthode mais plutôt comme un désavantage : "*The mutiplicity of possibilities, among which the user had to choose subjectively, caused some ecologists to wonder how objective the objective methods really were.*" (Noy-Meir & Whittaker 1977). Enfin, une imposante littérature s'est attachée à critiquer le caractère linéaire de la méthode (Beals 1973, Austin 1976, Noy-Meir & Whittaker 1977, Kessel & Whittaker 1976). En effet, les axes issus de l'ACP sont des combinaisons linéaires des variables initiales (les espèces) or, les courbes de réponse des espèces à une variable de milieu sont, la plupart du temps, curvilinéaires. Beals (1973) critique fermement l'ACP et incite au développement de méthodes d'ordination en adéquation avec les théories écologiques : "*Principal components analysis, as a method of ordination to detect environmental influences, makes many unreal assumptions about ecological data. It does not take into account the normal-curve relationship between species success and environment, nor the ecological ambiguity of species absence in a stand.*" (Beals 1973)

L'apparition de l'analyse factorielle des correspondances (AFC) a permis de concilier la rigueur mathématique à la théorie écologique. Elle a été popularisée parallèlement par Benzecri (1969) comme une méthode d'analyse aux vecteurs propres (*eigenvectors analysis*), cas particulier de l'ACP, et par Hill (1973), sous le nom de *reciprocal averaging*, comme une extension des travaux de Whittaker sur l'analyse directe de gradients. L'ordination des tableaux floro-faunistiques par l'AFC a permis d'annihiler certains problèmes inhérents à l'ACP. Alors que l'ACP peut être vue comme une

modélisation de l'abondance des espèces (f_{ij}), l'AFC travaille sur les abondances relatives ($\frac{f_{ij}}{f_{i\cdot} f_{\cdot j}} - 1$). Cette transformation a pour effet de linéariser les courbes de réponse des espèces à une variable de milieu (Lebreton *et al.* 1991). De plus, l'AFC réalise une ordination des espèces et des relevés de façon simultanée permettant de placer une espèce à la moyenne des relevés qu'elle occupe et un relevé à la moyenne des espèces qu'il contient. Selon Hill (1973), c'est le fait que l'AFC utilise une bonne ordination des espèces pour ordonner les relevés (et inversement) qui lui confère un avantage décisif par rapport à l'ACP. Le succès de l'utilisation de l'AFC en écologie a été reconnu par Digby & Kempton (1987) ("*Correspondence analysis (CA) is now probably the most popular ordination method among ecologists.*") et confirmé par l'interprétation de TerBraak (1985) qui présente l'AFC comme une approximation de modèle de réponse unimodale des espèces à une variable latente.

I.2. L'ordination directe

Les techniques destinées à coupler un tableau floro-faunistique à un tableau de variables mésologiques sont nombreuses et variées (Chessel & Mercier 1993). Parmi ces méthodes, on peut citer l'analyse canonique [des corrélations] (AC, Hotelling 1936) dont la procédure ainsi que des exemples appliqués à l'écologie sont décrits dans l'ouvrage de référence de Gittins (1985). L'AC permet d'examiner les liens existant entre deux tableaux appariés par ligne (ici, F et M) en cherchant des combinaisons linéaires des variables de chacun des deux tableaux de corrélation maximale. Bien que son concept théorique soit fondamental, le champ d'application de l'analyse canonique est assez réduit (TerBraak 1990). Une des premières applications de l'AC en écologie est attribuable à Austin (1968) qui juge qu'elle ne constitue pas une procédure satisfaisante et il poursuit : "*the mathematical model on which it is based, with its requirements for orthogonal correlations between vegetation and environment and complete linearity, appears to be too stringent.*". En effet, tout comme l'ACP, l'AC fait l'hypothèse de relations linéaires entre les espèces et le milieu. De nombreux auteurs ont donc affichés leur préférence pour l'ACP dont les résultats s'interprètent plus facilement que ceux issus de l'AC (Gauch & Wentworth 1976). De plus, l'analyse canonique nécessite un nombre important d'observations (relevés) par rapport au nombre de variables (espèces et variables de milieu) comme toutes les méthodes évoquant les régressions multiples. La recherche de combinaisons linéaires de corrélation maximale (critère optimisé en AC) pouvant s'interpréter comme une double régression multiple dont la variable dépendante est multivariée. Par conséquent, la validité de l'AC sera réduite pour des études dans des zones très diversifiées comportant un grand nombre d'espèces rares.

L'analyse canonique considère les deux tableaux de façon symétrique alors que dans le cas des relations espèces-milieu une stratégie dissymétrique apparaîtrait plus appropriée afin de déterminer quelles variables du tableau M (variables explicatives) influencent l'abondance des espèces contenue dans le tableau F (variables dépendantes). Cette dissymétrisation de prédiction a engendré les analyses sur variables instrumentales dont la version ACP/ACP est l'Analyse en Composantes Principales sur Variables Instrumentales (ACPVI, Rao 1964) ou Analyse des Redondances (Wollenberg 1977), et dont la version ACP/AFC est l'Analyse Canonique des Correspondances (ACC, TerBraak 1986) ou Analyse Factorielle des Correspondances sur Variables Instrumentales (AFCVI, Chessel *et al.* 1987).

Une très bonne présentation de ces différentes méthodes est donnée dans Lebreton *et al.* (1991). Le principe de l'ACPVI peut être résumé comme suit : les variables de F sont projetées sur le sous-espace engendré par M. On obtient alors, pour chacune des variables de F, un estimation par régression multiple sur les variables de M. On réalise ensuite une ACP sur les nouvelles variables. Cette procédure est donc plus robuste que l'analyse canonique quant au nombre de variables car elle ne met en jeu que les régressions des abondances par les variables de milieu. L'ACPVI requiert seulement que le nombre de variables mésologiques soit faible par rapport au nombre de relevés. Si le tableau F dérive d'une AFC, on pourra alors réaliser une AFCVI (dont une description plus approfondie est donnée par la suite) qui permet en outre de régler le problème des relations curvilinéaires.

Dans la pratique, l'AFC et sa version détendancée (Detrended Correspondence Analysis, Hill & Gauch 1980) sont les méthodes les plus utilisées pour l'ordination des tableaux floro-faunistiques. Pour l'ordination directe, c'est l'analyse canonique des correspondances et à travers elle un logiciel, CANOCO, qui s'est imposé comme standard. Birks *et al.* (1996) dénombre ainsi près de 800 références utilisant l'ACC ou une de ses variantes partielle (TerBraak 1988) ou détendancée (DCCA, Detrended Canonical Correspondence Analysis, TerBraak 1986).

I.3. L'analyse canonique des correspondances

L'analyse canonique des correspondances est présentée par TerBraak (1986) comme une extension du *reciprocal averaging* de Hill auquel est ajoutée une régression multiple. Elle est donc une alternative optimale à la pratique consistant à ordonner le tableau F par une AFC puis à interpréter les axes obtenus par régression sur des variables de milieu (Lebreton *et al.* 1988). L'ACC fait les deux en même temps et la procédure qu'elle met en oeuvre peut être résumée de la façon suivante :

1) Attribution de scores arbitraires inégaux aux différents relevés

2) Calcul des scores des espèces comme moyennes pondérées des nouveaux scores des relevés

3) Calcul de nouveaux scores des relevés comme moyennes pondérées des scores des espèces

4) Faire la régression multiple pondérée (les poids sont les marges par relevé du tableau relevés-espèces) des scores des relevés sur les m variables environnementales.

5) Utiliser les valeurs prédites par les régressions comme nouveaux scores des relevés

6) Centrer et réduire les valeurs obtenues

7) Réutiliser ces scores en 2) ou arrêter la procédure si il y a convergence (deux boucles donnent un résultat très proche).

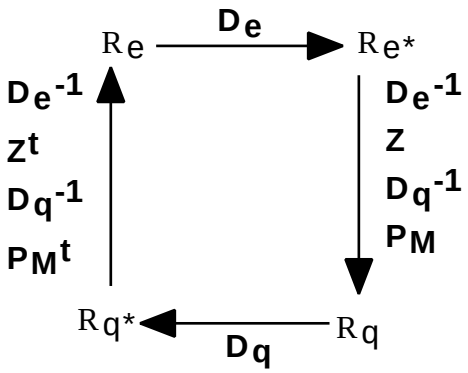
Cet algorithme correspond à celui de la méthode des moyennes réciproques présenté par Hill (1973) en y ajoutant les étapes 4 et 5. Il en résulte un code des relevés, combinaison linéaire des variables de milieu, maximisant la variance des positions moyennes des espèces (TerBraak 1987). On obtient alors une séparation optimale des niches des espèces par les variables mésologiques. Chessel *et al.* (1987) exposent les principales propriétés mathématiques de l'ACC et présentent cette procédure dans le cadre du schéma de dualité afin de la replacer parmi l'ensemble des méthodes existantes. L'ACC conduit à l'analyse d'un couple de tableaux :

- le premier (F) contient les abondances (f_{ij}) de e espèces (en colonnes) dans q relevés (en lignes). On notera Z le tableau de terme général $z_{ij}=f_{ij}/o$ avec $o=\sum f_{ij}$. On associe à ce tableau deux matrices diagonales Dq et De définies par :

$$Dq = diag(z_{1\cdot},\ldots,z_{q\cdot}) \qquad De = diag(z_{\cdot 1},\ldots,z_{\cdot e})$$

- le deuxième (M) contient les valeurs de m variables de milieu mesurées dans q relevés. Ce tableau subit un centrage pour la pondération des relevés D_q.

L'AFC du tableau F correspond à l'ACP du triplet ($D_q^{-1}ZD_e^{-1}-`_{qe}$, $D_e$, $D_q$). L'ACC vue comme AFCVI est alors l'ACP de ($\mathbf{P_M}(D_q^{-1}ZD_e^{-1}-`_{qe})$, D_e, D_q) où $\mathbf{P_M}$ définit le projecteur sur le sous-espace engendré par le tableau M. Si M est de rang plein, son expression matricielle est $\mathbf{P_M}=M(M^tD_qM)^{-1}M^tD_q$. Le centrage avec le terme $-`_{qe}$ permet, d'un point de vue technique, d'éliminer la première valeur propre triviale unité et dont les composants du vecteur propre associé sont égaux (Williams 1952). En effet, mis à part le premier facteur, l'analyse du nuage non centré conduit aux même facteurs que l'analyse du nuage centré. Si on néglige ce terme, le schéma de dualité de l'ACC s'écrit :



La matrice à diagonaliser est :

$D_e(D_e^{-1}Z^tD_q^{-1})\mathbf{P_M}^tD_q\mathbf{P_M}(D_q^{-1}ZD_e^{-1}) = D_e(D_e^{-1}Z^tD_q^{-1})D_q\mathbf{P_M}(D_q^{-1}ZD_e^{-1})$ car $\mathbf{P_M}^tD_q\mathbf{P_M} = D_q\mathbf{P_M}$ et si on remplace $\mathbf{P_M}$ par son expression matricielle, on obtient :

$$D_e(D_e^{-1}Z^tD_q^{-1})D_qM(M^tD_qM)^{-1}M^tD_q(D_q^{-1}ZD_e^{-1}) = D_e(D_e^{-1}Z^tM)(M^tD_qM)^{-1}(M^tZD_e^{-1})$$

Le tableau T_0 est défini par $T_0 = D_e^{-1}Z^tM$, il comporte e lignes et m colonnes et est formé par les moyennes par espèce des variables de M. L'ACC se résume donc à l'étude du triplet (T_0, $(M^tD_qM)^{-1}$, D_e) et diagonalise $T_0^tD_eT_0(M^tD_qM)^{-1}$. $T_0^tD_eT_0$ est une matrice de variances-covariances entre les moyennes par espèce des variables de milieu, c'est-à-dire une matrice de variances-covariances inter-espèces. La matrice à diagonaliser est donc du type (Inter)*(Totale)$^{-1}$. L'ACC permet de maximiser le rapport de l'inertie interclasse à l'inertie totale du nuage de points projetés et les valeurs propres obtenues correspondent à des rapports de corrélation et sont comprises entre 0 et 1. On retrouve donc ici l'interprétation de l'ACC comme analyse discriminante entre espèces. Le lien entre l'ACC et la théorie de la niche est alors très explicite : l'ACC permet de trouver des combinaisons de variables environnementales maximisant la séparation des niches des espèces. Les procédures de l'ACC et de l'AFC consistent, toutes les deux, à maximiser la variance entre les positions moyennes des espèces mais l'ACC réalise cet objectif avec une contrainte supplémentaire induite par les variables de milieu. Il en résulte que la première valeur propre de l'AFC ($\neg_1$) sera donc toujours supérieur ou égale à celle de l'ACC (μ_1). Le rapport $\mu_1/\neg_1$ fournit alors une mesure de la pertinence des variables de milieu à expliquer les variations de composition spécifique entre les relevés. L'ACC est une AFC sous

contrainte et paradoxalement ses résultats convergent vers ceux de l'AFC lorsque le nombre de variables de milieu augmente. Si $m \geq (q-1)$, l'ACC est alors simplement une AFC car la totalité de la structure de F est expliqué par M et l'analyse de $(D_q^{-1}ZD_e^{-1}-`_{qe})$ est équivalente à celle de $\mathbf{P_M}(D_q^{-1}ZD_e^{-1}-`_{qe})$.

Green (1971, 1974) présente une utilisation de l'analyse discriminante multiple (AD) dont l'objectif est d'identifier les facteurs environnementaux séparant au mieux les niches des espèces. L'AD constitue, selon Green, le modèle statistique approprié au concept de niche écologique multidimensionnelle développé par Hutchinson. Il présente ainsi les avantages de la méthode :

- L'AD est un modèle d'analyse multivariée et le seul adapté à ce type de problème.
- Les fonctions discriminantes sont des combinaisons linéaires des variables de milieu et aucune supposition n'est faite sur le type de relation espèces-milieu.
- L'analyse repose seulement sur les présences et ne fait aucune hypothèse concernant l'absence d'une espèce dans un relevé dont il précise l'ambiguïté :

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

Le lien entre l'ACC et l'AD est clairement explicité dans Lebreton *et al.* (1988). L'ACC correspond à l'AD par la variable nom d'espèce du tableau de variables environnementales dupliquées pour l'ensemble des correspondances (cases non nulles) du tableau F . TerBraak admet le lien entre l'AD et l'ACC et précise la différence fondamentale entre les deux méthode : "*In summary, the main difference between CCA and discriminant analysis is that the unit of the statistical analysis in discriminant analysis is the individual, whereas it is the site in CCA.*" (TerBraak & Verdonschot 1995). L'approche de Green n'est donc pas valide d'un point de vue statistique car les individus pris en compte dans son analyse sont issus de relevés et par conséquent ne sont pas indépendants les uns des autres.

II. Quand il n'y a plus de relevé, l'analyse des listes d'occurrences

Le traitement des listes d'occurrences conduit généralement à l'utilisation des méthodes d'ordination précédement citées en substituant un tableau espèces-quadrats au tableau espèces-relevés grâce à la superposition d'une grille sur la zone d'étude. Afin de conserver l'intégralité de l'information contenue dans une liste d'occurrences, nous avons développer une nouvelle méthode d'analyse en s'affranchissant de l'intermédiaire constitué par le quadrat. Les aspects théoriques de cette nouvelle approche sont présentés ici en s'appuyant sur l'écriture des schémas de dualité.

Une liste d'occurrences résulte d'une variable qualitative dont les e espèces constituent les modalités. Cette information s'écrit naturellement dans E un tableau disjonctif complet comportant o lignes et e colonnes. Sur chaque ligne, l'appartenance d'une occurrence à une espèce se traduit par une indicatrice dans la colonne concernée et des 0 ailleurs. Pour l'occurrence i appartenant à l'espèce j on aura donc (1<i<o, 1<j<e, et 1<k<e) :

$$\begin{cases} E_{ik} = 1 & \text{si} \quad k = j \\ E_{ik} = 0 & k \neq j \end{cases}$$

Chaque ligne du tableau E a le même poids 1/o, ce qui donne une matrice des poids D_O associé au tableau E avec $D_O=(1/o)I_O$. En face de cette liste d'occurrences apparaît de l'information mésologique (ou autre) associé à chacune des occurrences du tableau E. Selon la nature de cette information, diverses pratiques pourront être mises en place.

II.1. Une variable qualitative

La première situation consiste en une partition des occurrences en g classes induite par une variable qualitative y. Soit Y le tableau disjonctif complet associé à la variable y, il comporte o lignes et g colonnes (Figure 2).

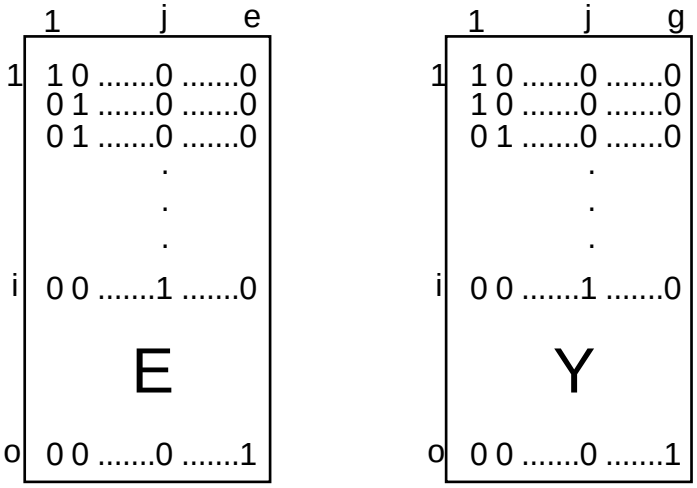
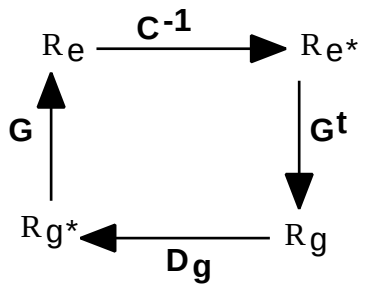
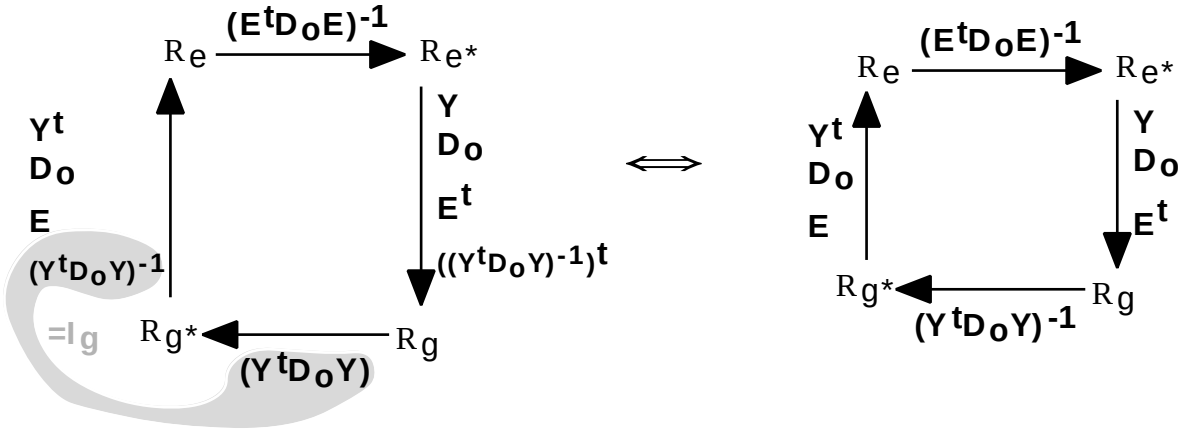


Figure 2 : Tableaux utilisées pour l'analyse de la liste d'occurrences face à une variable qualitative. Les tableaux E (liste d'occurrences) et Y (variables mésologiques) sont disjonctifs complets. Ils comportent o lignes et respectivement e et g colonnes.

