N° d'ordre 03-2003 Année 2003

THESE

présentée

devant l'UNIVERSITE CLAUDE BERNARD - LYON 1

pour l'obtention

du DIPLOME DE DOCTORAT
(arrêté du 30 mars 1992)

présentée et soutenue publiquement le 06 janvier 2003

par

Stéphane DRAY

Eléments d'interface entre analyses multivariées, systèmes d'information géographique et observations écologiques

Directeur de thèse : Daniel CHESSEL

JURY : M. Antoine GUISAN, Rapporteur
M. Francis LALOË, Rapporteur
M. Daniel CHESSEL
M. Jean-Dominique LEBRETON
M. Alain PAVE
Mme Dominique PONTIER

AVANT-PROPOS

Cette thèse a été préparée au sein de l'équipe d'Ecologie Statistique du laboratoire de Biométrie et Biologie Evolutive (UMR CNRS 5558).
Elle se caractérise par l'étude et la mise en relation de plusieurs domaines de compétences. Ceci s'est traduit notamment par la mise en œuvre de multiples collaborations avec des écologues. En biométrie, et plus particulièrement dans le cadre de cette thèse, ces diverses relations constituent l'essence du travail.

Je tiens à remercier l'ensemble des personnes ayant participé de près ou de loin à ce travail et particulièrement :

- *Raphaël Pélissier, Pierre Couteron, Clémentine Gimaret-Carpentier, Jean-Pierre Pascal, Nathalie Pettorelli, Jean-Michel Gaillard, Jean-François Michel qui m'ont permis de mettre des images de forêts, de chevreuils ou de mouches sur des tableaux de nombres. Leurs connaissances biologiques, leur intérêt méthodologique et leurs qualités humaines ont permis l'établissement de collaborations fructueuses. Rolland Alemand, Philippe Richoux et Joël Clary qui sont à l'origine de mon sujet de DEA et donc indirectement, de cette thèse. L'ensemble des organismes qui ont accepté de mettre leurs précieuses données à ma disposition.*

- *Les membres du Jury :*

  Françis Laloë et Antoine Guisan pour avoir accepter de juger ce travail en tant que rapporteur, pour leur intérêt, leurs remarques et leurs questions précises et très pertinentes.

  Jean-Dominique Lebreton, Alain Pavé et Dominique Pontier qui ont accepté de parcourir de (très) grandes distances pour participer à ce jury, pour la qualité de leurs remarques et les discussions que nous avons eues.

  Daniel Chessel pour avoir accepté de diriger cette thèse, pour sa confiance, sa passion, son intérêt, son écoute, son expérience et ses conseils, qu'il a su me transmettre depuis le DEA. Ces quatre années de collaboration ont été très très enrichissantes.

- *Jean Thioulouse et Anne-Béatrice Dufour de m'avoir accueilli dans leur équipe…et dans leur bureau. J'ai ainsi pu bénéficier d'un voisinage et d'un cadre de travail très agréable et d'une structure satisfaisant l'ensemble de mes demandes. Misou, Sophie (puis Anne-Lucie, puis Dominique) de m'avoir faciliter les différentes démarches administratives et Marie-Hélène pour, notamment, la préparation du pot.*

« Soit A un succès dans la vie. Alors A = x + y + z, où x = travailler, y = s'amuser, z = se taire.» (A . Einstein)

- *A tous les amis de Lyon, du Jura, de Montpellier ou d'ailleurs… pour le y.*

« C'est le devoir de chaque homme de rendre au monde au moins autant qu'il en a reçu. » (A . Einstein)

- *A ma famille, A Emeraude. A ma mère, mon père et mon frère qui m'ont beaucoup donné.*

Eléments d'interface entre analyses multivariées, systèmes d'information géographique et observations écologiques

Résumé

Dans une approche pluridisciplinaire basée sur 1) le contenu de méthodes statistiques multivariées et spatiales, 2) l'usage d'un système d'information géographique et 3) les objectifs expérimentaux définis par des écologues, cette thèse fournit des éléments nouveaux pour l'analyse de données écologiques multivariées et spatialisées. La première partie présente les différents éléments du dialogue. Il y figure 1) une synthèse, basée sur l'utilisation du schéma de dualité, des principales méthodes d'analyse de données, 2) une description des interfaces écrites entre trois logiciels largement utilisés (R, ADE-4 et ArcView), 3) une présentation des collaborations entreprises avec des écologues, en particulier N. Pettorelli, qui, au delà des prestations de la consultation statistique, permet d'accéder à des questions méthodologiques ouvertes.

La seconde partie présente les résultats obtenus lors de ces collaborations sous la forme de publications en anglais (6 acceptées et 4 soumises). L'utilisation de l'analyse canonique des corrélations sur des listes d'occurrences a permis d'étudier la structuration spatiale de peuplement forestier à une échelle locale ou régionale. Une modélisation de la prévalence d'une maladie à transmission vectorielle fondée sur des méthodes statistiques et des analyses spatiales est proposée. Les travaux sur la dynamique d'une population de chevreuils en relation avec l'habitat a conduit à définir deux nouvelles méthodes d'analyse multivariée. La première (analyse RLQ spatialisée) permet de coupler deux tableaux de données provenant de deux plans d'échantillonnage distincts d'un même espace, la seconde (analyse NIPALS spatialisée) permet de faire l'ACP de données provenant de *k* plans d'échantillonnage distincts d'un même espace. Enfin, un compromis entre analyse de co-inertie et rotation procustéenne a été introduit dans l'étude de la co-structure de deux tableaux.

Elements of interface between multivariate analyses, geographic information systems and ecological observations.

Abstract

Adopting a multidisciplinary approach based on 1) the content of spatial and multivariate analyses, 2) the use of a geographic information system, 3) experimental objectives defined by ecologists, this thesis gives new elements for the analysis of spatially referenced and multivariate ecological data. The first part presents the elements of the dialogue. It consists of 1) a synthesis, based on the duality diagram, of the main methods of multivariate analysis, 2) a description of written interfaces between three widely used software (R, ADE-4 and ArcView), 3) a presentation of the collaborations with ecologists, particularly N. Pettorelli, which, more than a simple statistical consulting, offers a number of methodological problems.

In the second part, the results obtained in these collaborations are presented with English publications (6 accepted and 4 submitted). The use of canonical correlation analysis of occurrences lists allows studying the spatial structure of forests at regional or local scale. Modeling the prevalence of a vector-transmitted disease, with statistical methods and spatial analyses, is proposed. The study of roe-deer population dynamic in relation with habitat leads to develop two new methods of data analysis. The first one (spatial RLQ analysis) links two datasets recorded from two different samplings of the same area, the second one (spatial NIPALS analysis) consists in PCA of data from *k* different samplings of the same area. Lastly, a compromise between co-inertia analysis and procrustean rotation is introduced for studying the co-structure of two datasets.

SOMMAIRE

LISTE DES ANNEXES

LISTE DES ABREVIATIONS

Abréviation	Nom complet
2B-PLS	Two-blocks Partial least squares
AC	Analyse canonique des corrélations
ACC	Analyse canonique des correspondances
ACI	Analyse de co-inertie
ACP	Analyse en composantes principales
ACPVI	Analyse en composantes principales sur variables instrumentales
AFC	Analyse factorielle des correspondances
AFCVI	Analyse factorielle des correspondances sur variables instrumentales
AFM	Analyse factorielle multiple
ANSC	Analyse non symétrique des correspondances
DCA	Detrended correspondence analysis
GPS	Global positionning system
ODBC	Open database connectivity
PLS	Partial least squares
SGBD	Système de gestion de base de données
SIG	Système d'information géographique
SQL	Structured query language

Introduction

Selon Ricklefs (1990), le zoologiste allemand Ernst Haeckel fut le premier, en 1870, a donné au mot « écologie » le sens que nous lui connaissons aujourd'hui. D'un point de vue étymologique, ce terme provient des mots grecques *oikos* (maison) et *logos* (science). L'écologie peut donc être définie comme l'étude des relations des organismes avec leur environnement ou bien, comme l'étude des interactions qui déterminent la distribution et l'abondance des organismes ou encore comme l'étude des écosystèmes (Barbault 1995). Legay (1980) affirme qu'« *Avoir un point de vue systématique, c'est admettre une série de relations entre éléments composant une situation ; et définir un système, c'est à la fois limiter le champ et en extraire un ensemble cohérent. Mais un système n'est jamais quelque chose de statique, si bien que, au-delà de sa description, le plus intéressant est évidemment l'étude de son fonctionnement et de sa dynamique. Parmi les systèmes biologiques, ceux qui ont été de beaucoup les plus étudiés au cours de ces dernières années sont les écosystèmes.* ». La description des écosystèmes par l'énumération de ses composants (*e.g.* biocénose) n'a que très peu d'intérêt. En effet, connaître l'ensemble des espèces vivantes d'un écosystème n'aidera pas forcément à comprendre son fonctionnement. L'écologie s'attache donc également à appréhender les interactions qu'entretiennent les espèces entre elles et avec le milieu. L'étude de l'évolution de ces relations dans l'espace et le temps est bien évidemment une question fondamentale (Blanc 2000). Cette tâche peut apparaître assez complexe car « *Aucun phénomène biologique n'a qu'une seule cause, aucun n'a qu'un seul effet. Il faut donc s'habituer à raisonner dans un univers multivarié, où règnent de nombreuses variables et leurs interactions d'ordre divers* » (Legay 1980). De nombreux processus interagissent dans le fonctionnement des écosystèmes et leur analyse pose naturellement de nouvelles problématiques biologiques, induisant parfois de nouvelles structures de données nécessitant des développements méthodologiques.

Les méthodes d'analyses de données permettent de rechercher les structures cachées dans les données et « *d'obtenir une description de nature statistique pour un certain phénomène qui a donné lieu au recueil de mesures ou observations trop nombreuses et dépendantes les unes des autres pour être interprétables en première lecture* » (Lebart *et al.* 1977). L'usage de ces méthodes pour l'analyse de tableaux écologiques est naturel et Gauch (1982 p. 1) en résume parfaitement les raisons : « *Community ecology concerns assemblages of plants and animals living together and the environmental and historical factors with which they interact. […] Community data are multivariate because each sample site is described by the abundances of a number of species, because numerous environmental factors affect communities, and so on. […] The application of multivariate analysis to community ecology is natural, routine and fruitful.* ». Ainsi, l'analyse d'un tableau floro-faunistique se fait classiquement avec une analyse factorielle des correspondances ou une analyse en composantes principales ; l'étude des relations espèces-milieu par une analyse canonique des correspondances (ter Braak 1986) ou une analyse de co-inertie (Dolédec & Chessel 1994). Les analyses inter-intra (Dolédec & Chessel 1987, 1989) permettent de prendre en compte une éventuelle partition des individus (date…) lors de l'analyse d'un tableau ou d'un couple de tableaux (Franquet & Chessel 1994, Franquet *et al.* 1995). Les méthodes multi-tableaux, dont le développement est plus récent, permettent d'analyser les variations spatio-temporelles d'une structure écologique (Chessel & Hanafi 1996) ou d'une co-structure (Simier *et al.* 1999). Ainsi, les développements méthodologiques ont aidé à répondre aux problématiques

biologiques et depuis plus de cinquante ans, les écologistes sont habitués à utiliser ces méthodes dont l'efficacité n'est plus à démontrer.

L'étude des séries temporelles et l'identification de structures spatiales sont des problématiques récurrentes en écologie (Cormack & Ord 1979, Smith 2002). En écologie végétationnelle notamment, un sujet de recherche central est l'étude des relations entre les phénomènes biologiques (croissance, compétition…) et les structures spatiales observées (Watt 1947). Ainsi, de nombreuses méthodes de géostatistique ont été utilisées en écologie avec succès (Dale 2000, Legendre & Fortin 1989). Les méthodes d'analyse des processus ponctuels (Besag 1977, Ripley 1976, 1977), les mesures d'autocorrélation spatiale (Cliff & Ord 1973), variogramme, covariogramme et corrélogramme permettent de décrire une structure spatiale ou de tester la présence d'une éventuelle corrélation spatiale. Des méthodes telles que la régression polynomiale (*trend surface analysis*, Gittins 1968) ou le krigeage (*e.g.* Bailey & Gatrell 1995) sont fréquemment utilisées pour estimer, interpoler et cartographier un processus spatialisé. Ainsi, Skalova *et al.* (1999) utilisent des mesures d'autocorrélation spatiale pour étudier l'influence du rayonnement lumineux sur la canopée. Cole & Syms (1999) emploient le même type de méthode pour identifier les causes de mortalité de micro-algues en Nouvelle-Zélande. Les méthodes d'analyse de processus ponctuels permettent de définir le mode de distribution spatiale de plantes ou d'arbres et ainsi de comprendre les processus de dynamique des communautés végétales (Couteron & Kokou 1997, Eccles *et al.* 1999). Le krigeage a permis d'étudier l'influence du recrutement des juvéniles sur la distribution spatiale d'une population de galathées (Roa & Tapia 2000) ou d'estimer l'abondance de poissons à partir de données provenant de surveillances acoustiques (Maravelias *et al.* 1996) ou de pêches systématiques (Rueda & Defeo 2001). Pour Legay (1980), « *dans le domaine spatial, l'une des difficultés majeures est celle de l'échelle* ». En effet, selon l'échelle spatiale choisie, les résultats concernant l'étude de la structure, des effets ou des causes d'un phénomène biologique peuvent varier fortement (Bohning-Gaese 1997, Dungan *et al.* 2002, Wiens 1989). Pour résoudre ce problème, les méthodes multi-échelles ont été développées et permettent de mettre en évidence une structure spatiale à différentes échelles d'observation (Goodall 1974, Greig-Smith 1952, Hill 1973a).

L'analyse de phénomènes biologiques requérant généralement la prise en compte de plusieurs variables, l'utilisation de ces méthodes présentait certaines limites car la plupart ont été développées uniquement dans le cas univarié. L'étude de structures spatiales dans un contexte multivarié a donc engendré le développement de nouvelles méthodes basées sur des principes de géostatistique et d'analyse de données. Le couplage de polynômes des coordonnées spatiales des individus et des variables quantitatives par une analyse canonique des corrélations (*canonical trend surface analysis*, Gittins 1985, Wartenberg 1985a) ou par une analyse canonique des correspondances (Borcard & Legendre 1994, Borcard *et al.* 1992, Méot *et al.* 1998, ter Braak 1987) permet d'identifier une structure spatiale multivariée. Les mesures d'autocorrélation spatiale ont également des extensions multivariées (Thioulouse *et al.* 1995, Wartenberg 1985b). L'analyse multi-échelles de données multivariées a entraîné le développement de nombreuses méthodes (Dale & Zbigniewicz 1995, Di Bella & Jona-Lasinio 1996, Noy-Meir & Anderson 1971, Schaefer & Messier 1994, Ver Hoef & Glenn-Lewin 1989, Ver Hoef *et al.* 1989) dont les applications en écologie demeurent peu nombreuses.

L'essor de l'informatique a constitué une véritable révolution technologique dont la biométrie et donc l'écologie ont grandement profité et « *il est certain qu'en biologie d'innombrables questions qu'on ne se posait même pas, ou qui n'étaient pas techniquement abordables, relèvent maintenant de démarches courantes, aussi bien dans la pratique et au niveau de la production qu'en théorie et au niveau de la recherche* » (Legay 1980). En effet, les développements méthodologiques ont été favorisés par les outils (matériel et logiciels)

disponibles et la distribution de ces nouvelles méthodes sous la forme de logiciels a permis d'améliorer et de faciliter le traitement de données écologiques. La mise à disposition de nombreux logiciels dédiés à l'analyse de données écologiques multivariées (ADE-4, CANOCO, DECORANA, TWINSPAN, PC-ORD, Progiciel R…) a permis de populariser et de favoriser l'usage des méthodes d'analyse multivariée en écologie. Dans le cadre du stockage, de la gestion et de l'analyse des données spatialisées, l'apport des systèmes d'information géographique (SIG) est indéniable. Les SIG facilitent notamment l'étude de phénomènes à grande échelle spatiale ou temporelle (Johnson 1990). L'écologie a adopté cet outil assez récemment et de nombreuses applications ont déjà été réalisées (Batek *et al.* 1999, Carmel & Kadmon 1999, Hirzel 2001, Kadmon & Danin 1999, Kadmon & Heller 1998, Yom-Tov & Kadmon 1998). L'utilisation conjointe d'un SIG avec des méthodes traditionnelles telles les géostatistiques (Anselin & Getis 1992, Burrough 2001, Ding & Fotheringham 1992, Goodchild *et al.* 1992) ou les méthodes d'analyses multivariées (Guisan *et al.* 1999, Guisan & Zimmermann 2000, Kadmon & Danin 1997, Ohmann & Spies 1998) offre un cadre très performant pour l'analyse de données spatialisées en raison de la complémentarité de ces différentes approches. Les analyses multivariées permettent d'identifier des structures, les géostatistiques permettent d'appréhender le caractère spatial de ces structures alors que le SIG est un outil de stockage, manipulation, création, et représentation de données spatialisées.

Ce travail de thèse s'inscrit dans la lignée des travaux biométriques réalisés dans le laboratoire (*e.g.* Blanc 2000, Chessel 1992, Gimaret-Carpentier 1999, Méot 1992, Thioulouse 1996, Yoccoz 1988). En privilégiant l'étude de processus spatiaux, ce travail se focalise notamment sur les apports des méthodes d'analyse de données et des systèmes d'information géographique pour l'analyse de données écologiques multivariées et spatialisées. Nous désirons insister plus particulièrement sur les bénéfices d'une approche pluridisciplinaire en nous situant à l'interface des outils, des méthodes et des données. L'intérêt d'une telle démarche est qu'elle génère la résolution de problématiques biologiques mais également le développement de nouvelles méthodes et la mise au point de nouveaux outils. Nous proposons ainsi quelques éléments pour améliorer le traitement de ce type de données en s'appuyant conjointement sur le cadre théorique de l'analyse de données, les caractéristiques des données collectées et l'utilisation des systèmes d'information géographique.

La première partie de ce mémoire présente l'environnement tripartite de ce travail. Elle vise à définir les différents éléments mis en relation lors de cette thèse. Un premier chapitre est consacré à la théorie de l'analyse de données. Cette présentation est basée sur l'utilisation du schéma de dualité et couvre un grand nombre de méthodes d'analyse d'un tableau et de couplage de deux tableaux. Elle permet de définir l'environnement théorique. Le deuxième chapitre fournit une description des outils utilisés. Un logiciel de statistique généraliste (R), un logiciel d'analyse de données (ADE-4) et un système d'information géographique (ArcView) y sont présentés. Ces trois logiciels ont été largement utilisés au cours de cette thèse et en constituent le cadre instrumental. Enfin, le troisième chapitre s'attache à définir l'environnement biologique de ce travail. Plusieurs collaborations scientifiques ont été entreprises avec des écologues lors de cet exercice de thèse. Nous nous attachons notamment à décrire les différentes problématiques biologiques, les objectifs de ces études ainsi que les caractéristiques des données.

La deuxième partie est consacrée aux résultats obtenus lors de ce travail. Ces résultats sont présentés sous forme de publications en anglais soumises ou acceptées dans des revues internationales. Le quatrième chapitre, basé sur trois articles acceptés dans *Plant Ecology*, *Ecology* et *Ecography*, concerne l'analyse des listes d'occurrences. Le cinquième chapitre présente la modélisation d'une maladie à transmission vectorielle. Ce travail a été présenté au

congrès GISVET (10-14/09/2001, Lancaster, Grande-Bretagne) et donnera lieu à une publication dans un numéro spécial de *Preventive Veterinary Medicine*. Le sixième chapitre aborde la problématique de la mise en relation de paramètres de dynamique de population avec les caractéristiques de l'habitat disponible. Il s'appuie sur deux publications soumises à *Oecologia* et *Wildlife Society Bulletin*. Le septième chapitre présente une nouvelle méthode d'analyse de données cartographiques incomplètes et reprend une publication soumise à *Transactions in GIS*. Le huitième chapitre est axé sur la problématique de couplage de deux jeux de données. Un premier article soumis à *Ecology* présente les avantages d'une approche basée sur le critère de la co-inertie. Un deuxième article accepté dans *Ecoscience* présente une nouvelle méthode de couplage utilisant les principes de l'analyse procustéenne et de l'analyse de co-inertie. Enfin, un troisième article accepté dans *Journal of Vegetation Science* présente une méthode originale permettant le couplage de données provenant de deux plans d'échantillonnage. Le neuvième chapitre présente deux outils développés au cours de cette thèse. CoCoAn est une librairie du logiciel R permettant notamment de réaliser une analyse canonique des correspondances et AVADE est une extension facilitant une utilisation d'ADE-4 à partir d'un SIG (ArcView).

Partie I : Les composantes du dialogue

Chapitre I : Théorie de l'analyse de données

« *L'analyse de données* » recouvre un ensemble de techniques ayant pour objectif la description statistique des grands tableaux. De par la quantité d'information analysée, ces méthodes nécessitent un usage quasi-impératif de l'informatique. Elles se caractérisent par leur objectif exploratoire et par l'abandon d'hypothèses probabilistes (contrairement à la statistique inférentielle) au profit de la géométrie euclidienne. En général, un tableau contient les mesures de p variables (colonnes) sur n individus (lignes). L'analyse exploratoire d'un tel tableau en fournit un résumé en insistant sur les représentations graphiques des individus qui sont considérés au même titre que les variables. Les domaines d'utilisation de l'analyse de données sont nombreux et diversifiés : biométrie, psychométrie, économétrie, … De ce fait, de nombreuses méthodes ont été découvertes, puis redécouvertes dans d'autres disciplines et possèdent ainsi différentes appellations. Par exemple, l'analyse d'une table de contingence à deux dimensions exprimant l'association entre deux variables qualitatives a été réalisée à de nombreuses occasions. Hirschfeld (1935) donne une formulation algébrique de la corrélation entre les lignes et les colonnes d'une table de contingence. Fisher (1940) analyse la relation entre la couleur des cheveux et la couleur des yeux pour un groupe d'écoliers écossais. En psychométrie, la « *scale analysis* » de Guttman (1941) et la « *method for quantification of qualitative data* » développée par Hayashi (1950) permettent d'assigner des valeurs numériques aux catégories afin de les discriminer de façon optimale (« *optimal scaling* » ou « *dual scaling* », (Nishisato 1980)). En biométrie, les travaux de Williams (1952) peuvent être cités et seront introduits en écologie par la suite (Hill 1973b, 1974) sous le nom de « *reciprocal averaging* ». Enfin, Benzécri (1969) développe l'analyse factorielle des correspondances (traduit en « *correpondence analysis* ») et introduit la géométrie euclidienne dans le cadre de cette méthode. Toutes les approches précitées sont équivalentes. Dès lors, on peut s'étonner du titre de l'article de Hill (1974), « *Correspondence analysis : A neglected multivariate method* », alors que les premières approches dataient de près de 40 ans. Ceci illustre bien les problèmes d'échanges interdisciplinaires induisant qu'une méthode, qui est certainement aujourd'hui la plus utilisée pour l'analyse d'un tableau espèces-relevés, a mis près de 40 ans pour être (re)connue en écologie. L'existence d'un cadre théorique général permettant de définir clairement les différentes méthodes et les liens qu'elles entretiennent apparaît alors nécessaire. La théorie du schéma de dualité initiée par Cailliez et Pagès (1976) et introduite en écologie dans Escoufier (1987) est une réponse adaptée à ce besoin. L'utilisation qu'en font Tenenhaus et Young (1985) dans le cadre de l'analyse factorielle multiple est certainement la démonstration la plus convaincante de la puissance d'un tel outil en analyse de données.

I.1. Schéma de dualité

Un schéma est constitué de trois éléments définissant un triplet $(\mathbf{X},\mathbf{Q},\mathbf{D})$. $\mathbf{X}$ est une matrice de données qui peut être issue d'une transformation préalable des données brutes. $\mathbf{X}$ a n lignes (individus) et p colonnes (variables). $\mathbf{Q}$ est un produit scalaire de $\mathbb{R}^p$ (matrice carrée symétrique) et $\mathbf{D}$ est un produit scalaire de $\mathbb{R}^n$. Les n lignes de $\mathbf{X}$ sont des vecteurs de $\mathbb{R}^p$ et les p colonnes de $\mathbf{X}$ sont des vecteurs de $\mathbb{R}^n$. On a alors deux représentations euclidiennes de $\mathbf{X}$ (figure 1) :

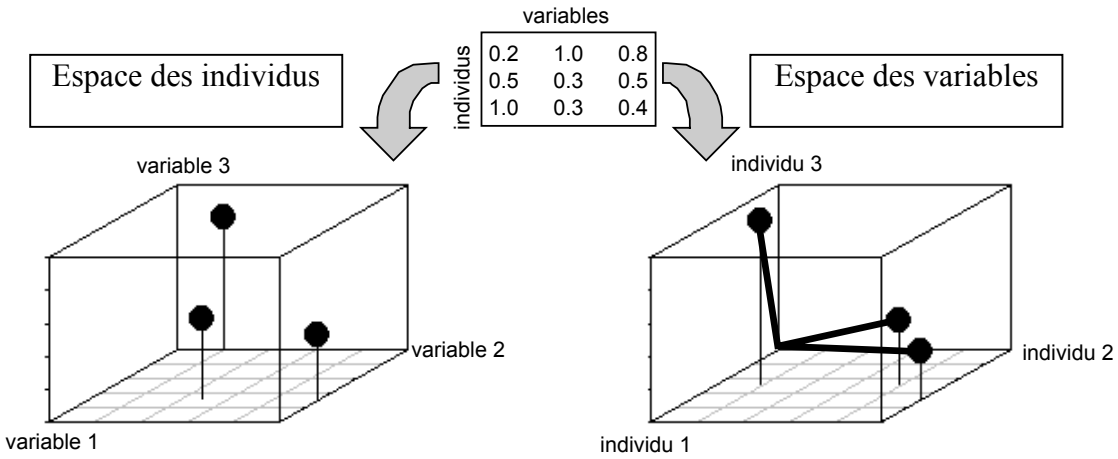


Figure 1 : Représentation géométrique d'un tableau comme un nuage de points dans l'espace des individus ou dans l'espace des variables.

On définit un schéma de dualité (figure 2) :

$$(\mathbf{X},\mathbf{Q},\mathbf{D}) \Leftrightarrow \begin{array}{ccc} \mathbb{R}^p & \xrightarrow{\mathbf{Q}} & \mathbb{R}^{p*} \\ \mathbf{X}^t \uparrow & & \downarrow \mathbf{X} \\ \mathbb{R}^{n*} & \xleftarrow[\mathbf{D}]{} & \mathbb{R}^n \end{array}$$

Figure 2 : Schéma de dualité associé au triplet $(\mathbf{X},\mathbf{Q},\mathbf{D})$.

$\mathbb{R}^{p*}$ est appelé dual de $\mathbb{R}^p$ (i.e. ensemble des applications linéaires de $\mathbb{R}^p$ dans $\mathbb{R}$) et $\mathbb{R}^{n*}$ est le dual de $\mathbb{R}^n$.

I.1.1. Diagonalisation d'un schéma de dualité

L'analyse de ce schéma consiste à déterminer les valeurs propres et vecteurs propres associés aux différentes possibilités de produits matriciels $\mathbf{X}^t\mathbf{DXQ}$, $\mathbf{DXQX}^t$, $\mathbf{XQX}^t\mathbf{D}$, et $\mathbf{QX}^t\mathbf{DX}$. Pour différentes valeurs des paramètres **X**, **Q** et **D**, le cadre théorique général fourni par le schéma de dualité permet d'appréhender de nombreuses méthodes telles que l'analyse en composantes principales (ACP) centrée et/ou normée, l'analyse factorielle des correspondances (AFC), l'analyse non symétrique des correspondances (ANSC), l'analyse factorielle multiple (AFM), l'analyse discriminante (AD), l'analyse de co-inertie (ACI), l'analyse canonique (AC)…

Au schéma $(\mathbf{X},\mathbf{Q},\mathbf{D})$ sont associées r valeurs propres non nulles et distinctes, avec $r \leq \min(n,p)$. La matrice $\mathbf{X}^t\mathbf{DXQ}$ (p x p) est **Q**-symétrique (i.e. $\mathbf{QX}^t\mathbf{DXQ} = \left(\mathbf{QX}^t\mathbf{DXQ}\right)^t$) et définit donc une base de vecteurs propres **Q**-orthonormée. Il existe alors une matrice **A** vérifiant :

$$\mathbf{X}^t\mathbf{DXQA} = \mathbf{A\Lambda} \text{ et } \mathbf{A}^t\mathbf{QA} = \mathbf{I}_p$$

$\mathbf{\Lambda}$ est une matrice diagonale (p x p) contenant les p valeurs propres rangées par ordre décroissant :

$$\mathbf{\Lambda} = diag(\lambda_1, \lambda_2, ..., \lambda_r, ..., \lambda_p)$$
$$\lambda_1 > \lambda_2 > ... > \lambda_r > \lambda_{r+1} = ... = \lambda_p = 0$$

Pour les r valeurs propres non nulles et distinctes, on a :

$$\mathbf{X}^t\mathbf{DXQA}_r = \mathbf{A}_r\mathbf{\Lambda}_r \text{ et } \mathbf{A}_r{}^t\mathbf{QA}_r = \mathbf{I}_r$$

La matrice $\mathbf{\Lambda}_r$ contient les r valeurs propres non nulles et la matrice $\mathbf{A}_r$ (p x r) contient les r vecteurs propres **Q**-normés associés.

De la même manière, la matrice $\mathbf{XQX}^t\mathbf{D}$, qui est **D**-symétrique, définit une base de vecteurs propres **D**-orthonormée. Il existe donc une matrice **B** telle que :

$$\mathbf{XQX}^t\mathbf{DB} = \mathbf{BM} \text{ et } \mathbf{B}^t\mathbf{DB} = \mathbf{I}_n$$

$\mathbf{M}$ est une matrice diagonale (n x n) contenant les n valeurs propres rangées par ordre décroissant :

$$\mathbf{M} = diag(\mu_1, \mu_2, ..., \mu_r, ..., \mu_n)$$
$$\mu_1 > \mu_2 > ... > \mu_r > \mu_{r+1} = ... = \mu_n = 0$$

Pour les r valeurs propres non nulles et distinctes, on a :

$$\mathbf{XQX}^t\mathbf{DB}_r = \mathbf{B}_r\mathbf{M}_r \text{ et } \mathbf{B}_r{}^t\mathbf{DB}_r = \mathbf{I}_r$$

La matrice $\mathbf{M}_r$ contient les r valeurs propres non nulles et la matrice $\mathbf{B}_r$ (n x r) contient les r vecteurs propres **D**-normés associés.

Les deux systèmes de valeurs propres sont identiques et on a donc :

$$\mu_1 = \lambda_1, \mu_2 = \lambda_2, ..., \mu_r = \lambda_r$$

Les matrices $\mathbf{QX}^t\mathbf{DX}$ et $\mathbf{DXQX}^t$ ont les mêmes valeurs propres non nulles que les précédentes. La matrice $\mathbf{QX}^t\mathbf{DX}$ est $\mathbf{Q}^{-1}$-symétrique et définit une base de vecteurs propres $\mathbf{Q}^{-1}$ orthonormée. Il existe donc une matrice $\mathbf{A}^*$ vérifiant :

$$\mathbf{QX}^t\mathbf{DXA}^* = \mathbf{A}^*\mathbf{\Lambda} \text{ et } \mathbf{A}^{*t}\mathbf{Q}^{-1}\mathbf{A}^* = \mathbf{I}_p$$

Pour les r valeurs propres non nulles et distinctes, on a :

$$\mathbf{QX}^t\mathbf{DXA}_r{}^* = \mathbf{A}_r{}^*\mathbf{\Lambda}_r \text{ et } \mathbf{A}_r{}^{*t}\mathbf{Q}^{-1}\mathbf{A}_r{}^* = \mathbf{I}_r$$

La matrice $\mathbf{DXQX}^t$ est $\mathbf{D}^{-1}$-symétrique et définit une base de vecteurs propres $\mathbf{D}^{-1}$ orthonormée. Il existe donc une matrice $\mathbf{B}^*$ vérifiant :

$$\mathbf{DXQX}^t\mathbf{B}^* = \mathbf{B}^*\mathbf{M} \text{ et } \mathbf{B}^{*t}\mathbf{D}^{-1}\mathbf{B}^* = \mathbf{I}_n$$

Pour les r valeurs propres non nulles et distinctes, on a :

$$\mathbf{DXQX}^t\mathbf{B}_r^* = \mathbf{B}_r^*\mathbf{\Lambda}_r \text{ et } \mathbf{B}_r^{*t}\mathbf{D}^{-1}\mathbf{B}_r^* = \mathbf{I}_r$$

Les vecteurs propres de $\mathbf{X}^t\mathbf{DXQ}$ contenus dans la matrice $\mathbf{A}_r$ sont les r axes principaux ($\mathbb{R}^p$). Les vecteurs propres de $\mathbf{XQX}^t\mathbf{D}$ contenus dans la matrice $\mathbf{B}_r$ sont les r composantes principales ($\mathbb{R}^n$). Les vecteurs propres de $\mathbf{QX}^t\mathbf{DX}$ contenus dans la matrice $\mathbf{A}_r^*$ sont les r facteurs principaux ($\mathbb{R}^{p*}$). Les vecteurs propres de $\mathbf{DXQX}^t$ contenus dans la matrice $\mathbf{B}_r^*$ sont les r cofacteurs principaux ($\mathbb{R}^{n*}$). Cette information peut être résumée dans le schéma suivant (figure 3) :

$$\begin{array}{ccc}
\mathbf{A}_r = [\mathbf{a}_1 | \cdots | \mathbf{a}_r] & & \mathbf{A}^*_r = [\mathbf{a}^*_1 | \cdots | \mathbf{a}^*_r] \\
\text{axes principaux} & & \text{facteurs principaux} \\
\mathbb{R}^p & \xrightarrow{\quad\mathbf{Q}\quad} & \mathbb{R}^{p*} \\
\mathbf{X}^t \uparrow & \Lambda_r = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_r \end{bmatrix} & \downarrow \mathbf{X} \\
\mathbb{R}^{n*} & \xleftarrow[\quad\mathbf{D}\quad]{} & \mathbb{R}^n \\
\mathbf{B}^*_r = [\mathbf{b}^*_1 | \cdots | \mathbf{b}^*_r] & & \mathbf{B}_r = [\mathbf{b}_1 | \cdots | \mathbf{b}_r] \\
\text{cofacteurs principaux} & & \text{composantes principales}
\end{array}$$

Figure 3 : Principaux éléments associés à la diagonalisation du schéma de dualité $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$.

Comme indiqué sur le schéma de dualité précédent, ces quatre systèmes de vecteurs propres entretiennent des relations très étroites. Dans la pratique, on ne réalisera qu'une seule diagonalisation (dans la dimension la plus petite) et l'on déduira les trois autres systèmes de vecteurs propres :

$$\mathbf{A}_r^* = \mathbf{QA}_r \qquad \mathbf{B}_r = \mathbf{XA}_r^*\mathbf{\Lambda}_r^{-\frac{1}{2}} \qquad \mathbf{B}_r^* = \mathbf{DB}_r \qquad \mathbf{A}_r = \mathbf{X}^t\mathbf{B}_r^*\mathbf{\Lambda}_r^{-\frac{1}{2}}$$

Par exemple, si $p < n$ alors on diagonalisera la matrice $\mathbf{X}^t\mathbf{DXQ}$ afin d'obtenir les vecteurs propres contenus dans $\mathbf{A}_r$. A l'aide des relations précédentes, on obtient :

$$\mathbf{A}_r^* = \mathbf{QA}_r \qquad \mathbf{B}_r = \mathbf{XQA}_r\mathbf{\Lambda}_r^{-\frac{1}{2}} \qquad \mathbf{B}_r^* = \mathbf{DXQA}_r\mathbf{\Lambda}_r^{-\frac{1}{2}}$$

Les représentations graphiques des individus et des variables sont alors obtenues par simple projection de $\mathbf{X}$ sur les composantes principales et axes principaux. Les coordonnées des lignes ($\mathbf{L}_r$) et les coordonnées des colonnes ($\mathbf{K}_r$) sont calculées comme suit :

$$\mathbf{K}_r = \mathbf{X}^t\mathbf{DB}_r = \mathbf{A}_r\mathbf{\Lambda}_r^{\frac{1}{2}}$$
$$\mathbf{L}_r = \mathbf{XQA}_r$$

I.1.2. Critères d'optimalité

Plusieurs critères sont optimisés par l'analyse d'un triplet $(\mathbf{X},\mathbf{Q},\mathbf{D})$:

- Les vecteurs $\mathbf{a}_1,...,\mathbf{a}_j,...,\mathbf{a}_r$ maximisent successivement sous contrainte d'orthonormalité au sens de **Q** la forme quadratique $\|\mathbf{XQa}_i\|_{\mathbf{D}}^2$. Le maximum atteint est λ_i.
- Les vecteurs $\mathbf{b}_1,...,\mathbf{b}_j,...,\mathbf{b}_r$ maximisent successivement sous contrainte d'orthonormalité au sens de **D** la forme quadratique $\|\mathbf{X}^t\mathbf{Db}_i\|_{\mathbf{Q}}^2$. Le maximum atteint est λ_i.
- Les couples de vecteurs $(\mathbf{a}_1,\mathbf{b}_1),...,(\mathbf{a}_j,\mathbf{b}_j),...,(\mathbf{a}_r,\mathbf{b}_r)$ maximisent successivement sous contrainte d'orthonormalité les produits scalaires $\langle \mathbf{XQa}_i | \mathbf{b}_i \rangle_{\mathbf{D}} = \langle \mathbf{X}^t\mathbf{Db}_i | \mathbf{a}_i \rangle_{\mathbf{Q}}$. Le maximum atteint est $\sqrt{\lambda_i}$.

I.1.3. Inertie d'un triplet

Il est possible de calculer des statistiques d'inertie lorsque l'on dispose d'une métrique (matrice positive symétrique) et d'une pondération (matrice diagonale positive). L'inertie d'un nuage de points est simplement la somme, pour l'ensemble des points, du produit du poids d'un point multiplié par la distance au carré entre ce point et l'origine. Considérons, par exemple, le nuage des lignes du tableau **X** contenu dans $\mathbb{R}^p$. Si **D** est une matrice diagonale positive, chaque ligne $\mathbf{X}_i$ est pondérée par $\mathbf{D}_{ii}$. Si **Q** est une métrique, on peut alors définir l'inertie du triplet $(\mathbf{X},\mathbf{Q},\mathbf{D})$:

$$I(\mathbf{X},\mathbf{Q},\mathbf{D}) = \sum_{i=1}^{n} \mathbf{D}_{ii} \|\mathbf{X}_i\|_{\mathbf{Q}}^2 = trace(\mathbf{XQX}^t\mathbf{D}) = \sum_{k=1}^{r} \lambda_k$$

Les lignes de **X** peuvent être projetées sur un vecteur **u** **Q**-normé et l'inertie projetée est exprimée par :

$$I(\mathbf{u}) = \mathbf{u}^t\mathbf{QX}^t\mathbf{DXQu}$$

L'inertie totale du triplet peut donc être aisément décomposée par projection sur un système de *p* vecteurs **Q**-orthonormés :

$$I(\mathbf{X},\mathbf{Q},\mathbf{D}) = \sum_{k=1}^{p} I(\mathbf{u}_k) = \sum_{k=1}^{p} \mathbf{u}_k{}^t\mathbf{QX}^t\mathbf{DXQu}_k = \sum_{k=1}^{p} \|\mathbf{XQu}_k\|_{\mathbf{D}}^2 = \sum_{k=1}^{r} \lambda_k$$

D'après les critères d'optimalité cités précédemment, il apparaît donc que la diagonalisation d'un schéma de dualité consiste à trouver un système de vecteurs orthonormés de $\mathbb{R}^p$ maximisant l'inertie projetée. Ces vecteurs sont contenus dans la matrice $\mathbf{A}_r$.

La décomposition de l'inertie projetée sur les axes principaux s'écrit donc :

$$I(\mathbf{X},\mathbf{Q},\mathbf{D}) = \sum_{k=1}^{r} I(\mathbf{a}_k) = \sum_{k=1}^{r} \mathbf{a}_k{}^t\mathbf{Q}\mathbf{X}^t\mathbf{D}\mathbf{X}\mathbf{Q}\mathbf{u}_k = \sum_{k=1}^{r} \left\|\mathbf{X}\mathbf{Q}\mathbf{a}_k\right\|_{\mathbf{D}}^2 = \sum_{k=1}^{r} \lambda_k$$

Deux indices basés sur l'inertie sont classiquement utilisés afin d'améliorer l'interprétation des résultats.

La contribution absolue exprime la part prise par un point dans la définition d'un axe. La contribution absolue d'un point $\mathbf{X}_i$ sur l'axe $\mathbf{a}_k$ est :

$$CA_k(\mathbf{X}_i) = \frac{\left\|\mathbf{X}_i\mathbf{Q}\mathbf{a}_k\right\|_{\mathbf{D}}^2}{\lambda_k}$$

La contribution relative ou cosinus carré quantifie la part prise par un axe dans la représentation d'un point. Autrement dit, elle mesure la qualité de représentation d'un point par sa projection sur un axe. La contribution relative de l'axe $\mathbf{a}_k$ sur $\mathbf{X}_i$ est :

$$CR_k(\mathbf{X}_i) = \frac{\left\|\mathbf{X}_i\mathbf{Q}\mathbf{a}_k\right\|_{\mathbf{D}}^2}{\mathbf{D}_{ii}\left\|\mathbf{X}_i\right\|_{\mathbf{Q}}^2}$$

La contribution relative cumulée des K premiers axes dans la représentation du point $\mathbf{X}_i$ s'écrit donc simplement :

$$CRC_k(\mathbf{X}_i) = \frac{\sum_{k=1}^{K}\left\|\mathbf{X}_i\mathbf{Q}\mathbf{a}_k\right\|_{\mathbf{D}}^2}{\mathbf{D}_{ii}\left\|\mathbf{X}_i\right\|_{\mathbf{Q}}^2}$$

Ces deux indices sont très utiles lorsque les pondérations des lignes ou des colonnes ne sont pas uniformes (*e.g.* analyse factorielle des correspondances) car dans cette situation la contribution d'un point à un axe ne se lit pas directement sur les représentations graphiques. En effet, un point proche de l'origine (distance faible) associé à un poids fort peut participer autant à la définition d'un axe qu'un point éloigné (distance grande) mais faiblement pondéré.

Dans cette partie consacrée à l'inertie d'un triplet, nous avons privilégié un point de vue en considérant uniquement la représentation d'un nuage de lignes. Il est évident que toutes ces observations peuvent être appliquées au nuage des colonnes si le triplet fournit une pondération des colonnes (**Q** diagonale positive) et une métrique de $\mathbb{R}^n$ (**D** matrice positive symétrique). En fonction du triplet $(\mathbf{X},\mathbf{Q},\mathbf{D})$, il pourra donc y avoir une double analyse d'inertie (**D** et **Q** sont des matrices diagonales positives), une analyse d'inertie simple sur le nuage des lignes (**D** est une matrice diagonale positive et **Q** est une matrice non diagonale mais positive et symétrique), une analyse d'inertie simple sur le nuage des colonnes (**Q** est une matrice diagonale positive et **D** est une matrice non diagonale mais positive et symétrique) ou bien aucune analyse d'inertie.

I.1.4. Reconstitution de données

La théorie du schéma de dualité est intimement liée à la décomposition d'une matrice en valeurs singulières. Il est ainsi possible d'obtenir une reconstitution du tableau **X** au rang m (Good 1969) :

$$\hat{\mathbf{X}} = \sum_{i=1}^{m} \sqrt{\lambda_i}\, \mathbf{b}_i \mathbf{a}_i^{\,t}$$

Si on applique cette formule dans le cas où $m = r$ alors la reconstitution du tableau **X** est complète.

I.1.5. Diagonalisation d'un schéma dissymétrique

Afin d'obtenir une base orthonormée, on s'attache à diagonaliser une matrice symétrique. Dans le cas où **Q** est symétrique mais non diagonale (*e.g.* métrique de Mahalanobis), les différents produits matriciels ne sont pas symétriques. Une solution consiste alors à « éclater » le schéma de dualité afin d'obtenir une matrice symétrique. On réalise la décomposition spectrale de **Q** ($\mathbf{Q} = \mathbf{U\Theta U}^t$) et on analyse le triplet $(\mathbf{XH}^t, \mathbf{I}_p, \mathbf{D})$ avec $\mathbf{H} = \mathbf{\Theta}^{\frac{1}{2}}\mathbf{U}^t$ (figure 4) :

$$\begin{array}{ccc}
\mathbb{R}^p & \xrightarrow{\mathbf{I}_p} & \mathbb{R}^{p*} \\
\mathbf{H} \uparrow & & \downarrow \mathbf{H}^t = \mathbf{U\Theta}^{\frac{1}{2}} \\
\mathbb{R}^p & \xrightarrow{\mathbf{Q}} & \mathbb{R}^{p*} \\
\mathbf{X}^t \uparrow & & \downarrow \mathbf{X} \\
\mathbb{R}^{n*} & \xleftarrow[\mathbf{D}]{} & \mathbb{R}^n
\end{array}$$

Figure 4 : Schéma de dualité associé au triplet $(\mathbf{XH}^t, \mathbf{I}_p, \mathbf{D})$.

On retrouvera les axes principaux de $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$ simplement par $\mathbf{A} = \mathbf{H}^{-1}\mathbf{Z}$ où **Z** est la matrice contenant les vecteurs propres de $\mathbf{HX}^t\mathbf{DXH}^t$.

Dans le cas où les deux produits scalaires **D** et **Q** ne sont pas diagonaux, on réalise deux décompositions spectrales permettant ainsi « d'éclater » le schéma des deux côtés.

I.1.6. Exemples de schéma de dualité

Considérons un tableau **Y** contenant les abondances de *p* espèces (colonnes) dans *n* relevés (lignes). Il existe de nombreuses possibilités pour analyser un tel tableau (*e.g.* Noy-Meir 1973, Noy-Meir *et al.* 1975) correspondant à différents schémas de dualité. On peut citer par exemple :

- ACP sur données originales : $\mathbf{X} = \left[y_{ij} \right], \mathbf{Q} = \mathbf{I}_p, \mathbf{D} = \frac{1}{n}\mathbf{I}_n$
- ACP centrée par espèce : $\mathbf{X} = \left[y_{ij} - \bar{y}_j \right], \mathbf{Q} = \mathbf{I}_p, \mathbf{D} = \frac{1}{n}\mathbf{I}_n$ avec $\bar{y}_j = \frac{1}{n}\sum_{i=1}^{n} y_{ij}$

- ACP centrée par site : $\mathbf{X} = \left[y_{ij} - \overline{y}_i \right], \mathbf{Q} = \mathbf{I}_p, \mathbf{D} = \frac{1}{n}\mathbf{I}_n$ avec $\overline{y}_i = \frac{1}{p}\sum_{j=1}^{p} y_{ij}$
- ACP normée-centrée par espèce : $\mathbf{X} = \left[\frac{y_{ij} - \overline{y}_j}{\sigma_j} \right], \mathbf{Q} = \mathbf{I}_p, \mathbf{D} = \frac{1}{n}\mathbf{I}_n$ avec

  $\sigma_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{ij} - \overline{y}_j)^2}$
- ACP centrée sur profils relevés : $\mathbf{X} = \left[\frac{y_{ij}}{y_{i+}} - m_j \right], \mathbf{Q} = \mathbf{I}_p, \mathbf{D} = \frac{1}{n}\mathbf{I}_n$ avec $m_j = \frac{1}{n}\sum_{i=1}^{n}\frac{y_{ij}}{y_{i+}}$
- AFC : $\mathbf{X} = \left[\frac{y_{ij}}{y_{i+}y_{+j}} - 1 \right], \mathbf{Q} = diag(\frac{y_{+1}}{y_{++}},...,\frac{y_{+j}}{y_{++}},...,\frac{y_{+p}}{y_{++}}), \mathbf{D} = diag(\frac{y_{1+}}{y_{++}},...,\frac{y_{i+}}{y_{++}},...,\frac{y_{n+}}{y_{++}})$
- ANSC sur profils relevés : $\mathbf{X} = \left[\frac{y_{ij}}{y_{i+}} - \frac{y_{+j}}{y_{++}} \right], \mathbf{Q} = \mathbf{I}_p, \mathbf{D} = diag(\frac{y_{1+}}{y_{++}},...,\frac{y_{i+}}{y_{++}},...,\frac{y_{n+}}{y_{++}})$
- ANSC sur profils espèces : $\mathbf{X} = \left[\frac{y_{ij}}{y_{+j}} - \frac{y_{i+}}{y_{++}} \right], \mathbf{Q} = diag(\frac{y_{+1}}{y_{++}},...,\frac{y_{+j}}{y_{++}},...,\frac{y_{+p}}{y_{++}}), \mathbf{D} = \mathbf{I}_n$

avec $y_{i+} = \sum_{j=1}^{p} y_{ij}, y_{+j} = \sum_{i=1}^{n} y_{ij}, y_{++} = \sum_{i=1}^{n}\sum_{j=1}^{p} y_{ij}$.

I.2. Couplage

On considère maintenant une deuxième matrice de données **Z** qui peut être issue d'une transformation préalable des données brutes. Cette matrice contient les mesures de *q* variables (colonnes) sur les mêmes *n* individus (lignes). On considère un nouveau schéma de dualité induisant le triplet $(\mathbf{Z}, \mathbf{R}, \mathbf{D})$ où **R** est un produit scalaire de $\mathbb{R}^q$ (matrice carrée symétrique) et **D** le produit scalaire de $\mathbb{R}^n$ identique à celui du triplet $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$. On a alors deux schémas de dualité appariés par les lignes (figure 5) :

$$\begin{array}{ccc} \mathbb{R}^p & \xrightarrow{\mathbf{Q}} & \mathbb{R}^{p*} \\ \mathbf{X}^t \uparrow & & \downarrow \mathbf{X} \\ \mathbb{R}^{n*} & \xleftarrow[\mathbf{D}]{} & \mathbb{R}^n \end{array} \qquad \begin{array}{ccc} \mathbb{R}^q & \xrightarrow{\mathbf{R}} & \mathbb{R}^{q*} \\ \mathbf{Z}^t \uparrow & & \downarrow \mathbf{Z} \\ \mathbb{R}^{n*} & \xleftarrow[\mathbf{D}]{} & \mathbb{R}^n \end{array}$$

Figure 5 : Schémas de dualité associés à deux triplets appariés par les lignes $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$ et $(\mathbf{Z}, \mathbf{R}, \mathbf{D})$.

Le schéma de dualité du couplage de **X** et **Z** s'écrit naturellement et définit le triplet $(\mathbf{Z}^t\mathbf{D}\mathbf{X}, \mathbf{Q}, \mathbf{R})$ équivalant à $(\mathbf{X}^t\mathbf{D}\mathbf{Z}, \mathbf{R}, \mathbf{Q})$ (figure 6) :

$$
\begin{array}{ccc}
\mathbb{R}^{p} & \xrightarrow{\mathbf{Q}} & \mathbb{R}^{p*} \\
\mathbf{X}^{t} \uparrow & & \downarrow \mathbf{X} \\
\mathbb{R}^{n*} & & \mathbb{R}^{n} \\
\mathbf{D} \uparrow & & \downarrow \mathbf{D} \\
\mathbb{R}^{n} & & \mathbb{R}^{n*} \\
\mathbf{Z} \uparrow & & \downarrow \mathbf{Z}^{t} \\
\mathbb{R}^{q*} & \xleftarrow[\mathbf{R}]{} & \mathbb{R}^{q}
\end{array}
$$

Figure 6 : Schéma de dualité associé au triplet $(\mathbf{Z}^t\mathbf{DX}, \mathbf{Q}, \mathbf{R})$.

La diagonalisation d'un tel schéma se fait par les mêmes principes que ceux énoncés pour un schéma de dualité simple. Selon les conditions numériques induites par le nombre de variables par rapport au nombre d'individus et le choix du critère à optimiser, il existe trois grandes stratégies de couplage.

I.2.1. Analyse canonique des corrélations

La première et la plus ancienne voie est celle des analyses canoniques initiées par Hotelling (1936) et popularisées en écologie par Gittins (1985). L'analyse canonique correspond au cas particulier du schéma de dualité précédent où **X** et **Z** sont des tableaux normés-centrés par colonne, $\mathbf{R} = \left(\mathbf{Z}^t\mathbf{DZ}\right)^{-1}$ et $\mathbf{Q} = \left(\mathbf{X}^t\mathbf{DX}\right)^{-1}$ les produits scalaires. D'après les propriétés générales du schéma de dualité, il ressort les propriétés particulières de cette analyse :

- les facteurs principaux sont les coefficients de combinaisons linéaires des variables de **X** de variance unité :

$$\left\|\mathbf{a}^*_i\right\|^2_{\left(\left(\mathbf{X}^t\mathbf{DX}\right)^{-1}\right)^{-1}} = \mathbf{a}^{*t}_i\mathbf{X}^t\mathbf{DXa}^*_i = \left\|\mathbf{Xa}^*_i\right\|^2_{\mathbf{D}}$$

- les cofacteurs principaux sont les coefficients de combinaisons linéaires des variables de **Z** de variance unité :

$$\left\|\mathbf{b}^*_i\right\|^2_{\left(\left(\mathbf{Z}^t\mathbf{DZ}\right)^{-1}\right)^{-1}} = \mathbf{b}^{*t}_i\mathbf{Z}^t\mathbf{DZb}^*_i = \left\|\mathbf{Zb}^*_i\right\|^2_{\mathbf{D}}$$

- les combinaisons linéaires variables de variance unité de **X** et **Z** sont de corrélation maximale :

$$\begin{aligned}
&\left\langle \mathbf{Z}^t\mathbf{DX}\left(\mathbf{X}^t\mathbf{DX}\right)^{-1}\mathbf{a}_i \middle| \mathbf{b}_i \right\rangle_{\left(\mathbf{Z}^t\mathbf{DZ}\right)^{-1}} = \left\langle \mathbf{X}^t\mathbf{DZ}\left(\mathbf{Z}^t\mathbf{DZ}\right)^{-1}\mathbf{a}_i \middle| \mathbf{b}_i \right\rangle_{\left(\mathbf{X}^t\mathbf{DX}\right)^{-1}} \\
&= \mathbf{a}_i^{\,t}\left(\mathbf{X}^t\mathbf{DX}\right)^{-1}\mathbf{X}^t\mathbf{DZ}\left(\mathbf{Z}^t\mathbf{DZ}\right)^{-1}\mathbf{b}_i = \mathbf{a}^{*t}_i\mathbf{X}^t\mathbf{DZb}^*_i = \left\langle \mathbf{Xa}^*_i \middle| \mathbf{Zb}^*_i \right\rangle_{\mathbf{D}} \\
&= cor\left(\mathbf{Xa}^*_i, \mathbf{Zb}^*_i\right)
\end{aligned}$$

L'analyse canonique cherche une combinaison linéaire de variance unité des variables de **X** ($\mathbf{Xa}^*_1$) et une combinaison linéaire de variance unité des variables de **Z** ($\mathbf{Zb}^*_1$) de corrélation maximale. Les autres couples de combinaisons linéaires sont également de corrélation maximale sous contrainte d'orthogonalité. Les combinaisons linéaires sont les variables canoniques et la valeur propre est le carré de corrélation canonique.

Le **D**-projecteur orthogonal sur le sous-espace engendré par **X** s'écrit (Takeuchi *et al.* 1982) :

$$\mathbf{P_X} = \mathbf{X}\left(\mathbf{X}^t\mathbf{DX}\right)^{-1}\mathbf{X}^t\mathbf{D}$$

Une matrice de projection **D**-orthogonal vérifie les deux conditions suivantes :

- elle est idempotente : $\mathbf{P_X}^2 = \mathbf{P_X}$.
- elle est **D**-symétrique : $\mathbf{P_X}^t\mathbf{D} = \mathbf{DP_X}$.

Le schéma de dualité de l'analyse canonique peut alors se réécrire sous la forme d'un produit de projecteurs (figure 7) :

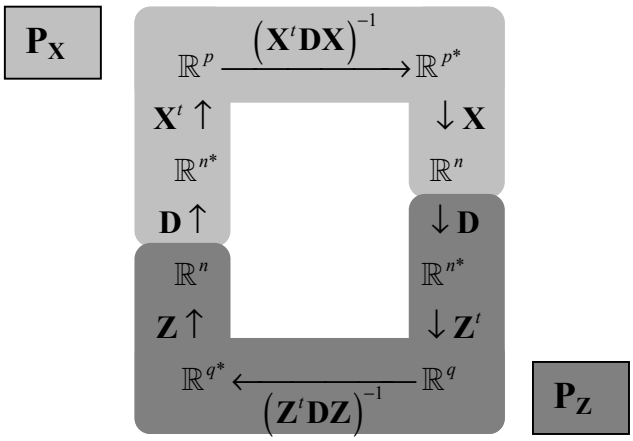


Figure 7 : Schéma de dualité de l'analyse canonique des corrélations entre **X** et **Z** correspondant au triplet $\left(\mathbf{Z}^t\mathbf{DX}, \left(\mathbf{X}^t\mathbf{DX}\right)^{-1}, \left(\mathbf{Z}^t\mathbf{DZ}\right)^{-1}\right)$.

Cette réécriture en terme de projecteurs fait apparaître clairement les deux régressions multiples sous-jacentes à l'analyse canonique. La variable canonique de **X** (respectivement **Z**) est une combinaison linéaire des variables de **X** (respectivement **Z**) qui est la plus prédictible par régression multiple sur les variables de **Z** (respectivement **X**). Les problèmes inhérents à la régression multiple sont donc présents en analyse canonique qui ne peut être utilisée que si le nombre d'individus est largement supérieur au nombre de variables ($\max(p,q) << n$).

Il est également remarquable que dans le cas où un des deux tableaux est disjonctif complet (indicatrices) induisant une partition des individus en plusieurs classes, l'analyse canonique est alors une analyse discriminante qui recherche des combinaisons linéaires des variables quantitatives maximisant le rapport de la variance interclasse sur la variance totale.

I.2.2. Analyses sur variables instrumentales

La deuxième grande stratégie est celles des analyses sur variables instrumentales (synthèse dans Lebreton *et al.* 1991). Parmi ces méthodes, on retrouve l'analyse canonique des correspondances (ACC, ter Braak 1986), appelée également analyse factorielle des correspondances sur variables instrumentales (AFCVI, Chessel *et al.* 1987, Lebreton *et al.* 1988a, Lebreton *et al.* 1988b), ainsi que l'analyse des redondances (Wollenberg 1977) ou analyse en composantes principales sur variables instrumentales (ACPVI, Rao 1964). Contrairement à l'analyse canonique, les analyses sur variables instrumentales sont foncièrement dissymétriques. Le tableau **X** contient les variables à prédire et le tableau **Z** les variables explicatives. D'un point de vue pratique, l'ACPVI peut être vue comme une ACP des prédictions du tableau **X** obtenues par régressions multiples sur les variables de **Z**. L'ACPVI est simplement l'analyse du triplet $(\mathbf{Z}^t\mathbf{DX},\mathbf{Q},\mathbf{R})$ avec $\mathbf{D}=\frac{1}{n}\mathbf{I}_n$, $\mathbf{R}=\left(\mathbf{Z}^t\mathbf{DZ}\right)^{-1}$ et $\mathbf{Q}=\mathbf{I}_p$. Le triplet précédent peut se réécrire $(\mathbf{P_Z X},\mathbf{Q},\mathbf{D})$. Cette nouvelle formulation permet d'identifier clairement l'étape de régression multiple qui s'exprime sous la forme d'une projection. D'un point de vue théorique, deux possibilités sont envisageables pour l'interprétation des résultats (figure 8) :

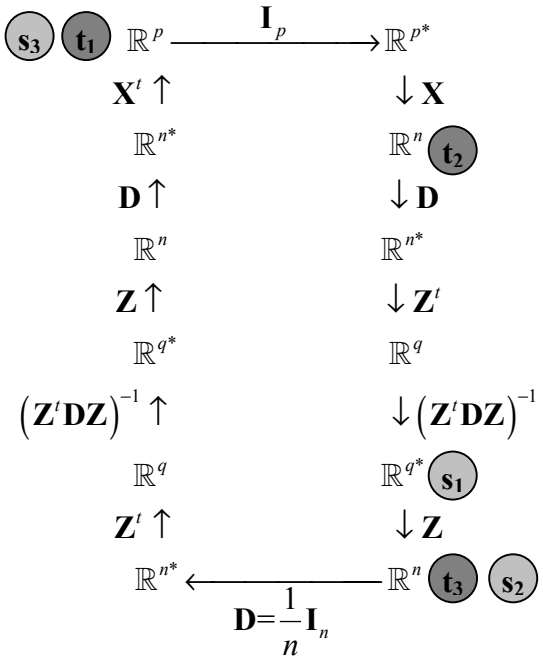


Figure 8 : Schéma de dualité de l'ACPVI correspondant au triplet $(\mathbf{P_Z X},\mathbf{I}_p,\frac{1}{n}\mathbf{I}_n)$.

L'ACPVI recherche des coefficients ($\mathbf{s}_1$) des variables de **Z**. La combinaison linéaire obtenue ($\mathbf{s}_2=\mathbf{Zs}_1$) est une composante principale sous contrainte ou composante explicative

(Obadia 1978). D'après les propriétés générales d'un schéma de dualité on sait que $\|\mathbf{s}_2\|^2_{\frac{1}{n}\mathbf{I}_n} = 1$ et que la quantité $\left\|(\mathbf{P_Z X})^t(\frac{1}{n}\mathbf{I}_n)\mathbf{s}_2\right\|^2_{\mathbf{I}_p} = \mathbf{s}_2{}^t(\frac{1}{n}\mathbf{I}_n)\mathbf{P_Z X I}_p\mathbf{X}^t\mathbf{P_Z}{}^t(\frac{1}{n}\mathbf{I}_n)\mathbf{s}_2$ est maximisée.

Les propriétés d'un projecteur orthogonal permettent de reformuler cette écriture :

$$\mathbf{s}_2{}^t(\frac{1}{n}\mathbf{I}_n)\mathbf{P_Z X I}_p\mathbf{X}^t\mathbf{P_Z}{}^t(\frac{1}{n}\mathbf{I}_n)\mathbf{s}_2 = \mathbf{s}_2{}^t\mathbf{P_Z}{}^t(\frac{1}{n}\mathbf{I}_n)\mathbf{X I}_p\mathbf{X}^t(\frac{1}{n}\mathbf{I}_n)\mathbf{P_Z s}_2$$

Par construction, on sait que $\mathbf{s}_2 = \mathbf{Z s}_1$ et $\mathbf{P_Z s}_2 = \mathbf{s}_2$ d'où :

$$\mathbf{s}_2{}^t\mathbf{P_Z}{}^t(\frac{1}{n}\mathbf{I}_n)\mathbf{X I}_p\mathbf{X}^t(\frac{1}{n}\mathbf{I}_n)\mathbf{P_Z s}_2 = \mathbf{s}_2{}^t(\frac{1}{n}\mathbf{I}_n)\mathbf{X I}_p\mathbf{X}^t(\frac{1}{n}\mathbf{I}_n)\mathbf{s}_2 = \left\|\mathbf{X}^t(\frac{1}{n}\mathbf{I}_n)\mathbf{s}_2\right\|^2_{\mathbf{I}_p} = \sum_{j=1}^{p} corr^2(\mathbf{X}_j, \mathbf{s}_2)$$

La composante explicative est donc une combinaison linéaire des variables de **Z** qui maximise la somme des carrés des corrélations avec les variables de **X** appelée critère de Stewart et Love (Stewart & Love 1968). Les variables à prédire (colonnes de **X**) sont alors représentées par leurs corrélations avec la composante explicative :

$$\mathbf{s}_3 = (\mathbf{P_Z X})^t(\frac{1}{n}\mathbf{I}_n)\mathbf{s}_2 = \mathbf{X}^t(\frac{1}{n}\mathbf{I}_n)\mathbf{P_Z s}_2 = \mathbf{X}^t(\frac{1}{n}\mathbf{I}_n)\mathbf{s}_2$$

La deuxième voie pour interpréter une ACPVI est de calculer un pseudo axe principal $\mathbf{t}_1$ vérifiant $\|\mathbf{t}_1\|^2_{\mathbf{I}_p} = 1$ et maximisant la quantité $\|\mathbf{P_Z X I}_p\mathbf{t}_1\|^2_{\frac{1}{n}\mathbf{I}_n} = \mathbf{t}_1{}^t\mathbf{I}_p\mathbf{X}^t\mathbf{P_Z}{}^t(\frac{1}{n}\mathbf{I}_n)\mathbf{P_Z X I}_p\mathbf{t}_1$. Ainsi, $\mathbf{t}_1$ contient les coefficients d'une combinaison linéaire des variables de **X** maximisant la variance expliquée par **Z** car $\mathbf{P_Z X I}_p\mathbf{t}_1$ contient les prédictions de la régression de $\mathbf{X I}_p\mathbf{t}_1$ sur les variables de **Z** et $\|\mathbf{P_Z X I}_p\mathbf{t}_1\|^2_{\frac{1}{n}\mathbf{I}_n}$ est donc la variance de ces prédictions (i.e. variance expliquée). Les projections des lignes de **X** sur le pseudo axe principal ($\mathbf{t}_2 = \mathbf{X I}_p\mathbf{t}_1$) contient les combinaisons des variables de **X** maximisant la variance expliquée par **Z**. Enfin, $\mathbf{t}_3 = \mathbf{P_Z X I}_p\mathbf{t}_1 = \mathbf{P_Z t}_2$ contient les prédictions obtenues par régression de $\mathbf{t}_2$ sur **Z**.

L'ACPVI fournit donc un compromis entre l'analyse canonique (maximisation du carré de la corrélation multiple) et l'analyse en composantes principales (maximisation de la variance) en maximisant la variance prédite (maximisation du produit). Contrairement au schéma de l'analyse canonique qui possède deux projecteurs, celui de l'ACPVI n'en contient qu'un seul. Par conséquent, l'ACPVI est moins sensible aux dimensions des tableaux et requiert seulement que le nombre de variables explicatives soit largement inférieur au nombre d'individus ($q << n$).

Dans l'article original de Rao (1964, pages 342-344), on trouve des propositions sur les analyses orthogonales, compléments naturels des analyses sur variables instrumentales. Les analyses orthogonales permettent de trouver des combinaisons linéaires de **X** qui soient à la fois indépendantes de **Z** (non corrélées aux variables de **Z**) et de variance maximum. Le projecteur **D**-orthogonal sur le sous-espace orthogonal à **Z** s'écrit :

$$\mathbf{P}_{\mathbf{Z}^\perp} = \mathbf{I}_n - \mathbf{Z}\left(\mathbf{Z}^t\mathbf{D}\mathbf{Z}\right)^{-1}\mathbf{Z}^t\mathbf{D}$$

Une ACPVI orthogonale définit le triplet $(\mathbf{P}_{\mathbf{Z}^\perp}\mathbf{X}, \mathbf{Q}, \mathbf{D})$ ou $(\mathbf{X} - \mathbf{P}_{\mathbf{Z}}\mathbf{X}, \mathbf{Q}, \mathbf{D})$ et recherche des combinaisons linéaires des variables de **X** indépendantes de **Z** (i.e. non corrélées aux variables de **Z**) de variance maximale.

Dans le cas où le tableau **Z** est disjonctif complet (indicatrices) induisant une partition des individus en plusieurs groupes, l'ACPVI est alors une analyse inter-classes qui recherche des combinaisons linéaires des variables quantitatives maximisant la variance interclasse et l'ACPVI orthogonale est une analyse intra-classes (Dolédec & Chessel 1987, 1989).

I.2.3. Analyse de co-inertie

La troisième approche est celle de la co-inertie (Dolédec & Chessel 1994) qui analyse le triplet $(\mathbf{Z}^t\mathbf{D}\mathbf{X}, \mathbf{Q}, \mathbf{R})$ sans la moindre contrainte sur les matrices **D**, **R** et **Q**. Le principe de cette méthode est donc très général et permet de coupler des triplets issus de diverses analyses simples. Dans le cas du couplage de deux triplets d'ACP normée, l'analyse inter-batterie (Tucker 1958) apparaît comme un cas particulier de l'analyse de co-inertie (Chessel & Mercier 1993). La co-inertie est une mesure de la structure commune des deux tableaux. Elle est définie par :

$$CoI(\mathbf{Z}^t\mathbf{D}\mathbf{X}, \mathbf{Q}, \mathbf{R}) = trace(\mathbf{Z}^t\mathbf{D}\mathbf{X}\mathbf{Q}\mathbf{X}^t\mathbf{D}\mathbf{Z}\mathbf{R})$$

Le coefficient *RV*, compris entre 0 et 1, permet de mesurer la corrélation entre deux typologies (Escoufier 1973). Il s'écrit simplement :

$$RV = \frac{CoI(\mathbf{Z}^t\mathbf{D}\mathbf{X}, \mathbf{Q}, \mathbf{R})}{\sqrt{CoI(\mathbf{X}^t\mathbf{D}\mathbf{X}, \mathbf{Q}, \mathbf{Q})}\sqrt{CoI(\mathbf{Z}^t\mathbf{D}\mathbf{Z}, \mathbf{R}, \mathbf{R})}}$$

La co-inertie est donc un critère qui mesure à la fois la corrélation entre **X** et **Z** mais également la variabilité dans **X** et **Z**. Les individus (lignes) induisent deux configurations de points dans $\mathbb{R}^p$ et $\mathbb{R}^q$. L'analyse de co-inertie cherche des couples de vecteurs $(\mathbf{a}_i \in \mathbb{R}^p, \mathbf{b}_i \in \mathbb{R}^q)$ maximisant la racine de la co-inertie projetée :

$$\left\langle \mathbf{Z}^t\mathbf{D}\mathbf{X}\mathbf{Q}\mathbf{a}_i \middle| \mathbf{b}_i \right\rangle_{\mathbf{R}} = \left\langle \mathbf{X}^t\mathbf{D}\mathbf{Z}\mathbf{R}\mathbf{b}_i \middle| \mathbf{a}_i \right\rangle_{\mathbf{Q}} = \mathbf{a}_i{}^t\mathbf{Q}\mathbf{X}^t\mathbf{D}\mathbf{Z}\mathbf{R}\mathbf{b}_i$$

Ces vecteurs sont appelés axes de co-inertie. Les vecteurs $\mathbf{a}_1, ..., \mathbf{a}_i, ..., \mathbf{a}_p$ sont orthogonaux (idem pour $\mathbf{b}_1, ..., \mathbf{b}_i, ..., \mathbf{b}_q$) et orthogonaux avec ceux de l'autre sous-espace à l'exception de celui de même rang. Deux jeux de coordonnées des lignes sont obtenus en projetant les lignes de **X** sur $\mathbf{a}_i$ et **Z** sur $\mathbf{b}_i$:

$$\mathbf{l}_{\mathbf{X}i} = \mathbf{X}\mathbf{Q}\mathbf{a}_i$$
$$\mathbf{l}_{\mathbf{Z}i} = \mathbf{Z}\mathbf{R}\mathbf{b}_i$$

Les coordonnées des colonnes de **X** et **Z** sont obtenues par les formules classiques du schéma de dualité :

$$\mathbf{c}_{\mathbf{X}_i} = \mathbf{X}^t\mathbf{DZRb}_i = \sqrt{\lambda_i}\mathbf{a}_i$$
$$\mathbf{c}_{\mathbf{Z}_i} = \sqrt{\lambda_i}\mathbf{b}_i$$

Contrairement au schéma de l'analyse canonique qui contient deux métriques de Mahalanobis et à celui des analyses sur variables instrumentales qui en contient une, les métriques de l'analyse de co-inertie proviennent des analyses simples. Par conséquent, il n'y a aucune contrainte en analyse de co-inertie concernant la dimension des tableaux analysés. Quand l'analyse de co-inertie fait le lien entre n points de $\mathbb{R}^p$ et n points de $\mathbb{R}^q$, l'analyse canonique fait le lien entre p points de $\mathbb{R}^n$ et q points de $\mathbb{R}^n$, et une analyse sur variable instrumentale fait le lien entre n points de $\mathbb{R}^p$ et q points de $\mathbb{R}^n$ (figure 9).

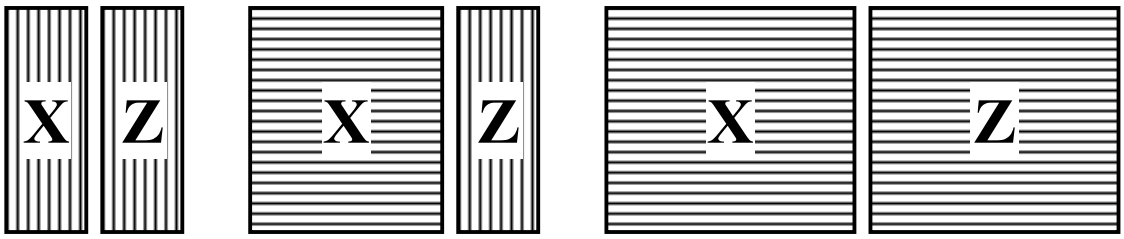


Figure 9 : Cadre d'utilisation de l'analyse canonique des corrélations, des analyses sur variables instrumentales et de l'analyse de co-inertie.

I.3. Correspondances et analyses canoniques

Pour commencer, voici quelques extraits de conversations qui se sont déroulées sur le forum ORDNEWS (http://www.okstate.edu/artsci/botany/ordinate/ordnews.htm) :

Mark Easter (6 Mars 1998) :

> *Dear ORDNEWS,*
>
> *Does anybody out there have any experience with using SAS for canonical correlation analysis? I'm trying to find alternatives to CANOCO for analysis of a dataset from a benthic macroinvertebrates study. I've used CANOCO before, however I want to avoid it primarily because of the limitations invoked by its interface and strict data input file format requirements.*
>
> *I've read in the SAS literature that the procedure can be accomplished using the PROC CANCORR procedure or using the CANPRINT option with the PROC REG procedure. Before I dive into it, I'd like to know if anybody else has done it that way, and whether the SAS PROC CANCORR procedure and CANOCO are analogous. I'm concerned that there are subtle (or not so subtle) differences in their implementation that could prove to be significant.*
>
> *Thanks for any help anyone can provide.*

Steve Bousquin (6 Mars 1998) :

Mark:

Canonical correlation (proc cancorr in SAS) and canonical correspondence analyis (in CANOCO) are not the same thing.

Pedro R. Peres-Neto (6 Mars 1998) :

Hi Mark, actually canoco and CCA (canonical correrlation analysis) are quite different, and SAS does no permit, as long as I know, to perform CANOCO. Please, let me know otherwise. There is a lot of confusion in terminology...some people refers to CANOCO as CCA, but I understand that CCA was coined initially for Canonical correlation and not for canonical correspondence. I assume that a lot of us make confusion among terms. In any case, the differences between these two techniques can be found in Jongman et al. (data analysis in community and landscape ecology).

Mike Palmer (6 Mars 1998) :

CANCORR is for canonical correlation, which is a multivariate multiple regression technique (i.e. there are multiple independent and dependent variables). It is a really great technique for some purposes. However, it is most definitely not the same thing as canonical correspondence analysis (CCA).

Un mois plus tard…

Tim Pearce (24 Avril 1998) :

Here is a summary of the numerous responses to my request for information on programs to perform Canonical Correspondence Analysis. […] One respondent is certain SAS does CCA, and believes SYSTAT will.

Mike Palmer (24 Avril 1998) :

Timothy et al: Please be aware that "CCA" is often used as an abbreviation for "Canonical Correlation Analysis", which is indeed performed by a number of software packages. It is a very useful technique, but is not the same thing as (nor should it be substituted for) Canonical Correspondence Analysis (which among ecologists, is the most common use of the acronym "CCA").

Enfin, un extrait d'échanges sur la liste R-HELP (http://www.r-project.org/mail.html) :

Patrick Foley (16 Février 2001) :

Is there an R function that does canonical correspondence analysis. Can it be done using the VR function corresp()? If not, how hard it be to write R code to do it? I am a population biologist with long but patchy programming experience in C, Smalltalk, Java and other languages.

Thanks,

Brian D. Ripley (17 Février 2001) :

What is `canonical correspondence analysis'? Correspondence analysis is a variation on canonical correlation analysis. None of my sources have `canonical correspondence analysis', but my guess is that it is a re-naming of something else.