La première stratégie consiste à réaliser l'analyse discriminante de E par Y. Cette analyse se résume dans le schéma suivant :



L'analyse discriminante considérée se résume à l'étude de (G, C^{-1}, D_g) avec G contenant les moyennes des variables (ici les espèces) par classe, $C^{-1}=(E^tD_oE)^{-1}$ et Dg les poids de chacune des classes. On a les relations suivantes : $G=Y^tD_oEDg^{-1}$, $Dg=(Y^tD_oY)$. Le schéma peut donc se réécrire :



Il s'avère donc que faire l'analyse discriminante de E par Y revient à faire l'analyse discriminante de Y par E qui n'est autre que l'analyse canonique de E et Y. Si on veut séparer les classes de Y par leur contenu spécifique on réalise implicitement la séparation des espèces selon leur distribution sur Y. Cette double analyse discriminante, qui est aussi une analyse canonique, est tout simplement l'AFC de la table de contingence croisant les deux variables qualitatives. Cette vision de l'AFC comme double discriminante ou analyse canonique de deux paquets d'indicatrices met en avant la notion de correspondances (cases non nulles de la table de contingence) et montre bien que l'AFC ne considère que les présences ; l'absence d'une espèce étant considérée comme une absence d'information. Ceci répond donc exactement au cadre théorique induit par une liste d'occurrences. Cependant, il semble nécessaire de s'interroger sur la nature de la partition entraînée par Y. L'analyse des listes d'occurrences conduit le plus souvent à une partition de l'espace en quadrats (Brown & Opler 1990, Ferrari *et al.* 1993, Legakis et Kypriotakis 1994). La partition ainsi obtenue n'a pas du tout le même statut que l'appartenance d'une occurrence à une espèce. En effet, alors que l'appartenance d'une occurrence à une espèce est un fait observé, l'appartenance à un quadrat est définie *a posteriori* par l'expérimentateur. Il serait donc légitime de tenir compte de cette dissymétrie dans l'analyse conjointe de ces deux partitions (espèces / quadrats). Comme nous l'avons précisé plus haut, les analyses sur variables instrumentales permettent de répondre à ce type de besoin. Dans le cas d'une partition en quadrats, l'analyse à considérer est celle de ($\mathbf{P_Y}E$, I_e, D_o) qui est l'analyse non symétrique des correspondances (ANSC) dans la version profils quadrats due à Lauro & D'Ambra (1984) et utilisée sur des listes d'occurrences

par Gimaret-Carpentier *et al.* (1998). Cette analyse est équivalente à celle de ($Dg^{-1}Y^tD_oE$, I_e, D_o) qui correspond à l'analyse inter-classe de E par Y. Ainsi, tout comme l'AFC est un cas particulier de l'analyse canonique, l'ANSC est un cas particulier de l'ACPVI. Face à une variable qualitative, l'analyse d'une liste d'occurrence conduit donc à deux stratégies dont l'une est symétrique (analyse canonique) et l'autre est dissymétrique (analyse sur variable instrumentale). Le choix entre ces différentes stratégies doit se faire selon la nature de la partition considérée et selon les objectifs de l'étude. Si on veut discriminer les quadrats par leur contenu en espèces, l'ANSC sur profils quadrats s'impose. Si on veut séparer les espèces par leur répartition, on réalisera une ANSC sur profils espèces. Si on veut faire les deux d'un coup, on fera une AFC.

II.2. Des variables quantitatives

On retrouve maintenant face à E un tableau X de m variables quantitatives mesurées ou estimées pour chaque occurrence (Figure 3).

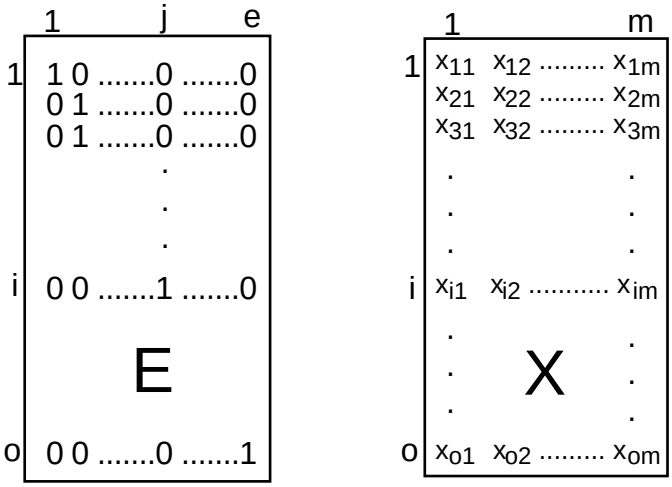


Figure 3 : Tableaux utilisées pour l'analyse de la liste d'occurrences face à des variables quantitatives. Le tableau E est disjonctif complet et il a o lignes et e colonnes. Le tableau X a o lignes et m colonnes correspondant à m variables mésologiques mesurées ou estimées pour les o occurrences.

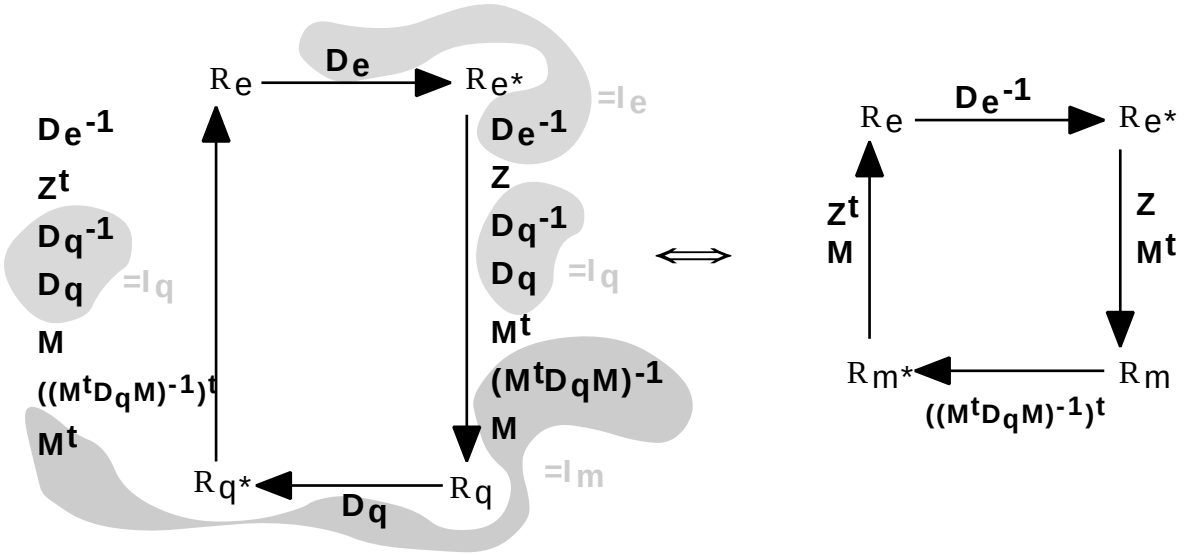
L'analyse de ce couple de tableaux apparaît comme une situation *a priori* favorable de l'utilisation de l'analyse canonique. Il n'y a pas de relevés et l'occurrence est le véritable individu statistique. Le nombre d'individus est très largement supérieur au nombre de variables : on a o»e et o»m. En écologie, il est rare de remplir les conditions numériques de l'AC qui est exclue dès qu'il y a des relevés. C'est pour cette raison que TerBraak invente, comme alternative de l'AC, l'ACC qui doit alors être pensé seulement en terme d'AFCVI. Lorsqu'il y a des relevés, l'occurrence est vue trois fois : par l'espèce, le milieu et le relevé. Dans le cas d'une vraie liste d'occurrences, l'occurrence n'est vue que deux fois : par l'espèce et le milieu. Il n'y a plus de relevé et il ne semble pas nécessaire d'utiliser un intermédiaire quadrat se substituant aux relevés.

Sur le plan théorique, il est intéressant de revenir sur le lien entre l'ACC et l'analyse discriminante utilisée par Green :

L'ACC a été définie comme une AFCVI (Chessel *et al.* 1987) soit :

$$\text{ACP de } (\mathbf{P_M}(D_q^{-1}ZD_e^{-1} - `_{qe}), D_e, D_q)$$

Sachant que $\mathbf{P_M} = M(M^tD_qM)^{-1}M^tD_q$ et si l'on néglige le terme $-`_{qe}$, on a :



On définit les tableaux E et N (variables de milieu dupliquées pour chaque correspondances, Figure 4) et on a les relations suivantes :

(1) : Z=QtDoE
(2) : Dq=QtDoQ
(3) : De=EtDoE
(4) : N=QM

avec Z de terme général $z_{ij}=f_{ij}/o$

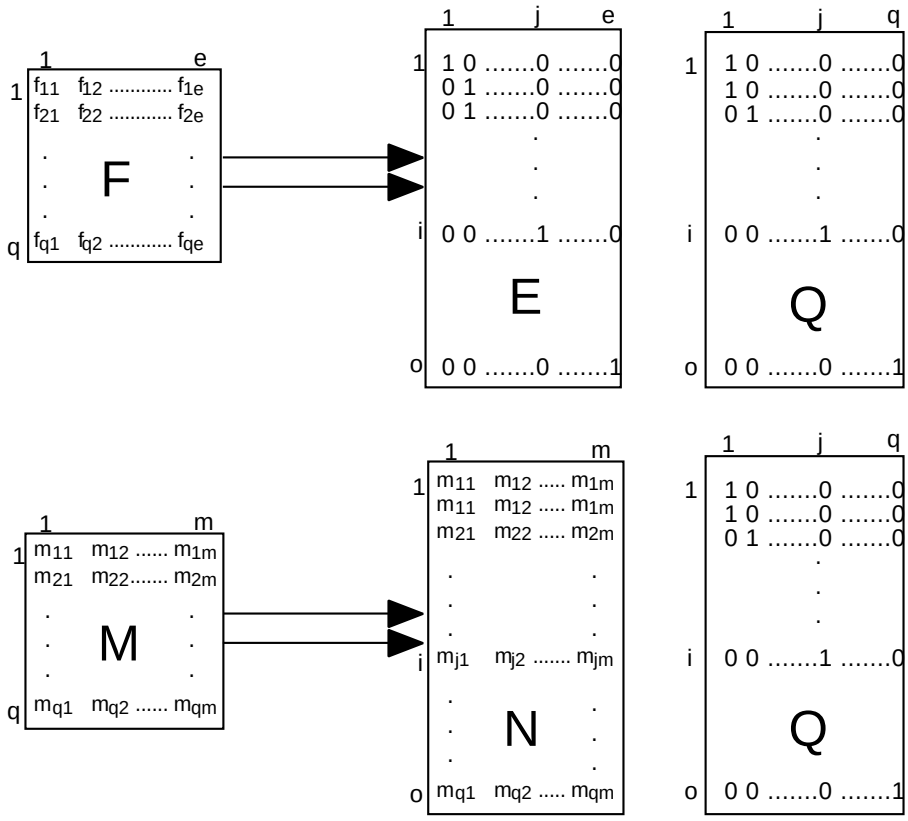


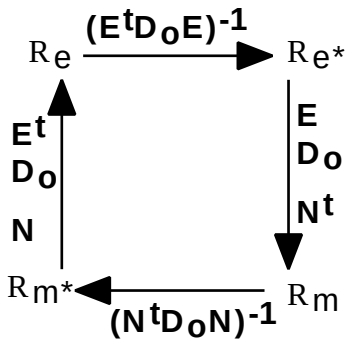
Figure 4 : Décomposition des tableaux F et M en tableaux disjonctifs complets. Les individus de la case fij sont alors les lignes de E et Q où l'on retrouve un 1 pour la colonne j de E **ET** pour la colonne i de Q. Pour le tableau M, les variables de milieu sont dupliquées pour chaque correspondance dans le tableau N.

Les éléments du schéma précédent peuvent se réécrire :

$M^tZ=(Q^{-1}N)^tZ=N^t(Q^{-1})^tQ^tD_oE=N^tD_oE$

$((M^tD_qM)^{-1})^t=(M^tD_qM)^{-1}=[(Q^{-1}N)^tD_q(Q^{-1}N)]^{-1}=[N^t(Q^{-1})^tQ^tD_oQQ^{-1}N]^{-1}=(N^tD_oN)^{-1}$

et après simplification on arrive à :



On reconnaît alors le schéma de l'analyse canonique des tableaux E et N défini par :

ACP de $(N^tD_oE, (E^tD_oE)^{-1}, (N^tD_oN)^{-1})$.

Cette procédure d'analyse canonique de deux tableaux dont l'un est formé par des indicatrices de classe est appelée *canonical variate analysis* (cf. §4 dans Gittins (1985)) et correspond à l'analyse discriminante des variables de milieu dupliquées pour chaque correspondance par la variable nom d'espèce. Dans le premier schéma, on a des applications de R^e vers R^q alors que dans le dernier schéma les applications vont de R^e vers R^m. On arrive donc à éliminer l'intermédiaire relevé en manipulant les schémas de dualité. Cette équivalence n'est cependant que théorique car la duplication des données dans le tableau N viole l'hypothèse d'indépendance des données et par conséquent ne permet pas l'utilisation de l'AC ou de l'AD sur ce type de données comme l'on fait remarquer Lebreton *et al.* (1988).

Dans le cas d'une liste d'occurrences, le tableau N est remplacé par X dont toutes les lignes sont indépendantes. Faire l'analyse discriminante de X par E est alors tout à fait légitime et cette analyse canonique particulière est pour la partie centrale du calcul une analyse canonique des correspondances. De plus, l'étape la plus critiquable de l'ACC qui consiste en un centrage des variables de milieu avec la richesse spécifique des relevés n'a pas lieu dans cette analyse. Les moyennes par espèce des variables de milieu sont calculées seulement à partir des observations réalisées (les présences) et les problèmes d'échantillonnage posés par l'utilisation de quadrats sont donc ignorés. L'étude des listes d'occurrences apparaît donc optimale par l'intermédiaire de l'analyse canonique. La méthode d'ordination étant définie, il est impératif de représenter les résultats obtenus à l'aide d'un diagramme d'ordination qui constitue l'outil principal pour résumer l'information extraite d'une analyse multivariée (TerBraak 1994). Les possibilités de représentation graphique des résultats d'une analyse canonique étant assez nombreuses (TerBraak 1990), il semble nécessaire d'explorer la géométrie inféodée à cette méthode afin de déterminer quelle représentation serait la plus appropriée dans notre cas.

II.3. Géométrie et représentation

L'analyses canonique considérée correspond à l'ACP de $(X^tD_oE, (E^tD_oE)^{-1}, (X^tD_oX)^{-1})$ et diagonalise $X(X^tD_oX)^{-1}X^tD_oE(E^tD_oE)^{-1}E^tD_o$. On reconnaît alors l'écriture des projecteurs sur E et X : $\mathbf{P_E} = E(E^tD_oE)^{-1}E^tD_o$ et $\mathbf{P_X} = X(X^tD_oX)^{-1}X^tD_o$. L'AC diagonalise donc le produit des projecteurs $\mathbf{P_X}\mathbf{P_E}$ ou $\mathbf{P_E}\mathbf{P_X}$ ce qui revient à diagonaliser leur somme $\mathbf{P_E}+\mathbf{P_X}$ ou $\mathbf{P_X}+\mathbf{P_E}$ (Foucart 1984) ce qui renvoie à la recherche de variables canoniques comme opération géométrique de R^o (Pages *et al.* 1979). En analyse canonique on pourra donc diagonaliser indifféremment une somme de projecteurs (stratégie $\sum$) ou un produit de projecteurs (stratégie $\prod$). La procédure de l'AC consiste à rechercher dans [X] et [E] les vecteurs η_1 et ξ_1 normés faisant un angle minimum α_1 ; puis à chercher, toujours dans ces sous-espaces, les vecteurs normés η_2 et ξ_2 orthogonaux aux précédents, faisant un angle minimum, etc… Les vecteurs η_i et ξ_i sont appelés variables canoniques, le carré du cosinus de l'angle α_i est un rapport de corrélation et le vecteur normé β_i porté par la bissectrice de cet angle est appelé score canonique (Figure 5). η_1 est une combinaison linéaire des variables de X et il est dans le sous-espace engendré [X]. De même, ξ_1 est une combinaison linéaire des variables de E et il est dans le sous-espace engendré [E]. η_1 et ξ_1 maximisent la corrélation, donc le cosinus de α_1, donc minimise α_1. η_1 est donc le vecteur de [X] le plus prédictible par [E] et ξ_1 est le vecteur de [E] le plus prédictible par [X]. Il en résulte donc que les vecteurs ξ_1 et $\mathbf{P_E}\eta_1$ sont colinéaires (respectivement η_1 et $\mathbf{P_X}\xi_1$). De plus, on aura β_1, bissectrice de l'angle formé par ξ_1 et η_1, colinéaire avec $\xi_1+\eta_1$. Les liens existants entre ξ_1, η_1 et β_1 sont clairement explicités par l'étude des stratégies $\prod$ et $\sum$.

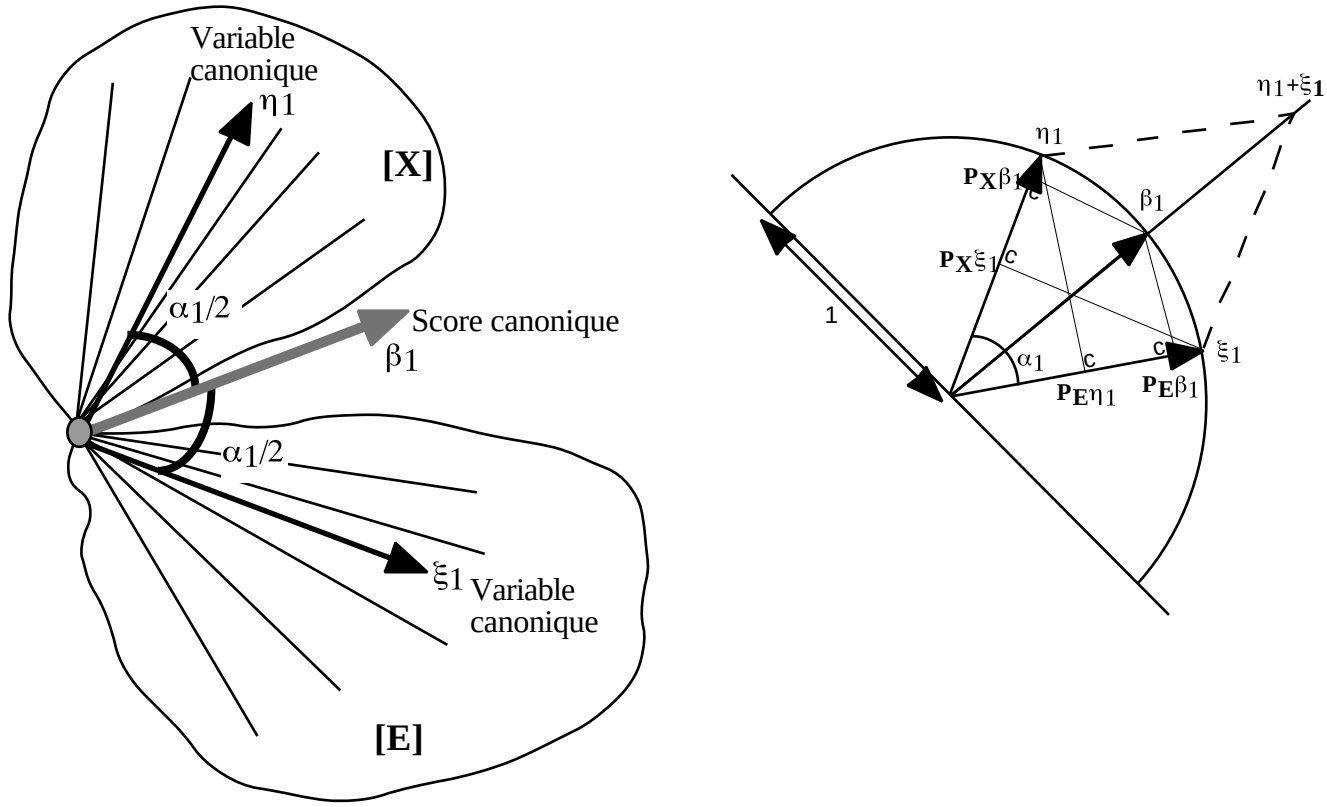


Figure 5 : Représentation schématique de la géométrie inhérente à l'analyse canonique pour le premier axe. Les variables canoniques sont des combinaisons linéaires des variables de [E] et [X] de corrélation maximale (angle α_1 minimal) et le score canonique est porté par la bissectrice de cet angle.