L'analyse canonique des correspondances a été proposée sous la forme d'une analyse des correspondances sous contrainte (Takane *et al.* 1991). Le nom choisi par ter Braak (1986, 1987) n'est donc pas très judicieux car il induit plus de symétrie qu'il n'y en a dans la méthode (Sabatier *et al.* 1989). De plus son abréviation (CCA) était déjà utilisée pour l'analyse canonique des corrélations. Les discussions sur les forums montrent bien les incompréhensions suscitées par ces problèmes de terminologie. Même s'il existe des similitudes dans leurs noms, l'analyse canonique, l'analyse canonique des correspondances et l'analyse des correspondances sont bel et bien des méthodes différentes. L'exploration de la théorie sous-jacente à ces méthodes, à l'aide du schéma de dualité, permet cependant d'identifier de nombreuses analogies.

I.3.1. Analyses des correspondances

Soit $\mathbf{N} = \left[n_{ij} \right]$ un tableau de nombres positifs avec I lignes et J colonnes. On définit les valeurs suivantes :

$$n_{i+} = \sum_{j=1}^{J} n_{ij} \text{ , } n_{+j} = \sum_{i=1}^{I} n_{ij} \text{ , et } n = \sum_{i=1}^{I} n_{i+} = \sum_{j=1}^{J} n_{+j} = n_{++}$$

Le tableau $\mathbf{P}$ des fréquences relatives s'écrit simplement $\mathbf{P} = \left[p_{ij} = \frac{n_{ij}}{n} \right]$. Les matrices des $\mathbf{D}_I = diag(p_{1+},...,p_{i+},...,p_{I+})$ et $\mathbf{D}_J = Diag(p_{+1},...,p_{+j},...,p_{+J})$ contiennent respectivement les poids des lignes et des colonnes (avec $p_{i+} = n_{i+}/n$ et $p_{+j} = n_{+j}/n$). Si on omet la première valeur propre triviale ($\lambda_0 = 1$), l'analyse factorielle des correspondances définit le triplet $\left(\mathbf{D}_I^{-1}\mathbf{P}\mathbf{D}_J^{-1}, \mathbf{D}_J, \mathbf{D}_I \right)$. Cazes *et al.* (1988) montre que cette analyse est équivalente à celle de $\left(\mathbf{P}\mathbf{D}_J^{-1}, \mathbf{D}_J, \mathbf{D}_I^{-1} \right)$ ou $\left(\mathbf{D}_I^{-1}\mathbf{P}, \mathbf{D}_J^{-1}, \mathbf{D}_I \right)$ ou encore $\left(\mathbf{P}, \mathbf{D}_J^{-1}, \mathbf{D}_I^{-1} \right)$ (figure 10).

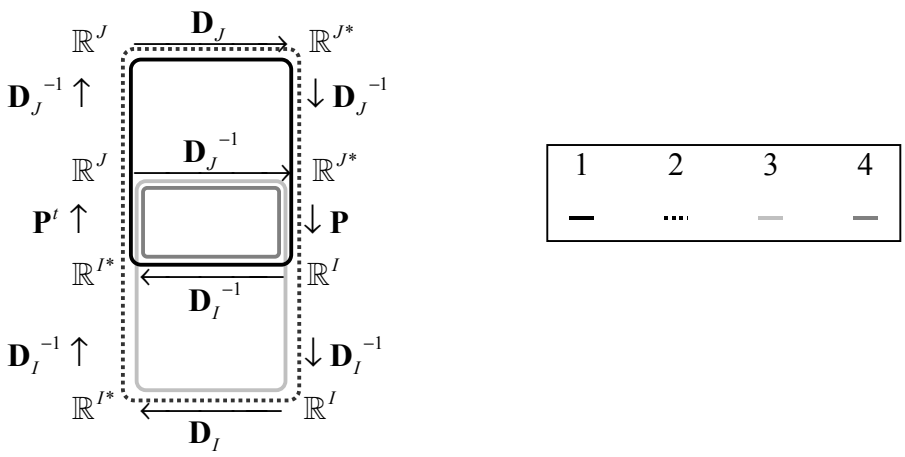


Figure 10 : Schéma de dualité de l'AFC d'après Cazes *et al.* (1988).

Considérons maintenant un tableau **E** $\mathbf{D}_I$-centré contenant les mesures de M variables (colonnes) pour les I catégories (lignes). L'analyse canonique des correspondances est une analyse des correspondances avec la contrainte que le score des lignes doit être une combinaison linéaire des variables externes contenues dans **E**. Cette contrainte peut être exprimée sous la forme d'un projecteur et le schéma de dualité de l'ACC s'écrit $\left(\mathbf{P}_{\mathbf{E}}\mathbf{D}_I^{-1}\mathbf{P}\mathbf{D}_J^{-1}, \mathbf{D}_J, \mathbf{D}_I\right)$ avec $\mathbf{P}_{\mathbf{E}} = \mathbf{E}\left(\mathbf{E}^t\mathbf{D}_I\mathbf{E}\right)^{-1}\mathbf{E}^t\mathbf{D}_I$. Cette analyse est équivalente à celle du schéma $\left(\mathbf{P}^t\mathbf{E}, \left(\mathbf{E}^t\mathbf{D}_I\mathbf{E}\right)^{-1}, \mathbf{D}_J^{-1}\right)$ (figure 11) :

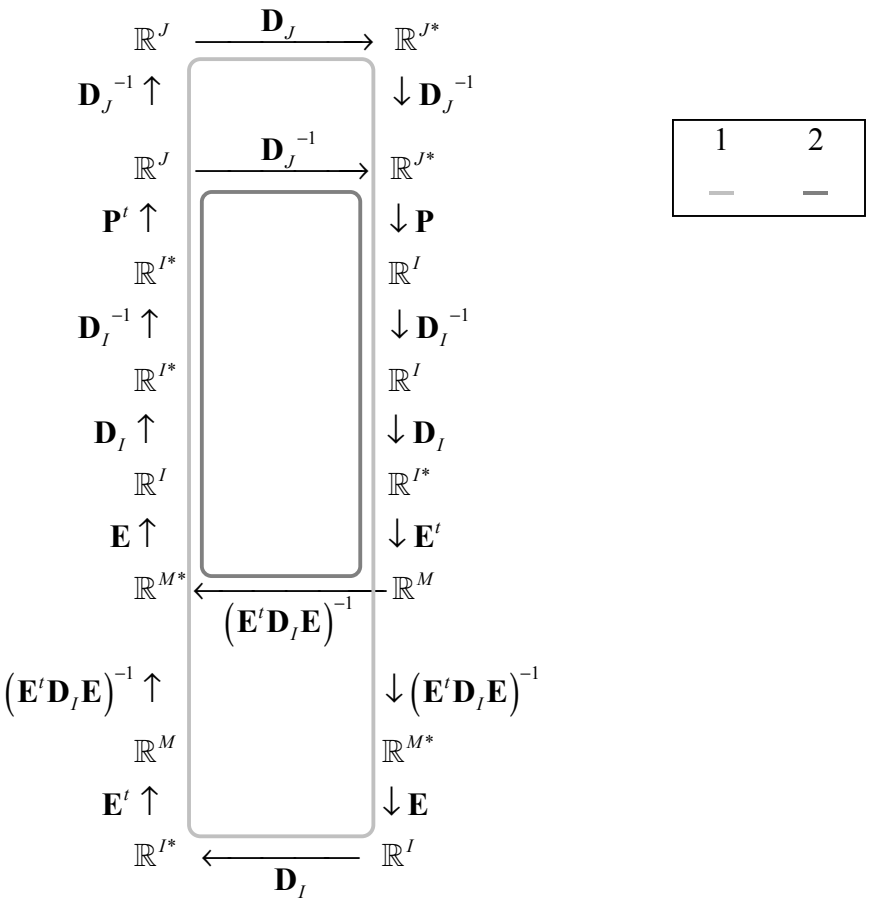


Figure 11 : Schéma de dualité de l'ACC.

Dans sa forme usuelle, l'ACC contient des contraintes sur le score des lignes mais elle peut s'accommoder de contraintes sur le score des colonnes (*e.g.* Ojeda *et al.* 1998). Si **F** est un tableau $\mathbf{D}_J$-centré avec J lignes et S colonnes, on analyse alors $\left(\mathbf{P}_{\mathbf{F}}\left(\mathbf{D}_I^{-1}\mathbf{P}\mathbf{D}_J^{-1}\right)^t, \mathbf{D}_I, \mathbf{D}_J\right)$. Les contraintes induites par les tableaux **E** et **F** peuvent être introduites simultanément dans une même analyse (Böckenholt & Böckenholt 1990). Cette méthode qui s'apparente à une variante, de type analyse sur variables instrumentales, de la méthode RLQ (Dolédec *et al.* 1996) a été dénommée double analyse canonique des correspondances par Lavorel *et al.* (1999, 1998). Dans ces articles, les auteurs introduisent cette méthode en écologie et proposent un algorithme basé sur deux analyses canoniques des correspondances successives (figure 12).

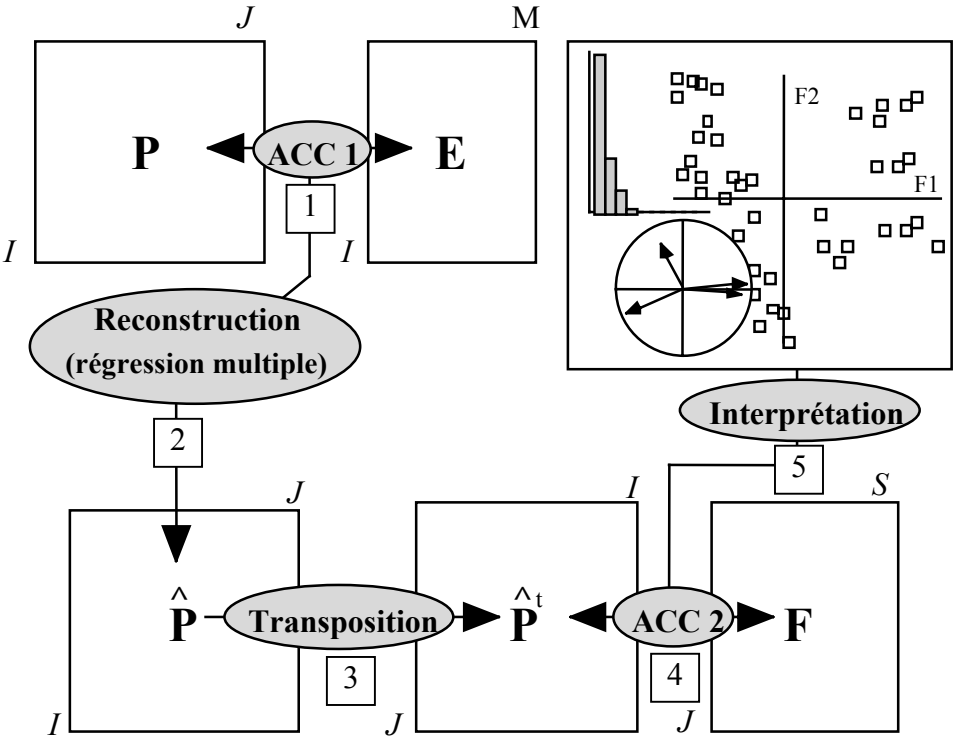


Figure 12 : Principe de la double ACC d'après Lavorel *et al.* (1999).

D'un point de vue formel, cette méthode se résume à l'analyse du schéma de dualité $\left(\mathbf{P_F}\left(\mathbf{P_E}\mathbf{D}_I^{-1}\mathbf{P}\mathbf{D}_J^{-1}\right)^t, \mathbf{D}_I, \mathbf{D}_J\right)$ qui contient les deux opérateurs de projection. La manipulation de ce schéma permet facilement d'identifier la double ACC comme l'analyse du tableau croisé avec les deux métriques de Mahalanobis $\left(\mathbf{F}^t\mathbf{P}^t\mathbf{E}, \left(\mathbf{E}^t\mathbf{D}_I\mathbf{E}\right)^{-1}, \left(\mathbf{F}^t\mathbf{D}_J\mathbf{F}\right)^{-1}\right)$.

I.3.2. Tableaux des correspondances

Selon Greenacre (1984), le terme « *correspondances* » a été utilisé par Benzécri pour signifier le système d'association entre les lignes et les colonnes d'un tableau. Le lien entre les lignes et les colonnes est mesuré uniquement à l'aide des cases non nulles de la table de contingence ; on appellera donc « correspondances » les cases non nulles d'une table de contingence dans la suite de ce document.

Considérons les C correspondances ou valeurs non nulles du tableau **N**. Il est aisé de construire deux tableaux des correspondances **X** et **Y** à partir du tableau **N** (figure 13). Pour la *c-ième* correspondance, on a :

$$\mathbf{X}_{cj} = \begin{cases} 1 \text{ si la } c-ième \text{ correspondance appartient à la } j-ième \text{ colonne } (1 \le j \le J) \\ 0 \text{ sinon} \end{cases}$$

$$\mathbf{Y}_{ci} = \begin{cases} 1 \text{ si la } c-ième \text{ correspondance appartient à la } i-ième \text{ ligne } (1 \le i \le I) \\ 0 \text{ sinon} \end{cases}$$

La matrice diagonale $\mathbf{D}_C$ contient le poids des correspondances.

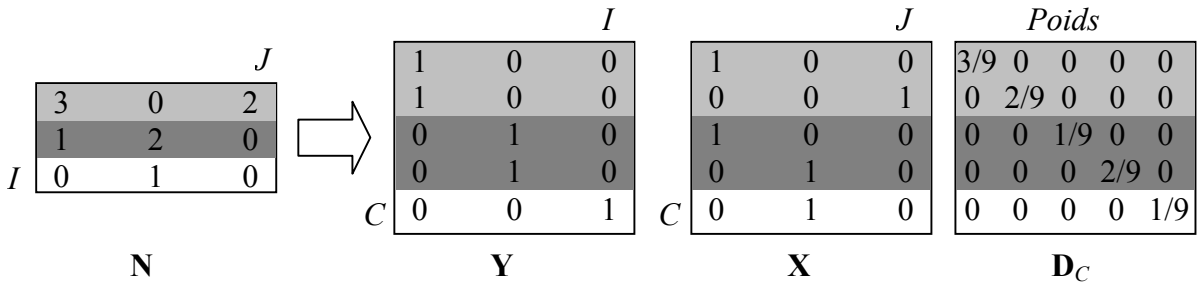


Figure 13 : Transformation d'un tableau espèces-relevés en tableaux de correspondances.

Par construction, on a les relations suivantes :

$$\mathbf{X}^t\mathbf{D}_C\mathbf{X} = \mathbf{D}_J, \mathbf{Y}^t\mathbf{D}_C\mathbf{Y} = \mathbf{D}_I, \mathbf{Y}^t\mathbf{X} = \mathbf{N} \text{ et } \mathbf{Y}^t\mathbf{D}_C\mathbf{X} = \mathbf{P}$$

Les données externes relatives aux lignes de **N** et contenues dans le tableau **E** peuvent être dupliquées afin d'obtenir cette information au niveau de la correspondance (figure 14) :

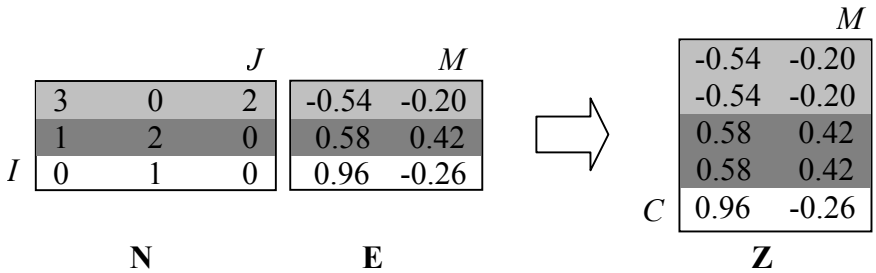


Figure 14 : Transformation d'un tableau relevés-variables de milieu en tableau de correspondances.

Le tableau **Z** vérifie les relations suivantes :

$$\mathbf{Z} = \mathbf{YE} \text{ et } \mathbf{E} = \left(\mathbf{Y}^t\mathbf{D}_C\mathbf{Y}\right)^{-1}\mathbf{Y}^t\mathbf{D}_C\mathbf{Z} = \mathbf{Y}^-\mathbf{Z}$$

où $\mathbf{Y}^-$ l'inverse généralisée de **Y** satisfait les conditions :

$$\mathbf{Y}^-\mathbf{Y}\mathbf{Y}^- = \mathbf{Y}^- \text{ et } \mathbf{Y}\mathbf{Y}^-\mathbf{Y} = \mathbf{Y}$$

De plus,

$$\mathbf{Y}\mathbf{Y}^- = \mathbf{P}_\mathbf{Y} \text{ et } \mathbf{Y}^-\mathbf{Y} = \mathbf{I}_C$$

La même opération est réalisée à partir du tableau **F** contenant l'information relative aux colonnes de **N**. La duplication des données permet de construire une matrice **W** vérifiant :

$$\mathbf{W} = \mathbf{XF} \text{ et } \mathbf{F} = \left(\mathbf{X}^t\mathbf{D}_C\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{D}_C\mathbf{W} = \mathbf{X}^-\mathbf{W}$$

$$\mathbf{X}^-\mathbf{X}\mathbf{X}^- = \mathbf{X}^- \text{ et } \mathbf{X}\mathbf{X}^-\mathbf{X} = \mathbf{X}$$

$$\mathbf{X}\mathbf{X}^- = \mathbf{P}_\mathbf{X} \text{ et } \mathbf{X}^-\mathbf{X} = \mathbf{I}_C$$

I.3.3. Analyses canoniques

L'AFC du tableau **N** a été définie comme l'analyse du triplet $\left(\mathbf{P},\mathbf{D}_J^{-1},\mathbf{D}_I^{-1}\right)$. A partir des relations établies ci-dessus, ce triplet est équivalent à $\left(\mathbf{Y}^t\mathbf{D}_C\mathbf{X},\left(\mathbf{X}^t\mathbf{D}_C\mathbf{X}\right)^{-1},\left(\mathbf{Y}^t\mathbf{D}_C\mathbf{Y}\right)^{-1}\right)$. On reconnaît le triplet de l'analyse canonique de **Y** et **X**. L'AFC cherche donc des combinaisons linéaires des indicatrices des lignes et des indicatrices des colonnes de corrélation maximale. C'est également une double analyse discriminante car les tableaux **X** et **Y** contiennent des indicatrices de classes. Si l'AFC est une analyse canonique, il est également intéressant de noter que l'analyse non symétrique des correspondances (Gimaret-Carpentier *et al.* 1998, Kroonenberg & Lombardo 1999, Lauro & D'ambra 1984) dans sa version profils lignes est équivalente à l'ACPVI qui cherche à expliquer les indicatrices colonnes (**X**) par les indicatrices lignes (**Y**). La version profils colonnes de l'ANSC est l'ACPVI qui cherche à expliquer les indicatrices lignes (**Y**) par les indicatrices colonnes (**X**).

L'ACC du couple **N**, **E** est l'analyse de $\left(\mathbf{P}^t\mathbf{E},\left(\mathbf{E}^t\mathbf{D}_I\mathbf{E}\right)^{-1},\mathbf{D}_J^{-1}\right)$. Par définition, $\mathbf{Z}=\mathbf{YE}$ donc le sous-espace engendré par **Z** est contenu dans celui engendré par **Y** et $\mathbf{P_Y Z}=\mathbf{Z}$. Le triplet de l'ACC est donc équivalent à $\left(\mathbf{X}^t\mathbf{D}_C\mathbf{Z},\left(\mathbf{Z}^t\mathbf{D}_C\mathbf{Z}\right)^{-1},\left(\mathbf{X}^t\mathbf{D}_C\mathbf{X}\right)^{-1}\right)$ car :

$$\mathbf{P}^t\mathbf{E}=\mathbf{X}^t\mathbf{D}_C\mathbf{Y}\mathbf{Y}^-\mathbf{Z}=\mathbf{X}^t\mathbf{D}_C\mathbf{P_Y Z}=\mathbf{X}^t\mathbf{D}_C\mathbf{Z}$$

$$\left(\mathbf{E}^t\mathbf{D}_I\mathbf{E}\right)^{-1}=\mathbf{Z}^t\left(\mathbf{Y}^-\right)^t\mathbf{Y}^t\mathbf{D}_C\mathbf{Y}\mathbf{Y}^-\mathbf{Z}=\mathbf{Z}^t\left(\mathbf{P_Y}\right)^t\mathbf{D}_C\mathbf{P_Y Z}=\mathbf{Z}^t\mathbf{D}_C\mathbf{P_Y Z}=\mathbf{Z}^t\mathbf{D}_C\mathbf{Z}$$

$$\mathbf{D}_J^{-1}=\left(\mathbf{X}^t\mathbf{D}_C\mathbf{X}\right)^{-1}$$

L'analyse canonique des correspondances est donc une analyse canonique ! Elle recherche des combinaisons linéaires des variables de **Z** de corrélation maximale avec les indicatrices des colonnes. Autrement dit, c'est une analyse discriminante qui recherche des combinaisons linéaires des variables de **Z** maximisant la séparation des colonnes de **N**. Comme l'ont admis ter Braak et Verdonschot (1995), l'analyse discriminante de Green (1971, 1974) peut être considérée comme l'ancêtre de l'ACC. Cependant, il ne faut pas omettre que les individus statistiques sont les *C* correspondances dans l'analyse canonique alors que ce sont les *I* catégories-lignes en analyse canonique des correspondances.

La double ACC définie par le triplet $\left(\mathbf{F}^t\mathbf{P}^t\mathbf{E},\left(\mathbf{E}^t\mathbf{D}_I\mathbf{E}\right)^{-1},\left(\mathbf{F}^t\mathbf{D}_J\mathbf{F}\right)^{-1}\right)$ coïncide avec l'analyse canonique entre les tableaux **Z** et **W** identifiée simplement par le triplet $\left(\mathbf{W}^t\mathbf{D}_C\mathbf{Z},\left(\mathbf{Z}^t\mathbf{D}_C\mathbf{Z}\right)^{-1},\left(\mathbf{W}^t\mathbf{D}_C\mathbf{W}\right)^{-1}\right)$.

Le fait d'identifier les analyses des correspondances avec ou sans contraintes comme des analyses canoniques présente un intérêt théorique certain mais également plusieurs avantages pratiques. Considérons par exemple le cas de l'AFC vue comme analyse canonique. Pour la *k–ième* valeur propre, l'analyse recherche un couple de variables canoniques ($\mathbf{Xa}^*_k,\mathbf{Yb}^*_k$) de variance unité et de corrélation maximale. Si λ_k est la valeur propre associée à ce couple de vecteurs, leur corrélation vaut $\sqrt{\lambda_k}$ et correspond géométriquement au cosinus de l'angle formé par les deux variables canoniques ($\cos(\alpha_k)=\sqrt{\lambda_k}$). Comme l'analyse

canonique travaille simultanément sur deux sous-espaces, la représentation des résultats n'est pas simple si l'on veut conserver l'aspect symétrique de la méthode (ter Braak 1990). Une solution proposée est l'utilisation du score canonique défini comme le vecteur unité porté par la bissectrice de l'angle formé par les deux variables canoniques (figure 15). Le score canonique donne une position propre à chaque correspondance induite à la fois par le score de sa ligne et par celui de sa colonne. Son calcul se fait simplement (Thioulouse & Chessel 1992) :

$$\mathbf{s}_k = \frac{\mathbf{Xa}^*_k + \mathbf{Yb}^*_k}{\left\|\mathbf{Xa}^*_k + \mathbf{Yb}^*_k\right\|_{\mathbf{D}_C}} = \frac{\mathbf{Xa}^*_k + \mathbf{Yb}^*_k}{\sqrt{2\left(1+\sqrt{\lambda_k}\right)}}$$

car $\left\|\mathbf{Xa}^*_k + \mathbf{Yb}^*_k\right\|^2_{\mathbf{D}_C} = \left\|\mathbf{Xa}^*_k\right\|^2_{\mathbf{D}_C} + \left\|\mathbf{Yb}^*_k\right\|^2_{\mathbf{D}_C} + 2\left\langle \mathbf{Xa}^*_k \middle| \mathbf{Yb}^*_k \right\rangle_{\mathbf{D}_C} = 1+1+2\cos\left(\mathbf{Xa}^*_k, \mathbf{Yb}^*_k\right) = 2+2\sqrt{\lambda_k}$

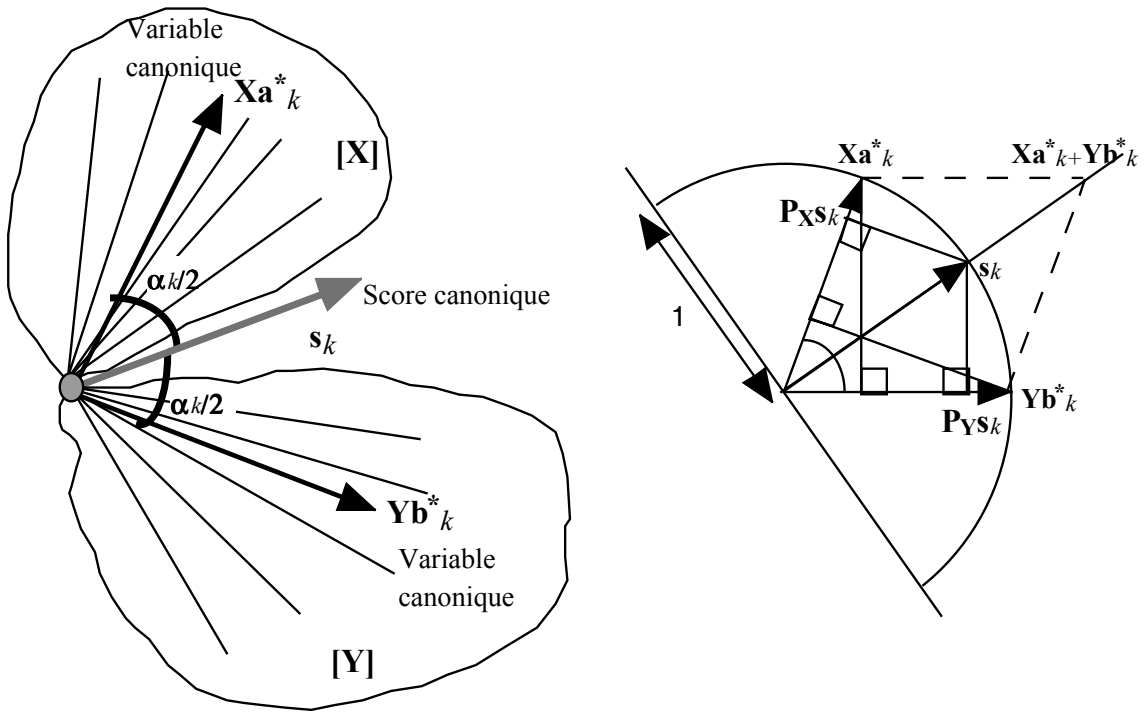


Figure 15 : Géométrie de l'analyse canonique des corrélations.

Les lignes et les colonnes de **N** sont alors représentées par projection du score canonique sur les deux sous-espaces. Dans le cas de l'AFC, les deux sous-espaces sont engendrés par des tableaux disjonctifs complets et cette opération de projection correspond à un calcul de moyennes. Les lignes et colonnes de **N** sont alors placées au centre de gravité de leurs correspondances. De plus, il est alors possible de calculer pour chaque ligne et/ou colonne la variance du score de ses correspondances. En écologie, cette interprétation a été utilisée sur des tableaux sites-espèces (Thioulouse & Chessel 1992) afin de représenter sur le même plan factoriel la tolérance des espèces (variances du score des correspondances par colonne) et la diversité des sites (variances du score des correspondances par lignes). La notion de positions duales (« *Reciprocal averaging* ») est alors étendue à celle de variabilités duales (« *Reciprocal scaling* »).

Dans le cas de l'ACC, le score canonique sera projeté sur les indicatrices des colonnes (**X**) et chaque colonne sera donc à la moyenne de ses correspondances (figure 16a). La projection sur le tableau de variables dupliquées (**Z**) fournit un score des lignes par régression multiple sur ces variables externes. Il peut être également envisageable de projeter le score canonique sur les indicatrices des lignes (**Y**), les lignes seront alors placées au centre de gravité de leurs correspondances. Dans l'ACC classique, on retrouve également deux types de scores connus sous les noms de « LC score » (Linear combination) et « WA score » (Weighted Averaging). Si le « LC score » de l'ACC classique est strictement équivalent à la projection du score canonique sur **Z** ($\mathbf{P_Z s}_k$), le « WA score » correspond à la projection de la variable canonique issue de **X** ($\mathbf{Xa}^*_k$) sur les indicatrices des lignes ($\mathbf{P_Y Xa}^*_k$). D'un point de vue théorique, il est clair que l'utilisation du score obtenu par régression multiple est justifiée car il est obtenu par projection sur le sous-espace qui est analysé (**Z**). On pourra tout de même observer l'adéquation entre ces deux scores afin de déterminer dans quelles mesures les variables externes permettent de prédire les moyennes par ligne (figure 16b). Dans le cas de l'ACC classique, les travaux empiriques de Palmer (1993) conduisent à la même conclusion (utilisation préférentielle des « LC scores ») mais les arguments utilisés par l'auteur apparaissent peu convaincants.

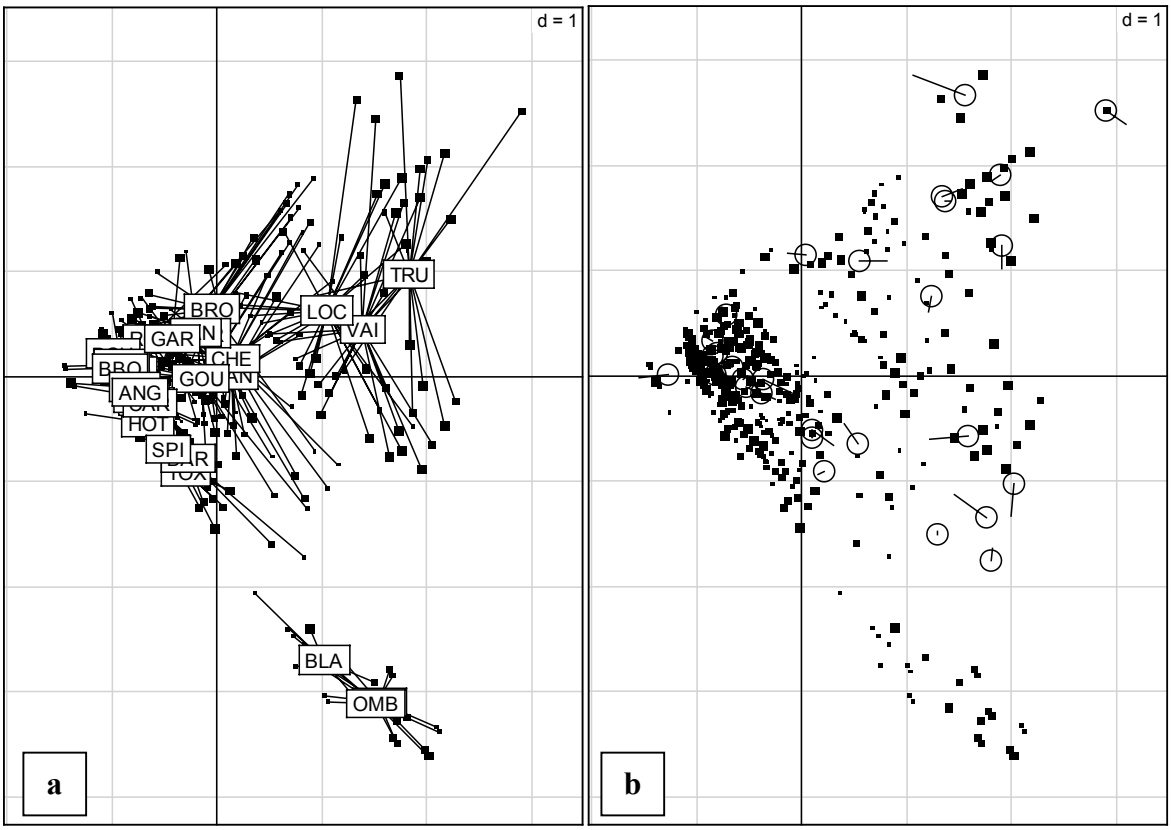


Figure 16 : Analyse canonique des correspondances vue comme analyse canonique des corrélations (données Doubs). (a) Espèces (colonnes) à la moyenne des correspondances. (b) Concordance entre les deux systèmes de représentation des relevés-lignes (WA au centre du cercle et LC à l'extrémité du segment). Chaque correspondance est représentée par un carré dont la taille est proportionnelle à l'abondance (poids).

.

La manipulation des schémas de dualité permet d'identifier facilement les relations existant entre les méthodes. L'analyse factorielle des correspondances, l'analyse canonique des correspondances et la double analyse canonique des correspondances sont des analyses canoniques. Les analyses non symétriques des correspondances sont des analyses en composantes principales sur variables instrumentales. De nouvelles méthodes d'ordination sont envisageables comme par exemple l'analyse non symétrique canonique des correspondances. La version profil ligne de cette méthode de couplage de tableaux définit un triplet $\left(\mathbf{P_E}\mathbf{D}_I^{-1}\mathbf{P}, \mathbf{I}_J, \mathbf{D}_I\right)$ équivalent à celui de l'ACPVI $\left(\mathbf{X}^t\mathbf{D}_C\mathbf{Z}, \left(\mathbf{Z}^t\mathbf{D}_C\mathbf{Z}\right)^{-1}, \mathbf{I}_J\right)$ qui cherche à expliquer **X** par **Z**. Les propriétés d'une telle méthode restent à développer.

Chapitre II : Supports du savoir-faire

L'échantillonnage est souvent la partie la plus fastidieuse et la plus coûteuse de la recherche écologique. Cette étape n'est pas une fin en soi. Les données collectées sont ensuite stockées et analysées. Les résultats des analyses numériques permettent alors d'établir des conclusions, de valider ou non les hypothèses testées, d'en induire de nouvelles, et donc d'influencer les futures problématiques de recherche. Les phénomènes étudiés en écologie sont généralement de nature complexe. Les appréhender requiert souvent la collecte d'un volume important de données répertoriant des mesures de descripteurs qualifiant les objets de l'étude. Les objets auxquels s'intéresse l'écologiste sont les échantillons, parcelles, localités, prélèvements… Les descripteurs utilisés sont de nature très différente : conditions expérimentales (date, lieu…), caractéristiques chimiques, physiques, écologiques ou biologiques… Certains attributs seront considérés comme descripteurs ou comme objets. Par exemple, une espèce peut servir à décrire un site et sera décrite par des traits biologiques. La diversité des descripteurs induit une grande variabilité des formats de données (caractère qualitatif, quantitatif, flou…). De plus, un même descripteur pourra être mesuré de différentes manières : présence, abondance, biomasse, pourcentage de recouvrement… sont des indices permettant de caractériser l'occurrence d'une espèce dans un site. Il en résulte donc des données dont la structure est souvent très complexe. Il apparaît alors nécessaire de posséder des outils logiciels capables de prendre en compte cette complexité pour le stockage et l'analyse des données écologiques. Trois logiciels largement utilisés au cours de ce travail de thèse sont présentés : ADE-4 qui est axé essentiellement sur les méthodes d'analyse multivariée, R qui est un logiciel de statistique généraliste et enfin ArcView qui est le système d'information géographique le plus vendu au monde.

II.1 ADE-4

Jackson (1995) propose un test de permutation basé sur l'analyse procustéenne (Gower 1971) afin de tester la concordance entre deux jeux de données multivariées. Dans l'introduction de cet article l'auteur affirme que « *Procrustean methods are used infrequently in ecology. This lack of use likely reflects the previously limited availability of the procedure* ». D'autres facteurs sont probablement à l'origine du faible intérêt manifesté par les écologistes à l'égard des rotations procustéennes (Dray *et al.* sous presse). Cependant, il est évident que la mise à disposition d'une méthode par l'intermédiaire d'un logiciel facilite grandement sa diffusion et donc son utilisation. Même si la valorisation scientifique associée au développement d'un logiciel peut sembler restreinte (Thioulouse 1996), cette démarche fait partie intégrante de l'activité du biométricien. De plus, de nombreuses revues sont maintenant dédiées à cet aspect de la recherche (Statistics and Computing, Computational Statistics and Data Analysis…) et certaines revues proposent des sections consacrées à l'évaluation de logiciels (*e.g.* Palmer 2002).

ADE-4 (Thioulouse *et al.* 1997) est un logiciel permettant notamment de réaliser des analyses multivariées et des représentations graphiques sur Macintosh ou PC sous Windows. Ce logiciel est gratuit et téléchargeable sur internet à http://pbil.univ-lyon1.fr/ADE-4/ADE-4F.html. Afin de faciliter l'utilisation du logiciel, de nombreuses documentations ont été rédigées et un forum de discussion (ADELIST) permet également de favoriser les échanges entre utilisateurs et développeurs. ADE-4 est formé de plusieurs applications indépendantes

(modules) qui peuvent fonctionner séparément. Chacun de ces modules regroupe un ensemble cohérent d'outils et/ou de méthodes (*e.g.* le module PCA permet de réaliser 10 ACP différentes). Les modules ont été programmés en langage C, ce qui leur confère une grande vitesse d'exécution permettant ainsi l'analyse de grands tableaux de données. Dans le même objectif d'efficacité, le logiciel ADE-4 fonctionne avec des tableaux de données enregistrés en format binaire. Ce choix démontre que ce logiciel est orienté exclusivement vers l'analyse statistique et la représentation graphique au détriment du stockage et de la manipulation des données brutes. Ainsi, avant de réaliser une analyse il est impératif de convertir le tableau de données du format texte au format binaire. Cette étape peut être réalisée par simple « glisser-déposer » à l'aide du module « ADETrans ». Chaque module est utilisable indépendamment mais une interface utilisateur basée sur le logiciel Metacard en facilite l'exécution et permet en outre de naviguer sur les piles contenant des jeux de données, ou des références bibliographiques (figure 17).

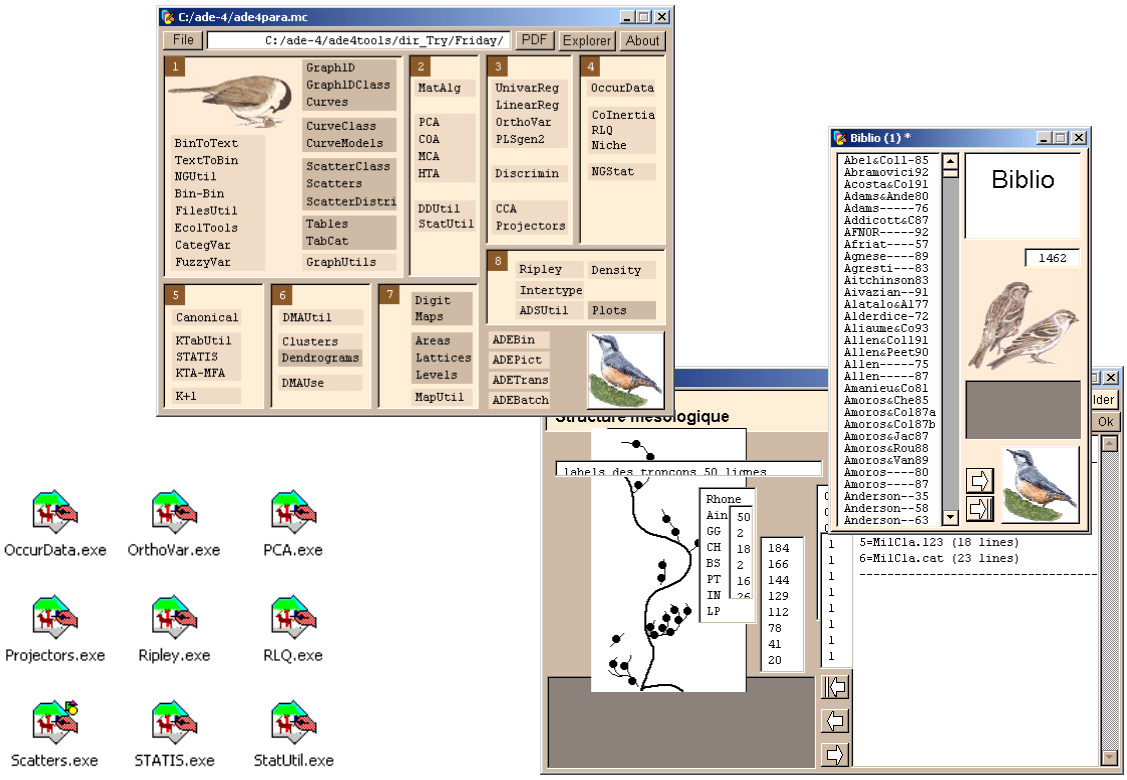


Figure 17 : Le logiciel ADE-4 (version Windows). Copies d'écran de l'interface utilisateur, de la pile de données, de la pile de bibliographie. Les modules d'analyse sont des fichiers exécutables indépendants.

Le lancement des modules peut également se faire de manière automatisée afin d'éviter les fenêtres de dialogue nécessaire lors de l'utilisation classique. Sur PC, sous Windows, cette option permet de lancer des analyses paramétrées à l'aide d'un fichier batch (.bat) qui sera exécuté par l'interpréteur de commandes de Windows (*e.g.* cmd.exe).

L'implémentation des méthodes multivariées dans ADE-4 est basée sur l'utilisation du schéma de dualité. Ainsi, quelle que soit la méthode choisie, l'analyse du tableau *data* produira les fichiers :

data.--ta (tableau modifié)
data.--pl (poids des lignes)
data.--pc (poids des colonnes)
data.--vp (valeurs propres)
data.--li (score des lignes)
data.--co (score des colonnes)

En fonction de l'analyse choisie, des fichiers supplémentaires correspondant à de nouvelles aides à l'interprétation, peuvent être créés. Les deux premiers caractères de l'extension des fichiers de sortie sont définis par le type d'analyse choisie (tableau 1). Par exemple, les poids des lignes provenant de l'AFC de *data* sont contenus dans le fichier « *data.fcpl* ».

Tableau 1 : Exemples d'analyses et extensions correspondantes dans ADE-4

Module	Option	Extension
PCA	Covariance matrix PCA (ACP centrée)	**.cp**
PCA	Correlation matrix PCA (ACP normée)	**.cn**
COA	Correspondence analysis (AFC)	**.fc**
MCA	Multiple correspondence analysis (AFM)	**.cm**

ADE-4 permet naturellement le couplage des deux triplets d'analyse simple appariés par les lignes avec une analyse de co-inertie, une analyse canonique des corrélations ou par une stratégie de variables instrumentales.

Si, par exemple, on veut coupler un tableau faunistique (*Fau*) à un tableau de milieu (*Mil*), la première étape consiste à réaliser deux analyses simples. Le tableau faunistique est traité par une AFC (figure 18) :

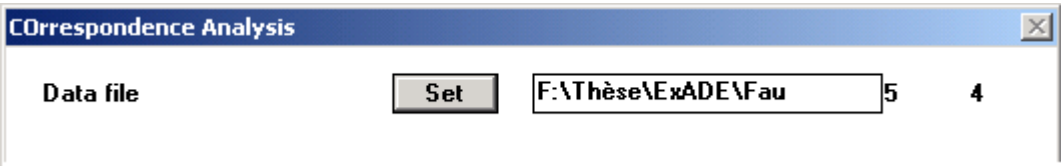


Figure 18 : Copie d'écran (ADE-4, version Windows) du module COA.

Afin de conserver l'appariement des lignes, le tableau de variables mésologiques est traité par une ACP normée avec la pondération des lignes issue de l'AFC (figure 19) :

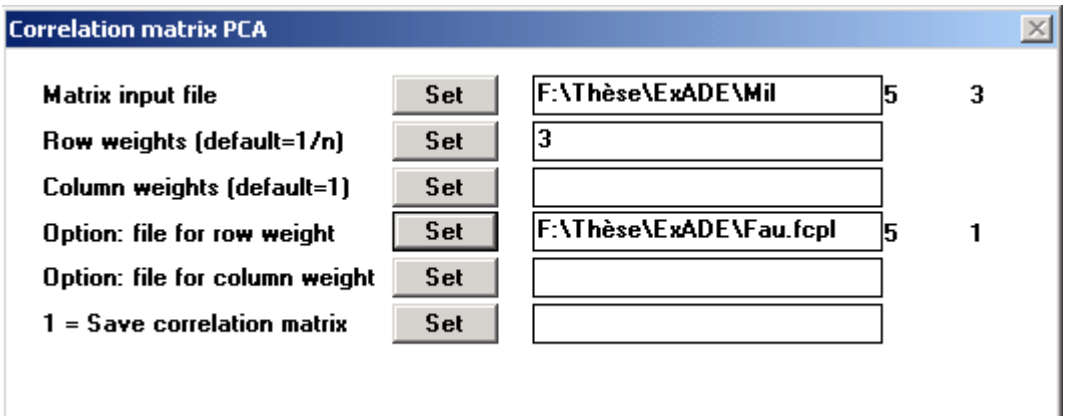


Figure 19 : Copie d'écran (ADE-4, version Windows) du module PCA.

Le couplage de ces deux triplets par une analyse de co-inertie se fait alors très simplement à l'aide du module « CoInertia » (figure 20) :

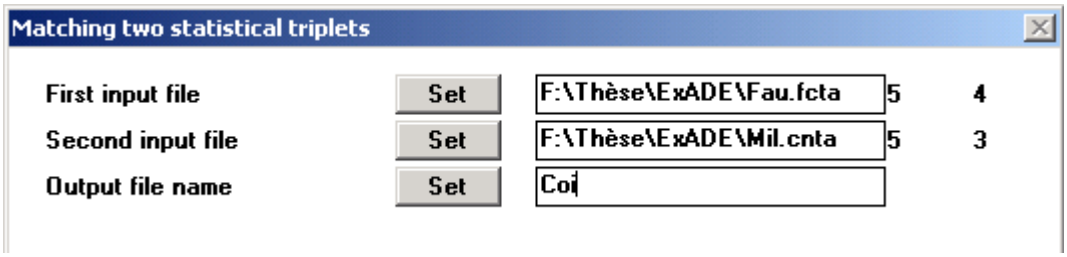


Figure 20 : Copie d'écran (ADE-4, version Windows) du module CoInertia.

On pourra également réaliser une AFCVI en adoptant une stratégie de variables instrumentales en couplant les mêmes triplets à l'aide du module « Projectors ». Comme les méthodes sur variables instrumentales comportent une étape de régression linéaire multiple, il est nécessaire d'obtenir une base orthonormée des variables explicatives afin d'éviter les problèmes de colinéarité (figure 21). Dans ADE-4, cette décorrélation est réalisée grâce à la méthode d'orthonormalisation de Gram-Schmidt (Harville 1997, p. 63-66) :

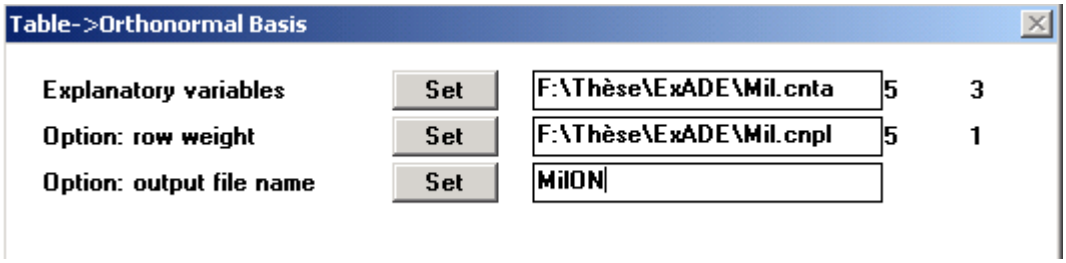


Figure 21 : Copie d'écran (ADE-4, version Windows) du module Projectors.

L'AFCVI peut alors être réalisée (figure 22) :

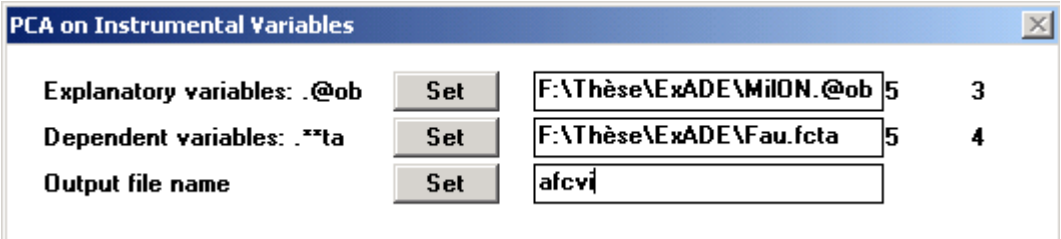


Figure 22 : Copie d'écran (ADE-4, version Windows) du module Projectors.

De la même manière, on peut coupler un triplet à une partition des lignes par une analyse discriminante ou par des analyses inter-intra. Ce type de fonctionnement offre une grande souplesse et de nombreuses options d'analyse à l'utilisateur. Ainsi, certains enchaînements de modules peuvent produire des méthodes complètement originales aux propriétés intéressantes. Cependant, les possibilités d'assemblage des procédures ouvertes dans ADE-4 impliquent une certaine prudence et une connaissance préalable de la théorie de l'analyse de données. En effet, même si un enchaînement de modules est possible du point de vue du logiciel, ceci ne garantit en aucun cas la validité mathématique ou écologique de l'analyse réalisée. Néanmoins, quelques messages recueillis sur le forum ADELIST montrent que pour certains utilisateurs, les problèmes théoriques se posent seulement dans le cas où des difficultés techniques apparaissent lors d'un enchaînement douteux de modules :

Yorick REYJOL (25 Juin 2002) : « *Et là, pb, mon dernier fichier centré-réduit contenant des valeurs négatives, la CoA requise pour la CCA ne fonctionne pas. Comment procéder ?* ». Le problème n'est plus technique. L'AFC est dédiée à l'analyse des tables de contingences qui, par définition, ne contiennent pas de valeurs négatives.

La méthode RLQ (Dolédec *et al.* 1996), de nombreuses méthodes multi-tableaux (*e.g.* Chessel & Hanafi 1996, Lafosse & Hanafi 1997, Simier *et al.* 1999), ainsi que les méthodes d'analyse sous contraintes spatiales (Thioulouse *et al.* 1995) sont également implémentées dans ADE-4. Des modules sont dédiés aux régressions linéaires simples, multiples, orthogonales, polynomiales, lowess (Cleveland 1979, Cleveland & Devlin 1988) et PLS de première et deuxième génération (Tenenhaus 1998).

Par essence, les résultats issus d'une analyse multivariée conduisent à des représentations graphiques. L'analyse permet de synthétiser les données et le graphique fournit alors un moyen d'expression simple et efficace de reproduire l'information résumée. ADE-4 contient de nombreux modules consacrés aux représentations graphiques. La plupart de ces modules permettent un multifenêtrage correspondant à une partition des lignes ou aux différentes variables d'un jeu de données. Différents types de graphiques peuvent être réalisés avec ADE-4. Le module « Graph 1D » est utile pour la représentation de données univariées (figure 23) par courbe de Gauss, histogramme …

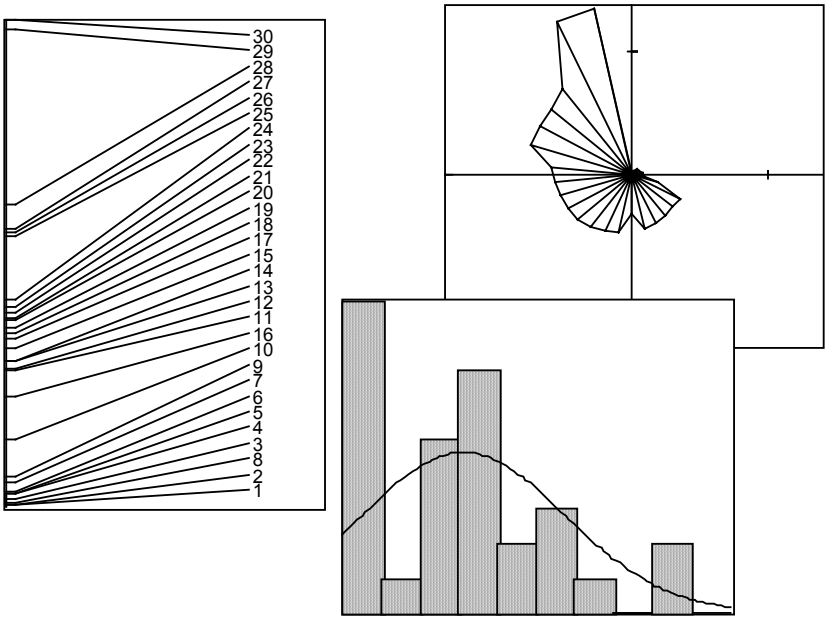


Figure 23 : Exemples de graphiques obtenus par le module Graph1D (ADE-4, version Windows). Représentation unidimensionnelle, graphique en étoile, histogramme et courbe de Gauss.

Le module « Curves » permet de tracer des courbes (i.e. série de valeurs en ordonnée le long d'un axe en abscisse) sous différentes formes (lignes, barres, boîte à moustaches …) (figure 24).

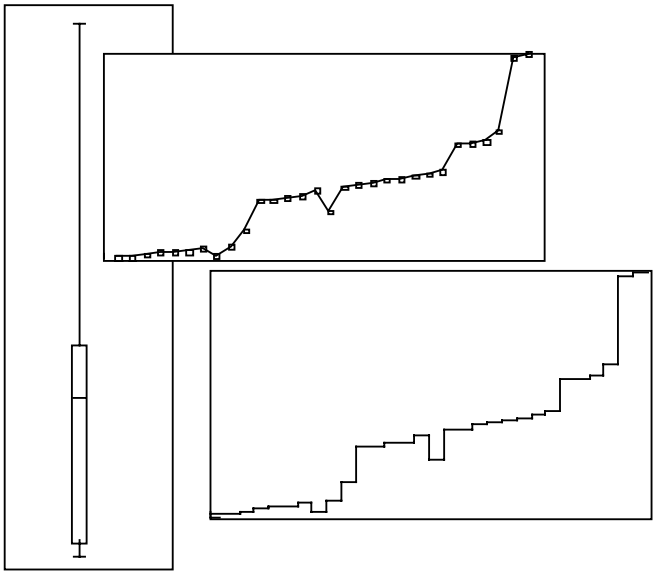


Figure 24 : Exemples de graphiques obtenus par le module Curves (ADE-4, version Windows). Boîte à moustaches, courbe et graphique en escalier.

La carte factorielle est le graphique le plus utilisé pour représenter les résultats d'une analyse multivariée. Le module « Scatters » est destiné à réaliser de telles cartes avec de nombreuses options (figure 25). La représentation la plus simple consiste à indiquer les noms des individus et variables sur le plan factoriel. Il est également possible de tracer une ligne

entre les individus afin de rendre compte d'une éventuelle relation spatiale ou temporelle. L'utilisation de symboles permet de représenter une troisième variable sur un plan factoriel et peut faciliter l'interprétation. La représentation de la concordance entre deux systèmes de coordonnées est également très utile pour les méthodes de couplages.

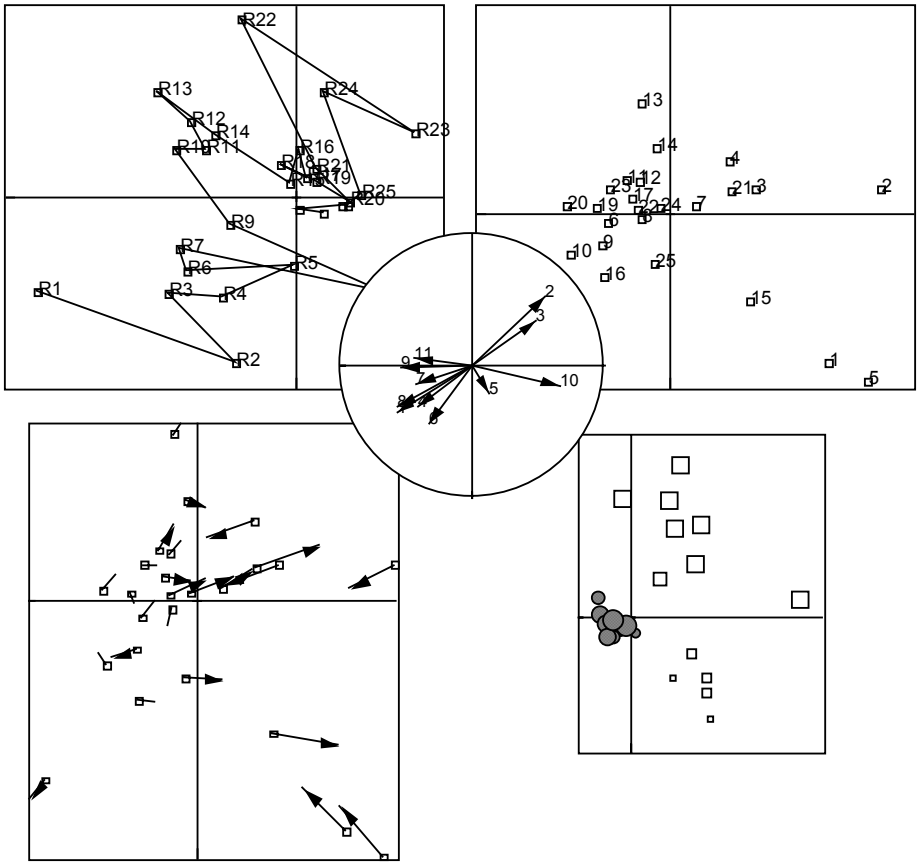


Figure 25 : Exemples de graphiques obtenus par le module Scatters (ADE-4, version Windows). Trajectoires, nuage de points, cercle des corrélations, représentation de la concordance entre deux nuages et représentation d'une troisième variable sur un nuage de points.

Le module « ScatterClass » permet de représenter une partition, sur un plan factoriel, sous la forme de polygones, d'étoiles ou d'ellipses (figure 26). Les ellipses sont alors représentées à l'aide des moyennes, variances et covariances intra-groupes.

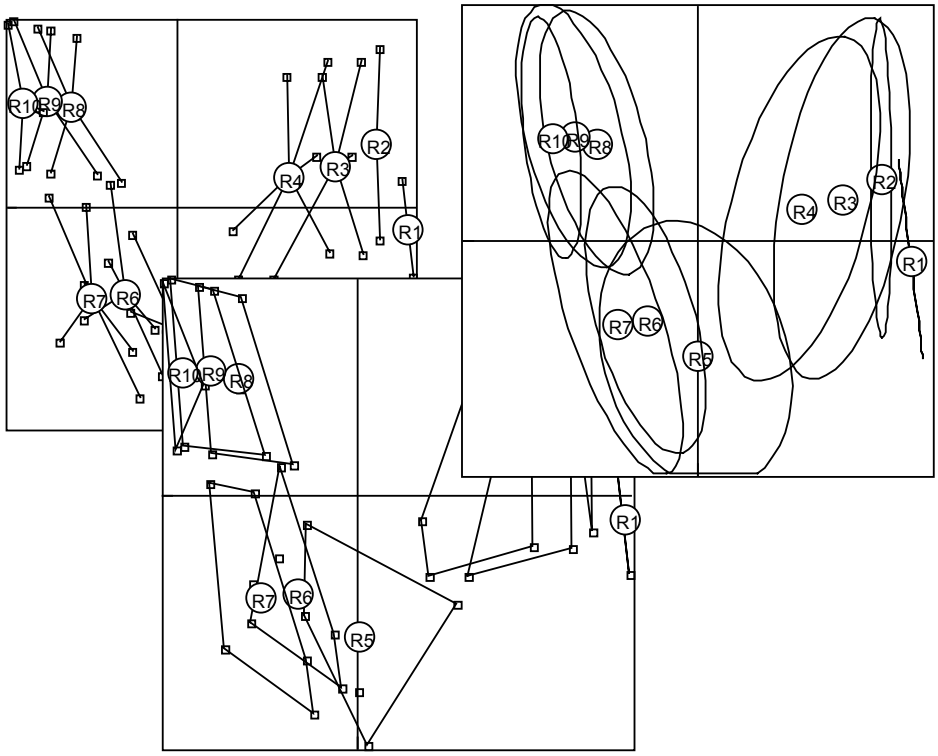


Figure 26 : Exemples de graphiques obtenus par le module ScatterClass (ADE-4, version Windows). Graphiques en étoiles, ellipses et polygones convexes.

Enfin, quelques modules fournissent des fonctions élémentaires de cartographie afin de représenter les résultats d'une analyse ou les données brutes dans l'espace géographique (figure 27). La cartographie de relations de voisinages, la réalisation de courbes de niveaux ou de choroplèthes (Jenks & Caspall 1971) en sont les fonctions les plus significatives.

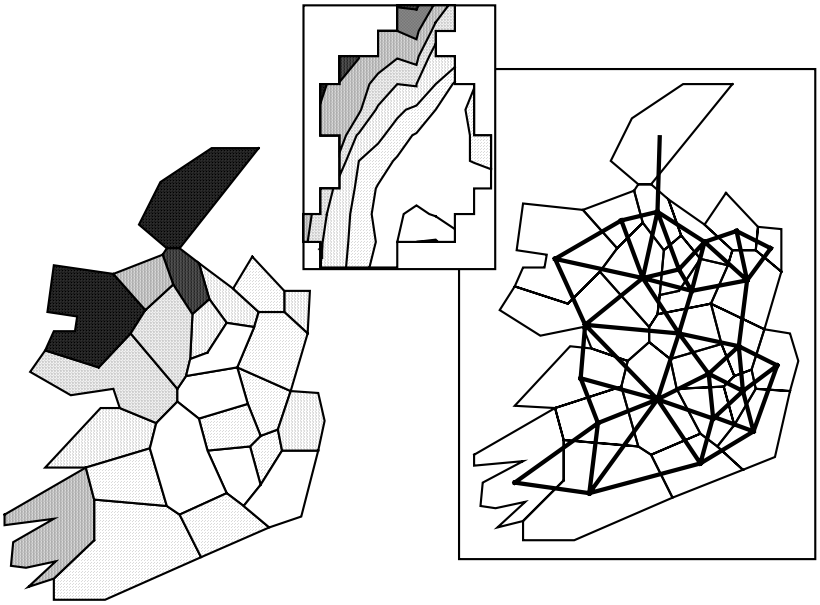


Figure 27 : Exemples de graphiques obtenus par les modules Areas, Levels et Maps (ADE-4, version Windows). Chloroplèthe, courbes de niveaux et relation de voisinage.

II.2. R

R (Ihaka & Gentleman 1996) est un langage et un environnement pour les calculs statistiques et représentations graphiques. **R** est un logiciel libre (GNU) et l'accès à son code source est donc libre. Le langage **R** peut être considéré comme un dialecte du langage **S** commercialisé sous la forme du logiciel S-PLUS par la société Insightful Corporation. Il existe quelques différences entre la conception de **R** et celle de **S** (Hornik 2002) sur lesquelles nous ne nous attarderons pas dans ce document. La plupart des fonctions visibles pour l'utilisateur dans **R** sont écrites en **R**. Des liens avec des procédures écrites en C, C++ ou Fortran sont également possibles. Le système de logiciel libre permet un développement rapide et fiable de **R** grâce aux nombreux échanges entre développeurs et utilisateurs du monde entier. De plus, chacun peut développer et soumettre des modules supplémentaires (librairies) contenant des fonctions spécifiques et qui pourront s'ajouter aux fonctions de base du logiciel. Analyses multivariées (librairie *mva*), bootstrap (*boot*), classification (*cluster*), géostatistiques (*geoR*)… sont ainsi disponibles dans le logiciel **R**.

R est un langage orienté-objet, c'est-à-dire que les variables, données, matrices, fonctions… sont stockées dans la mémoire vive de l'ordinateur sous forme d'objets qui ont un nom. Il suffit alors de taper le nom de l'objet pour avoir son contenu. Ainsi, lorsque d'autres logiciels statistiques affichent une masse importante d'information à la suite d'une analyse, **R** se contente de les stocker dans un objet. L'utilisateur peut donc facilement sélectionner et afficher l'information qui l'intéresse. Il est alors très simple de comparer différents modèles ou de combiner plusieurs fonctions statistiques pour réaliser des analyses plus complexes. Pour attribuer un nom à un objet, il suffit d'utiliser le signe « <- » :

```
> n<-25 # création
> n #affichage
[1] 25
```

L'importation, la gestion et la manipulation des données sont très performantes avec **R**. Le format le plus approprié pour stocker un tableau de données est le `data.frame`. On peut créer ce type de tableau directement dans **R** ou en important un fichier texte par exemple :

```
> milieu_read.table("DoubsMil.txt",sep="\t")
> milieu[1:5,]
  DtS Alt   Slp Flw pH Har Pho Nit Amm Oxy Bod
1   3 934 6.176  84 79  45   1  20   0 122  27
2  22 932 3.434 100 80  40   2  20  10 103  19
3 102 914 3.638 180 83  52   5  22   5 105  35
4 185 854 3.497 253 80  72  10  21   0 110  13
5 215 849 3.178 264 81  84  38  52  20  80  62
```

Les bases de données sont de plus en plus répandues à l'heure actuelle. Elles permettent de stocker une masse importante de données d'une manière ordonnée et logique. Il existe plusieurs librairies permettant une interface de **R** avec un système de gestion de base de données (SGBD). La librairie fournissant la plus grande souplesse est certainement *RODBC*. Ce module est basé sur la technologie ODBC (*Open Database Connectivity*), développée par Microsoft, qui permet la communication entre des clients bases de données fonctionnant sous Windows et les SGBD du marché. La technologie ODBC permet d'interfacer de façon standard une application à n'importe quel serveur de bases de données, pour peu que celui-ci possède un driver ODBC (la quasi-totalité des SGBD possèdent un tel pilote, dont tous les

principaux SGBD du marché). Il est ainsi possible d'importer directement des tables contenues dans un classeur Excel :

```
> library(RODBC) #chargement de la librairie
> connection<-odbcConnect("Excel Files",case="toupper") #
ouverture de la connexion
> sqlTables(connection) # liste des tables
         TABLE_CAT TABLE_SCHEM TABLE_NAME   TABLE_TYPE REMARKS
1 F:\\Thèse\\R\\Doubs          NA     Faune$ SYSTEM TABLE      NA
2 F:\\Thèse\\R\\Doubs          NA    Milieu$ SYSTEM TABLE      NA
> milieu<-sqlQuery(connection,"select * from [milieu$]")
> milieu[1:5,]
  DtS Alt   Slp Flw pH Har Pho Nit Amm Oxy Bod
1   3 934 6.176  84 79  45   1  20   0 122  27
2  22 932 3.434 100 80  40   2  20  10 103  19
3 102 914 3.638 180 83  52   5  22   5 105  35
4 185 854 3.497 253 80  72  10  21   0 110  13
5 215 849 3.178 264 81  84  38  52  20  80  62
```

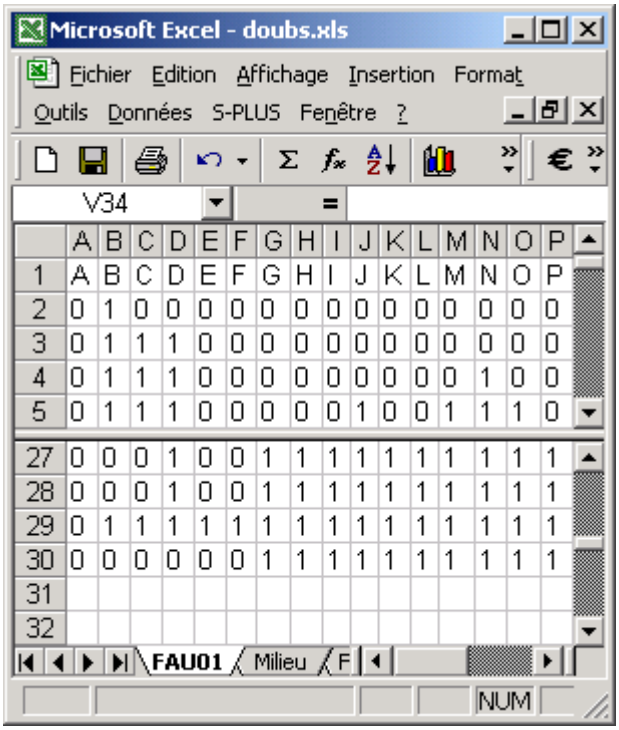
Il est également possible de réaliser dans **R** des requêtes SQL sur la base de données connectée :

```
> Fau1<-sqlQuery(connection,"select * from [Faune$] where a>0
order by b") # selection des sites pour lesquels l'espèce a est
présente et triés en fonction de l'abondance de l'espèce b
> Fau1
  a b c d e f g h i j k l m n o p q r s t u v w x y z zz
1 1 1 3 3 1 1 1 3 2 3 3 3 2 1 3 2 1 0 1 1 0 0 1 2 0 2 1
2 1 2 4 4 1 2 1 4 3 2 3 4 1 1 2 1 1 0 1 1 0 0 0 2 0 2 1
3 2 3 3 5 0 5 0 4 5 2 2 1 2 1 1 0 1 0 1 1 0 0 0 1 0 0 0
4 1 3 4 1 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
5 3 4 4 5 2 4 0 0 3 3 2 0 2 0 0 0 0 0 0 1 0 0 0 0 0 0 0
6 3 5 5 4 4 3 0 0 0 1 1 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0
7 2 5 5 2 3 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
8 2 5 4 4 2 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

De la même manière, un tableau de données créé avec **R** peut être exporté dans la base de données connectée :

```
> Fau<-sqlQuery(connection,"select * from [Faune$]")
> Fau01<-as.data.frame(ifelse(Fau>0,1,0)) # passage en présence
absence
> Fau01[1:5,]
  a b c d e f g h i j k l m n o p q r s t u v w x y z zz
1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
2 0 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
3 0 1 1 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0
4 0 1 1 1 0 0 0 0 0 1 0 0 1 1 1 0 0 0 0 1 0 0 0 0 0 0 0
5 0 1 1 1 0 0 0 0 1 1 0 0 1 1 1 0 0 1 0 1 0 0 0 1 0 0 0
> sqlSave(connection,Fau01) # sauvegarde de la table dans le
classeur Excel
```

Une feuille s'est ajoutée dans le classeur Excel (figure 28) :



Microsoft Excel - doubs.xls

Fichier Edition Affichage Insertion Format Outils Données S-PLUS Fenêtre ?

V34 =

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
4	0	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0
5	0	1	1	1	0	0	0	0	0	1	0	0	1	1	1	0
27	0	0	0	1	0	0	1	1	1	1	1	1	1	1	1	1
28	0	0	0	1	0	0	1	1	1	1	1	1	1	1	1	1
29	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
30	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
31																
32																

FAU01 Milieu F

NUM

Figure 28 : Copie d'écran du logiciel Excel. La Feuille FAU01 a été rajoutée grâce à la fonction SqlSave de la librairie RODBC de R.

R possède également de nombreuses fonctions permettant notamment de faire du calcul matriciel. Ces fonctions sont très efficaces et permettent d'analyser de grands tableaux de données. La multiplication de deux matrices de dimension 500 par 500 contenant 250000 nombres prend moins de 5 secondes sur PC (Pentium III) :

```
> mat1_matrix(rnorm(250000),500,500) # création de la matrice 1
> mat2_matrix(rnorm(250000),500,500) # création de la matrice 2
> system.time(mat1%*%mat2) # temps de calcul de la multiplication
[1] 4.59 0.00 4.62   NA   NA
```

De même, la décomposition en valeurs singulières d'une matrice de dimension 500 par 500 prend moins de 20 secondes :

```
> system.time(svd(mat1))
[1] 18.92  0.00 19.18    NA    NA
```

De nombreuses fonctions statistiques de base sont également disponibles (tests, régression linéaire, analyse de la variance, modèle linéaire généralisé…).

```
# Exemple de Test t (comparaison de moyennes)
> y1_c(120,107,110,116,114,111,113,117,114,112)
> y2_c(110,111,107,108,110,105,107,106,111,111)
> t.test(y1,y2,var.eq=T)

        Two Sample t-test

data:  y1 and y2
t = 3.4843, df = 18, p-value = 0.002647
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
```

```
 1.905773 7.694227
sample estimates:
mean of x mean of y
    113.4     108.6

# Exemple d'ANOVA :
# Source : Dagnélie, P. (1981) Théorie et méthodes statistiques.
Exercices. Les Presses Agronomiques de Gembloux, Gembloux, 186 p.
>
gain_c(37.7,44.6,42.1,45.1,43.2,45.2,54.2,38.1,48.3,55.1,48.3,44.1
,56.9,42.2,54.0)
> ali_as.factor(rep(c("t1","t2","t3"),c(5,5,5)))
> gain
 [1] 37.7 44.6 42.1 45.1 43.2 45.2 54.2 38.1 48.3 55.1 48.3 44.1
56.9 42.2 54.0
> ali
 [1] t1 t1 t1 t1 t1 t2 t2 t2 t2 t2 t3 t3 t3 t3 t3
Levels:  t1 t2 t3
> anova(lm(gain~ali))
Analysis of Variance Table

Response: gain
          Df Sum Sq Mean Sq F value Pr(>F)
ali        2 126.15   63.07  1.9529 0.1844
Residuals 12 387.58   32.30
```

Enfin, **R** possède un grand nombre d'outils graphiques particulièrement flexibles. La librairie *lattice* fournit notamment la possibilité de réaliser des graphiques dits « *trellis* ». Le principe de ce type de graphiques est basé sur la décomposition d'un graphique simple (nuage de points, courbe, histogramme…) par rapport aux valeurs prises par des variables externes. Ce type de graphiques a été créé par Cleveland (1993) dans le cas de deux variables externes puis étendu aux situations plus complexes. Les « *trellis* » sont très utiles pour observer la structure des données et notamment pour identifier les phénomènes d'interaction entre variables explicatives et une variable réponse (figure 29).

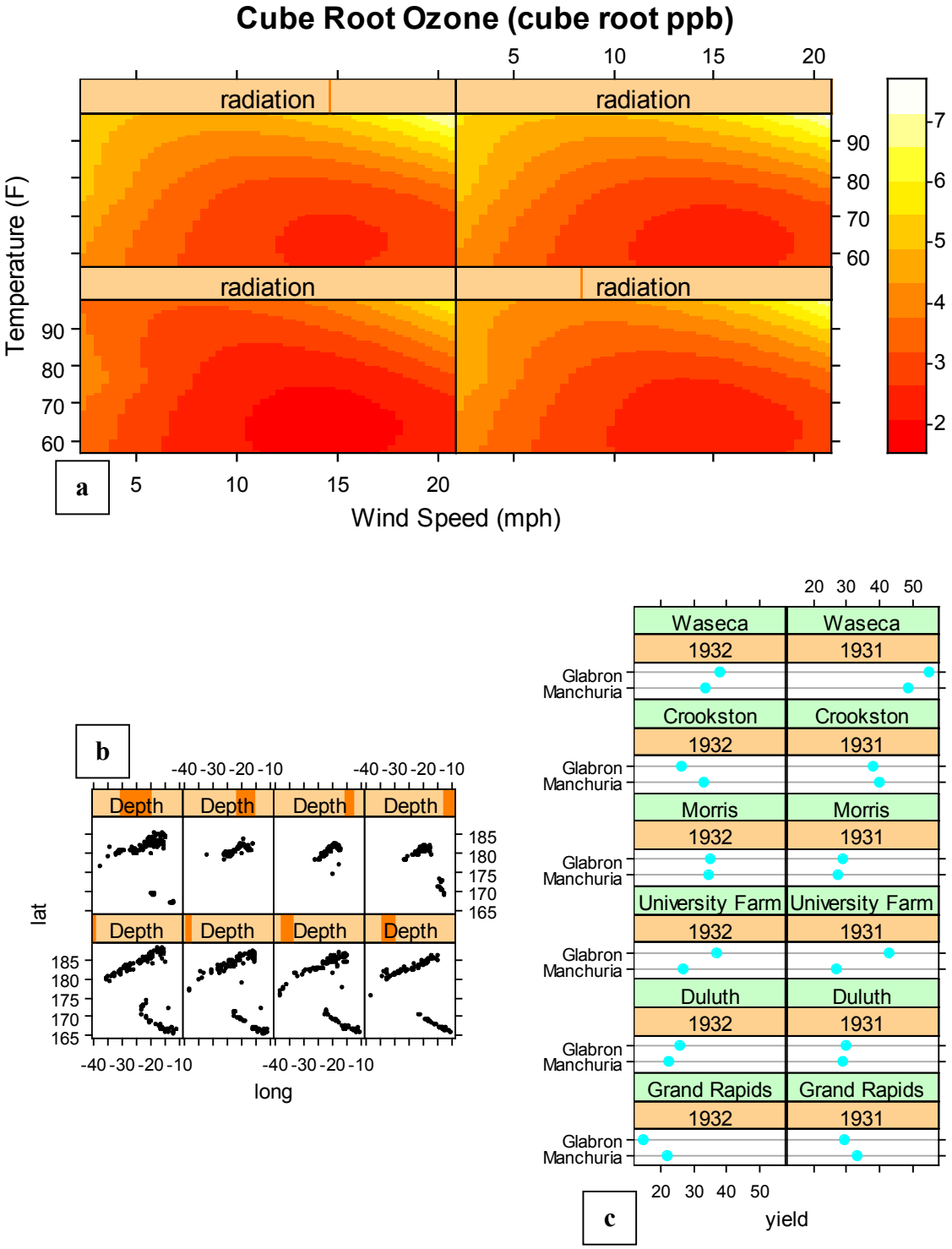


Figure 29 : Graphiques « *trellis* » réalisés dans R. Les données sont contenues dans la librairie lattice. (a) Concentration d'ozone en fonction de la vitesse du vent, de la température et des radiations solaires. (b) Localisation d'événements sismiques par tranche de profondeur. (c) Rendements de deux variétés d'orge en fonction de l'année et du site.

Dans le cadre de son développement, une librairie a été développée afin d'inclure la plupart des méthodes disponibles dans le logiciel ADE-4 dans **R**. Même s'il avait été possible d'appeler dans **R** le code source d'ADE-4 programmé en langage C, le choix a été fait de reprogrammer l'ensemble des méthodes directement dans **R** favorisant ainsi la lisibilité des fonctions. La programmation de l'ensemble des méthodes est basée sur la théorie du schéma de dualité. Chaque méthode d'analyse multivariée fait appel à la fonction *as.dudi* qui accepte comme argument un triplet défini par un tableau et deux métriques ou pondération et qui renvoie un objet de la classe *dudi* (duality diagram) :

```
> args(as.dudi)
function (df, col.w, row.w, scannf, nf, call, type, tol = 1e-07,
full = FALSE)
```

Un objet de la classe *dudi* est une liste contenant les composants :

$tab	data frame à *n* lignes et *p* colonnes
$cw	poids des lignes, vecteur à *n* composantes
$lw	poids des colonnes, vecteur à *p* composantes
$eig	valeurs propres, vecteur à *min*(*n*,*p*) composantes
$nf	entier, nombre de facteurs conservés
$c1	axes principaux, data frame à *p* lignes et *nf* colonnes
$l1	composantes principales, data frame à *n* lignes et *nf* colonnes
$co	coordonnées des colonnes, data frame à *p* lignes et *nf* colonnes
$li	coordonnées des lignes, data frame à *n* lignes et *nf* colonnes
$call	contient l'origine de l'appel

En fonction de la méthode choisie, d'autres composants peuvent être ajoutés à cette liste. Des fonctions génériques sont associées à la classe *dudi* afin d'afficher les résultats (*print*), de réaliser un graphique (*plot*), de relancer une analyse (*redo*)…

Les qualités inhérentes à **R** relatives à la gestion des objets, aux performances de calculs ou aux outils graphiques font de la librairie ade4 un outil simple et très efficace pour l'analyse de données. Quelques lignes de commandes suffisent pour réaliser une analyse de co-inertie et les représentations graphiques associées :

```
> library(ade4) #chargement de la librairie
> data(doubs) # chargement des données
> dudi1_dudi.pca(doubs$mil,scale=T,scan=F,nf=3) # acp normée
> dudi2_dudi.pca(doubs$poi,scale=F,scan=F,nf=2) # acp centrée
> coin1_coinertia(dudi1,dudi2,scan=F,nf=2) # analyse de co-inertie


> print(coin1) # affichage relatif à l'objet coin1
Coinertia analysis
call: coinertia(dudiX = dudi1, dudiY = dudi2, scannf = F, nf = 2)
class: coinertia dudi

$rank (rank)      : 11
$nf (axis saved)  : 2
$RV (RV coeff)    : 0.4505569

eigen values: 119 13.87 0.7566 0.5278 0.2709 ...

  vector length mode    content
1 $eig   11     numeric eigen values
2 $lw    27     numeric row weigths (crossed array)
3 $cw    11     numeric col weigths (crossed array)
```

```
   data.frame nrow ncol content
1  $tab       27   11   crossed array (CA)
2  $li        27   2    Y col = CA row: coordinates
3  $l1        27   2    Y col = CA row: normed scores
4  $co        11   2    X col = CA column: coordinates
5  $c1        11   2    X col = CA column: normed scores
6  $lX        30   2    row coordinates (X)
7  $mX        30   2    normed row scores (X)
8  $lY        30   2    row coordinates (Y)
9  $mY        30   2    normed row scores (Y)
10 $aX        3    2    axis onto co-inertia axis (X)
11 $aY        3    2    axis onto co-inertia axis (Y)


> summary(coin1) # résumé de l'objet coin1

Eigenvalues decomposition:
        eig     covar      sdX      sdY      corr
1 119.01942 10.909602 2.326324 6.422570 0.7301798
2  13.87137  3.724429 1.685078 2.863743 0.7718017

Inertia & coinertia X:
    inertia      max     ratio
1  5.411785 6.321624 0.8560752
12 8.251272 8.553220 0.9646978

Inertia & coinertia Y:
    inertia      max     ratio
1  41.24940 42.74627 0.9649824
12 49.45042 50.90461 0.9714331