La stratégie $\prod$ conduit aux équations suivantes :

$$\begin{cases} P_X P_E \eta_i = \lambda_i \eta_i \\ P_E P_X \xi_i = \lambda_i \xi_i \end{cases} \qquad \begin{cases} \sqrt{\lambda_i}\,\eta_i = P_X \xi_i \\ \sqrt{\lambda_i}\,\xi_i = P_E \eta_i \end{cases} \qquad \begin{cases} \|\eta_i\| = 1 \\ \|\xi_i\| = 1 \end{cases} \quad \text{avec } \lambda_i = \boldsymbol{cos}^2(\alpha_i)$$

La stratégie $\sum$ conduit à :

$$\begin{cases} (P_X + P_E)\beta_i = \mu_i \beta_i \\ \|\beta_i\| = 1 \end{cases}$$

En multipliant l'équation précédente par $\mathbf{P_E}$ ou $\mathbf{P_X}$ on arrive à :

$$\begin{cases} P_X(P_X + P_E)\beta_i = \mu_i P_X \beta_i \\ P_E(P_X + P_E)\beta_i = \mu_i P_E \beta_i \end{cases} \text{ soit } \begin{cases} P_X P_E \beta_i = (\mu_i - 1) P_X \beta_i \\ P_E P_X \beta_i = (\mu_i - 1) P_E \beta_i \end{cases} \text{ ou encore } \begin{cases} P_X P_E P_X \beta_i = (\mu_i - 1)^2 P_X \beta_i \\ P_E P_X P_E \beta_i = (\mu_i - 1)^2 P_E \beta_i \end{cases}$$

On a donc les relations suivantes à un coefficient multiplicateur près :

$$\begin{cases} \eta_i = P_X \beta_i \\ \xi_i = P_E \beta_i \end{cases}$$

Cette multiplicité de points de vue inhérents à la géométrie de l'AC engendrent de nombreuses possibilités de représentation des résultats. On pourra par exemple utiliser η_i et $\mathbf{P_X}\xi_i$, ou $\mathbf{P_E}\eta_i$ et ξ_i ou bien $\eta_i + \xi_i$. C'est pour cette raison qu'en AFC (analyse canonique de E et Q), on peut indifféremment :

-mettre les classes engendrées par Q à la moyenne des espèces qu'elles abritent

-mettre les espèces à la moyennes des classes de Q qu'elles occupent

-placer les espèces et les classes de Q à la moyenne de leurs occurrences.

Ces différents points de vue sont utiles principalement dans le cas de deux variables qualitatives (nom d'espèce et milieu). Dans le cas d'une variable qualitative (nom d'espèce) et de plusieurs variables quantitatives (milieu) on choisira parmi l'une des trois possibilités. Le couplage d'un tableau issu d'une liste d'occurrences à un tableau de variables mésologiques par analyse canonique conduit à une recherche de combinaison linéaire des indicatrices des espèces et des variables de milieu de corrélation maximale. Les variables canoniques de E forment un codage des occurrences constant par espèce. Par contre, chaque occurrence placée par le milieu a sa position propre. La stratégie la plus claire consiste donc à placer les occurrences par des scores non corrélés de variance unité et de placer les espèces à la moyenne de leurs occurrences. Dans le cas de deux variables qualitatives, pour garder une cohérence générale au raisonnement et pour ne pas perdre la réalité des occurrences, on placera les occurrences par des scores non corrélés de variance unité et on mettra espèces et classes à la moyenne de leurs occurrences. Cette vision conduit au *reciprocal scaling* (Thioulouse & Chessel 1992) qui permet de définir de façon symétrique les amplitudes des niches des espèces et la diversité spécifique des relevés dans l'analyse des tableaux classiques espèces-relevés. Utiliser les variables canoniques comme moyen de représentation graphique c'est privilégier le point de vue de la corrélation entre les deux groupes de variables. Il est préférable, dans notre cas, d'utiliser un score unique pour les occurrences (score canonique) qui maximise la variance de positions moyennes des espèces. Cette stratégie apparaît plus adaptée car elle est en totale adéquation avec les concepts écologiques. Pour conclure, dans le cas des listes d'occurrences on choisira un score des occurrences maximisant la séparation des niches plutôt que deux scores de corrélation maximale.

III. Applications

Afin d'illustrer la théorie présentée dans la partie précédente, nous nous sommes intéressés à l'analyse de deux listes d'occurrences de nature bien différente. Le premier exemple constitue une application de l'analyse d'une liste dont l'ensemble des individus est répertorié sur une parcelle de 1,65 ha. Dans le deuxième exemple, la base de données est formée par la compilation de données de sources diverses à une échelle régionale. La surface de la zone étudiée est d'environ 45 ha et les spécimens collectés ne forment qu'un échantillon de l'ensemble des individus réellement présents. De plus, cette base se distingue de la précédente car elle a mis à contribution un grand nombre de collecteurs (environ 800) sur une période assez longue (de 1822 à 1998).

III.1. Analyse d'une liste exhaustive

Les listes d'occurrences émanent essentiellement de données issues d'herbiers, de collections privées ou muséographiques ou d'atlas. Parallèlement à ces compilations régionales, certaines pratiques propres à la foresterie constituent une autre source importante de listes d'occurrences. Il s'agit de la cartographie d'arbres sur une parcelle, travail indispensable pour l'étude des structures horizontales et verticales du peuplement forestier. Ainsi chaque arbre est identifié par son nom d'espèce, ses coordonnées spatiales, ainsi que sa circonférence, sa hauteur, etc… Il en résulte une liste d'occurrences particulière dont tous des individus sont connus et répertoriés. Le dispositif d'Uppangala (Kadamakal Reserve Forest) géré par l'Institut Français de Pondichéry se situe dans l'état du Karnataka (Inde) au pied des Ghâts occidentaux et est soumis à un climat de type tropical. Cette parcelle de forêt naturelle couvre au total 28 hectares et comporte plusieurs systèmes d'échantillonnage dont un placeau rectangulaire de 110 mètres de large sur 150 mètres de long auquel nous nous sommes intéressés. Tous les arbres de plus de 30 centimètres de circonférence à 1,30 mètre du sol sont identifiés et localisés et certains paramètres tels que l'altitude ou la pente sont mesurés. Il en résulte une liste qui se présente de la façon suivante (Figure 6) :

PLOTHC~1.TXT

	A	B	C	D	E	F	G
1	RefNo	SpCode	XCoord	YCoord	Alti	Slope	Exposition
2	C243	stce	112,6	109,55	38,3	27,9	North-West
3	Hk85	vain	109,85	109,45	37,8	29,1	North-West
4	Hk84	myda	107,95	109,45	38,4	29,1	North-West
5	Hd63	hubr	32	109,45	37,6	14,7	North-East

Figure 6 : Extrait de la liste d'occurrences réalisée à partir de la cartographie d'une parcelle de forêt.

Le jeu de données est constitué de 1097 arbres référencés spatialement appartenant à 69 espèces. Parmi ces 69 espèces, trois d'entre elles (*Vateria indica, Humboldtia brunonis, Myristica dactyloides*) représentent près de la moitié des occurrences (48,77%). Le profil des abondances des espèces met en avant le grand nombre d'espèces rares (43 espèces comportent moins de 7 individus) caractéristique des forêts tropicales (Figure 7).

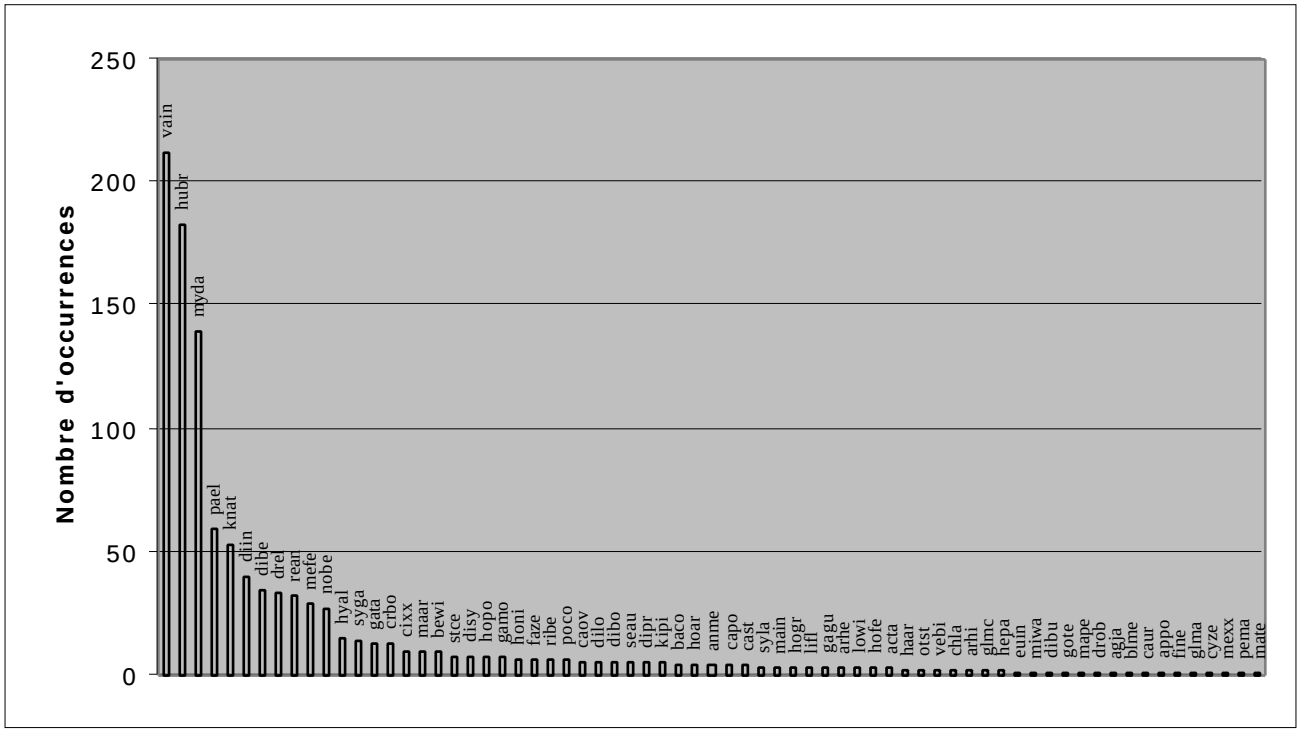


Figure 7 : Distribution des abondances des 69 espèces

Par la suite, notre analyse sera axée exclusivement sur l'étude de la structure horizontale du peuplement. Les méthodes destinées à décrire une structure multispécifique à partir d'occurrences spatialisées sont assez rares malgré l'ancienneté des listes d'occurrences en foresterie. Gittins (1968) utilise des coordonnées (x,y) et introduit la *trend surface analysis* en écologie. Cette méthode permet, à l'aide de polynômes des coordonnées (x, y, xy, x^2, …) de mettre en évidence l'existence de structures spatiales. Wartenberg (1985) emploie ces polynômes dans une analyse canonique (*canonical trend surface analysis*) afin d'étudier les relations existantes entre des variables spatiales et biologiques. Borcard *et al.* (1992) réalisent une analyse canonique des correspondances, comme le suggère TerBraak (1987), et une analyse des redondances à partir de ces polynômes. Ce type d'approche apparaît satisfaisant lorsque la zone d'étude est globalement homogène et lorsque l'échantillonnage est assez régulier (Thioulouse *et al.* 1995). Pour analyser la structure de la parcelle étudiée, une première approche consiste donc à confronter la liste d'occurrences à un tableau formé par des polynômes des coordonnées de chacun des arbres et ainsi réaliser ce que l'on appellera une *canonical correspondence trend surface analysis.* Le polynôme choisi est de degré 4 et on obtient alors 14 variables qui sont : x, y, x^2, xy, y^2, x^3, x^2y, xy^2, y^3, x^4, x^3y, x^2y^2, xy^3, y^4. Ces variables sont centrées-réduites et se distribuent en décrivant diverses structures spatiales (Annexe A3). Par exemple, les formes cubiques et quadratiques ainsi que leur produits croisés permettent de décrire des structures complexes telles que des zones d'agrégation locale. Le choix du degré du polynôme est subjectif mais il est possible de réduire ce nombre de variables à l'aide d'une extension multivariée d'une régression pas à pas (Borcard *et al.* 1992). Dans notre cas, nous sommes largement dans le domaine de validité de l'analyse canonique (1097 observations pour 69 et 14 variables) et nous allons donc garder l'ensemble du polynôme afin de conserver une bonne précision pour analyser ces patterns spatiaux. Un test de permutation multivarié est réalisé et justifie la description des différences entre les répartitions spatiales des espèces (Manly 1991, Figure 8). Des analyses multi-échelles de la répartition spatiale (Ripley

1976 et 1977, Besag 1977) des espèces les plus abondantes confirment l'existence de structures spatiales pour la plupart d'entre elles (Annexe A4).

```
|
|
|**
|********
|******************
|*******************************
|*****************************************
|*************************************************
|**********************************************
|*********************************
|************************
|******************
|********  number of random permutations: 5000   Observed:  1.123e+00
|******    Histogram:  minimum =  6.455e-01, maximum =  1.196e+00
|**        number of simulations X<Obs: 4993 (frequency:  9.986e-01)
|*         number of simulations X>=Obs: 7 (frequency:  1.400e-03)
|
•->|  Observed
|
|
```

Figure 8 : Test de permutation multivarié. On calcule pour chaque répartition aléatoire des individus dans les groupes (espèces) l'inertie interclasse totale avec la métrique de Mahalanobis. Une valeur significativement grande indique une ressemblance globalement plus forte entre individus d'une même espèce. En ordonnée est indiqué les différentes valeurs d'inertie interclasse obtenues pour les permutations (*) et pour les données observées (->).

Les résultats de l'analyse canonique entre les indicatrices des 69 espèces et les polynômes formés par leur coordonnées laissent apparaître une structure que l'on interprétera seulement sur le premier axe (Figure 9a). La Figure 9b représente le premier score factoriel de l'analyse. Les courbes de niveaux sont tracées grâce à une régression locale à deux dimensions sur 19 voisins. Le nombre de voisins est choisi grâce à la fonction E(n) (Figure 9c). Cette fonction calcule l'erreur commise quand on estime par cette méthode les données à l'aide des n points voisins (Cleveland & Devlin 1988). Le premier axe de l'analyse met en opposition la zone centrale de la parcelle aux zones latérales.

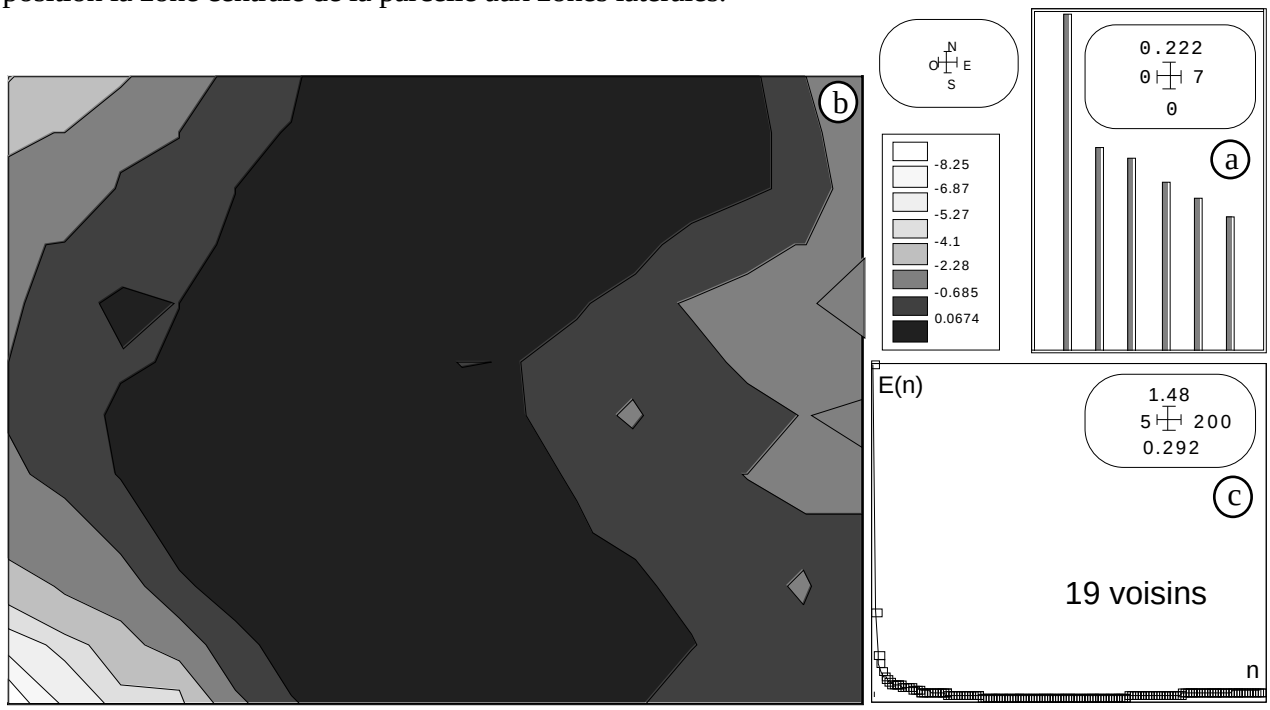


Figure 9 : Résultats de l'analyse canonique entre le tableau disjonctif complet formé par les indicatrices des espèces et les variables induite par le polynôme. Le graphe des valeurs propres (a) indique que l'interprétation ne peut se faire que

sur le premier axe. Les scores factoriels des occurrences sur le premier axe sont représentés (b) par régression locale sur 19 voisins. Le nombre de voisins choisi est la valeur n minimisant la fonction E(n) (c, cf. texte).

La zone à l'extrême Sud-Ouest possède une structure particulière caractérisée par la présence de nombreuses espèces rares qui occupent exclusivement cette partie de la parcelle. La figure 10 présente l'ordination des espèces obtenue sur le premier axe. Cette typologie peut conduire à la définition de 3 groupes :

- 1 : Les espèces occupant exclusivement la zone centrale,
- 2 : Les espèces présentes sur l'ensemble de la parcelle,
- 3 : Les espèces se distribuant sur les zones Est et/ou Ouest.

Les résultats obtenus concordent avec ceux provenant des analyses multi-échelle de la répartition spatiale réalisées sur les espèces les plus abondantes (Annexe A4) : les espèces des groupes 1 et 3 ont des répartitions non aléatoires. Pour les espèces du groupe 2, quatre d'entre elles ont des répartitions aléatoires et quatre autres ont des répartitions agrégées (*Dimorphocalyx beddomei*, *Humboldtia brunonis*, *Myristica dactyloides*, *Vateria indica*).