RV:
 0.4505569


> plot(coin1) # représentation graphique de l'objet coin1
```

Les résultats d'une analyse de co-inertie sont facilement accessibles et représentés (figure 30).

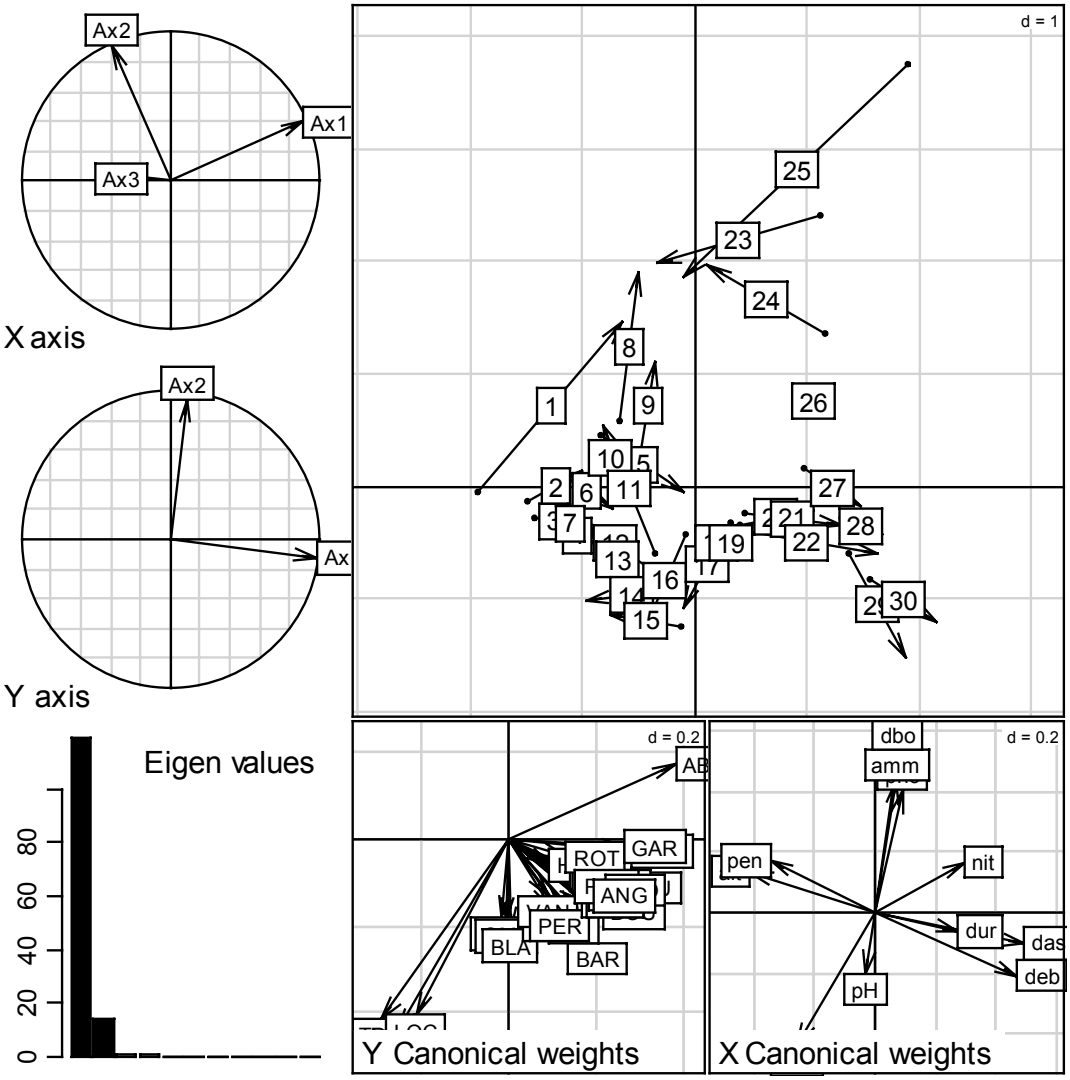


Figure 30 : Représentation graphique des résultats d'une analyse de co-inertie (données Doubs). Projection des axes principaux dans le plan formé par les axes de co-inertie, valeurs propres, double représentation des sites et projections des variables dans le plan formé par les axes de co-inertie.

II.3. ArcView

ArcView est un système d'information géographique (SIG) commercialisé par la société ESRI. Un SIG peut être défini comme un système informatisé (matériels et logiciels) capable de stocker, gérer, manipuler, analyser, modéliser, représenter des données à références spatiales. Pour Didier (1991), un SIG est un « *ensemble de données repérées dans l'espace, structuré de façon à pouvoir en extraire commodément des synthèses utiles à la décision* ». Ces deux définitions se complètent en insistant sur les caractéristiques techniques d'un SIG et sur la finalité d'un tel système. Un SIG est un SGBD prenant en compte l'information spatiale. Un objet est alors défini par ses caractéristiques spatiales (où ?) et par ses attributs (quoi ?) permettant ainsi d'obtenir une base de données spatialisées (figure 31). Selon Poidevin (1999), près de 85 % des bases de données contiennent un composant géographique associé à un lieu précis. Cette masse « astronomique » de données spatialisées a entraîné un

très fort développement de l'utilisation des SIG dans de nombreux domaines d'activités tels que l'environnement, l'agriculture, le marketing, les transports ou la gestion du territoire…

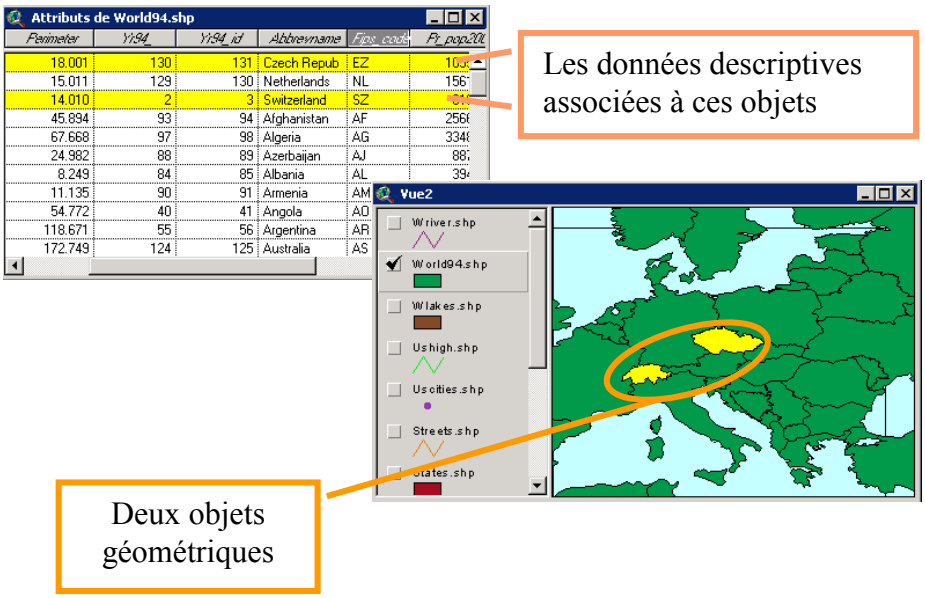


Figure 31 : Base de données spatialisées d'un SIG. La sélection de deux entités géographiques entraîne la sélection des données relatives à ces objets (et inversement).

Afin d'étudier un phénomène, l'ensemble de l'information relative aux objets contenus dans une zone géographique définie est stocké sous forme de couche ou thème. Généralement, chaque thème correspond à un type de données et l'assemblage des différents thèmes permet alors d'obtenir un modèle du monde réel (figure 32).

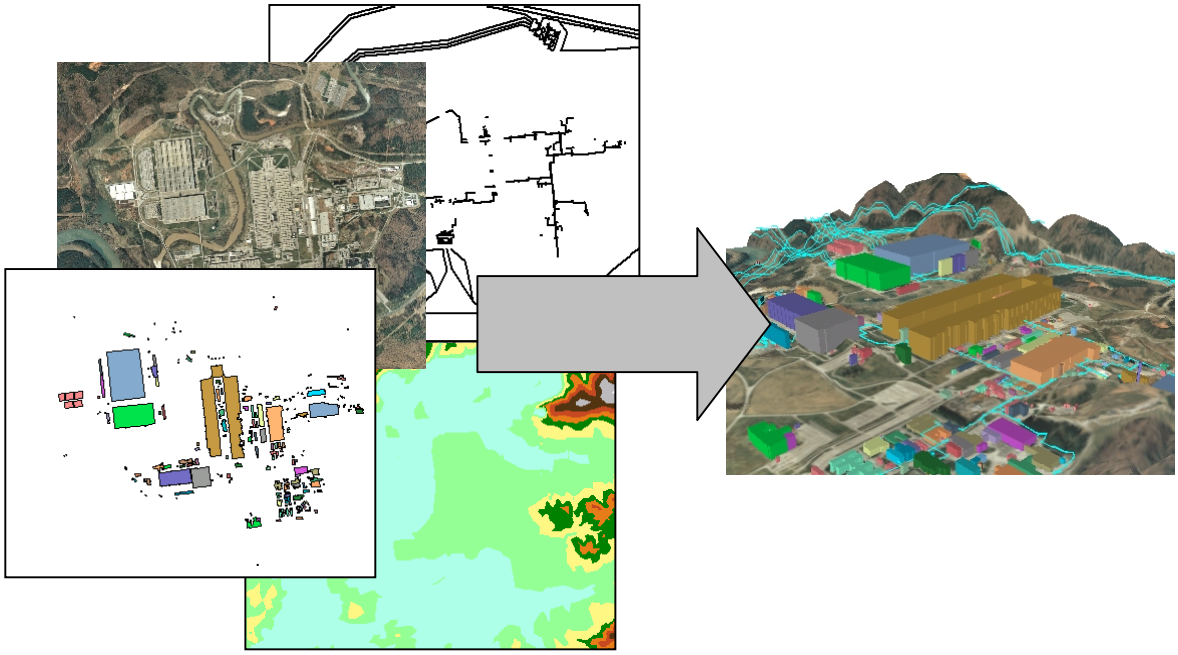


Figure 32 : Ensemble des couches d'informations d'un SIG représentant les différents objets et phénomènes du monde réel.

Il existe deux grands modes de représentation des données dans un SIG (figure 33). Le mode vecteur est un système de représentation orienté objet. Les attributs concernent des objets dont les coordonnées spatiales sont définies précisément. Ces objets sont des points, des lignes, des lignes brisées, des courbes, des polygones, des cercles… Le mode vecteur est adapté à la représentation de points de capture, parcelles, cours d'eau, lacs, routes… pour lesquels un grand nombre d'attributs peut être identifié (*e.g.* nom, date de capture, nombre d'individus…). La deuxième stratégie de représentation est orientée image et appelée mode raster. La zone d'étude est alors divisée à l'aide d'une grille de cellules rectangulaires ou carrées. Chaque couche contient alors un seul type d'information et une valeur est attribuée à chaque cellule. Le mode raster sert à représenter de l'information spatiale continue (*e.g.* altitude, température, image satellite…).

Figure 33 : Modes de représentation vecteur et raster.

L'acquisition des données dans ArcView peut se faire de différentes façons. De nombreuses données sont disponibles au format d'ArcView. Certaines d'entre elles sont gratuites mais elles sont en général d'une faible précision. En revanche, un grand nombre d'éditeurs (*e.g.* IGN, BRGM…) commercialise des cartes thématiques de grande qualité et dont les coûts ont tendance à diminuer. Il est également possible de générer de nouvelles données directement dans ArcView en important un fichier texte contenant les coordonnées géographiques des objets. Une deuxième stratégie est également envisageable car ArcView propose des outils pour tracer des points, des lignes et des polygones… Ainsi, lorsque l'on dispose d'une carte papier scannée que l'on désire numériser, il suffit d'importer le fond de carte et de s'appuyer dessus pour créer les objets correspondants dans ArcView. Certains logiciels permettent de réaliser ce type de numérisation automatiquement. Le logiciel MapScan développé par les Nation-Unies est gratuit et disponible sur internet à l'adresse http://unsd.ics.trieste.it/software/mapscan.htm. Il permet notamment de transformer un fichier image (bitmap, TIFF…) en un thème au format ArcView (*vectorisation*) sans aucune intervention de l'utilisateur. Il peut être alors nécessaire de modifier quelques erreurs ou de géoréférencer les données, c'est-à-dire transformer le thème créé afin de le rendre compatible avec un système de coordonnées géographique usuel.

La cartographie est une des fonctions primordiales d'un SIG. En fonction du type de données à représenter (variables qualitatives ou quantitatives) et des objets de l'étude (points, lignes, polygones), de nombreuses possibilités de représentation sont disponibles dans ArcView (variation de la forme, de la taille ou de la couleur des symboles). Les cartes réalisées peuvent contenir une légende, une échelle, une flèche nord ou du texte et être exportées sous forme de graphiques dans un traitement de texte. Il est évident qu'une carte est

un moyen simple et très efficace d'exprimer les résultats d'une analyse à caractère spatial. Goodall (1954) représentait déjà, sous forme de courbes de niveaux, des coordonnées factorielles dans l'espace géographique afin de mettre en évidence la structure spatiale des phénomènes écologiques étudiés. La souplesse d'utilisation d'un SIG comme ArcView autorise la création de cartes dont la taille, la zone d'étude et l'information symbolisée sont en adéquation parfaite avec les besoins de l'utilisateur et permet donc d'améliorer sensiblement la représentation des résultats d'une étude (figure 34).

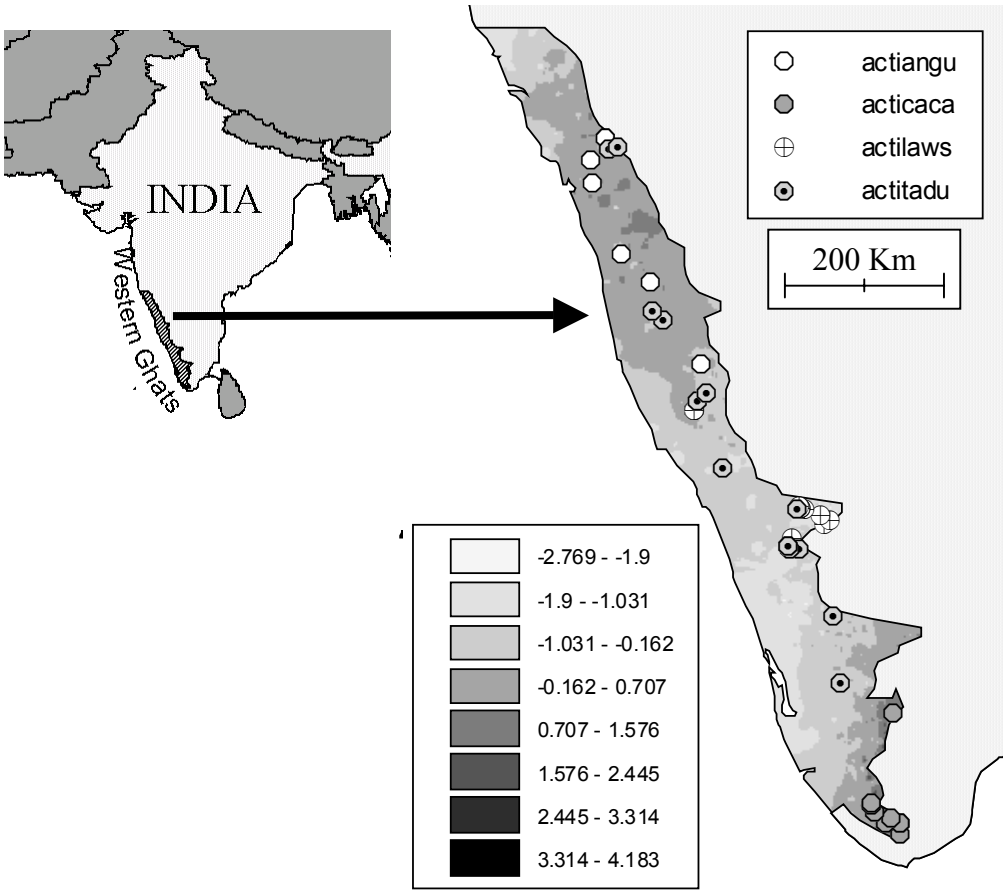


Figure 34 : Cartographie des spécimens d'herbiers de quatre espèces dans les Ghâts occidentaux.

ArcView possède également de nombreuses fonctions de base d'un SGBD (figure 35). Il est ainsi possible de sélectionner des individus à l'aide de requêtes SQL, de joindre deux tables ayant un champ (colonne) en commun ou d'obtenir des statistiques usuelles. De plus, ArcView offre la possibilité de se connecter à des bases de données externes par l'intermédiaire de la technologie ODBC.

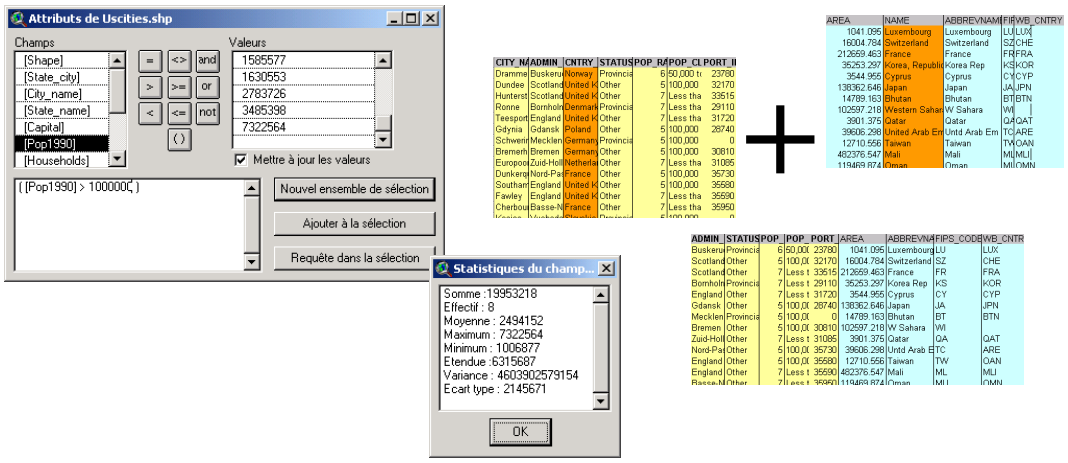


Figure 35 : Exemples de fonctions de SGBD d'ArcView. Il est possible de réaliser des requêtes SQL, des résumés à l'aide de statistiques et de joindre des tables par une information commune.

L'originalité d'un SIG par rapport aux SGBD classiques est la prise en compte de l'information spatiale. Il est donc possible de réaliser des requêtes concernant les relations spatiales qu'entretiennent les objets et de les coupler avec des requêtes classiques : « Sélectionner les arbres contenus dans des parcelles dont le sol est argileux et se situant à moins d'un kilomètre d'un cours d'eau ». De nombreux outils sont également disponibles pour mesurer des distances, aires ou densités de points ; agréger des objets en fonction de la valeur d'un champ ; affecter des données par jointure spatiale ; réaliser l'intersection de deux thèmes ; générer de nouvelles données par le croisement de différents thèmes (figure 36)…

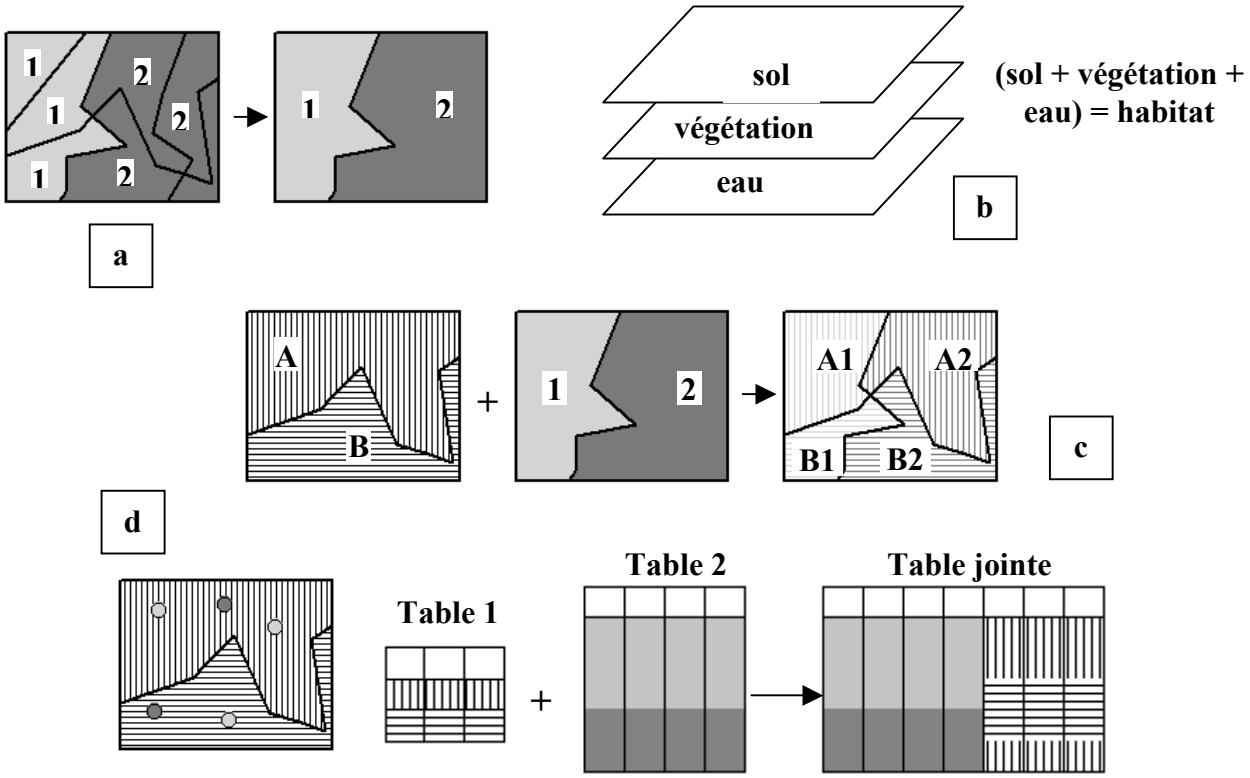


Figure 36 : Exemples de fonctions spatiales disponibles dans un SIG. (a) Agrégation d'entités. (b) Calcul cartographique. (c) Intersection de deux thèmes. (d) Jointure spatiale.

Enfin, une des caractéristiques les plus intéressantes d'ArcView est l'intégration d'un système de programmation dans le logiciel. Les versions 3.x étaient basées sur le langage Avenue développé par ESRI alors que VisualBasic a été adopté dans la version 8. L'intégration d'un outil de développement dans le logiciel permet de créer des fonctions sous la forme de scripts. Cette possibilité repousse les limites d'utilisation d'un tel logiciel en permettant à l'utilisateur de développer ses propres outils pour l'analyse, la gestion ou la création de données. Ces scripts peuvent être exécutés de différentes manières. La plus simple consiste à utiliser la fonction « Exécuter » de l'éditeur de script. Il est également possible de personnaliser ArcView en ajoutant des boutons ou menus auxquels sont liés des scripts. Un simple clic sur le bouton permet alors d'exécuter le script. Enfin, le dernier moyen d'intégrer des scripts dans ArcView est basé sur l'utilisation des extensions. Une extension est un ensemble de scripts compilés ajoutant des fonctionnalités au logiciel. Pour activer l'extension désirée, il suffit de la sélectionner à l'aide du menu « Fichier –> Extensions… » (figure 37) :

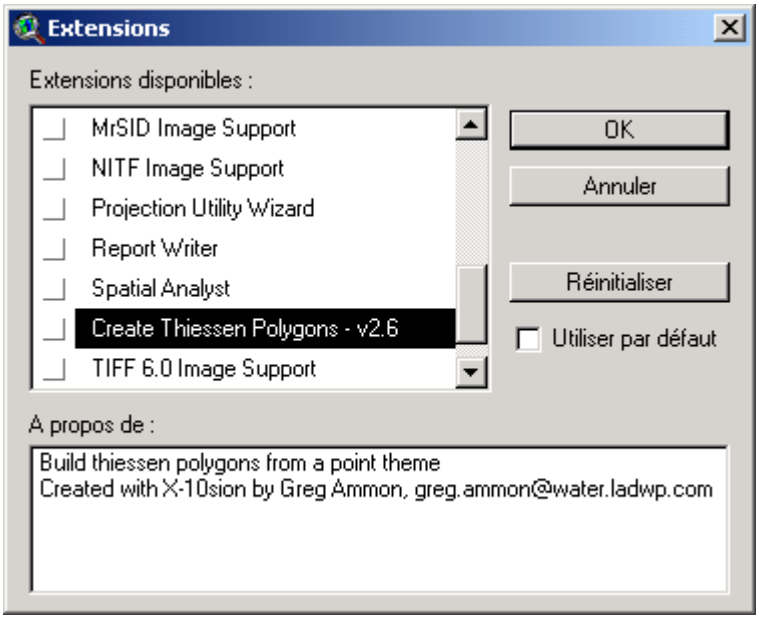


Figure 37 : Copie d'écran illustrant le chargement d'une extension dans ArcView.

Dès lors, des menus ou boutons correspondant à de nouvelles fonctions sont automatiquement ajoutés dans ArcView. Par exemple, l'extension de Greg Ammon ajoute un bouton et permet de réaliser une tesselation à partir d'un thème de points (figure 38) :

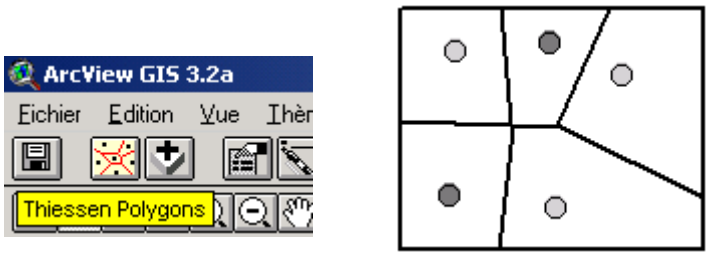


Figure 38 : Tesselation réalisée dans ArcView à l'aide de l'extension « *Create Thiessen Polygons* » de Greg Ammon.

La création d'extension est très simple et réalisée à partir de l'éditeur de scripts. De nombreuses extensions sont commercialisées pour l'analyse de données à trois dimensions (*3D Analyst*), l'analyse de données raster (*Spatial Analyst*), la publication de cartes interactives sur internet (*HTML ImageMapper*)… De plus, l'extension permet aux utilisateurs de développer leurs outils et de les rendre disponibles à la communauté dans un format standard et simple d'utilisation. Une partie du site d'ESRI est dédiée à l'échange de scripts entre utilisateurs fournissant ainsi un ensemble important d'outils gratuits et prêts à l'emploi (figure 39) :

Figure 39 : Rubrique ArcScripts du site d'ESRI permettant le téléchargement de scripts.

Ainsi, même si le nombre d'outils inclus dans ArcView est assez réduit, l'intégration d'un outil de développement rend ce logiciel très évolutif et permet d'accroître sensiblement ses fonctionnalités. De plus, il est possible de programmer, en Avenue, des fonctions destinées à l'interpréteur de commandes de Windows et ainsi de lancer d'autres application à partir d'ArcView. Cette fonctionnalité ouvre donc la possibilité d'interfacer ArcView avec d'autres logiciels.

Chapitre III : La consultation des écologues

La consultation statistique constitue le point de rencontre fondamental entre le biométricien et l'écologue (Chessel 1992). Pour Escoufier (1983), la consultation « *réside dans la prestation de service qui englobe aussi bien la réponse favorable au coup de téléphone « Pouvez-vous réaliser une analyse de la variance pour moi ? », que l'acceptation de contrats visant à la mise en œuvre de méthodes sophistiquées sur des ensembles de données éventuellement très volumineux…* ». Lorsque la problématique est simple et l'intervention du biométricien réduite, l'échange est informel, a lieu dans un couloir ou à la salle café et sera valorisé éventuellement par un remerciement à la fin d'une publication. Dans le cas d'une situation plus complexe, l'implication du biostatisticien sera nécessaire pour mettre au point une méthodologie adaptée. La consultation est alors essentielle pour définir le cadre de la future collaboration. Au cours de cette entrevue, l'écologue présente sa problématique, les caractéristiques des organismes étudiés, le plan d'échantillonnage, le type de données collectées et ses objectifs. Le biostatisticien questionne et écoute. Il tente de formaliser l'information, le dialogue peut s'avérer quelquefois difficile : « *L'écologue situe donc, sans le savoir, sa demande à un niveau mathématiquement très élevé et l'exprime suffisamment mal pour que le mathématicien ne reconnaisse pas les problèmes.* » (Legay 1984). Lorsque la collaboration est fructueuse, l'écologue et le biométricien sont comblés. Le premier a répondu à ses questions biologiques, le second a développé de nouvelles méthodes dont « *la mise au point […] ne se fait pas, ne peut pas se faire dans l'abstrait, mais bien à l'occasion de problèmes réels dans toute leur complexité et parfois même toute l'imprécision de leur formulation originelle* » (Legay 1976).

Ce travail de thèse s'inscrit dans le cadre d'une activité de biométricien. Différentes collaborations ont donc été entreprises avec des chercheurs du laboratoire ou d'organismes extérieurs. Au travers de ces collaborations, plusieurs problématiques écologiques m'ont été présentées. Bien que foncièrement différentes d'un point de vue biologique, ces questions présentent certaines similarités lorsque l'on s'intéresse aux outils et méthodes à mettre en œuvre pour le traitement statistique.

III.1. Les listes d'occurrences

Mon travail de DEA avait pour sujet l'analyse des listes d'occurrences (Dray 1999a, 1999b) et faisait suite à la thèse de Clémentine Gimaret-Carpentier (1999). Ces listes constituent une classe de données particulières dont l'unité de base est l'enregistrement de la présence d'un individu appartenant à une espèce à une date et en un lieu donné. Les collections muséographiques ou privées, les spécimens d'herbiers, les atlas existent depuis des siècles et fournissent une masse importante de listes d'occurrences. Le développement des ordinateurs a permis le stockage et le traitement de grands jeux de données. De nombreuses collections ont donc été informatisées et ces bases de données réservées initialement aux travaux concernant la systématique des organismes sont maintenant exploitables dans d'autres domaines tels que l'écologie ou la biogéographie.

Hawksworth (1995) estime à environ 280 millions le nombre d'exemplaires contenus dans les collections des Muséums à travers le monde. Il est évident que cette grande quantité d'information peut être un support privilégié pour étudier la répartition d'espèces à une large échelle spatiale et à moindre coût. Le caractère fondamental de ce type de données est

reconnu : « *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) mais peu de méthodes d'analyse réellement adaptées à ce type de données ont été développées. En effet, la démarche la plus usuelle consiste à diviser la zone d'étude à l'aide d'une grille de quadrats, de dénombrer le nombre d'occurrences par quadrat et d'utiliser des méthodes classiques d'analyse de données (*e.g.* AFC, ACC…) en considérant les quadrats comme de véritables relevés (figure 40). L'assimilation de quadrats définis *a posteriori* à de véritables relevés posent de nombreux problèmes d'un point de vue méthodologique. Il semblait alors nécessaire de tenir compte des caractéristiques des listes d'occurrences afin d'offrir des outils adaptés pour leur analyse.

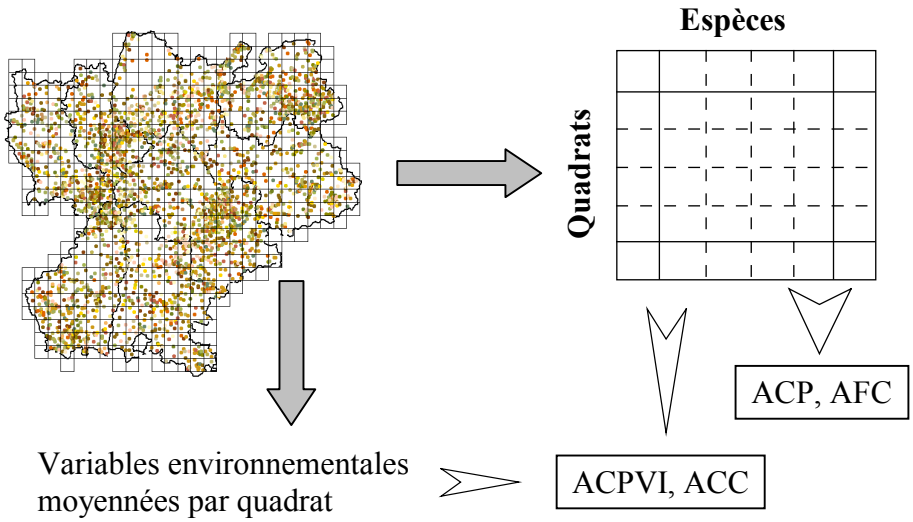


Figure 40 : Approche classique pour le traitement des listes d'occurrences. Un tableau espèces-quadrats est construit et analysé par une méthode classique (AFC…). Si de l'information mésologique est disponible, les variables environnementales sont moyennées par quadrat et liées au tableau faunistique par une ACC ou une ACPVI.

Pour réaliser un atlas sur les coléoptères carabiques de la région Rhône-Alpes (Coulon *et al.* 2001), le Muséum d'Histoire Naturelle de Lyon a recensé les spécimens de collection accumulés par de nombreux entomologistes depuis un siècle et demi. Le groupe des carabiques rassemble un grand nombre d'espèces (plus de 1000 pour la France) dont la plupart sont prédatrices d'autres organismes (vers,…) mais certaines sont omnivores ou même granivores. Les carabiques occupent pratiquement tous les milieux terrestres mais chaque espèce occupe un habitat particulier. Ce groupe d'insectes se révèle donc particulièrement intéressant pour appréhender des phénomènes écologiques : nombreuses espèces, statut de prédateurs situés au sommet de la chaîne trophique (reflétant *a priori* les niveaux inférieurs), présence dans tous les milieux et exigences pour des habitats particuliers. La base de données constituée par le Muséum regroupe plus de 35 000 occurrences représentant 549 espèces et provenant de plus de 200 collections. Le logiciel Biogeographica développé par L. Delaunay a été utilisé pour stocker cette base de données (figure 41). Des travaux préliminaires ont été réalisés sur ces données au cours du DEA mais n'ont pas été poursuivis à cause notamment de la difficulté à obtenir des données environnementales pour l'ensemble de la région Rhône-Alpes. Les résultats méthodologiques obtenus lors du DEA ont cependant été employés au cours d'autres travaux.

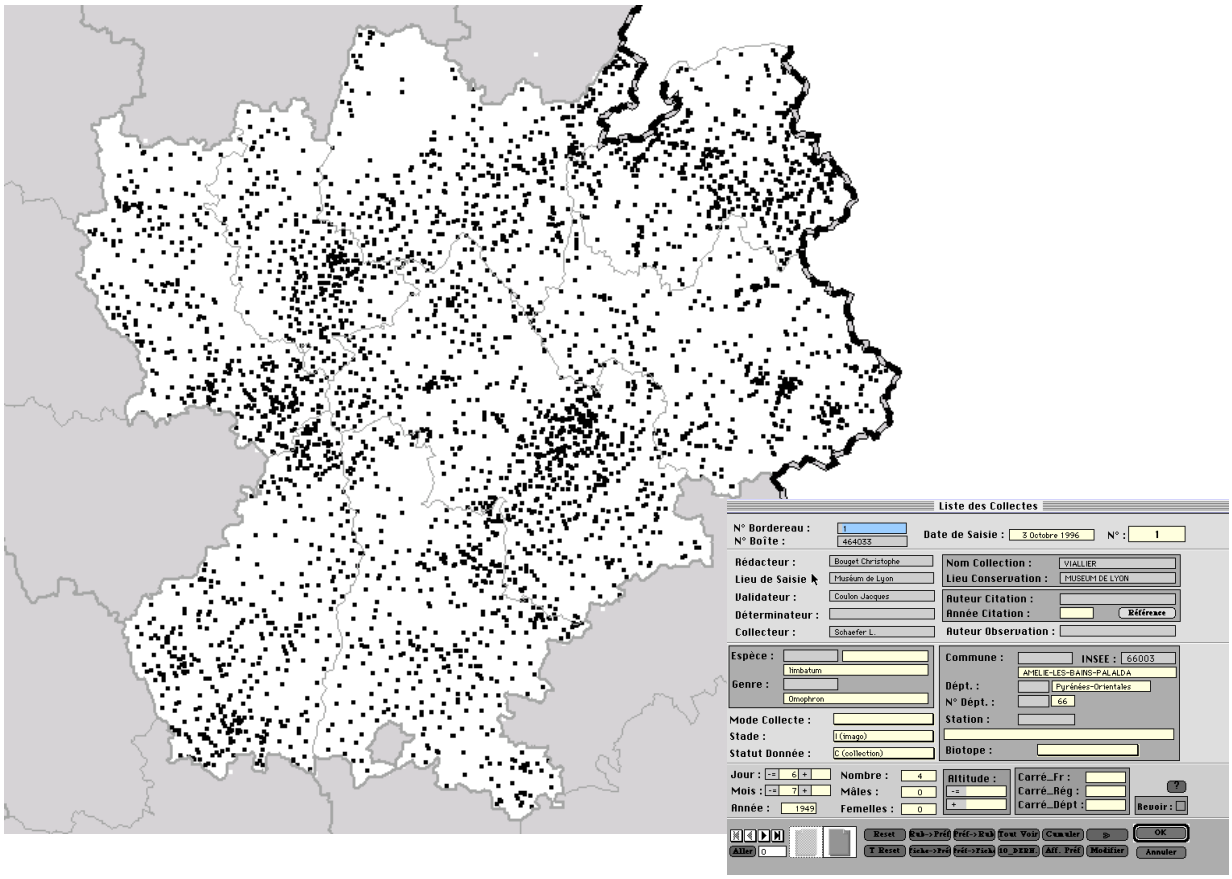


Figure 41 : Répartition spatiale des 35000 occurrences de carabiques en région Rhône-Alpes et copie d'écran du logiciel Biogeographica.

Dans le cadre de l'analyse des listes d'occurrences, une collaboration a été entreprise avec C. Gimaret-Carpentier et J.-P. Pascal (Laboratoire de Biométrie et Biologie Évolutive). L'étude portait sur des données tirées de l'atlas des espèces endémiques d'arbres des Ghâts occidentaux en Inde (Ramesh & Pascal 1997). Pour réaliser cet atlas, les auteurs ont répertorié de nombreux spécimens déposés dans des herbiers nationaux et internationaux. Afin d'étudier la distribution des espèces endémiques, une zone d'étude a été définie *a priori* (Gimaret-Carpentier 1999). Cette zone correspond à l'aire d'implantation naturelle des forêts sempervirentes humides et sèches des Ghâts occidentaux. Cette région a été reconnue comme l'un des « points chauds » de la biodiversité à l'échelle mondiale (Myers 1990) et constitue une zone de forte richesse spécifique et d'endémisme pour plusieurs groupes, tels que les arbres, les papillons ou les amphibiens (Ranjit Daniels 1992). De plus, les Ghâts forment un continuum forestier soumis à de forts gradients climatiques, si bien que les arbres de ces forêts constituent un cadre approprié pour étudier l'influence des facteurs climatiques sur la diversité des espèces.

Des cartes de topographie et de bioclimats ont été utilisées afin d'affecter à chaque occurrence des mesures de variables environnementales (température, durée de saison sèche et pluviométrie). La base de données regroupe plus de 5000 occurrences représentant plus de 300 espèces (figure 42). Environ cent espèces rares (moins de 5 occurrences) ont été éliminés de la base lors des analyses. La nouvelle base contenant 4920 occurrences correspondant à 224 espèces endémiques a été stockée à l'aide du logiciel ArcView. L'objectif de cette collaboration était d'étudier la distribution spatiale de la biodiversité à différentes échelles

taxonomiques, de la mettre en relation avec les facteurs environnementaux et d'émettre des hypothèses quant aux phénomènes de spéciation observés dans cette région.

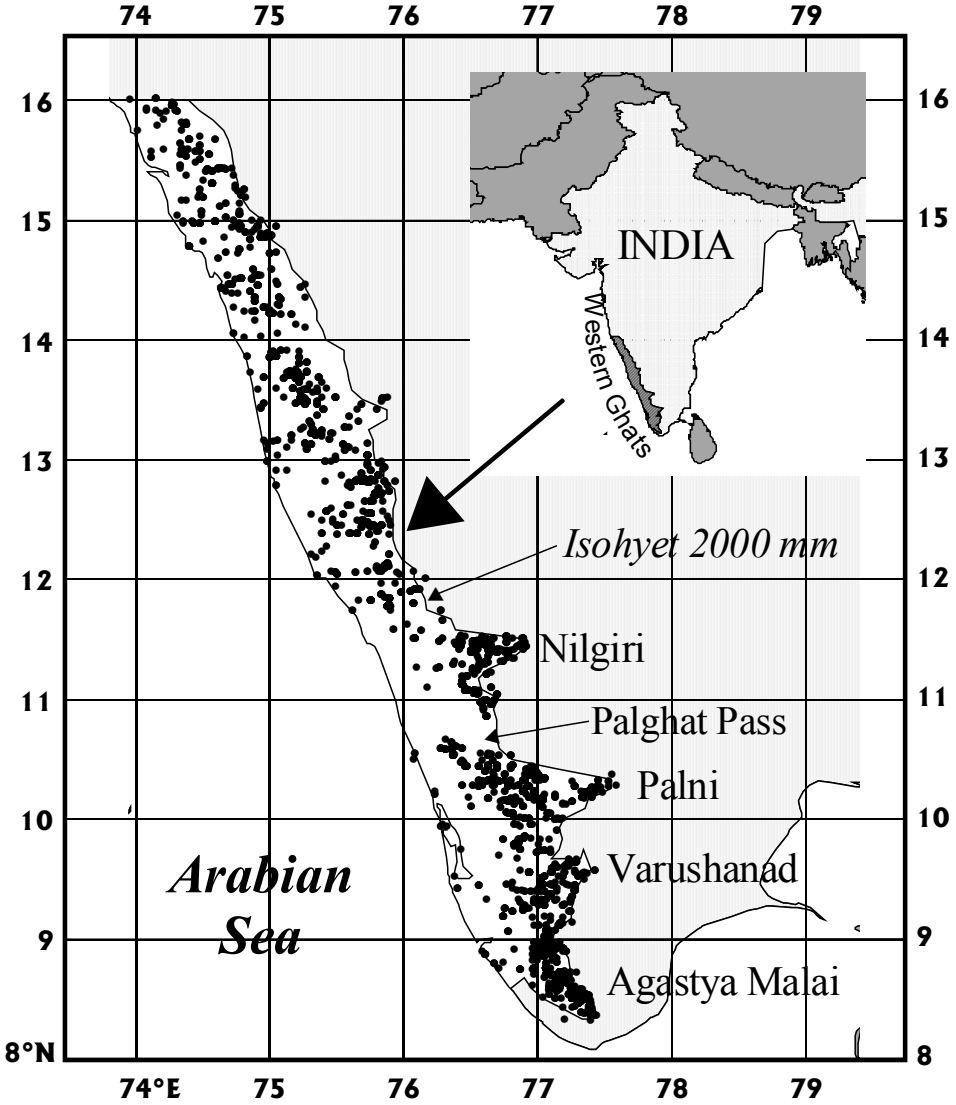


Figure 42 : Répartition spatiale des 5000 occurrences d'espèces endémiques d'arbres dans les Ghâts occidentaux.

Les compilations régionales émanant d'herbiers ou de collections muséographiques constituent des listes d'occurrences échantillonnées. En effet, les spécimens collectés ne forment qu'un échantillon des individus réellement présents. Certaines pratiques propres à la foresterie fournissent une autre source importante de listes d'occurrences. Afin d'étudier les structures horizontales et verticales du peuplement forestier, une pratique courante consiste à cartographier les arbres d'une parcelle. Chaque arbre est alors identifié par son nom d'espèce, ses coordonnées spatiales, sa circonférence, sa hauteur, son milieu local… Il en résulte une liste exhaustive d'occurrences dont tous les 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. Cette parcelle naturelle de forêt couvre au total 28 hectares et comporte plusieurs systèmes d'échantillonnage dont un

placeau rectangulaire de 100 mètres de large sur 150 mètres de long qui a été étudié lors du DEA. Ces données avaient été préalablement analysées, notamment lors de la thèse de Raphaël Pélissier (1995).

Les travaux entamés en DEA sur les listes exhaustives d'occurrences ont été poursuivis en collaboration avec Raphaël Pélissier (IRD) et Pierre Couteron (ENGREF). Les données provenant du dispositif de la Piste de St-Elie (Guyane française) constituaient le matériel biologique pour ce travail. L'ensemble des arbres (plus de 6000), d'un diamètre supérieur à 10 cm, contenu sur une parcelle de 10 hectares (100 mètres de large et 1000 mètres de long) a été cartographié. Une typologie des sols basée sur des caractéristiques hydromorphologiques a également été réalisée. Ainsi, chaque arbre est identifié par son espèce, son milieu local (type de sol) et ses coordonnées spatiales (figure 43). Une analyse factorielle des correspondances des profils écologiques avait permis une première exploration de ces données (Sabatier *et al.* 1997). En tenant compte des caractéristiques des listes d'occurrences, une méthodologie a été mise en œuvre afin d'étudier la répartition spatiale des espèces et de la diversité en espèces en relation avec l'hydromorphologie du sol.

StElie

	A	B	C	D
1	Espece	X	Y	Sol
2	Abarjupu	17.5	56.8	Alt
3	Abarjupu	20.3	437.3	UhS+DC
4	Abarjupu	61.6	628.1	DVD
5	Abarjupu	85.8	789.6	Alt
6	Abarmata	98.4	579.7	DVD
6125	Zygirace	38.3	49.4	SLD2
6126	Zygirace	44.2	64.3	DVD
6127	Zygirace	59.7	142.1	Alt
6128	Zygirace	31	121.6	Alt
6129	Zygirace	69.1	378.4	UhS
6130	Zygirace	80.9	328.5	UhS
6131	Zygirace	64.7	572.6	DhS
6132	Zygirace	60.5	622.8	DVD

Figure 43: Base de données de la Piste de St-Elie et cartographie du type de sol pour chaque arbre.

III.2. Maladie à transmission vectorielle

Une seconde collaboration, concernant la modélisation du risque trypanosomien, a été entreprise avec Jean-François Michel et Stéphane de La Rocque (CIRAD EMVT) dans le cadre de l'action thématique programmée « Santé-Environnement » (de La Rocque *et al.* 2001). Ce travail a été réalisé lors d'un stage au Centre International de Recherche Développement sur l'Élevage en zone Sub-humide (CIRDES) à Bobo-Dioulasso (Burkina-Faso) du 27 juillet au 24 août 2000. Ce stage avait pour objectif d'apporter un appui statistique pour le traitement de données d'enquêtes parasitologiques et sérologiques concernant les trypanosomes des bovins.

En Afrique sub-saharienne, les trypanosomoses (« maladie du sommeil » chez l'homme et « Nagana » chez le bétail) constituent des contraintes majeures pour la santé humaine et animale. Le système pathogène met en jeu trois acteurs (figure 44) : le parasite

(trypanosome), le vecteur (glossine) et l'hôte (bétail dans notre cas). Le cycle parasitaire peut être interrompu par une lutte contre ces vecteurs en utilisant des pièges mécaniques ou vivants (bétail imprégné d'insecticides). Afin de la rendre compatible avec les capacités techniques et financières des populations rurales, cette lutte doit être spatialement la plus ciblée possible.

Figure 44 : Système pathogène de la trypanosomose. Trypanosome (*Trypansoma brucei*, parasite), glossine (vecteur), bétail (hôte).

L'étude a été réalisée sur la zone agropastorale de Sidéradougou située au sud de Bobo-Dioulasso. Cette région s'étend sur environ 1200 km^2 et est traversée par plusieurs rivières. Une démarche pluridisciplinaire a été adoptée afin d'évaluer le risque trypanosomien. L'analyse d'images satellitales a permis de définir différents types de paysages. Sur les 120 kilomètres de réseau hydrographique, des pièges ont été posés afin d'estimer l'abondance de glossines. Une identification des milieux favorables à la présence de glossines a été obtenue en croisant la carte des paysages et celle concernant la répartition des glossines. Un recensement terrestre exhaustif a été réalisé dans cette zone et plus de 800 troupeaux (près de 16 500 bovins) ont été comptabilisés. Pour chaque troupeau, les coordonnées spatiales du camp et des points d'eau ont été relevées à l'aide d'un système GPS (Global Positionning System) et diverses informations relatives aux pratiques d'élevage ont été consignées (*e.g.* type de point d'eau, taille du troupeau, transhumance…). Un modèle d'occupation de l'espace par les troupeaux de bovins a été mis en place à partir de ces données (Michel *et al.* 1999). La superposition, à l'aide d'un SIG, des cartes concernant l'occupation de l'espace par les bovins et la distribution spatiale des glossines a permis de définir des zones de fort risque de transmission (i.e. zones de forte densité de vecteurs et d'hôtes, figure 45).

Une enquête sérologique a également été réalisée sur un échantillon de près de 1800 bovins représentant 216 troupeaux de la zone. Un modèle statistique a été mis en œuvre afin d'étudier le statut sérologique des animaux en fonction des pratiques d'élevage. Pour ce faire, il a été tout d'abord nécessaire de sélectionner les variables d'intérêt en accord avec les vétérinaires. En effet, les données collectées correspondaient à plus de 50 variables explicatives et il n'était ni possible ni raisonnable d'inclure l'ensemble de cette information dans le modèle. Une fois le modèle statistique mis en place, il a été possible d'identifier des pratiques d'élevage à risque et de représenter une prévalence prédite pour l'ensemble de la zone d'étude. L'identification des zones à fort risque de transmission et des zones de forte séroprévalence devait permettre de cibler une action de lutte contre les glossines.

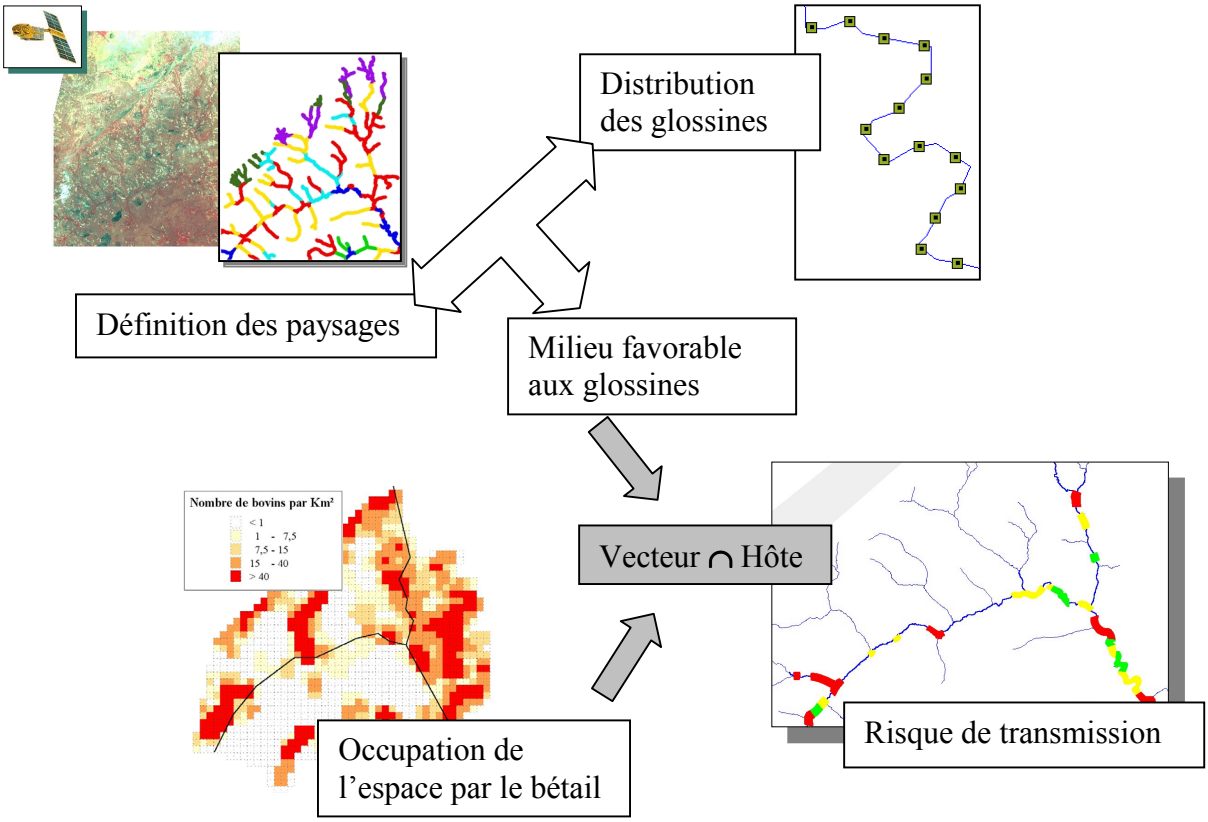


Figure 45 : Détermination des zones à risque. Différents types de milieux ont été définis par interprétation de photos satellitales. La carte des milieux a été mise en relation avec l'abondance des glossines afin de déterminer les milieux favorables aux glossines. La superposition des cartes des milieux favorables et de l'occupation de l'espace par le bétail permet d'identifier des zones à risque (i.e. favorables aux glossines et occupées par le bétail).

III.3. Dynamique de population

La forêt domaniale de Chizé est située à 20 kilomètres au sud de Niort (département des Deux-Sèvres) et s'étend sur une surface d'environ 5000 hectares. Une partie de cette forêt (2600 hectares) a été transformée en réserve nationale de la chasse et de la faune sauvage et est cogérée par l'ONF et l'ONC. Depuis 1978, la population de chevreuils est suivie à l'aide de méthodes de capture-recapture et chaque année environ la moitié de la réserve est échantillonnée (Gaillard *et al.* 1993). L'échantillonnage se fait en encerclant un groupement de parcelles de forêt choisies à l'aide de grands filets. Les individus occupant la zone sont alors capturés et marqués. Les chevrettes forment souvent des petits groupes « sédentaires » avec leurs faons et il est donc concevable d'assimiler le lieu de capture d'un jeune faon (8 mois). Depuis 1978, plus de 1200 faons ont ainsi été capturés. Pour chaque individu, son poids, son sexe et l'identifiant de la capture sont notés.

Une collaboration a été engagée avec Nathalie Pettorelli (NP) et Jean-Michel Gaillard du Laboratoire de Biométrie et Biologie Évolutive afin d'étudier les effets de l'environnement sur la dynamique de la population de chevreuils. Les analyses avaient pour objectif la mise en évidence d'une structuration spatiale des poids des faons en Janvier-Février et la mise en

relation de cette structure avec des caractéristiques de l'habitat. Ce trait biologique a été choisi parce qu'il est adapté pour décrire la dynamique de population. En effet, le poids des faons en Janvier-Février est étroitement lié au poids adulte, à la survie hivernale mais également à la densité de la population. Une partition grossière de la réserve en deux zones (Nord/Sud) basée sur la géologie, l'observation des types de forêts et le réseau routier avait été réalisée et des différences avaient été observées entre ces deux zones concernant le type de plantes, la qualité nutritive de l'habitat et les poids des faons (Pettorelli *et al.* 2001).

Un SIG a été mis en place par l'ONF concernant le peuplement forestier de la réserve en 1993 (figure 46a). La réserve a été divisée en 648 parcelles. Pour chaque parcelle, la composition floristique de la futaie est caractérisée par les quatre essences dominantes et celle du taillis par trois essences. Le pourcentage de chacune des espèces est également indiqué. Les données concernant les captures de chevreuils étaient consignées dans des fichiers Excel. Des petits utilitaires ont été programmés dans ArcView afin de pouvoir introduire les données concernant les chevreuils dans le SIG (figure 46b). Enfin, NP a réalisé un inventaire de la flore disponible pour le chevreuil (i.e. hauteur inférieure à 1,20 m) en Mai 2001. Pour ce faire, l'ensemble de la réserve a été couvert à l'aide de 578 points d'échantillonnage (figure 46c). Pour chaque station, les végétaux présents dans une placette de 1 m^2 ont été répertoriés en présence-absence. L'information taxonomique a été collectée au niveau du genre et la position de chaque point d'échantillonnage a été calculée à l'aide d'un GPS. Ces données ont également été introduites dans le SIG.

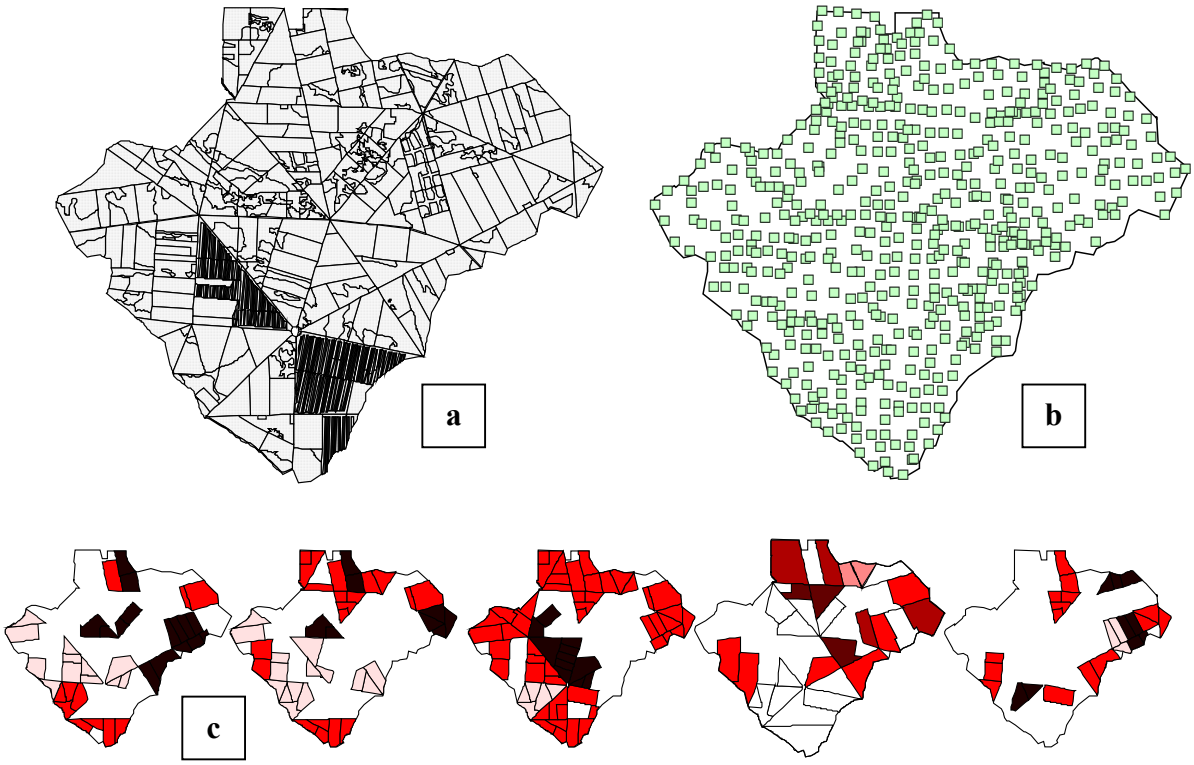


Figure 46 : La réserve de Chizé dans un SIG. (a) Parcelles de gestion forestière (futaie et taillis). (b) Points d'échantillonnage, réalisés par NP, de la végétation accessible au chevreuil. (c) Quelques exemples de représentation du poids moyen des faons à l'aide des données provenant d'échantillonnages annuels.

L'objectif principal de cette collaboration était d'établir une typologie de l'habitat afin de connaître les différents types de milieux de la réserve. Parallèlement, il fallait étudier une éventuelle structuration spatiale de la distribution des poids des faons et la mettre en relation avec les caractéristiques de l'habitat. L'introduction de l'ensemble des données dans le SIG a fait apparaître de nouvelles problématiques dont la résolution a nécessité des développements méthodologiques. Ainsi, l'étude de la concordance entre le couvert forestier et la végétation au sol a entraîné le développement d'une méthode statistique permettant de coupler des données provenant de deux plans d'échantillonnage distincts (parcelles ONF et stations NP). De même, une structure spatiale dans la distribution des poids des faons à partir des données collectées en 2000 et 2001 a été observée et mise en relation avec l'habitat disponible pour le chevreuil. Il s'est alors posé la question de la variation temporelle de cette structure spatiale. Une méthode a donc été mise au point pour comparer plusieurs cartes correspondant à différents plans d'échantillonnage.

Partie II : Résultats

Chapitre IV : L'analyse des listes d'occurrences

Ce chapitre est composé de trois publications acceptées. Dans le cas des listes d'occurrences, il s'avère que l'utilisation de l'analyse canonique des corrélations (AC) est adaptée. Cette méthode, qui présente de nombreuses contraintes, a été rapidement abandonnée en écologie après de sévères critiques (Austin 1968, Gauch & Wentworth 1976). Lors de l'analyse des listes d'occurrences, les principales contraintes de l'AC n'ont plus lieu d'être (conditions numériques, linéarité des relations). L'AC permet alors une ordination sans relevés avec les mêmes qualités que l'analyse canonique des correspondances. Des variantes de cette analyse sont proposées (analyses partielles, *trend surface analysis*…) afin d'identifier les principales structures écologiques de deux jeux de données.

La première publication est le fruit de la collaboration établie avec C. Gimaret-Carpentier et J.-P. Pascal. Le jeu de données, constitué par 5142 spécimens d'herbiers, fournit une liste d'occurrences échantillonnée à une échelle régionale (Ghâts occidentaux). Ces données permettent d'étudier des processus écologiques influant sur la biodiversité. Au moyen de l'analyse canonique des corrélations, l'habitat des espèces est défini d'un point de vue spatial et écologique. De nombreuses analyses, dont certaines tiennent compte de l'information taxonomique à l'aide d'analyses partielles, suggèrent l'existence de deux phénomènes qui pourrait expliquer l'origine de la composition spécifique actuelle : spéciation allopatrique (divergence provoquée par une séparation géographique) entre les deux côtés des Ghâts et radiation adaptative (diversification rapide d'un groupe en liaison avec la colonisation d'habitats nouveaux) le long des gradients environnementaux.

Les deux autres articles concernent également l'analyse des listes d'occurrences. Les données provenant du dispositif de la Piste de St-Elie (Guyane française) fournissent une liste exhaustive d'occurrences pour laquelle tous les individus de la zone sont connus et répertoriés. L'information relative aux coordonnées spatiales et à l'hydromorphologie du sol a été collectée pour 4992 arbres, représentant 120 espèces, distribués sur une parcelle de 10 hectares.

Dans la seconde publication, les relations entre la distribution spatiale des espèces et les contraintes hydrologiques du sol sont étudiées grâce à l'analyse canonique des corrélations. La représentation graphique des résultats adoptée dans cet article permet de figurer l'optimum (moyenne) et la tolérance (variance) des espèces, par rapport à un gradient environnemental, sur une carte factorielle. Une première analyse révèle que, mis à part quelques espèces privilégiant les milieux saturés durablement en eau, les espèces les plus abondantes peuvent être ordonnées selon deux gradients de tolérance limitant l'amplitude des niches. Le premier est un gradient de tolérance à la saturation prolongée en eau. Le second, de moindre importance, concerne uniquement les espèces intolérantes à une saturation prolongée. Il les sépare selon leur capacité à supporter à une saturation temporaire. Il semble donc que les contraintes induites par une saturation en eau ont plus d'impact sur la distribution spatiale des espèces que celles provoquées par un manque d'eau. Une seconde analyse fondée sur les coordonnées géographiques confirme l'influence prépondérante de la perméabilité du sol sur les structures spatiales multispécifiques observées. Une seconde structure spatiale, indépendante de l'effet du sol, est également observée et pourrait refléter l'influence de facteurs endogènes résultant de la dynamique de population.

Le troisième article aborde le lien entre analyses multivariées et mesures de diversité. En effet, l'approche par les listes d'occurrences permet de réaliser une partition de la biodiversité en fonction des variables explicatives. Dans le cas d'une seule variable qualitative, on retrouve la décomposition de la diversité en deux composantes (Whittaker 1972) : diversité intra-biotope (α) et diversité inter-biotope (β). Dans ce cadre, un document est fourni en annexe (Annexe 1) afin d'expliciter la marche à suivre pour réaliser ce type d'analyse à partir du logiciel ADE-4.

Running head: Western Ghats Biodiversity Pattern Analysis

BROAD-SCALE BIODIVERSITY PATTERN ANALYSES OF THE ENDEMIC TREE FLORA OF THE WESTERN GHATS (INDIA) USING CANONICAL CORRELATION ANALYSIS OF HERBARIUM RECORDS

C. GIMARET-CARPENTIER †, 1

S. DRAY †, 2

and

J.-P. PASCAL †, 3

Abstract: A crucial step in understanding the origin and maintenance of biological diversity is the assessment of its distribution over space and time and across environmental gradients. At the regional scale, two important attributes of species can be assessed that provide insight into speciation processes: species geographical and environmental ranges. The endemic tree flora of the Western Ghats is an interesting case for analyzing broad-scale biodiversity patterns because of the steep environmental gradients that characterize this tropical region of India. Species geographical and environmental ranges were analyzed by Canonical Correlation Analysis of point data from herbarium collections. This method enabled us to identify different levels of organization in the distribution of endemic species in the Western Ghats, and suggested two general processes that could explain its origin: allopatric speciation between the two sides of the Ghats and adaptive radiation along environmental gradients.

Key words: broad-scale biodiversity patterns, point data, herbarium specimen records, environmental gradients, canonical correlation analysis, macroevolution, Indian evergreen forests.

INTRODUCTION

Increasing awareness of the ongoing loss of biological diversity has heightened interest in its estimation, origin, function, and maintenance (Heywood and Watson 1995, Hawksworth 1995). The assessment of biodiversity patterns over space and time and across environmental gradients is a crucial step in the understanding of these issues. With the development of macroecology, broad-scale biodiversity patterns have been recognized (Brown 1995). These approaches have emphasized the importance of processes that act at the regional scale, combining ecological, historical and evolutionary factors (Ricklefs and Schluter 1993). Species geographical and environmental ranges are two important attributes of species that can be assessed at such a scale and can provide insights into speciation processes. Indeed, allopatric and sympatric modes of speciation have different effects on the geographical ranges of sister taxa (Taylor and Gotelli 1994, Barraclough and Vogler 2000), whereas species niche differentiation may indicate adaptive radiation (Knox and Palmer 1995, Schluter 1996, Losos et al. 1998).

The endemic tree flora of the Western Ghats is an interesting case for investigating large-scale biodiversity patterns. The Western Ghats form a 1500 km-long escarpment parallel to the southwestern coast of the Indian Peninsula, recognized as one of the world biodiversity hotspots (Myers et al. 2000). The interaction of the summer monsoon winds with the relief of the Ghats results in two steep environmental gradients, a west-east decrease in rainfall and a south-north increase in dry season length, that determine strong changes in the vegetation. The western side of the Ghats supports wet evergreen forests, whereas the eastern side supports deciduous forests, with the exception of a belt of dry evergreen forests in the south (Pascal 1988). In the north, the wet evergreen forests are displaced by deciduous formations because of the lengthening of the dry season. As a consequence, the evergreen forests of the Western Ghats are isolated from their counterpart in the north-eastern part of India, a feature that may explain their high level of endemism: 63% of the evergreen tree species are endemic to this region (Ramesh and Pascal 1991). The uplift of the Western Ghats would date from the Pliocene; the concomitant decrease in rainfall and increase of in dry season length would have induced the replacement of the wet evergreen forests of the Indian Peninsula by deciduous forests and the formation of dry evergreen forests in the south on the eastern side of the Ghats, while reducing the area formerly occupied by wet evergreen forests to the western side of the Ghats and the north-eastern region of Assam (Legris and Meher-Homji 1982). So the endemic tree species of the Western Ghats share a common biogeographic history since the isolation of the evergreen forests, and their geographical and ecological ranges can be compared to provide insights into speciation processes (Taylor and Gotelli 1994).

One way to assess species distributions consists of gathering available records from the collections in museums or herbaria (Mourelle and Ezcurra 1996, Skov and Borchsenius 1997, Shaffer et al. 1998, Parmesan et al. 1999, Peterson et al. 1999). However, as stressed by Ponder et al. (2001), these data suffer from several biases: they can provide reliable data about species presence, but species absence is generally dubious due to geographical gaps in the sampling effort. Moreover, each record is uniquely characterized by the identity of the collector, the place and the date of record, so it corresponds to a distinct sampling unit which does not provide any information about the presence of other species: local species diversity cannot be assessed directly from such data.

Two sets of methods have been developed to deal with collection data. The first set of methods aims at modeling individual species distribution, coupling point data with abiotic variables to define species habitat requirements (Stockwell and Peters 1999, Bonn and Schröder 2001, Ponder et al. 2001). The second set of methods aims at modeling species communities, analyzing the spatial and environmental distribution of species diversity (Guisan et al. 1999). In the context of unimodal species response curves, Canonical Correspondence Analysis (CCA; ter Braak 1986) is very efficient to investigate niche separation along environmental gradients. CCA requires a table containing abundances or presence-absence of species at a series of sites and measurements of environmental variables for the same sites where a site is "*the basic sampling unit, separated in space or time from other sites*" (ter Braak 1986, p.1167). In principle, CCA can be extended in the case of herbarium data by dividing the study area into smaller units, such as grid cells. Number of records are then computed for each cell and variations in species diversity are analyzed from these units (Hill 1991, Bolognini and Nimis 1993, Heikkinen 1996). However, if one uses spatial units defined a posteriori, one cannot control the sampling effort, which may biase the analyses (species absences are more likely to result from a poor sampling effort). Moreover, the size of the cells can influence the results (Gaston 1994, Böhning-Gaese 1997, Donald and Fuller 1998). In particular, the relationships between species occurrences and environment will not be tightly coupled because environmental data are then averaged at the cell scale.

In this paper, we use a general method based on Canonical Correlation Analysis (CANCOR, Hotelling 1936) to assess variations in species diversity directly from a list of independent species occurrences, such as geo-referenced herbarium specimens. We show that CANCOR, which was found ineffective for ordination of real relevés (Gauch and Wentworth 1976), is especially suitable for collection data. Our aim was to analyze the distribution of tree endemism in the Western Ghats; we assessed broad-scale biodiversity patterns using different kinds of analyses in order to reveal different levels of organization.

MATERIAL AND METHODS

Presentation of the data set

To depict the geographical ranges of the endemic evergreen tree species of the Western Ghats, Ramesh and Pascal (1997) gathered occurrences from three sources: (i) specimens in herbaria, (ii) published data and (iii) results of field surveys by the French Institute of Pondicherry and other botanists. From this database we extracted the 5142 occurrences that correspond to herbarium specimens, forming a set of reliable and independent observations, verified by the same expert (a botanist from the French Institute). Taxa represented by fewer than 5 occurrences were not considered, as their geographical range could not be assessed properly. A total of 224 endemic tree species, belonging to 104 genera and 38 families, were documented.

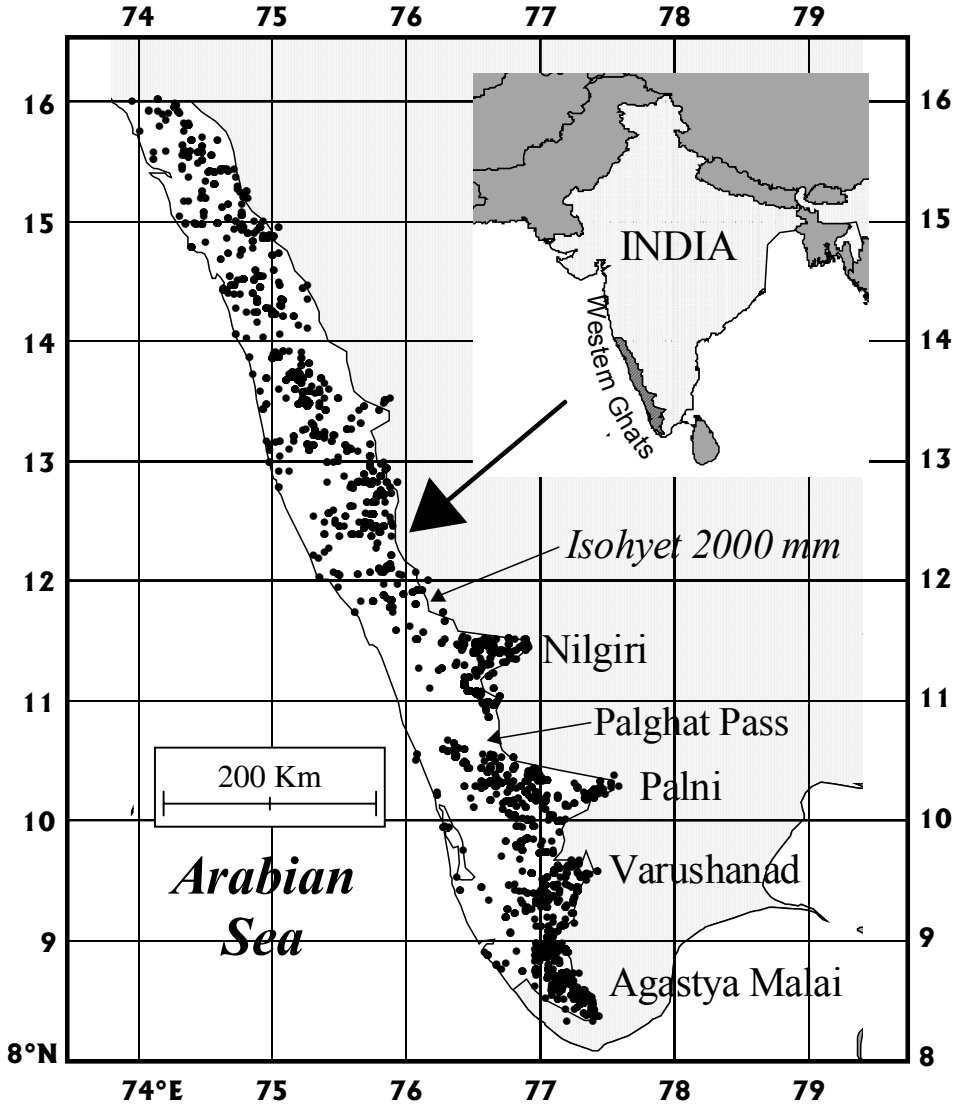


Figure 1: Study area and location of the 5142 herbarium specimens.

The 5142 species occurrences occupy 8 degrees of latitude, from 8° to 16°, the area covered with evergreen forests delimited towards the east by the isohyet 2000 mm (fig. 1). In this region, the escarpment of the Ghats consists almost exclusively of archean rocks with varying degrees of metamorphism, and the soils (saturated acidic soils) appear homogeneous (Bourgeon 1989). Studies carried out by the French Institute of Pondicherry (e.g. Pascal 1992, Ramesh et al. 1997) have shown that, at the regional scale, species turnover is driven by three climatic factors: variation in total annual rainfall, rainfall seasonality (dry season length) and temperature, associated with altitude. So, we assessed species environmental ranges by compiling these data for each occurrence (fig. 2). Altitude was recorded for each specimen in the database, while mean total annual rainfall and mean annual dry season length were deduced from bioclimatic maps of the Western Ghats (Pascal 1982). The dry season is defined by the number of months during which the amount of rainfall measured in mm is lower than twice the temperature expressed in °C (Bagnouls and Gaussen 1953).

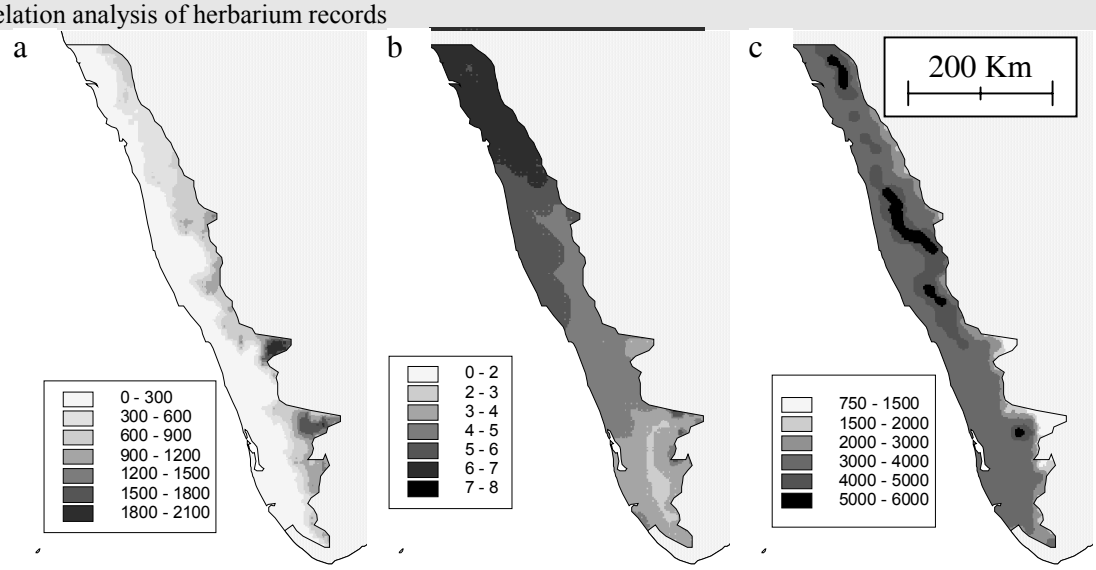


Figure 2: Variation of the three main environmental factors in the Western Ghats. Mapping of occurrence values by contour curves computed over the 1000 nearest occurrences for a) altitude (m), b) mean annual dry season (months), c) mean annual rainfall (mm).

Canonical Correlation Analysis of species occurrences

Species information concerning the 5142 occurrences gathered from natural history institutions is contained in a species by occurrences table X with $x_{ij} = 1$ if the *i*th occurrence belongs to the *j*th species, otherwise $x_{ij} = 0$. A table Y describes the environmental features at each locality where y_{ik} represents the measurements of the *k*th environmental variable recorded for the *i*th occurrence (fig. 3). As noted above, species occurrences such as herbarium specimens are independent observations, so they represent the actual statistical units. As a consequence, tables X and Y represent two sets of variables measured on the same statistical units, the 5142 occurrences. In this case, the number of observations is much higher than the number of variables. Such characteristics correspond exactly to the requirements of CANCOR, a method that finds the linear combinations of one set of variables that are maximally correlated with linear combinations of another set of variables (Gittins 1985).

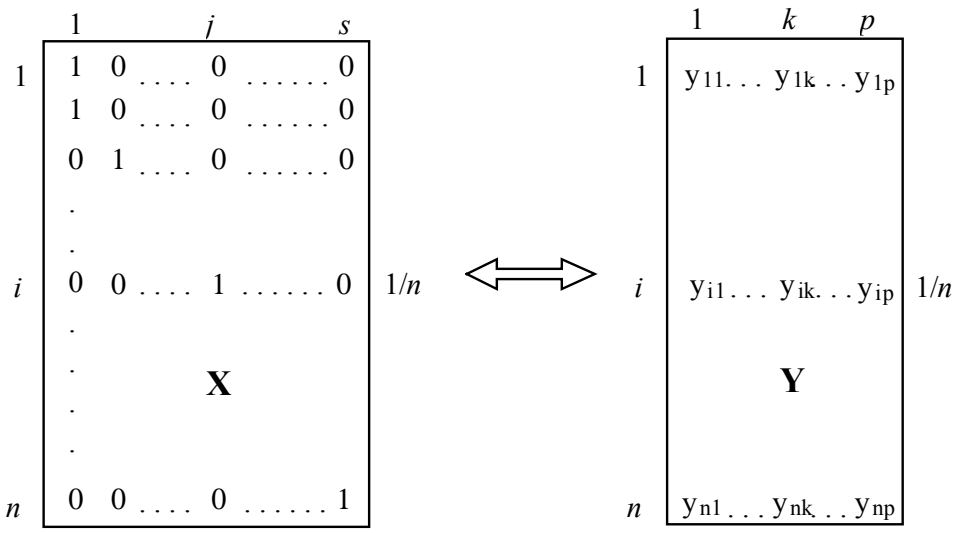


Figure 3: Occurrences-by-species (X) and occurrences-by-environment (Y) tables obtained from a list of *n* occurrences distributed among *s* species and characterized by *p* environmental variables. Species appear as dummy variables indicating species identity, while the environmental factors are considered as quantitative variables.

As X is formed by variables indicating species membership, the CANCOR of X and Y finds linear combinations of environmental variables that best discriminate the species. This is equivalent to Discriminant Analysis (DA) of Y by the specific groups of occurrences in X (Lebart et al. 1984), also known as Canonical Variates Analysis (James and McCulloch 1990). DA has been used in ecology by Green (1971), and Lebreton et al. (1988) demonstrate that Green's DA is mathematically equivalent to the CCA of ter Braak (1986). ter Braak and Verdonschot (1995) admit this link and specify that "*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*". Hence, just like CCA, our CANCOR approach provides linear combinations of environmental variables that maximize species niche separation. CANCOR of occurrence data can be viewed as a CCA without any sampling site. Contrary to multivariate analyses based on a partition of the study area into grid cells, CANCOR computes species environmental means directly from the list of occurrences, i.e. only from species presences. Sampling bias associated with the use of cells is thus removed. The analysis allows a direct comparison of species environmental ranges, i.e. abiotic realized niches. We defined species niches by ellipses containing 50 percent of species occurrences on the canonical planes, following the procedure described by Thioulouse and Chessel (1992) and Pélissier et al. (in press).

We analyzed species geographical ranges in a similar way, with the table Y containing now polynomial functions of the geographic coordinates of species occurrences, i.e. the variables x, y, x^2, xy, y^2, etc., where x and y are, respectively, occurrence longitude and latitude. The CANCOR of X and Y becomes a Canonical Correlation Trend Surface Analysis (CCTSA) (cf. Gittins 1968, Wartenberg 1985). This analysis identifies multispecies spatial patterns, i.e. gradients of species composition or areas of high endemism, according to the degree of overlap among species ranges. In order to visualize these patterns, we computed contour curves from the canonical scores mapped on the study area, following Thioulouse et al. (1995).

Partial analyses

CANCOR permits partial analyses to study the relationships between X and Y. In this case, Y contains variables from which the effect of another set of explanatory variables has been removed. Principles of the method are the same as those used for partial CCA (ter Braak 1988). The mathematical details are explained in Pélisser et al. (2001), following Méot et al. (1998). We used these partial analyses in two ways: (i) to remove the effect of environmental factors on the multispecies spatial patterns and identify other correlates of species distribution, such as geographical barriers to dispersal, and (ii) to analyze the contribution of higher taxonomic levels to species geographical and environmental ranges. In the latter case, the partial analyses are similar to within-group analysis, depicting species-environment relationships or multispecies spatial patterns within genera or within families.

Permutation test

The statistical significance of the results of CANCOR can be evaluated by a random permutation test (Manly 1991). First, the number of eigenvalues of interest (l) is determined from the graph of the decrease in CANCOR eigenvalues, and the sum of the l first eigenvalues is computed; CANCOR eigenvalues are equal to the square of the canonical correlation coefficients, so the sum of eigenvalues measures the strength of the correlation between X and Y. Then the rows of Y are permuted randomly and CANCOR analysis is performed between X and the permuted table Y. We performed 1000 permutations and computed the sum of the l first eigenvalues for each one. We compared the results obtained from the original data set with those obtained after permutations of the rows of Y to obtain a p-value.

All calculations and related graphs were generated with ADE-4 software (Thioulouse et al. 1997), available on the internet at URL: http://pbil.univ-lyon1.fr/ADE-4/). The maps were generated with Arcview 3.2.

RESULTS

Multispecies spatial patterns

To obtain a general picture of the distribution of endemism in the Western Ghats, we first analyzed the changes in endemic species composition in the study area by means of CCTSA, using polynomial functions of the geographic coordinates of order less than four, i.e. the 9 variables x, y, x^2, xy, y^2, x^3, x^2y, xy^2 and y^3 (x and y are, respectively, occurrence longitude and latitude).

The analysis produced two highly significant factors representing 52% of total variance ($p < 0.001$). The canonical correlation coefficients of these first factors were respectively 0.747 and 0.655. The first factor reveals a strong change in endemic species composition from the coast towards the top of the Ghats, and the progressive disappearance of this β-diversity gradient in the northern part of the study area (fig. 4a). The second factor highlights the floristic distinctness of the southeastern part of the study area, revealing that several endemic species are restricted to this region (fig. 4b). However, the variation in endemic species composition from the southeast is not uniform, with steep gradients of β-diversity observed towards the coast and the tops of the Ghats, whereas the change in endemic species composition towards the north appears much less marked.

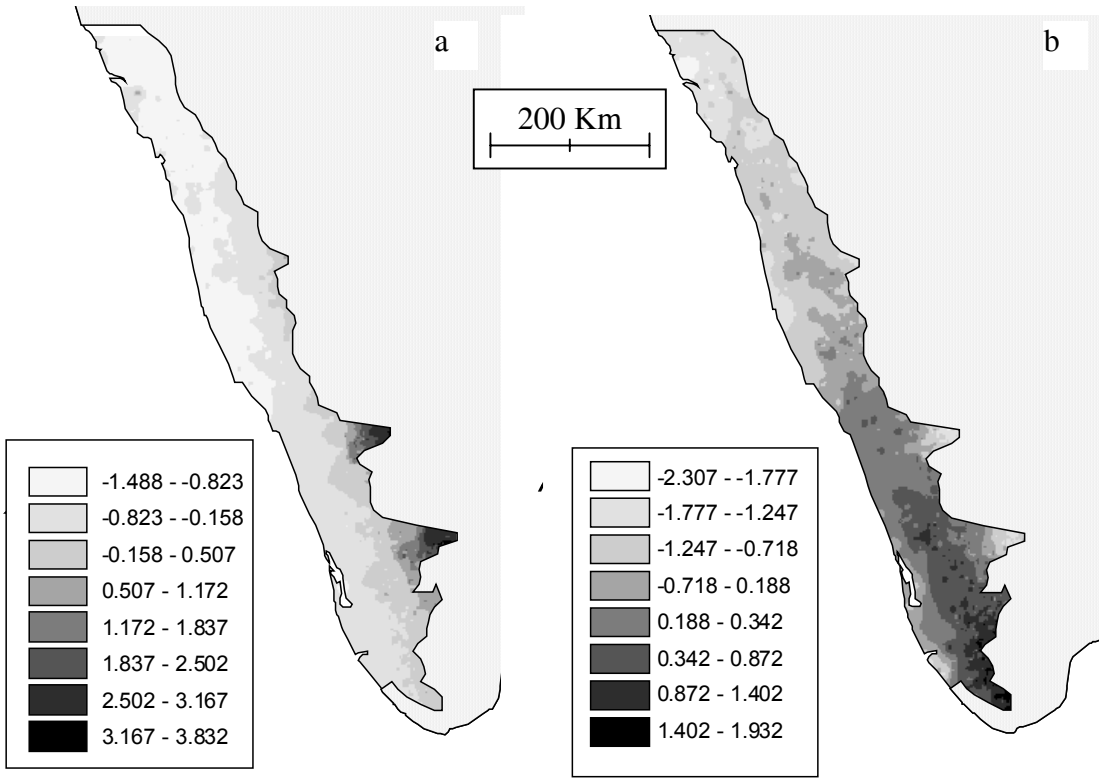


Figure 4: Multispecies spatial patterns obtained by CCTSA. Mapping of occurrence canonical scores by contour curves computed over the 500 nearest neighbours. a) first factor, b) second factor. The colours reflect variation in occurrence scores (cf. insert).

Multispecies spatial patterns independent of environmental factors

We performed a partial CCTSA to remove the effect of environmental variables and analyze the residual component of species distribution (the same polynomial functions of geographical coordinates were used).

The first two factors were highly significant, representing 52% of total variance ($p < 0.001$). The canonical correlation coefficients were smaller than in the previous analysis, 0.483 and 0.466. They reveal more complicated spatial patterns, confirming the strong correlation between environmental factors and species distribution (fig. 5). Species canonical scores (obtained as the centroid of their occurrence scores) allows the identification of species groups sharing the same spatial features; the occurrences of species representing each group can then be mapped by their physical coordinates.

The first axis highlights species with narrow latitudinal ranges (fig. 5a). At the negative end, there are species whose distributions are limited to the central part of the study area between 10 and 12°N, such as *Pithecolobium gracile* at low elevation or *Euonymus angulatus* at medium elevation, whereas at the positive end, there are two kinds of species: (i) species whose geographical ranges are limited to the southern mountains, and more specifically occurring on their eastern side, such as *Garcinia travancorica* and *Diospyros foliolosa*, and (ii) species restricted to the northwestern part of the study area, such as *Mammea suriga*. Three floristic regions are thus delineated, each of which harbours a specific endemic tree flora (fig. 5a).

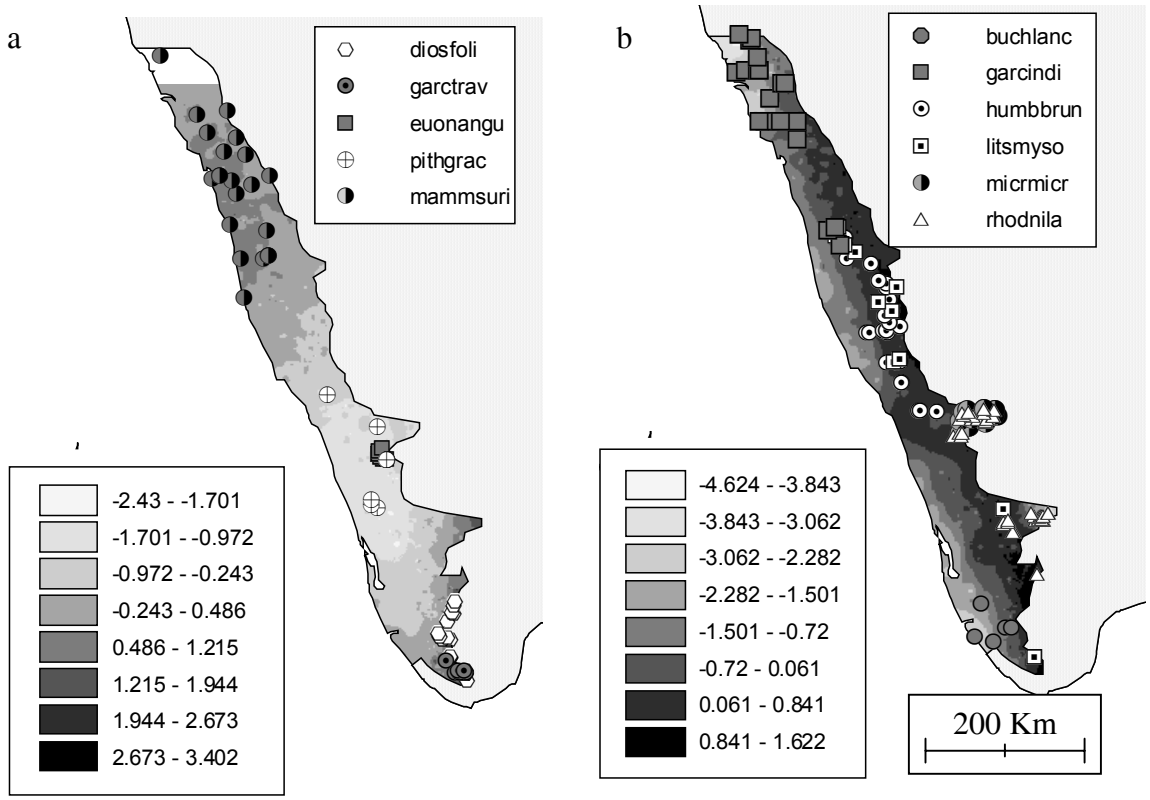


Figure 5: Multispecies spatial patterns obtained by partial CCTSA removing the effect of environmental variables. Mapping of occurrence canonical scores by contour curves computed over the 500 nearest neighbours, and projection of the occurrences of species characteristic of the spatial patterns identified. The colours reflect variation in occurrence scores (cf. insert). See appendix for species names.

The second axis highlights species with narrow longitudinal ranges, i.e. with different altitudinal ranges, pointing to the floristic homogeneity of the western slopes of the Ghats

(fig. 5b). At the positive end, are species whose distributions correspond to these middle elevation slopes, such as *Litsea mysorensis* and *Humboldtia brunonis*, whereas at the negative end, are two kinds of species: (i) species whose distributions are restricted to the coastal area, such as *Buchanania lanceolata* in the south and *Garcinia indica* in the north, and (ii) species located on the top and/or the eastern slopes of Palni and Nilgiri mountains, such as *Microtropis microcarpa* and *Rhododendron nilagiricum*. An overlap of species geographical ranges is observed on the western slopes of the Ghats, as shown by the mapping of species occurrences (fig. 5b).

Environmental ranges of endemic species

We then analyzed the role of environmental factors by assessing species niche differentiation along the three main environmental gradients. This was done by performing a CANCOR between the occurrences-by-species and occurrences-by-environment tables (cf. fig. 3).

The first two factors were highly significant, representing 87% of total variance ($p < 0.001$). The canonical correlation coefficients were equal to 0.803 and 0.573. The first factor represents the altitudinal gradient, whereas the second factor corresponds to the gradient of dry season length and, to a less extent, the rainfall gradient (fig. 6a). On the first canonical plane, the projection of species at the centroid of their occurrences forms a triangle, showing the correlation between altitude and dry season length (fig. 6b). The tips of this triangle are occupied by three kinds of species with a narrow ecological range: (i) species restricted to the lowlands with a short dry season, such as *Myristica fatua* var. *magnifica* and *Gymnacranthera canarica*, two species typically associated with swamps, (ii) species restricted to the high elevations, such as the unique species of Magnoliaceae *Michelia nilagirica* and *Viburnum hebanthum*, and (iii) species restricted to the lowlands with a long dry season, such as *Diospyros angustifolia* and *Garcinia indica*. Between these species, are species with a broader environmental range, with the generalists located near the origin of the canonical plane.

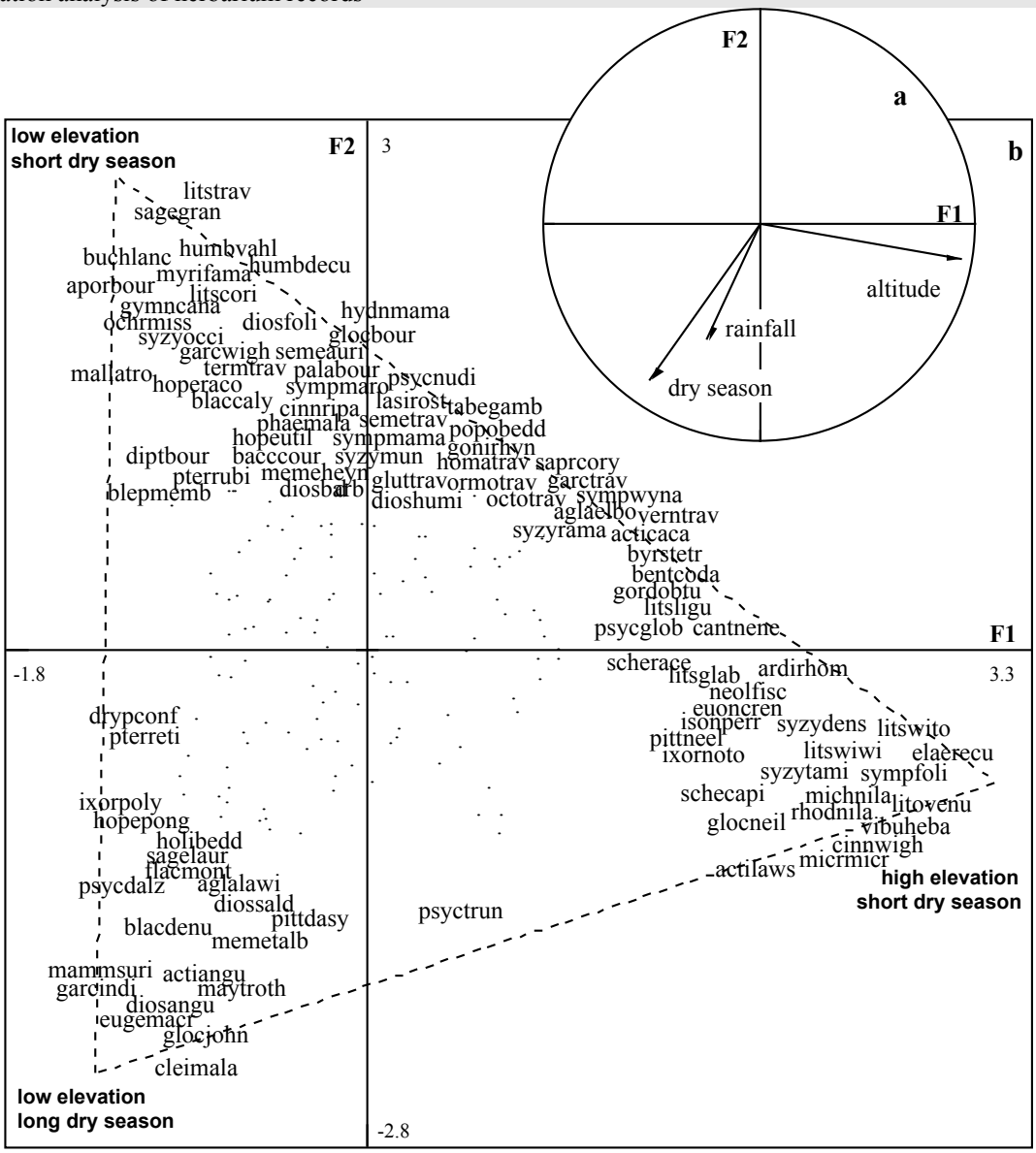


Figure 6: Results of CANCOR of occurrences-by-endemic tree species and occurrences-by-environment tables. a) correlation of the environmental variables with the first two canonical factors, b) projection of species at the centroid of their occurrences. Species with the highest scores are represented by their names, with the other species represented by dots. The dotted area shows the triangle formed by species niches and corresponding to three combinations of environmental factors. See appendix for species names.

Species niche separation among congeneric species

In order to assess the importance of species niche separation among congeneric species, we plotted the results separately for the three most species-rich genera, *Litsea* (Lauraceae), *Diospyros* (Ebenaceae) and *Syzygium* (Myrtaceae); the low number of species per genus allowed the visualization of species niches by ellipses containing 50% of their occurrences (fig. 7).

Litsea species occupy most of the environmental space, displaying rather wide ecological amplitudes, except three species with a narrow altitudinal range, *L. travancorica*, restricted to the lowlands, *L. keralana*, occurring at middle elevation, and *L. wightiana* var. *tomentosa*, inhabiting the tops of the mountains (fig. 7a); the highest niche overlap is

observed at middle elevation (800 to 1400 m) under medium dry season length (3 to 4 months).

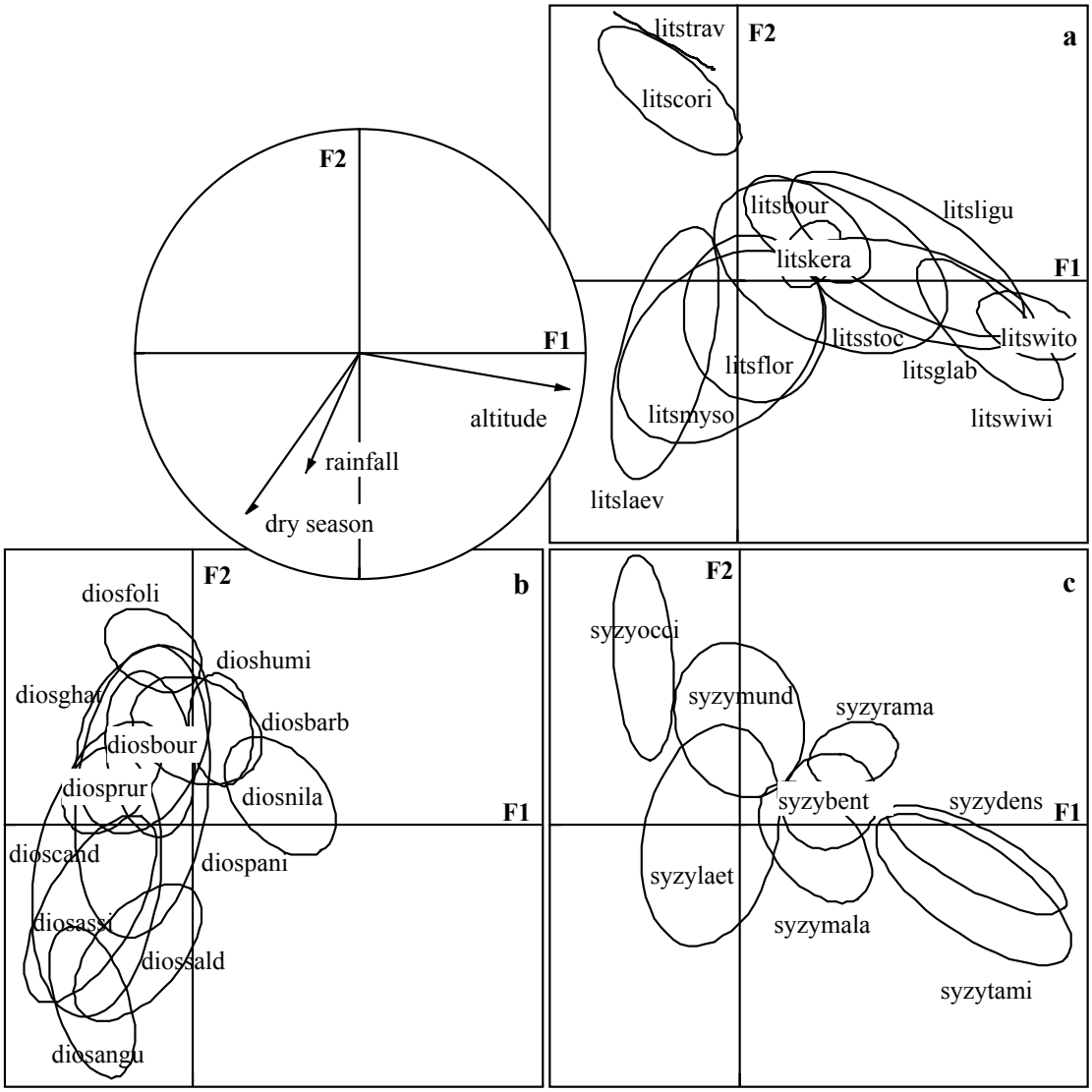


Figure 7: Species environmental ranges within the most species-rich genera. a) *Litsea*; b) *Diospyros*; c) *Syzygium*. Species represented by ellipses containing 50% of their occurrences. See appendix for species names.

Contrary to *Litsea*, *Diospyros* is primarily located in the lowlands (fig. 7b). Only one species is observed at medium elevation, *D. nilagirica*, and none occurs at high elevation. The length of the dry season in addition to the amount of rainfall appears as the main factor of species niche differentiation. For example, the second canonical factor separates species tolerating a long dry season, such as *D. angustifolia*, from species inhabiting the lowlands under a short dry season, such as *D. bourdillonii*. Within this last group, it distinguishes *D. foliolosa*, a species located on the eastern side of the Ghats, where the annual rainfall is lower than on the western side. The highest niche overlap is observed at low elevation under a short dry season.

With 20 endemic species, *Syzygium* represents the most species-rich genus of the endemic tree flora of the Western Ghats, but the majority of its species are rare. Compared with the two previous genera, species niche differentiation among the 8 most abundant *Syzygium* species is especially high (fig. 7c).

Contribution of higher taxonomic ranks to species distribution

Species geographical and environmental ranges are expected to differ among genera and families, as these higher taxonomic ranks may be characterized by a series of traits influencing species dispersal capacities or environmental tolerance. We analyzed the contribution of genus and family to species ranges by means of partial CANCOR. If genera (or families) are not randomly distributed, the results of these partial analyses should differ from the previous ones.

After partialling out the effect of genus or family on species ranges, the canonical correlation coefficients decrease, indicating that these higher taxonomic levels partly determine species environmental and geographical ranges (tables 1 and 2). In both cases, the influence of family is lower than the influence of genus. In terms of patterns however, no difference is observed regarding species-environment relationships: niche separation along the environmental gradients occurs primarily within genera, i.e. at the species level. Regarding the multispecies spatial patterns independent of environment, slight differences are observed after removing the effect of genus (fig. 5 and 8). On the first factor, of the three floristic regions identified previously, the southeastern region remains unchanged whereas the others are less clear, being divided into different subregions (fig. 8a). On the second factor, the pattern is almost unchanged, with the floristic homogeneity of the middle elevation slopes being even more strongly underlined than in the previous analysis (fig. 8b).

Table 1. Contribution of higher taxonomic levels to species ecological ranges. Canonical correlation coefficient (r) and percentage of total variance (% V) of the first two factors compared for three analyses: CANCOR between the occurrences-by-species and occurrences-by-environment tables (SE), and same analysis after partialling out the effect of genus (SE-G) or family (SE-F) (see text for further details).

Axis	Analysis	SE	SE-G	SE-F
1	r	0.803	0.651	0.756
	% V	57.85	59.48	58.52
2	r	0.573	0.453	0.539
	% V	29.51	28.87	29.02

Table 2. Contribution of higher taxonomic levels to species geographical ranges. Canonical correlation coefficient (r) and percentage of total variance (% V) of the first two factors compared for three analyses: Partial CCTSA removing the effect of environmental variables (Sxy), and same analysis after partialling out the effect of genus (Sxy-G) or family (Sxy-F) (see text for further details).

Axis	Analysis	Sxy	Sxy-G	Sxy-F
1	r	0.483	0.392	0.459
	% V	26.92	29.28	27.02
2	r	0.466	0.356	0.444
	% V	25.06	24.24	25.29

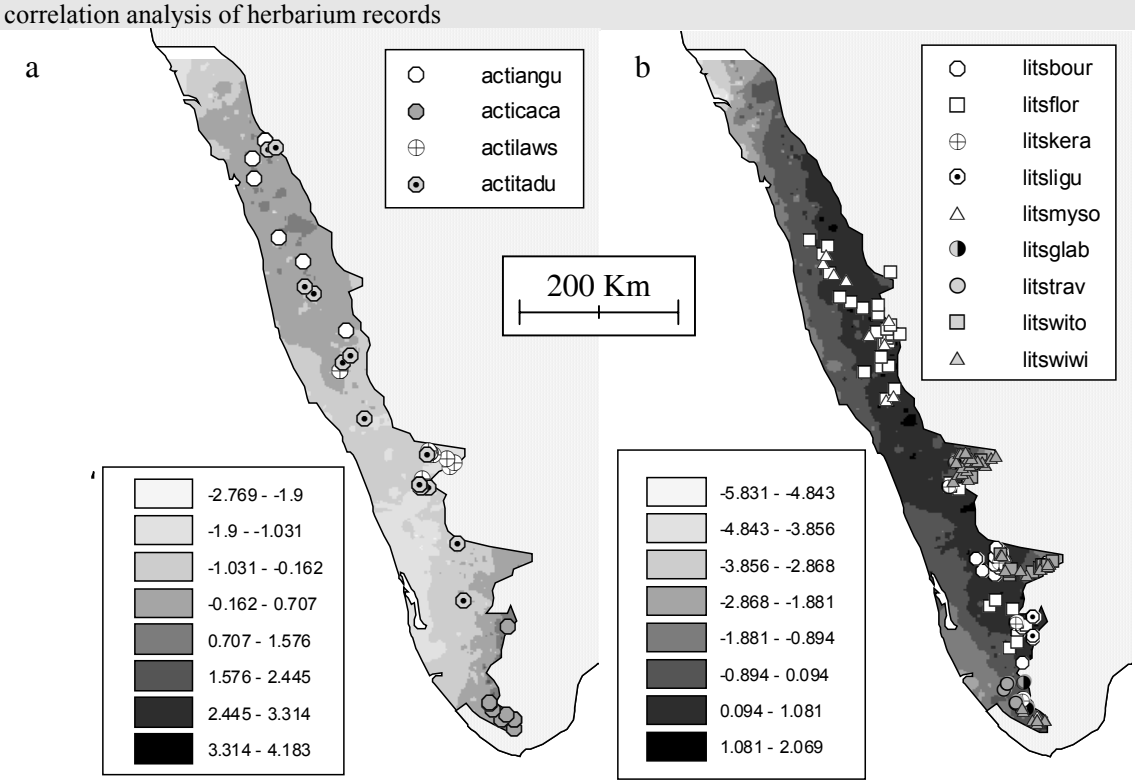


Figure 8: Multispecies spatial patterns obtained by partial CCTSA removing the effect of environment and genus. Mapping of canonical scores by contour curves computed over the 500 nearest neighbours, and location of species within genera that have high and contrasted factorial scores. a) first factor, b) second factor. The colours reflect variation in occurrence scores (cf. insert). See appendix for species names.

In order to compare the multispecies spatial patterns within genera, we joined the canonical scores of congeneric species by a polygon (fig. 9). The size of the polygon decreases with the overlapping of geographical ranges among congeneric species, whereas its shape reflects the diversity of species ranges (congeneric species may have narrow latitudinal extents and/or narrow longitudinal extents). Four types of patterns were identified: (i) genera represented by a large polygon, such as *Diospyros*, *Garcinia* and *Syzygium*, (ii) genera with a large variance of species scores on the first factor, such as *Actinodaphne*, *Goniothalamus* and *Symplocos*, (iii) genera with a large variance of species scores on the second factor, such as *Litsea* and *Microtropis*, and (iv) genera represented by a small polygon, such as *Psychotria*.

These multispecies spatial patterns were visualized by mapping the occurrences of congeneric species with contrasted factorial scores. For instance, if one considers the genus *Actinodaphne*, at the positive end of the first factor one finds *A. angustifolia*, a species occupying the northern part of the Ghats, and *A. campanulata* var. *campanulata*, a species restricted to the southeast, while at the negative end one finds *A. lawsonii*, a species restricted to the central part of the Ghats, essentially the Nilgiri Mountains, and *A. tadulingami*, a species present both in the northern and in the central regions but absent from the southeastern area (fig. 8a). If one considers the genus *Litsea*, at the positive end of the second factor one finds *L. bourdillonii*, *L. floribunda*, *L. keralana*, *L. ligustrina* and *L. mysorensis*, five species occupying the middle elevation slopes of the Ghats, while at the negative end one finds *L. glabrata*, *L. wightiana* var. *tomentosa* and *L. wightiana* var. *wightiana*, three taxa occurring on the tops of the Ghats, and *L. travancorica*, a species occupying the low elevation slopes of the southern part of the Ghats (fig. 9b).

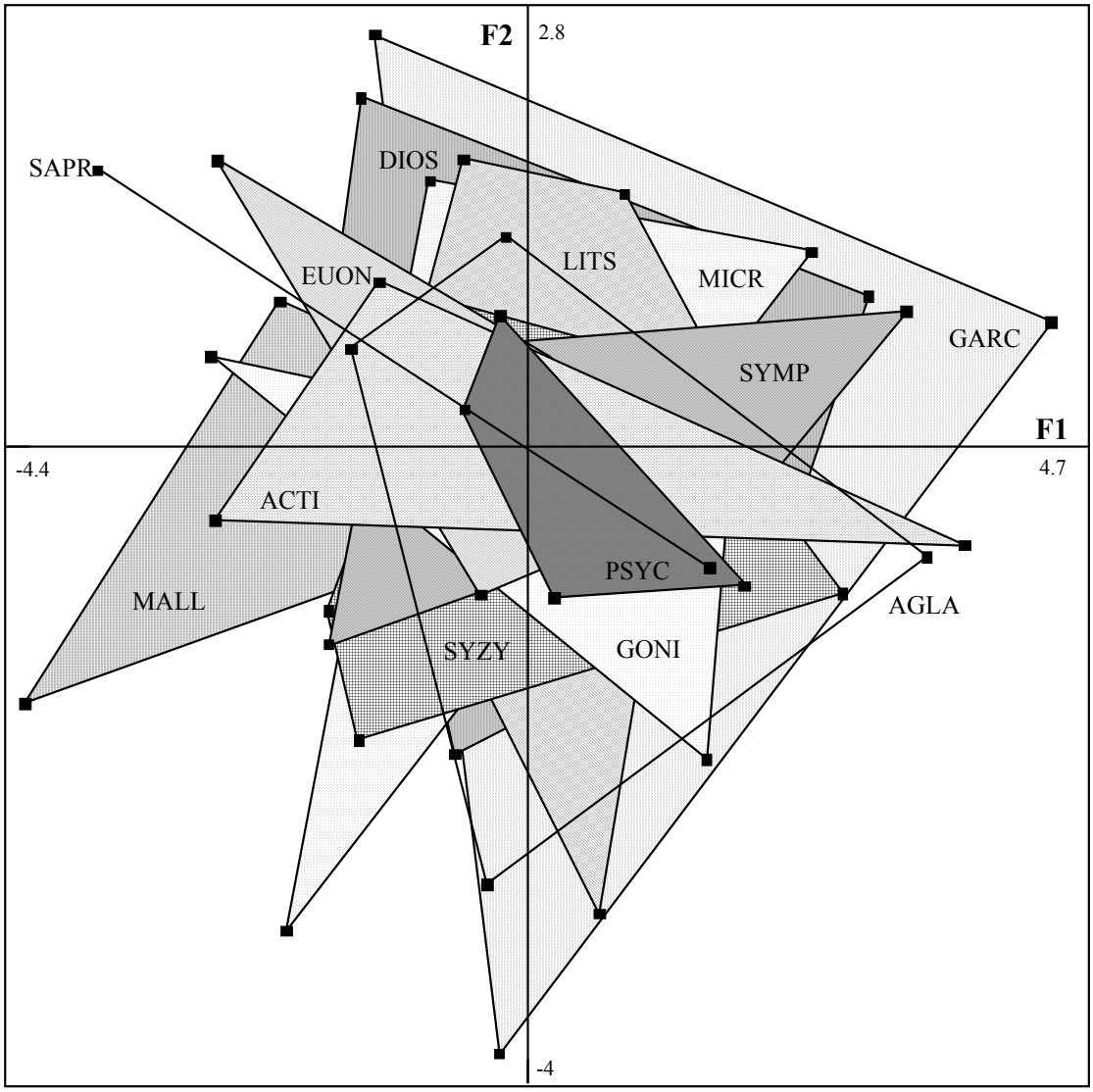


Figure 9: Species factorial map obtained by a partial CCTSA removing the effect of environment and genus. The species represented at the centroid of their occurrences (black squares) are joined within each genus. The size of the polygon reflects the importance of spatial segregation among species belonging to the same genus. ACTI, *Actinodaphne*; AGLA, *Aglaia*; DIOS, *Diospyros*; EUON, *Euonymus;* GARC, *Garcinia*; GONI, *Goniothalamus*; LITS, *Litsea*; MALL, *Mallotus*; MICR, *Microtropis*; PSYC, *Psychotria*; SAPR, *Saprosma*; SYMP, *Symplocos*; SYZY, *Syzygium*.

Environmental profiles of the endemic tree families

Another way to assess the contribution of family to the observed species ranges is to analyze family environmental profiles. This was done by CANCOR of occurrences-by-families and occurrences-by-environment tables.

The analysis produced two highly significant factors representing 93% of total variance ($p < 0.001$). As expected, the correlation between taxa and environment was lower for families than for species: the canonical correlation coefficients were 0.548 and 0.269. The patterns obtained for both analyses were similar (fig. 6 and 10).

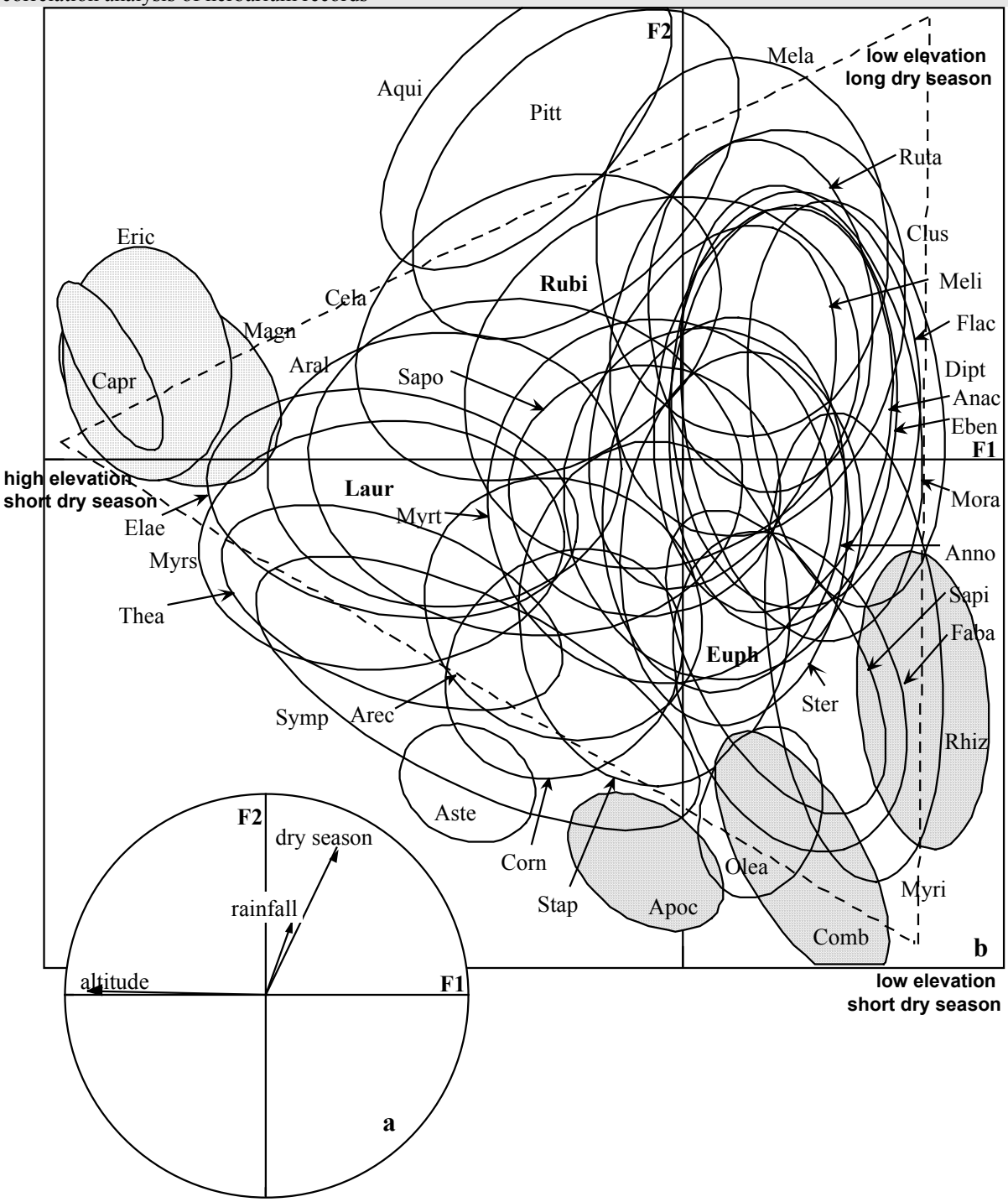


Figure 10: Results of CANCOR of occurrences-by-families and occurrences-by-environment. a) correlation of the environmental variables with the first two canonical factors, b) representation of the 38 families by ellipses containing 50% of their occurrences. Anac, Anacardiaceae; Anno, Annonaceae; Apoc, Apocynaceae; Aqui, Aquifoliaceae; Aral, Araliaceae; Arec, Arecaceae; Aste, Asteraceae; Capr, Caprifoliacae; Cela, Celastraceae; Clus, Clusiaceae; Comb, Combretaceae; Corn, Cornaceae; Dipt, Dipterocarpaceae; Eben, Ebenaceae; Elae, Elaeocarpaceae; Eric, Ericaceae; Euph, Euphorbiaceae; Faba, Fabaceae; Flac, Flacourtiaceae; Laur, Lauraceae; Magn, Magnoliaceae; Mela, Melastomataceae; Meli, Meliaceae; Mora, Moraceae; Myri, Myristicaceae; Myrs, Myrsinaceae; Myrt, Myrtaceae; Olea, Oleaceae; Pitt, Pittosporaceae; Rhiz, Rhizophoraceae; Rubi, Rubiaceae; Ruta, Rutaceae; Sapi, Sapindaceae; Sapo, Sapotaceae; Stap, Staphylaceae; Ster, Sterculiaceae; Symp, Symplocaceae; Thea, Theaceae.

Within the triangle formed by the projection of occurrences on the first canonical plane, the overlap of family ecological profiles is not uniform and indicates that most of the

families are present at low to medium elevation under a short to medium dry season (fig. 10b). The narrow profiles observed at the tips of the triangle often correspond to 'monospecific' families (i.e. families with only one tree species endemic to the Western Ghats), such as Caprifoliaceae, Ericaceae and Magnoliaceae at high elevation, and Apocynaceae, Combretaceae and Rhizophoraceae at low elevation under a short dry season. Conversely, the families having the largest environmental profiles correspond to the most species-rich families, notably Euphorbiaceae, Lauraceae and Rubiaceae, suggesting a link between clade diversification and species niche differentiation.

DISCUSSION

Distribution of endemism in the Western Ghats

The comparison of the pattern identified by the first factor of CCTSA (fig. 4a) with variation in altitude (fig. 2a) indicates that the change in endemic species composition is mainly driven by the altitudinal gradient and the related decrease in temperature, while the homogenization of floristic composition observed in the northwestern part of the study area can be explained by the effect of an increase in dry season length being counterbalanced by the effect of an increase in the amount of occult precipitation with elevation (fig. 2b). The second factor emphasizes the concentration of endemic species in the southeastern part of the Ghats and the concomitant decrease of endemic species richness along the altitudinal and the dry season length gradients. Indeed, the steep gradient of β-diversity towards the coast in the southern part, can be explained by a severe decrease of endemic species richness due to the high anthropogenic pressure that characterizes the coastal area (fig. 4b).

Biogeographic history of the Western Ghats

The analysis of family environmental profiles by CANCOR can be interpreted in the light of the biogeographic history of the Western Ghats. Indeed, the fact that most of the families are present at low to medium elevation under a short dry season (fig. 10) indicates that this habitat is associated with a larger pool of endemic evergreen species than high elevation areas or lowlands with a long dry season, and this could be related to its older age, as compared to the other habitats (cf. Taylor et al. 1990). After the uplift of the Western Ghats, wet evergreen taxa would have been faced with a physiological barrier to occupy the new habitats, *i.e.* high elevation areas and lowlands with a long dry season. As a consequence, the current latitudinal and altitudinal gradients of endemic species richness (fig. 4) could result from these historical and evolutionary factors. A similar hypothesis has been formulated by Latham and Ricklefs (1993) regarding the northern temperate tree flora.

The environmental profiles of the 'monospecific' families can be related to the Quaternary history of the Western Ghats. On the one hand, families restricted to high elevation areas (> 1800 m) have a Laurasian affinity (cf. Raven and Axelrod 1974), so these species most probably migrated from Himalaya during glacial periods and took refuge at the highest elevations of the Ghats, as postulated by Vishnu-Mittre (1974) for the Ericaceae *Rhododendron nilagiricum* (fig. 10). Indeed, the Western Ghats experienced a cooler and drier climate during the glacial cooling in the Quaternary (cf. Rostek et al. 1997). On the other hand, families restricted to low and middle elevations with a short dry season have a Gondwanan affinity, so these species may be relictual or recently evolved taxa, with the southern part of the Ghats representing a former refuge for species of the wet evergreen forest, or alternatively an active centre of speciation (cf. De Franceschi 1993). Indeed, summer

rainfall would have persisted in this region in the Quaternary (Van Campo 1986), and the topographic complexity of the middle elevation reliefs of the southern part of the Ghats could have favoured species survival as well as species formation (cf. Fjeldsa and Lovett 1997).

Speciation processes in the endemic tree flora of the Western Ghats

The analysis of multispecies spatial patterns by partial CCTSA allowed the identification of three groups of species with a narrow latitudinal range that define three floristic areas, with the strongest floristic differentiation observed between the southeastern and the central part of the study area (fig. 5a). Variation in the three environmental factors measured here shows that the northern region is characterized by a long dry season (fig. 2b), while the central region is located around Palghat Pass, an interruption in the relief of the Ghats isolating Nilgiri from Palni Mountains (fig. 2a), and the southern region mainly corresponds to the eastern side of the Ghats, also characterized by a low total annual amount of rainfall (fig. 2c). So the existence of distinct endemic floras can be explained by physical and/or climatic barriers that would have limited species migration or adaptation, also enhancing range contraction and species extinction processes.

The second pattern identified by this analysis concerns the floristic homogeneity of the western slopes of the Ghats, in contrast to the coastal area and the top of Palni and Nilgiri Mountains (fig. 5b). As shown by the overlap in *Litsea* species ranges at middle elevation (fig. 8b), this pattern seems to result from the intermediate position of the slopes of the Ghats along the altitudinal gradient. This allows the co-existence of species from the lowlands or from the tops of the Ghats with species typical of middle elevation forests, a phenomenon emphasized by the decrease in endemic species richness towards the coast and the tops of the Ghats (fig. 4b). So the middle elevation slopes of the Ghats appear as an area of species accumulation rather than a centre of speciation.

Allopatric speciation

The analysis of multispecies spatial patterns within genera provides an insight into the origin of the three floristic regions. Indeed, the limits of the southeastern region remain unchanged while the other regions become less clearly delimited (fig. 8a). This result suggests a general phenomenon of geographical vicariance for evergreen species between the eastern and the western sides of the Ghats, whereas independent adaptations to the long dry season would have occurred on the western side, as shown by the relative distributions of *Actinodaphne* species (fig. 8a). The phylogenetic analyses of Indian Sapotaceae and Ebenaceae by De Franceschi (1993) are consistent with the hypothesis of allopatric speciation: *Diospyros foliolosa* and *D. paniculata* would derived from a widely distributed ancestor that underwent differential selection on each side of the Western Ghats.

Adaptive radiation

The analysis of species environmental ranges revealed the importance of species niche separation along the altitudinal and the dry season length gradients (fig. 6). This general feature holds among congeneric species, suggesting a link between speciation and niche evolution. While altitude appears as the main factor of species niche separation within *Syzygium* (fig. 7c), the length of the dry season plays this role within *Diospyros*, primarily located at low elevation (fig. 10b), and both factors contribute to species niche separation within *Litsea*. The steepness of the environmental gradients in the Western Ghats could have favoured the diversification of the evergreen tree flora by parapatric or sympatric speciation (cf. Endler 1982, Hodges and Arnold 1994, Schneider et al. 1999).

However, the comparison of congeneric species spatial patterns shows that the hypothesized phenomena of geographical and environmental vicariances have not affected all genera equally (fig. 9). Indeed, the six species of *Psychotria* have largely overlapping ranges, compared with the five species of *Garcinia* (fig. 12). This means that diversification of congeners is not necessarily correlated with species separation in physical or environmental spaces. Several factors could explain these discrepancies. For instance, sympatric speciation could have occurred by polyploidy. An alternative hypothesis would be a high dispersal ability or low competition among congeners, as noted for *Psychotria* (Valladares et al. 2000).

Phylogenetic analyses would be necessary to test the hypotheses formulated above to explain the diversification of the endemic tree flora of the Western Ghats, notably by plotting species ecological profiles and species geographical ranges on the phylogenetic trees. The link between speciation and niche evolution could then be tested (cf. Barraclough et al. 1999). The data available for *Diospyros* suggest that the Indian species do not form a monophyletic group (De Franceschi 1993). If this were true for the other genera as well, Indian endemic congeneric species should be split into several groups of closely related taxa, but this would not alter the patterns identified in the present study. More information is also needed regarding the genetics and the life-history traits of Indian taxa to investigate speciation processes. For instance, De Franceschi (1993) suggests that the adaptive flexibility of *Diospyros* could result from its high number of chromosomes allowing frequent rearrangements. Lastly, the mean phylogenetic relatedness of taxa could be assessed along the environmental gradients of the Western Ghats, following Webb (2000) who noted that species of tropical tree communities in a stressful environment are less closely related than species of communities in a favorable environment.

Methodological issues: perspectives

Since its first use in ecology (Austin 1968), CANCOR has been a neglected ordination method, notably because of collinearity problems and related difficulties in presentation and interpretation (ter Braak 1990). Our analyses confirm the value of CANCOR as the suitable method for collection data analysis. Conversely, herbarium records offer the most accurate information about species distribution, provided that they are numerous and evenly distributed in the area under study. Moreover, such data could be used to analyze the correlation between biological attributes of taxa and environmental factors, allowing the identification of functional diversity patterns (cf. Lavorel et al. 1999). More generally, the unifying framework of CANCOR would be especially suitable for assessing the relative roles of biological and environmental factors in the variation of species-richness between clades (cf. Barraclough et al. 1998).

ACKNOWLEDGEMENTS

We thank D. Chessel for his valuable contribution to the statistical part of this study, and Brian Enquist, Susan Mazer and Horacio Paz for helpful comments on the manuscript. We are grateful to B. R. Ramesh, the French Institute of Pondicherry and its successive directors who permitted us to use the data on the endemic tree species of the Western Ghats.

REFERENCES

Austin, M. P. 1968. An ordination study of a chalk grassland community. – J. Ecol. 56: 739-757.

Bagnouls, F. and Gaussen, H. 1953. Saison sèche et régime xérothermique. Documents pour les cartes des productions végétales. - Toulouse, France.

Barraclough, T. G., Vogler, A. P. and Harvey, P. H. 1998. Revealing the factors that promote speciation. – Phil. Trans. R. Soc. Lond. B 353: 241-249.

Barraclough, T. G. and Vogler, A. P. 2000. Detecting the geographical pattern of speciation from species-level phylogenies. – Am. Nat. 155: 419-434.

Böhning-Gaese, K. 1997. Determinants of avian species richness at different spatial scales. - J. Biogeogr. 24: 49-60.

Bolognini, G. and Nimis, P. L. 1993. Phytogeography of Italian deciduous oak woods based on numerical classification of plant distribution ranges. - J. Veg. Sci. 4: 847-860.

Bonn, A. and Schröder, B. 2001. Habitat models and their transfer for single- and multi-species groups: a case study of carabids in an alluvial forest. – Ecography 24: 483-496.

Bourgeon, G. 1989. Explanatory booklet on the reconnaisance soil map of forest area (Western Karnataka and Goa). - Institut Français de Pondichéry, India.

Brown, J. H. 1995. Macroecology. - The University of Chicago Press.

De Franceschi, D. 1993. Phylogénie des Ebénales. Analyse de l'ordre et origine biogéographique des espèces indiennes. - Institut Français de Pondichéry, India.

Donald, P. F. and Fuller, R. J. 1998. Ornithological atlas data: a review of uses and limitations. - Bird Study 45: 129-145.

Endler, J. A. 1982. Pleistocene forest refuges: fact or fancy? - In: Prance, G. T. (ed.), Biological Diversification in the Tropics. Columbia Univ. Press, pp. 641-657.

Fjeldsa, J. and Lovett, J. C. 1997. Geographical patterns of old and young species in African forest biota: the significance of specific montane areas as evolutionary centres. - Biodiversity and Conservation 6: 325-346.

Gaston, K. J. 1994. Measuring geographic range sizes. - Ecography 17: 198-205.

Gauch, H.G. Jr. and Wentworth, T. R. 1976. Canonical correlation analysis as an ordination technique. - Vegetatio 33: 17-22.

Gittins, R. 1985. Canonical analysis, a review with applications in ecology. - Springer-Verlag.

Gittins, R. 1968. Trend-surface analysis of ecological data. – J. Ecol. 56: 845-869.

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

Guisan, A., Weiss, S. B. and Weiss, A. D. 1999. GLM versus CCA spatial modeling of plant species distribution. - Plant Ecology 143, 107-122.

Hawksworth, D. L. (ed.). 1995. Biodiversity, measurement and estimation. - Chapman and Hall.

Heikkinen, R. K. 1996. Predicting patterns of vascular plant species richness with composite variables: a meso-scale study in Finnish Lapland. - Vegetatio 126: 151-165.

Heywood, V. H. and Watson, R. T. (eds). 1995. Global biodiversity assessment. Cambridge University Press.

Hill, M. O. 1991. Patterns of species distribution in Britain elucidated by canonical correspondence analysis. - J. Biogeogr.18: 247-255.

Hodges, S. A. and Arnold, M. L. 1994. Floral and ecological isolation between *Aquilegia formosa* and *Aquilegia pubescens*. – Proc. Nat. Acad. Sci. USA 91: 2493-2496.

Hotelling, H. 1936. Relations between two sets of variates. - Biometrika 28: 321-377.

Knox, E. B., and Palmer, J. B. 1995. Chloroplast DNA variation and the recent radiation of the giant senecios (Asteraceae) on the tall mountains of eastern Africa. - Proc. Nat. Acad. Sci. USA 92: 10349-10353.

James, F. C. and McCulloch, C. E. 1990. Mutivariate analysis in ecology and systematics: panacea or pandora's box? - Ann. Rev. Ecol. Syst. 21: 129-166.

Latham, R. E. and Ricklefs, R. E. 1993. Continental comparisons of temperate-zone tree species diversity. – In: Ricklefs, R. E. and Schluter, D. (eds.), Species diversity in ecological communities. The University of Chicago Press, pp. 294-314.

Lavorel, S., Rochette, C. and Lebreton, J.-D. 1999. Functional groups for response to disturbance in Mediterranean old fields. - Oikos 84: 480-498.

Lebart, L., A. Morineau and Warwick, K. 1984. Multivariate Descriptive Statistical Analysis. - Wiley.

Lebreton, J.-D., Chessel, D., Prodon, R., and Yoccoz, N. 1988. L'analyse des relations espèces-milieu par l'analyse canonique des correspondances. I. Variables de milieu quantitatives. - Acta Œcologica 9: 53-67.

Legris, P. and Meher-Homji, V. M. 1982. History of India's flora and vegetation. - Scientific Reviews on Arid Zone Research 1:145-171.

Losos, J. B., T. R. Jackman, A. Larson, K. de Queiroz and Rodriguez-Schettino, L. 1998. Contingency and determinism in replicated adaptive radiations of island lizards. - Science 279: 2115-2118.

Manly, B. J. F. 1991. Randomization and Monte Carlo methods in biology. - Chapman and Hall.

Méot, A., P. Legendre and Borcard, D. 1998. Partialling out the spatial component of ecological variation: questions and propositions in the linear modelling framework. - Environmental and Ecological Statistics 5: 1-27.

Mourelle, C. and Ezcurra, E. 1996. Species richness of Argentine cacti: a test of biogeographic hypotheses. - J. Veg. Sci. 7: 667-680.

Myers, N., R. A. Mittermeier, C. G. Mittermeier, G. A. B. da Fonseca and Kent, J. 2000. Biodiversity hotspots for conservation priorities. - Nature 403: 853-858.

Parmesan, P., N. Ryrholm, C. Stefanescu, J. K. Hill, C. D. Thomas, H. Descimon, B. Huntley, L. Kaila, J. Kullberg, T. Tammaru, W. J. Tennent, J. A. Thomas, and Warren, M. 1999. Poleward shifts in geographical ranges of butterfly species associated with regional warming. - Nature 399: 579-583.

Pascal, J.-P. 1982. Bioclimates of the Western Ghats at 1:500 000 (2 sheets). - Institut Français de Pondichéry, India.

Pascal, J.-P. 1988. Wet evergreen forests of the Western Ghats of India: ecology, structure, floristic composition and succession. - Travaux de la Section Scientifique et Technique 20 bis, Institut Français de Pondichéry, India.

Pascal, J.-P. 1992. Forest map of South India (1 : 250 000). Bangalore-Salem sheet. - Karnataka Forest Dept., Tamil Nadu Forest Dept. and Institut Français de Pondichéry, India.

Pélissier, R., S. Dray and Sabatier, D. Within-plot relationships between tree species occurrences and hydrological soil constraints: an example in French Guiana investigated through correlation analysis. - Plant Ecology (*in press*).

Peterson, A. T., J. Soberón, and Sánchez-Cordero, V. 1999. Conservatism of ecological niches in evolutionary time. - Science 285: 1265-1267.

Ponder, W. F., G. A. Carter, P. Flemons, and Chapman, R. R. 2001. Evaluation of museum collection data for use in biodiversity assessment. Cons. Biol. 15: 648-657.

Ramesh, B. R., and Pascal, J.-P. 1991. Distribution of endemic arborescent evergreen species in the Western Ghats. - In: Kerala Forest Dept., Symposium on rare, endangered and endemic plants of Western Ghats. Thiruvananthapuram, India, pp. 20-29.

Ramesh, B. R., and 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, India.

Ramesh, B. R., D. de Franceschi, and Pascal, J.-P. 1997. Forest map of South India (1: 250 000). Thiruvananthapuram-Tirunelveli sheet. - Kerala Forest Dept., Tamil Nadu Forest Dept., Kerala Forest Research Institute and Institut Français de Pondichéry, India.

Raven, P. H. and Axelrod, D. I. 1974. Angiosperm biogeography and past continental movements. - Ann. Mo. Bot. Gard. 61: 539-673.

Ricklefs, R. E. and Schluter, D. (eds). 1993. Species diversity in ecological communities. - The University of Chicago Press.

Rostek, F., E., Bard, L. Beaufort, C. Sonzogni and Ganssen, G. 1997. Sea surface temperature and productivity records the past 240 kyr in the Arabian Sea. - Deep-Sea Research II 44: 1461-1480

Schluter, D. 1996. Ecological causes of adaptive radiation. – Am. Nat. 148: S40-S64.

Schneider, C. J., Smith, T. B., Larison, B. and Moritz, C. 1999. A test of alternative models of diversification in tropical rainforests: ecological gradients vs. rainforest refugia. – Proc. Nat. Acad. Sci. USA 96:13869-13873.

Shaffer, H. B., R. N. Fisher, and Davidson, C. 1998. The role of natural history collections in documenting species declines. – Trends Ecol. Evol. 13:27-30.

Skov, F. and Borchsenius, F. 1997. Predicting plant species distribution patterns using simple climatic parameters: a case study of Ecuadorian palms. - Ecography 20: 347-355.

Stockwell, D. and Peters, D. 1999. The GARP modeling system: problems and solutions to automated spatial prediction. - International Journal of Geographical Information Science 13:143-158.

Taylor, C. M., and Gotelli, N. J. 1994. The macroecology of *Cyprinella*: correlates of phylogeny, body size, and geographical range. - Am. Nat. 144: 549-569.

Taylor, D. R., L. W. Aarssen, and Loehle, C. 1990. On the relationship between r/K selection and environmental carrying capacity: a new habitat templet for plant life history strategies. - Oikos 58:239-250.

ter Braak, C. J. F. 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. - Ecology 67: 1167-1179.

ter Braak, C. J. F. 1988. Partial canonical correspondence analysis. – In: Bock, H. H. (ed.), Classification and related methods of data anlysis. Elsevier Science Publishers B. V.

ter Braak, C. J. F. 1990. Interpreting canonical correlation analysis through biplots of structure correlations and weights. - Psychometrika 55: 519-531.

ter Braak, C. J. F. and Verdonschot, P. F. M.. 1995. Canonical correspondence analysis and related multivariate methods in aquatic ecology. - Aquatic Sciences 57: 255-289.

Thioulouse, J. and Chessel, D. 1992. A method for reciprocal scaling of species tolerance and sample diversity. - Ecology 73: 670-680.

Thioulouse, J., D. Chessel and Champély, S. 1995. Multivariate analysis of spatial patterns:a unified approach to local and global structures. - Environmental and Ecological Statistics 2: 1-14.

Thioulouse, J., D. Chessel, S. Dolédec and Olivier, J. M. 1997. ADE-4: a multivariate analysis and graphical display software. - Statistics and Computing 7: 75-83.

Valladares, F., Wright, S. J., Lasso, E., Kitajima, K. and Pearcy, R. W. 2000. Plastic phenotypic response to light of 16 congeneric shrubs from a Panamanian rainforest. Ecology 81: 1925-1936.

Van Campo, E. 1986. Monsoon fluctuations in two 20000-yr B.P. oxygen-isotope/pollen records off southwest India. - Quaternary Research 26: 378-388.

Vishnu-Mittre. 1974. Late Quaternary paleobotany and palynology in India — an appraisement. – In: Symposium on Late Quaternary vegetational developments in extra european areas. Birbal Sahni Institute of Palaeobotany vol. 5, pp. 16-51.

Wartenberg, D. E. 1985. Canonical trend surface analysis:a method for describing geographic pattern. - Syst. Zool. 34: 259-279.

Webb, C. O. 2000. Exploring the phylogenetic structure of ecological communities: an example for rainforest trees. - Am. Nat. 156: 145-155.

APPENDIX

List of species names identified by Canonical Correlation Analysis (fig. 5, 6, 7 and 8).

Species code	Species name	Family
actiangu	*Actinodaphne angustifolia* (Bl.) Nees	Lauraceae
acticaca	*Actinodaphne campanulata* var. *campanulata* J. Hk.	Lauraceae
actilaws	*Actinodaphne lawsonii* Gamble	Lauraceae
actitadu	*Actinodaphne tadulingamii* Gamble	Lauraceae
aglaelbo	*Aglaia elaeagnoidea* (Juss.) Benth. var. *bourdillonii* (Gamble) K.K.N. Nair	Meliaceae
aglalawi	*Aglaia lawii* (Wt.) Saldanha	Meliaceae
aporbour	*Aporosa bourdillonii* Stapf.	Euphorbiaceae
ardirhom	*Ardisia rhomboidea* Wt.	Myrsinaceae
bacccour	*Baccaurea courtallensis* M. Arg.	Euphorbiaceae
bentcoda	*Bentinckia codapanna* Berry	Arecaceae
blaccaly	*Blachia calycina* Benth.	Euphorbiaceae
blacdenu	*Blachia denudata* Benth.	Euphorbiaceae
blepmemb	*Blepharistemma membranifolia* (Miq.) Ding Hou	Rhizophoraceae
buchlanc	*Buchanania lanceolata* Wt.	Anacardiaceae
byrstetr	*Byrsophyllum tetrandrum* (Bedd.) J. Hk. ex Bedd.	Rubiaceae
cantnene	*Canthium neilgherrense* Wt. var. *neilgherrense*	Rubiaceae
casewyna	*Casearia wynadensis* Bedd.	Flacourtiaceae
cinnripa	*Cinnamomum riparium* Gamble	Lauraceae
cinnwigh	*Cinnamomum wightii* Meissn.	Lauraceae
cleimala	*Cleistanthus malabaricus* M. Arg.	Euphorbiaceae
diosangu	*Diospyros angustifolia* (Miq.) Kosterm.	Ebenaceae
diosassi	*Diospyros assimilis* Bedd.	Ebenaceae
diosbarb	*Diospyros barberi* Ramas.	Ebenaceae
diosbour	*Diospyros bourdillonii* Brand.	Ebenaceae
dioscand	*Diospyros candolleana* Wt.	Ebenaceae
diosfoli	*Diospyros foliolosa* Wall. ex DC.	Ebenaceae
diosghat	*Diospyros ghatensis* Ramesh & De Franceschi	Ebenaceae
dioshumi	*Diospyros humilis* Bourd.	Ebenaceae
diosnila	*Diospyros nilagirica* Bedd.	Ebenaceae
diospani	*Diospyros paniculata* Dalz.	Ebenaceae
diosprur	*Diospyros pruriens* Dalz.	Ebenaceae

diossald	*Diospyros saldanhae* Kosterm.	Ebenaceae
diptbour	*Dipterocarpus bourdilloni* Brandis	Dipterocarpaceae
drypconf	*Drypetes confertiflorus* (J. Hk.) Pax & Hoffm.	Euphorbiaceae
elaerecu	*Elaeocarpus recurvatus* Corner	Elaeocarpaceae
eugemacr	*Eugenia macrosepala* Duthie	Myrtaceae
euonangu	*Euonymus angulatus* Wt.	Celastraceae
euoncren	*Euonymus crenulatus* Wall. ex Wt. & Arn.	Celastraceae
flacmont	*Flacourtia montana* Grah.	Flacourtiaceae
garcindi	*Garcinia indica* (Thouras) Choisy	Clusiaceae
garctrav	*Garcinia travancorica* Bedd.	Clusiaceae
garcwigh	*Garcinia wightii* T. Andr.	Clusiaceae
glocbour	*Glochidion bourdillonii* Gamble	Euphorbiaceae
glocjohn	*Glochidion johnstonei* J. Hk.	Euphorbiaceae
glocneil	*Glochidion neilgherrense* Wt.	Euphorbiaceae
gluttrav	*Gluta travancorica* Bedd.	Anacardiaceae
gonirhyn	*Goniothalamus rhynchantherus* Dunn.	Annonaceae
gordobtu	*Gordonia obtusa* Wall. ex Wt. & Arn.	Theaceae
gymncana	*Gymnacranthera canarica* (King) Warb.	Myristicaceae
holibedd	*Holigarna beddomei* J. Hk.	Anacardiaceae
homatrav	*Homalium travancoricum* Bedd.	Flacourtiaceae
hopepong	*Hopea ponga* (Dennst.) Mabb.	Dipterocarpaceae
hoperaco	*Hopea racophloea* Dyer	Dipterocarpaceae
hopeutil	*Hopea utilis* (Bedd.) Bole	Dipterocarpaceae
humbbrun	*Humboldtia brunonis* Wall.	Fabaceae
humbdecu	*Humboldtia decurrens* Bedd. ex Oliv.	Fabaceae
humbvahl	*Humboldtia vahliana* Wt.	Fabaceae
hydnmama	*Hydnocarpus macrocarpa* Bedd. Warb. ssp. *macrocarpa*	Flacourtiaceae
isonperr	*Isonandra perrottetiana* DC.	Sapotaceae
ixornoto	*Ixora notoniana* Wall. ex G. Don.	Rubiaceae
ixorpoly	*Ixora polyantha* Wt.	Rubiaceae
lasirost	*Lasianthus rostratus* Wt.	Rubiaceae
litovenu	*Litosanthes venulosus* (Wt. & Arn.) Deb. & Gang.	Rubiaceae
litsbour	*Litsea bourdillonii* Gamble	Lauraceae
litscori	*Litsea coriacea* J. Hk.	Lauraceae
litsflor	*Litsea floribunda* Gamble	Lauraceae

litsglab	*Litsea glabrata* J. Hk.	Lauraceae
litskera	*Litsea keralana* Kosterm.	Lauraceae
litslaev	*Litsea laevigata* Gamble	Lauraceae
litsligu	*Litsea ligustrina* J. Hk.	Lauraceae
litsmyso	*Litsea mysorensis* Gamble	Lauraceae
litsstoc	*Litsea stocksii* J. Hk.	Lauraceae
litstrav	*Litsea travancorica* Gamble	Lauraceae
litswito	*Litsea wightiana* (Nees) J. Hk. var. *tomentosa* Meissn.	Lauraceae
litswiwi	*Litsea wightiana* (Nees) J. Hk. var. *wightiana*	Lauraceae
mallatro	*Mallotus atrovirens* J. Hk.	Euphorbiaceae
mammsuri	*Mammea suriga* (Buch.-Ham. ex Roxb.)	Clusiaceae
maytroth	*Maytenus rothiana* (Walp.) Ramam.	Celastraceae
memeheyn	*Memecylon heyneanum* Benth. ex Wt. & Arn.	Melastomataceae
memetalb	*Memecylon talbotianum* Brandis	Melastomataceae
michnila	*Michelia nilagirica* Zenk.	Magnoliaceae
micrmicr	*Microtropis microcarpa* Wt.	Celastraceae
myrifama	*Myristica fatua* Houtt. var. *magnifica* (Bedd.) Sinclair	Myristicaceae
neolfisc	*Neolitsea fischeri* Gamble	Lauraceae
ochrmiss	*Ochrinauclea missionis* (Wall. ex G. Don.) Ridsd.	Rubiaceae
octotrav	*Octotropis travancorica* Bedd.	Rubiaceae
ormotrav	*Ormosia travancorica* Bedd.	Fabaceae
palabour	*Palaquium bourdilloni* Brand.	Sapotaceae
phaemala	*Phaeanthus malabaricus* Bedd.	Annonaceae
pithgrac	*Pithecolobium gracile* Bedd.	Fabaceae
pittdasy	*Pittosporum dasycaulon* Miq.	Pittosporaceae
pittneel	*Pittosporum neelgherrense* Wt. & Arn.	Pittosporaceae
popobedd	*Popowia beddomeana* J. Hk. & Thoms.	Annonaceae
psycdalz	*Psychotria dalzellii* J. Hk.	Rubiaceae
psycglob	*Psychotria globicephala* Gamble	Rubiaceae
psycnudi	*Psychotria nudiflora* Wt. & Arn.	Rubiaceae
psyctrun	*Psychotria truncata* Wall.	Rubiaceae
pterreti	*Pterospermum reticulatum* Wt. & Arn.	Sterculiaceae
pterrubi	*Pterospermum rubiginosum* Heyne ex Wt. & Arn.	Sterculiaceae
rhodnila	*Rhododendron nilagiricum* Zenk.	Ericaceae
sagegran	*Sageraea grandiflora* Dunn.	Annonaceae

sagelaur	*Sageraea laurifolia* (Grah.) Blatt. & Mc. Cann.	Annonaceae
saprcory	*Saprosma corymbosum* Bedd.	Rubiaceae
schecapi	*Schefflera capitata* (Wt. & Arn.) Harms	Araliaceae
scherace	*Schefflera racemosa* Harms	Araliaceae
semeauri	*Semecarpus auriculata* Bedd.	Anacardiaceae
semetrav	*Semecarpus travancorica* Bedd.	Anacardiaceae
sympfoli	*Symplocos foliosa* Wt.	Symplocaceae
sympmama	*Symplocos macrocarpa* Wt. ex Cl. ssp. *macrocarpa*	Symplocaceae
sympmaro	*Symplocos macrophylla* Wall. ex A. DC. ssp. *rosea* (Bedd.) Noot.	Symplocaceae
sympwyna	*Symplocos wynadense* (O. K.) Noot.	Symplocaceae
syzybent	*Syzygium benthamianum* (Wt. ex Duthie) Gamble	Myrtaceae
syzydens	*Syzygium densiflorum* Wall. ex Wt. & Arn.	Myrtaceae
syzylaet	*Syzygium laetum* (Buch.-Ham.) Gandhi	Myrtaceae
syzymala	*Syzygium malabaricum* (Bedd.) Gamble	Myrtaceae
syzymund	*Syzygium mundagam* (Bourd.) Chithra	Myrtaceae
syzyocci	*Syzygium occidentale* (Bourd.) Gandhi	Myrtaceae
syzyrama	*Syzygium rama-varma* (Bourd.) Chithra	Myrtaceae
syzytami	*Syzygium tamilnadensis* Rathak. & Chithra	Myrtaceae
tabegamb	*Tabernaemontana gamblei* Subramanyam & Henry	Apocynaceae
termtrav	*Terminalia travancorensis* Wt. & Arn.	Combretaceae
verntrav	*Vernonia travancorica* J. Hk.	Asteraceae
vibuheba	*Viburnum hebanthum* Wt. & Arn.	Caprifoliaceae

Within-plot relationships between tree species occurrences and hydrological soil constraints: an example in French Guiana investigated through canonical correlation analysis

Raphaël Pélissier[1], Stéphane Dray[2] & Daniel Sabatier[3]

Running head: Species-soil relationships in a Guianan rainforest plot

Abstract

Spatial relationships between tree species and hydrological soil constraints are analysed within a 10-ha rainforest plot at Piste de St Elie in French Guiana. We used canonical correlation analysis to cross directly the occurrence-by-species table of 4 992 individuals (d.b.h. ≥ 10 cm) belonging to 120 species with qualitative soil variables and quantitative spatial data.

Firstly, the list of species occurrences was confronted to nine soil descriptors characterising a weathering sequence from the initial well-drained ferralitic cover to transformed hydromorphic soil conditions. This analysis revealed that, apart from some specialised species restricted to the swamps that experience prolonged water saturation, the most abundant species can be ordered along two intermingled gradients of tolerance limiting their niche amplitude: a main gradient of tolerance to prolonged water saturation that appears down slope during the weathering sequence; a second gradient of less importance, displaying the species intolerant of prolonged water saturation according to their tolerance to temporary confinement of the uphill transformed soil systems due to the late appearance of a perched water-table. The results support the hypothesis that at Piste de St Elie, the constraining soil conditions imposed by surface water saturation are more important determinants for tree zonation of many tree species than water shortage.

Secondly, the list of species occurrences was confronted to a spatial data table built from a trend surface regression of the tree coordinates. This analysis indicated that soil drainage is the main structuring factor of the local multispecies spatial pattern. After partialling out the soil effect, the multispecies pattern revealed a broader scale of heterogeneity that we supposed to be linked to endogenous factors resulting from population dynamics.

Implications of the results are then discussed in the perspective of future research on tree zonation, local diversity pattern and community structuring in tropical rainforests.

Key words: canonical ordination, multispecies spatial pattern, rainforest, species tolerance, trend surface analysis, tropical soil hydromorphy.

Introduction

Complementary environmental and biotic hypotheses are commonly invoked for explaining the spatial structuring of natural ecosystems (see for example Borcard *et al.* 1992 for an introduction). In tropical rainforests, gap disturbances, species competition, topographic, and edaphic variations are among the driving forces explaining the maintenance of high tree species diversity (Denslow 1987, 1995; He *et al.* 1996; Gimaret-Carpentier *et al.* 1998a). These factors are, moreover, expected to act at various spatial scales. The influence of soil conditions (water availability, nutrient content) on the species composition have been emphasised from regional (Baillie *et al.* 1987; Ashton & Hall 1992) to landscape (Gartlan *et al.* 1986; Clark *et al.* 1998, 1999) and to local scales (Newbery & Proctor 1986; Lescure & Boulet 1985; Basnet 1992; Ter Steege *et al.* 1993; Sabatier *et al.* 1997).

In Amazonia, the soil organisation shows a currently unbalanced status regarding the present tectonic and climatic constraints, which results in a supergene weathering of the initial ferralitic cover, and consequently, in a continuous expansion of the hydromorphic and podzolic conditions (Fritsch *et al.* 1986; Lucas *et al.* 1986; Veillon & Soria-Solano 1988). The current transformation process of the initial ferralitic cover has been demonstrated at the local scale in French Guiana by studying hydrological, structural and geochemical modifications of the soil along topographical catenas (Humbel 1978; Guelhl 1984; Grimaldi & Boulet 1990; Grimaldi *et al.* 1994). This transformation can be summarised in four stages (see Fritsch *et al.* 1986 or Sabatier *et al.* 1997 for a more detailed introduction). Stage I corresponds to the thick and vertically well-drained initial ferralitic cover. At this stage, only the soils restricted to the edges of the main watercourses experience prolonged water saturation in the upper horizons (surface hydromorphy). At stage II, the profile gets thinner under mechanical and chemical erosion and a compact weathering horizon appears down-slope at very low depth (< 1 m). At stage III the weathering horizon, which is "dry to the touch" in all seasons, is present all along the slope and modifies the soil drainage from deep and vertical to superficial and lateral. Depending on the slope angle, the weathering horizon induces a more or less fleeting water saturation of the upper horizons. Stage IV marks the extension of two transformed hydromorphic systems: (i) a uphill system associated with the formation of closed depressions on the flattop hills (the so called "djougoung-pétés" in French Guiana; Fritsch *et al.* 1986), where a perched water-table that appears during the rainy season at the bottom of the profile, determines a temporary confined environment in the upper horizons; (ii) a more open downhill system where the upper horizons are more or less permanently saturated by a slow lateral water flow coming from the slope, and in connection with the main drainage axes. These transformed hydromorphic systems represent the first stages of a podzolisation process (Fritsch *et al.* 1986). The time scale of this transformation is considered to be considerably longer than the lifetime of trees (probably on the order of several thousand years; Grimaldi, M. unpublished data), and therefore affects the spatial pattern of local stand structure and composition (Lescure & Boulet 1985; Sabatier *et al.* 1997).

On another hand, ecophysiological studies carried out in plantations (Huc *et al.* 1994, Guelh *et al.* 1998, Bonal *et al*. 2000a) as well as in natural conditions (Guelh *et al.* 1998, Bonal *et al.* 2000b) in French Guiana, mainly focused on water-use efficiency (WUE), but did not enable to clearly elucidate the distribution pattern of species regarding soil conditions. For instance, the very common canopy species *Eperua falcata*, shows high WUE which should allow it to growth in conditions of reduced soil moisture (Bonal *et al.* 2000a, b), while this species is less abundant than expected on the well-drained initial ferralitic cover which exhibits low water potential in the dry season (Sabatier *et al.* 1997). There is consequently a need to clarify the species position and tolerance regarding both the weathering

transformation sequence of the initial ferralitic cover and the gradient of increasing surface hydromorphy demonstrated by Sabatier *et al.* (1997).

On the basis of a re-examination of the data from a 10-ha rainforest plot in French Guiana previously studied by Sabatier *et al.* (1997), we propose in this paper to analyse within-plot relationships between species occurrences and soil factors using direct canonical correlations (CANCOR; Hotelling 1936). This multivariate constrained ordination technique, briefly presented in a first section, is particularly well-designed for the analysis of occurrence data. As does the classical Canonical Correspondence Analysis (CCA; Ter Braak 1986, 1987) from species-by-relevés tables, CANCOR allows finding the axes that maximise the linear correlation between two sets of variables, namely a taxonomic occurrence table and a table of qualitative soil variables. Furthermore, we developed an algorithm to compute from this analysis, canonical scores leading to the dual scaling of species amplitude or tolerance (niche breadth; Carnes & Slade 1982) and habitat diversity, a strategy equivalent to reciprocal scaling proposed by Thioulouse & Chessel (1992) for species-site correspondence analysis. In a last section, CANCOR is applied to our entirely mapped data and enables to deal with local multispecies spatial pattern in a Canonical Correspondence Trend Surface Analysis (CCTSA) using an environmental table directly built from the geographical coordinates of trees (trend surface; Gittins 1968). Results are discussed with special emphasis on species tolerance, habitat diversity and multispecies spatial patterns.

Material and methods

Study site and data

The site studied is a 10-ha forest plot of the station Piste de St Elie (5°18'N; 53°3'W) in French Guiana. The forest is a lowland tropical rainforest that grows on the Armina series of the volcano-sedimentary substratum of the platform of the Guianas, in Northeast Amazonia (Milési *et al.* 1995). The site is subjected to a humid tropical climate with annual rainfall ranging from 2 500 to 4 000 mm (Boyé *et al.* 1979).

A total of 6 134 trees with diameter at breast height (d.b.h.) ≥ 10 cm and belonging to 459 taxons were mapped and identified (nomenclature follows Boggan *et al.* 1992). Three species, *Eperua falcata*, *Lecythis idatimon* and *Lecythis persistens* (species' authorities are given in Appendix) were represented by more than 300 individuals, while 400 species were represented by less than 20 individuals.

Following the method introduced by Boulet *et al.* (1982; see also Lescure & Boulet 1985), nine hydrological soil types in the 10-ha plot at Piste de St Elie, were entirely mapped (Sabatier *et al.* 1997). Height soil classes characterising the transformation of the initial ferralitic cover were considered: DVD = deep vertical drainage; Alt = weathered material (red alloterite) at a depth less than 1.2 m; SLD1 = superficial lateral drainage with "dry to the touch" character (DC) between 1 and 1.2 m; SLD2 = superficial lateral drainage with DC at less than 1 m; UhS = uphill system; UhS+DC = uphill system with DC at less than 1.2 m; DhS = downhill system; DhS+DC = downhill system with DC at less than 1.2 m. The ninth soil class, called SH for surface hydromorphy, characterises thalwegs experiencing prolonged periods of surface water saturation, while being relatively independent of the weathering process. From the superimposition of the soil and tree maps, each tree was then attached to one of the nine soil descriptors. Figure 1 gives the distribution of trees among the soil classes. The DVD class, which represents the initial ferralitic cover, is the most abundant one, while the transformed stages and the SH class are less abundant.

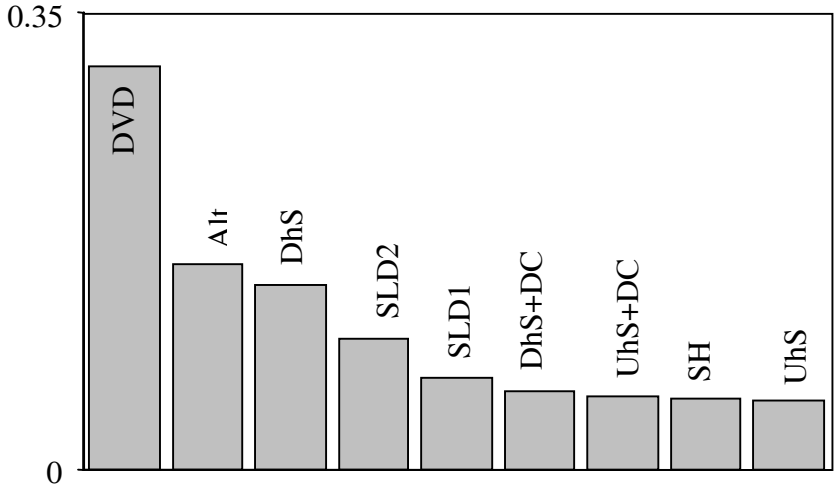


Figure 1. Frequency distribution of 6 134 trees (d.b.h. ≥ 10 cm) among 9 hydrological soil classes in a 10-ha forest plot at Piste de St Elie, French Guiana. DVD = deep vertical drainage; Alt = weathered material (red alloterite) at a depth less than 1.2 m; SLD1 = superficial lateral drainage with "dry to the touch" character (DC) between 1 and 1.2 m; SLD2 = superficial lateral drainage with DC at less than 1 m; UhS = uphill system; UhS+DC = uphill system with DC at less than 1.2 m; DhS = downhill system; DhS+DC = downhill system with DC at less than 1.2 m; SH = surface hydromorphy.

Canonical correlation analysis

Our aim is to perform an ordination of species with the help of environmental variables. When sampling units are real relevés, such as spatially independent plots, CCA (Ter Braak 1986, 1987) is a well-suited technique for that purpose which, instead of correlating *a posteriori* a set of environmental variables with the main ordination axes, extracts axes that represent the best linear combination of the environmental variables separating the mean species position (see the bibliographic review of Birks *et al.* 1996 and the recent paper of Clark *et al.* 1999 for an example in tropical forest ecology). In principle, CCA can be extended to studying within-plot species-environment relationships by partitioning the plot into more or less arbitrary contiguous quadrats of equal size, a common practice in forest ecology (*e.g.*, Newbery *et al.* 1996; Pascal & Pélissier 1996; Sabatier *et al.* 1997). However, because quadrats defined *a posteriori* are not real relevés, Gimaret-Carpentier *et al.* (1998b) stressed the interest to consider certain floristic data as lists of species occurrences, where each tree, known by its botanical name and location, is the basic statistical sampling unit. Such lists have been used for a long time in forest ecology for within-plot analysis of tree spatial patterns through methods based on inter-individual distances (e.g., Ripley 1981; Diggle 1983), but they have been rarely considered as such in ordination techniques. CANCOR (Hotelling 1936), first used in ecology by Austin (1968), gives however an appropriate framework to analyse within-plot relationships between species occurrences and environmental factors.

Let us consider a list of n trees. Its botanical name, spatial position and environment characterise each tree. The taxonomic information is arranged in a table **T** with n rows (trees) and s columns (species). For the i-th tree we have:

$$\mathbf{T}_{ij} = \begin{cases} 1 \text{ if the } i\text{ - th tree belongs to the } j\text{ - th species } (1 \leq j \leq s) \\ 0 \text{ otherwise} \end{cases}$$

Let **E** be the trees-by-environment table with the measurements or estimations of m environmental variables as columns for the n individuals (rows). Moreover, let $\mathbf{D}_n$ be the n by n diagonal matrix of weights:

$$\begin{cases} \mathbf{D}_n(i,i) = 1/n \\ \mathbf{D}_n(i,j) = 0 \end{cases} \quad \text{for } i,j \in [1,n],\ i \neq j$$

CANCOR of **T** and **E** corresponds to the analysis of the statistical triplet ($\mathbf{T}^t\mathbf{D}_n\mathbf{E}$, $(\mathbf{E}^t\mathbf{D}_n\mathbf{E})^{-1}$, $(\mathbf{T}^t\mathbf{D}_n\mathbf{T})^{-1}$) (see Escoufier 1987 for a comprehensive explanation of statistical triplet analysis). CANCOR requires numerical conditions that are well satisfied by our data set: number of trees (6 134) largely bigger than the number of species (459) and the number of environmental variables (9).

CANCOR of (**T**, **E**) consists in finding the normalised canonical variates **t*** and **e*** representing the linear combinations of each set of variables which are the most correlated (see for example Gittins 1985 for further explanations and examples in ecology). In our case, table **T** gives a taxonomic information. Therefore, rows of this table are identical for all trees of the same species, to which the canonical variate **t*** consequently gives the same scores. **e*** represents the linear combination of the variables of **E** which is the most correlated with the variables of **T**. As table **T** is engendered by a dummy variable, it follows that **e*** is a linear combination of environmental variables maximising the separation of the species centroids. This part of the analysis is known under the name of canonical variate analysis (Gittins 1985) or discriminant analysis (DA; Lebart *et al.* 1984). Lebreton *et al.* (1988) proved that DA is, in a theoretical point of view, exactly the CCA of Ter Braak (1986), but taking as statistical unit the individual instead of the site (Ter Braak & Verdonschot 1995). The averages of environmental variables by species are thus only computed with realised observations (presence), so that sampling bias caused by the use of quadrats is excluded and that the much-debated step in classical CCA, which consists in a centring of environmental variables with the species richness of the sampling units, is avoided.

Based on individual occurrences, CANCOR permits also partial analyses in which the effects of covariables are factored out (Ter Braak 1988). Considering a table $\mathbf{E_1}$ of variables of interest and a table $\mathbf{E_2}$ of covariables from which the effects must be eliminated, the principle of partial analysis consists in using the residuals of the regression of $\mathbf{E_1}$ on $\mathbf{E_2}$ in table **E**. In this paper, we performed partial analyses as explained in Méot *et al.* (1998), i.e. by constraining the residuals to be linear combinations of the variables of $\mathbf{E_1}$.

Ordination diagram

The difficulty in representing results issued from CANCOR, which concerns two sets of variables, is mentioned in several papers (Gauch & Wentworth 1976; Ter Braak 1990). However, we propose an ordination diagram that seems to be very suitable for species occurrence data.

In a geometrical point of view, CANCOR finds a vector **t*** in the species space, which is the nearest of the environment space and, simultaneously, a vector **e*** in the environment space, which is the nearest of the species space. But it is also possible to define CANCOR by finding in the environment + species space, a normalised vector **s*** which is the nearest of the two spaces. This vector is the bisector of the angle formed by the two canonical variates **t*** and **e*** and can be expressed as: $\mathbf{s}^* = (\mathbf{e}^* + \mathbf{t}^*) / \sqrt{\text{var}\ (\mathbf{e}^* + \mathbf{t}^*)}$.

The canonical score **s*** maximises simultaneously, and in the same proportion, the between-species variance and the correlation with the environmental variables. When the environmental variables are qualitative, the proposed graphical representation allows to position the occurrences with uncorrelated scores of unit variance, and to put by averaging, species and environmental categories at the centroid of their occurrences. This strategy corresponds to reciprocal scaling (Thioulouse & Chessel 1992) introduced in the context of traditional species-by-sites contingency tables and that allows representing on the same graph,

the notions of species weighted optimum and tolerance (niche width), and within- and between- sample diversity.

Computation
D. Chessel implemented CANCOR analysis in the OccurData module of ADE-4 software (Thioulouse *et al.* 1997) available with documentation at http://pbil.univ-lyon1.fr/ADE-4.html. A multivariate randomisation test (Manly 1991) was also implemented in order to test the statistical significance of the canonical axes obtained by CANCOR. It consists in a comparison of the eigenvalues (or sum of eigenvalues) obtained from the original data set, with the ones obtained after row permutations of the environmental table **E**. We used 1 000 row permutations of the environmental table.

Results

Species-soil relationships
As the rare species induced *de facto* strong correlations with the soil variables, we retained from the initial data, 4 992 trees accounting for 120 species with 10 individuals or more in the 10-ha plot. A multivariate randomisation test (Manly 1991) on the total inertia among the soil classes, computed from the centred-by-species table using the Mahalanobis norm and 1 000 permutations ($P < 0.001$), justified further investigation of the relationships between species occurrences and soil factors.

An analysis of the data set by CANCOR suggested that this relationship was clearly expressed by the two first canonical axes: the sum of the two first eigenvalues representing 63.7% of the total inertia, was highly significant ($P < 0.001$). The first canonical plane (not shown) showed outliers corresponding to the occurrences of *Tabebuia insignis* ($N = 11$) and *Euterpe oleracea* ($N = 33$) exclusively found on SH soil.

In a second step, we thus removed these two species in order to examine the ordination of the remaining species from a reduced data set of 4 948 trees belonging to 118 species. As expected, the canonical correlation coefficients were less than in the previous analysis ($r_1 = 0.380$ *vs.* 0.507; $r_2 = 0.298$ *vs.* 0.318), but the sum of the two first eigenvalues representing 54.8% of the total inertia, remained highly significant ($P < 0.001$). The dual scaling property of CANCOR allows the representation of species and soil classes at the weighted mean position of their occurrences on the first canonical plane (Fig. 2). This shows that species were scattered between the vertices of a triangle which oppose three poles of constraining soil conditions: prolonged water saturation (SH), temporary confinement during the rainy season (UhS and UhS+DC), and low water potential in the dry season (DVD) (see Fig. 4 in Sabatier *et al.* 1997).

The between-habitat distance, which corresponds to soil β-diversity, indicates that the species composition in SH differed markedly from the other soil classes, while the transformed soil complex was floristically close to the initial ferralitic cover (DVD). Computation of the multivariate among-group distances using the Mahalanobis norm, confirmed this visual impression (Table 1).

Table 1. Among-group distances (Mahalanobis norm) between nine hydrological soil classes containing 4 948 trees (d.b.h. ≥ 10 cm) belonging to 118 species in a 10-ha forest plot at Piste de St Elie, French Guiana. Key for the soil classes is the same as in Figure 1.

	SH	DVD	Alt	SLD1	SLD2	DhS+DC	DhS	UhS+DC	UhS
SH	0.00	3.66	2.77	3.33	2.96	4.07	2.35	3.59	4.14
DVD		0.00	0.50	0.89	0.84	1.09	0.91	1.18	1.29
Alt			0.00	0.67	0.59	1.05	0.49	0.95	1.23
SLD1				0.00	0.84	1.24	0.59	0.71	1.15
SLD2					0.00	0.99	0.67	0.92	1.16
DhS+DC						0.00	1.19	1.39	1.49
DhS							0.00	0.82	1.17
UhS+DC								0.00	0.83
UhS									0.00

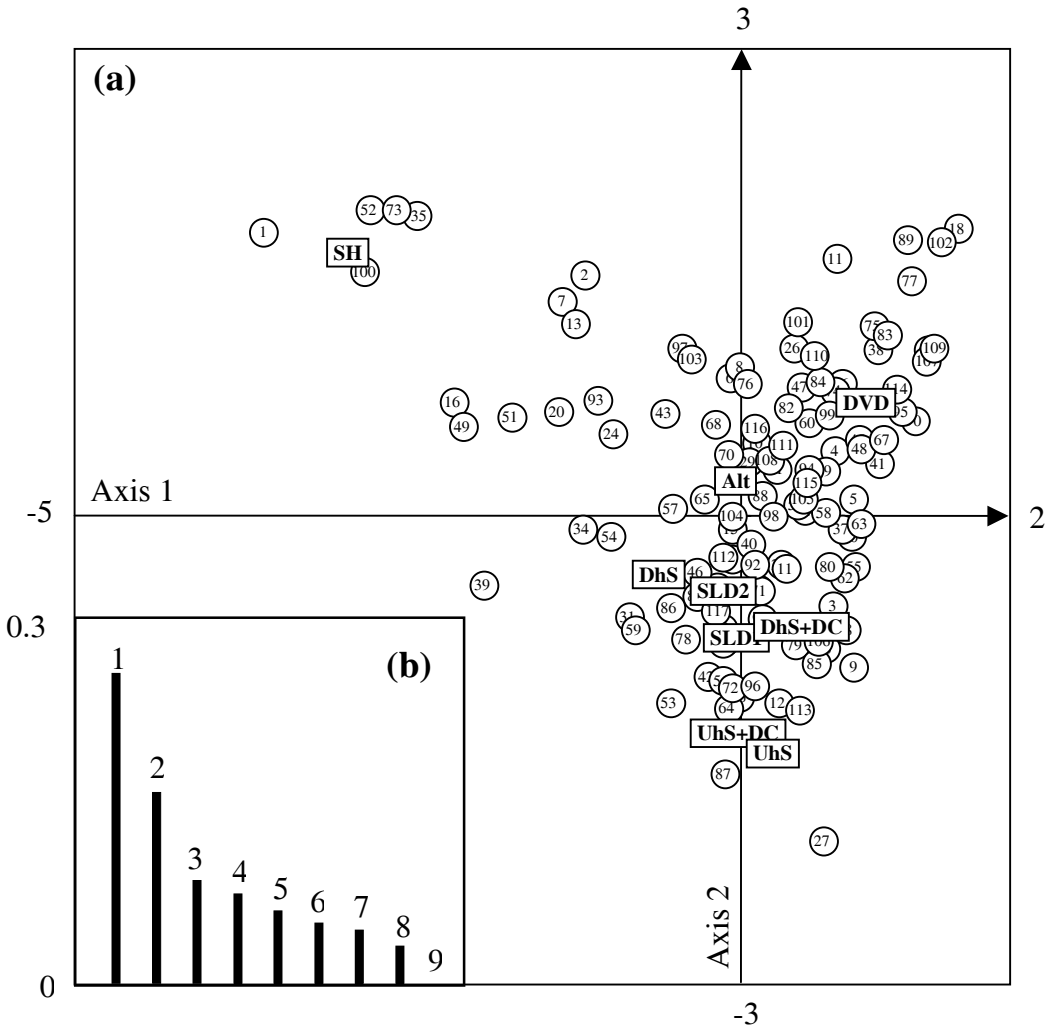


Figure 2. Canonical correlation analysis crossing 4 948 occurrences of trees (d.b.h. ≥ 10 cm) belonging to 118 species with 9 hydrological soil classes in a 10-ha forest plot at Piste de St Elie, French Guiana. (a) The circles and squares represent the weighted mean position of the occurrences of the 118 species and the 9 soil classes in the first canonical plane. Species codes are given in Appendix. Key for the soil classes is the same as in Figure 1. (b) Percentage of inertia absorbed by the 9 canonical axes.

Species optimum and amplitude regarding soil conditions, which can be viewed through the mean and variance of their canonical scores along the ordination axes (given in Appendix), are usually summarised by Gaussian curves (see Thioulouse & Chessel 1992). However, for the sake of clarity, we preferred a different representation (Figure 3) showing separately for axes 1 and 2, the species mean position along a diagonal, with horizontal bars proportional to their amplitude. Figure 3a reveals that species ordination along axis 1 corresponds to a gradient of increasing tolerance to prolonged water saturation from positive to negative values: species with high negative scores on axis 1 showed high variances along this axis (they predominated in SH but could also be found in other soil conditions), while species with high positive scores had low variances because they avoided soils experiencing surface hydromorphy (SH, DhS, DhS+DC). This main gradient was combined with a second one (Fig. 3b), which displayed along axis 2 the species that avoided surface hydromorphy, according to an increasing gradient of tolerance to temporary confinement from positive to negative values. Species with high positive scores on both axes exhibited low variances along these two axes: they were intolerant of prolonged water saturation as well as of temporary confinement. They were thus confined to the well-drained soil types (DVD and Alt). Species with negative scores on axis 2 showed high variances along this axis: they were intolerant of prolonged water saturation but could suffer temporary confinement. Frequency distributions among the nine soil classes of the three characteristic species numbered in Figure 3 are given in Figure 4.

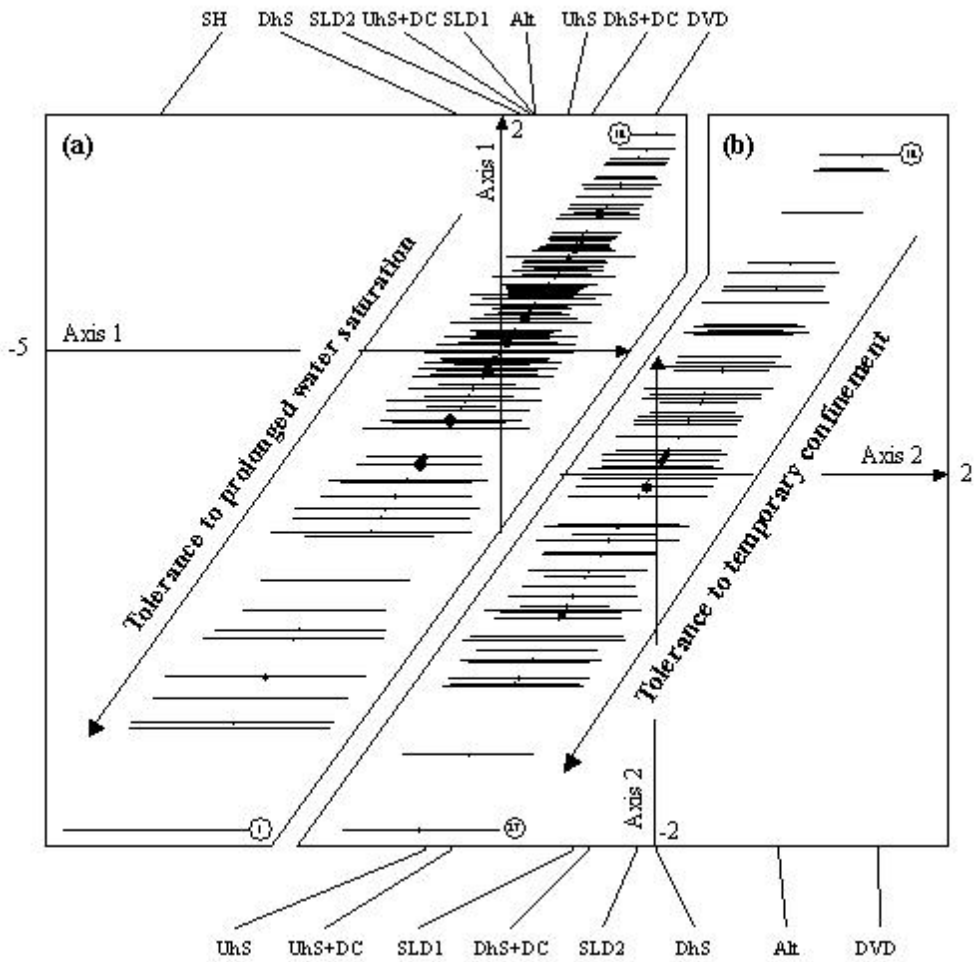


Figure 3. Species weighted mean position (species optimum) and variance (species amplitude) along axes 1 (a) and 2 (b) issued from the canonical correlation analysis crossing 4 948 occurrences of trees (d.b.h. ≥ 10 cm) belonging to 118 species with 9 hydrological soil classes in a 10-ha forest plot at Piste de St Elie, French Guiana. Only species avoiding SH soil type have been represented in (b). The circles are proportional to species

frequency and the horizontal bars to species amplitude. Frequency distributions among the soil classes of the numbered species are given in Figure 4.

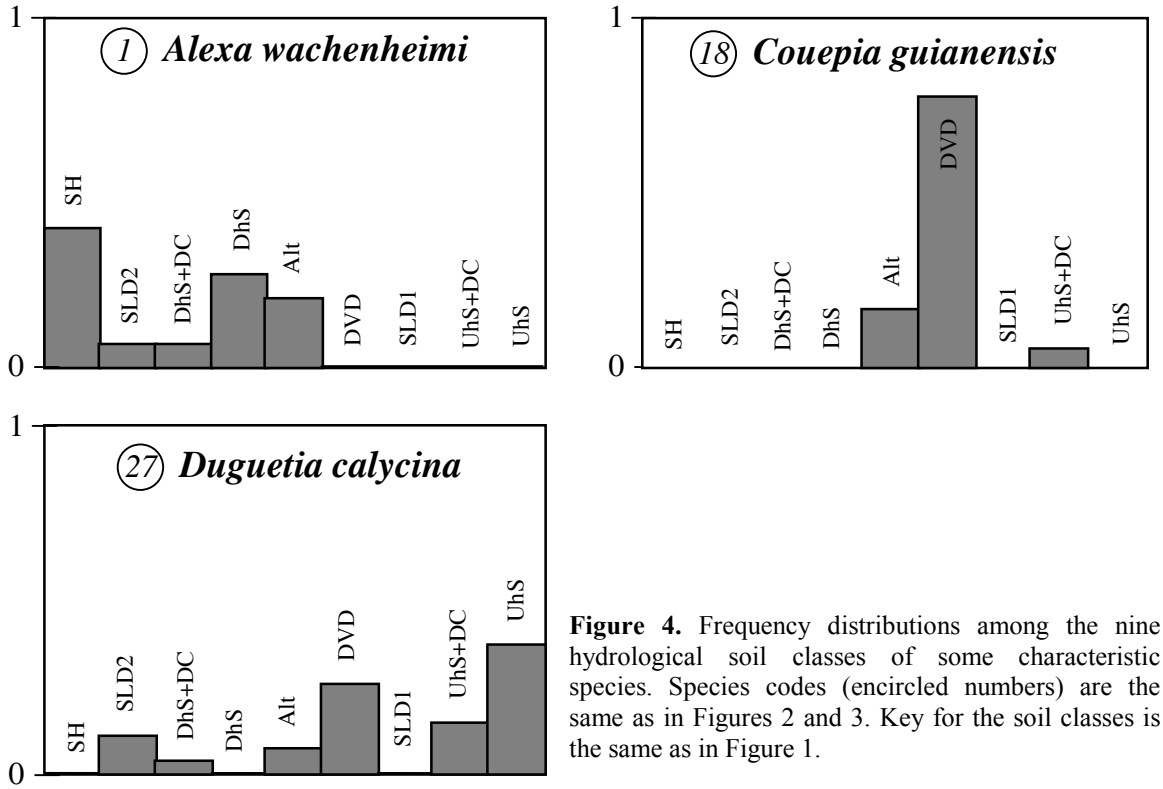


Figure 4. Frequency distributions among the nine hydrological soil classes of some characteristic species. Species codes (encircled numbers) are the same as in Figures 2 and 3. Key for the soil classes is the same as in Figure 1.

Multispecies spatial pattern

The existence of tangible relationships between species occurrences and soil factors is expected to influence the multispecies spatial pattern of the community. A first assessment of this pattern can be made through trend surface analysis (Gittins 1968; Wartenberg 1985; Borcard *et al.* 1992). We thus confronted in a Canonical Correspondence Trend Surface Analysis (CCTSA), the list of species occurrences to an environmental table containing the terms of a 2-dimensional polynomial regression of the standardised *x* and *y* tree coordinates (trend surface). According to the shape of the plot (100 x 1 000 m), we retained only 32 orthonormal polynomial variables of order less than 2 on *x* and less than 20 on *y*. The analysis was conducted on the reduced data set of 4 948 trees belonging to 118 species. A preliminary randomisation test on the total inertia among species indicated a highly significant multispecies spatial pattern (1 000 permutations; $P < 0.001$).

The sum of the first two eigenvalues of the CCTSA, representing 24.9% of the total inertia, appeared highly significant ($P < 0.001$). A partial-CCTSA was then computed in order to partial out the soil effect. When the soil effect was removed, there remained only one significant axis ($P < 0.001$) representing 13.0% of the total inertia (Table 2). The corresponding multispecies spatial patterns can be viewed as maps by drawing contour lines from the canonical scores of each tree positioned in (x,y) coordinates (Thioulouse *et al.* 1995). The contour lines were computed by interpolation of values estimated at each node of a 10 x 10 m systematic grid using bidimensional lowess regressions (Cleveland 1979) of the canonical scores over the 25 nearest trees, a value minimising the mean smoothing error (Fig. 5; Cleveland & Delvin 1988). Comparing these maps with the one of the soil classes (Fig. 6a) revealed that the multispecies pattern displayed by axis 1 of the CCTSA mainly expressed the soil effect (Fig. 6b), while the pattern displayed by axis 2 exhibited a broader scale of floristic

heterogeneity (Fig. 6c). The map computed from scores on the first axis of the partial-CCTSA, i.e. after elimination of the soil effect (Fig. 6d), showed a pattern very similar to the one displayed by axis 2 in the previous analysis. This means that soil heterogeneity plays a major role on the spatial structuring of the species mixture, but that another factor, still unidentified, is responsible for a broader significant floristic heterogeneity in this plot. Scrutinising the five most correlated species with the first canonical axis of the partial-CCTSA — *Couepia guianensis* ($N = 18$); *Pourouma villosa* ($N = 12$); *Pouteria egregia* ($N = 28$); *Siparuna decipiens* ($N = 17$); *Theobroma subincanum* ($N = 14$) — revealed that they were mainly present in the upper quarter part of the plot.

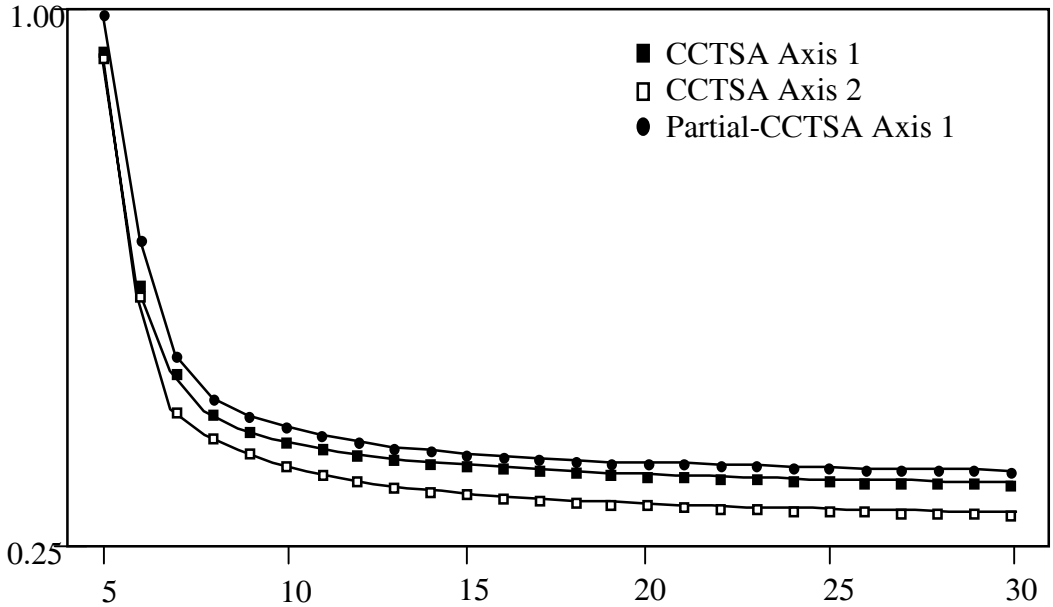


Figure 5. Mean smoothing error (sum of the squared differences between the observed and estimated values) as a function of the number of nearest neighbours taken into account in the lowess regressions conducted on the canonical scores issued from CCTSA and Partial-CCTSA (see Table 2).

Table 2. Canonical correspondence trend surface analysis (CCTSA) of 4 948 occurrences of trees (d.b.h. ≥ 10 cm) and orthonormal polynomial variables of the standardised *x* and *y* tree coordinates (trend surface) in a 10-ha forest plot at Piste de St Elie, French Guiana. Partial-CCTSA corresponds to the same analysis after partialling out the soil effect. r_k^2 = squared canonical correlation coefficient for axis *k*; % inertia = percentage of total inertia absorbed by axis *k*.

	Axis (*k*)			
	1	2	3	4
CCTSA (32 orthonormal variables)				
r_k^2	0.199	0.142	0.093	0.091
% inertia	14.55	10.37	6.79	6.62
Partial-CCTSA (24 orthonormal variables)				
r_k^2	0.147	0.072	0.062	0.057
% inertia	13.06	8.20	7.11	6.54

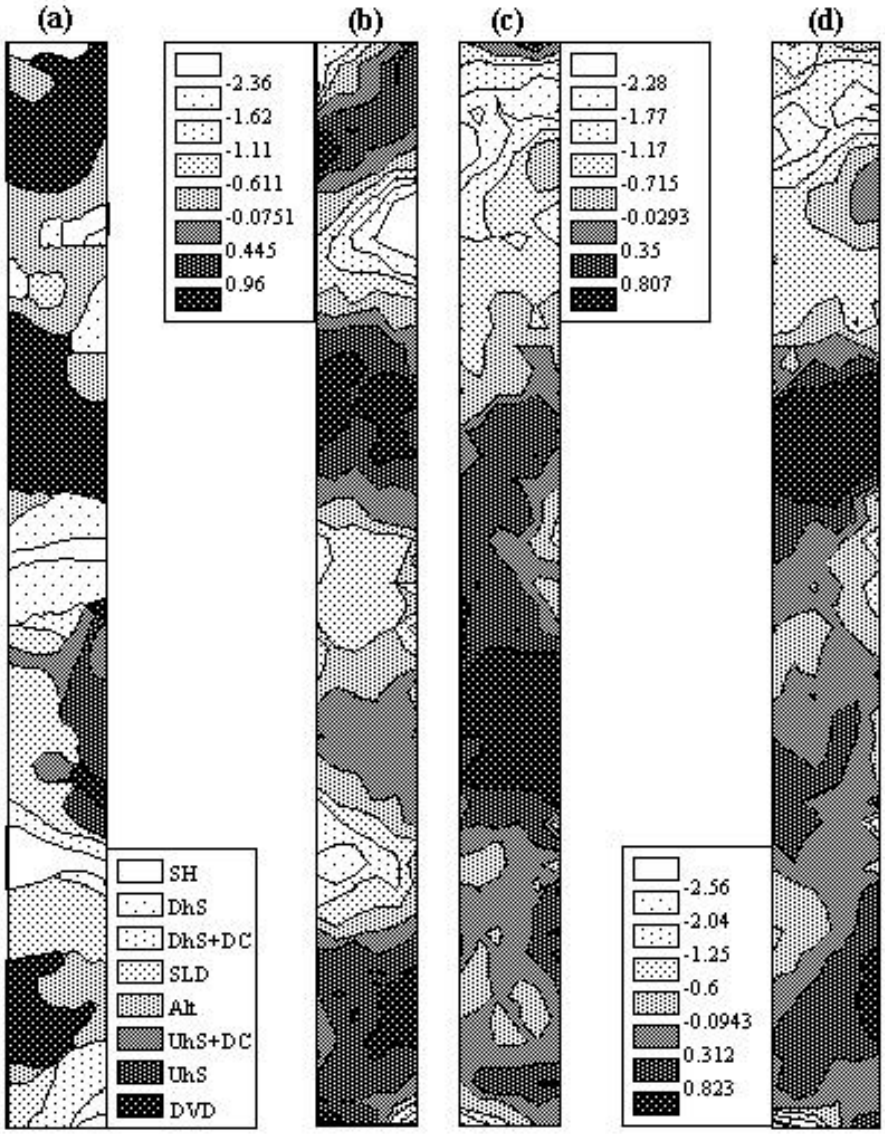


Fig. 6. Map of the soil classes (a) and of the multispecies spatial patterns displayed by axes 1 (b) and 2 (c) of the CCTSA and axis 1 (d) of the partial-CCTSA (see Table 2). The contour lines of maps (b) to (d) were drawn by interpolation of values estimated at each node of 10 x 10 m systematic grid using bidimensional lowess regressions of the canonical scores over the 25 nearest trees.

Discussion

Species tolerance to soil water saturation

Direct analysis of species-soil relationships using canonical correlations confirmed the previous results obtained at Piste de St Elie by Lescure & Boulet (1985) and Sabatier *et al.* (1997) concerning the major effect of the hydrological soil characteristics on species composition at the local scale. Apart from some specialised species (*Tabebuia insignis* and *Euterpe oleracea*, exclusively found in hydromorphic thalwegs), the most abundant species can be ordered along two intermingled soil gradients. The predominant one opposes the soils experiencing prolonged surface hydromorphy to the other soil types, which are ordered along a second gradient displaying the weathering sequence from the initial ferralitic cover to the transformed hydromorphic soil systems. These two gradients are organised between three poles of constraining hydrological soil conditions: prolonged surface water saturation (SH),

temporary confinement during the rainy season (UhS and UhS+DC), and low water potential in the dry season (DVD).

Our results completes this figure in an interesting way. Indeed, CANCOR leads to an ordination diagram so that species mean canonical scores and their variances represent species optimum and amplitude along the soil gradients. From the point of view of species, canonical axis 1 corresponds to a gradient of increasing tolerance to prolonged surface water saturation: intolerant species, *i.e.* the ones exhibiting a low amplitude, are found at one extreme of the gradient because they were never found in prolonged hydromorphic soil conditions (SH or even DhS and DhS+DC); tolerant species found in more varied soil conditions, including SH, exhibit a higher amplitude and are found at the opposite extreme of the gradient (*e.g.*, *Alexa wachenheimi, Eschweilera coriacea, Jessenia bataua, Myrcia decorticans, Sclerolobium melonii*). Species intolerant of prolonged surface water saturation display a similar pattern along canonical axis 2 according to their tolerance to temporary confinement of the transformed hydromorphic soil systems: species intolerant of prolonged water saturation and intolerant of temporary confinement, *i.e.* the ones exhibiting a low amplitude along axis 1 as well as along axis 2 (*e.g.*, *Couepia guianensis, Sloanea sp., Thyrsodium guianense, Tetragastris panamensis, Micropholis obscura*), are found at one extreme of the gradient because their distribution is restricted to the well-drained initial ferralitic cover (DVD or DVD and Alt); species intolerant of prolonged surface water saturation but tolerant of temporary confinement (low amplitude along axis 1 but high amplitude along axis 2; *e.g.*, *Duguetia calycina, Pouteria grandis*) are at the opposite extreme of this gradient.

It is notable that the species restricted to the initial ferralitic cover (DVD) does not seem to endure water shortage during the dry season, because in the reverse case they should also be found on the moister soils of the weathering sequence. We can thus support the hypothesis that, at Piste de St Elie, the soil conditions imposed by surface water saturation are more important determinants for niche breadth limitation of many tree species than water shortage. Water saturation excludes oxygen from the soil pore space and imposes constraining, and sometimes lethal, anoxic conditions for seed establishment and plant growth (Kozlowski 1986; Vartapetian & Jackson 1997; Siebel & Blom 1998). It can also induce modifications of the soil chemical properties along the drainage gradients (Sabatier *et al.* 1997).

Since excessively wet soils are common in tropical rainforests, poor soil aeration is recognised to be an important factor for tree zonation (Joly 1991; Ter Steege 1994). But the effects of water excess have been explored much less in tropical rainforests than has the effects of water shortage. From an ecophysiological point of view, this perspective is surely an interesting direction for future research that could help in explaining the species zonation in natural conditions.

Pattern of diversity and community structuring

Soil heterogeneity has been demonstrated to be an important factor influencing the local distribution of tree species in several tropical rainforests (Basnet 1992; Ter Steege *et al.* 1993; Newbery *et al.* 1996; Sabatier *et al.* 1997). At Piste de St Elie, prolonged water saturation is a major discriminant factor of species composition that leads to a high β-diversity between SH and the other soil types. This confirms the floristic singularity of the extreme environments experiencing water excess in tropical forests (Clark *et al.* 1998; Clark *et al.* 1999). This floristic singularity seems, in French Guiana, to go hand in hand with a singular canopy structure that may be viewed from aerial photographs (M. Pain-Orcet & P. Couteron *pers. com.*). This has been little investigated.

The influence of soil heterogeneity on diversity patterns has also not been clearly stated. Soil conditions are often linked to topography (Basnet 1992; Newbery *et al.* 1996;

Sabatier *et al.* 1997) and topography is recognised to increase the frequency of the natural disturbances (treefall gaps), which in turn increases species diversity (Denslow 1987; Gimaret-Carpentier *et al.* 1998a). Our results suggest that well-drained soils contain more species (both species tolerant and intolerant of surface water saturation) than transformed and hydromorphic soils, which could lead to higher diversity on hilltops. This means that different factors could have opposite effects on local diversity patterns. It could thus be very informative to quantify the relative importance of these factors in order to decompose the diversity measure accordingly.

From our results, it is already plausible to support that soil heterogeneity, and probably the depth and the length of the period of soil water saturation, is a predominant factor playing on the spatial structuring of the multispecies community. Indeed, we showed that the first axis of the Canonical Correspondence Trend Surface Analysis (CCTSA) displays a significant multispecies spatial pattern at a scale very similar to the pattern of soil heterogeneity. More than that, this pattern disappears when the soil effect is removed by partial-CCTSA. On the contrary, the second axis of the CCTSA displays a pattern whose scale does not correspond to the scale of soil heterogeneity. Moreover, a very similar pattern was displayed by the first canonical axis of the partial-CCTSA, after partialling out the soil effect. Some specific spatial patterns give flimsy arguments to hypothesise that this broad scale of floristic heterogeneity could be due to endogenous factors resulting from temporal fluctuations of population abundance (Condit *et al.* 1996). Further studies on population dynamics are however needed in this way.

Acknowledgements

This work is part of the AME-project supported by the "GIS Silvolab-Guyane" under the "XIème Contrat de Plan Etat-Région pour la Guyane". The authors are very grateful to Daniel Chessel, Clémentine Gimaret-Carpentier, Michel Grimaldi and the reviewers of the journal for their helpful comments on the manuscript, and to Sarah Blanquère for correcting the English.

References

Ashton, P.S. & Hall, P. 1992. Comparisons of structure among mixed dipterocarp forests of north-west Borneo. Journal of Ecology 80:459–481.

Austin, M.P. 1968. An ordination study of a chalk grassland community. Journal of Ecology 56:739–757.

Baillie, I.C., Ashton, P.S., Court, M.N., Anderson, J.A.R., Fitzpatrick, E.A. & Tinsley, J. 1987. Site characteristics and the distribution of tree species in mixed dipterocarp forest on tertiary sediments in Central Sarawak, Malaysia. Journal of Tropical Ecology 3:201–220.

Basnet, K. 1992. Effect of topography on the pattern of trees in *Tabonuco (Dacryodes excelsa*) dominated rain forest of Puerto Rico. Biotropica 24:31–42.

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.

Boggan, J., Funk, V., Kelloff, C., Hoff, M., Cremers, G. & Feuillet, C. 1992. Cheklist of the plants of the Guianas (Guyana, Surinam, French Guiana). National Museum of Natural History, Smithoninan Institution, Washington.

Bonal, D., Barigah, T.S., Granier, A. & Guelh, J.M. 2000a. Late-stage canopy tree species with extremly low delta-^{13}C and high stomatal sensitivity to seasonal soil drought in the tropical rainforest of French Guiana. Plant, Cell and Environment 23:445–459.

Bonal, D., Sabatier, D., Montpied, P., Tremeaux, D. & Guelh, J.M. 2000b. Interspecific variability of delta-^{13}C among trees in rainforests of French Guiana: functional groups and canopy integration. Oecologia 124:454–468.

Borcard, D., Legendre, P. & Drapeau, P. 1992. Partialling out the spatial component of ecological variation. Ecology 73:1045–1055.

Boulet, R., Humbel, F.X. & Lucas, Y. 1982. Analyse structurale et cartographie en pédologie. II: une méthode d'analyse prenant en compte l'organisation tridimentionelle des couvertures pédologiques. Cahiers ORSTOM, série pédologie 19:323–339.

Boyé, M., Cabaussel, G. & Perrot, Y. 1979. Climatologie. pp. 1–4. In: Atlas de la Guyane. CNRS & ORSTOM, Paris.

Carnes, B.A. & Slade, N.A. 1982. Some comments on niche analysis in canonical space. Ecology 63:888–893.

Clark, D.B., Clark, D.A. & Read, J.M. 1998. Edaphic variation and the mesoscale distribution of tree species in a neotropical rain forest. Journal of Ecology 86:101–112.

Clark, D.B., Palmer, M.W. & Clark, D.A. 1999. Edaphic factors and the landscape-scale distributions of tropical rain forest trees. Ecology 80:2662–2675.

Cleveland, W.S. 1979. Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association 74:829–836.

Cleveland, W.S. & Delvin, S.J. 1988. Locally weighted regression: an approach to regression analysis by local fitting. Journal of the American Statistical Association 83:596–610.

Condit, R., Hubbell, S.P., LaFrankie, J.V., Sukumar, R., Manokaran, N., Foster, R.B. & Ashton, P.S. 1996. Species-area and species-individual relationships for tropical trees: a comparison of three 50-ha plots. Journal of Ecology 84:549–562.

Denslow, J.S. 1987. Tropical rain forest gaps and tree species diversity. Annual Review of Ecology and Systematics 18:431–451.

Denslow, J.S. 1995. Disturbance and diversity in tropical rain forests: the density effect. Ecological Applications 5:962–965.

Diggle, P.J. 1983. Statistical analysis of spatial point patterns. Academic Press, London.

Escoufier, Y. 1987. The duality diagram: a means of better practical applications. pp. 139–156. In: Legendre, P. & Legendre, L. (Eds.), Development in numerical ecology. Springer Verlag, Berlin.

Fritsch, E., Boulet, R., Bocquier, G., Dosso, M. & Humbel, F.X. 1986. Les systèmes transformants d'une formation supergène de Guyane française et leurs mode de représentation. Cahiers ORSTOM, série Pédologie 22:361–395.

Gartlan, J.S., Newbery, D.M., Thomas, D.W. & Waterman, P.G. 1986. The influence of topography and soil phosphorus on the vegetation of Korup Forest Reserve, Cameroun. Vegetatio 65:131–148.

Gauch, H.G. Jr. & Wentworth, T.R. 1976. Canonical correlation analysis as an ordination technique. Vegetatio 33:17–22.

Gimaret-Carpentier, C., Pélissier, R., Pascal, J.-P. & Houllier, F. 1998a. Sampling strategies for the assessment of tree species diversity. Journal of Vegetation Science 9:161–172.

Gimaret-Carpentier, C., Chessel, D. & Pascal, J.-P. 1998b. Non-symmetric correspondence analysis: an alternative for species occurrences data. Plant Ecology 138:97–112.

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.

Grimaldi, M. & Boulet, R. 1990. Relation entre l'espace poral et le fonctionnement hydrodynamique d'une couverture pédologique sur socle en Guyane française. Cahiers ORSTOM, série pédologie 25:263–275.

Grimaldi, C., Fritsch, E. & Boulet, R. 1994. Composition chimique des eaux de nappe et évolution d'un matériau ferrallitique en présence du système muscovite-kaolinite-quartz. Compte-Rendu de l'Académie des Sciences de Paris, série III 319:1383–1389.

Guelh, J.M. 1984. Dynamique de l'eau dans le sol en forêt tropicale humide guyanaise. Influence de la couverture pédologique. Annales des Sciences Forestières 41:195–236.

Guelh, J.M., Domenach, A.M., Bereau, M., Barigah, T.S., Casabianca, H., Ferhi, A. & Garbaye, J. 1998. Functional diversity in an Amazonian rainforest of French Guiana: a dual isotope approach ($\delta^{15}N$ and $\delta^{13}C$). Oecologia 116:316–330.

He, F., Legendre, P. & LaFrankie, J.V. 1996. Spatial pattern of diversity in a tropical rain forest in Malaysia. Journal of Biogeography 23:57–74.

Hotelling, H. 1936. Relation between two sets of variates. Biometrika 28:321–377.

Huc, R., Ferhi, A. & Guelh, J.M. 1994. Pioneer and late stage tropical rainforest tree species (French Guiana) growing under common conditions differ in leaf gas exchange regulation, carbon isotope discrimination and leaf water potential. Oecologia 99:297-305.

Humbel, F.X. 1978. Caractérisation par des mesures physiques, hydriques et d'enracinement des sols de Guyane française à dynamique de l'eau superficielle. Sciences du Sol 2: 83–94.

Joly, C.A. 1991. Flooding tolerance in tropical trees. pp. 23–34. In: Jackson, M.B., Davies, D.D. & Lambers, H. (eds.), Plant life under oxygen stress. SPB Academic Publishing, The Hague.

Kozlowski, T.T. 1986. Soil aeration and growth of forest trees. Scandinavian Journal of Forest Research 1:113–123.

Lebart, L., Morineau, A. & Warwick, K. 1984. Multivariate descriptive statistical analysis. J. Wiley, New-York.

Lebreton, J.-D., Chessel, D., Prodon, R. & Yoccoz, N. 1988. L'analyse des relations espèces-milieu par l'analyse canonique des correpondances. I-Variables de milieu qualitatives. Acta Oecologica 9:53–67.

Lescure, J.P. & Boulet, R. 1985. Relationships between soil and vegetation in a tropical rain forest in French Guiana. Biotropica 17:155–164.

Lucas, Y., Boulet, R., Chauvel, A. & Veillon, L. 1986. Systèmes sols ferrallitiques-podzols en région amazonienne. pp. 53–65. In: Butt, C.R.M. & Zeegers, H. (eds), Handbook of exploration geochemistry, Vol. 4. Elsevier, Amsterdam.

Manly, B.J.F. 1991. Randomization and Monte Carlo methods in biology. Chapman and Hall, London.

Méot, A, Legendre, P. & Borcard, D. 1998. Partialling out the spatial component of ecological variation: questions and propositions in the linear modelling framework. Environmental and Ecological Statistics 5:1–27.