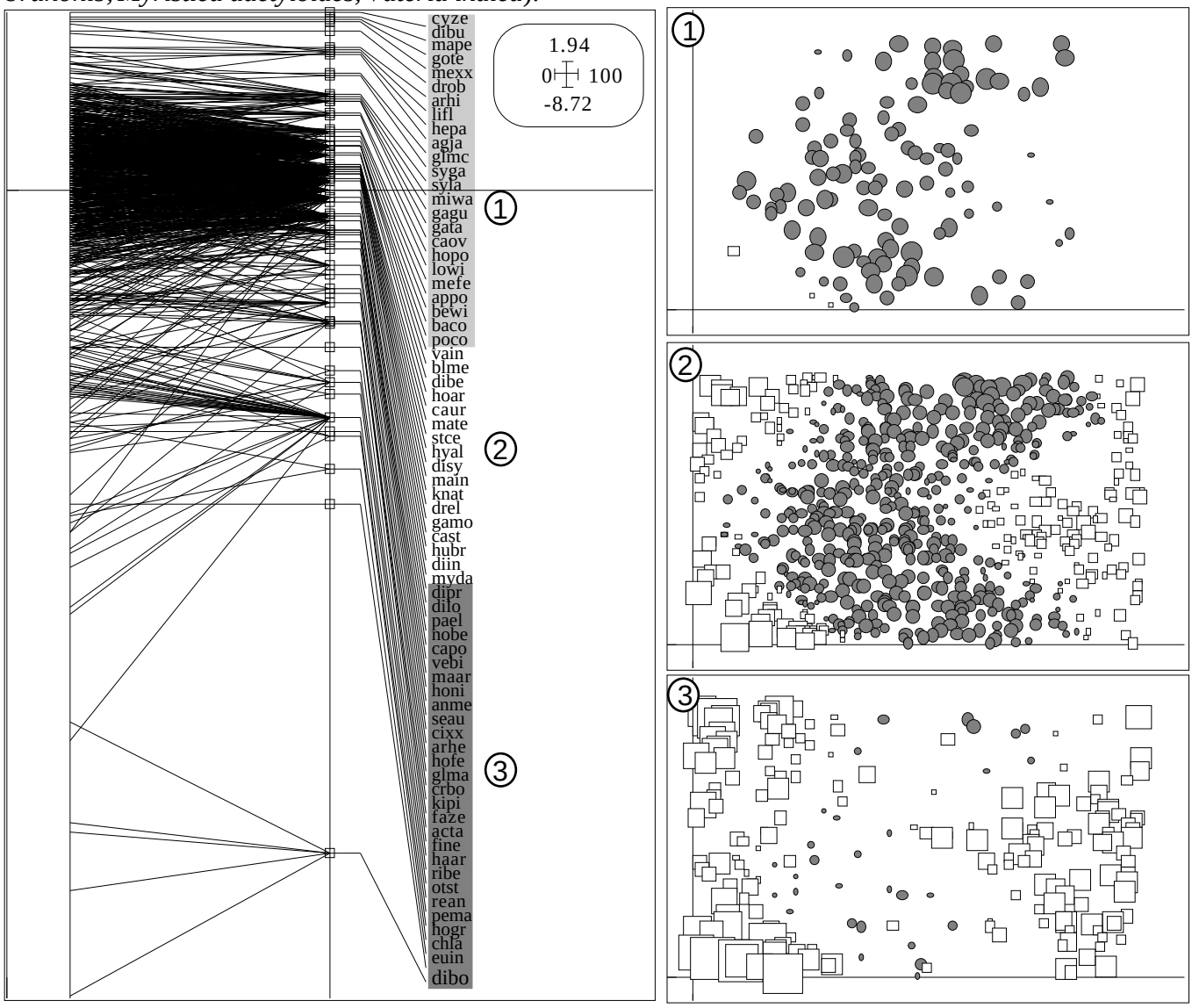


Figure 10 : Ordination des espèces sur le premier axe de l'analyse canonique entre le tableau disjonctif complet formé par les indicatrices des espèces et les variables induite par le polynôme. Les espèces sont placées à la moyenne de leurs occurrences. Les espèces sont alors divisées en 3 groupes : (1) : Les espèces occupant exclusivement la zone centrale, (2) : Les espèces présentes sur l'ensemble de la parcelle, (3) : Les espèces se distribuant sur les zones Est et/ou Ouest. Les distributions spatiales des occurrences de chacun des groupes sont indiquées ainsi que leurs score sur le premier axe

factoriel (les ronds représentent les scores positifs et les carrés les scores négatifs ; la taille du symbole est proportionnelle à la valeur absolue du score de l'occurrence).

Il s'avère que ces quatre espèces se situent très près des limites du groupe 2 : elles occupent la zone centrale de la parcelle et une seule autre partie à l'Est ou à l'Ouest de cette zone (Annexe A2). Leur répartition n'est donc pas aléatoire sur l'ensemble de la parcelle mais elles occupent en partie les deux grandes zones définies par l'analyse.

L'existence d'une structuration spatiale est donc évidente et les facteurs qui pourrait en être à l'origine sont essentiellement de deux types :

- écologiques (exigences édaphiques de l'espèce)
- sociologiques (mode de développement de l'espèce et relations inter-espèces).

De plus, il semble indéniable que ces deux facteurs interagissent et que d'autres interviennent (facteurs historiques notamment). Pour l'ensemble des arbres, nous disposons dans la base des mesure d'altitude et de pente (calculée par carré de 10 m de côté). Ainsi, il est possible d'établir des cartes de la parcelle pour ces deux variables afin de mettre en évidence une éventuelle influence sur la répartition des différentes espèces. En comparant les figures 9b et 11a,11c on remarque que les limites de la zone centrale correspondent à des zones de forte pente. L'effet de ces deux variables topographiques semble donc bien réel mais il est plus que probable qu'il soit plutôt indirect : la pente et l'altitude influent sur l'exposition et le taux d'humidité favorisant (ou non) ainsi le développement de certaines espèces. C'est certainement une combinaison de plusieurs variables (dont la pente et l'altitude) qui contraint la répartition des espèces et pas seulement ces deux variables étudiées. Il serait donc intéressant de pouvoir disposer d'autres mesures mésologiques afin d'estimer l'importance de chacun de ces facteurs sur la distribution de ces espèces. Si l'influence des conditions environnementales est mieux connue, il serait alors possible d'étudier les relations "sociologiques" que peuvent entretenir différentes espèces entre elles (mise à distance, compétition,…).

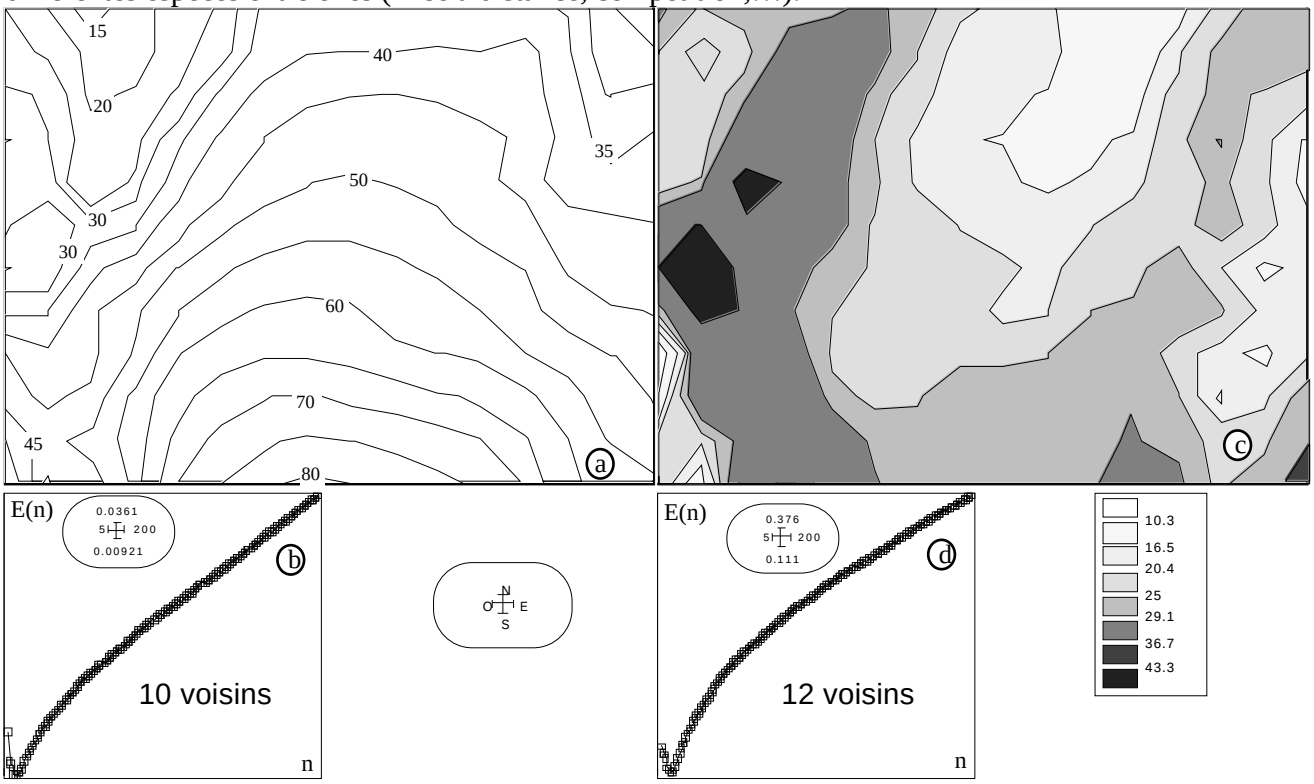


Figure 11 : Représentation de l'altitude (a) et de la pente (c) sur la parcelle grâce à des régressions locales sur 10 et 12 voisins (valeurs de n minimisant E(n)). Pour l'altitude, les courbes de niveaux sont équidistantes de 5 mètres. Les pentes

sont indiquées en pourcentage ex : une pente de 3% définit une différence de niveaux de 3 m sur 100 m de distance horizontale).

L'analyse de patterns multispécifiques, en appliquant l'analyse canonique sur des listes d'occurrences issues de la foresterie, fournit des cartes résumant l'organisation horizontale d'un peuplement forestier. L'utilisation de polynôme permet d'identifier des patterns existants à différentes échelles (plus le degré est grand, plus l'échelle est locale). L'introduction de variables mésologiques dans l'analyse pourrait permettre de mieux comprendre l'influence de chacune d'entre elles et il serait ainsi possible d'étudier les relations inter-espèces en éliminant l'influence du milieu.

III.2. Analyse d'une liste échantillonnée

Cette deuxième application fournit une illustration de l'utilisation de l'analyse canonique sur des listes d'occurrences à l'échelle régionale. Les données proviennent de l'inventaire des Carabidae de la région Rhône-Alpes (environ 45 000 km^2) réalisé par le Réseau Entomologique Rhône-Alpin (R.E.R.A.) et le Muséum d'Histoire Naturelle de Lyon en collaboration avec la région Rhône-Alpes. Cet inventaire a été élaboré grâce à la compilation de données issues de collections privées et muséographiques en mettant à contribution un grand nombre d'entomologistes amateurs de la région Rhône-Alpes. Il en résulte une base de données contenant près de 35000 occurrences appartenant à 578 espèces dont l'objectif principal est la réalisation d'un atlas à paraître sous peu. L'analyse de ces données dans le cadre d'un DEA n'ayant pas été prévue lors de la réalisation de la base, certaines données ont été censurées par respect envers les collecteurs qui n'étaient pas prévenus. Ainsi, il a été convenu qu'aucun nom de personne physique ne devait être mentionné dans mon travail et que l'utilisation d'un nom d'espèce était autorisée sous réserve que les positions de ses occurrences n'étaient pas indiquées avec précision.

Ces données sont issues de diverses sources et il apparaît évident que la pression d'échantillonnage n'est pas uniforme sur l'ensemble de la région d'étude. Afin de rendre compte de cette hétérogénéité, la région a été divisée en 156 quadrats et le nombre d'occurrences répertoriées pour chaque quadrat a été calculé. D'après la figure 12, les alentours de la région lyonnaise forment une zone de très forte densité d'occurrences. Cette situation est certainement due à une forte intensité de prospection de cette partie de la région mais également à une plus grande participation des entomologistes lyonnais à l'élaboration de cette base de données. Cependant, il ne nous est pas possible de confirmer (ou d'infirmer) ces hypothèses et il est donc évident que notre étude ne doit considérer que les présences notifiées car l'absence d'une espèce dans un quadrat peut avoir une origine écologique (le milieu ne satisfait pas les exigences de l'espèce) ou d'autres causes qu'il est impossible de prendre en compte dans une analyse. Ceci montre bien les biais que pourraient induire l'utilisation de quadrats : les zones fréquemment visitées présenteraient certainement une richesse spécifique exagérément grande par rapport à des zones négligées par les entomologistes (Figure 12a). En effet, la relation entre le nombre d'occurrences et la richesse spécifique d'un quadrat est très étroite (coefficient de corrélation de Spearmann ajusté pour les ex-aequos, ρ=0,9631 (p=0,0001), Figure 12b) et de la forme des fonctions d'accumulation décrites par Soberón & Llorente (1993). L'utilisation de l'analyse canonique directement sur les

occurrences semble donc une alternative indispensable pour éliminer ces problèmes posés par l'utilisation de quadrats.

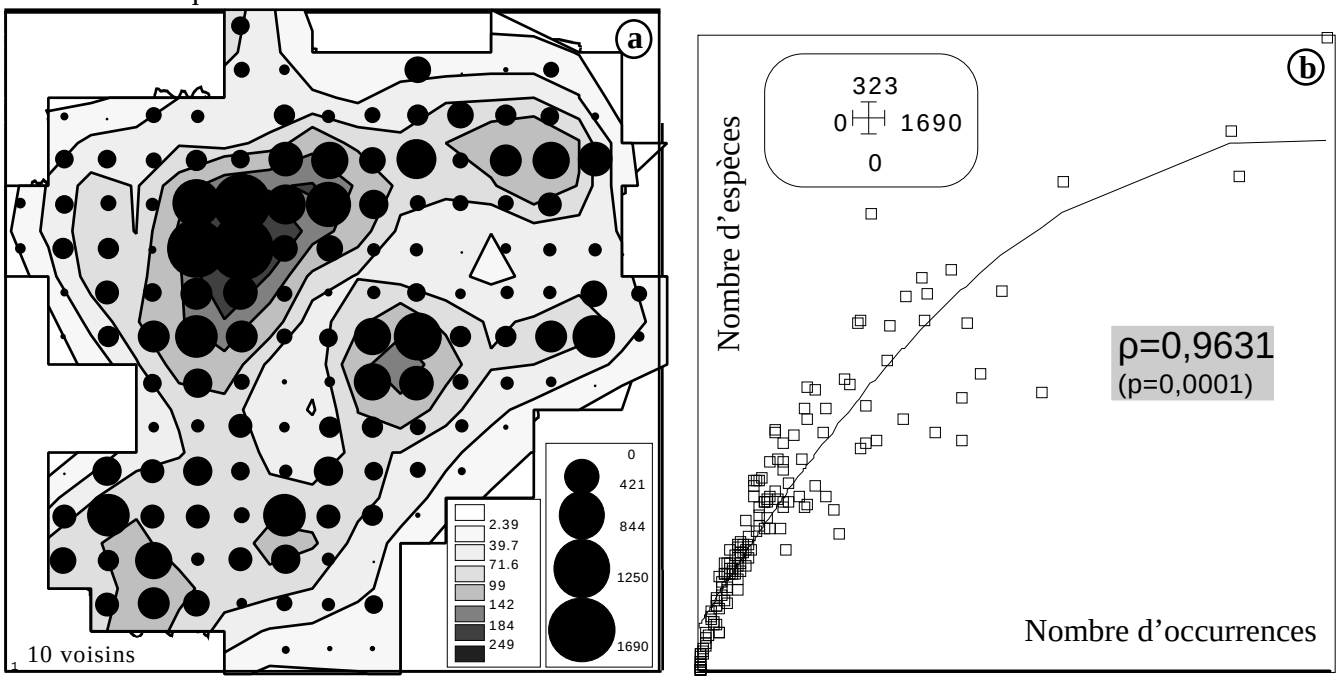


Figure 12 : Lien entre le nombre d'occurrences et le nombre d'espèces. La région est divisée en 156 quadrats dans lesquels le nombre d'occurrences notifiées et le nombre d'espèces sont calculées. Le nombre d'occurrences par quadrat est représenté par des cercles noirs et la richesse spécifique est représentée par des courbes de niveaux calculées par régression locale à deux dimensions sur 10 voisins (a). En (b), est indiquée la relation significative (coefficient de corrélation de Spearmann) entre le nombre d'occurrences et le nombre d'espèces par quadrat.

La problématique induite par les listes échantillonnées est radicalement différente de celle inhérente aux listes issues des pratiques de foresterie. Les relations inter-espèces sont, dans le cas présent, totalement ignorées du fait de la taille de la région étudiée et du caractère particulier des données. Notre étude est donc axée uniquement sur les relations espèces-milieu en négligeant totalement l'existence de communautés. Comme dans l'exemple précédent, nous avons réalisé une *canonical correspondence trend surface analysis* à l'aide d'un polynôme de degré 5 dont les variables sont représentées en Annexe B1. Les résultats de cette analyse nous ont amené à observer les structures décrites sur les trois premiers axes dont les valeurs propres associées valent, respectivement, 0.51, 0.27 et 0.25 (Figure 13a). Comme dans l'exemple précédent, les scores factoriels sont représentés à l'aide de régression locale à deux dimensions (cf. III.1.). Le premier axe met en opposition la zone centrale de la région aux parties Est et Ouest (Figure 13b). La structure ainsi définie correspond exactement à la topographie observée dans la région (Figure 15a).