Milési, J.P., Egal, E., Ledru, P., Vernhet, Y., Thiéblemont, D. Cocherie, A., Tegyet, M. Martel-Jantin, B. & Lagny, P. 1995. Les minéralisations du Nord de la Guyane française dans leur cadre géologique. Chronique de la Recherche Minière 518:5–58.

Newbery, D.M. & Proctor, J. 1984. Ecological studies in four contrasting lowland rain forests in Gunung Mulu National Park, Sarawak. IV-Association between tree distribution and soil factors Journal of Ecology 72:475–493.

Newbery, D.M., Campbell, E.J.F., Proctor, J. & Still, M.J. 1996. Primary lowland dipterocarp forest at Danum Valley, Sabah, Malaysia. Species composition and patterns in the understorey. Vegetatio 122:193–220.

Pascal, J.-P. & Pélissier, R. 1996. Structure and floristic composition of a tropical evergreen forest in south-west India. Journal of Tropical Ecology 12:191–214.

Ripley, B.D. 1981. Spatial statistics. J. Wiley, New York.
Sabatier, D., Grimaldi, M., Prévost, M.-F., Guillaume, J., Godron, M., Dosso, M. & Curmi, P. 1997. The influence of soil cover organization on the floristic and structural heterogeneity of a Guianan rain forest. Plant Ecology 131 81–108.
Siebel, H.N. & Blom, C.W.P.M. 1998. Effects of irregular flooding on the establishment of tree species. Acta Botanica Neerlandica 47:231–240.
Ter Braak, C.J.F. 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology 67:1167–1179.
Ter Braak, C.J.F. 1987. The analysis of vegetation-environment relationships by canonical correspondence analysis. Vegetatio 69:69–77.
Ter Braak, C.J.F. 1988. Partial canonical correspondence analysis. pp. 551–558. In: Bock, H.H. (Ed.), Classification and related methods of data analysis. North Holland Press, Amsterdam.
Ter Braak, C.J.F. 1990. Interpreting canonical correlation analysis through biplots of structure correlations and weights. Psychometrika 55:519–531.
Ter Braak, C.J.F. & Verdonschot, P.F.M. 1995. Canonical correspondence analysis and related multivariate methods in aquatic ecology. Aquatic Sciences 57:255–289.
Ter Steege, H. 1994. Flooding and drought tolerance in seeds and seedlings of two Mora species segregated along a soil gradient in the tropical rain forest of Guyana. Oecologia 100:356–357.
Ter Steege, H., Jetten, V.G., Polak, A.M. & Werger, M.J.A. 1993. Tropical rain forest types and soil factors in a watershed area in Guyana. Journal of Vegetation Science 4:705–716.
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.
Thioulouse, J. Chessel, D., Dolédec, S. & Olivier, J.-M. 1997. ADE-4: a multivariate analysis and graphical display software. Statistics and Computing 7:75–83.
Vartapetian, B.B. & Jackson, M.B. 1997. Plant adaptations to anaerobic stress. Annals of Botany 73:3–20.
Veillon, L. & Soria-Solano, B. 1988. Transition sol ferrallitique-podzol : cas d'une terrasse sédimentaire de l'Ucayali (Pérou). Cahier ORSTOM, série Pédologie 24:97–113.
Wartenberg, D.E. 1985. Canonical trend surface analysis: a method for describing geographic pattern. Systematic Zoology 34:259–279.

Appendix. Species codes, names and authorities for 118 species in a 10-ha plot at Piste de St Elie, French Guiana. *N* corresponds to the number of trees. *MCS* and *VCS* correspond to the mean species canonical scores (species optimum) and their variances (species amplitude) on axes 1 (*A1*) and 2 (*A2*) issued from canonical correlation analysis crossing species occurrences with nine hydrological soil classes (see text).

Code	TAXON	*N*	*MCS A1*	*MCS A2*	*VSC A1*	*VSC A2*
1	*Alexa wachenheimi* Benoist	15	-3.5779	1.8209	0.9508	0.5722
2	*Ambelania acida* Aubl.	11	-1.1778	1.541	0.9983	0.3621
3	*Andira coriacea* Pulle	11	0.6869	-0.5853	0.0914	0.3327
4	*Astrocaryum sciophilum* (Miq.) Pulle	113	0.7069	0.408	0.1696	0.2799
5	*Bocoa prouacensis* Aubl.	69	0.838	0.0975	0.1233	0.3196
6	*Brosimum guianense* (Aubl.) Huber	13	-0.0844	0.8827	0.5482	0.3041
7	*Carapa procera* A. DC.	47	-1.3451	1.3748	1.0948	0.5551
8	*Caryocar glabrum* (Aubl.) Pers.	20	-0.0098	0.9583	0.6238	0.4452
9	*Casearia javitensis* Kunth	12	0.8396	-0.9832	0.1152	0.6005
10	*Cassipourea guianensis* Aubl.	58	0.1219	0.46	0.4591	0.4025
11	*Catostemma fragrans* Benth.	17	0.7178	1.646	0.4632	0.2471
12	*Chaetocarpus schomburgkianus* (Kuntze) Pax & K. Hoffm.	35	0.2905	-1.2059	0.1021	0.4152
13	*Chrysophyllum argenteum* Jacq. *ssp. Nitidum* (G. Mey.) Pennington	17	-1.2413	1.2305	0.8901	0.4302
14	*Chrysophyllum prieurii* A. DC.	10	-0.118	-0.8146	0.1777	0.2738
15	*Chrysophyllum sanguinolentum* (Pierre) Baehni	17	-0.0677	-0.0882	0.1237	0.1671
16	*Conceveiba guianensis* Aubl.	12	-2.1562	0.7313	0.9497	0.544
17	*Couepia caryophylloides* Benoist	12	-0.1419	-0.8225	0.1456	0.2518
18	*Couepia guianensis* Aubl.	18	1.6294	1.8442	0.0818	0.1586
19	*Couratari multiflora* (J.E. Smith) Eyma	12	0.6367	0.2784	0.1912	0.2392
20	*Crudia aromatica* (Aubl.) Willd.	10	-1.3727	0.6691	0.4916	0.2425
21	*Crudia bracteata* Benth.	171	-0.0663	-0.2827	0.2818	0.3613
22	*Cupania scrobiculata* L.C. Rich.	12	0.1249	-0.4764	0.1548	0.2808
23	*Dacryodes nitens* Cuatrec.	19	0.6321	-0.8584	0.1226	0.3415
24	*Dendrobangia boliviana* Rusby	44	-0.9581	0.5208	0.6859	0.4904
25	*Dicorynia guianensis* Amsh.	79	-0.1798	-0.4635	0.2978	0.4095
26	*Drypetes variabilis* Uittien	25	0.3962	1.0751	0.3538	0.2112
27	*Duguetia calycina* Benoist	27	0.6219	-2.0971	0.0781	0.5686
28	*Duguetia surinamensis* R.E. Fries	32	0.795	-0.7334	0.1124	0.4278
29	*Duroia aquatica* (Aubl.) Bremek.	14	0.0566	0.3444	0.196	0.1623
30	*Ecclinusa guianensis* Eyma	25	0.8291	-0.1423	0.1074	0.4097
31	*Eperua falcata* Aubl.	378	-0.8407	-0.6553	0.4083	0.4114
32	*Eperua grandiflora* (Aubl.) Benth.	78	-0.1372	-0.7279	0.2364	0.3522
33	*Eschweilera apiculata* (Miers) A.C. Smith	16	0.2959	-0.3201	0.1144	0.3241
34	*Eschweilera cf. chartaceifolia* Mori	25	-1.179	-0.0913	0.6033	0.4779
35	*Eschweilera coriacea* (A.P. DC.) Mori	67	-2.4319	1.9269	1.0953	0.4343
36	*Eschweilera decolorans* Sandw.	47	-0.003	-1.1669	0.2368	0.4079
37	*Eschweilera micrantha* (Berg) Miers	199	0.7655	-0.0863	0.1438	0.3781
38	*Eschweilera parviflora* (Aubl.) Miers	269	1.0256	1.061	0.1561	0.2132
39	*Eschweilera pedicellata* (Rich.) Mori	12	-1.9328	-0.4479	0.5824	0.5092
40	*Eschweilera sagotiana* Miers	176	0.0685	-0.1911	0.2698	0.3375
41	*Eugenia sp. 10*	10	1.0398	0.326	0.0918	0.2596
42	*Goupia glabra* Aubl.	10	-0.2461	-1.0437	0.168	0.3419
43	*Gustavia hexapetala* (Aubl.) J.E. Smith	15	-0.5691	0.6516	0.8048	0.4938
44	*Hebepetalum humiriifolium* (Planch.) Benth.	14	0.4812	0.0306	0.1923	0.275
45	*Hirtella bicornis* Mart. & Zucc.	15	0.8901	0.4835	0.1251	0.2563

46	*Humiriastrum subcrenatum* (Benth.) Cuatrec.	11	-0.3281	-0.3668	0.1478	0.1613
47	*Inga fanchoniana* O. Poncy	10	0.4478	0.8251	0.1394	0.1323
48	*Inga sp.4*	20	0.8941	0.4245	0.1287	0.3044
49	*Iryanthera hostmanni* (Benth.) Warb.	26	-2.0779	0.5698	0.8053	0.5724
50	*Iryanthera sagotiana* (Benth.) Warb.	107	0.418	0.0654	0.1552	0.287
51	*Jacaranda copaia* (Aubl.) D. Don	13	-1.7124	0.6273	0.621	0.3547
52	*Jessenia bataua* (Mart.) Burret	54	-2.7765	1.9654	1.1049	0.481
53	*Lacmellea floribunda* (Poepp.) Benth. & J.D. Hook.	21	-0.5328	-1.2091	0.0845	0.1383
54	*Laetia procera* (Poepp.) Eichl.	11	-0.9781	-0.1384	0.4966	0.3077
55	*Lecythis holcogyne* (Sandw.) Mori	23	0.8642	-0.3267	0.1249	0.4566
56	*Lecythis idatimon* Aubl.	375	-0.1406	-1.0571	0.2638	0.4051
57	*Lecythis persistens* Sagot *ssp. persistens*	306	-0.5173	0.0425	0.5214	0.4595
58	*Lecythis poiteaui* Berg	30	0.6321	0.0128	0.1503	0.3937
59	*Licania alba* Bernoulli) Cuatrec.	122	-0.7956	-0.7383	0.3558	0.382
60	*Licania canescens* Benoist	53	0.5047	0.5969	0.1587	0.1964
61	*Licania densiflora* Kleinh.	21	0.2735	0.2997	0.2038	0.2522
62	*Licania granvillei* Prance	39	0.773	-0.4071	0.1273	0.4327
63	*Licania heteromorpha* Benth. *var. heteromorpha*	41	0.8958	-0.0485	0.1096	0.3812
64	*Licania laxiflora* Fritsch	23	-0.0945	-1.2395	0.1448	0.3515
65	*Licania membranacea* Sagot ex Laness.	26	-0.2755	0.1078	0.4818	0.3692
66	*Licania parvifructa* Fanshawe & Maguire	12	0.7592	0.8507	0.1957	0.2874
67	*Maquira guianensis* Aubl.	12	1.0678	0.4871	0.1408	0.3673
68	*Micropholis guyanensis* (A. DC.) Pierre	39	-0.1848	0.5792	0.5498	0.3932
69	*Micropholis obscura* Pennington	13	1.4093	1.0604	0.1067	0.3415
70	*Minquartia guianensis* Aubl.	13	-0.0949	0.3931	0.1579	0.13
71	*Moronobea coccinea* Aubl.	17	0.1437	-0.4859	0.1626	0.2965
72	*Mouriri crassifolia* Sagot	12	-0.0589	-1.1154	0.1461	0.3303
73	*Myrcia decorticans* DC.	12	-2.5896	1.9611	1.1069	0.4517
74	*Ocotea ceanothifolia* (Nees) Mez	11	0.7036	0.8004	0.1651	0.2184
75	*Ocotea rubra* Mez	12	0.998	1.2187	0.1975	0.174
76	*Ocotea schomburgkiana* (Nees) Mez	13	0.0502	0.8483	0.5503	0.378
77	*Oenocarpus bacaba* Mart.	14	1.285	1.5072	0.1153	0.1479
78	*Ouratea melinonii* (Tiegh.) Lemee	17	-0.411	-0.7922	0.124	0.1782
79	*Oxandra asbeckii* (Pulle) R.E. Fries	74	0.4117	-0.8328	0.1085	0.2997
80	*Pachira dolichocalyx* Robyns	19	0.6676	-0.3246	0.1816	0.2743
81	*Poraqueiba guianensis* Aubl.	34	-0.3321	-0.5275	0.2239	0.2326
82	*Posoqueria latifolia* (Rudge) Roem. & Schult.	32	0.3529	0.6891	0.465	0.3644
83	*Pourouma laevis* Benoist	12	1.0929	1.1599	0.1537	0.2753
84	*Pouteria egregia* Sandw.	28	0.5952	0.8534	0.2274	0.1868
85	*Pouteria eugeniifolia* (Pierre) Baehni	10	0.5691	-0.9587	0.1002	0.5428
86	*Pouteria gongrijpii* Eyma	37	-0.5279	-0.5926	0.3401	0.3218
87	*Pouteria grandis* Eyma	28	-0.1204	-1.6575	0.0998	0.3641
88	*Pouteria guianensis* Aubl.	20	0.1555	0.1292	0.2315	0.2449
89	*Pouteria sp.21*	33	1.2488	1.7653	0.1671	0.0877
90	*Pouteria torta* (Mart.) Radlk.	17	1.3031	0.6088	0.0942	0.4091
91	*Pradosia cochlearia* (Lecomte) Pennington	13	0.1538	-0.6711	0.1421	0.2914
92	*Pradosia ptychandra* (Eyma) Pennington	28	0.1021	-0.3139	0.1679	0.303
93	*Protium opacum* Swart *var. rabelianum* Daly	16	-1.0798	0.7437	0.6481	0.3195
94	*Protium sp.1*	52	0.5097	0.2932	0.2506	0.3309
95	*Protium subserratum* (Engl.) Engl.	12	1.2143	0.6701	0.1409	0.3512
96	*Rinorea pectino-squamata* Hekking	40	0.1077	-1.0985	0.0907	0.4066
97	*Ruitzerania albiflora*	11	-0.4369	1.0719	0.5959	0.2557
98	*Sandwithia guianensis* Lanj.	144	0.2453	-0.0023	0.3061	0.3979

99	*Schefflera decaphylla* (Seem.) Harms	11	0.6669	0.6381	0.2001	0.2294
100	*Sclerolobium melinonii* Harms	11	-2.8188	1.5674	1.1654	0.7025
101	*Siparuna decipiens* (Tul.) A. DC.	17	0.4281	1.243	0.5552	0.2832
102	*Sloanea sp.*	11	1.5096	1.7528	0.0967	0.132
103	*Sterculia pruriens* (Aubl.) K. Schum.	10	-0.3754	1.0007	0.7028	0.3452
104	*Swartzia polyphylla* DC.	25	-0.0613	-0.0007	0.3296	0.3473
105	*Symphonia sp.1*	35	0.4622	0.1007	0.1592	0.2608
106	*Talisia microphylla* Uittien	14	0.5773	-0.8122	0.1351	0.5362
107	*Tetragastris panamensis* (Engl.) Kuntze	12	1.3911	0.987	0.1037	0.3565
108	*Theobroma subincanum* Mart.	14	0.2117	0.3581	0.5753	0.5964
109	*Thyrsodium guianense* Sagot ex March.	16	1.4483	1.0734	0.0945	0.3737
110	*Tovomita sp.*	40	0.5553	1.0308	0.413	0.3362
111	*Tovomita sp.1*	20	0.3177	0.4549	0.1799	0.2714
112	*Tovomita sp.5*	11	-0.1304	-0.2706	0.5713	0.5698
113	*Trymatococcus oligandrus* (Benoist) Lanj.	11	0.441	-1.2573	0.1331	0.4481
114	*Unonopsis rufescens* (Baill.) R.E. Fries	45	1.1686	0.8153	0.1445	0.3242
115	*Virola michelii* Heckel	46	0.4965	0.2048	0.1502	0.3101
116	*Vouacapoua americana* Aubl.	71	0.1087	0.5539	0.4773	0.3887
117	*Xylopia nitida* Dunal	17	-0.1858	-0.618	0.1234	0.1469
118	*Zygia racemosa* (Ducke) Barneby & Grimes	14	0.3417	-0.3473	0.1303	0.3565

CONSISTENCY BETWEEN ORDINATION TECHNIQUES AND DIVERSITY MEASUREMENTS: TWO STRATEGIES FOR SPECIES OCCURRENCE DATA

RAPHAËL PÉLISSIER,[1] PIERRE COUTERON,[2] STÉPHANE DRAY,[3] AND DANIEL SABATIER[1]

Running head: Ordination techniques and diversity measurements

Abstract. Both the ordination of taxonomic tables and the measurements of species diversity aim to capture the prominent features of the species composition of a community. However, inter-relations between ordination techniques and diversity measurements are seldom explicated and are mainly ignored by many field ecologists. This paper starts from the notion of the species occurrence table which provides a unifying formulation for different kinds of taxonomic data. Here it is demonstrated that alternative species weightings can be used to equate the total inertia of a centred-by-species occurrence table with common diversity indices, such as species richness, Simpson diversity or Shannon information. Such an equation defines two main ordination strategies related to two different but consistent measures of species diversity. The first places emphasis on scarce species and is based on Correspondence Analysis and species richness (CA-richness strategy). The second, in which abundant species are prominent, relies on Non-Symmetric Correspondence Analysis and Simpson diversity (NSCA-Simpson strategy). Both strategies are suitable for measuring α- and β-diversity by analyzing the centred-by-species occurrence table with respect to external environmental or instrumental variables.

In this paper, these two strategies are applied to ecological data obtained in a neotropical rainforest plot. The results are then discussed with respect to the intrinsic characteristics of the community under analysis, and also to the broad classes of floro-faunistic data used in ecology, i.e. data gathered from Museum or Herbarium collections, exhaustive inventories in a reference plot or enumeration through species-by-relevés tables. The approach encompasses several well-known techniques such as Correspondence Analysis, Non-Symmetric Correspondence Analysis, Canonical Correspondence Analysis and Redundancy Analysis, and provides greater insight into inter-relations between ordination methods and diversity studies.

Key words: α- and β-diversity; inertia decomposition; multivariate analysis; species-environment relationships; species richness; species weight; species occurrence table.

INTRODUCTION

The measurement of species diversity is a central topic in community ecology (Ricklefs 1990, Krebs 1994, Begon et al. 1996). The simplest measure of diversity is species richness, which corresponds to the number of species present in a community. But numerous other indices, based on the idea that species frequency distribution is more informative than simple species richness, can be found in ecological literature (e.g. Magurran 1988). Of these, the famous non-parametric Shannon (1948) information, H, Simpson (1949) concentration, λ and Simpson diversity, $D = 1-\lambda$ (Greenberg 1956) are the most commonly employed. Pielou (1969) has shown that D is an unbiased sample estimator of population diversity that is insensitive to very rare species (which are generally poorly sampled) but sensitive to changes in the abundance of the few prevalent species, so that Hill (1973) recommended to use $1/\lambda$instead. Following Patil and Taillie (1982), Lande (1996) noted however that Simpson's $D = 1-\lambda$ can be expressed as the total variance in species identity within a community.

Whittaker (1972) introduced the important concept of diversity partitioning into within (α) and between (β) components, which prefigured an inter-relation between ordination techniques and diversity studies (Gauch and Whittaker 1972, Gauch 1973). However, for most field ecologists, this link remained somewhat abstract because it was not clearly related to the commonly-used indices of species diversity. Later, Ter Braak (1983) emphasized that Principal Component Analysis (PCA) based on species profiles can be interpreted in terms of α- and β-diversity related to Simpson's D . More recently, Gimaret-Carpentier et al. (1998a) pointed out that total inertia computed by Non-Symmetric Correspondence Analysis (NSCA; Lauro and D'Ambra 1984) corresponds exactly to Simpson diversity. As far as we know, however, no paper has ever been published in the ecological literature to really unify the notions of ordination and diversity measures.

The paper here starts from the notion of the species occurrence table, which is a unifying formulation for different kinds of taxonomic data. It will be shown that the choice of an appropriate species weighting modifies the metric of the eigenanalysis of a centred-by-species occurrence table and enables its total inertia to be equated with the common diversity indices. Such an equation will be extended to α- and β-diversity by considering a constraint analysis (Dray 2001) of the occurrence table with respect to environmental or instrumental (*sensu* Rao 1964) variables. Our aim is to underline how the analysis of species-environment relationships may rely on two main alternative strategies that encompass simultaneously an ordination technique and a consistent measurement of taxonomic diversity. Both strategies will be applied to ecological data obtained in a neotropical rainforest plot, before being discussed with respect to the broad classes of floro-faunistic data used in ecology.

MEASURING TOTAL DIVERSITY FROM A TABLE OF SPECIES OCCURRENCES

A table of species occurrences is a very simple, thought unusual, form of presentation of taxonomic data, where each individual recorded is allocated to a taxonomic category (usually a particular species). In terms of data analysis, it corresponds to a complete binary table **T** with n rows (individuals or occurrences) and p columns (species), such that:

$$\mathbf{T}_{(n\times p)} = \left[t_{ij}\right] = \begin{cases} 1 \text{ if the } i\text{ - th occurrence belongs to the } j\text{ - th species} \\ 0 \text{ otherwise} \end{cases}$$

Table **T** differs from the usual site-by-species ecological table since it is irrespective of sampling units: each row represents a single species presence (or occurrence) observed

either at a single site or throughout a collection of sites. Table **T** has some analogy with the *inflated matrix* (Legendre & Legendre 1998, p. 595) used in Canonical Correspondence Analysis (CCA; Ter Braak 1986) to get a weighted regression. Such a species occurrence table may originate from data gathered from Museum or Herbarium collections (e.g. Gimaret et al. 1998a), exhaustive inventories in a reference plot (e.g. Pélissier et al. 2002a) or species enumeration through several relevés.

Column means of **T** are simply the species relative frequencies, noted $f_j = t_{+j} / t_{++}$, where t_{+j} is the sum of values in column *j*, i.e. the number of occurrences belonging to species *j*, and t_{++} is the sum of all values in the whole table, i.e. *n*, the total number of occurrences. Let us consider **Tc** of size (*n* x *p*), the centred-by-columns (species) table derived from **T**. It contains either $t_{ij} - f_j = 1 - f_j$ when the occurrence belongs to species *j* ($t_{ij} = 1$) or $t_{ij} - f_j = -f_j$ when the occurrence does not belong to species *j* ($t_{ij} = 0$). Column means of **Tc** are 0 and column variances are $f_j(1 - f_j)$.

Let us now consider a (*n* x *n*) diagonal matrix of row weights, noted $\mathbf{D}_n$, and a (*p* x *p*) diagonal matrix of species weights, noted Δ_p. Matrix $\mathbf{D}_n$ is defined such that:

$$\mathbf{D}_n = \left[d_{n_{ij}}\right] = \begin{cases} f_i = t_{i+} / t_{++} & i = j \\ 0 & i \neq j \end{cases}$$

For the table of species occurrences **T** defined above, $t_{++} = n$ and $\forall i,\ t_{i+} = 1$, hence $\mathbf{D}_n$ contains on its diagonal the natural weights of the occurrences, i.e. $f_i = 1/n$. Our purpose is first to show that varying the species weights contained in Δ_p changes the metric of the eigenanalysis of table **Tc** and gives it some interesting properties with respect to diversity measurements.

For the sake of clarity, the statistical triplet notation of Escoufier (1987) will be adopted below to specify the type of analysis under consideration (see also Dolédec et al. 1996, 2000 for an introduction to triplet notation in ecology). For instance, the eigenanalysis of the statistical triplet (**Tc**, $\mathbf{\Delta}_p$, $\mathbf{D}_n$) indicates that we perform a generalized singular value decomposition (GSVD; Greenacre 1984) of table **Tc**, using column weights given by $\mathbf{\Delta}_p$ and row weights given by $\mathbf{D}_n$ (*see Appendix*).

Let us consider first the classical χ^2 metric of the Correspondence Analysis (CA; Hill 1974), which defines a diagonal matrix $\mathbf{\Delta}_p = \mathbf{D}_p^{-1}$ containing species weights corresponding to:

$$\mathbf{D}_p^{-1} = \left[d_{p_{ij}}\right] = \begin{cases} 1/f_j & i = j \\ 0 & i \neq j \end{cases}$$

The CA of **T** can be defined from the statistical triplet (**Tc**, $\mathbf{D}_p^{-1}$, $\mathbf{D}_n$), whose eigenanalysis has total inertia corresponding to the weighted sum of the column variances of **Tc**:

$$I_T = \sum_{j=1}^{p} 1/f_j \cdot f_j(1 - f_j) = \sum_{j=1}^{p} (1 - f_j) = p - 1$$

where *p* corresponds to the total number of species (columns of **Tc**). Species richness expressed as *p* - 1 is obviously a measure of diversity (Hill 1973, Patil and Taillie 1982) since a community containing a single species has a diversity of 0. Note also that the lower the value of f_j, the greater the contribution by a given species (Fig. 1).

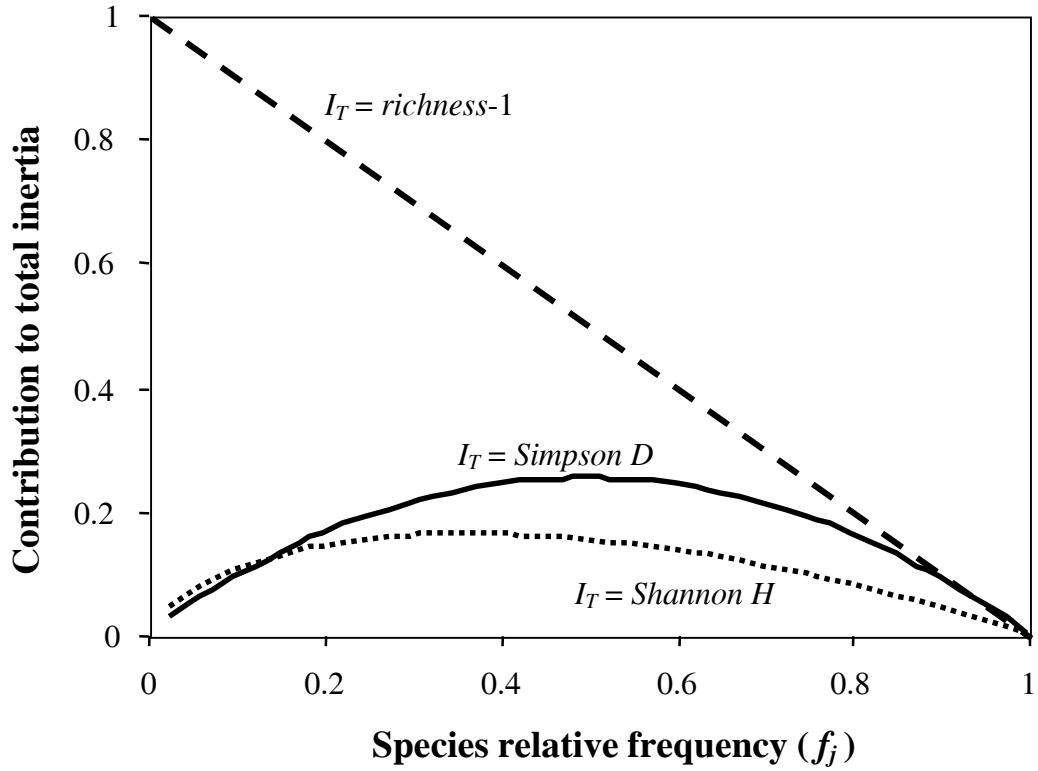


Figure 1. Species contributions to the total inertia (I_T) of the centred-by-species occurrence table (**Tc**) in relation to the relative frequency of species (f_j) when species weights are: $1/f_j$ (I_T = species richness-1); $\log[1/f_j]/(1-f_j)$ (I_T = Shannon information H); or 1 (I_T = Simpson diversity D).

Consider now a diagonal identity matrix $\mathbf{\Delta}_p = \mathbf{I}_p$ that contains uniform species weights:

$$\mathbf{I}_p = \left[i_{p_{ij}}\right] = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

The NSCA of **T** can be defined from the statistical triplet (**Tc**, $\mathbf{I}_p$, $\mathbf{D}_n$), whose eigenanalysis uses the Euclidian metric, and corresponds simply to the centred-PCA of **T**. The total inertia of this analysis computed as the weighted sum of the column variances of **Tc** is:

$$I_T = \sum_{j=1}^{p} f_j(1-f_j) = 1 - \sum_{j=1}^{p} f_j^2$$

which corresponds exactly to Simpson's diversity. The contribution of a given species increases here with its relative frequency on condition that f_j does not reach 0.5 (Fig. 1).

It should also be noted that $\mathbf{\Delta}_p$ containing species weights of $\log[1/f_j]/(1-f_j)$ leads to an intermediate situation between species richness and Simpson index (Fig. 1, see also Gimaret-Carpentier et al. 1998b), where the total inertia of the eigenanalysis of (**Tc**, $\mathbf{\Delta}_p$, $\mathbf{D}_n$) would correspond to Shannon $H = -\sum_{j=1}^{p} f_j \log(f_j)$. However, for the sake of simplicity, the Shannon index will not be considered further in this paper.

EXPLAINING THE OCCURRENCES BY ENVIRONMENT

Analysis of the statistical triplet based on **Tc** is of no interest by itself except that it links total inertia to a diversity index through the choice of species weights. Various meaningful analyses may nevertheless be conducted in the general framework of PCA on Instrumental Variables (PCAIV; Rao 1964), more widely known as the Redundancy Analysis (RDA; Wollenberg 1977), by analyzing **Tc** with respect to external variables that provide biological, geographical or environmental information for each occurrence.

Let us consider a table **X**, with n rows (occurrences) and m columns (external variables). Each row of **X** represents a vector of environmental data facing a single species occurrence of the taxonomic table. Multiple linear regressions of each column in **Tc** (response variables) on all columns of **X** (explanatory variables) result in a table $\mathbf{Tc_X}$ containing the fitted values and a table $\mathbf{Tc_{|X}}$ containing the residuals $\mathbf{Tc} - \mathbf{Tc_X}$. The statistical triplet ($\mathbf{Tc}$, $\mathbf{\Delta}_p$, $\mathbf{D}_n$) can thus be broken down into two additive parts providing two new statistical triplets (*see Appendix*): ($\mathbf{Tc_X}$, $\mathbf{\Delta}_p$, $\mathbf{D}_n$), which concerns the part of **Tc** explained by the environment and whose eigenanalysis corresponds to PCAIV, and ($\mathbf{Tc_{|X}}$, $\mathbf{\Delta}_p$, $\mathbf{D}_n$), which concerns the part independent from **X** and whose eigenanalysis corresponds to Orthogonal PCAIV (Sabatier 1984, Sabatier et al. 1989), also known as partial-RDA (Davies and Tso 1982). This decomposition divides the total inertia of table **Tc** into a part explained and a part unexplained by the variables contained in table **X**:

$$Iner\,(\mathbf{Tc}, \mathbf{\Delta}_p, \mathbf{D}_n) = Iner\,(\mathbf{Tc_X}, \mathbf{\Delta}_p, \mathbf{D}_n) + Iner\,(\mathbf{Tc_{|X}}, \mathbf{\Delta}_p, \mathbf{D}_n)$$

PARTITIONING THE OCCURRENCES: MEASUREMENTS OF α- AND β-DIVERSITY

Let us consider for instance a qualitative variable that partitions the occurrences into m classes. Such a variable may be truly environmental (e.g. soil classes) or more instrumental (e.g. sampling units or relevés). Table **X**, with n rows (occurrences) and m columns (classes of the qualitative variable), contains dummy variables, such that:

$$\mathbf{X}_{(n \times m)} = [x_{ik}] = \begin{cases} 1 \text{ if the } i\text{-th occurrence belongs to the } k\text{-th class} \\ 0 \text{ otherwise} \end{cases}$$

As **X** contains dummy variables, $\mathbf{Tc_X}$ contains the average species profiles for the classes, i.e., $f_{j/k} - f_j$, where $f_{j/k}$ is the relative frequency of the j-th species in the k-th class. Consequently, the rows in this table are identical for all occurrences that belong to the same class.

The statistical triplets ($\mathbf{Tc_X}$, $\mathbf{\Delta}_p$, $\mathbf{D}_n$) and ($\mathbf{Tc_{|X}}$, $\mathbf{\Delta}_p$, $\mathbf{D}_n$) can now be analyzed using alternatively one of the two matrices of column weights given by $\mathbf{\Delta}_p = \mathbf{D}_p^{-1}$ or $\mathbf{\Delta}_p = \mathbf{I}_p$. These analyses, which are identical to PCAIV and Orthogonal PCAIV with qualitative instrumental variables, are called between- and within-class analyses in the context of partitioned occurrence data (Dolédec and Chessel 1989). The between-class inertia (I_B) is the part of the total inertia of table **Tc** explained by **X**, while the within-class inertia (I_W) is the part of the total inertia of table **Tc** unexplained by **X**. Hence, partitioning the occurrences according to a qualitative external variable gives the following decomposition: $I_T = I_B + I_W$, where I_T measures the total species diversity of the community (i.e. the γ-diversity of Whittaker 1972), broken down into between (β-diversity) and within (α-diversity) additive components (Lande 1996).

It can be demonstrated (*see Appendix*) that the eigenanalysis of ($\mathbf{Tc_X}$, $\mathbf{\Delta}_p$, $\mathbf{D}_n$) is a Correspondence Analysis (CA) when $\mathbf{\Delta}_p = \mathbf{D}_p^{-1}$, and a Non-Symmetric Correspondence Analysis (NSCA) when $\mathbf{\Delta}_p = \mathbf{I}_p$. It follows that the eigenanalyses of ($\mathbf{Tc}$, $\mathbf{\Delta}_p$, $\mathbf{D}_n$), ($\mathbf{Tc_X}$, $\mathbf{\Delta}_p$, $\mathbf{D}_n$) and ($\mathbf{Tc}_{|\mathbf{X}}$, $\mathbf{\Delta}_p$, $\mathbf{D}_n$) compute I_T, I_B and I_W, respectively, either in the χ^2 ($\mathbf{\Delta}_p = \mathbf{D}_p^{-1}$) or the Euclidian metric ($\mathbf{\Delta}_p = \mathbf{I}_p$), i.e. measuring species diversity by species richness or Simpson index. For the sake of simplicity, the text below will refer to a CA-richness strategy as an analysis that uses $\mathbf{D}_p^{-1}$ as species weights, and a NSCA-Simpson strategy as an analysis that uses $\mathbf{I}_p$ as species weights.

The weighted average α-diversity within the classes of $\mathbf{X}$, is $I_W = \sum_{k=1}^{m} f_k I_k$, where I_k is the within-class inertia of k (Lande 1996). Here it is noteworthy that the NSCA-Simpson strategy gives $I_k = \sum_{j=1}^{p} f_{j/k}(1 - f_{j/k}) = 1 - \sum_{j=1}^{p} f_{j/k}^2$, which is exactly Simpson's diversity within the k-th class of $\mathbf{X}$. Conversely, the CA-richness strategy gives $I_k = \sum_{j=1}^{p} 1/f_j \cdot f_{j/k}(1 - f_{j/k}) = \sum_{j=1}^{p} (f_{j/k} - f_{j/k}^2)/f_j$, and therefore more weight to the species that are rare in the overall community but abundant in class k, and less weight to the species that are common in the overall community but rare in class k.

APPLICATION TO FOREST ECOLOGICAL DATA

The data were obtained in a 10-ha forest plot at the Piste de St-Elie station in the lowland rainforest of French Guiana (transect B in Sabatier et al. 1997). In this paper we have considered all the trees (381 individuals) with a diameter at breast height (d.b.h.) ≥ 50 cm. These corresponded to 113 species (species nomenclature follows Boggan et al. 1997) of which 2 comprised more than 20 individuals and 97 less than 5 individuals. Taxonomic data were arranged as a complete binary species occurrence table, with 381 tree occurrences as rows and 113 species as columns, secondarily centred by columns (species) in table **Tc**.

Break down of diversity with respect to soil classes

A soil map of the plot (Guillaume 1992) was used to allocate each tree to one of 9 soil classes determined in relation to both a weathering transformation sequence of the initial ferralitic cover and a gradient of increasing hydromorphy (Table 1). Soil data were arranged in a complete binary table **S**, with 381 tree occurrences as rows and 9 soil classes as columns.

We compared both the CA-richness and the NSCA-Simpson strategies applied to the centred-by-species occurrence table **Tc**, and to the approximated and residual tables, $\mathbf{Tc_S}$ and $\mathbf{Tc}_{|\mathbf{S}}$, obtained from multiple linear regressions of the columns in **Tc** on all columns of table **S** containing the soil variables. The results are given in Table 2. The inertia of **Tc** measures the total species diversity of the community, i.e. species richness – 1 = 112 from the CA-richness strategy and Simpson's $D = 0.9331$ from the NSCA-Simpson strategy. The inertias of $\mathbf{Tc_S}$ and $\mathbf{Tc}_{|\mathbf{S}}$ measure β- and α-diversity defined by the soil classes.

In both analyses, the ratio $I(\mathbf{Tc_S})/I(\mathbf{Tc})$, i.e. the proportion of the total species diversity of the community explained by the soil classes, was low (< 5%) but highly statistically significant using the NSCA-Simpson strategy (row permutation test: $P < 0.001$). In this very diverse forest characterized by an important cortège of scarce species, the NSCA-Simpson

strategy places emphasis on the most abundant species and allowed the soil classes to explain a portion of the total species diversity that was about 2 fold that given by the CA-richness strategy (4.02 *vs.* 2.08%). This led to a more accurate characterization of the floristic structure on the two first factorial axes in the former case than in the later (axes 1 and 2 correspond to 71.25% *vs.* 44.62% of the inertia of table $\mathbf{Tc_S}$).

Table 1. Key for the soil classes. The soil sequence from the initial ferralitic cover (*DVD*) up to the transformed hydromorphic systems (*DhS* and *UhS*) characterized a weathering transformation process under mechanical erosion. *SH* corresponds to the periodically flooded bottomlands and is relatively independent of the weathering process (see Sabatier et al. 1997 for more details). *fr* = relative frequency of trees ≥ 50 cm d.b.h. per soil type in a 10-ha rainforest plot in French Guiana.

Code	Description	*fr*
DVD	deep vertical drainage	0.312
Alt	weathered material at a depth less than 1.2 m	0.194
SLD1	superficial lateral drainage with "dry to the touch" character (DC) between 1 and 1.2 m depth	0.0604
SLD2	superficial lateral drainage with DC at less than 1 m depth	0.0814
DhS	downhill transformed hydromorphic system	0.139
DhS+DC	DhS with DC at less than 1.2 m depth	0.0446
UhS	uphill transformed hydromorphic system	0.0656
UhS+DC	UhS with DC at less than 1.2 m depth	0.0472
SH	prolonged surface water saturation	0.0551

Table 2. Inertia decomposition of a centred-by-species occurrence table (**Tc**, with 381 occurrences of trees ≥ 50 cm d.b.h. as rows and 113 species as columns) analyzed with respect to 9 soil types (table **S** of qualitative explanatory variables) in a 10-ha rainforest plot in French Guiana, using the CA-richness ($\mathbf{D}_p^{-1}$ as species weights) and the NSCA-Simpson strategy ($\mathbf{I}_p$ as species weights). $I(\mathbf{Tc_S})/I(\mathbf{Tc})$ is the proportion of the total species diversity explained by the soil variables. Row permutation tests (Manly 1991): *n.s.* = non-significant; *** $P < 0.001$.

Ordination strategy	Total species diversity $I(\mathbf{Tc})$	Diversity explained by soil $I(\mathbf{Tc_S})$	Diversity unexplained by soil $I(\mathbf{Tc_{\vert S}})$	$\frac{I(\mathbf{Tc_S})}{I(\mathbf{Tc})}$
CA-richness	112	2.33 *n.s.*	109.67	2.08%
NSCA-Simpson	0.9331	0.0376 ***	0.896	4.02%

Changing species weights also reversed the hierarchy between the main axes obtained through the analysis of $\mathbf{Tc_S}$ (Fig. 2). Indeed, the first axis resulting from the CA-richness strategy (26.3% of $\mathbf{Tc_S}$ inertia) underlined the originality of the periodically flooded bottomlands (SH) while the second axis (18.3% of $\mathbf{Tc_S}$ inertia) contrasted the ferralitic soils (DVD) with the weathered soil classes (Fig. 2a). On the other hand, the NSCA-Simpson strategy (Fig. 2b) yielded a very prominent axis 1 (60.3% of $\mathbf{Tc_S}$ inertia) expressing more legibly the sequence of soil weathering that led from untransformed DVD to highly weathered soil classes (Uhs, UhS+DC). It is noteworthy that the α-diversity of the soil classes (represented by gray circles in Fig. 2a-b left) displayed more variations between soil classes with the CA-richness than with the NSCA-Simpson strategy. In particular, SLD1 and

UhS+DC exhibited very low α-diversity in the CA-richness strategy, indicating that they harbored less rare species than the other soil classes. The relative species loadings on axis 1 (represented by gray circles in Fig. 2a-b right) highlighted the fact that influential species in the CA-richness strategy were mainly rare species confined to SH. By contrast, more common species played a prominent role in the NSCA-Simpson strategy. In particular, the distribution of *Eperua falcata*, an abundant and ubiquitous species in this region of French Guiana, corresponded in the study plot to the weathering soil sequence. This points towards a floristic pattern of a broader significance than the sole specificity of the flooded locations (Fig. 3).

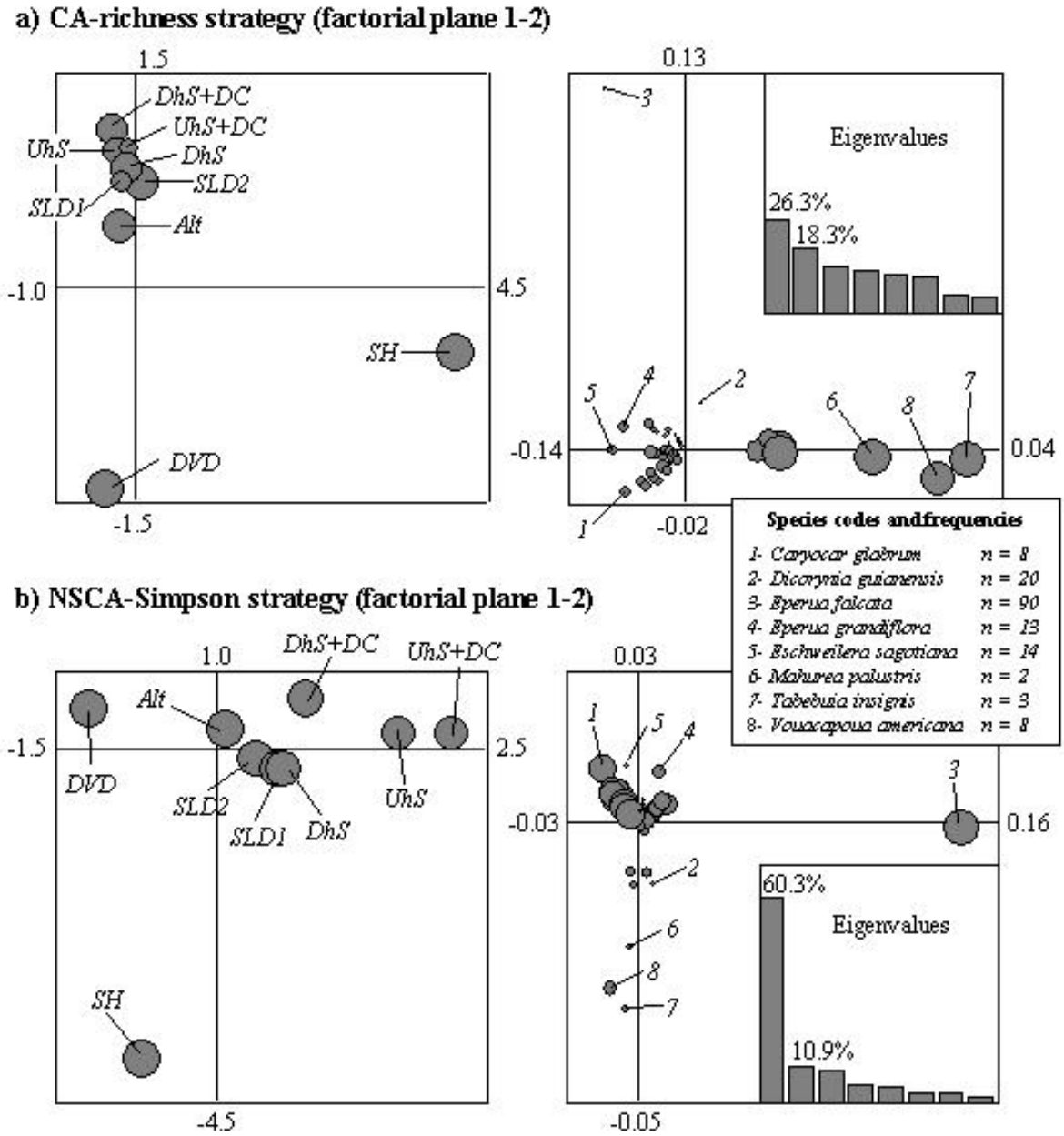


Figure 2. Ordination of the approximated table $\mathbf{Tc_S}$ (with 381 occurrences of trees ≥ 50 cm d.b.h. as rows and 113 species as columns) derived from the analysis of a centred-by-columns occurrence table (**Tc**) with respect to 9 qualitative explanatory soil variables (table **S**) in a 10-ha rainforest plot in French Guiana. **a)** CA-richness strategy ($\mathbf{D}_p^{-1}$as species weights); **b)** NSCA-Simpson strategy ($\mathbf{I}_p$ as species weights). The soil classes and species are positioned by averaging at the weighted mean of their occurrences. Gray circles: within-class α-diversity of the soil classes (left figures); species relative contributions to axis 1 (right figures). The key for the soil classes is given in Table 1.

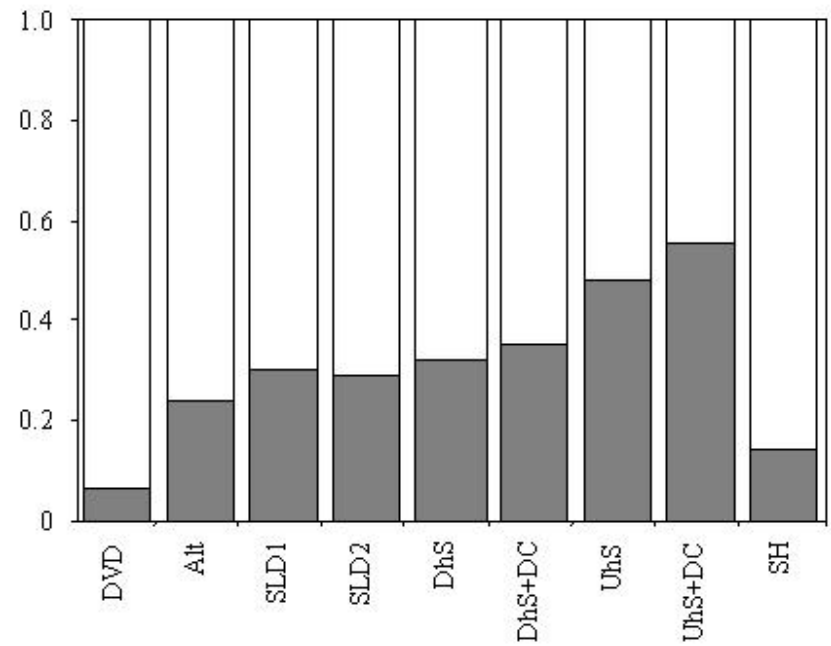


Figure 3. Relative frequency of trees ≥ 50 cm d.b.h. belonging to *Eperua falcata* (gray) and to other species (white) within 9 soil classes in a 10-ha rainforest plot in French Guiana. The key for the soil classes is given in Table 1.

Partitioning the occurrences into quadrats

In order to illustrate the possible extension of the method to classical ecological relevés, we partitioned the plot into contiguous quadrats. These quadrats may be considered to be homologous to sampling units of a limited area that are often used to collect individual trees in forest inventories. Three partitions with quadrat sizes of 20 m x 20 m, 25 m x 25 m and 25 m x 50 m were used. Each gave a complete binary table **Q** of explanatory variables, with 381 tree occurrences as rows and as many columns as quadrats (i.e. 250, 160 and 80 columns for the three partitions). The results of the analysis of **Tc**, $\mathbf{Tc_Q}$ and $\mathbf{Tc_{|Q}}$ using the CA-richness and the NSCA-Simpson strategies are given in Table 3.

Table 3. Inertia decomposition of a centred-by-species occurrence table (**Tc**, with 381 occurrences of trees ≥ 50 cm d.b.h. as rows and 113 species as columns) analyzed with respect to explanatory spatial variables (table **Q**) in a 10-ha rainforest plot in French Guiana, using the CA-richness ($\mathbf{D}_p^{-1}$ as species weights) and the NSCA-Simpson strategy ($\mathbf{I}_p$ as species weights). $I(\mathbf{Tc_Q})/I(\mathbf{Tc})$ is the proportion of the total species diversity explained by quadrats. Row permutation tests (Manly 1991): *n.s.* = non-significant; ** $P < 0.01$.

Ordination strategy	Total species diversity $I(\mathbf{Tc})$	Diversity explained by quadrats $I(\mathbf{Tc_Q})$	Diversity unexplained by quadrats $I(\mathbf{Tc_{\vert Q}})$	$\frac{I(\mathbf{Tc_Q})}{I(\mathbf{Tc})}$
80 quadrats of 50 m x 25 m				
CA-richness	112	23.63 *n.s.*	88.37	21.10%
NSCA-Simpson	0.9331	0.2137 **	0.7194	22.90%
160 quadrats of 25 m x 25 m				
CA-richness	112	44.90 *n.s.*	67.10	40.09%
NSCA-Simpson	0.9331	0.3847 **	0.5483	41.23%
250 quadrats of 20 m x 20 m				
CA-richness	112	60.63 *n.s.*	51.37	54.13%
NSCA-Simpson	0.9331	0.5049 *n.s.*	0.4282	54.11%

The proportion of the total species diversity of the community explained by the partition into quadrats was high and very similar with both strategies: $I(\mathbf{Tc_Q})/I(\mathbf{Tc})$ ranged from 21.10% to 54.13% using the CA-richness strategy and from 22.90% to 54.11% using the NSCA-Simpson strategy. Logically, this proportion always decreased with quadrat area since α-diversity increased with quadrat size. However, $I(\mathbf{Tc_Q})$ was only statistically significant for quadrats of 50 m x 25 m and 25 m x 25 m using the NSCA-Simpson strategy (row permutation tests: $P < 0.01$). Hence, the only significant feature of β-diversity among quadrats consisted of variations in the abundance of the most common species, and required a quadrat area in excess of 25 m x 25 m.

Factorial axes were computed from the $\mathbf{Tc_Q}$ table approximated by the partition into 80 quadrats measuring 50 m x 25 m (Fig. 4). Although ordinations were not constrained by the soil variables, the factorial planes – when featuring projections of soil classes – were remarkably similar to those obtained from the analyses of $\mathbf{Tc_S}$ (see Fig. 2 and Fig. 4). This was also the case when using a partition into smaller quadrats (not shown). Hence, despite $I(\mathbf{Tc_Q}) > I(\mathbf{Tc_S})$, the analysis of $\mathbf{Tc_Q}$ was not more informative in terms of floristic structures. The distribution of eigenvalues demonstrated, moreover, that the CA-richness approach was far less efficient at extracting meaningful patterns from $\mathbf{Tc_Q}$ than from $\mathbf{Tc_S}$, though the NSCA-Simpson strategy performed fairly well, even on $\mathbf{Tc_Q}$. This latter analysis yielded the most notable departure from Fig. 2 through its second axis (Fig. 4b) which contrasted upland quadrats (negative part) from slopes and lowlands quadrats (positive part) on the basis of the abundance of the second most common species, *Dicorynia guianensis*, whose aggregated spatial distribution (Collinet 1997) was more largely explained by the quadrat partition (20.43%) than the soil classes (5.15%).

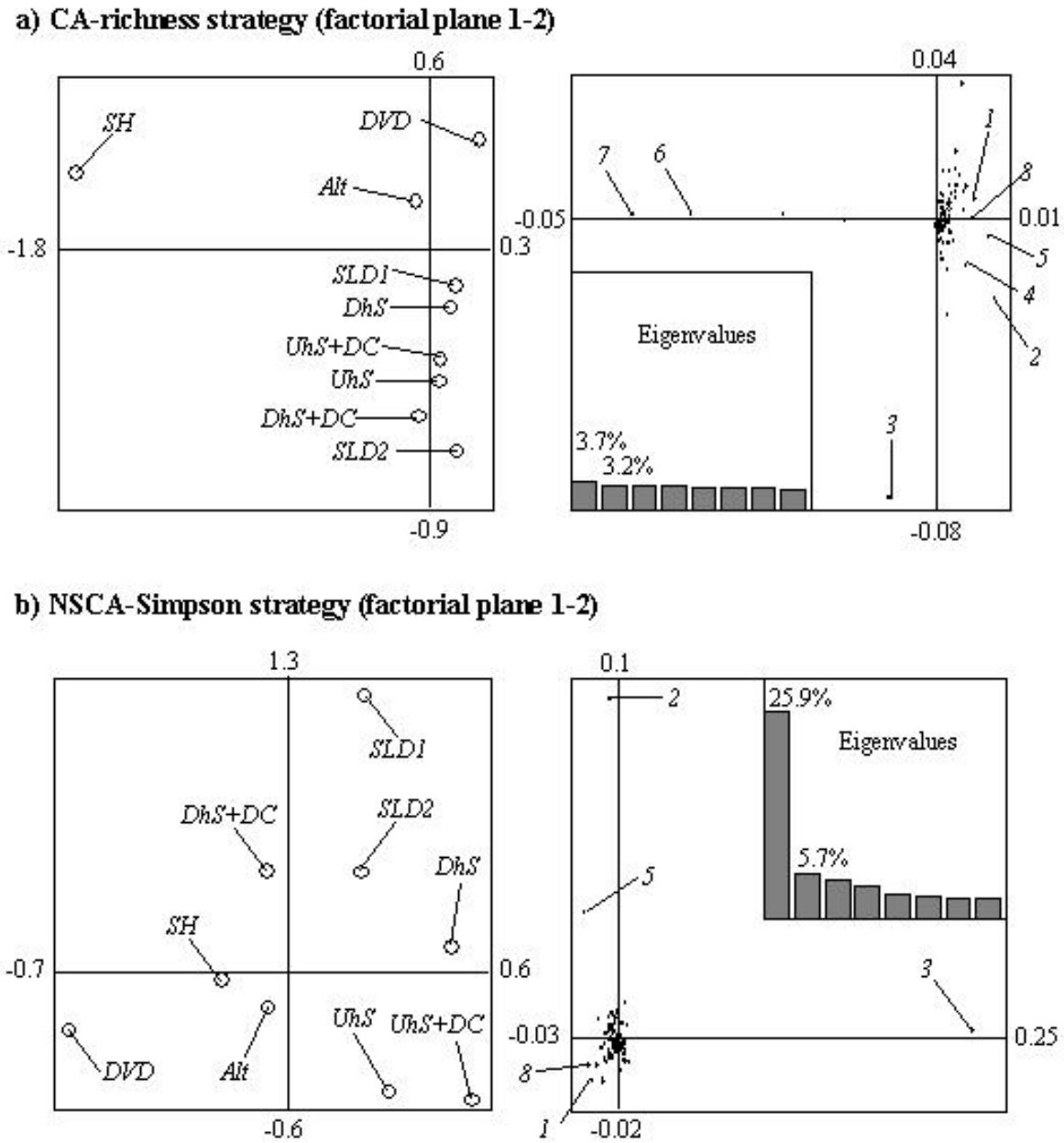


Figure 4. Ordination of the approximated table $\mathbf{Tc_Q}$ (with 381 occurrences of trees ≥ 50 cm d.b.h. as rows and 113 species as columns) derived from the analysis of a centred-by-columns occurrence table (**Tc**) with respect to 80 qualitative explanatory spatial variables (table **Q** of quadrats of 50 m x 25 m) in a 10-ha rainforest plot in French Guiana. **a)** CA-richness strategy ($\mathbf{D}_p^{-1}$as species weights); **b)** NSCA-Simpson strategy ($\mathbf{I}_p$ as species weights). The soil classes and species are positioned by averaging at the weighted mean of their occurrences. The key for the soil classes is given in Table 1. Key for species is the same as in Fig. 2.

Comparison with classical ordinations based on quadrats

We demonstrated (*Appendix*) that analyzing $\mathbf{Tc_Q}$ through either CA or NSCA is strictly equivalent to performing these analyses on a more classical quadrats-by-species contingency table. Furthermore, such a contingency table can be coupled with an ancillary table containing soil information, via Canonical Correspondence Analysis (CCA; Ter Braak 1986) or via a form of RDA consistent with NSCA. A table conveying soil information can be constructed from our data by assigning to each quadrat its modal soil type (binary table) or the relative importance of soil types within it (quantitative table based on occurrences). By

accepting such a loss of information we are now back in a framework that will appear more usual to most vegetation scientists, and this yields the inertia decomposition presented in Table 4.

Table 4. Inertia decomposition of a quadrats-by-species contingency table (with 80 quadrats of 50 m x 25 m as rows and 113 species as columns) analyzed with respect to an explanatory soil table (with 80 quadrats as rows and 9 soil classes as columns) in a 10-ha rainforest plot in French Guiana, using the CA-richness and NSCA-Simpson strategies. The soil table is either a binary table assigning to each quadrat its modal soil type (Bin) or a quantitative table assigning to each quadrat the occurrence relative frequency per soil type (Quant). The proportion of explained inertia is given in parentheses. CA: Correspondence Analysis; NSCA: Non-Symmetric Correspondence Analysis; CCA: Canonical Correspondence Analysis; RDA: Redundancy Analysis.

CA-richness strategy			NSCA-Simpson strategy		
CA	CCA		NSCA	RDA	
	Bin	Quant		Bin	Quant
23.63	2.47 (10.43%)	2.57 (10.89%)	0.2137	0.0300 (14.05%)	0.0365 (17.08%)

The proportion of the inertia explained by the soil variables was 10.43% (binary table) and 10.89% (quantitative table) from CCA, and 14.05% (binary table) and 17.08% (quantitative table) from RDA. The explanatory power of the soil factor may appear more convincing from these values than from the results given in Table 2. However, the proportion of the inertia explained by the soil table in CCA and RDA conveys the same intuitive meaning as the proportion of the inertia of the approximation by **S** of the approximation of **Tc** by **Q**, i.e. of $\mathbf{Tc_{QonS}}$ obtained from the multiple linear regressions of the columns in $\mathbf{Tc_Q}$ on all columns of **S** (there is no strict mathematical equivalence here since there are at least two ways to define $\mathbf{Tc_{QonS}}$ from CCA; Méot et al. 1998). Anyhow, it is clear that CCA and RDA, which start from the quadrats-by-species table, measure β-diversity between quadrats, i.e. $I(\mathbf{Tc_Q})$, instead of the total species diversity of the community, and thus ignore quadrats α-diversity. One advantage of the occurrence-based approach, even for species-by-relevés data, is therefore to provide a fully explicit breakdown of the total floristic diversity of a community, $I(\mathbf{Tc})$, as quantified by usual indices.

DISCUSSION

We have presented two main strategies for the analysis of species occurrence data with respect to environmental or instrumental information. Each strategy encompasses an ordination technique (CA *vs.* NSCA) and a consistent measurement of species diversity (species richness *vs.* Simpson diversity). The CA-richness strategy places emphasis on scarce species while the NSCA-Simpson strategy primarily relies on abundant species. Both strategies enable successive decompositions of the total inertia of a species occurrence table, and thus of the total diversity of the community, between fractions that are explained by certain environmental variables and fractions that are not. A similar scheme of successive decompositions based on CA has already been proposed by Ter Braak (1986, 1987, 1988) for taxonomic data originating from field plots or relevés. This is usually referred to as the Canonical Correspondence Analysis (CCA), though Correspondence Analysis on

Instrumental Variables (CAIV) may be preferable (Sabatier et al. 1989, Dray 2001). However, this approach does not rely directly on the floristic diversity of a community as quantified by usual indices.

The paper given here provides a wider scope based on the notion of species occurrences. Indeed, through the choice of species weighting, our approach allowed us equating the taxonomic table inertia with common diversity indices. It also permits the use of a wide range of sampling schemes as individuals enumerated in independent field plots (e.g. quadrats, relevés), can be considered as particular types of occurrence data grouped according to an external qualitative spatial variable that codes the plots. This formulation, though unusual, is strictly equivalent to the well-known species-by-relevés table usually considered through CA or CCA when additional environmental variables are available. Subsequent analyses with respect to environmental variables can then be carried out on the fraction of the total inertia explained by the between-relevés analysis (diversity within relevés is ignored as in CCA). This is fully equivalent to directly starting the process from the species-by-relevés table, as via CCA. Even at this stage, the choice between the CA-richness and the NSCA-Simpson strategy remains since CCA can be paralleled by a RDA using uniform species weights. Each strategy has strengths and weaknesses. Hence, the choice between the two depends on community characteristics and on the nature of the information to be analyzed.

In fact, a species occurrence table is a unifying mathematical formulation that can be constructed from very diverse field data. Occurrences may correspond to species observations made by successive generations of field investigators without any consistent sampling scheme (e.g. data from Museum or Herbarium collections; see for example Gimaret-Carpentier et al. 1998a). Exhaustive sampling of large field plots, such as the 10-ha plot presented in this paper, also produces a table for which an occurrence corresponds to an individual tree. The last kind of occurrence data derives from the widely used sampling through plots or relevés within which individuals-by-species are enumerated.

The relevance of the CA-richness strategy *vs.* NSCA-Simpson strategy greatly depends on the aspect of the plant (or animal) community that one may want to emphasize and also on the broad characteristics of the community under study. For instance, in very rich vegetation types such as the tropical rainforest, an estimation of species richness can neither be precise nor unbiased on a local scale: all scarce species that can grow in a given context would not be present in a sampling plot of limited area (Condit et al. 1996, Gimaret-Carpentier et al. 1998b). Consequently, the relative abundance of scarce species and the total inertia of the CA-richness strategy are ill-defined. Hence, the NSCA-Simpson strategy is likely to provide more robust species-by-environment relationships than the CA-richness approach. In other words, scarce species are less reliable than abundant species when partitioning local diversity into α and β components. Analyzing occurrences of scarce species is nevertheless useful when assessing and comparing γ-diversity on a regional scale. This can be done by compiling existing information from local ecological studies as well as large-scale botanical surveys. At this level, the heterogeneity of the data casts doubts on the relevance of the NSCA-Simpson strategy since in the absence of a consistent sampling design there is no guaranty that the most frequent species in the data set are also the most frequent in the region under consideration.

Indeed, an important criterion when choosing species weightings is the reliability of the information conveyed by the absence of a given species in a certain environmental situation (soil class, geographical quadrat, etc.). In the absence of a thorough sampling scheme, such information is dubious since some species could have been missed. Furthermore, the relative abundance of species in the data set is likely to be biased with respect to their relative abundance in the field. The use of the χ^2 metric, through species weights given by $\mathbf{D}_p^{-1}$, is hence highly reasonable and the CA-richness, which guarantees that

the species absent from a given environmental modality do not contribute to its ordination score (Legendre and Legendre 1998), is obviously to be preferred. On the other hand, several sampling schemes (either random or systematic; Cochran 1977) may warrant a fair approximation of the species relative abundance in the community under study. This is also true with exhaustive sampling within a large plot, at least for the plot itself. In such cases, the absence of a species in a given environmental context is truly informative and the closer the relative frequency of a species to 0.5 the more information it conveys. This is fully consistent with a measurement of diversity through the Simpson index, as well as with the use of the Euclidian metric, i.e. of uniform species weights given by $\mathbf{I}_p$ instead of $\mathbf{D}_p^{-1}$. Hence, the NSCA-Simpson strategy appears to be more relevant than the CA-richness strategy, and is likely to provide greater insight into the data set.

The link between the sampling scheme and the choice of an ordination technique has already been emphasized with respect to the weighting of sampling units and to its incidence on species niche measurement by Dolédec et al. (2000). Thus, it has been recognized that CA (or CCA) can be used to compensate for unequal sampling efforts, while data from equitable sampling efforts can be analyzed by other ordination techniques, of which NSCA on species profiles.

Although this paper grew out of efforts to analyze floristic data based on species distributions, it should be emphasized that the overall principles apply to any kind of organism and/or taxonomic level.

COMPUTATION

All analyses presented in this paper have been performed using ADE-4 package (Thioulouse et al. 1997) freely available at http://pbil.univ-lyon1.fr/ADE-4/. A user's manual to CA-richness and NSCA-Simpson strategies (Pélissier et al. 2002b) is also available in PDF format from the web site mentioned above as volume 3.9 of the topic documentation of ADE-4.

ACKNOWLEDGEMENTS

This paper stems from important insights provided by D. Chessel through informal discussions and also from documentation accompanying ADE-4 software. This work was partly supported by the GIS Silvolab-Guyane through the AME-Project directed by D. Sabatier. The authors are very grateful to all the botanists and soil scientists who participated in the enumeration and mapping of the 10-ha plot at Piste de St-Elie. They also wish to thank their colleagues, and particularly D. W. Roberts, P. Legendre, B. McCune and two anonymous referees who made very useful comments on the manuscript.

LITERATURE CITED

Begon, M., J. L. Harper, and C. R. Townsend. 1996. Ecology: individuals, populations and communities. Blackwell Science Ltd, Oxford, UK.

Boggan, J., V. Funk, C. Kelloff, M. Hoff, G. Cremers, and C. Feuillet. 1997. Cheklist of the plants of the Guianas (Guyana, Surinam, French Guiana). National Museum of Natural History, Smithonian Institution, Washington, USA.

Cochran, W. G. 1977. Sampling techniques. Wiley, New York, USA.

Condit, R., S. P. Hubell, J. V. LaFrankie, R. Sukumar, N. Manokaran, R. B. Foster, and P. S. Ashton. 1996. Species-area and species-individual relationships for tropical trees: a comparison of three 50-ha plots. Journal of Ecology **84**: 549–562.

Collinet, F. 1997. Essai de regroupement des principales espèces structurantes d'une forêt dense humide d'après l'analyse de leur répartition spatiale (Forêt de Paracou, Guyane). Thèse de Doctorat, Université Claude Bernard, Lyon, France.

Davies, P. T., and M. K.-S. Tso. 1982. Procedures for reduced-rank regression. Applied Statistics **31**: 244-255.

Dolédec, S., and D. Chessel. 1989. Rythmes saisonniers et composantes stationnelles en milieu aquatique. II-Prise en compte et élimination d'effets dans un tableau faunistique. Acta Oecologica **10**: 207–232.

Dolédec, S., D. Chessel, C. J. F. Ter Braak, and S. Champely. 1996. Matching species traits to environmental variables: a new three-table ordination method. Environmental and Ecological Statistics **3**: 143–166.

Dolédec, S., D. Chessel, and C. Gimaret-Carpentier. 2000. Niche separation in community analysis: a new method. Ecology **81**: 2914–2927.

Dray, S. 2001. The CoCoAn (Constrained Correspondence Analysis) R Package. http://lib.stat.cmu.edu/R/CRAN/src/contrib/PACKAGES.html#CoCoAn.

Escoufier, Y. 1987. The duality diagram: a means of better practical applications. Pages 139–156 *in* P. Legendre. and L. Legendre, editors. Development in numerical ecology, Springer-Verlag, Berlin, Germany.

Gauch, H. G., Jr. 1973. The relationship between sample similarity and ecological distance. Ecology **54**: 618–622.

Gauch, H. G., Jr., and R. H. Whittaker. 1972. Coenocline simulation. Ecology **53**: 446–451.

Gimaret-Carpentier, C., D. Chessel, and J.-P. Pascal. 1998a. Non-symmetric correspondence analysis: an alternative for species occurrences data. Plant Ecology **138**: 97–112.

Gimaret-Carpentier, C., R. Pélissier, J.-P. Pascal, and F. Houllier. 1998b. Sampling strategies for the assessment of tree species diversity. Journal of Vegetation Science **9**: 161–172.

Greenberg, J. H. 1956. The measurement of linguistic diversity. Language **32**: 109–115.

Greenacre, M. J. 1984. Theory and applications of correspondence analysis. Academic Press, London, UK.

Guillaume, J. 1992. Cartographie du sol sous forêt naturelle en Guyane française. Influence des caractères pédologiques sur la structure de la forêt : étude préliminaire. Mémoire de DEA, ENSA, Rennes, France.

Hill, M. O. 1973. Diversity and evenness: a unifying notation and its consequences. Ecology **54**: 427–432.

Hill, M. O. 1974. Correspondence analysis: a neglected multivariate method. Journal of the Royal Statistical Society **23**: 340–354.

Krebs, C. J. 1994. Ecology: the experimental analysis of distribution and abundance. Harper Collins College Publishers, New-York, USA.

Lande, R. 1996. Statistics and partitioning of species diversity, and similarity among multiple community. Oikos **76**: 5–13.

Lauro, N., and L. D'Ambra. 1984. L'analyse non symétrique des correspondances. Pages 433–446 *in* E. Diday, editor. Data analysis and informatics *III*, Elsevier, The Netherlands.

Legendre, P., and L. Legendre. 1998. Numerical ecology. Elsevier, Amsterdam, The Netherlands.

Magurran, A. E. 1988. Ecological diversity and its measurement. Croom Helm Ltd, London, UK.

Manly, B. J. F. 1991. Randomization and Monte Carlo methods in biology. Chapman and Hall, London, UK.

Méot, A., P. Legendre, and D Borcard. 1998. Partialling out the spatial component of ecological variation: questions and propositions in the linear modelling framework. Environmental and Ecological Statistics **5**: 1-27.

Patil, G. P., and C. Taillie. 1982. Diversity as a concept and its measurement. Journal of the American Statistical Association **77**: 548–561. Pélissier, R., S. Dray, and D. Sabatier. 2002a. Within-plot relationships between tree species occurrences and hydrological soil constraints: an example in French Guiana investigated through canonical correlation analysis. Plant Ecology: in press.

Pélissier, R., S. Dray, and P. Couteron. 2002b. User's manual to CA-richness and NSCA-Simpson strategies. ADE-4 topic documentation, Vol. 3.9, http://pbil.univ-lyon1.fr/ADE-4/.

Pielou, E. C. 1969. An introduction to mathematical ecology. John Wiley and Sons, New-York, USA.

Rao, C. R. 1964. The use and interpretation of principal component analysis in applied research. Sankhya **A26**: 329–357.

Ricklefs, E. 1990. Ecology. W.H. Freeman and Company, New-York, USA.

Sabatier, D., M. Grimaldi, M.-F. Prévost, J. Guillaume, M. Godron, M. Dosso, and P. Curmi. 1997. The influence of soil cover organization on the floristic and structural heterogeneity of a Guianan rain forest. Plant Ecology **131**: 81–108.

Sabatier, R. 1984. Quelques généralisations de l'analyse en composantes principales de variables instrumentales. Statistique et Analyse de Données **9**: 75–103.

Sabatier, R., J.-D. Lebreton, and D. Chessel. 1989. Principal component analysis with instrumental variables as a tool for modelling composition data. Pages 341–352 *in* R. Coppi and S. Bolasco, Editors. Multiway data analysis, Elsevier Science Publishers, The Netherlands.

Shannon, C. E. 1948. A mathematical theory of communication. Bell System Technical Journal **27**: 379–423.

Simpson, E. H. 1949. Measurement of diversity. Nature **163**: 688.

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

Ter Braak, C.J.F. 1983. Principal component biplots and alpha and beta diversity. Ecology **64**: 454-462.

Ter Braak, C. J. F. 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology **67**: 1167–1179.

Ter Braak, C. J. F. 1987. The analysis of vegetation-environment relationships by canonical correspondence analysis. Vegetatio **69**: 69–77.

Ter Braak, C. J. F. 1988. Partial canonical correspondence analysis. Pages 551–558 *in* H. H. Bock, Editor. Classification and related methods of data analysis, North Holland Press, Amsterdam, The Netherlands.

Thioulouse, J., D. Chessel, S. Dolédec, and J.-M. Olivier. 1997. ADE-4: a multivariate analysis and graphical display software. Statistics and Computing **7**: 75-83.

Whittaker, C. B. 1972. Evolution and measurement of species diversity. Taxon **21**: 213–251.

Wollenberg, A. L. 1977. Redundancy analysis. An alternative for canonical correlation analysis. Psychometrika **42**: 207-219.

APPENDIX

1) *Triplet notation*. Let **Tc** be a $(n \times p)$ matrix derived from any data table, $\mathbf{D}_n$ be a $(n \times n)$ diagonal matrix containing the associated row weights, and Δ_p be a $(p \times p)$ diagonal matrix containing the associated metric (i.e. column weights). The resulting statistical triplet (*sensu* Escoufier 1987) is (**Tc**, $\mathbf{\Delta}_p$, $\mathbf{D}_n$). The generalized singular value decomposition (GSVD; Greenacre 1984) of (**Tc**, $\mathbf{\Delta}_p$, $\mathbf{D}_n$) consists in finding right-hand eigenvectors of $\mathbf{Tc}^t\mathbf{D}_n\mathbf{Tc}\mathbf{\Delta}_p$ and $\mathbf{Tc}\mathbf{\Delta}_p\mathbf{Tc}^t\mathbf{D}_n$:

$$svd\left(\mathbf{Tc}^t\mathbf{D}_n\mathbf{Tc}\mathbf{\Delta}_p\right) = \mathbf{U}_p\mathbf{D}_\lambda\mathbf{V}_p^t$$

$$svd\left(\mathbf{Tc}\mathbf{\Delta}_p\mathbf{Tc}^t\mathbf{D}_n\right) = \mathbf{U}_n\mathbf{D}_\lambda\mathbf{V}_n^t$$

where, $\mathbf{D}_\lambda$ is the diagonal matrix of eigenvalues, such that $Iner\left(\mathbf{Tc}, \mathbf{\Delta}_p, \mathbf{D}_n\right) = tr\left(\mathbf{D}_\lambda\right)$.

Rows (**y**) and columns (**x**) scores are then computed as:

$$\mathbf{x} = \mathbf{Tc}\mathbf{\Delta}_p\mathbf{V}_p \text{ and } \mathbf{y} = \mathbf{Tc}^t\mathbf{D}_n\mathbf{V}_n$$

2) *Triplet decomposition*. Consider **X** a $(n \times m)$ matrix of explanatory variables. The general case of multiple linear regressions can be expressed in terms of orthogonal projection operators (Takeuchi et al. 1982):

$$\mathbf{Tc} = \mathbf{Tc_X} + \mathbf{Tc_{|X}} = \mathbf{P_X}(\mathbf{Tc}) + (\mathbf{Tc} - \mathbf{P_X}(\mathbf{Tc}))$$

where $\mathbf{P_X}$ is the projection operator onto **X** and is equal to:

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

The GSVD of (**Tc**, $\mathbf{\Delta}_p$, $\mathbf{D}_n$) can be partitioned as:

$$\begin{aligned}\mathbf{Tc}^t\,\mathbf{D}_n\,\mathbf{Tc}\,\Delta_p &= (\mathbf{Tc_X} + \mathbf{Tc_{|X}})^t\mathbf{D}_n(\mathbf{Tc_X} + \mathbf{Tc_{|X}})\Delta_p \\ &= \mathbf{Tc_X}^t\mathbf{D}_n\mathbf{Tc_X}\Delta_p + \mathbf{Tc_X}^t\mathbf{D}_n\mathbf{Tc_{|X}}\Delta_p + \mathbf{Tc_{|X}}^t\mathbf{D}_n\mathbf{Tc_X}\Delta_p + \mathbf{Tc_{|X}}^t\mathbf{D}_n\mathbf{Tc_{|X}}\Delta_p \\ &= \mathbf{Tc_X}^t\mathbf{D}_n\mathbf{Tc_X}\Delta_p + \mathbf{Tc_{|X}}^t\mathbf{D}_n\mathbf{Tc_{|X}}\Delta_p\end{aligned}$$

This decomposition yields two statistical triplets ($\mathbf{Tc_X}$, $\mathbf{\Delta}_p$, $\mathbf{D}_n$) and ($\mathbf{Tc_{|X}}$, $\mathbf{\Delta}_p$, $\mathbf{D}_n$).

3) *Relationships between occurrence and contingency tables*. If **Tc** is a centred-by-columns occurrence table and **X** contains dummy variables corresponding to the modalities of a single qualitative variable, then between-analysis is the GSVD of ($\mathbf{Tc_X}$, $\mathbf{\Delta}_p$, $\mathbf{D}_n$), which can be expressed by introducing the projection operator onto **X** as:

$$\begin{aligned}\mathbf{Tc_X}^t\mathbf{D}_n\mathbf{Tc_X}\Delta_p &= \mathbf{P_X}(\mathbf{Tc})^t\mathbf{D}_n\mathbf{P_X}(\mathbf{Tc})\Delta_p \\ &= (\mathbf{X}(\mathbf{X}^t\mathbf{D}_n\mathbf{X})^{-1}\mathbf{X}^t\mathbf{D}_n\mathbf{Tc})^t\mathbf{D}_n\mathbf{X}(\mathbf{X}^t\mathbf{D}_n\mathbf{X})^{-1}\mathbf{X}^t\mathbf{D}_n\mathbf{Tc}\Delta_p \\ &= \mathbf{Tc}^t\mathbf{D}_n\mathbf{X}(\mathbf{X}^t\mathbf{D}_n\mathbf{X})^{-1}\mathbf{X}^t\mathbf{D}_n\mathbf{X}(\mathbf{X}^t\mathbf{D}_n\mathbf{X})^{-1}\mathbf{X}^t\mathbf{D}_n\mathbf{Tc}\Delta_p \\ &= \mathbf{Tc}^t\mathbf{D}_n\mathbf{X}(\mathbf{X}^t\mathbf{D}_n\mathbf{X})^{-1}\mathbf{X}^t\mathbf{D}_n\mathbf{Tc}\Delta_p \\ &= \mathbf{P_0}^t\mathbf{D}_m^{-1}\mathbf{P_0}\Delta_p\end{aligned}$$

where: $\mathbf{P_0} = \mathbf{X}^t\mathbf{D}_n\mathbf{Tc} = \left[f_{kj} - f_k f_j\right]$ is the $(m \times p)$ centred table of the relative frequencies f_{kj} of the occurrences that belong to modality k and species j; f_k is the marginal frequency of modality k and f_j the marginal frequency of species j; $\mathbf{D}_m^{-1} = (\mathbf{X}^t\mathbf{D}_n\mathbf{X})^{-1}$ is a $(m \times m)$ diagonal matrix containing modality weights equal to $1/f_k$.

The GSVD of ($\mathbf{P_0}$, $\mathbf{D}_p^{-1}$, $\mathbf{D}_m^{-1}$) is a Correspondence Analysis (CA; Hill 1974), while the GSVD of ($\mathbf{P_0}$, $\mathbf{I}_p$, $\mathbf{D}_m^{-1}$) is a Non-Symmetric Correspondence Analysis (NSCA) on row profiles (Lauro and D'Ambra 1984).

Chapitre V : Modélisation d'une maladie à transmission vectorielle

Ce chapitre est constitué d'une publication à paraître dans un numéro spécial de la revue *Preventive Veterinary Medicine*. Une modélisation de la distribution spatiale de la prévalence de trypanosomose bovine, dans une zone agropastorale du Burkina-Faso, y est proposée. Cette modélisation est fondée sur l'utilisation d'analyses spatiales et de méthodes statistiques. En s'appuyant sur un recensement de l'ensemble des animaux et des fermes de la zone d'étude, des échantillons sanguins ont été prélevés sur plus de 2000 bovins choisis aléatoirement. Des données relatives aux pratiques d'élevage ont également été collectées durant cette étude. La totalité de l'information a été introduite dans un SIG et de nouvelles variables, représentant des contraintes spatiales de la zone, ont été générées.

La prévalence sérologique a été modélisée au moyen d'une régression logistique. Il a été ainsi possible d'identifier et de quantifier le risque associé à certaines pratiques. La proximité du réseau hydrographique et le type de point d'eau fréquenté, par exemple, ont une influence sur le statut sérologique. Dans une seconde étape, le modèle statistique a été utilisé afin de prédire la prévalence sérologique pour l'ensemble des exploitations de la zone. Une valeur de prévalence prédite a donc été assignée à chaque troupeau en fonction de sa localisation spatiale et des pratiques d'élevage. En tenant compte des mouvements journaliers des troupeaux, les prévalences prédites ont été représentées à l'aide du SIG. Des zones à fort risque épidémiologique ont ainsi pu être identifiées.

Le modèle statistique a été réalisé à l'aide du logiciel R et le SIG a permis la modélisation spatiale ainsi que la représentation cartographique des données et des résultats. L'annexe 2 fournit les éléments techniques de la mise en œuvre de cette modélisation dans R.

Modelling bovine trypanosomosis spatial distribution by GIS in an agro-pastoral zone of Burkina Faso

Jean-François Michel[1], Stéphane Dray[2], Stéphane de La Rocque[1], Marc Desquesnes[1], Philippe Solano[3], Gérard De Wispelaere[4], Dominique Cuisance[5].

[1] : CIRDES/CIRAD-EMVT, Bobo Dioulasso, Burkina Faso
[2] : Université Claude Bernard Lyon I, Villeurbanne, France
[3] : IRD/IPR, Bouake, Côte d'Ivoire
[4] : CIRAD-EMVT, Maison de la télédetection, Montpellier, France
[5] : CIRAD-EMVT, Campus International de Baillarguet, Montpellier, France

Keywords: GIS, spatial modelling, logistic regression, trypanosomosis, epidemiology

Abstract
Modelling of the spatial distribution of Bovine Trypanosomosis prevalence in Sideradougou district Burkina Faso was performed by using a combination of spatial and statistical analysis. Based on a comprehensive and geographically representative census of herds and farms in the area, more than 2000 bovines were randomly chosen and their blood sampled during field survey. Data on livestock farming practices were recorded for each farm. All data were mapped within a GIS to generate new information on spatial constraints in the area.
Surveys results were analysed and serological prevalence data were modelled using logistic regression. The model allowed identification and quantification of risk factors. In a second step the statistical model was used predictively on the entire farm population in the area. This method was successful in predicting the serological prevalence for each individual herd in the sample, from their livestock management patterns and spatial location. Predicted prevalences were represented within the GIS, taking daily movements of animals into account. Spatial distribution of prevalence would illustrate specific locations at risk from an epidemiological viewpoint. It gives evidence that the hydrological network and land occupation patterns in the savannah-type countryside are playing an important part when structuring a so-called Trypanosomosis Space.

1. Introduction

Animal trypanosomosis are one of the main pathological constraints on the development of animal production in sub-Saharan Africa (Hursey and Slinghenbergh, 1995), and cause annual losses estimated at US$ 1 billion (De Haan and Bekure, 1991). Tsetse flies are the main vectors. The risk of transmission is primarily linked to the intensity of the encounters between vectors and hosts, and depends on the spatial and temporal interfaces between the protagonists in the pathogen system (host-vector-parasite) (Laveissière et al., 1986; de La Rocque et al., 1999). High-risk areas have been identified on this basis in an agro-pastoral zone of southern Burkina Faso, taking environmental and socio-economic factors into account. The available data were georeferenced, included into a geographic information system (GIS), and high-risk areas were identified by spatial modelling (de La Rocque et al., 2001), as it was performed at the other scales (Hendrikx et al., 2001).
The serological prevalence of the disease (prevalence of antibodies directed against trypanosomal antigens) was studied on a sample of cattle farms in the study area, to validate the list of epidemiological risk areas identified. However, the data obtained were both partial and spatially disjointed. The method described here was subsequently developed for estimating and modelling disease spatial distribution, with a view to making the data compatible with the layers of geographic data available for the study zone as a whole.

2. Material and methods

2.1. Study zone

The study was conducted in part (1 200 km²) of the Sidéradougou agro-pastoral zone south of Bobo-Dioulasso (Burkina Faso), 11°N and 4°W (Fig. 1). The zone has 1 000 to 1 100 mm of rainfall per year, with a dry season from November to April and a rainy season from May to October. It is typical of the Sudanian tropical climate zone, with bushy savannas and forest stands along its watercourses. These types of riverside vegetation are the preferred biotopes of the tsetse flies found in the zone (*Glossina tachinoides* and *Glossina palpalis gambiensis*) (Challier, 1973; Gruvel, 1975).

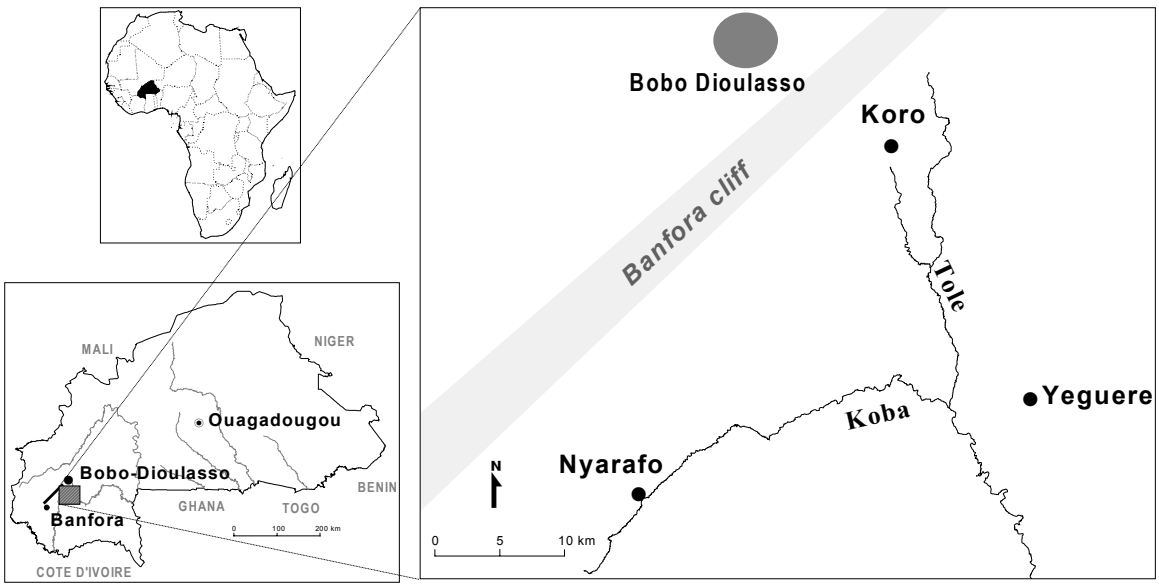


Figure 1: Location of the study zone. The Sidéradougou agro-pastoral zone is located in the south of Bobo-Dioulasso (Burkina Faso), at 11°N and 4°W.

2.2. Population, sampling and diagnosis

The cattle in the zone were counted exhaustively, based on the dwellings by which they are penned during the night (Michel et al., 1999). For each dwelling, the number of head, their watering points at the end of the dry season, and information on transhumance were recorded. The geographic positions of each dwelling and the watering point or points (two at most) were determined by GPS (Global Positioning System, Garmin™).