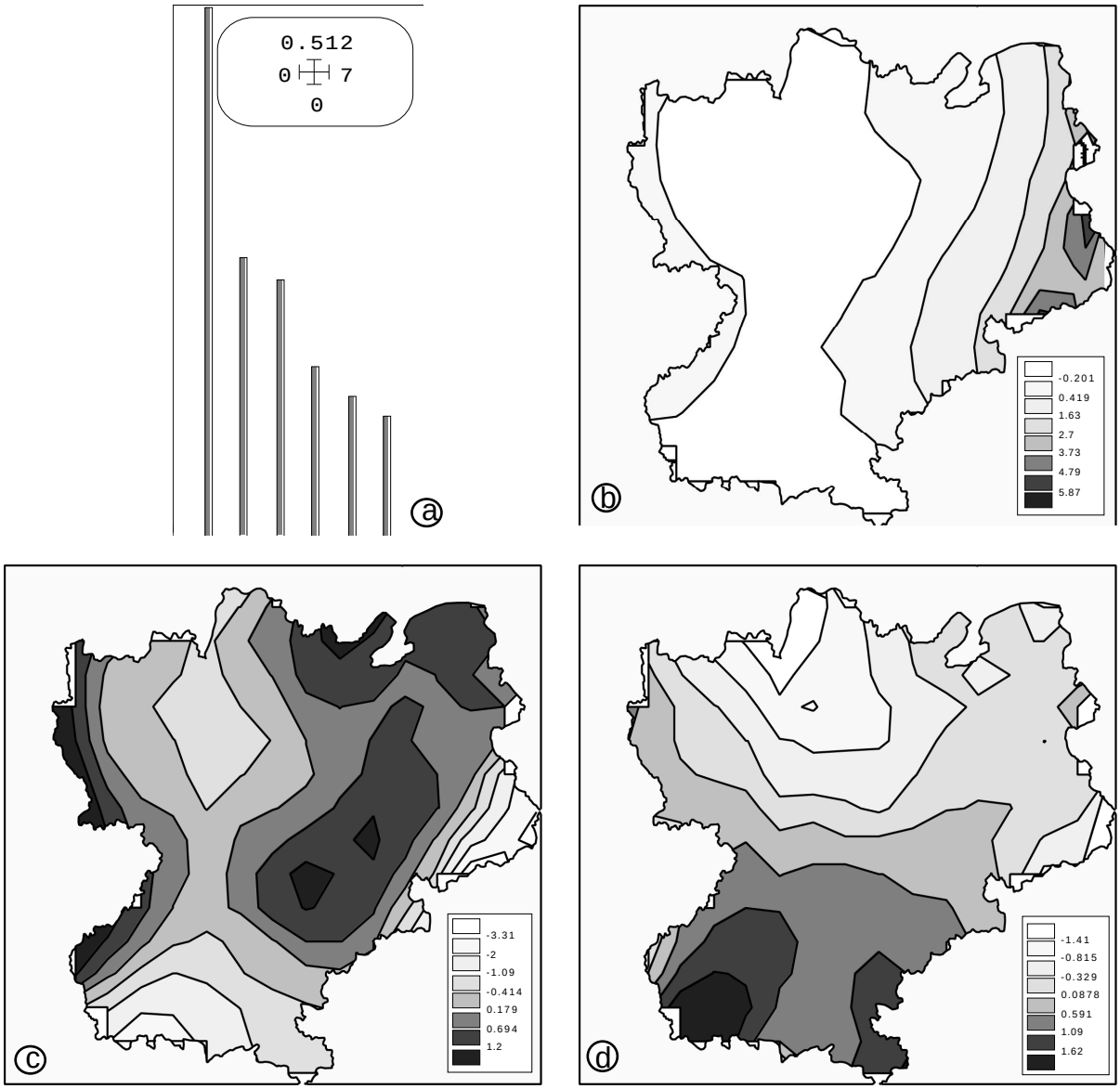


Figure 13 : Résultats de la *canonical correspondence trend surface analysis.* Le graphe des valeurs propres (a) conduit à une description des structures observées sur les trois premiers axes. Les scores des occurrences sur l'axe 1 (b), 2 (c) et 3 (d) de l'analyse sont représentés par régression locale à deux dimensions sur 600 voisins.

Sur la partie Est s'étend le massif alpin alors qu'à l'Ouest on retrouve une partie des Cévennes (Sud-Ouest) ainsi que les monts du Forez (Nord-Ouest). Ce premier facteur reflète donc l'influence déterminante de l'altitude sur la structuration du peuplement de Carabidae. Ainsi, il permet de séparer les espèces de plaine telles que *Asaphidion curtum* (score moyen des occurrences sur le premier axe S1=-0.65) ou *Microlestes minutulus* (S1=-0.49) des espèces de montagnes comme *Platycarabus depressus* (S1=4.4) ou *Cicindela gallica* (S1=2.81).Ce fait a déjà été mis en évidence sur des populations de Carabidae en Wallonie par Dufrêne (1992). Le plan formé par les deux premiers axes laisse apparaître une structure en fer à cheval (effet Guttman) très marquée s'expliquant par une corrélation non linéaire entre ces deux axes (Figure 14). L'existence d'une telle structure dans une analyse multivariée est souvent due à un artefact mathématique (Hill & Gauch 1980). Dans le cas présent, il est probable que cet effet Guttman n'est pas un artefact mais provienne de la structure existante dans les données. Les scores obtenus sur le deuxième axe laisse présager une influence des précipitations sur les variations de composition spécifique (Figure 15b). Si tel est le cas, il est alors tout à fait logique que le premier axe (gradient altitudinal) et le deuxième (influence des précipitations)

soit corrélés. Ainsi, le rôle des précipitations est observé surtout dans les zones de moyennes et hautes altitudes permettant notamment de séparer le Sud-Est des Alpes (de 800 à 1400 mm de précipitations annuelles) de la Chartreuse, Belledonne, Bauges, Chablais-Giffre, Bornes, Haut-Jura (la situation de ces différents districts est indiquée en Annexe B2) caractérisés par de fortes précipitations annuelles (jusqu'à 2000 mm).

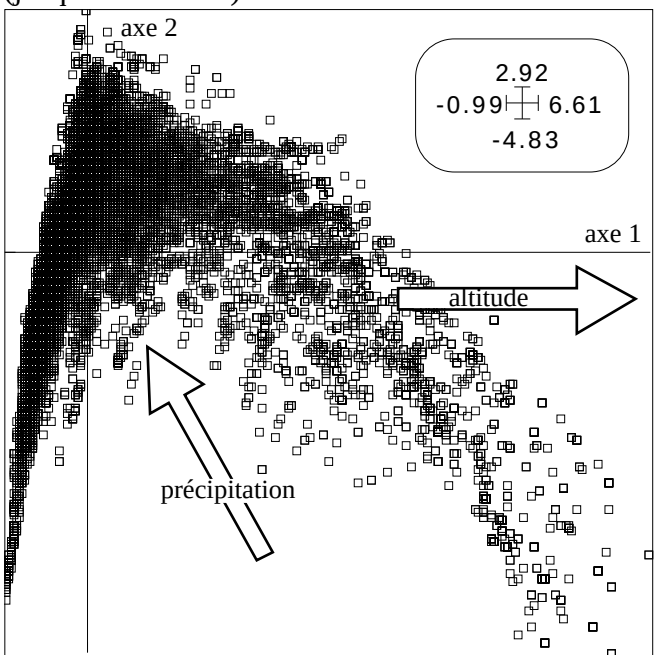


Figure 14 : Représentation des scores des occurrences sur le plan formé par les deux premiers axes de l'analyse. On observe un effet d'arche signifiant une liaison non linéaire entre les scores sur les deux axes.

Dans la partie Ouest de la région, cet axe permet également d'identifier certaines zones de fortes précipitations (Cévennes et une partie du Vivarais notamment). Au niveau des espèces, ce deuxième facteur sépare les espèces préférant, à une échelle régionale, les zones de fortes précipitations (*Pterostichus hagenbachi* (S2=1.75) ou *Cicindela sylvicola* (S2=1.57)) de celles préférant les secteurs plus arides (*Cicindela maroccana* ssp. *pseudomaroccana* (S2=-1.39)).

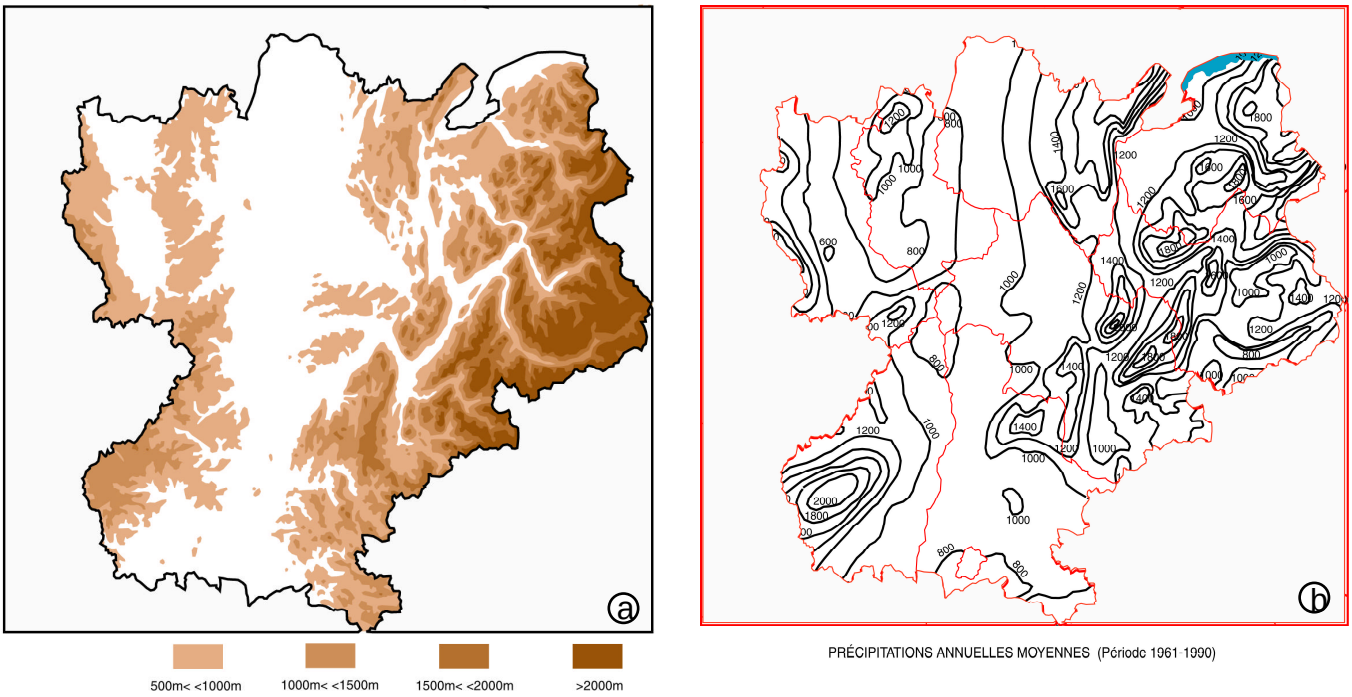


Figure 15 : Relief (a) et précipitations (b) de la région Rhône-Alpes. (d'après Richoux *et al.* (1999) et Blanchet & Richoux (1999)).

Enfin, le troisième axe décrit une structure qui s'étale dans le sens latitudinal et qui reflète certainement l'opposition entre des espèces de type méditerranéen (*Chrysocarabus hispanus* (S3=2.14), *Speotrechus mayeti* (S3=3.47), *Cicindela maroccana* ssp. *pseudomaroccana* (S3=2.7)) qui occupent le sud de la région et des espèces vivant sous un climat continental (*Melanius anthracinus* (S3=-1.31), *Eupetedromus dentellum* (S3=-1.28)) présentes dans le reste de la région (Figure 13d). Dans la zone d'influence méditerranéenne, la séparation entre le centre et les deux parties extérieurs permet également de dissocier les espèces méditerranéennes de moyenne altitude (Cévennes, Ventoux) des espèces de basse altitude.

L'analyse de l'inventaire des Carabidae par une *canonical correspondence trend surface analysis* a permis de faire ressortir les principales structures observées dans le peuplement rhône-alpin. L'influence de paramètres environnementaux, tels que l'altitude ou les précipitations, semble indéniable mais l'utilisation de polynômes ne permet pas de quantifier l'importance de chacun d'entre eux. Il aurait été beaucoup plus judicieux d'utiliser les variables de milieu en lieu et place des polynômes afin d'obtenir une ordination directement basée sur les variables environnementales et ainsi, d'éviter l'étape de comparaison des cartes factorielles et des cartes de milieu. Pour ce faire, il est nécessaire que chaque occurrence possède des mesures pour l'ensemble des variables étudiées or ce n'était pas le cas pour cette base de données. L'utilisation du Système d'Information Géographique (SIG) permettrait d'obtenir des mesures estimées pour l'ensemble des occurrences en croisant la base des Carabidae avec d'autres bases déjà établies pour la région (climatologie, végétation, …). De plus, le SIG autorise une représentation graphique des résultats plus performante que celle obtenue à partir de logiciels de statistique. Le couplage du SIG et de la statistique multivariée semble donc une alternative très prometteuse pour l'étude des relations espèces-milieu à partir de listes d'occurrences.

Conclusions

Une liste d'occurrences introduit une variable qualitative (nom d'espèce) qui se résume dans un tableau disjonctif complet. Ce type de données possède un certain nombre de particularités qu'il est nécessaire de prendre en compte lors de leur étude. La confrontation d'une liste d'occurrences à de l'information externe relève de la statistique multivariée. Lorsque cette information est une variable qualitative, plusieurs stratégies sont envisageables. Deux de ces analyses sont dissymétriques (analyses sur variables instrumentales) et l'autre est symétrique (analyse canonique des deux tableaux d'indicatrices). Cette analyse canonique est tout simplement l'AFC de la table croisant les deux variables qualitatives qui est également une double analyse discriminante.

Lorsque l'on confronte la liste d'occurrences à un tableau de variables quantitatives (milieu), la stratégie d'analyse canonique est alors parfaitement valide. Une partie de cette analyse correspond à l'analyse discriminante du tableau des variables de milieu par la variable nom d'espèce. L'analyse canonique considérée cherche à séparer au mieux les niches des espèces et sa procédure correspond alors parfaitement avec celle de l'analyse canonique des correspondances. L'ACC est donc une véritable analyse canonique dans le cas des listes d'occurrences et son nom se justifie alors *a posteriori*. L'analyse canonique trouve dès lors un domaine d'application, en écologie, pour lequel elle est parfaitement adaptée théoriquement et pratiquement.

La représentation graphique utilisée dans ce type d'analyse s'appuie alors sur une double projection du score canonique sur les deux espaces engendrés et donne le *reciprocal scaling* dans le cas de deux paquets d'indicatrices.

L'introduction des listes d'occurrences et leur étude en analyse de données remet la notion de correspondances au centre du débat. Le lien entre analyse canonique et liste d'occurrences est très fort et implique certaines conséquences sur les méthodes classiquement appliquées aux tableaux espèces-relevés. Utiliser une stratégie d'analyse des correspondances (AFC ou ACC) revient à ne considérer que les occurrences réellement observées. L'analyse des tableaux floro-faunistiques classiques par ces méthodes remet en cause la notion de relevé comme individu statistique. Par exemple, si les relevés sont issus de différentes techniques de capture ou fonction des conditions météorologiques l'emploi d'une analyse des correspondances se justifie pour n'utiliser que les individus présents. Par contre, dans le cas de facteurs limitants ou de pollution, l'utilisation de l'AFC ou de l'ACC peut s'avérer catastrophique car on ne tiendra pas compte des absences d'une ou plusieurs espèces dues à ce facteur. Voir dans un tableau espèces-relevés une liste d'occurrences, en utilisant une analyse des correspondances, induit des conséquences théoriques assez importantes. En effet, l'individu statistique à considérer est alors l'occurrence et, par conséquent, on nie l'existence de communautés puisque l'on fait alors l'hypothèse d'indépendance des individus.

Cette situation révèle l'importance de l'introduction des caractéristiques d'un type de données dans le modèle inhérent à la méthode que l'on veut appliquer. Il est alors possible de faire évoluer les méthodes et de remettre en cause celles existantes. L'existence d'un langage universel permettant de comparer l'ensemble des méthodes semble alors inévitable. Le schéma de dualité et l'utilisation des projecteurs apparaît naturellement comme un moyen simple et efficace pour pallier ce besoin.

L'intérêt pour les listes d'occurrences en écologie ou en biogéographie est indéniable. La constitution de listes d'occurrences à partir de compilations de données muséographiques permet d'étudier la composition spécifique et les aires de distribution des espèces à une échelle régionale voire continentale et l'utilisation de l'analyse canonique apparaît comme un moyen efficace de s'affranchir des intermédiaires (quadrats) habituellement employés dans ce genre d'étude. Ce type d'analyse revalorise les collections contenues dans les muséums et autorise des études spatio-temporelles de grande ampleur pour un coût raisonnable. L'étude de la biodiversité et de ses variations à travers le temps et l'espace, problématique majeure de l'écologie actuelle, pourra alors être effectuée en s'affranchissant des problèmes de pression d'échantillonnage présents dès que l'on utilise des quadrats. Pour réaliser ce type d'étude, le couplage de la statistique multivariée et du Système d'Information Géographique (SIG) semble très prometteur afin d'incorporer de nouvelles variables dans l'analyse et d'obtenir une représentation efficace des résultats.

Dans le cas des listes exhaustives d'occurrences, l'étude des processus de structuration horizontale des forêts par une *canonical correspondence trend surface analysis* permet d'obtenir un résumé des variations de composition spécifique à travers l'espace. L'utilisation de l'analyse canonique partielle et la prise en compte des relations interspécifiques pourrait permettre une meilleure compréhension des processus responsables des structures observées. L'analyse canonique partielle permettrait, en effet, de dissocier l'effet de variables environnementales d'un effet purement spatial. Il serait alors possible d'incorporer des facteurs sociologiques identifiant le voisinage d'un individu. Cependant, la

quantification du voisinage d'un individu semble assez délicate car elle doit tenir compte à la fois de la distance et de l'espèce des voisins. L'emploi de polynômes apparaît comme une solution intéressante mais il ne permet pas d'analyser les patterns existants à toutes les échelles d'observation car les phénomènes locaux ne peuvent être appréhendés qu'à l'aide de polynôme de grand degré. Le nombre de variables augmente alors très rapidement et le risque de sortir du domaine de validité de l'analyse canonique semble, dans ce cas-là, bien réel. Le développement de métriques d'échelle apparaît comme une alternative, que nous souhaitons développer, à l'utilisation de polynômes avec un grand nombre de variables. Ainsi, en définissant pour chaque échelle d'étude une métrique différente, il serait alors possible d'étudier simultanément les phénomènes locaux et globaux.

Bibliographie

AUSTIN, M.P. (1968) An ordination study of a chalk grassland community. *Journal of Ecology :* 56, 739-757.

AUSTIN, M.P. (1976) On non-linear species response models in ordination. *Vegetatio :* 33, 33-41.

BEALS, E.W. (1973) Ordination : mathematical elegance and ecological naïveté. *Journal of Ecology :* 61, 23-35.

BENZECRI, J.P. (1969) Statistical analysis as a tool to make patterns emerge from data. *In : Methodologies of pattern recognition.* **WATANABE, S. (ED.)** Academic Press, New-York. 35-60.

BESAG, J. E. (1977) Comments on Ripley's paper. *Journal of the Royal Statistical Society, B :* 39, 193-195.

BIRKS, H.J.B., PEGLAR, S.M. & AUSTIN, H.A. (1996) An annotated bibliography of canonical correspondence analysis and related constrained ordination methods 1986-1993. *Abstracta Botanica :* 20, 17-36.

BLANCHET, G & RICHOUX, P. (1999) Quelques aspects du climat de la région Rhône-Alpes. *Bulletin mensuel de la Société linéenne de Lyon* (sous presse).

BORCARD, D., LEGENDRE, P. & DRAPEAU, P. (1992) Partialing out the spatial component of ecological variation. *Ecology :* 73, 1045-1055.

BROWN, J. W. & OPLER, P. A. (1990) Patterns of butterfly species density in peninsular Florida. *Journal of Biogeography :* 17, 615-622.

CAILLIEZ, F. & PAGES, J.P. (1976) *Introduction à l'analyse des données.* SMASH, 9 rue Duban, 75016 Paris. 1-616.

CHESSEL, D., LEBRETON, J.D. & YOCCOZ, N. (1987) Propriétés de l'analyse canonique des correspondances. Une utilisation en hydrobiologie. *Revue de Statistique Appliquée :* 35, 55-72.

CHESSEL, D. & MERCIER, P. (1993) Couplage de triplets statistiques et liaisons espèces-environnement. *In : Biométrie et Environnement.* **LEBRETON, J.D. & ASSELAIN, B. (EDS.)** Masson, Paris. 15-43.

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

CRONK, Q. C. B. (1989) The past and the present vegetation of Saint Helena. *Journal of Biogeography :* 16, 47-64.

DIGBY, P. G. N. & KEMPTON, R. A. . (1987) *Multivariate Analysis of Ecological Communities.* Chapman and Hall, Population and Community Biology Series, London. 1-205.

DUFRENE, M. (1992) *Biogéographie et écologie des communautés de Carabidae en Wallonie.* Thèse de doctorat, Université Catholique de Louvain. 1-197 + Figures.

ESCOUFIER, Y. (1987) The duality diagramm : a means of better practical applications. *In : Development in numerical ecology.* **LEGENDRE, P. & LEGENDRE, L. (EDS.)** NATO advanced Institute , Serie G .Springer Verlag, Berlin. 139-156.

FERRARI, C., BONAFEDE, F. & ALESSANDRINI, A. (1993) Rare plants of the Emilia-Romagna region (Northern Italy): a data bank and computer-mapped atlas for conservation purposes. *Biological Conservation :* 64, 11-18.

GAUCH, H.G. JR. & WENTWORTH, T.R. (1976) Canonical correlation analysis as an ordination technique. *Vegetatio :* 33(1), 17-22.

GIMARET-CARPENTIER, C., CHESSEL, D. & PASCAL, J.P. (1998a) Non-symetric correspondence analysis: an alternative for species occurrences data. *Plant Ecology :* 138, 97-112.

GIMARET-CARPENTIER, C. (1999) *Analyse de la biodiversité à partir d'une liste d'occurrences d'espèces : nouvelles méthodes d'ordination appliquées à l'étude de l'endémisme dans les Ghats occidentaux.* Thèse de doctorat, Université Lyon 1. 1-235 + Annexes.

GITTINS, R. (1968) Trend-surface analysis of ecological data. *Journal of Ecology :* 56, 845-869.

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

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

GREEN, R.H. . (1974) Multivariate niche analysis with temporally varying environmental factors. *Ecology :* 55, 73-83.

HILL, M.O. & GAUCH, H.G. . (1980) Detrended correspondence analysis: an improved ordination technique. *Vegetatio :* 42, 47-58.

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

HILL, M. O. (1991) Patterns of species distribution in Britain elucidated by canonical correspondence analysis. *Journal of biogeography :* 18, 247-255.

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

HOTELLING, H. (1936) Relations between two sets of variates. *Biometrika :* 28, 321-377.

JONGMAN, R.H., TERBRAAK, C.J.F. & VAN TONGEREN, O.F.R. (1987) *Data analysis in community and landscape ecology.* Pudoc, Wageningen. 1-298.

KADMON, R. & DANIN, A. (1997) Floristic variation in Israel: a GIS analysis. *Flora :* 192, 341-345.

KESSEL, S. R. & WHITTAKER, R. H. (1976) Comparisons of three ordination techniques. *Vegetatio :* 32(1), 21-29.

LAURO, N. & D'AMBRA, L. (1984) L'analyse non symétrique des correspondances. *In : Data Analysis and Informatics III.* **DIDAY, E. & COLL. (ED.)** Elsevier, North-Holland. 433-446.

LEBRETON, J.D., CHESSEL, D., PRODON, R. & YOCCOZ, N. (1988) L'analyse des relations espèces-milieu par l'analyse canonique des correspondances. I. Variables de milieu quantitatives. *Acta Œcologica, Œcologia Generalis :* 9(1), 53-67.

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 instudies of structure - activity and species - environment relatioships. *In : Applied Multivariate Analysis in SAR and Environmental Studies.* **DEVILLERS, J. & KARCHER, W. (EDS.)** Kluwer Academic Publishers. 85-114.

LEGAKIS, A. & KYPRIOTAKIS, Z. (1994) A biogeographical analysis of the island of Crete, Greece. *Journal of Biogeography :* 21, 441-445.

MANLY, B.F.J. (1991) *Randomization and Monte Carlo methods in biology.* Chapman and Hall, London. 1-281.

MOURELLE, C. & EZCURRA, E. (1996) Species richness of Argentine cacti: a test of biogeographic hypotheses. *Journal of Vegetation Science :* 7, 667-680.

NOY-MEIR, I. & WHITTAKER, R.H. (1977) Continuous multivariate methods in community analysis: some problems and developments. *Vegetatio :* 33, 79-98.

PAGES, J.P., CAILLIEZ, F. & ESCOUFIER, Y. (1979) Analyse factorielle : un peu d'histoire et de géométrie. *Revue de Statistique Appliquée :* 27, 5-28.

PEARSON, D. L. & GHORPADE, K. (1989) Geographical distribution and ecological history of tiger beetles (Coleoptera: Cicindelidae) of the Indian subcontinent. *Journal of Biogeography :* 16, 333-344.

RAMESH, B. R. & PASCAL, J.-P. (1997) *Atlas of endemics of the Western Ghats (India). Distribution of tree species in the evergreen and semi-evergreen forests.* Institut Français de Pondichéry, Inde. 403 p.

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

RICHOUX, P., ALLEMAND, R. & COLLOMB, G. (1999) Ecogéographie de la région Rhône-Alpes : définition de districts naturels pour la cartographie de l'entomofaune. *Bulletin mensuel de la Société linéenne de Lyon* (sous presse).

RIPLEY, B. D. (1976) The second-order analysis of stationary point processes. *Journal of Applied Probability :* 13, 255-266.

RIPLEY, B. D. (1977) Modelling spatial patterns (with discussion). *Journal of the Royal Statistical Society, B :* 39, 172-212.

SHAFFER, H. B., FISHER, R. N. & DAVIDSON, C. (1998) The role of natural history collections in documenting species declines. *Trends in Ecology and Evolution :* 13, 27-30.

SOBERON, J. M. & LLORENTE, J. B. (1993) The use of species accumulation functions for the prediction of species richness. *Conservation Biology :* 7, 480-488.

TAKEUCHI, K., YANAI, H. & MUKHERJEE, B.N. (1982) *The foundations of multivariate analysis. A unified approach by means of projection onto linear subspaces.* John Wiley and Sons, New York. 1-458.

TERBRAAK, C. J. F. & VERDONSCHOT, P. F. M. (1995) Canonical correspondence analysis and related multivariate methods in aquatic ecology. *Aquatic Sciences :* 57, 255-289.

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

TERBRAAK, C. J. F. (1986) Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology :* 67, 1167-1179.

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

TERBRAAK, C. J. F. (1988) Partial canonical correspondence analysis. *In : Classification and related methods of data anlysis.* **BOCK, H. H. (ED.)** Elsevier Science Publishers B. V. 551-558.

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

TERBRAAK, C.J.F. (1994) Canonical community ordination. Part I: Basic theory and linear methods. *Ecoscience :* 1, 127-140.

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

THIOULOUSE, J., CHESSEL, D. & CHAMPELY, S. (1995) Multivariate analysis of spatial patterns: a unified approach to local and global structures. *Environmental and Ecological Statistics :* 2, 1-14.

WARTENBERG, D.E. (1985) Canonical trend surface analysis: a method for describing geographic pattern. *Systematic Zoology :* 34(3), 259-279.

WHITTAKER, R.H. (1967) Gradient analysis of vegetation. *Biological Reviews :* 42, 207-264.

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

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

WRIGHT, D. H., PATTERSON, B. D., MIKKELSON, G. M., CUTLER, A. & ATMAR, W. (1998) A comparative analysis of nested subset patterns of species composition. *Oecologia :* 113, 1-20.

YOCCOZ, N. (1988) *Le rôle du modèle euclidien d'analyse des données en biologie évolutive.* Thèse de doctorat, Université Lyon 1. 1-254.

ANNEXES

Annexe A : Documents annexés à l'étude d'une liste exhaustive :

Annexe A1 : Liste des 69 espèces et des codes espèces

Annexe A2 : Cartes de distribution des espèces

Annexe A3 : Variables utilisées dans le polynôme

Annexe A4 : Analyse multi-échelle de la répartition spatiale

Annexe B : Documents annexés à l'étude d'une liste échantillonnée :

Annexe B1 : Carte de distribution des 34806 occurrences

Annexe B2 : Variables utilisées dans le polynôme

Annexe B3 : Districts écogéographiques

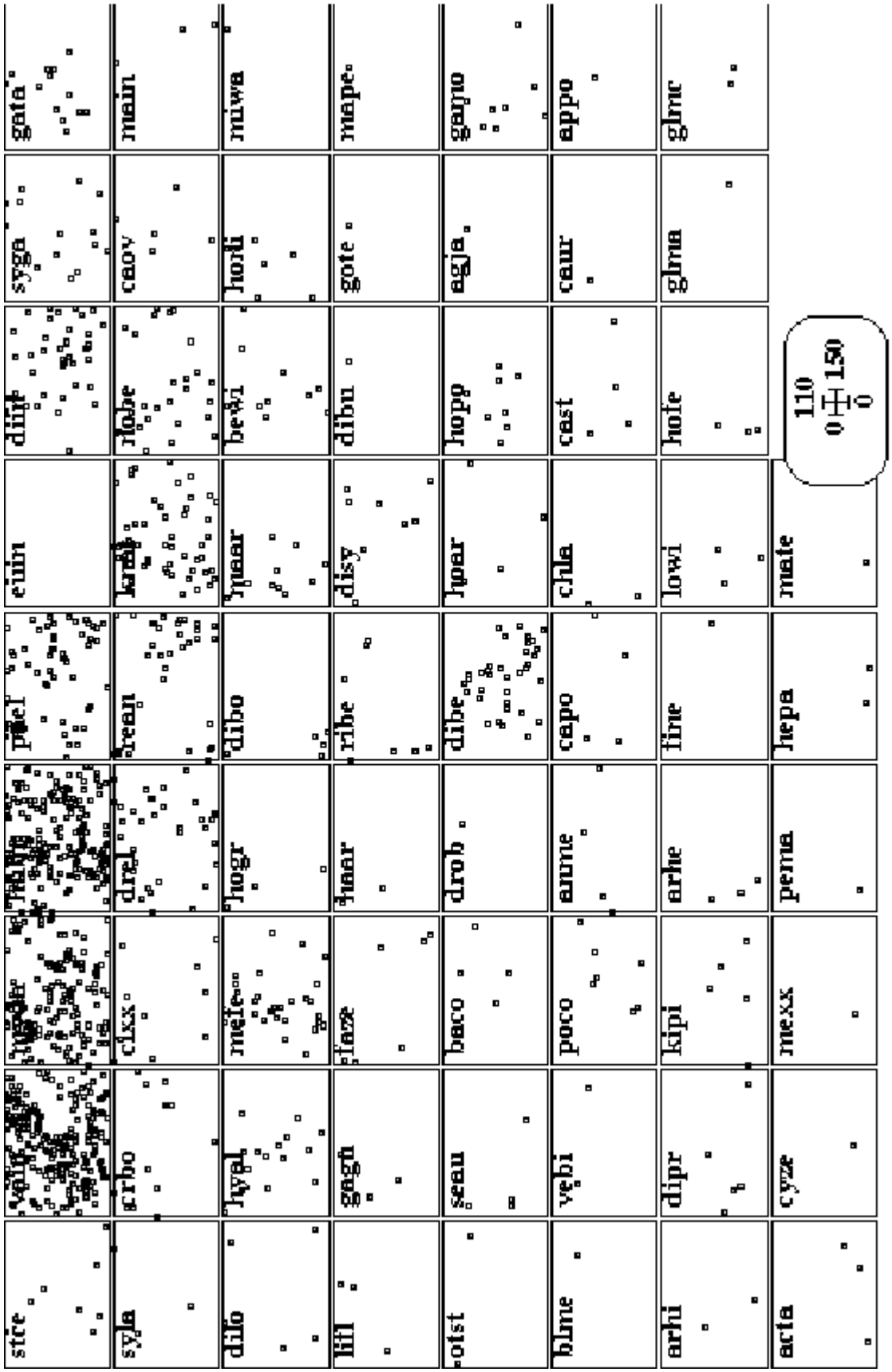
Annexe B4 : Cartes de distribution des espèces citées dans le texte

Annexe A1 : Liste des 69 espèces et des codes espèces

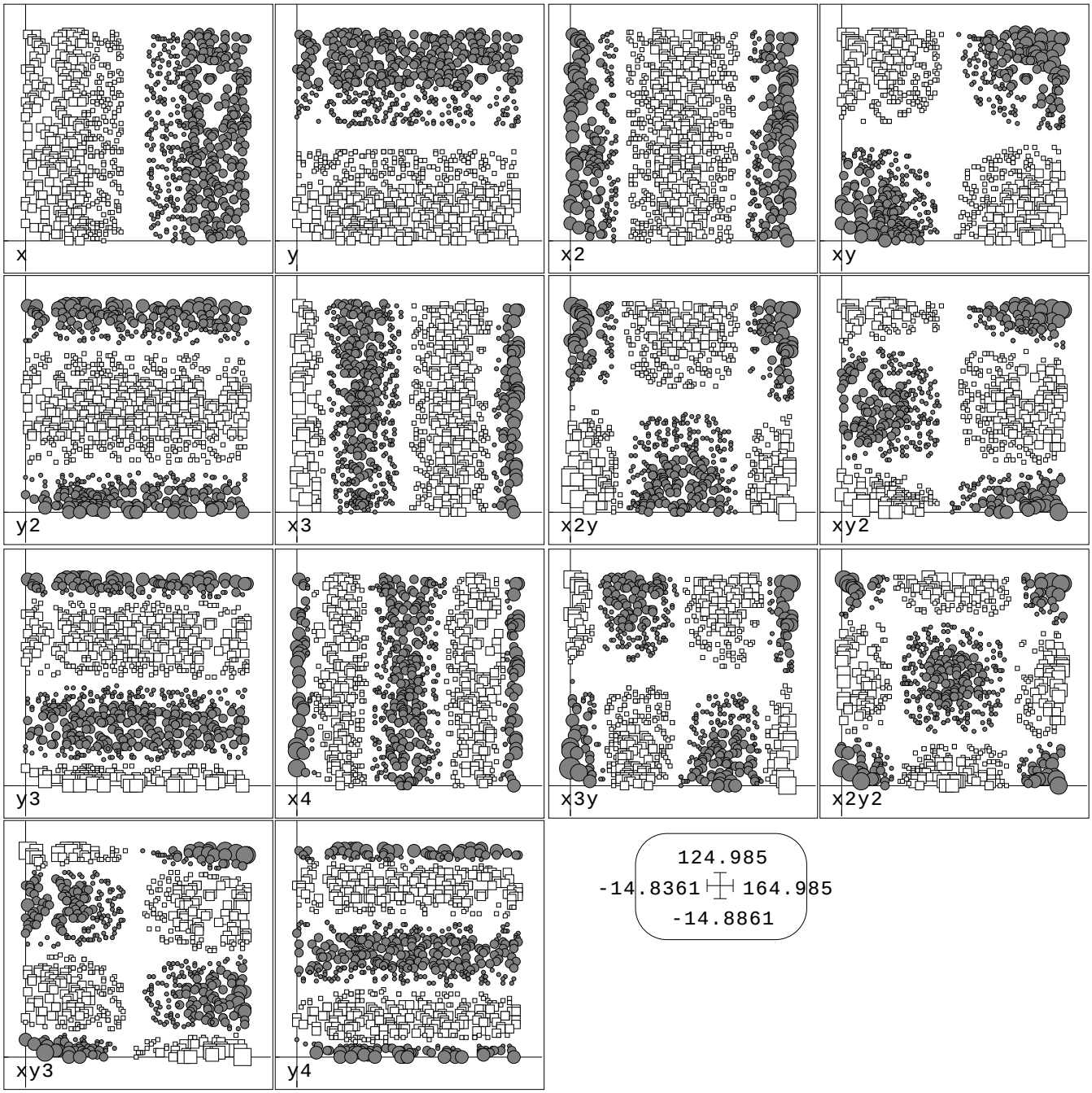
Code espèce	Nombre d' occurrences	Nom complet
acta	3	*Actinodaphne tadulingami*
agja	1	*Aglaia jainii*
anme	4	*Antidesma menasu*
appo	1	*Aphanamixis polystachya*
arhe	3	*Artocarpus heterophyllus*
arhi	2	*Artocarpus hirsutus*
baco	4	*Baccaurea courtallensis*
bewi	10	*Beilschmiedia wightii*
blme	1	*Blepharistemma membranifolia*
caov	5	*Casearia ovata*
capo	4	*Calophyllum polyanthum*
cast	4	*Canarium strictum*
caur	1	*Caryota urens*
chla	2	*Chrysophyllum lanceolatum*
cixx	10	*Cinnamomum sp*
crbo	13	*Cryptocarya bourdillonii*
cyze	1	*Cyathocalyx zeylanicus*
dibe	35	*Dimorphocalyx beddomei*
dibo	5	*Diospyros bourdillonii*
dibu	1	*Diospyros buxifolia*
diin	40	*Dipterocarpus indicus*
dilo	5	*Dimocarpus longan*
dipr	5	*Diospyros pruriens*
disy	8	*Diospyros sylvatica*
drel	34	*Drypetes elata*
drob	1	*Drypetes oblongifolia*
euin	1	*Euonymus indicus*
faze	7	*Fahrenheitia zeylanica*
fine	1	*Ficus nervosa*
gagu	3	*Garcinia gummi-gutta*
gamo	8	*Garcinia morella*
gata	13	*Garcinia talbotti*
glma	1	*Glochidion malabaricum*
glmc	2	*Glycosmis macrocarpa*
gote	1	*Gomphandra tetrandra*
haar	2	*Harpullia arborea*
hepa	2	*Heritiera papilio*
hoar	4	*Holigarna arnottiana*
hofe	3	*Holigarna ferruginea*
hogr	3	*Holigarna grahamii*
honi	7	*Holigarna nigra*
hopo	8	*Hopea ponga*
hubr	183	*Humboldtia brunonis*
hyal	15	*Hydnocarpus alpina*
kipi	5	*Kingiodendron pinnatum*
knat	53	*Knema attenuata*
lifl	3	*Litsea floribunda*
lowi	3	*Lophopetalum wightianum*

maar	10	*Mastixia arborea*
main	3	*Mangifera indica*
mape	1	*Macaranga peltata*
mate	1	*Mallotus tetracoccus*
mefe	29	*Mesua ferrea*
mexx	1	*Memecylon sp*
miwa	1	*Microtropis wallichiana*
myda	140	*Myristica dactyloides*
nobe	27	*Nothopegia beddomei*
otst	2	*Otonephelium stipulaceum*
pael	59	*Palaquium ellipticum*
pema	1	*Persea macrantha*
poco	7	*Polyathia coffeoides*
rean	33	*Reinwardtiodendron anaimalaiense*
ribe	7	*Rinorea bengalensis*
seau	5	*Semecarpus auriculata*
stce	8	*Strombosia ceylanica*
syga	14	*Syzygium gardneri*
syla	3	*Syzygium laetum*
vain	212	*Vateria indica*
vebi	2	*Vepris bilocularis*

Annexe A2 : Cartes de distribution des espèces



Annexe A3 : Variables utilisées dans le polynôme



Annexe A4 : Analyse multi-échelle de la répartition spatiale

L'analyse spatiale proposée par Ripley (1976, 1977) est basée sur la statistique K(d). L'analyse consiste à effectuer un dénombrement des j voisins situés à une distance inférieure ou égale à d de l'individu i. On calcule la fonction grâce à la formule suivante :

$$K(d) = \left(A \sum_{i=1}^{N} \sum_{j=1}^{N} k_{ij}\right) / N^2$$

avec :

d distance en m,

A surface de la parcelle en m^2,

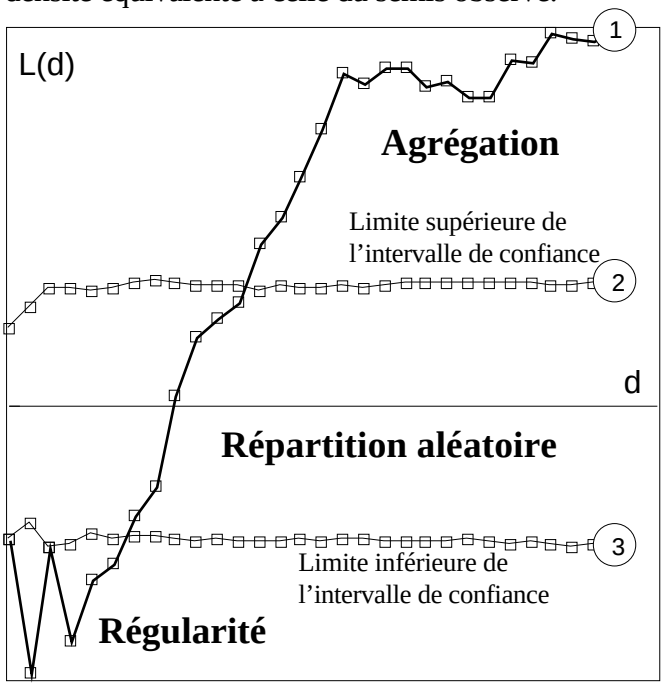
N le nombre total d'individus

k_{ij} prend la valeur 1 si la distance entre i et j est inférieure ou égale à d, 0 sinon.