Over 800 dwellings, with 16 576 head, were visited. The herds were split into three categories: (i) small units with one or two pairs of draught oxen; (ii) mixed units, generally with fewer than 20 head, including draught oxen and a few breeders; (iii) large herds of several dozen head, with transhumance often practised during the dry season. In this zone, where livestock are a major component of farming systems practised, there are many small and medium-sized herds, which account for over 80% of farms but only 20% of the animals. On the other hand, 80% of the cattle in the zone are owned by 17% of the farmers (Table 1).

The herds are found in three main zones (Fig. 2): (i) an agricultural zone in which animal production is closely integrated into the farming system, with medium-sized herds, in the east; (ii) a mixed agricultural and pastoral zone in the west, with small and large herds; (iii) an almost exclusively pastoral zone in the south, with large herds (zone 3). This distribution corresponds to the pattern for crops (de La Rocque et al., 2001). In the whole study area, there are very few trading and non-trading exchanges of cattle.

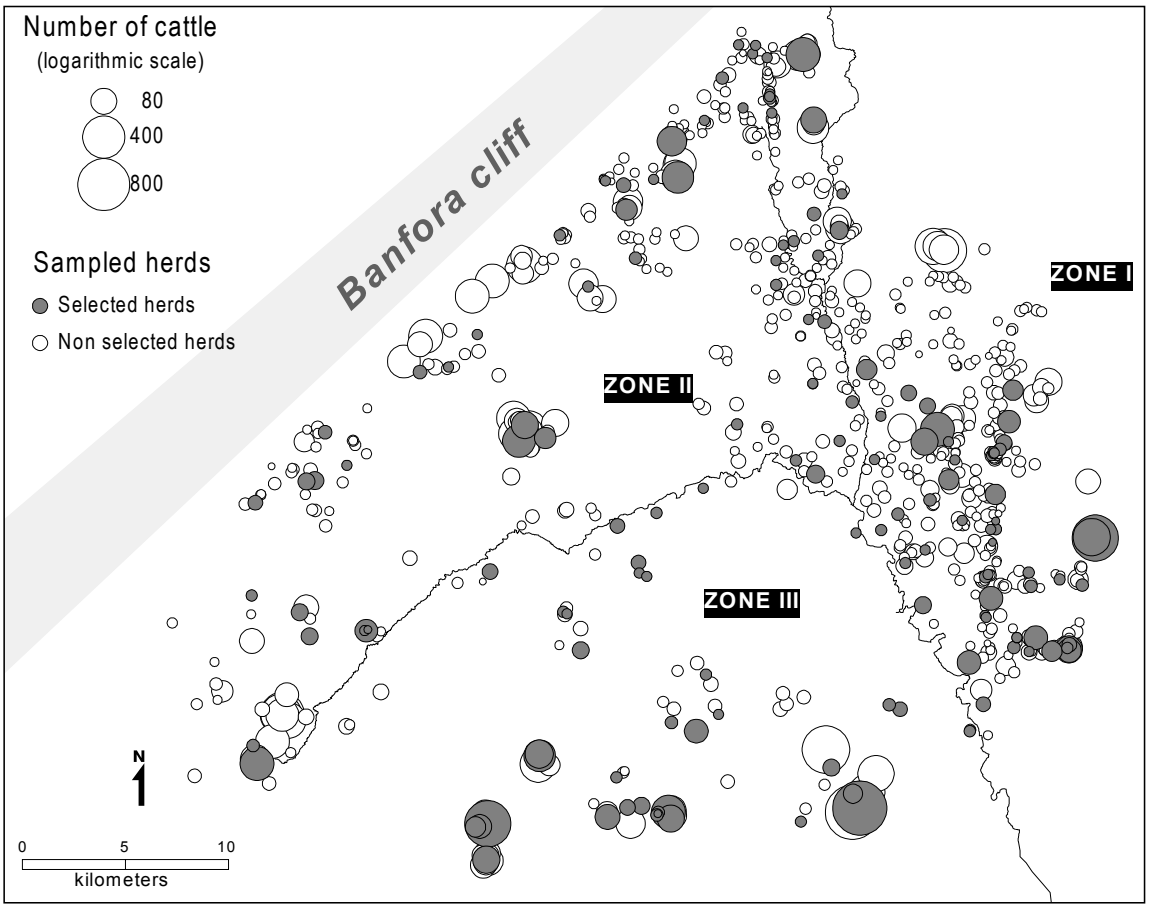


Figure 2: Herds, sampled herds and agricultural distribution in the study zone. *The size of the points varies according to the $\log_e$ of herd size.* The herds are found in three main zones which are delineated by the hydrographic network: (i) an agricultural zone in which animal production is closely integrated into the farming system, with medium-sized herds, in the East; (ii) a mixed agricultural and pastoral zone in the West, with small and large herds; (iii) an almost exclusively pastoral zone in the South, with large herds. This distribution corresponds to the pattern for crops.

Table 1: Herd size and head number in the local population

	Herd size			
	Under 5 head	5 to 20 head	Over 20 head	Total
Number of herds	476 (59%)	188 (24%)	137 (17%)	801
Number of animals	1 372 (9%)	1 861 (11%)	13 343 (80%)	16 576

A two-stage sampling was performed. The first sampling unit was on herd, i.e. an animal management unit subject to common animal production practices. It was easily identifiable in the field and corresponds to an epidemiological entity. The herds were drawn at random. The second sampling unit was animals which were chosen as follow: (i) exhaustive sampling in small herds (fewer than five head); (ii) ten head at most in medium-sized herds (between five and 20 head); (iii) 20 head at most in large herds (over 20 head). Within the herds, the head were drawn at random, without replacement. For logistical reasons it was decided to sample 2 000 head spread over 15% of the herds in the zone.
A questionnaire on animal production practices was filled in for each herd.
Blood samples were taken from the jugular vein. The plasma was analysed in the laboratory using three indirect ELISA systems (*T. vivax*, *T. brucei* and *T. congolense*), revealing antibodies against *Trypanosoma* spp. (Desquesnes et al., 2000).

2.3. Available data and statistical model

Several types of data were used to analyse and model trypanosomosis seroprevalence in the herds:

- Serological data corresponding to the variable to be explained.
- Animal husbandry data obtained from the field survey: herd size, transhumance practices and the type of watering point used at the end of the dry season.
- Spatial data generated by the GIS from the geographic position of the different units: distance between dwelling and watering point, and proximity of dwellings to the hydrological network.

The descriptive variables were classified according to knowledge of practices (Lhoste et al., 1993), and their epidemiological significance (Table 2). The type of watering point was divided into springs and rivers (which are propitious to tsetse flies), and wells and boreholes (which are generally found in zones not favourable to the flies). The zone classed as *neighbouring* on the hydrographic network was set at 2km, based on known data on the tsetse fly's ability to spread (Cuisance et al., 1985).
The serological prevalence for each herd was modelled using logistic regression, since the response variable is a proportion and the error function is assumed to follow the binomial law (McCullagh and Nelder, 1989). The link function used was the logit function, defined as $\text{logit}(p) = \log_e\left(\frac{p}{1-p}\right)$. An over-dispersion phenomenon often appears in using generalized linear model with a logit link when the response variable is a proportion. Over-dispersion means that the variance of the response variable exceeds the binomial variance and this problem is very common in large-scale epidemiological studies (McCullagh and Nelder, 1989). Taking into account the over-dispersion problem, we used a quasi-likelihood approach (McCullagh and Nelder, 1989) in the place of likelihood function. This led to wider confidence interval of parameters than the classical approach. For the same reasons, to test the contribution of the different descriptive variables in the model, we conducted a deviance analysis with F-test in the place of χ^2 test (Collett, 1991). The coefficients obtained in the

model were interpreted by calculating the odds-ratios and their confidence interval (Bouyer et al., 1995). This enabled us to quantify the risk factors associated with the levels of each of the explanatory variables in relation to a reference level.

Table 2: Descriptive variables used for modelling. Reference levels for the model are shown in bold

Code	Variable	Levels
typew	Type of watering point used in dry season	**art: artificial** nat: natural
distw	Distance between farm and watering point used in dry season	**1: <1000 m** 2: 1000 to 4000 m 3: > 4000 m
hrdsz	Herd size	**1: small (<5)** 2: medium (5 to 20) 3: large (>20)
Hy2km	Farm less than 2 km from hydrographic network	**No** Yes
allyr	Animals kept by dwelling all year round	**No** Yes

As the model was not spatialized, it was necessary to look for autocorrelation among residuals. If the presence of autocorrelation was detected, it could imply the omission of regressor variables, the presence of non-linear relationships or that the regression model should have an autoregressive structure (Cliff and Ord, 1973). To test the autocorrelation, we firstly established neighbourhood relationships between herds using a Delaunay triangulation as proposed by Schmoyer (1994). In the second step, Geary (Geary, 1954) and Moran's (Moran, 1950) statistics (see also Cliff and Ord, 1973) were computed for residuals. By permuting the values of the residual map, we computed new values of autocorrelation statistics and the observed value is tested by comparing to the set of values obtained for the permutations. As the number of possible permutation was very large, we used a Monte-Carlo (Manly, 1991) version of the test. The same procedure has been carried out for observed and predicted prevalence values. This kind of procedure has been recently used in Kleinschmidt et al. (2000) with the non-parametric D-statistic (Walter, 1992) to measure autocorrelation of predictions from a logistic regression. When these indices were applied to observed and predicted prevalence values and the residuals of the model, they enable to test the capacity of a model to take account the spatial nature of data.
The statistical model was then inverted to estimate the serological prevalence for all the herds in the zone, using the explanatory variables shared with the survey, which were the same than those used to generate the model.
All calculations were made using the R software (Ihaka and Gentleman, 1996).

2.4. Spatial model
Specific problems linked to the geographical nature of mapped objects such as herds and the "in herd" variability of measured variables have to be considered when mapping seroprevalence : (i) herds are points that may be superimposed if they are close to one another, hence masking information, (ii) the meaning of prevalence within a herd varies with the number of head in the herd, a prevalence of 50 % in a two head herd has not the same significance than a prevalence of 50 % in a one hundred head herd, (iii) mapping only the points corresponding to the pens used at night provides only a partial representation of reality. Animals move and occupy a continuous space, (iv) spatial information on prevalence has to be compatible with the other information available in the GIS if they are to be compared.
To overcome these problems, a spatial model of land occupation by cattle and of disease distribution was developed. All spatial object manipulations used the Mapinfo™ software.

The representation of land occupation by cattle in a zone as a whole is based on modelling the daily movements of the animals in each herd. In savanna zones, water availability is the main constraint at the end of the dry season, and governs movements (Boutrais, 1994). Herd movements were therefore modelled by representing the direct route between the night pen and the watering point or points, and drawing a buffer zone around the route, corresponding to the area occupied by the cattle during the day (Michel et al., 1999). This zone of daily use by the herd varies in size. The wider the herd and the nearer it is to its watering point, the larger the zone of frequentation (Fig. 3). This model was validated by monitoring the movements of a sample of herds.

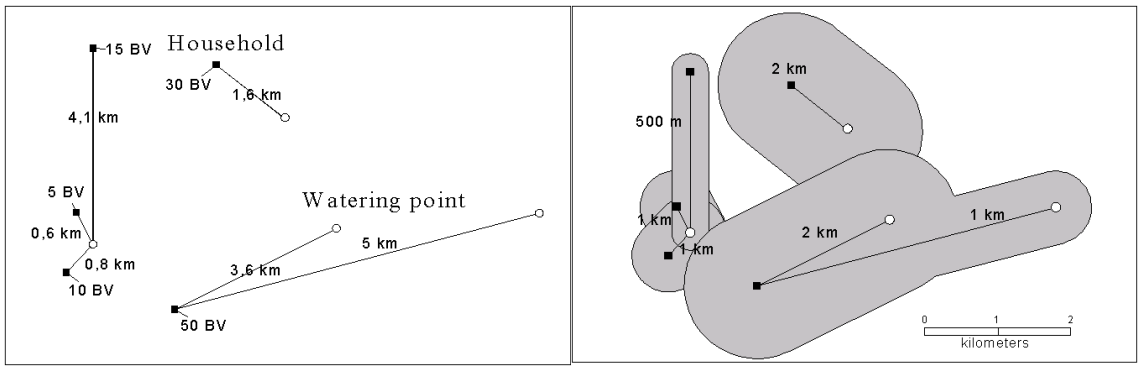


Figure 3: Modelling of daily herd movements. Modelling herd movements consists in representing the direct route between the night pen and the watering point or points, and drawing a buffer zone around the route, corresponding to the area occupied by the cattle during the day. This zone of daily use by the herd varies in size. The larger the herd and the nearer it is to its watering point, the larger the zone of frequentation is.

The predicted prevalence for all the herds was applied to their zones of daily use. To synthesize this information, which was not yet very easy to resolve due to the superimposition of polygons, it proved necessary to aggregate it so as to shift to a smaller scale. This was done by projecting all the zones of use and the corresponding prevalences onto a regular geographic grid of 1-km squares (Raynal et al., 1996). The cumulated distribution of antibody-prevalence in the study zone was then represented by assigning to each square the mean value of the prevalences for the herd polygons, impinging on it, so as to produce a map of average prevalence (Fig. 4). The calculus of mean value of prevalence was weighted by the size of herds in order to take into account for problems (ii) cited above. Smoothing by two-dimensional weighted local regression (Cleveland and Devlin, 1988) on the centroids of the squares in the grid made the maps more realistic.

3. Results

3.1. Sampling and observed seroprevalence

In total, 216 herds and 1 784 head were sampled. Herd and cattle distribution in the sample showed that small herds were slightly under-represented, in favour of medium-sized herds (Table 3). On the other hand, small herds were over-represented near the hydrographic network (Fig. 2). The differences in relation to the sample initially planned can be attributed to field constraints such as herds in an out-of-the-way place or cattle breeder absent or not in agreement with taking a blood sample.

The average serological prevalence observed among the cattle was 73.4%. The map of herd prevalence, shown as points according to the corresponding dwelling, showed case distribution but was difficult to interpret (Fig. 5).

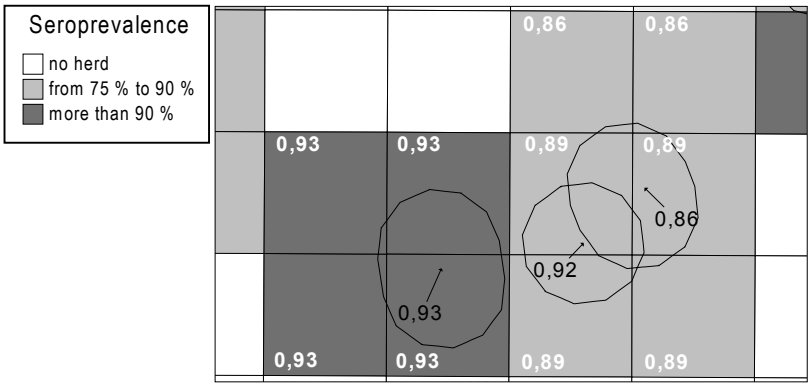


Figure 4: Data aggregation and spatial distribution of mean prevalence. This was done by projecting all the zones of use and the corresponding prevalences onto a regular geographic grid of 1-km squares. The cumulated distribution of antibody-prevalence in the study zone was then represented by assigning to each square the mean value of the prevalences for the herd polygons impinging on it, so as to produce a map of average prevalence.

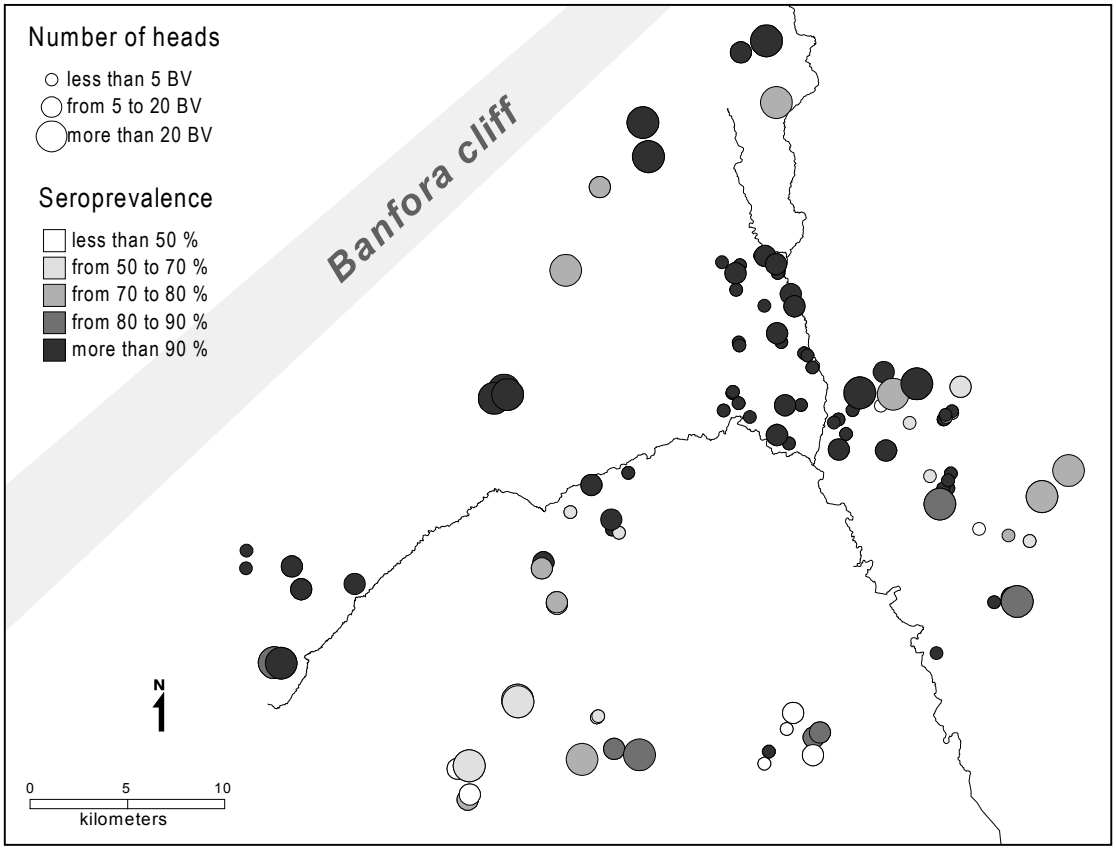


Figure 5: Distribution of serological prevalence among the herds sampled. The map of seroprevalence for the herds sampled, shown as points according to the corresponding dwelling, indicated case distribution.

Table 3: Herd size and head number in the sample

	Herd size			
	Under 5 head	5 to 20 head	Over 20 head	Total
Number of herds	110 (51%)	70 (32%)	36 (17%)	216
Number of animals	327 (18%)	736 (41%)	721 (41%)	1 784

3.2. Statistical modelling: identification of risk factors

The deviance analysis showed that only the distance between cattle pen and watering point was not significant and this variable was excluded from the model. All the other variables were significant (Table 4). The dispersion parameter for the model was 2.48. The relation between the numbers of observed and predicted positives (Spearman's rank correlation $\rho=0.45$, $p<0.0001$) showed that the statistical model has a good fit. The spatial auto-correlation tests revealed a positive correlation between observed and predicted prevalences, whereas the residuals of the model were not correlated (Table 5). The variables used in the model thus take account of the spatial factor. The odds-ratios calculated with the coefficients estimated by the model showed that proximity to the hydrographic network, frequentation of natural watering points, large herd size and the fact of keeping animals near dwellings all year round were all risk factors (Table 6).

Table 4: Deviance analysis of the model (DF: degree of freedom, Resid. DF: residual degree of freedom, Resid. Dev.: residual deviance).

	DF	Deviance	Resid. DF	Resid.Dev	p(>F)
NULL			215	666.51	
typew	1	17.66	214	648.86	<0.0082
hrdsz	2	37.37	212	611.49	<0.0006
allyr	1	59.07	211	552.41	<0.0001
hy2km	1	54.17	210	498.24	<0.0001

Table 5: Spatial auto-correlation tests for the observed and the predicted prevalences and the residues of the model, using Moran's (I) and Geary's (c) indexes

Variable	I	p values	c	p values
obs. prevalence	0.196	<0.001	0.799	<0.001
pred. prevalence	0.542	<0.001	0.465	<0.001
residuals of model	0.020	0.248	0.972	0.288

Table 6: Odds-ratios (OR) calculated from the coefficients of the serological model

Variable	Lower confidence interval OR	OR	Upper confidence interval OR
(Intercept)	0.53	1.08	2.19
typewnat	0.88	1.34	2.04
hrdsz2	0.38	0.67	1.16
hrdsz3	1.03	1.99	3.84
allyryes	1.69	2.70	4.31
hy2kmyes	1.94	3.43	6.07

3.3. Modelling of spatial distribution of prevalence

The map of predicted serological prevalences, obtained by spatial modelling on a whole study zone scale, showed that mean serological prevalence is distributed along the hydrographic network, with focal points of high values, and that it spreads radially into the neighbouring savannas (Fig. 6).

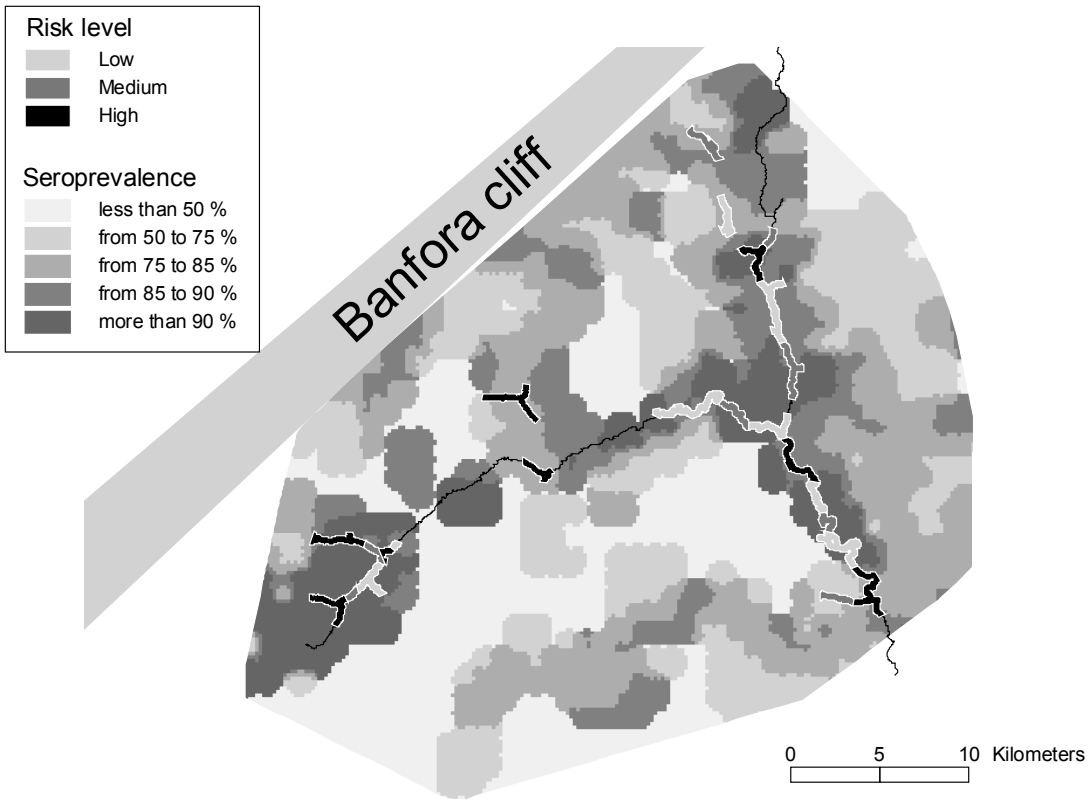


Figure 6: Distribution of mean serological prevalence in the zone and high-transmission-risk zones. The high - transmission risk zones are delineated by a white outline. This map shows that mean serological prevalence is structured linearly along the hydrographic network, with focal points of high values, and that it spreads radially into the neighbouring savannas.

4. Discussion

4.1. Observed prevalence and statistical model

The serological results obtained from the sample confirm the enzootic situation that had already been observed for trypanosomosis in the Sidéradougou zone, with high infection levels among vectors (de La Rocque, 1997). On an animal production zone scale, parasite pressure can be evaluated more accurately by the number of antibody carriers than by direct detection of parasites (Desquesnes et al., 2000). With serological data, the statistical model proved to be of high quality, and confirmed the risk factors conventionally identified in terms of animal trypanosomosis, such as spatial and temporal proximity to vector-propitious biotopes and intensity of contact with tsetse flies, particularly through watering practices (de La Rocque et al., 1999). The risk associated with large herds can be put down to the fact that such herds use natural watering points with sufficient capacity, which are generally located in preferred biotopes of tsetse flies.

4.2. Spatial modelling of trypanosome prevalence

Daily movements of cattle were modelled at the end of the dry season, since at such times, animal movements centre on specific sites: dwellings and watering points. Moreover, the end of the dry season is a key period for bovine trypanosomosis epidemiology: it is the period with the highest risk of parasite transmission (Rowlands et al., 1993). Lastly, modelling was conducted with a specific aim: to study the relations between cattle trypanosomosis vectors and hosts on a whole-zone scale. The model predicts the presence of cattle around the crucial

points of contact between cattle and tsetse flies. It would have been possible to introduce spatial constraints into the model, to take account of the relief or of zones occupied by crops. However, these constraints are very limited at the end of the dry season, and projection onto a geographic grid would reduce its accuracy.

The attribution of prevalences to zones of use by cattle and their aggregation within a geographic grid provides continuous information in spatial terms. The map of mean prevalences, which takes account of all the predicted prevalences, is well suited to serological data.

Superimposing the map of predicted prevalences and the epidemiological risks zones identified elsewhere (de La Rocque et al., 2001) shows that prevalence distribution corresponds roughly to the zones with a high risk of disease transmission (Fig. 6). The zones in the Northeast, at the foot of the Banfora Cliff and in the extreme North, which had high prevalence rates among cattle, were not subjected to the high-transmission-risk site identification procedure. The existence of high serological prevalences well away from the hydrographic network can be explained by (i) tsetse fly dispersion during the rainy season: the flies infect cattle, which may still have antibodies when the next dry season arrives (Desquesnes 1997); (ii) the assignment of prevalences to cattle use zones that stretch well into the savannas. It is thus crucial to take account of animal movements and land occupation if we are to obtain a realistic picture of bovine trypanosomosis prevalence distribution in the zone.

Projecting the land occupied by each herd onto the geographic grid enabled us to improve the resolution of disease representation in the zone. The choice of the size of the elementary squares in the grid is crucial, as it governs the change in scale. We chose a size of 1 km, since it corresponds (i) to a reasonable scale for taking account of variations in daily herd movements and (ii) to the scale of the study and integration in the GIS of the other topics covered in the study of disease transmission.

5. Conclusion

By taking account of animal movements and modelling the prevalence of the disease, the method described enabled us to convert specific, partial spatial information on a herd scale into continuous information on a whole-zone scale. Spatial modelling is a robust method, and produces a realistic picture of disease epidemiology. It showed that the natural environment and animal production practices account for structure in trypanosome distribution. The data layer obtained was integrated into a GIS with a view to validating the zones with a high risk of animal trypanosomosis transmission. This opens the way for the identification of spatial indicators of a trypanosomal risk, such as the presence of crops, the spatial structure of habitats, and soil characteristics.

The approach described was based on detailed field data whose acquisition is time-consuming, such as exhaustive, georeferenced counts of herds and their watering points. One essential improvement would be to identify simple, easy to obtain indicators of the presence of cattle and their movements.

The spatialization of data, their integration into a GIS, the coupling of conventional statistical models and spatial models, and the methods available for changing scale offer new methodological and thematic prospects for studying the epidemiology of directly and indirectly transmitted diseases on different scales (Hay et al., 2000; Hendrickx et al., 2001), but also for understanding the interactions between animal production and the surrounding environment.

Acknowledgements
This work was funded by CIRAD-EMVT (Action Thématique Programmée « Santé-Environnement ») and CNRS (Programme Interdisciplinaire "Environnement, Vie et Société"). We are indebted to the directors of CIRDES, S.M. TOURE and A. GOURO for allowing us to do the work.

References

Boutrais, J., 1994. Eleveurs, bétail et environnement. *In* : A la croisée des parcours, Pasteurs, éleveurs, cultivateurs. Editions de l'ORSTOM, Paris. 304-319.
Bouyer, J., Hémon, D., Cordier, S. Derriennic, F., Stücker, I., Stengel, B., Clavel, J., 1995. Epidémiologie, principes et méthodes quantitatives. INSERM, Paris. 498pp.
Challier, A., 1973. Ecologie de *Glossina palpalis gambiensis* Vanderplanck 1949 en savane d'Afrique Occidentale. ORSTOM, Montpellier. 274pp.
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.
Cliff, A.D., Ord, J.K., 1973. Spatial autocorrelation. Monographs in spatial and environmental systems analysis. Pion Limited, London. 178pp.
Collet, D., 1991. Modelling Binary Data. Chapman & Hall, London. 384pp.
Cuisance, D., Fevrier, J., Déjardin, J., Filledier, J., 1985. Dispersion linéaire de *Glossina palpalis gambiensis* et *G. tachinoides* dans une galerie forestière en zone soudano-guinéenne (Burkina Faso). Rev. Elev. Med. Vet. Pays Trop. 47, 69-75.
De Haan, C., Bekure, S., 1991. Animal health services in Sub-Saharan Africa: initial experiences with new approaches. World Bank, Washington. 88pp.
De La Rocque, S., 1997. Facteurs discriminants majeurs de la présence des glossines dans une zone agro-pastorale du Burkina Faso. Intérêt dans la prévision du risque trypanosomien. PhD thesis, Université de Montpellier, Montpellier. 162pp.
De La Rocque, S., Bengaly, Z., Michel, J.F., Solano, P., Sidibé, I., Cuisance, D., 1999. Importance des interfaces spatiales et temporelles entre les bovins et les glossines dans la transmission de la trypanosomose animale en Afrique de l'Ouest. Rev. Elev. Med. Vet. Pays Trop. 52, 215-222.
De La Rocque, S., Michel, J.F., Cuisance, D., De Wispeleare, G., Solano, P., Augusseau, X., Guillobez, S., Arnaud, M., 2001. Le risque trypanosomien, une approche globale pour une décision locale. CIRAD, Montpellier, France, 150pp.
Desquesnes, M., 1997, Les trypanosomoses du bétail en Amérique Latine, étude spéciale dans le Plateau des Guyanes. PhD Thesis, Université de Lille, Lille. 409pp.
Desquesnes, M., Michel, J.F., De La Rocque, S., Solano, P., Millogo, L., Sidibé, I., Cuisance, D., 2000. Enquête parasitologique et sérologique (ELISA-indirectes) sur les trypanosomoses des bovins dans la zone de Sidéradougou, Burkina Faso. Rev. Elev. Med. Vet. Pays Trop. 52, 223-232.
Geary, R.C., 1954. The contiguity ratio and statistical mapping. The incorporated Statistician. 5, 115-145.
Gruvel, J., 1975. Données générales sur l'écologie de *Glossina tachinoides* Westwood, 1850, dans la réserve de Kalamaloué, vallée du Bas –Chari. Rev. Elev. Med. Vet. Pays Trop. 28, 27-40.
Hay, S.I., Randolph, S.E., Rogers, D.J., 2000. Remote sensing and geographical information systems in epidemiology. Advances in parasitology, Vol 47. Baker, J.R., Muller, R., Rollinson, D., Eds.
Academic Press, London.

Hendrickx, G., De La Rocque, S., Reid, R., Wint, W., 2001. Spatial trypanosomosis management: From data-layers to decision making. *Parasitology Today*. 17, 35-41.
Hursey, B.S., Slingenbergh, J., 1995. The tsetse fly and its effects on agriculture in sub-Saharan Africa. Revue Mondiale de Zootechnie. 84, 67-73.
Ihaka, R., Gentleman, R., 1996. R: A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics. 5, 299-314.
Kleinschmidt, I., M. Bagayoko, Clarke, G.P.Y., Craig, M., Le Sueur, D., 2000. A spatial statistical approach to malaria mapping. International Journal of Epidemiology. 29, 355-361.
Laveissière, C., Couret, D., Hervouët, J.P., 1986. Localisation et fréquence du contact homme/glossines en secteur forestier de Côte d'Ivoire. 1. Recherche des points épidémiologiquement dangereux dans l'environnement végétal. Cahiers de l'ORSTOM, Série Entomologie médicale et Parasitologie. 14, 21-35.
Lhoste, P, V Dollé, J Rousseau, and D Soltner. 1993. *Manuel de zootechnie des régions chaudes. Les systèmes d'élevage*, *Manuels et précis d'élevage*: Coopération française. 288pp.
Manly, B.F.J., 1991. Randomization and Monte Carlo methods in biology. London, Chapman and Hall. 281pp.
McCullagh, P. & Nelder, J. A. (1989): Generalized Linear Models. Second Edition. Chapman & Hall. 512pp.
Michel, J.F., Michel, V., De La Rocque, S., Touré, I., Richard, D., 1999. Modélisation de l'occupation de l'espace par les bovins. Applications à l'épidémiologie des trypanosomoses animales. Rev. Elev. Med. Vet. Pays Trop. 52, 25-33.
Moran, P.A.P., 1948. The interpretation of statistical maps. Journal of the Royal Statistical Society Series B-Methodological. 10, 243-251.
Raynal, L., Dumolard, P., D'Aubigny, G., Weber, C., Rigaux, P., Scholl, M., Larcena, D., 1996. Gérer et générer des données spatiales hiérarchisées. Revue Internationale de Géomatique. 6, 365-382.
Rowlands, G.J., Mulatu, W., Authié, E., D'Ieteren, G.D.M., Leak, S.G.A., Nagda, S.M., Peregrine, A.S., 1993. Epidemiology of bovine trypanosomiasis in the Ghibe valley, Southwest Ethiopia. 2. Factors associated with variations in trypanosome prevalence, incidence of new infections and prevalence of recurrent infections. Acta Tropica. 53, 135-150.
Schmoyer, R.L., 1994. Permutation tests for correlation in regression errors. Journal of the American Statistical Association. 89, 1507-1516.
Walter, S.D., 1992. The analysis of regional patterns in health data, Part 1. American Journal of Epidemiology. 136, 730-741.

Chapitre VI : Dynamique de population et habitat disponible

Ce chapitre est basé sur deux publications soumises. Les principaux résultats biologiques obtenus lors de la collaboration entreprise avec N. Pettorelli et J.-M. Gaillard sont présentés. La structuration spatiale d'une population de chevreuils et ses liens avec les paramètres environnementaux sont analysés.

Dans la première publication, un exemple très simple présente l'avantage d'une approche couplant système d'information géographique et analyse multivariée. L'objectif est de décrire les patterns de végétation, à une échelle spatiale cohérente, afin de définir les différents types d'habitats disponibles pour le chevreuil. Les données analysées ont été collectées en 1993 dans la réserve de Chizé et décrivent la composition en essences du taillis et de la futaie. L'utilisation d'analyses multivariées et d'un SIG a permis d'extraire les principales structures spatiales de la flore de Chizé. Trois grands types ont été définis : chênaie avec taillis d'érable dans le nord-ouest, chênaie avec taillis de charme dans le nord-est et hêtraie sans taillis dans le sud de la réserve. Les analyses ont été réalisées dans le SIG à l'aide l'extension AVADE (cf chapitre IX).

Le deuxième article s'attache à mesurer l'importance de l'espace dans la variabilité de la masse des faons et à identifier quelles ressources peuvent être responsables de cette variabilité spatiale. Pour ce faire, la structuration spatiale de la végétation et celle de la masse corporelle ont été analysées à une échelle cohérente avec les objectifs de l'étude. La part des effets spatiaux et temporels dans la variabilité de la masse corporelle a été mesurée grâce à un modèle linéaire construit à l'aide de données s'étalant sur 24 années de capture (1235 individus). Parallèlement, les grands patterns de distribution de la végétation ont été appréhendés à partir d'une ACP globale réalisée à partir de 578 sites d'échantillonnage avec l'extension AVADE. L'information relative à la végétation et celle concernant les poids des faons ont ensuite été reliées qualitativement. Les effets spatiaux-temporels expliquent 36.8% et 39.2% de la variabilité individuelle en poids pour les mâles et les femelles. L'effet de cohorte (temporel) représente seulement 20.4% (mâles) et 20% (femelles) de la variabilité. La structure spatiale des poids des faons est constante durant les 24 années et similaire pour les deux sexes. Une différence maximale d'environ 2 kg est observée selon la localisation des animaux dans la réserve. La distribution de trois plantes (charme, jacinthe, ornithogale), connues pour être des ressources alimentaires importantes pour le chevreuil au printemps, est positivement corrélée à la distribution des poids, pour les deux sexes, en hiver. La distribution de plantes évitées par les faons (fragon piquant, hêtre) est corrélée négativement à la distribution des poids. Il semble donc que la distribution spatiale des plantes, qui sont sélectionnées durant le printemps et l'été, est un facteur important qui influe sur la distribution spatiale des masses des faons en hiver et donc sur la dynamique de population de chevreuils. La mise en place du modèle linéaire, à l'aide du logiciel R, est présentée en annexe 3.

Nathalie Pettorelli, Stéphane Dray, Daniel Maillard, Stéphane Villarubias

Coupling multidimensional analysis and GIS to classify and map deer habitats

Abstract: We aimed to define at a relevant scale the spatial pattern of major vegetation types, in order to characterize major habitat types available to deer. We analyzed data on timber stands and coppice collected in 1993 in the 2,614ha Chizé reserve, situated in western France. Multidimensional analyses (Principal Component Analysis and biplot) and Geographic Information System (GIS) were used to extract most of the variation in vegetation data collected at the 4 ha level. Three vegetation types occurred within the reserve: maple dominated coppices in the oak stand in the northwest, hornbeam dominated coppices in the oak stand in the northeast, and a beech stand with no coppice in the south. The coupled use of multivariate analysis and GIS, which allowed us to assess the classification of forest habitats, seems to be promising for use in wildlife management.

Key words*:* deer, GIS, habitat classification, management, multidimensional statistics

Introduction

The importance of temporal and large-scale spatial fluctuations of ecological factors in population dynamics has been heavily underlined over the last decades (Gilpin and Hanski 1991, Tilman and Kareiva 1997, Tuljapurkar and Caswell 1996). However, recent studies have demonstrated the overwhelming importance of individual variability in mammalian population dynamics (Gaillard et al. 2000). In this research of factors structuring individual variability, the importance of spatial variation in fitness components has often been overlooked. However, recent studies have demonstrated that spatial variation in habitat quality affects population dynamics in ungulates (Coulson et al. 1997 and 1999; Milner-Gulland et al. 2000), as individuals within a population do not always distribute themselves according to an ideal free model (Fretwell and Lucas 1970, Conradt et al. 1999).

To precise the importance of space as an attribute of individual variability in mammalian population dynamics, a detailed knowledge of population habitat (in particular spatial distribution of food resources) is needed. This knowledge generally implies to work with numerous variables, and multivariate analyses are a natural tool to deal with such databases. However, this knowledge also implies some precise spatial representation of the variability. Geographical Information Systems (GIS) have recently been showed as a powerful tool when dealing with spatial representation. Nevertheless, few studies have underlined the interest for managers and foresters to couple both tools.

In the Chizé reserve (situated in western France), habitat specific life history traits such as fawn body weight have been found from a long term monitoring of the roe deer population (Pettorelli et al. 2001). Based on soil characteristics and coarse data on dominant tree species, 2 broad vegetation types were distinguished within this 2,614 ha managed forest: the northern part of the reserve is composed of an oak stand on soils with a high clay content, and the southern part of the reserve is composed of a beech dominated stand on limestone and chalky soils associated with marl. This difference of woodland type was found to have consequences on the repartition of main resources used by roe deer in spring and summer: this 2 periods are associated with high energy requirements (Andersen et al. 1998) for this species which show little body mass variation throughout the year ("income breeder strategy"; Andersen et al. 2000). The principal food plants in both seasons occurred more frequently in the oak woodland that in the beech woodland. However, this rather crude approach to habitat definition was based on coarse data on soil structure and forest management, and might therefore be far from optimal. In particular, the 2 broad types did not take into account the shrub layer, which is one of the most important factors affecting habitat selection by roe deer (Cibien and Sempéré 1989; Mysterud et al. 1999).

We have developed a general method that allows spatial patterns of deer habitats to be assessed from vegetation data. A common way of dealing with spatial data on habitat description is to consider only some information among the global database in possession. It thus involves lumping the population's habitat into a few blocks, generally based on the knowledge of animal diet (Coulson et al. 1997) or on assumptions of plant quality such as phenology (Mysterud et al. 2001). Here we present a method coupling multivariate analyses and GIS that allow considering all the descriptive variables collected in order to separate major plant communities occurring within the habitat of the population.

We thus assessed spatial structure at a scale that could be related to roe deer, in order to determine the effects of habitat structure and quality on roe deer population dynamics. At Chizé, the roe deer population is monitored for over 20 years for research and management purposes: each year, animals are caught in winter by drive-netting in blocks of about 100 ha on average during capture sessions (Gaillard et al. 1993) and thus the location of individuals can be made only at a broad scale. Moreover, many roads have been constructed within the

reserve, and they often constitute natural limits for roe deer home ranges (Hewison et al. 1998). Accordingly, we decided to assess major types of vegetation community by using roads to delimit the groups. This choice would allow us to define major pattern of variability in vegetation community at a relevant scale.

Methods

Data on forest structure were collected in 1993 by foresters of the Office National des Forêts (ONF) in the Chizé reserve, situated in western France (46°05'N, 0°25'W). The climate is oceanic with Mediterranean influences, and characterized by mild winters (mean temperature per month (from December to February) of 6.16 °C, with a minimum temperature of 3.85 °C and a maximum temperature of 7.87 °C; mean precipitations per month of 90.67 mm with minimum precipitations of 34.43 mm and maximum precipitations of 139.3 mm) and hot, dry summers (mean temperature (from June to August) of 19.38 °C, with a minimum temperature of 17.45 °C and a maximum temperature of 21.27 °C; mean precipitations of 52.63 mm with minimum precipitations of 28.63 mm and maximum precipitations of 104.4 mm; data collected from 1977 to 1998).

In France, state forests are divided into numbered plots, the limits between them being forest trails. However, the forest is generally managed at the finer sub-plot scale (4 ha on average in the Chizé reserve). At this particular scale, foresters in the Chizé reserve have determined in 1993 the 4 coppices' dominant species and the 3 timber stands' dominant species, and their cover (%). For each sub-plot, the total of cover for each stratum (timber stand and coppices) was 100%.

4 oak subspecies (*Quercus robur, pubescens, petraea,* and *cerris*) occurred in the timber stand, and 2 (*Quercus pubescens* and *robur*) in the coppices. 4 pine species (*Pinus pinaster, sylvestris, nigra,* and *laricio*) occurred in the timber stand, and finally 2 maple species (*Acer campestre* and *monspessulanum*) occurred in both strata. The species were not separated in all sub-plots and thus were pooled into 3 groups (oaks, maples, pines). Rare species of coniferous and deciduous trees were pooled in 2 different categories. 10 categories of timber stand were thus considered (maples, pines, oaks, hornbeam [*Carpinus betulus*], beech [*Fagus sylvatica*], other deciduous, other coniferous, plum [*Prunus avium*], cedar [*Cedrus sp.*], Douglas fir [*Abies douglasi*]) for 629 sub-plots. 5 categories of coppices (maples, hornbeam, oaks, beech, and other deciduous) for 621 sub-plots were considered. Data were available for more than 97% of sub-plots (n=648 in the reserve) for the timber, and 96% for the coppices.

We used multivariate statistical analyses to extract the main sources of variation in the data set. Principal Component Analysis (PCA) on a table of proportions has been introduced in general ecology in the last decades (Ter braak 1983, Aitchison 1983). This method and the associated graphical representation based on biplot theory (Gabriel 1971, De Crespin de Billy et al. 2000) are very suitable for table of proportions data.

On the biplot representation, the cover type categories are positioned relative to the principal axes of unit variance (mean = 0) and represented by arrows. Sub-plots are plotted according to an averaging procedure based on proportion data. This averaging procedure allows the user to display the mean composition of the sub-plots, which differentiates this analysis from the usual PCA procedure (Hotelling 1933). The position of cover type category on the biplot representation depends on their relative abundance and on the variability of vegetation composition among sub-plots. Hence, rare cover type categories are located around the origin whereas major cover type categories are more distant. Thus, the ordination of sub-plots allows the direct examination of their vegetation composition and gives the main structure of the reserve. Sub-plots that are close to the origin contain all dominant cover type categories or are characterized by rare cover type categories whereas sub-plots that occur near the end of

an identified arrow are characterized essentially by the cover type category the arrow represents.

Two separate PCAs were performed on the timber stand and coppice tree data, with each sub-plot being weighted by its area. All analyses were performed using ADE-4 software (Thioulouse et al. 1997).

A spatial representation of sub-plot scores on the PCA factorial axes (Goodall 1954) was then performed using ArcView GIS (Mitchell 1999) to firstly observe whether variations in vegetation composition were spatially structured and then to determine limits between such structures in the reserve. Jenks' Goodness of Variance Fit statistic (Jenks and Caspall 1971) was used to identify breakpoints between classes of scores. This iterative method allows the sum of the variance within each of the classes to be minimized, in order to detect the best groupings and patterns structuring the data. The number of classes was determined using score distribution. The histogram representing the frequency distribution of scores was first plotted, and then a non-parametric estimate of the function of probability's density was computed with S-PLUS software (Venables and Ripley 1997). The general trend of this function (number of peaks) has permitted to determine the number of classes of PCA scores in order to improve the readability. To determine the proportion of the total variability explained by the final classification, we calculated for both the timber stand and the coppice data the ratio between the inter-class inertia and the total inertia.

Results

Timber stand

Timber stands occurred in 574 out of the 621 sub-plots. The PCA on row profiles showed a structure with two axes: the first axis (representing 65% of the total inertia) opposed oak and beech, the two main tree species in the reserve (256 oak dominant and 224 beech dominant sub-plots in the reserve). The second axis of the PCA (representing 23% of the total inertia) opposed the oak-beech stands to pines (pines was the dominant species in only 65 out of the 574 sub-plots considered).

According to the biplot (Fig.1), the sub-plots distributed themselves in a triangular structure around the three main cover type categories (beech, oaks and pines). Most of the sub-plots were aggregated in the corners, which means that the gradient structures between different timber stand types were weak.

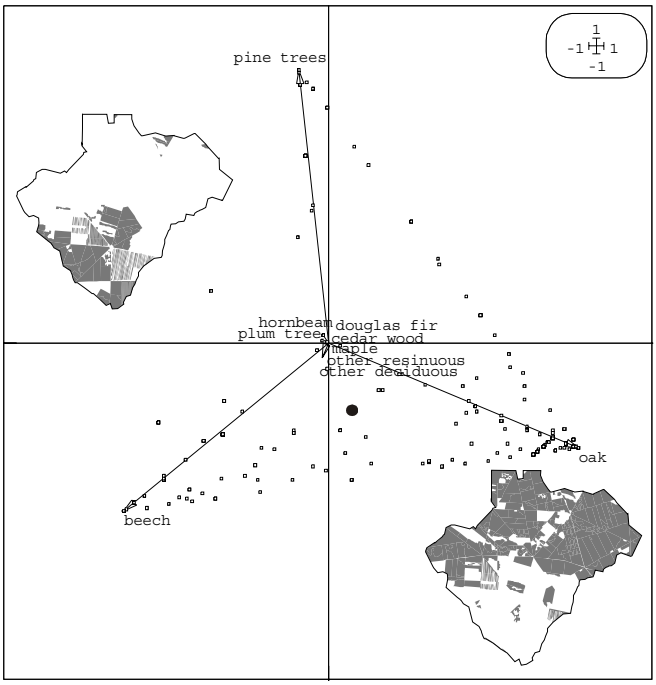


Figure 1: Biplot of the PCA of the timber stand data. Arrows represent the ten cover type categories considered, and open squares represent the sub-plots. The black circle represents the average timber stand sub-plot in the reserve, with its average vegetation composition. Oak (in the right side) and beech (in the left side) dominant sub-plots are also represented.

The spatial representation of sub-plots scores on the first axis of the PCA (Fig.2) supported the conclusion that the different timber stand types occurred in different areas in the reserve. The beech stand occurred essentially in the south of the reserve: 215 out of the 332 sub-plots of the south (71% of the area) were beech dominant. The oak stand occurred essentially in the north: oaks dominated 213 out of the 242 sub-plots of the north (90.7%). Pine dominant sub-plots were non-aggregated and were distributed approximately equally in both parts of the reserve. The reserve was therefore divided into two groups (Fig.2): this classification accounted for 29.6% of the total variability of the timber stand database.

Figure 2: Spatial representation of sub-plots scores on the first axis of the PCA performed on the proportion table of timber stand data. The dashed line represents the division performed between the two timber stand types (oak and beech stand). The sub-plots with no data available on timber stand are represented in white. The sub-plots with no timber stand are represented in black.

Coppices

Coppices were present only in 237 out of 621 sub-plots. The PCA on row profiles performed on the sub-plots also showed a two axes structure. The first axis (representing 64% of the total inertia) opposed sub-plots dominated by hornbeam to sub-plots dominated by maples. The second axis (representing 22% of the total inertia) opposed the latter sub-plots to beech dominant sub-plots.

According to the biplot (Fig.3), the sub-plots distributed themselves in a triangular structure around the three cover type categories (beech, hornbeam, and maples). Contrary to the timber stand results, most of the sub-plots were not aggregated around the corners, meaning that there was a gradient structure between different coppice types (particularly between hornbeam and maple coppice types).

The spatial representation of sub-plot scores (Fig.4) confirms the spatial structure in coppice types: hornbeam dominant coppices occurred essentially in the northeast while maple dominant coppices occurred mostly in the northwest part of the reserve. Few sub-plots were beech dominant (5 sub-plots in the reserve). Moreover, only 15 sub-plots in the beech stand had coppices.

Finally, as sub-plots with coppices in the beech stand were not numerous and differed in dominant cover type categories, we decided to discard them when dividing the reserve. As the gradient structure was marked between both coppice types, we had to use both extreme

classes of scores on the first axis of the PCA to determine coppice types delimitation. Two types were thus distinguished (Fig.4): the maple coppices in the northwest and the hornbeam coppices in the northeast part of the reserve. This classification accounted for 32% of the total inertia of the coppices data set. Because of the gradient structure between both coppice types, intra-group variability was relatively high: 64 out of the 73 sub-plots in the northwest part (71%) were maple dominant, and 99 out of the 149 sub-plots in the northeast part (67%) were hornbeam dominant sub-plots.

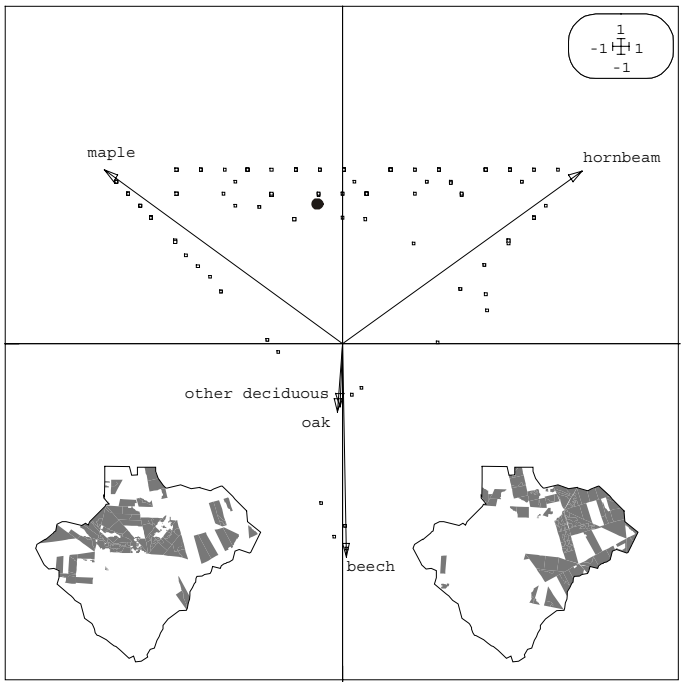


Figure 3: Biplot of the PCA of the coppices data set. Arrows represent the five cover type categories considered, and open squares represent the sub-plots. The black circle represents the average coppice sub-plot, with its average vegetation composition. Hornbeam (on the right side) and maple (on the left side) dominant sub-plots are also represented.

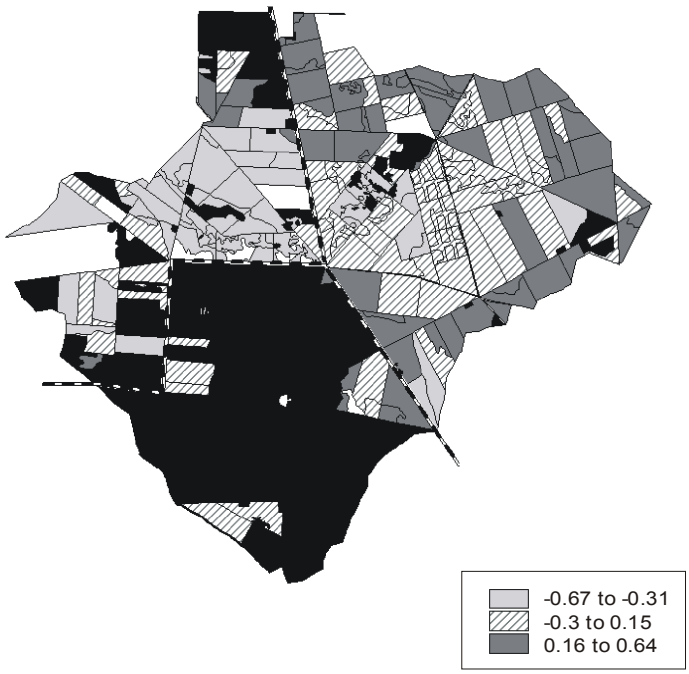


Figure 4: Spatial representation of sub-plots scores on the first axis of the PCA performed on the proportion table of coppices data set. The sub-plots with no data available on coppices are represented in white. The sub-plots with no coppices are represented in black. The dashed line represents the division performed between the two timber stands types (oak and beech stands) and the two coppices types (hornbeam vs. maple dominant coppices).

Discussion

Different spatially structured vegetation communities occurred in the Chizé reserve. This result was expected because the forest has been managed by the ONF for more than 200 years. Three habitats can thus be identified within the reserve (Fig.4): the maple dominant coppices in the northwest part of the oak stand, the hornbeam dominant coppices in the northeast part the oak stand, and the beech stand without coppices in the south. This result somewhat validates those obtained in our previous study (Pettorelli et al. 2001), where two habitat types were recognized, the north zone corresponding approximately to the oak stands and the south zone corresponding to the beech stand. However, our results underline the heterogeneity of the oak stand quality for roe deer, as hornbeam is a principal and preferred component of roe deer diet in spring and summer (Duncan et al. 1998), two periods of high-energy requirements for roe deer (Andersen et al. 1998). The distinction of separate coppices types in the oak stand could thus improve the importance of habitat quality in determining roe deer condition indexes like body mass. We found no coppices in the beech stand. Coppices was identified as a major determinant of roe deer habitat quality, as coppices are directly linked to food availability and hiding possibilities (Cien and Sempéré 1989; Mysterud et al. 1999). This results could thus explain why lower fawn and adult body masses were found in the beech stand (Pettorelli et al. 2001; Pettorelli et al. 2002).

Coupling multivariate analysis and Geographic Information system appears to be a powerful for dealing with vegetation structure analysis. There are numerous studies in landscape ecology over the last decade, which have coupled multivariate analyses and GIS to provide some good description of vegetation structure in particular regions (Kadmon and Danin 1997; Ohmann and Spies 1998; Kadmon and Heller 1998; Guisan et al. 1999). Generally, those approaches are coupled with clustering methods, in order to determine major vegetation communities occurring in the considered region.

In our case, clustering methods were not conceivable for different reasons. Firstly, clustering methods generally do not consider spatial proximity, which means that categories are defined without the consideration of spatial location of the sampled stations. Thus, defining for example three clusters could imply to define several spatial entities belonging to each of the three defined vegetal communities. This implies that the scale of the definition of major patterns is reduced, and thus that the connection with deer data is compromised. Spatially constrained clustering procedures are available, but were not usable as those procedures essentially concern binomial or Poissonian data provided by health science (Kulldorff 1997). Secondly, clustering methods (which are spatially constrained or not) do not take into account natural frontiers like roads, which are of major interests for roe deer home ranges delimitation (Hewison et al. 1998). As our purpose was to define major spatial structure of vegetation at a scale that is relevant for coupling vegetation and deer data, we needed to consider such components of deer habitats. Finally, there was a strong pattern in vegetation community distribution and a simple spatial representation of PCA scores was clearly enough to define the major coppices and stand types occurring in the reserve. The goal of our study was more to define efficiently the major vegetation communities occurring in the reserve, in order to better understand the role of habitat on deer population dynamics. Our aim was not to develop a robust method for apprehending fine spatial vegetation structures.

However, this method could suffer from the fact that delimitation between different classes is free from an objective procedure. This is particularly clear when dealing with gradient structures, like the hornbeam-maple gradient occurring in our oak stand. But this simplification could also be viewed as the important point in our study. Forestry data are both abundant and underused. Moreover, there are more and more evidence that habitat quality is a major attribute of individual variability, like sex or age. The need of coupling vegetation

and animal data is thus growing up, like the need for tools. We did not provide any new methodology, as PCA, GIS and the couple use of both tools do exist for a while. What is new is the importation of such landscape ecology methodologies in management purposes. The connection between vegetation and animal data is clearly a new challenge for ecologists but also for managers who need simple and efficient tools strong enough to integrate sites particularities. We present a simple use of GIS and PCA, which allow obtaining at the relevant scale simple and efficient results on habitat structure. Those results can then be directly connected with deer data like weight or fecundity in management purposes.
We are conscious that our approach could be improved, as such biological situations could constitute some very interesting departure points for biometry researchers.

Acknowledgments
We are grateful to the Forestry National Organization and to all field assistants and volunteers that spent time collecting data on the study site. Special thanks to Jean Michel Gaillard, Daniel Chessel, Patrick Duncan, Anne Loison, Erling Solberg, Noël and Nadine Guillon.

References
Andersen, R., P. Duncan, and J. D. C. Linnell. (1998) The european roe deer: the biology of success. Scandinavian University Press, Oslo, Norway.
Andersen, R., J. M. Gaillard, J. D. C. Linnell, and P. Duncan. 2000. Factors affecting maternal care in an income breeder, the European roe deer. Journal of Animal Ecology 69:672-682.
Aitchison, J. 1983. Principal component analysis of compositional data. Biometrika 70:57-65.
Cibien, C., and A. Sempéré. 1989. Food availability as a factor in habitat use by roe deer. Acta Theriologica 34:111-123.
Conradt, L., T. H. Clutton-Brock, and F. E. Guinness. 1999. The relationship between habitat choice and lifetime reproductive success in female red deer. Oecologia 120:218-224.
Coulson, T., S. D. Albon, F. E. Guinness, J. Pemberton, and T. H. Clutton-Brock. 1997. Population substructure, local density, and calf winter survival in red deer. Ecology 78:852-863.
Coulson, T., S. D. Albon, J. Pilkington, and T. H. Clutton-Brock. 1999. Small-scale spatial dynamics in a fluctuating ungulate population. Journal of Animal Ecology 68:658-671.
De Crespin de Billy, V., S. Dolédec, and D. Chessel. 2000. Biplot presentation of diet composition data: an alternative for fish stomach content analysis. Journal of Fish Biology 56:961-973.
Duncan, P., H. Tixier, R.R. Hofman, and M. Lechner-Doll. 1998. Feeding strategies and the physiology of digestion in roe deer. Pages 91-116 *in* R. Andersen, P. Duncan, and J. D. C. Linnell, editors. The european roe deer : the biology of success. Scandinavian University Press, Oslo, Norway.
Fretwell, S. D., and H. L. Lucas. 1970. On territorial behavior and other factors influencing habitat distribution in birds I. Theoretical development. Acta Biotheoriologica 19:16-36.
Gabriel, K. R. 1971. The biplot graphic display of matrices with application to principal component analysis. Biometrika 58:453-467.
Gaillard, J. M., D. Delorme, J. M. Boutin, G. Van Laere, B. Boisaubert, and R. Pradel. 1993. Roe deer survival patterns: a comparative analysis of contrasting populations. Journal of Animal Ecology 62:778-791.
Gaillard, J. M., M. Festa-Bianchet, N. G. Yoccoz, A. Loison, and C. Toigo. 2000. Temporal

variation in fitness components and population dynamics of large herbivores. Annual Review of Ecology and Systematics 31:367-393.
Gilpin, M., and I. Hanski. 1991. Metapopulation dynamics: empirical and theoretical investigations. Academic Press, London, UK.
Goodall, D. W. 1954. Objective methods for the classification of vegetation III. An essay in the use of factor analysis. Australian Journal of Botany 2:304-324.
Guisan, A., S. B. Weiss, and A. D. Weiss. 1999. GLM versus CCA spatial modeling of plant species distribution. Plant Ecology 143:107-122.
Hewison, A. J. M., J. P. Vincent, and D. Reby. 1998. Social organization of European roe deer. Pages 189-220 *in* R. Andersen, P. Duncan, and J. D. C. Linnell, editors. The european roe deer : the biology of success. Scandinavian University Press, Oslo, Norway.
Hotelling, H. 1933. Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology 24:417-441.
Jenks, G. F., and F. C. Caspall. 1971. Error on choropleth maps: definition, measurement, reduction. Annals of the Association of American Geographers 61:217-244.
Kadmon, R., and A. Danin. 1997. Floristic variation in Israel: a GIS analysis. Flora 192:341-345.
Kadmon, R., and J. Heller. 1998. Modelling faunal responses to climatic gradients with GIS: land snails as a case study. Journal of Biogeography 25:527-539.
Kulldorf, M. 1997. A spatial scan statistic. Communications in statistics: theory and methods 26:1481-1496.
Milner-Gulland, E. J., T. N. Coulson, and T. H. Clutton-Brock. 2000. On harvesting a structured ungulate population. Oikos 88:592-602.
Mitchell, A. 1999. The ESRI Guide to GIS Analysis. ESRI Press, USA.
Mysterud, A., P. K. Larsen, R. A. Ims, and E. Ostbye. 1999. Habitat selection by roe deer and sheep : does habitat ranking reflect resource availability? Canadian Journal of Zoology 77:776-783.
Mysterud, A., R. Langvatn, N. G. Yoccoz, and N. C. Stenseth. 2001. Plant phenology, migration and geographic variation in body weight of a large herbivore: the effect of a variable topography. Journal of Animal Ecology 70:915-923.
Ohmann, J. L., and T. A. Spies. 1998. Regional gradient analysis and spatial pattern of woody plant communities of Oregon forests. Ecological Monographs 68:151-182.
Pettorelli, N., J. M. Gaillard, P. Duncan, J. P. Ouellet, and G. Van Laere. 2001. Spatial variation in habitat quality, local density and phenotypic quality in roe deer. Oecologia 128:400-405.
Pettorelli, N., J. M. Gaillard, P. Duncan, G. Van Laere, P. Kjellander, O. Liberg, D. Delorme, D. Maillard. 2002. Variations in adult body mass in roe deer: the effects of population density at birth and of habitat quality. Proceedings of the Royal Society of London (B) 269:747-754.
Ter braak, C. J. F. 1983. Principal components biplots and alpha and beta diversity. Ecology 64: 454-462.
Thioulouse, J., D. Chessel, S. Dolédec, and J. M. Olivier. 1997. ADE-4: a multivariate analysis and graphical display software. Statistics and Computing 7:75-83.
Tilman, D., and P. Kareiva. 1997. Spatial ecology: the role of space in population dynamics and interspecific interactions. Princeton University Press, Princeton, New Jersey.
Tuljapurkar, S., and H. Caswell. 1996. Structured population models in marine, terrestrial, and freshwater systems. Chapman and Hall, New York, USA.
Venables, W. N., and B. D. Ripley. 1997. Modern Applied Statistics with S-PLUS, Second Edition. Springer, Berlin, Germany.

NATHALIE PETTORELLI · STEPHANE DRAY · JEAN-MICHEL GAILLARD · DANIEL CHESSEL · PATRICK DUNCAN · ANDREW ILLIUS · NADINE GUILLON· FRANCOIS KLEIN · GUY VAN LAERE

The distribution of preferred plant species in spring determines spatial variation in the body mass of roe deer fawns in winter

Abstract In order both to quantify the importance of the spatial component in the determination of body mass and to identify precisely which resources induce spatial variation in this trait, we determined at a fine scale the spatial structure of both vegetation and body mass. We analysed the relative importance of temporal and spatial variation in the winter body mass of roe deer fawns over 24 years using data on 1235 individuals captured at known locations. We then qualitatively related the spatial variation in body mass with plant species composition from 578 sites in the forest. The combination of temporal and spatial effects accounted for 36.8% and 39.2% of the among-individual variability observed in the body mass of male and female eight-month fawns, while the cohort effect alone accounted for 20.4% and 20% of the observed variance in males and females respectively. The spatial distribution of fawn body mass was perennial over the 24 years considered and similar in the two sexes, and predicted values showed a 2 kg range according to location in the reserve. The occurrence of three plant species that are known to be important food items in spring/summer roe deer diets, hornbeam (*Carpinus betulus*), bluebell (*Hyacinthoides sp*.) and wild asparagus (*Ornithogalum sp.*) was positively related to winter fawn body mass of both sexes. The occurrence of species known to be avoided in spring/summer roe deer diets (*e.g.* butcher's broom, (*Ruscus aculeatus*), and beech, (*Fagus sylvatica*)), was negatively related to fawn body mass. We conclude that the distribution of plant species that are actively selected during spring and summer is an important determinant of spatial variation in winter fawn body mass, and therefore is a likely critical factor of roe deer population dynamics.

Key words: food supply variation, key resources, population dynamics, spatial heterogeneity, spatial scale

Introduction

Ecological factors such as density or weather generally account for most of the temporal variation observed in the dynamics of mammalian populations (Caughley 1977; Gilpin and Hanski 1991; Tilman and Kareiva 1997). Furthermore, the variability associated with phenotypic differences such as sex (Clutton-Brock et al. 1982), age (Charlesworth 1980) or social status (Lott 1991) strongly affects the dynamics of particular populations (Gaillard et al. 2000). Habitat quality has recently been identified as a significant source of individual variability in ungulates (Coulson et al. 1997; Conradt et al. 1999; Coulson et al. 1999). Recent studies have also demonstrated that ignoring spatial variation in performances by *e.g.* harvesting two contrasting sub-areas at the same rate, could induce a significant loss in the proportion of managers' potential profits (Milner-Gulland et al. 2000).

Roe deer *(Capreolus capreolus*) are small and widespread cervids (20-30 kg; Andersen et al. 1998), the populations of which have increased strongly during the last decades in Western Europe (Andersen et al. 1998). They are highly sedentary ungulates with little body reserves and show little within-year variation in body mass (income breeder strategy; Jönsson 1997, Andersen et al. 2000). The adults and particularly the fawns are thus very sensitive to variations in the availability of resources.

Fawn body mass can be considered as a reliable proxy for population condition (Hanks 1981). Moreover, fawn body mass in January-February is closely related to subsequent adult body weight (Gaillard 1994), winter survival (Gaillard et al. 1993), and age at maturity (Gaillard et al. 1992). Understanding which factors influence this fundamental life history trait is thus of prime importance.

We aim here first to determine the spatial structure of both vegetation and winter body mass of roe deer fawns, and then to test for links between the two.

A common way of dealing with spatial data on individual performances involves lumping the population's habitat into a few blocks, generally based on vegetation analysis (Pettorelli et al. 2001) or on assumptions about plant quality such as phenology (Mysterud et al. 2001) and then analysing variation in animal performance among the pre-defined areas.

Previous results in this population have showed that spatial variation in timber stands influenced the body mass of fawns (eight-month fawns were on average 0.6 kg heavier in the oak stand than in the beech stand; Pettorelli et al. 2001) and adults (males and females were respectively 0.9 kg and 0.5kg heavier in the oak stand, Pettorelli et al. 2002). We have also demonstrated that principal and preferred plant species in roe deer diet (Duncan et al. 1998) are more common in the oak stand than in the beech stand (Pettorelli et al. 2001).

The scale considered in those analyses based on vegetation data was defined by an *a priori* stratification of the reserve. However, scale in spatial ecology may be a key determinant of the strength of patterns and processes observed (Levin 1992; Ray and Hastings 1996; Donaldson and Nisbet 1999). Changing the scale of our analysis by separating vegetation and life history traits would allow us to access finer processes like the possible occurrence of major variations within each stand type, and to work at a scale fine enough to reveal the importance of individual plant species on fawn body mass. We aimed here to go beyond the correlation between food distribution and fawn body mass, and test whether the spatial distribution of preferred plant species induces a spatial structure in fawn body mass distribution.

Material and methods

The study site

This study was carried out in the 2614 ha fenced Chizé reserve situated in western France (46°05'N, 0°25'W), which has an oceanic climate with Mediterranean influences,

characterized by mild winters and hot, dry summers. Mean monthly temperatures vary from 5.5°C (January) to 20.5°C (July) and precipitation from 49mm (August) to 102mm (December); summer droughts are common (Gaillard et al. 1996).
The forest is managed by the Office National des Forêts, and is divided by forest trails into plots. The major tree species are oak (*Quercus sp*.), beech (*Fagus sylvatica*), hornbeam (*Carpinus betulus*), maple (*Acer campestris* and *A. monspessulanum*), dogwood (*Cornus mas, C. sanguinea*) and hawthorn (*Crataegus sp*.) (Cibien and Sempéré 1989).

On a coarse scale, the reserve presents contrasting habitats, according to the timber stand and the nature of the coppices (Pettorelli et al. 2001; Pettorelli, Dray, Maillard and Villarubias unpublished data). The dominant tree of the northern part of the reserve (1,397 ha) is oak. Conversely, the southern plots (1,143 ha) are dominated by beech. Oak stands can be further distinguished according to the shrub-layer: the eastern part is dominated by hornbeam, the western part by maple.

Data collection

Fawn body mass

The roe deer population has been intensively monitored by capture-mark-recapture methods since 1978 (Gaillard et al. 1993). Ten days of capture in January and February allow 150-350 roe deer to be caught each year. Most animals are released with individual collars and the remainder are exported. In a particular capture session, >100 people drive animals into two to five km of nets, which enclose chosen forestry plots.

About one third to one half of the area of the reserve is sampled each year, capture areas being varied over time. A total of 1235 fawns (639 males and 596 females) were captured between January and February 1978 to 2001, and weighed using an electronic balance. As females form small groups with their fawns and occupy overlapping home ranges (Hewison et al. 1998), we can assume that capture location of eight-month fawns reveals the place they were reared in. The site of capture and the sex were noted. We attributed to each fawn the coordinates of the centre of the capture site.
All the information was transferred into a GIS (Geographic Information System; Mitchell 1999).

The Vegetation

We assessed the distribution of the vegetation by determining areas of species occurrence encountered during the sampling (without a pre-established list).
Five hundred and seventy eight random plots of 1 m^2 and on average 200 m-apart were sampled by one of us (N.P.) in 2001 from May 15 until mid-June (Fig. 1), when all herbaceous and woody genera accessible to roe deer (<1.20 m) are recognizable.
The coordinates of all sampling plots were calculated by a Global Positioning System (Magellan GPS 315, 12 parallel channels, 15m RMS accuracy), and were transferred to the GIS. To sample a particular plot, a quadrat was thrown at chance (generally in front of us) when approaching the defined average distance between two plots (indicated by the G.P.S.), and the presence of any herbaceous and woody plant was noted. Data were collected to the generic level (98 genera).

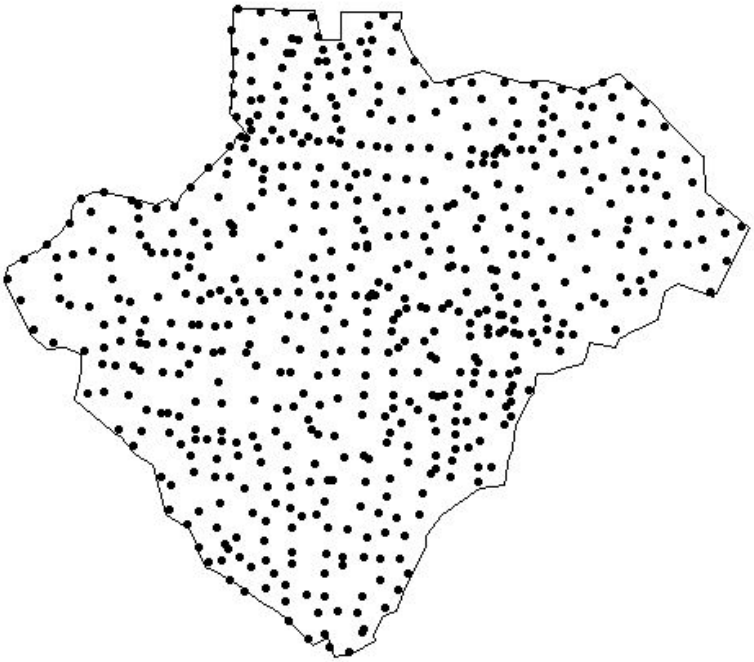


Figure 1: The spatial distribution of the 578 plots sampled from May 15 to June 15 2001.

Statistical procedures

The distribution of the centres of the capture sites during our 24 years monitoring is discrete (Fig. 2). We first performed a hierarchical cluster analysis (Everitt 1974, see Coulson et al. 1997 for an application on red deer *Cervus elaphus*) on capture site centres' spatial coordinates in the Chizé reserve. Fourteen clusters, which well summarized the 24 years of monitoring in Chizé, were considered in our analysis (Fig. 2): they provided a good trade-off between the number of individuals per year and per clusters and the number of clusters to be considered. As we do not consider any information on habitat prior to the analysis of fawn body masses, and as we do not expect any particular spatial distribution of this variable, we consider here the fourteen clusters as independent and we search for any spatial autocorrelation (any spatial pattern) among clusters. For that, we have performed a spatial representation of the expected variable in the different clusters (Figure 3) that will allow us to define the spatial pattern in fawn body mass distribution.

Time was previously reported in this site as a major structuring factor in fawn body mass, with a 5 kg range between extreme cohorts (Gaillard et al. 1996). As there is a strong auto-correlation in the effect of time on fawn body mass due to the effect of density dependence on this life history trait (Gaillard et al. 1996), we decided to model time using a five-degree polynomial (Diggle 1990). We consider five degree, as when more degrees were added, the Akaike Information Criterion (Sokal and Rohlf 1995) of the fit did not change by more than 5%.

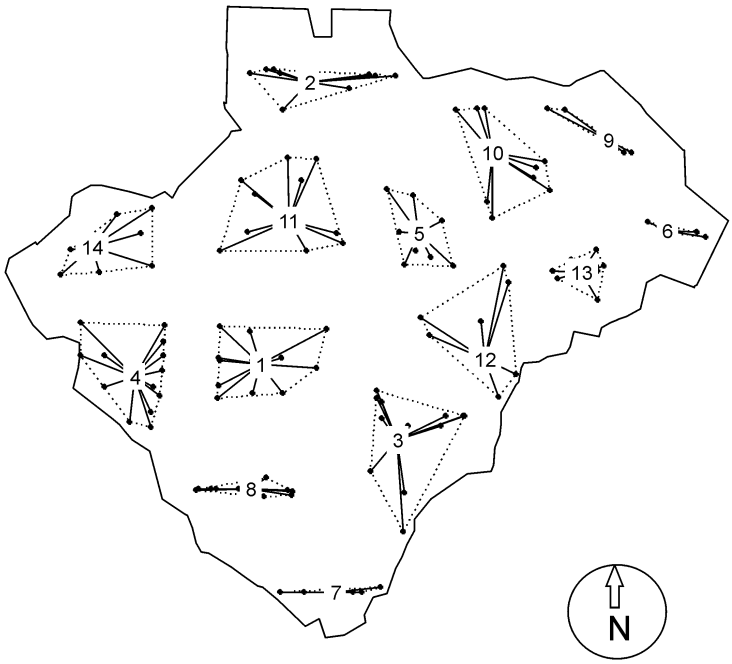


Figure 2: Capture site centers' distribution during the 24 years monitoring. Each dot represents a particular capture site center. From this discrete distribution, fourteen clusters (numbered in the figure) were defined using a hierarchical cluster procedure. The black lines represent the attribution of capture centers to particular clusters. The dashed lines only give an idea of the considered cluster area.

Male fawns were consistently heavier than female fawns (Gaillard et al. 1996), so the analysis was replicated for each sex. For each sex, one model considering only the temporal variation, and another one considering both temporal and spatial variations were run. We tested using ANOVA procedure (Sokal and Rohlf 1995) whether these temporal and spatial variables and the interactions between them were significant.

Linear regression (Sokal and Rohlf 1995) between observed and predicted values of fawn body mass from the interactive model was used to quantify the relative importance of year and location on the among-individual variation in this life history trait.

Finally, we analysed the spatial distribution of plant genera collected in May 2001 by using a global PCA (Thioulouse et al. 1995). This procedure is highly recommended when trying to extract the major plant community spatial patterns occurring in the reserve. The procedure of the Global PCA is very near from the standard PCA procedure. However, contrary to the standard PCA in which a linear combination of variables (the plant genera here) that maximizes the variance among the 578 sampled stations is looked for, the global PCA maximizes the spatial correlation among those stations, introducing thus the neighbourhood matrix among the sampled stations (instead of the usual identity matrix) in the statistical triplet to be analysed (Thioulouse et al. 1995). All genera with a probability of occurrence <1% were removed from the analysis, which left 57 genera.

Results

As expected, space was a major structuring factor of fawn body masses in Chizé.