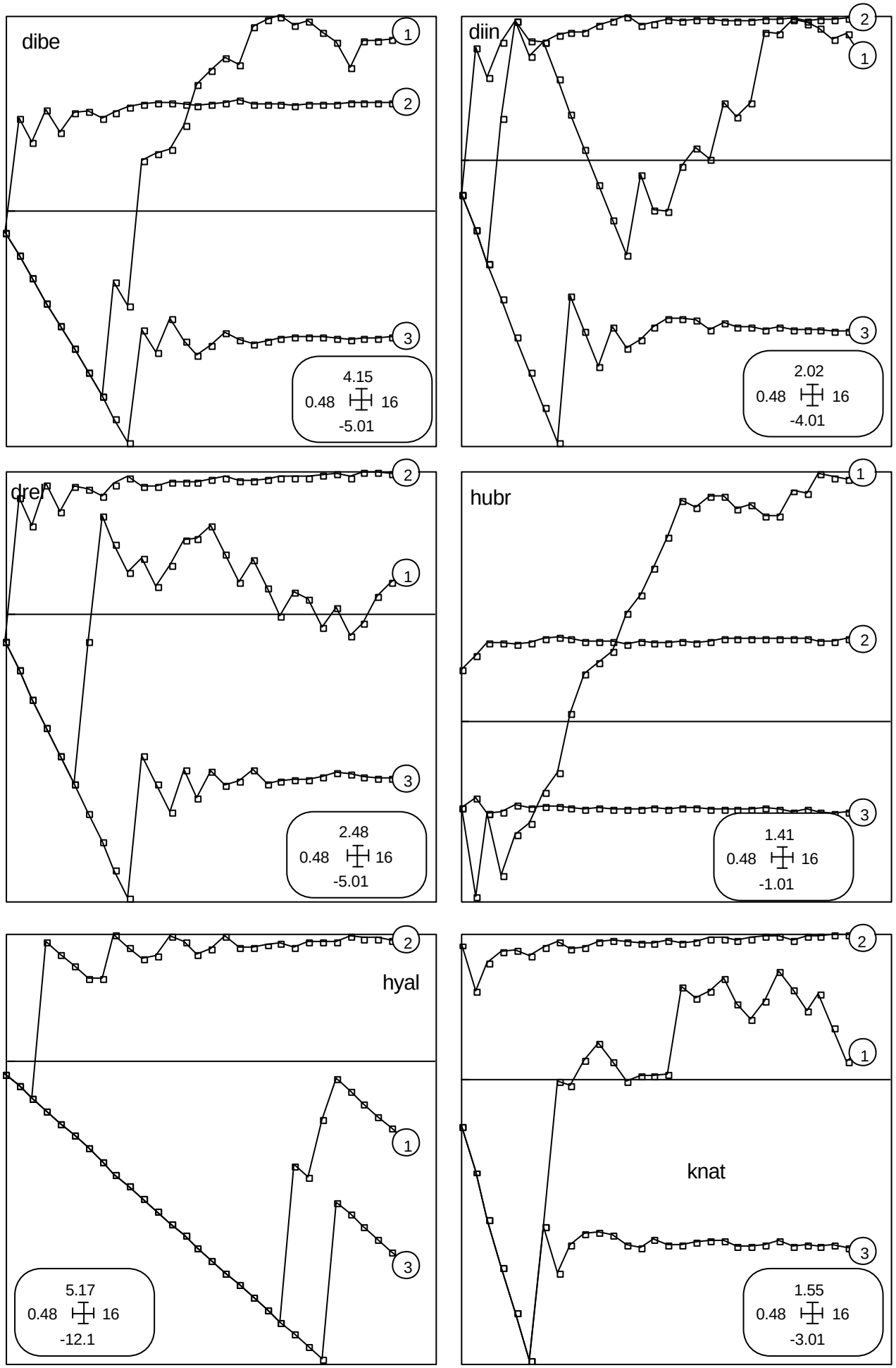
On utilisera la fonction modifiée de Besag (1977) :

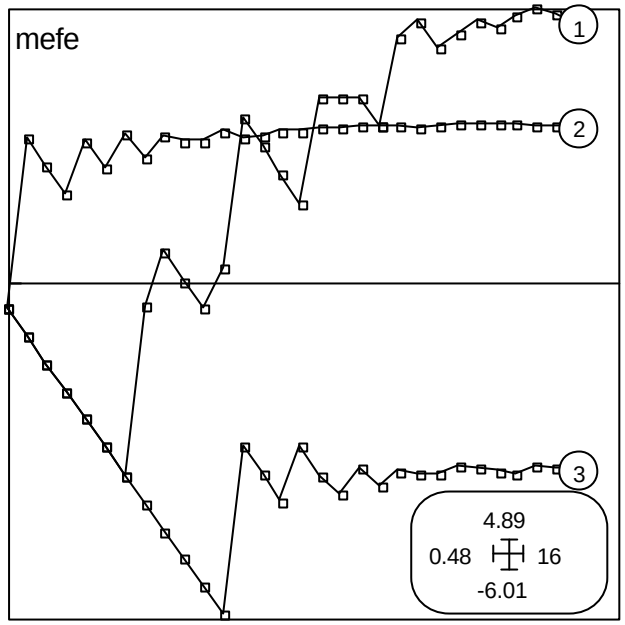
$$L(d) = \sqrt{K(d)/\pi} - d$$

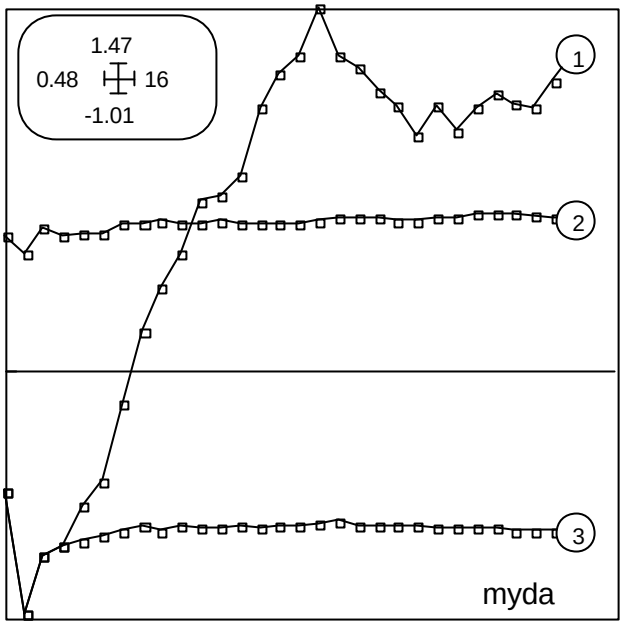
qui présente l'avantage d'être facilement interprétable. L(d)=0 pour un processus aléatoire (loi de Poisson), L(d)>0 pour un semis agrégé, et L(d)<0 pour un semis régulier. Un intervalle de confiance à 90% est calculé par la méhode de Monte Carlo pour l'hypothèse nulle d'un semis de Poisson (répartition aléatoire) de densité équivalente à celle du semis observé.

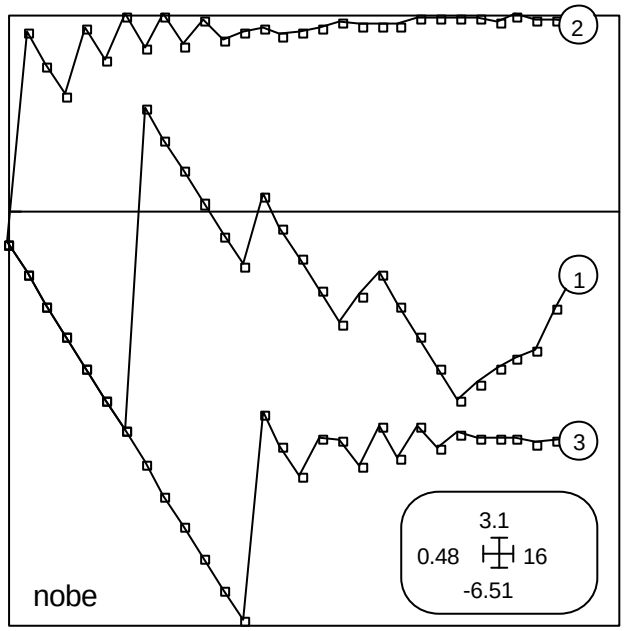


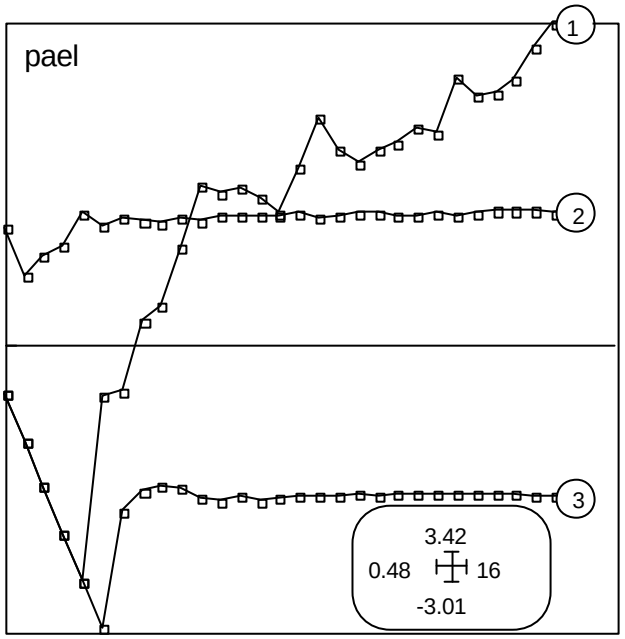
Pour les espèces les plus abondantes (au moins 15 occurrences), la statistique L(d) a été calculée pour d compris entre 0 et 15 m avec un pas de 0,5 m. L'intervalle de confiance est établi avec 5000 simulations. Les graphes représentent L (ordonnée) en fonction de d (abcisse).

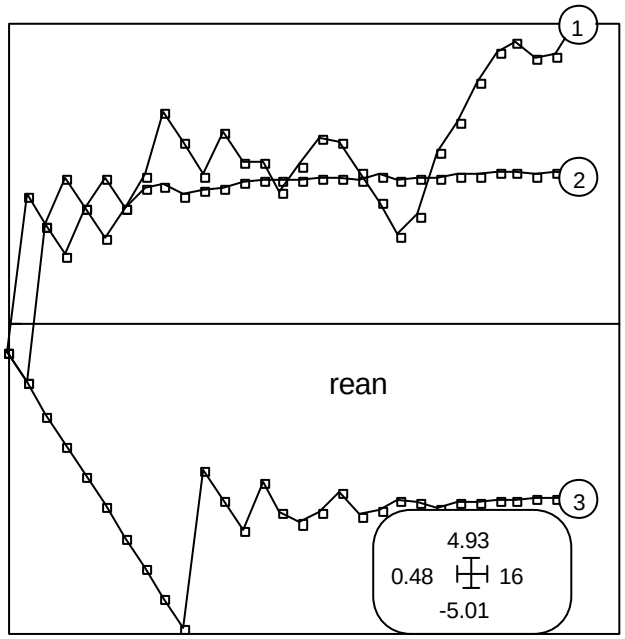

dibe
1
2
3
4.15
0.48
16
-5.01
diin
2
1
3
2.02
0.48
16
-4.01
drel
2
1
3
2.48
0.48
16
-5.01
hubr
1
2
3
1.41
0.48
16
-1.01
hyal
2
1
3
5.17
0.48
16
-12.1
knat
2
1
3
1.55
0.48
16
-3.01

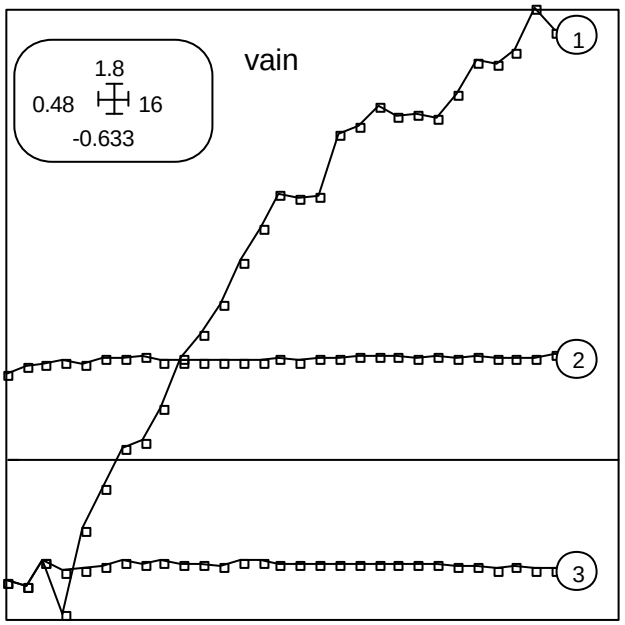

mefe
1
2
3
4.89
0.48
16
-6.01


1.47
0.48
16
-1.01
1
2
3
myda


2
1
3
3.1
0.48
16
-6.51
nobe

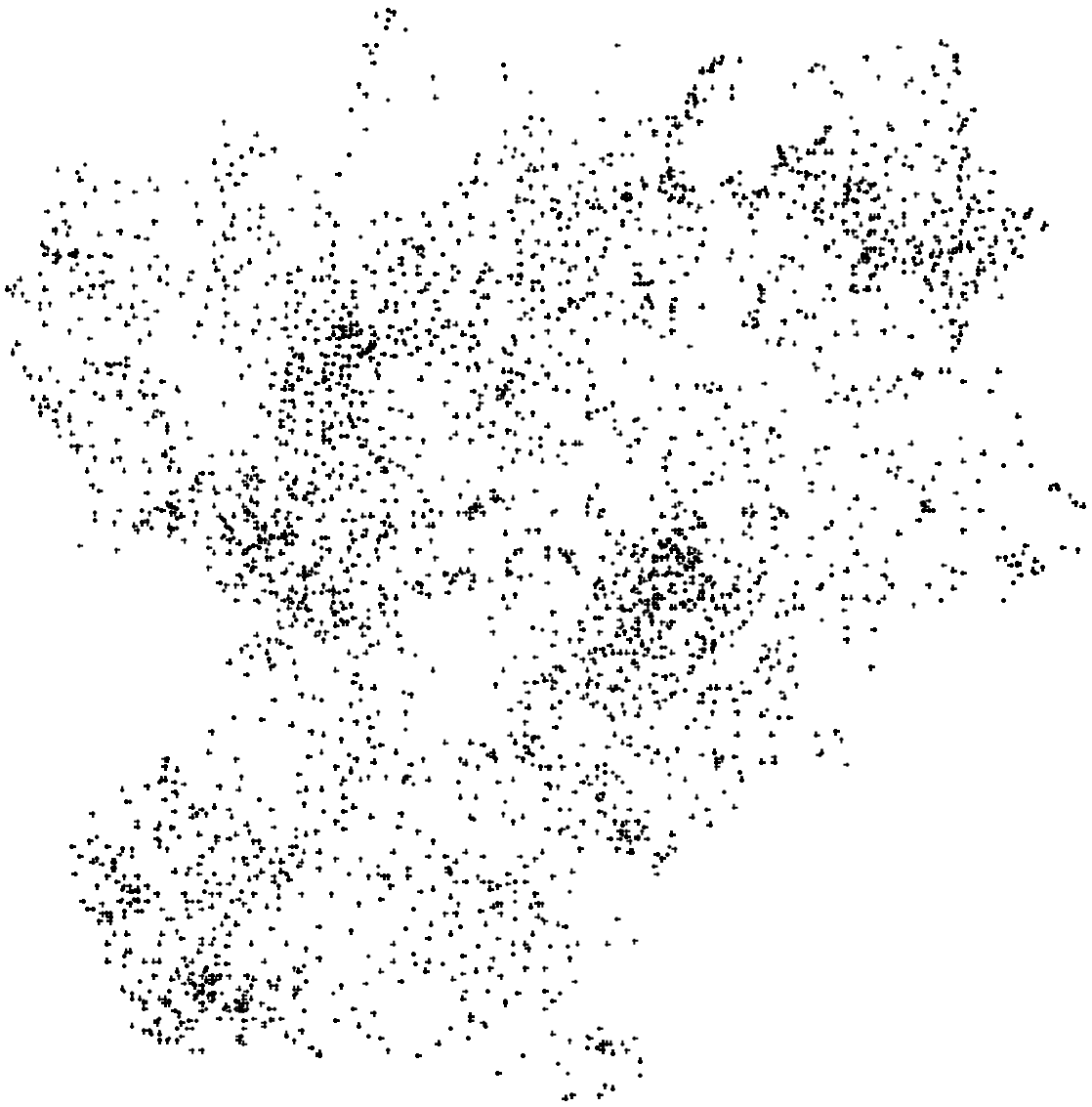

pael
1
2
3
3.42
0.48
16
-3.01


1
2
rean
3
4.93
0.48
16
-5.01


1.8
0.48
16
-0.633
vain
1
2
3

Annexe B1 : Carte de distribution des 34806 occurrences

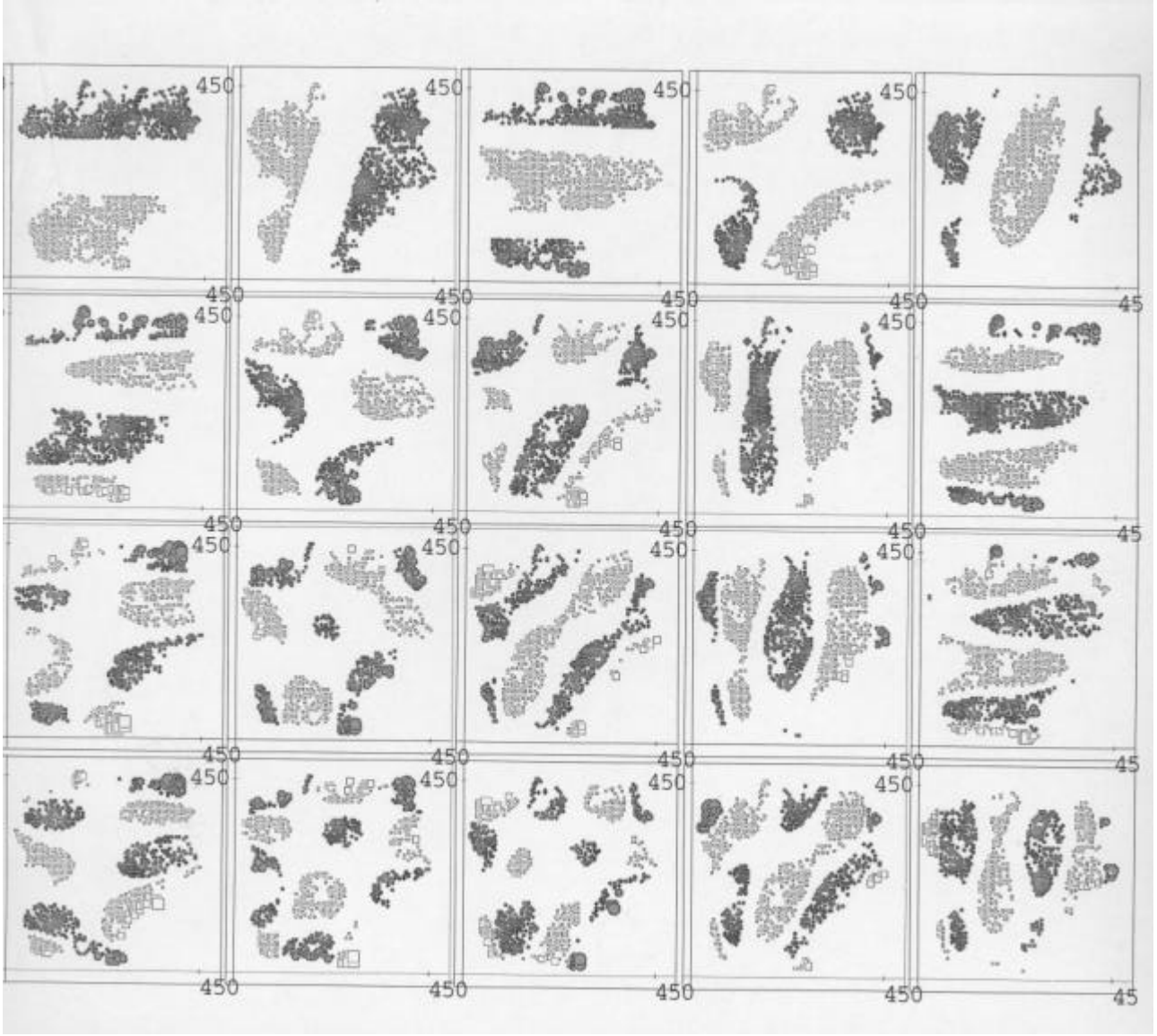
Sur la carte sont indiquées les différents lieux de collectes des 34806 spécimens (578 espèces) répertoriés dans la base.



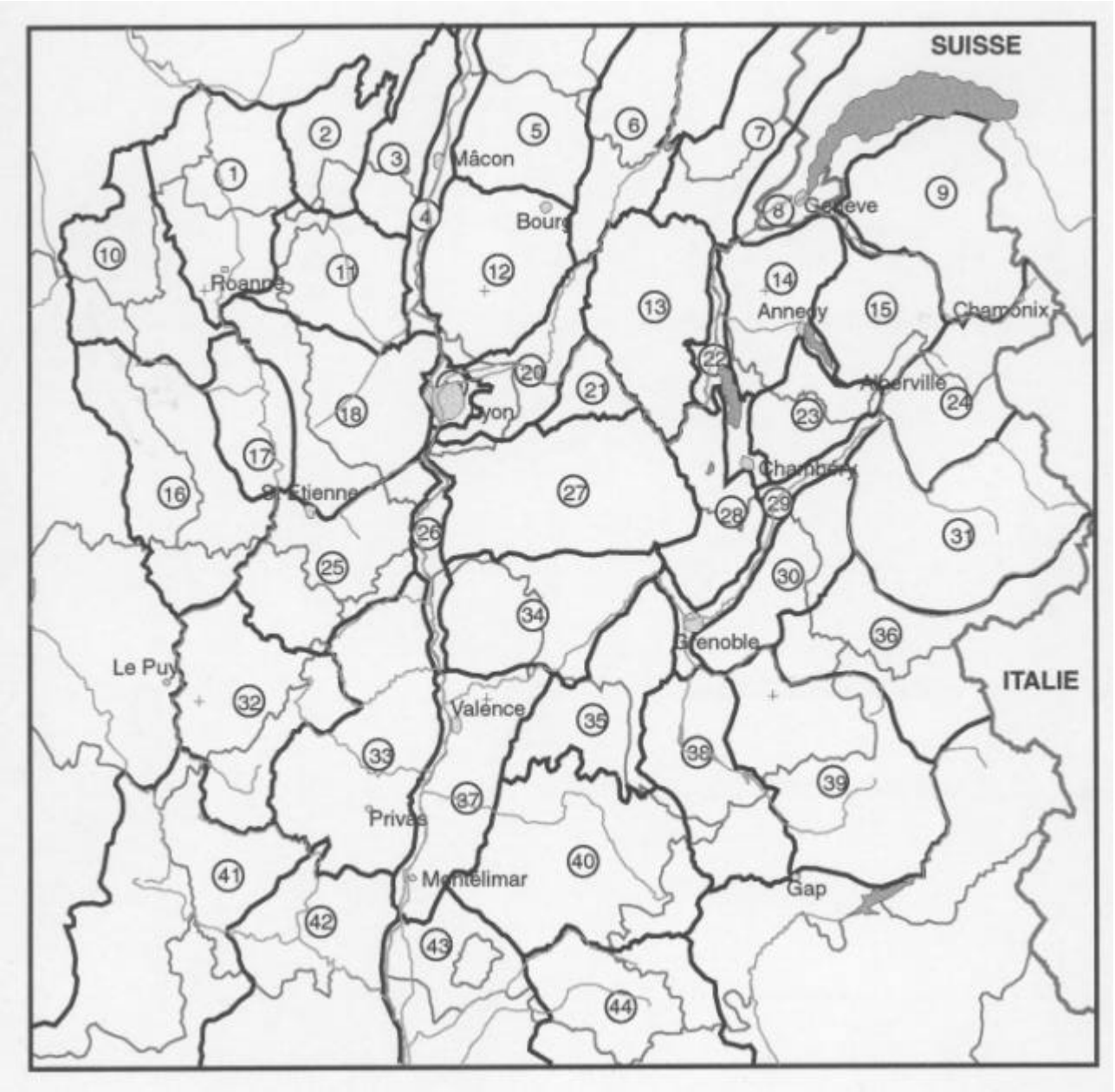
Annexe B2 : Variables utilisées dans le polynôme

Les variables sont centrées et réduites. La valeur de la variable est représentée par un rond (valeur positive) ou un carré (valeur négative) de taille proportionnelle à sa valeur absolue. Les variables sont repérées par les coordonnées des occurrences.

Les variables représentées sont respectivement (de gauche à droite puis de haut en bas de la page) :
x, y, x2, xy, y2, x3, x2y, xy2, y3, x4, x3y, x2y2, xy3, y4, x5, x4y, x3y2, x2y3, xy4, y5 .

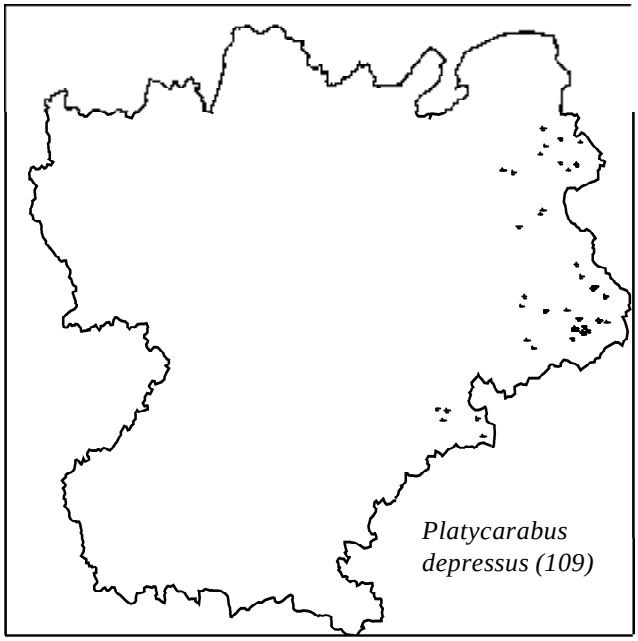


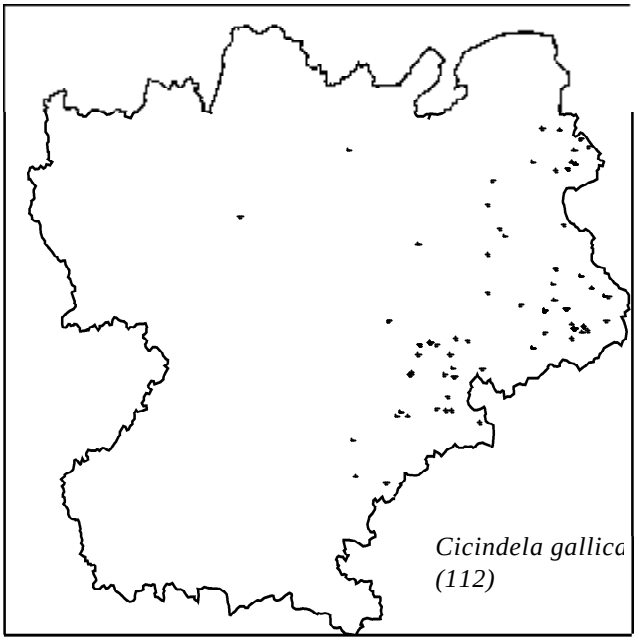
Annexe B3 : Districts écogéographiques

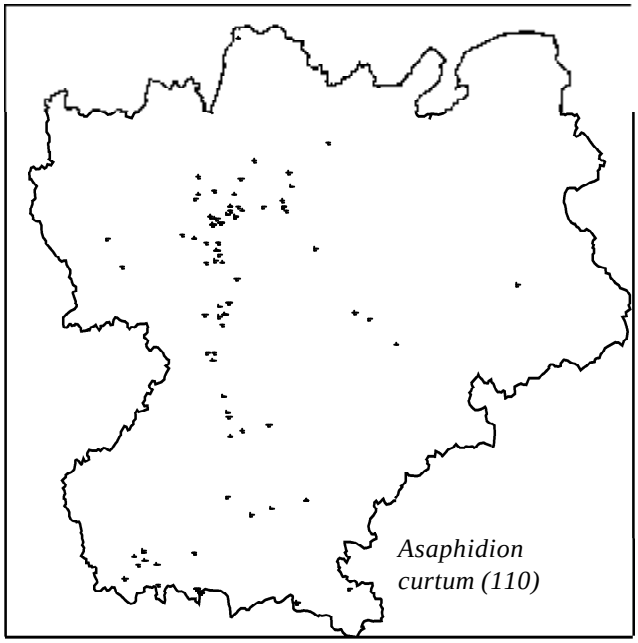


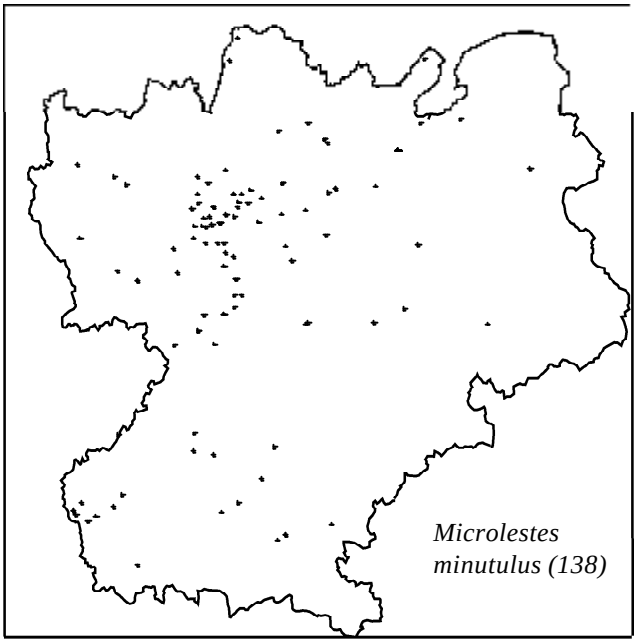
Annexe B4 : Cartes de distribution des espèces citées dans le texte

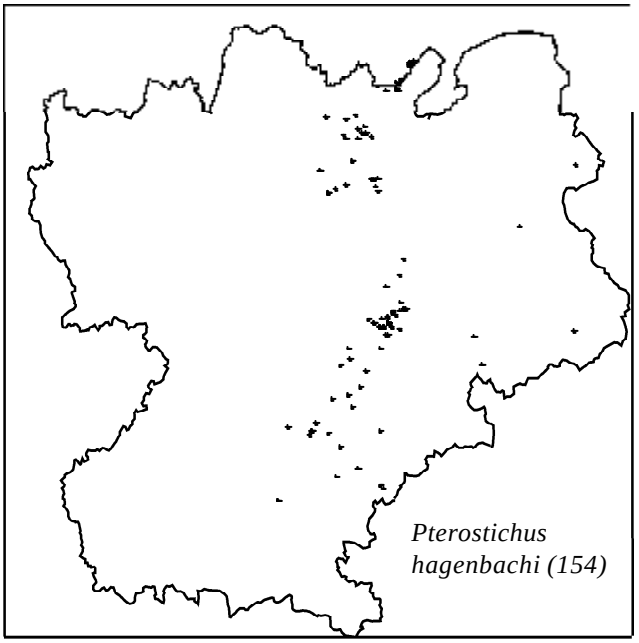
Le nombre d'occurrences de chaque espèce est indiqué entre parenthèses

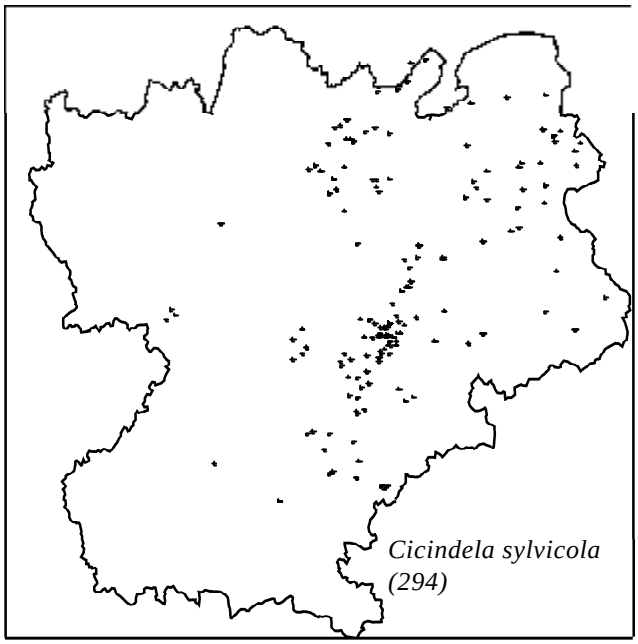


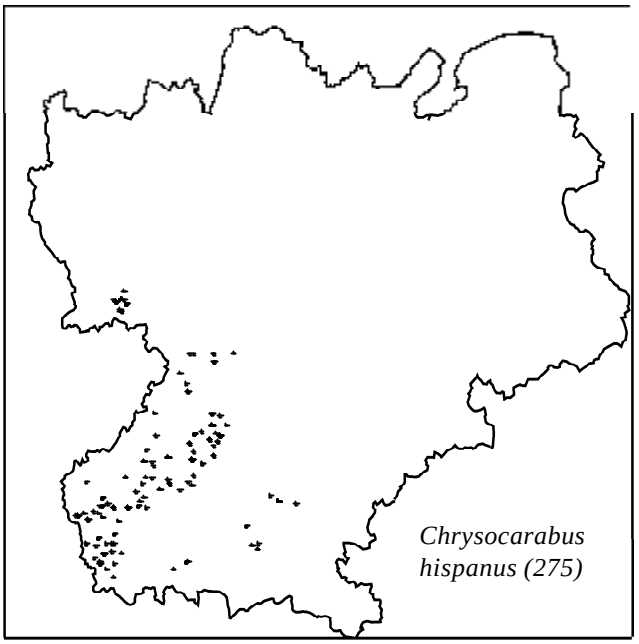


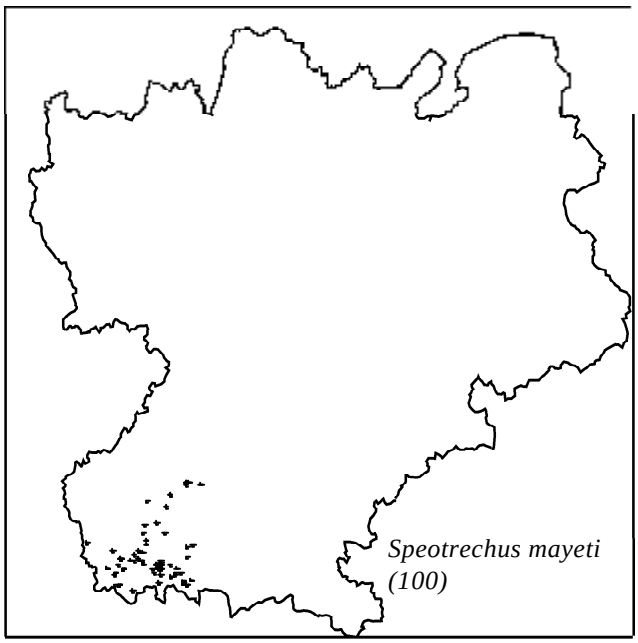


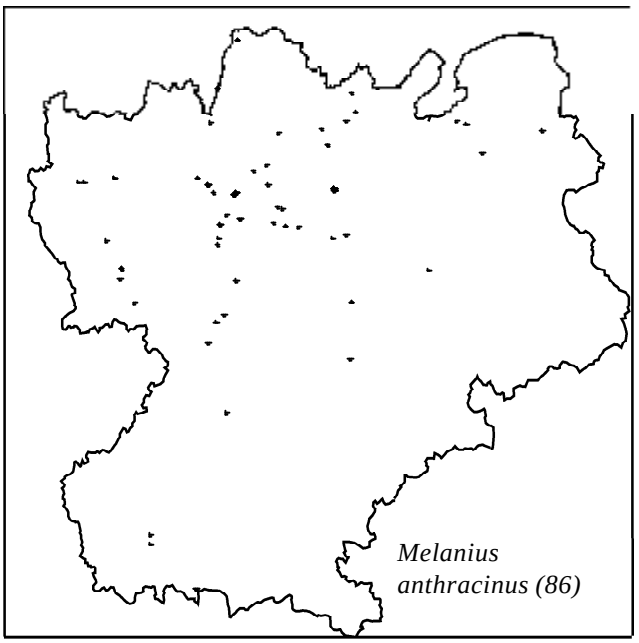


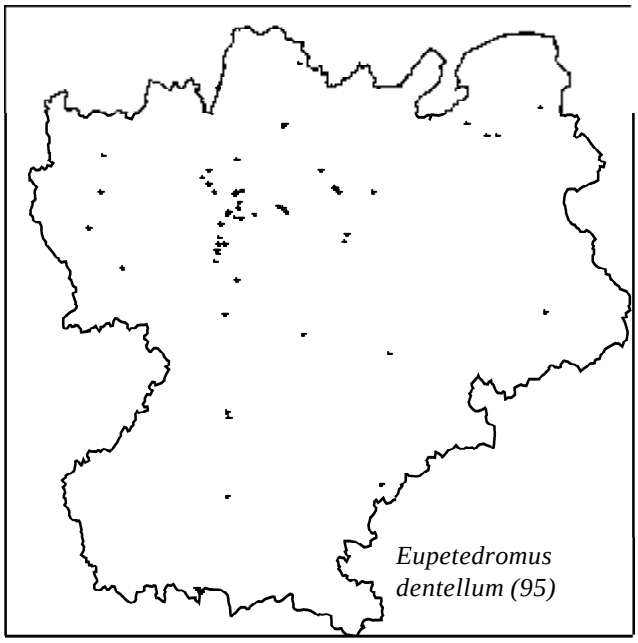



Cicindela sylvicola (294)


Chrysocarabus hispanus (275)


Speotrechus mayeti (100)


Melanius anthracinus (86)


Eupetedromus dentellum (95)