Considering the same interactive model for both sexes, the integration of spatial location of captured fawns allowed us to account for 36.75% and 39.2% of variation in winter body mass of fawns at the individual level, for males and females respectively. Thus, nearly 40% of the variance in male and female eight-month fawn body mass was accounted for by cohort and the broad geographical location of the mothers' home range. Time alone accounted

for 20.36% and 19.96% of the variance in fawn body mass, for males and females respectively.

Considering both temporal and spatial factors on male and female body mass, we observed no significant interaction between them in males (F=1,27, df=65, 555; P=0,08), the main effects being, however, highly significant (space: F=4,67, df=13, 555, P<0,001; time: F=35,72, df=5, 555; P<0,001). A significant interaction between spatial and temporal factors was observed in females (F=1,62, df=65, 512, P=0,002), contrary to what we found in males. However when we repeated the analysis by removing only 3 extreme individuals, the interaction between space and time was no longer significant (F=1,25, df=65, 509, P=0.097), the main effects being highly significant (space: F=4.88, df=13, 509, P<0,001; time: F=34.78, df=5, 509, P<0,001).

For both sexes, the same additive model considering the effects of space and time on fawn body mass was thus retained. In both cases, space and time play a nearly equal role in determining their masses (because time alone accounts for about 20% of variance, and because the interaction between both temporal and spatial factors accounts for a negligible part of the variance). Under this additive model, we obtain the same perennial spatial distribution patterns in both sexes, as well as the same range of variation (Fig. 3). Predicted values from the additive model including space and time show a 2kg range (from the poorest cluster to the better one) for males and females. There is thus some spatial autocorrelation in fawn body masses in winter among clusters, the pattern showing a North East-South West gradient.

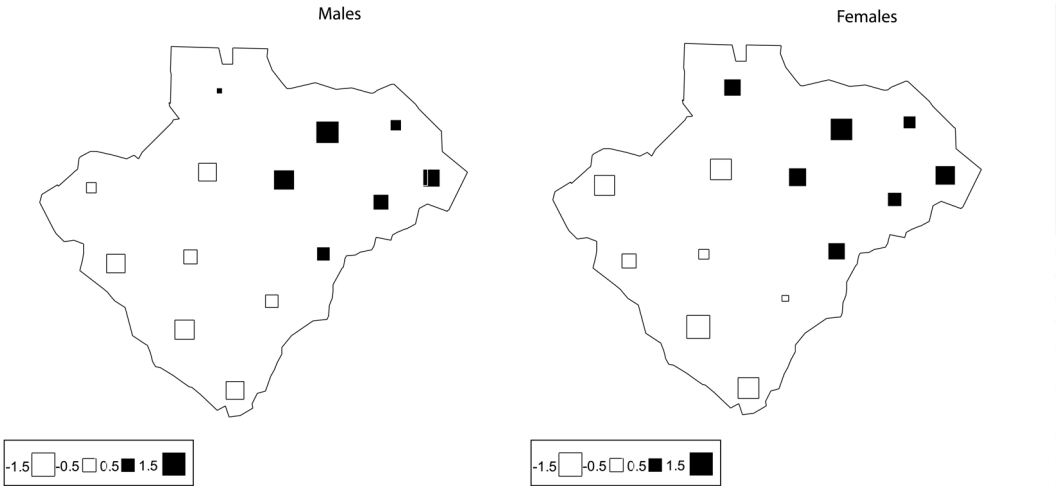


Figure 3: Spatial distribution of predicted male (a) and female (b) body masses according to the additive model considering a five-degree polynomial of time and clusters. Predicted values were normalized. Squares represent the fourteen clusters. Black squares represent clusters with higher expected body mass than the average cluster, white squares with lower expected body mass.

Considering the distribution of the eigenvalues of the global PCA, we mainly retained the first axis as the major structuring axis of the data set. Indeed the first eigenvalue was twice higher than the second one. Plotting the scores of the stations shows a pronounced spatial structure occurring in plant community distribution. This axis clearly opposed the scores of the stations sampled in the North East part to the scores of the stations belonging to the South part of the reserve (Fig. 4). The distributions of butcher's broom (*Ruscus aculeatus*), beech, and brambles (*Rubus sp*.) were highly positively related to the first axis of this PCA (Table 1), meaning thus that those species mainly occur in the South part of the reserve where higher scores are essentially found. On the contrary, the distributions of hornbeam, bluebells (*Hyacinthoides sp*.) and wild asparagus (*Ornithogalum sp*.) were highly negatively related to this first axis (Table 1), meaning that those ones mainly occur in the North East part of the reserve where lower scores are found. Genera like pines (Pinus sp.), ivy (Hedera helix), ash

(Fraxinus excelsior), hazel (Corylus avellana), dogwood (Cornus sp.) and sorb (Sorbus domestica) were poorly related with this first axis (Table 1) and thus occur indifferently in those two parts.

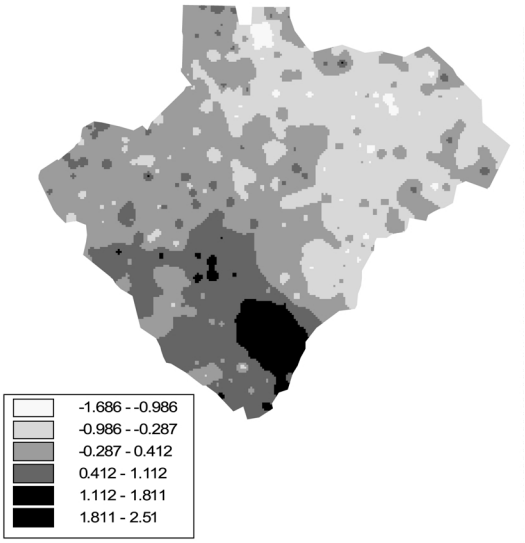


Figure 4: The scores of the sampled stations in May 2001 on the first axis of the Global PCA procedure. High scores are found essentially in the Northeast of the Reserve, whereas low scores are found in the Southwestern part.

Table 1: Scores (ranging from 0.341 to –0.315) of the genera on the first axis of the global PCA. Only highly correlated genera and genera described as important or avoided in the roe deer diet were represented (Duncan et al. 1998; Tixier et al. 1997).

Species	coordinates
Carpinus betulus	-0.315
Ornithogalum sp.	-0.243
Hyacinthoides sp.	-0.185
Anemone nemorosa	-0.140
Evonymus europeatus	-0.097
Arum sp.	-0.078
Crataegus sp.	-0.077
Viola sp.	-0.071
Ulmus sp.	-0.042
Sorbus torminalis	-0.025
Fraxinus excelsior	-0.009
Hedera helix	-0.003
Sorbus domestica	-0.001
Pinus sp.	0.033
Corylus avellana	0.060
Cornus sp.	0.077
Prunus spinosa	0.121
Lonicera periclymenum	0.149
Acer sp.	0.153
Rosa sp.	0.170
Rubia peregrina	0.172
Euphorbia sp.	0.185
Quercus sp.	0.187
Ligustrum vulgare	0.214
Clematis vitalba	0.221
Fagus sylvatica	0.262
Ruscus aculeatus	0.328
Rubus sp.	0.341

In conclusion, two main plant communities with a North East–South gradient structure among them occur in the reserve according to this data set: the first one is dominated by hornbeam, wild asparagus and bluebells and occur in the oak stand in the North East part of the reserve. The other community is dominated by brambles, beech and butcher's broom and occur in the beech stand in the South part of the reserve.

Discussion

We have demonstrated here that the spatial component plays a fundamental role in explaining individual variability in winter body mass of fawns. Winter mass is a reliable proxy of population condition (Maillard et al. 1989; Vincent et al. 1995; Gaillard et al. 1996), and is related to future individual performance (Gaillard 1994; Pettorelli et al. 2002). By separating information on vegetation and life history trait components' analysis we have shown that the spatial component is a major factor in shaping variation in winter body mass of fawns in Chizé. We have previously found (Pettorelli et al. 2001) a 0.6 kg difference between male and female fawns raised in the oak stand or in the beech stand. Here, predicted values show a 2 kg (about 15% of the average body mass of fawns in winter) range according to habitat location. Considering this constant 2kg range in body masses of eight-month fawns, there is, under harsh conditions, a non-negligible risk that such a difference leads fawns living in poorest habitats to be under a threshold body mass for surviving over their first winter (Gaillard et al. 1993).

The spatial structure in the distribution of fawn body mass is perennial over 24 years, which means that no events like summer droughts (1993 for example), density increase (between 1982 to 1987), or the storm Lothar (2001) have interacted with this spatial pattern. The spatial constraints on body mass seem thus to be strong, and linked to perennial spatial structures in the reserve. As there is a weak predation on juveniles (Gaillard et al. 1993), it underlines thus the interest of searching such perennial structure in the plant community distribution.

In most ungulate species male fawns generally suffer more under harsh conditions than females (Glucksman 1974). We should therefore not expect to obtain the same range and spatial pattern of predicted body masses for both sexes. However, in this study we obtained a similar spatial distribution and the same range of predicted body masses in the two sexes. Contrary to many other ungulate species (see Clutton-Brock et al. 1982), males and females have similar birth weights, post-natal growth rates and juvenile survival rates in roe deer (Gaillard et al. 1993, Gaillard et al. 1998). Our results are thus consistent with the view that roe deer are close to the ungulates whose life history traits are only weakly influenced by sexual selection.

A strong result in this study is that we obtained the same spatial pattern in fawn body mass distribution of both sexes (Fig. 3) and in the distribution of floristic composition (Fig. 4). But the limitation of this study is that different time scales were used for measuring the variables we analyzed and it seems thus difficult to relate the two kinds of information (*i.e.* vegetation and body mass): contrary to the body mass database collected over 24 years, the vegetation data were collected in only one sampling session (2001), after the hurricane Lothar hit France in December 1999. However, according to all information gathered concerning the spatial distribution of plant communities over the last ten years, the important floristic community structures seem to be relatively perennial. A previous study has revealed that there was an important positive relationship between our soil-vegetation data collected in 2001 after the storm and stand/coppices data collected in 1993 by foresters (Dray, Pettorelli and Chessel in press). Considering the database collected in 1993 by GVL (Pettorelli et al. 2001) we also observe a good correlation between the distribution of butcher's broom, brambles and hornbeam in 1993 and 2001 (only woody species were sampled by GVL in 1993). Additional

surveys available for 1995 and 1997 (Pettorelli et al. 2001) confirm also this pattern (although these were performed by different observers and may be less reliable; Morellet 1998). Finally, bluebells and wild asparagus are bulb plant species (Rameau et al. 1989), which implies a good autocorrelation between yearly spatial distributions. Although we cannot assess the generality of our results, we can thus assume that the different time scales used in this study should not affect our conclusions.

Roe deer are generalist feeders but are highly selective (Duncan et al. 1998). The spring and summer seasons are critical for roe deer nutrition. In this period of high energy requirements the animals select food items, which are highly digestible and rich in soluble carbohydrates (Maizeret and Tran Manh Sung 1984) like hornbeam and bluebells (Tixier et al. 1997; Maizeret et al. 1991). Wild asparagus are also highly preferred (G.V.L., personal observations). These species induced the major spatial structure observed in the floristic composition at Chizé because they are present only in the Northeast part of the reserve, contrary to the other principal food resources of roe deer (like oak, dogwood, ivy, hawthorn or maple; Duncan et al. 1998). Moreover higher body masses were found in this particular part of the reserve. It is therefore likely that the difference in the distribution of these highly preferred plants affected the food supply of the breeding females, which in turn led to a high degree of spatial variation in fawn body mass.

Brambles are consumed all year round by roe deer (Duncan et al. 1998, Tixier et al. 1997). It may therefore seem surprising that the relative availability of brambles was negatively related to the PCA first axis and thus to fawn body mass. We found in another study that fawns in the North East part of the reserve survived better over their first winter than those in other locations (Pettorelli, Gaillard, Duncan, Maillard, Van Laere and Delorme unpublished data). In addition, the positive relationship between hornbeam, bluebells, asparagus, and body mass distributions we report here suggests that these plants rather than brambles are the important resources. These plants are present and heavily used by roe deer in spring (for bluebells and asparagus) and summer (hornbeam). Therefore, spatial differences in mass might be due mainly to differences in growth rates in the first months of life. 65% of adult mass is usually reached when fawns are eight-months old (Gaillard 1988), but growth rate is much lower in autumn than in spring and summer. Growth in roe deer follows a monomolecular model characterized by a growth rate declining continuously from birth (Portier et al. 2000). Resources heavily used during the autumn and winter periods like brambles could therefore be used for maintenance rather than for growth. The spatial distribution of fawn body mass should therefore be detectable from the end of summer.

In temperate ungulates, survival of fawns is the key factor determining the dynamics of the population (Gaillard et al. 2000). The factors causing variation in this parameter are, therefore, of considerable interest. A recent study modelling ungulate population dynamics in arid and semi-arid grazing systems distinguished key resources, whose supply was shown to be crucial to dry-season survival, from other resources that did not affect the key factor (Illius and O'Connor 2000). Key resources were thus defined as *resources whose supply determines the size of the key factor of the population*. Of course, one would not expect to find a single resource whose abundance completely determines the population dynamic of temperate generalist herbivore like *i.e.* roe deer (Duncan et al. 1998). Nevertheless, our results do suggest that the abundance of resources which are preferred during this key season could play a crucial role in affecting shaping survival of fawns in populations of this ungulate.

Acknowledgments

We thank the Office National de la Chasse for organizing all the captures of roe deer at the Chizé reserve. We are grateful to all the students, field assistants, and volunteers that spent time catching and monitoring the roe deer fawns on the study site. Special thanks to Atle Mysterud, Daniel Maillard and Anne Loison for ideas, comments and suggestions on previous drafts of this work.

References

Andersen R, Duncan P, Linnell JDC (1998) The european roe deer: the biology of success. Scandinavian University Press, Oslo

Andersen R, Gaillard JM, Linnell, JDC, Duncan P (2000) Factors affecting maternal care in an income breeder, the European Roe deer. J Anim Ecol 69:672-682

Caughley G (1977) Analysis of vertebrate populations. John Wiley and sons (eds), New York

Charlesworth B (1980) Evolution in age-structured populations. Cambridge University Press, Cambridge

Cibien C, Sempéré A (1989) Food availability as a factor in habitat use by roe deer. Acta Theriol 34:111-123

Clutton-Brock TH, Guinness FE, Albon SD (1982) Red deer: behaviour and ecology of two sexes. University Chicago Press, Chicago

Conradt L, Clutton-Brock TH, Guinness FE (1999) The relationship between habitat choice and lifetime reproductive success in female red deer. Oecologia 120:218-224

Coulson T, Albon SD, Guinness FE, Pemberton J, Clutton-Brock TH (1997) Population substructure, local density, and calf winter survival in red deer. Ecology 78:852-863

Coulson T, Albon SD, Pilkington J, Clutton-Brock TH (1999) Small-scale spatial dynamics in a fluctuating ungulate population. J Anim Ecol 68:658-671

Diggle PJ (1990) Time series: a biostatistical introduction. Clarendon Press, Oxford, UK

Donaldson DD, Nisbet RM (1999) Population dynamics and spatial scale: effects of system size on population persistence. Ecology 80:2492-2507

Duncan P, Tixier H, Hofman RR, Lechner-Doll M (1998) Feeding strategies and the physiology of digestion in roe deer. In: Andersen R, Duncan P, Linnell JDC (eds) The european roe deer: the biology of success, Scandinavian University Press, Oslo, pp 91-116

Everitt B (1974) Cluster analysis. Heinemann educational Books, London, UK

Gaillard JM (1988) Contribution à la dynamique des populations de grands mammifères : l'exemple du chevreuil. PhD thesis, University of Lyon, France

Gaillard JM, Sempéré AJ, Van Laere G, Boutin JM, Boisaubert B (1992) Effects of age and body weight on the proportion of females breeding in a population of roe deer. Can J Zool 70:1541-1545

Gaillard JM, Delorme D, Boutin JM, Van Laere G, Boisaubert B, Pradel R (1993) Roe deer survival patterns: a comparative analysis of contrasting populations. J Anim Ecol 62:778-791

Gaillard JM (1994) Réflexions sur la variabilité biodémographique des mammifères. Mémoire d'Habilitation à Diriger des Recherches, University of Lyon, France

Gaillard JM, Delorme D, Boutin JM, Van Laere G, Boisaubert B (1996) Body mass of roe deer fawns during winter in 2 contrasting populations. J Wildl Manage 60:29-36

Gaillard JM, Liberg O, Andersen R, Hewison AJM, Cederlund G (1998) Population dynamics of roe deer. In: Andersen R, Duncan P, Linnell JDC (eds) The european roe deer: the biology of success, Scandinavian University Press, Oslo, pp 309-336

Gaillard JM, Festa Bianchet M, Yoccoz NG, Loison A, Toigo C (2000) Temporal variation in fitness components and population dynamics of large herbivores. Ann Rev Ecol Syst 31:367-393
Gilpin M, Hanski I (1991) Metapopulation dynamics: empirical and theoretical investigations. Academic Press, London, UK
Glucksmann A (1974) Sexual dimorphism in mammals. Biol Rev 49:423-475
Hanks J (1981) Characterization of population condition. In: Fowler CW, Smith TD (eds) Dynamics of large mammal populations, Wiley, New York, pp 47-73
Hewison AJM, Vincent JP, Reby D (1998) Social organization of European roe deer. In: Andersen R, Duncan P, Linnell JDC (eds) The european roe deer: the biology of success, Scandinavian University Press, Oslo, pp 189-220
Illius AW, O'Connor TG (2000) Resource heterogeneity and ungulate population dynamics. Oikos 89:283-294
Jönsson KI (1997) Capital and income breeding as alternative tactics of resource use in reproduction. Oikos 78:57-66
Levin SA (1992) The problem of pattern and scale in ecology. Ecology 73:1943-1967
Lott DF (1991) Intraspecific variation in the social systems of wild vertebrates. Cambridge University Press, Cambridge, UK
Maillard D, Boisaubert B, Gaillard JM (1989) La masse corporelle : un bioindicateur possible pour le suivi des populations de chevreuils. G Faun Sauv 6:57-68
Maizeret C, Tran Manh Sung D (1984) Etude du régime alimentaire et recherche du déterminisme fonctionnel de la sélectivité chez le chevreuil des landes de Gascogne. G Faun Sauv 3:63-103
Maizeret C, Bidet F, Boutin JM, Carlino JP (1991) Influence de la composition chimique des végétaux sur les choix alimentaires des chevreuils. Rev Ecol Terre Vie 46:39-52
Milner-Gulland EJ, Coulson TN, Clutton-Brock TH (2000) On harvesting a structured ungulate population. Oikos 88:592-602
Mitchell A (1999) The ESRI Guide to GIS Analysis. ESRI Press, USA
Morellet N (1998) Des outils biométriques appliqués aux suivis des populations animales : l'exemple des cervidés. PhD Thesis, University of Lyon, France
Mysterud A, Langvatn R, Yoccoz NG, Stenseth, NC (2001) Plant phenology, migration and geographic variation in body weight of a large herbivore: the effect of a variable topography. J Anim Ecol 70:915-923
Pettorelli N, Gaillard JM, Duncan P, Ouellet JP, Van Laere G (2001) Spatial variations in habitat quality, local density and phenotypic quality in roe deer. Oecologia 128:400-405
Pettorelli N, Gaillard JM, Duncan P, Kjellander P, Liberg O, Delorme D, Maillard D, Van Laere G (2002) Variations in adult body mass in roe deer: the effects of population density at birth and of habitat quality. Proc R Soc Lond (B) 269:747-754
Portier C, Duncan P, Gaillard JM, Guillon N, Sempéré A (2000) Growth of European roe deer: Patterns and rates. Acta Theriol 45:87-94
Rameau JC, Mansion D, Dumé G (1989) Flore forestière française. Ministère de l'agriculture et de la forêt.
Ray C, Hastings A (1996) Density-dependence: are we searching at the wrong spatial scale? J Anim Ecol 65:556-566
Sokal RR, Rohlf FJ (1995) Biometry: third edition. Freeman WH and Company (eds), New York, USA
Thioulouse J, Chessel D, Champely S (1995) Multivariate analysis of spatial patterns: a unified approach to local and global structures. Env Ecol Stat 2:1-14
Tilman D, Kareiva P (1997) Spatial ecology: the role of space in population dynamics and interspecific interactions. Princeton University Press, Princeton, New Jersey

Tixier H, Duncan P, Scehovic J, Yani A, Gleizes M, Lila M. (1997) Food selection by european roe deer: effects of plant chemistry, and consequences for the nutritional value of their diets. J Zool (Lond) 242:229-245

Vincent JP, Bideau E, Hewison AJM, Angibault JM (1995) The influence of increasing density on body weight, kid production, home range and winter grouping in roe deer. J Zool (Lond) 236:371-382

Chapitre VII : Analyse de données incomplètes

Ce chapitre contient une publication présentant une nouvelle méthode d'analyse de données. Les variations temporelles d'une structure spatiale sont analysées à partir de données provenant de différents types d'échantillonnages d'un même espace. Il en résulte un tableau de données avec de nombreuses valeurs manquantes.

Les méthodes classiques d'analyse de données (ACP, AFC…) sont basées sur des principes d'algèbre linéaire et nécessitent la diagonalisation d'une matrice. La situation évoquée dans ce chapitre conduit à un tableau de données avec valeurs manquantes qui ne peut donc être analysé par une méthode classique nécessitant une diagonalisation. Nous proposons une nouvelle méthode permettant de s'affranchir de ce problème induit par les caractéristiques des données. Cette méthode est basée sur un usage conjoint des méthodes PLS (*Partial Least Squares*, Tenenhaus 1998) et de la technologie SIG. La méthode NIPALS, qui permet de faire une ACP par l'intermédiaire d'un algorithme itératif, a été implémentée dans le logiciel R afin de réaliser les calculs (annexe 4). Une application s'intéressant à la stabilité temporelle de la distribution spatiale des poids des faons dans la réserve de Chizé est proposée.

Multivariate analysis of incomplete mapped data

Stéphane Dray, Nathalie Pettorelli, Daniel Chessel

UMR CNRS 5558, Laboratoire de Biométrie et Biologie Evolutive, Université Claude Bernard Lyon 1, 69622 Villeurbanne Cedex, France.

Running title: *analysis of misaligned data*

Abstract: Classical multivariate analyses are based on matrix algebra and enable to summarize a table containing measurements of a set of variables for a set of sites. Incomplete mapped data consist in measurements in the same area of a set of variables recorded at different locations and so cannot be analysed with usual methods. We propose a new approach using GIS technology and NIPALS, an iterative multivariate method, to analyse the spatial patterns of this kind of data. We illustrate the method in studying data concerning the distribution of roe deer weights over years in a reserve.

Keywords: Partial least squares, spatio-temporal variations, PCA, areal interpolation

1. Introduction

Multivariate analysis is a natural tool to analyse the spatial patterns of a set of variables. Standard methods such as principal component analysis (Hotelling 1933) or spatially constrained method such as local or global PCA (Thioulouse et al 1995) are commonly used with GIS (Guisan et al 1999, Kadmon and Danin 1997, Zhang and Selinus 1998) to identify and represent multivariate spatial structures. Multivariate analyses are devoted to the analysis of the variations in a table containing the measurements of a set of variables for some locations (sites). Representation of scores from the analysis in the geographical space allows to identify the common spatial patterns of variation from the data (Goodall 1954). Multivariate analyses are based on matrix algebra (singular value decomposition) and data must be contained in a matrix where each column represent a variable and each row represent a site (e.g. Greenacre 1984). So, each site must be sampled for each variable otherwise data with missing values must be estimated or excluded from the analysis.
The purpose of this paper is the analysis of incomplete cartographic data. This deals with data collected in the same geographical region but where not all variables are measured in all sites. The analysis of spatial variations of this kind of data is not possible with usual multivariate analyses because the data cannot be entered in a variable-by-individual matrix. We propose a new methodology to study the spatial patterns of different variables measured in the same area but not exactly at the same locations. The approach is based on a joint use of Geographic Information System (GIS) technology and multivariate analyses. We analyse a data set on the distribution of roe deer weights over years in the Chizé reserve (France) to illustrate the method. The sampling scheme adopted for this study involves that the sampling locations where the data have been collected are not considered as points (like for most studies on spatial data) but as polygons. Indeed, in a particular capture session, people fenced some given forestry plots with 2-5 km of nets and animals enclosed in this area were counted and weighted. In our example, variables correspond to different dates and so we analyse the temporal variations of a spatial structure.

2. Spatial linkage

We consider an area with defined boundaries. Some parts of this area are sampled at the first date (figure 1). For the next dates (e.g. year 2, year 3…), others areas are sampled which could be different or could overlap the previous sampled areas. The first step of our procedure is to create a reference layer of spatial units. Administrative units, or other kind of space partitioning can define the spatial units. In this paper, we chose to define the spatial units as the quadrats of a grid. The choice of the quadrat size is discussed below. Then, for each year, it is easy to construct neighbouring relationships between the quadrats of the reference grid and the sampling areas (figure 2).
Let us consider, for the year p, that k_p areas have been sampled. We construct a grid of n quadrats. The easiest way to establish neighbouring relationships is to construct a matrix $\mathbf{A_p}$ with n rows and k_p columns where:

$$\mathbf{A_p}_{ij}=1 \text{ if quadrat } i \text{ intersects the sampling area } j$$
$$\mathbf{A_p}_{ij}=0 \text{ otherwise}$$

This neighbourhood matrix represents the strength of the potential interaction between quadrats and sampling areas. A more elegant and realistic way to fill this spatial weight matrix is to take into account the surface of overlap between sampling areas and quadrats. The matrix $\mathbf{A_p}$ is then filled as follow:

$$\mathbf{A_p}_{ij}=\frac{S_{i\cap j}}{\sum_{j=1}^{k_p} S_{i\cap j}}$$

where $S_{i \cap j}$ is the surface of the intersection between quadrat *i* and sampling area *j*. Data of figure 2 can be used to illustrate this statement. The link between quadrat Q5 and sampling area P1 is simply expressed by $\frac{a_1}{a_2+a_1}$ and by $\frac{a_2}{a_2+a_1}$ for Q5 and P2.

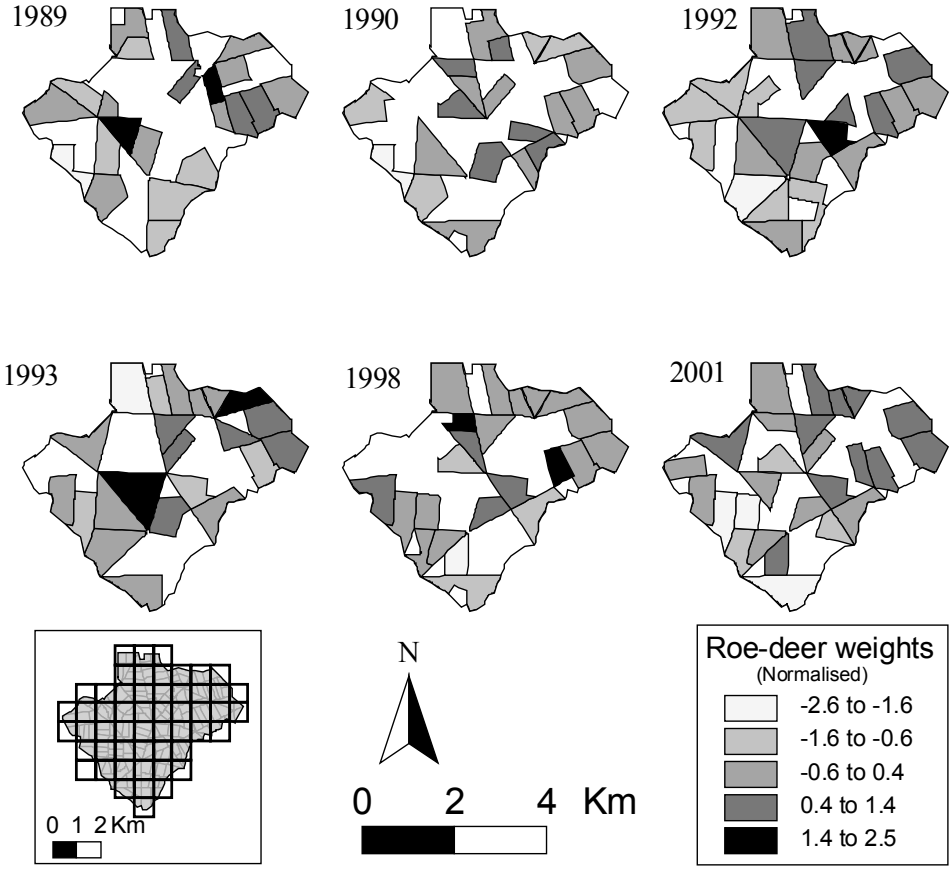


Figure 1: Sampling schemes and distribution of roe deer weights in the Chizé reserve. The number, the size, the shape and the location of sampling areas are different among years. Weights have been normalised by year. The reference grid that has been chosen for the analysis is also represented on the map of forestry plots.

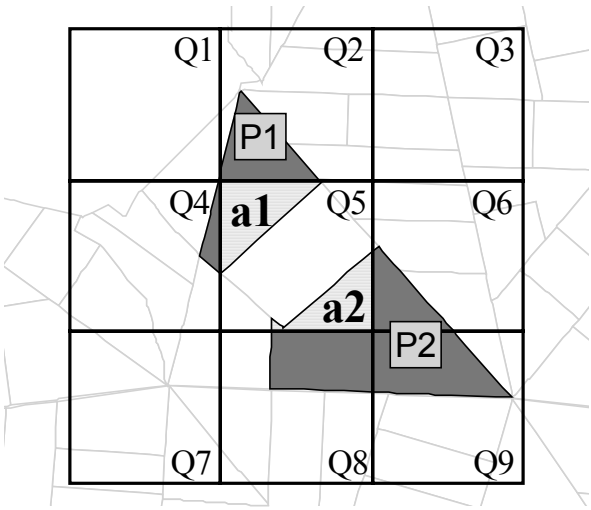


Figure 2: Crossing of the reference grid and sampling area. Establishment of the neighbourhood matrix between quadrats and sampling area is made in computing the area of the intersection between a quadrat and a sampling area. For example, the link between the quadrat Q5 and the sampling area P1 is simply expressed by a1/(a1+a2)

3. Construction of the data table

Let us consider a quantitative variable X measured in all sampling areas for each date. Hence for each year *p*, data consist in a vector $\mathbf{X_p}$ with k_p rows (figure 3). For each year, averages of X, weighted by overlapped area, can be computed for each quadrat and result in a vector $\mathbf{Z_p}$ with *n* rows. Computation of weighted average for the *i*-th quadrat for year *p* (i.e. *i*-th element of $\mathbf{Z_p}$) is simply the product of the *i*-th row of $\mathbf{A_p}$ by vector $\mathbf{X_p}$. But, if for year *p* no sampling area intersects a quadrat *i* (i.e. sum of elements of the *i*-th row of $\mathbf{A_p}$ is null) then a missing value is assigned for the *i*-th element of vector $\mathbf{Z_p}$. A matrix **Z** with *n* rows and *N* (total number of years) columns is then constructed in binding the *N* vectors $\mathbf{Z_p}$. Applying classical multivariate analysis using singular value decomposition on table **Z** is not possible because of the existence of missing values. An alternative is the use of NIPALS algorithm (Wold 1966) in the place of singular value decomposition.

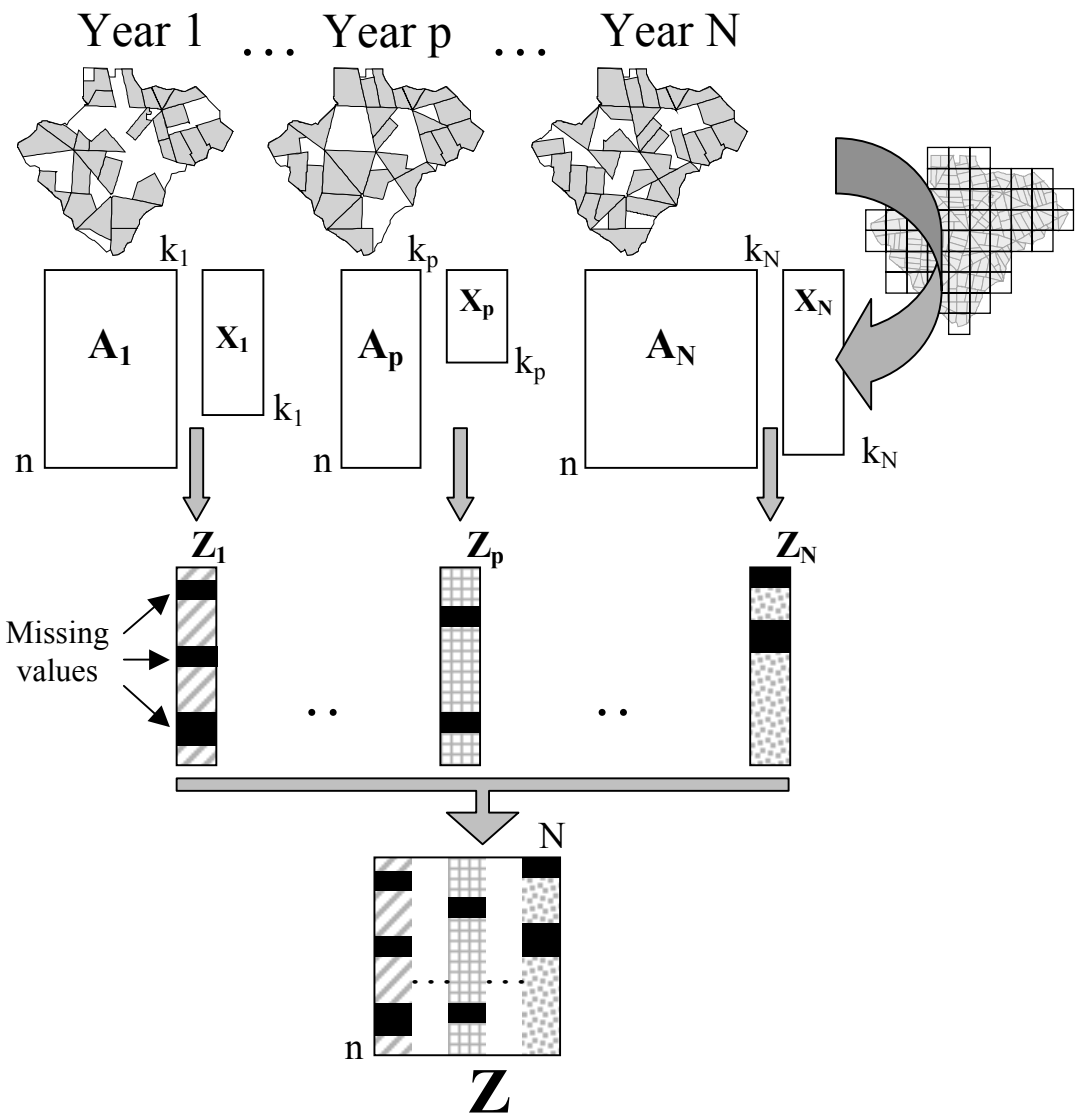


Figure 3: Schematic representation of the method. Crossing the sampling areas and the reference grid allows to obtain a neighbourhood matrix ($\mathbf{A_i}$) for each year *i*. For year *i,* measurements of the variable ($\mathbf{X_i}$) and neighbourhood matrix ($\mathbf{A_i}$) are used to compute average for each quadrat. A new table crossing quadrats and years is then created in binding all vectors $\mathbf{Z_i}$. The table **Z** will be analysed using NIPALS algorithm.

4. NIPALS Analysis

NIPALS (Nonlinear estimation by iterative partial least squares) is an algorithm, which is at the root of PLS regression. Wold 1966) presented this algorithm under the name of NILES (Nonlinear estimation by Iterative Least Square) in the case of PCA. NIPALS allows performing a PCA (principal component analysis) with missing values without deleting individuals with missing data or estimating the missing data. The algorithm is iterative and based on successive linear regressions (Tenenhaus 1998). A general presentation of NIPALS algorithm is given in Wold et al (1987). The algorithm of NIPALS analysis with missing data is defined as follow:

Step 1: Normalisation of $\mathbf{Z}$ ($\mathbf{Z}_0$)

Step 2: For h = 1, 2, …, a (with a ≤ N):

Step 2.1: $\mathbf{t}_h = \mathbf{Z}_{h-1}[,1]$

Step 2.2: Repeat until the convergence of $\mathbf{p}_h$

Step 2.2.1: For j = 1, 2, …, N:

$$\mathbf{p}_h[j] = \frac{\sum_{i=1}^{n} {}_{\{\text{if } \mathbf{Z}[i,j] \text{ and } \mathbf{t}_h[i] \text{ exist}\}} \mathbf{Z}_{h-1}[i,j]\mathbf{t}_h[i]}{\sum_{i=1}^{n} {}_{\{\text{if } \mathbf{Z}[i,j] \text{ and } \mathbf{t}_h[i] \text{ exist}\}} \mathbf{t}_h[i]^2}$$

Step 2.2.2: Normalise $\mathbf{p}_h$

Step 2.2.3: For i = 1, 2, …, n:

$$\mathbf{t}_h[i] = \frac{\sum_{j=1}^{N} {}_{\{\text{if } \mathbf{Z}[i,j] \text{ exists}\}} \mathbf{Z}_{h-1}[i,j]\mathbf{p}_h[j]}{\sum_{j=1}^{N} {}_{\{\text{if } \mathbf{Z}[i,j] \text{ exists}\}} \mathbf{p}_h[j]^2}$$

Step 2.3: $\mathbf{Z}_h = \mathbf{Z}_{h-1} - \mathbf{t}_h\mathbf{p}_h'$

So, NIPALS enables to perform PCA with missing values without estimating or deleting empty records. As for classical PCA, NIPALS allows to compute row ($\mathbf{t}_h$) and columns ($\mathbf{p}_h$) coordinates as well as eigenvalues for the h-th axis:

$$\lambda_h = \frac{1}{n-1}\mathbf{t}_h'\mathbf{t}_h$$

Moreover, missing values can be estimated using classical reconstitution formulas at the h-th order (Good 1969):

$$\hat{\mathbf{Z}}_0[i,j] = \sum_{l=1}^{h} \mathbf{p}_l[j]\mathbf{t}_l[i]$$

PCA of table $\mathbf{Z}$ with NIPALS algorithm aimed to find a row score, which is a compromise of the different spatial patterns observed for all variables (i.e. years).

5. Application: spatio-temporal variation of roe-deer weights

This study was carried out in the 2614 ha fenced Chizé reserve situated in western France (46°05'N, 0°25'W). The roe deer population has been intensively monitored by capture-mark-recapture methods since 1978 (Gaillard et al 1993). Ten days of capture in January and February allow 150-350 roe deer to be caught each year. In a particular capture session, >100 people are involved and drive animals into 2-5 km of nets, which enclose some given forestry plots. Most animals are released with individual collars and the remainder is exported.

Fawns were captured between January and February and weighed using an electronic balance. The site of capture and the sex were noted. All information was transferred into a GIS. Sampling areas vary for one year to one other (figure 1). Male fawns are slightly heavier than

female fawns (Gaillard et al 1996), so adjusted weights for males were computed with an ANOVA in order to include in the analysis all individuals that have been captured.
The reference layer was chosen as a grid of 58 quadrats, with length of side was 800 m (figure 4). We choose this size to be consistent with the scale of the data because the area of a quadrat (0.64 km^2) corresponds roughly to the average sampling area (0.654 km^2). GIS was used to compute the table of weight means crossing the 58 quadrats and the 6 years. In this table, there are more than 6 % of missing values. The year 1998 was the poorest sampled area (10 % of missing values) and 3 quadrats contained only values for 4 years out of the 6. Convergence was obtained in NIPALS analysis and results in a one-axis structure (figure 4a). In order to make the maps more legible, quadrats scores were assigned to their centroids and surface was computed (figure 4c) by two-dimensional weighted local regression (Cleveland 1979, Cleveland and Devlin 1988). All years are positively correlated with axis 1 which indicated that for these years heavy roe deer are found in the Northeast part of the reserve whereas light roe deer are in the South (figure 4b). Reconstitution formulas based on the first axis of NIPALS analysis have been used to obtain estimation of weight distribution for all years for the quadrats. We used GIS to compute averages of predicted weights for each sampled area. For this purpose, we used predictions for quadrats and computation was made in taking into account the area of intersection between quadrat and sampled area. The results are satisfying if we consider that reconstitution of the data have been made in using only one axis (figure 5). The total sum of squares of differences between observed data and estimation is 222.57 (table 1).

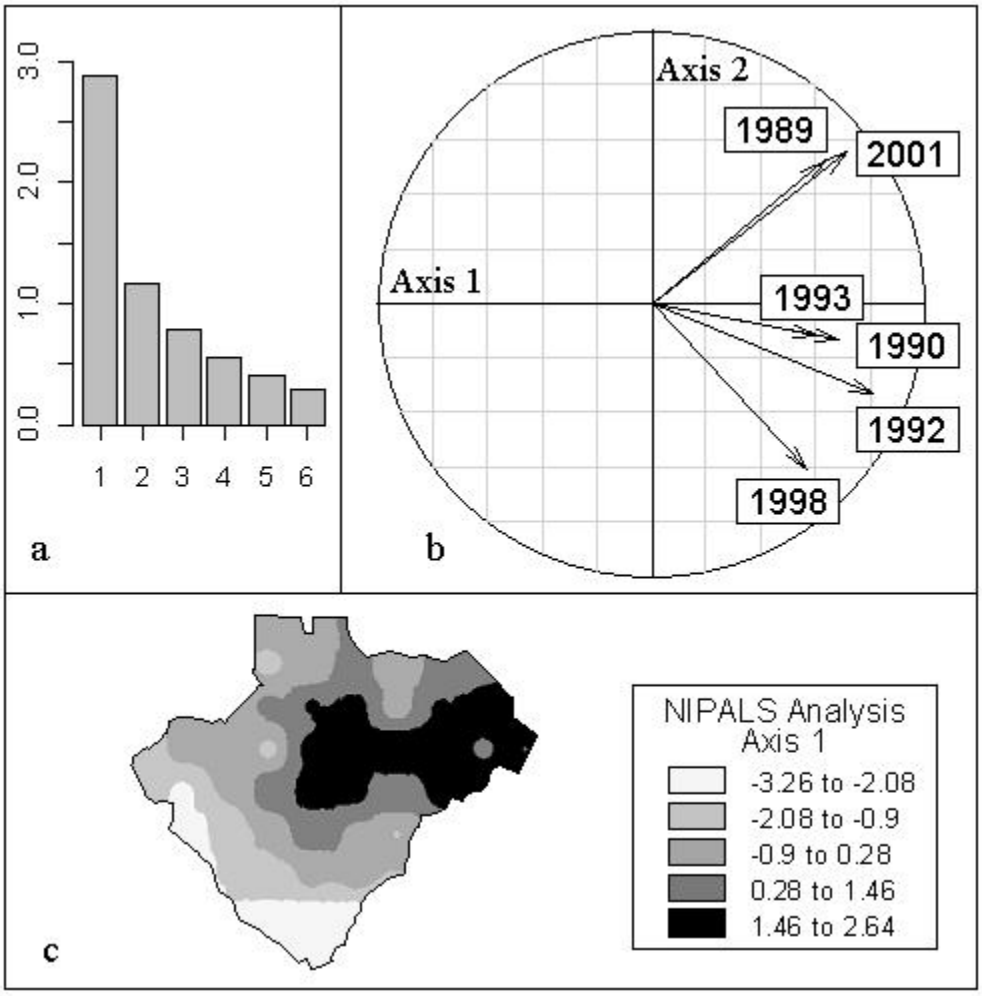


Figure 4: NIPALS analysis of roe deer weights data. (a) Eigenvalues. (b) Correlation circle. (c) Spatial mapping of factorial scores for the first axis.

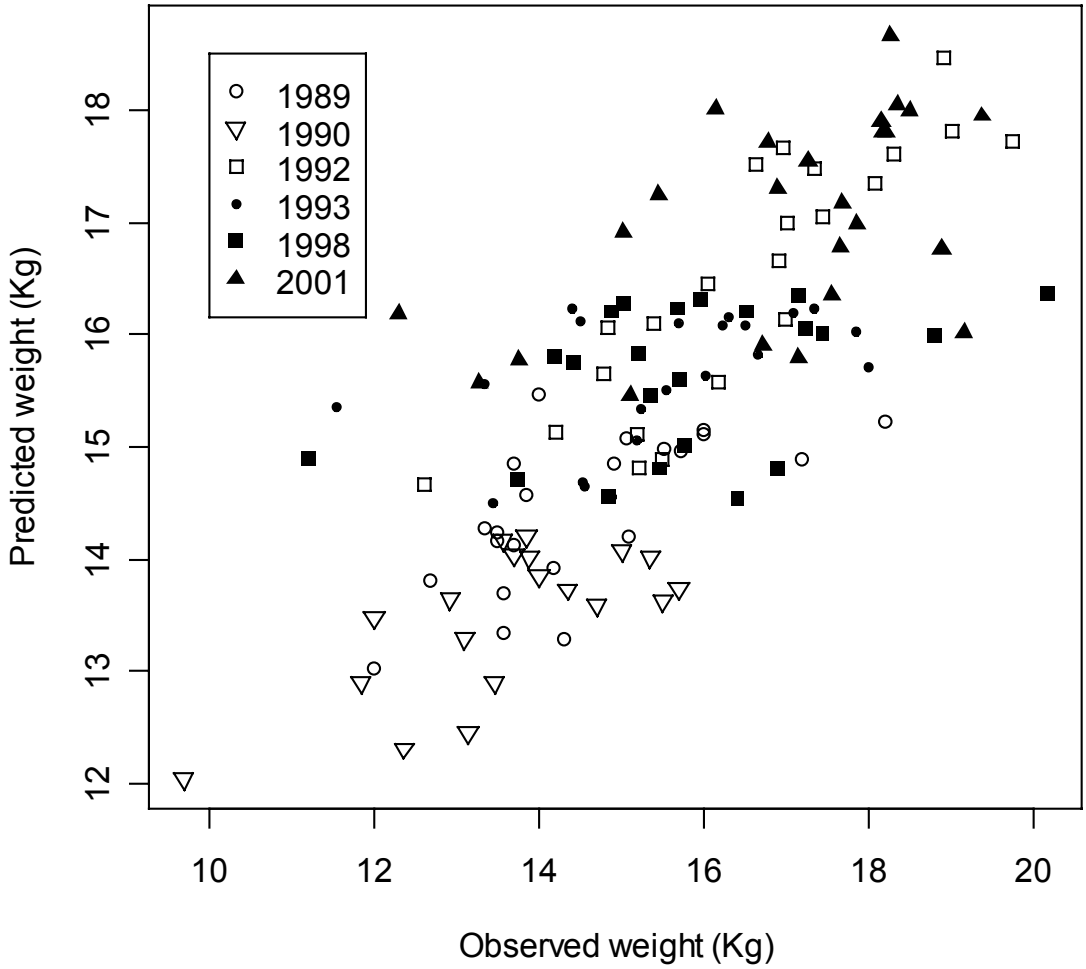


Figure 5: Estimations and observed roe deer weights. Estimations of weights are computed for each quadrat using reconstitution formulas after NIPALS analysis. Then, these estimations are used to compute the averages, weighted by area of intersection, for each sampled area.

Table 1: NIPALS analysis of roe-deer weights data with quadrats of varying size. The total number of quadrats and the percentage of missing values is given for each quadrat size. Moreover, the first eigenvalue and the number of iterations from NIPALS analysis is noted. The error sum of squares (ESSQ) between observed data and estimation computed using first axis of NIPALS analysis and GIS are given.

Quadrat size (m)	100	200	400	600	800	1000	1200
Number of quadrats	2819	739	203	99	58	41	30
% of missing values	33.1	25.4	16.2	10.6	6	4.4	4.4
Number of iterations	20	17	15	13	10	12	12
First eigenvalue	3.41	3.19	3.03	2.85	2.88	2.86	2.73
ESSQ	175.37	181.35	195.45	213.99	222.57	233.14	244.60

Variations of quadrat size (table 1) influence NIPALS analysis. With the decrease of the size, the number of quadrats and the percentage of missing values increase. If the number of missing values increases, then the convergence is attained with difficulty and the number of iterations increases. But the use of small quadrats induced a finer-scale study and local patterns of variations can be detected. So, the first eigenvalue increases and estimation are better adjusted with the detection of local patterns. In our example, the influence of the quadrat size on the results is minor because the structure observed in the roe-deer weights distribution is simple and strong and so is detected easily at each spatial scale.

6. Discussion and Conclusions

Our approach requires that the user specify a reference layer of spatial units. Theses spatial units are the statistical individuals in the NIPALS analysis. The choice of this layer can be induced by the data but in most cases the user must create this space partitioning and so the simplest way is to create a grid of quadrats. The choice of the size of quadrats has to be consistent with the spatial scale of the study. In our case, the area of quadrat is roughly the area of sampled plots. Moreover, the size of quadrats influences the number of missing values in the new data table. The smaller the quadrats are, the larger the number of missing values are, because the number of intersections will decrease. If there are too many missing values, convergence will not be attained. In the other way, the use of large quadrats will decrease the efficiency of the method to detect local structure and thereby the quality of the estimation. But it is obvious that the choice of the size has to be decided by considering the data a priori. In our case, sampling locations are polygons and so it is easy to establish the neighbourhood relationships in considering intersection across polygons. If the sampling locations are points, the use of buffer zone or more sophisticated methods such as tessellation (Green and Sibson 1978) can be used to assign a polygon to each sampling location and define neighbours.

Estimations of the data by reconstitution formulas are not very satisfactory but this fact is not surprising. Indeed, estimations have been made using only one axis. The first aim of our method is to identify the most important structures in the data and not to predict data for new locations. This approach enables to obtain information on spatial patterns for the whole region study from partial spatial data. Estimation of the data is made on the basis of the structures identified by the first axes of the analysis. So, we estimate a year in taking into account the global structure of all years and considering only the common spatial variability among years. Predicting data for one year for the whole study region is not the major goal of our approach. Spatial techniques for interpolation such as trend surface analysis (Gittins 1968), kriging or cokriging (Bailey and Gatrell 1995) would be probably more efficient for this task but they require that the sites should be fairly evenly distributed in the space and considered as points (i.e. centroid of sampled area) which is not the case for the data we consider here. Areal interpolation methods (Flowerdew and Green 1992, Lam 1983, Xie 1995) are more appropriate for this task because they consider the spatial characteristics of the areal units. If the first part of our approach (spatial linkage and construction of the data table) can be seen as an areal interpolation technique, it does not require a complete overlapping of the different sampling plans like most of the classical areal interpolation methods. Moreover, our approach do not require any assumption on the distribution of the variables. Furthermore, NIPALS analysis gives more results than a simple interpolation approach.

In this study, GIS appears as the central part of a multidisciplinary problematic. GIS has been used to capture, manage and display the data. The representation of the data in the GIS has motivated new biological questions. The joint use of GIS and statistical analyses results in the elaboration of new methodology that allows resolving biological problems. So the integration of GIS allows us to improve the statistical methodology as well as the biological knowledge. GIS users get to know spatial analysis tools from geostatistics. Problems related to the integration of geostatistics in GIS softwares have been discussed since a long time (Anselin and Getis 1992, Goodchild et al 1992). These reflections have led to a lot of packages making easier the interface between geostatistics and GIS (Bivand 2001). GIS and geostatistics are now considered rightly as essential partners for spatial analysis (Burrough 2001). In the same way, we think that the analysis of spatial data would probably benefit for the improvement of the links between GIS and multivariate analyses.

Acknowledgments
We thank the Office National de la Chasse for organizing all the captures of roe deer at the Chizé reserve. We are grateful to all the students, field assistants, and volunteers that spent time catching and monitoring the roe deer fawns on the study site. Special thanks to Roger Bivand and Jean-Michel Gaillard for ideas, comments and suggestions on previous drafts of this work.

References

Anselin L and Getis A 1992 Spatial statistical analysis and geographic information systems. *Annals of Regional Science* **26**: 19-33
Bailey T C and Gatrell A C 1995 *Interactive spatial data analysis*. Harlow, Longman
Bivand R 2001 More on spatial data analysis. *R News* **1**: 13-17
Burrough P A 2001 GIS and geostatistics: Essential partners for spatial analysis. *Environmental and Ecological Statistics* **8**: 361-377
Cleveland W S 1979 Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* **74**: 829-836
Cleveland W S and Devlin S J 1988 Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association* **83**: 596-610
Flowerdew R and Green M 1992 Developments in areal interpolation methods and GIS. *Annals of Regional Science* **26**: 67-78
Gaillard J M, Delorme D, Boutin J M, Van Laere G and Boisaubert B 1996 Body mass of roe deer fawns during winter in 2 contrasting populations. *Journal of Wildlife Management* **60**: 29-36
Gaillard J M, Delorme D, Boutin J M, Van Laere G, Boisaubert B and Pradel R 1993 Roe deer survival patterns: a comparative analysis of contrasting populations. *Journal of Animal Ecology* **62**: 778-791
Gittins R 1968 Trend-surface analysis of ecological data. *Journal of Ecology* **56**: 845-869
Good I J 1969 Some applications of the singular decomposition of a matrix. *Technometrics* **11**: 823-831
Goodall D W 1954 Objective methods for the classification of vegetation III. An essay on the use of factor analysis. *Australian Journal of Botany* **2**: 304-324
Goodchild M, Haining R and Wise S 1992 Integrating GIS and spatial data analysis: problems and possibilities. *International Journal of Geographical Information Systems* **6**: 407-423
Green P and Sibson R 1978 Computing Dirichlet tessellations in the plane. *The computer journal* **21**: 168-173
Greenacre M J 1984 *Theory and Applications of Correspondence Analysis*. London, Academic Press
Guisan A, Weiss S B and Weiss A D 1999 GLM versus CCA spatial modeling of plant species distribution. *Plant Ecology* **143**: 107-122
Hotelling H 1933 Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* **24**: 417-441
Kadmon R and Danin A 1997 Floristic variation in Israel: a GIS analysis. *Flora* **192**: 341-345
Lam N S 1983 Spatial interpolation methods: a review. *The American Cartographer* **10**: 129-149
Tenenhaus M 1998 *La régression PLS. Théorie et Pratique*. Paris, Editions Technip
Thioulouse J, Chessel D and Champely S 1995 Multivariate analysis of spatial patterns: a unified approach to local and global structures. *Environmental and Ecological Statistics* **2**: 1-14

Wold H 1966 Estimation of principal components and related models by iterative least squares. In Krishnaiah P R (eds) *Multivariate Analysis*. New-York, Academic Press: 391-420
Wold H, Esbensen K and Geladi P 1987 Principal component analysis. *Chemometrics and intelligent laboratory systems* **2**: 37-52
Xie Y 1995 The overlaid network algorithms for areal interpolation problem. *Computers, Environment and Urban Systems* **19**: 287-306
Zhang C S and Selinus O 1998 Statistics and GIS in environmental geochemistry - some problems and solutions. *Journal of Geochemical Exploration* **64**: 339-354

Chapitre VIII : Couplage

Trois publications composent ce chapitre et décrivent des méthodes permettant d'étudier la structure commune induite par deux (ou plusieurs) jeux de données. Le premier article, soumis à *Ecology*, présente l'intérêt du critère de co-inertie pour étudier des co-structures. Le couplage de deux tableaux par l'analyse de co-inertie et ses relations avec d'autres méthodes de couplage est bien évidemment évoqué. L'analyse de co-inertie est plus robuste que l'analyse canonique des correspondances et que l'analyse canonique des corrélations. De plus, cette méthode est très souple et autorise divers types de variables (qualitatives, quantitatives ou floues), ainsi que de nombreuses possibilités de pondération, de centrage et de standardisation des données. Nous insistons également sur la généralité de cette approche qui a déjà été étendue à de nombreuses situations concernant plus de deux tableaux (Chessel & Hanafi 1996, Dolédec *et al.* 1996, Lafosse & Hanafi 1997, Simier *et al.* 1999). Une illustration écologique est proposée.

Deux méthodes complètement originales sont également présentées. La première (analyse de co-inertie procustéenne) permet d'étudier la co-structure d'un couple de tableaux. Il s'agit d'un compromis entre analyse de co-inertie classique et rotation procustéenne. L'analyse procustéenne est une méthode permettant d'ajuster par rotation, translation, et homothétie, un nuage de points sur une configuration cible de points. Les résultats d'une telle analyse se résument, la plupart du temps, à une mesure numérique de l'adéquation entre les deux nuages et à une représentation graphique des deux nuages de points lorsque les données sont contenues dans un plan (deux variables). Nous fournissons une solution pour représenter graphiquement les résultats d'une rotation procustéenne lorsque le nombre de variables de chaque jeu de données est supérieur à 2. Le couplage de deux matrices de distances est également envisageable par cette méthode. Enfin, un test de permutation est proposé comme alternative du test de Mantel ou de PROTEST (Jackson 1995). L'analyse de co-inertie procustéenne a été implémentée dans le logiciel R et fait partie de la librairie ade4 (annexe 5).

Le troisième article propose une nouvelle méthode permettant le couplage de deux jeux de données provenant de deux échantillonnages distincts d'un même espace. Les méthodes de couplage sont bien connues et souvent utilisées en écologie. Toutes ces méthodes nécessitent que les variables de deux jeux de données (*e.g.* espèces et variables environnementales) soient enregistrées pour les mêmes individus (sites). Cependant, il peut advenir que l'on souhaite coupler des données provenant de différents systèmes d'échantillonnage d'un même espace et, par conséquent, correspondant à différents sites. Dans ce cas, la solution la plus utilisée consiste à transformer un jeu de données afin de le rendre compatible avec le second qui représente alors le plan d'échantillonnage de référence. La méthode proposée, l'analyse RLQ spatialisée, permet de s'affranchir de cette transformation initiale des données et autorise donc à coupler des données provenant de deux échantillonnages d'un même espace. L'analyse RLQ spatialisée est une extension de l'analyse de co-inertie. Elle est basée sur les principes de la méthode RLQ et sur la théorie des graphes de voisinage. Cette analyse recherche des combinaisons linéaires des variables de chaque jeu de données maximisant la covariance spatiale. Il en résulte une ordination commune des deux tableaux prenant en compte les relations de voisinage. Une application s'intéressant à l'étude des relations entre la végétation au sol et le couvert forestier, dans la réserve de Chizé, est présentée. Dans ce cadre, un script programmé en Avenue est fourni en annexe 6 afin de faciliter la mise en œuvre de la méthode.

Running head: co-inertia analysis

Still putting things in order: co-inertia analysis and the linking of ecological tables.

Stéphane Dray, Daniel Chessel, Jean Thioulouse

UMR CNRS 5558, Laboratoire de Biométrie et Biologie Evolutive, Université Claude Bernard Lyon 1, 69622 Villeurbanne Cedex, France.

Abstract: Co-inertia analysis is a multivariate method for coupling two tables. It is often neglected by ecologists who prefer the widely used canonical correspondence analysis. We present the co-inertia criterion for measuring the adequacy between two data sets. Co-inertia analysis is based on this criterion and appears as a more robust coupling method than canonical correspondence analysis or canonical correlation analysis. Co-inertia analysis is very flexible and allows many possibilities for coupling. Co-inertia analysis is suitable for quantitative and/or qualitative or fuzzy environmental variables. Moreover, various weighting of sites and various standardizations and/or centering of species data are available for this method. Hence, more ecological considerations can be taken into account in the statistical procedures. Moreover, the principle of this method is very general and can be easily extended to the case of distance matrices or to the case of more than two tables. An ecological illustration is proposed.

Keywords: co-inertia, ordination, canonical correspondence analysis, multi-tables analysis, redundancy analysis, canonical correlation analysis

INTRODUCTION

Problems in applied or theoretical ecology often deal with the study of a pair of numerical data tables. Some relate species traits to species composition (Ojeda et al. 1998) or habitat utilization to species traits (Willby et al. 2000), while others link experimental conditions (e.g. geographic locations or date) to species composition (Dolédec and Chessel 1987). One of the major tasks of ecological studies remains to analyze the response of community composition to environmental conditions and often requires the use of multivariate analyses. Gauch summarizes the reasons for this choice in a clear and concise way. "*Community ecology concerns assemblages of plants and animals living together and the environmental and historical factors with which they interact. […] Community data are multivariate because each sample site is described by the abundances of a number of species, because numerous environmental factors affect communities, and so on. […] The application of multivariate analysis to community ecology is natural, routine and fruitful*." (Gauch 1982). Ordination methods allow detecting the underlying data structure but "*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 and Wentworth 1976). Canonical correlation analysis (CANCOR, Hotelling 1936), principal component analysis with instrumental variables (Rao 1964) also named redundancy analysis (RDA), co-inertia analysis (CIA, Dolédec and Chessel 1994) and canonical correspondence analysis (CCA, ter Braak 1986) are some of the available methods for this task.

Various kinds of data such as numbers of individuals, presence-absence, abundance indexes, biomass, etc. can fill the species table. Environmental conditions are recorded at each site by the way of quantitative, qualitative or fuzzy variables (Chevenet et al. 1994). Moreover, ecologists have learnt to juggle with various transformations of environmental data as well as species data (Noy-Meir 1973, Noy-Meir et al. 1975) before performing multivariate analysis. The diversity of biological questions has lead to the study of various living creatures, implying a diversity of data types, a diversity of numerical conditions, and a diversity of statistical approaches. In a given situation, characterized by the properties of collected data and the objectives of the study, the choice of the "good" statistical method is very difficult for ecologists. Unfortunately, this choice is often guided by practical considerations such as the possibilities proposed by a statistical software, and the theoretical considerations concerning the characteristics of the data and the objectives of the study are neglected. Co-inertia analysis, which has been implemented only in the ADE-4 software (Thioulouse et al. 1997), is much less used than CCA (Birks et al. 1996). The various methods available for ecological studies do not provide optimal results for all ecological situations and that is why "*no single method has emerged as a solution to all problems of describing and explaining patterns of compositional variation in natural communities*" (Noy-Meir and Whittaker 1977). Almost 10 years after Palmer's paper in Ecology (Palmer 1993), the controversy about ecological tables analysis methods (Wartenberg et al. 1987, Peet et al. 1988, Jackson and Somers 1991) has nearly settled down. Indeed, RDA and CCA have become the most widely used methods. The success of CCA (more than 800 references versus 80 for CIA) is probably due to the availability of this method in many statistical packages and to the previous success of correspondence analysis (Hill 1974). But this success should not hide the fact that CCA is only devoted to gradient analysis (Palmer 1993) and is not always appropriate for the coupling of two tables. Indeed, CCA is very stringent and requires that the species table is analyzed by correspondence analysis (CA) and that the sites are weighted by their richness. These considerations are not suitable for all situations.

In this paper, we present the principles of CIA, showing the numerous possibilities available for coupling two tables. The co-inertia criterion, that has been often neglected, is presented and appears as a central concept when analyzing a pair of tables. Moreover, we emphasize the generality of the principle of CIA that can be extended to the case of linking more than two tables.

A GLOBAL MEASURE OF CO-STRUCTURE

Different statistics such as Pearson correlation coefficient or covariance can be used to measure the relation between two variables. The purpose of this part is to define a statistic that measures the relation between two (or more) sets of variables. Let **X** be a table containing the values of p environmental variables (columns) measured at n sites (rows). Each site can be represented as a point in an ecological hyperspace with p dimensions where each axis represents an environmental variable. If **D** is the diagonal matrix (n x n) of site weights ($\mathbf{D} = diag(w_1,...,w_n)$, e.g. w_i=1/n) and if **Q** (p x p) is a metric of this hyperspace (e.g. **Q**=*diag*(1,…,1) for the Euclidean metric), then the inertia of the "*cloud of sites*" is simply :

$$I_0 = \sum_{i=1}^{n} w_i \left\|\mathbf{X}_i\right\|_{\mathbf{Q}}^2 = trace\left(\mathbf{XQX}^{\mathrm{T}}\mathbf{D}\right)$$

This total inertia is a global measure of the variability of the data. It is the weighted sum of square distances measured with **Q,** between the points of **X** (n sites) and the origin. The sites $\mathbf{X}_i$ can be projected on a **Q**-normed vector **u** and the projected inertia is express by :

$$I(\mathbf{u}) = \mathbf{u}^{\mathrm{T}}\mathbf{QX}^{\mathrm{T}}\mathbf{DXQu}$$

The total inertia can be easily decomposed on a set of p orthogonal **Q**-normed vectors $\mathbf{u}_k$:

$$I_0 = \sum_{k=1}^{p} I(\mathbf{u}_k) = \sum_{k=1}^{p} \mathbf{u}_k{}^{\mathrm{T}}\mathbf{QX}^{\mathrm{T}}\mathbf{DXQu}_k = \sum_{k=1}^{p} \mathbf{u}_k{}^{\mathrm{T}}\mathbf{QX}^{\mathrm{T}}\mathbf{DXQu}_k = \sum_{k=1}^{p} \left\|\mathbf{XQu}_k\right\|_{\mathbf{D}}^2$$

Let **Y** be the species table, with n rows (sites) and q columns (species). In the species hyperspace, each site $\mathbf{Y}_i$ is represented by a point. If each site has the same weight ($\mathbf{D} = diag(w_1,...,w_n)$) and if metric **R** is used in the species hyperspace, then the inertia is :

$$J_0 = trace\left(\mathbf{YRY}^{\mathrm{T}}\mathbf{D}\right)$$

and can be decomposed as above on a set of vectors $\mathbf{v}_k$.

It is not much more difficult to study the common geometry of the two clouds. Co-inertia is a global measure of the co-structure of sites in the environmental and species hyperspaces : it is high when the two structures vary accordingly (and also when they vary inversely), and low when they vary independently, or when they do not vary. It is defined by :

$$CoI = \sum_{k=1}^{p}\sum_{j=1}^{q}\left(\mathbf{u}_k{}^{\mathrm{T}}\mathbf{QX}^{\mathrm{T}}\mathbf{DYRv}_j\right)^2 = \sum_{k=1}^{p}\sum_{j=1}^{q}\left(\mathbf{X}^{k\mathrm{T}}\mathbf{DY}^j\right)^2 = trace\left(\mathbf{XQX}^{\mathrm{T}}\mathbf{DYRY}^{\mathrm{T}}\mathbf{D}\right)$$

If the clouds are centered, then inertia is a sum of variances and co-inertia is a sum of square covariances.

PRINCIPLES OF CO-INERTIA ANALYSIS

The co-inertia criterion measures the concordance between two data sets, and a multivariate method based on this statistic has been developed. Co-inertia analysis (Dolédec and Chessel 1994) is a symmetric coupling method which provides a decomposition of the co-inertia criterion on a set of orthogonal vectors. It is defined by the analysis of statistical triplet ($\mathbf{Y}^{\mathrm{T}}\mathbf{DX}$, **Q**, **R**). Different types of data lead to different transformations (centering, normalization, …) of **X** and **Y** and to different metrics (weights) **Q** and **R**. Co-inertia analysis aims to find a vector $\mathbf{v}_1$ in the species space and a vector $\mathbf{u}_1$ in the environmental space with

maximal co-inertia. If **X** and **Y** are centered, then co-inertia analysis maximizes the square covariance between the projection of **X** on $\mathbf{u}_1$ and the projection of **Y** on $\mathbf{v}_1$:

$$\mathrm{cov}^2\left(\mathbf{XQu}_1, \mathbf{YRv}_1\right) = \mathrm{corr}^2\left(\mathbf{XQu}_1, \mathbf{YRv}_1\right) * \mathrm{var}(\mathbf{XQu}_1) * \mathrm{var}(\mathbf{YRv}_1)$$

(a) = (b) * (c) * (d)

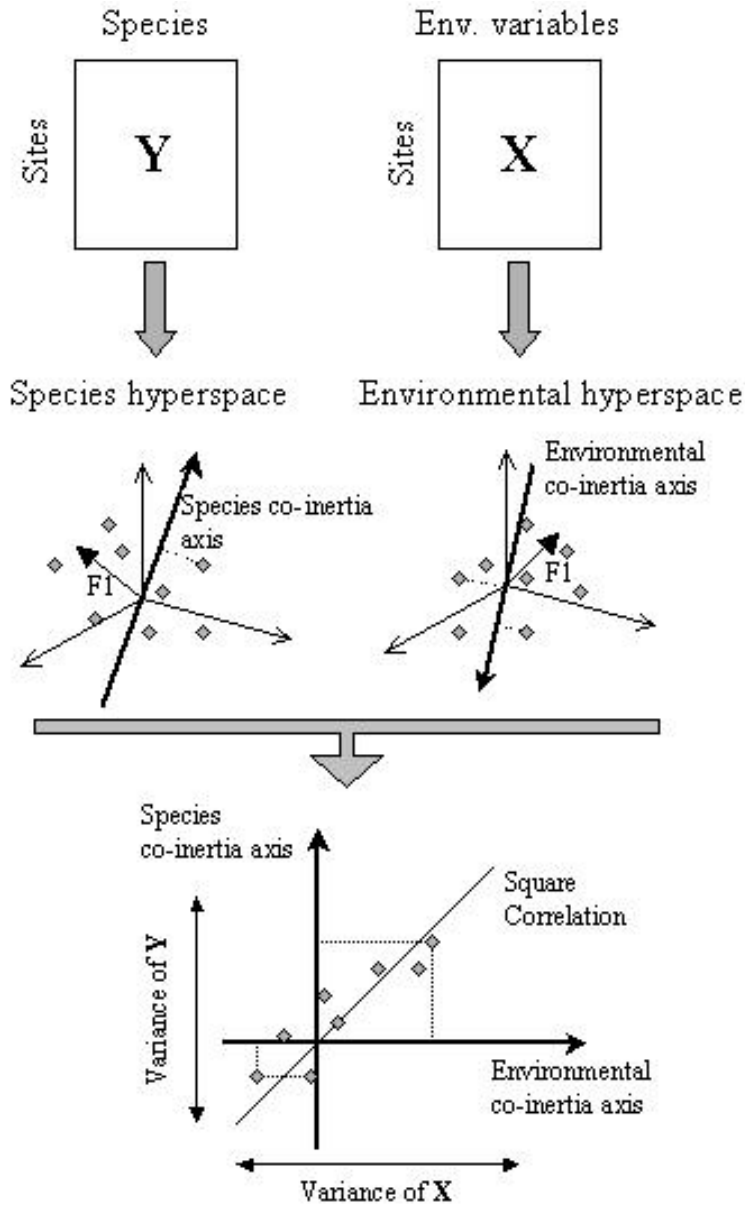


Figure 1: Principles of co-inertia analysis (CIA). The two tables **X** and **Y** produce two representations of the sites in two hyperspaces. Separate analyses find axes maximizing inertia of in each hyperspace (F1). CIA aims to find a couple of co-inertia axes on which the sites are projected. CIA maximizes the square covariance between the projections of the sites on the co-inertia axes.

This square covariance (a) can be easily decomposed, showing that CIA finds a compromise between the correlation (b), the variance of sites in the species viewpoint (d) and the variance of sites in the environmental viewpoint (c) (figure 1). The second and further pairs of vectors ($\mathbf{u}_2$, $\mathbf{v}_2$…) maximize the same quantity but are subject to extra-constraints of orthogonality. CANCOR, CCA and RDA also maximize the square covariance but with additional constraints (figure 2) influencing the robustness of the analysis relative to the

number of variables. CANCOR is defined by the use of two Mahalanobis metrics $\left(\mathbf{X}^t\mathbf{DX}\right)^{-1}$ and $\left(\mathbf{Y}^t\mathbf{DY}\right)^{-1}$, CCA and RDA have only one Mahalanobis metric $\left(\mathbf{X}^t\mathbf{DX}\right)^{-1}$. The Mahalanobis metric takes into account the correlation in the data since it is calculated using the inverse of the variance-covariance matrix (De Maesschalck et al. 2000). The use of this metric adds constraints in the analysis and its calculation implies precautions concerning the dimensions of the tables when using CCA, RDA or CANCOR.

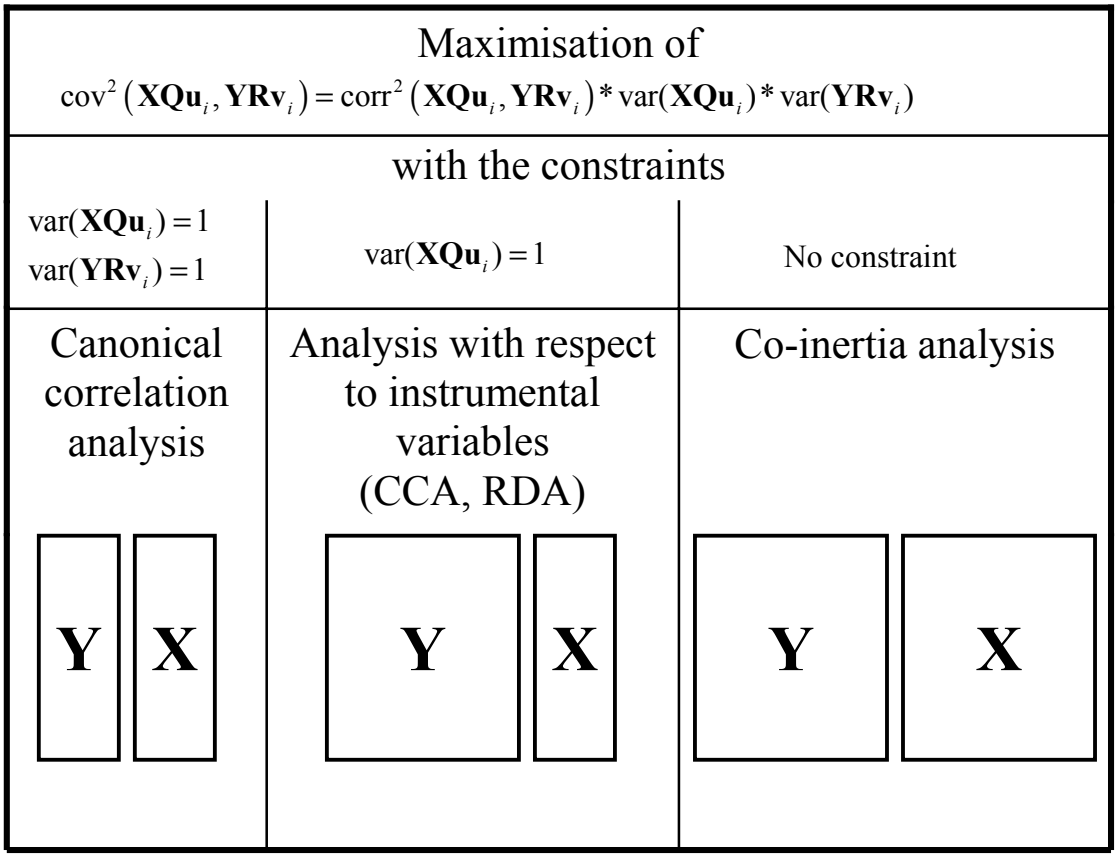


Figure 2: Criteria maximized in canonical correlation analysis, co-inertia analysis and analysis with respect to instrumental variables. Canonical correlation analysis requires few species and few environmental variables compared to site number. Analysis with respect to instrumental variables (e.g. CCA or RDA) requires only few environmental variables, while there is no constraint in the co-inertia approach.

Co-inertia analysis is a general coupling method that maximizes the co-inertia between the variables of two tables. Separate tables **X** and **Y** can be analyzed by various methods, leading to different coupling methods. Hence, Reynaud and Thioulouse (2000) perform a CA on **Y** and a MCA on **X** and link these two analyses through CIA (CA-MCA CIA). Various couplings have been performed in ecological studies, such as PCA-PCA CIA (Cadet and Thioulouse 1998) or CA-PCA CIA (Lods-Crozet et al. 2001) but many other possibilities have not yet been used. The CA-PCA or CA-MCA CIA is very similar to CCA and the two approaches aim to find a site score which is a linear combination of environmental variables maximizing the variance of species centroid (i.e. separation of species niches). The only difference is that CCA has an additional constraint (the total variance must be equal to one), and must be avoided in the case of numerous environmental variables.

CIA is very general and some existing methods appear as special cases of it. Inter-battery analysis (Tucker 1958) is mathematically equivalent to a PCA-PCA CIA. The MCA-MCA CIA is equivalent to the correspondence analysis of the Burt's table crossing two qualitative

tables (Cazes 1980). When table **X** contains qualitative variables and **Y** contains species numbers, it is usual to cross tables **X** and **Y** to obtain a matrix containing the distribution of species among the environmental variables categories. A simple CA of this new table allows to ordinate the species and the environmental classes (analysis of ecological profiles, Romane 1972, Montana and Greig-Smith 1990). Binary discriminant analysis (Strahler 1978) which has been used in ecology (Del Moral 1982, Huang and Del Moral 1988) is mathematically equivalent to Romane's CA. Although this approach allows plotting of species and environmental classes (Ben-Shahar 1987, Ben-Shahar and Skinner 1988), no information about ordination of sites is available. It is easy to demonstrate that CA of the table of ecological profiles is a CA-MCA CIA (Mercier et al. 1992).

THE FLEXIBILITY OF CO-INERTIA ANALYSIS

As seen before, co-inertia analysis allows various coupling of ecological data. The two tables can be analyzed by various analyses (e.g. CA, centered PCA, normed PCA…) with the only constraint that the sites are weighted in the same way for the two separate analyses. Hence, CIA can analyze quantitative, qualitative or even fuzzy environmental variables. For quantitative variables, the environmental table can be analyzed by at least ten different PCA options. In the case of qualitative variables, MCA is applied on the environmental table. Fuzzy correspondence analysis (FCA, Chevenet et al 1994) is suitable for fuzzy variables. Hill & Smith analysis (Hill and Smith 1976) allows to analyze qualitative and quantitative environmental variables simultaneously. All types of environmental variables can be incorporated in CIA and this flexibility is also available for the species data.

The choice of the analysis for species data is decisive for the coupling because different analyses imply different ecological consideration. The first element to take into account is the shape of the species response in relation to environmental variables. In the case of unimodal response, the niche centroid (i.e. average of environmental variables per species) is a good summary of species distribution (ter Braak and Looman 1986). Computations of niche centroids are based on the weighted average principle and are computed in table $\mathbf{Y}^T\mathbf{DX}$ when **Y** has been transformed by $\frac{y_{ij}}{y_{+j}} = p_{i/j}$. This transformation is introduced when **Y** is analyzed by a CA, by a PCA on species profiles, or by a non-symmetric correspondence analysis (Lauro and D'ambra 1984, Gimaret-Carpentier et al. 1998, Kroonenberg and Lombardo 1999).

The use of CA in the instrumental variables approach leads to CCA, and the use of PCA on species profiles in the co-inertia approach leads to OMI analysis (Outlying Mean Index analysis, Dolédec et al. 2000). But other coupling possibilities exist, such as a CA of **Y** in the co-inertia approach. When species response is assumed to be linear, the species-environment relation is well summarized by correlation coefficients. A simple PCA on **Y** can then be applied, and leads to RDA in the instrumental variables approach, and to inter-battery analysis (Tucker 1958) in the co-inertia approach. Table **Y** can also be transformed to a table of site profiles (i.e. $\frac{y_{ij}}{y_{i+}} = p_{j/i}$). Other standardizations are available, such as double standardization by species-and-sites totals ($\frac{y_{ij}}{y_{i+}y_{j+}}$), and standardization by norm or standard deviation (Noy-Meir 1973).

Standardization of data clearly has an influence on the orientation of the study. In some particular situations, standardization is dictated by data but in most cases the decision is made

by ecologists, and based on the study objectives. Standardization by sites implies that the site is the unit of interest and that the study is focused on the species composition of the site. Standardization by species implies that the species is the unit of interest and that the study is focused on the distribution of species over the sites. If data are not standardized then the unit of interest is *the occurrence* of a species in a site. The importance of a site is proportional to its richness, and the importance of a species is proportional to its abundance. Double standardization of correspondence analysis is a compromise between sites-orientated studies and species-orientated studies.

Various centerings can also be applied to table **Y**. Mathematically and geometrically, and also ecologically, centering involves a point of reference for the study. In the case of non-centered data, the point of reference is the all-zero record : an empty site or a species that is always absent. Information is all that deviates from this absolute zero and study is focused on absolute variations and not on the deviations from a simple model.

Centering by species implies that the reference point is an hypothetical site where the species composition is simply the mean species composition computed for all sites. Information is given by a site when it deviates from this hypothetical site and a species is taken into account if it departs from a uniform distribution over all sites. The study is focused on composition differences between sites and species and not on absolute composition.

Centering by site implies that the reference point is an average species for which the abundance in a site is a constant proportion of the species total abundance in this site. Information is given by a species only if its distribution differs from the distribution of the total abundance. A site is informative if its composition deviates from equal proportions of all species.

All these considerations have to be taken into account and the choice must and can only be made by ecologists.

The third element to take into account is the sampling effort, and therefore the significance of a null value in table **Y** (i.e. absence of a species). As Green (1971) pointed out "*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*." When the sampling effort is not constant (e.g. different sampling size, influence of meteorological conditions, influence of the number and the quality of observers…), the absence of a species is ambiguous and variations of abundances among sites have no meaning. Weighting sites with $\frac{y_{i+}}{y_{++}}$ removes the differences of abundances between sites and only the presences of species are taken into account. This option is used in CA, so it must be underlined that CCA removes the abundance effect, and the information given by species absences is not considered. When the aim of the study is, for example, to analyze limiting factors or pollution effects, species absence is an information and CCA should not be used. Assigning equal weights to all sites allows to take into account species absence and thus to reveal a limiting factor. This option is used in classical PCA (linear model) and in species profile PCA (unimodal model) and leads to OMI analysis (Dolédec et al. 2000) in the co-inertia approach.

EXTENSIONS OF CO-INERTIA ANALYSIS

The concept of co-inertia appears as a central part in two-tables coupling methods, and is general enough to be extended to other cases (Esposito Vinzi 2001). If the use of Euclidean or Chi-square distances is not satisfactory, ecological distances computed on species data can be preferred (Legendre and Anderson 1999, Legendre and Gallagher 2001, McArdle and

Anderson 2001). This strategy is also available in CIA (figure 3b) and distance matrices can be analyzed to improve the ecological sense of the study (Dray et al. submitted-a). Moreover, the possibility to extend the co-inertia criterion to the case of more than two tables provides new efficient tools for the ecologist (figure 3). For example, RLQ analysis (figure 3c, Dolédec et al. 1996) is simply the extension of co-inertia analysis to analyze the relationships between species traits and environmental variables through a species by sites table (Legendre et al. 1997). Dray et al. (submitted-b) are adapting RLQ analysis to study the relationships between two data sets arising from different sampling schemes with different sampling locations. If the sites are partitioned, within-class or between-class co-inertia analysis (figure 3f, Franquet and Chessel 1994, Franquet et al. 1995) allows to analyze the species-environment relationships while taking into account this partition. Co-inertia has also been extended to the case of coupling k tables (k>2), under the name of multiple co-inertia analysis (figure 3d, Chessel and Hanafi 1996) to study spatio-temporal variations of species composition. Another extension, concordance analysis (figure 3e, Lafosse and Hanafi 1997) is suitable for coupling k tables with a reference table and can be used to study the relationships between the environment and various groups of species. Lastly, co-inertia is useful in the case of the analysis of a series of k pairs of tables (STATICO, figure 3g, Simier et al. 1999). The STATICO method is particularly well adapted to the study of the modifications of species-environment relationships during several sampling years.

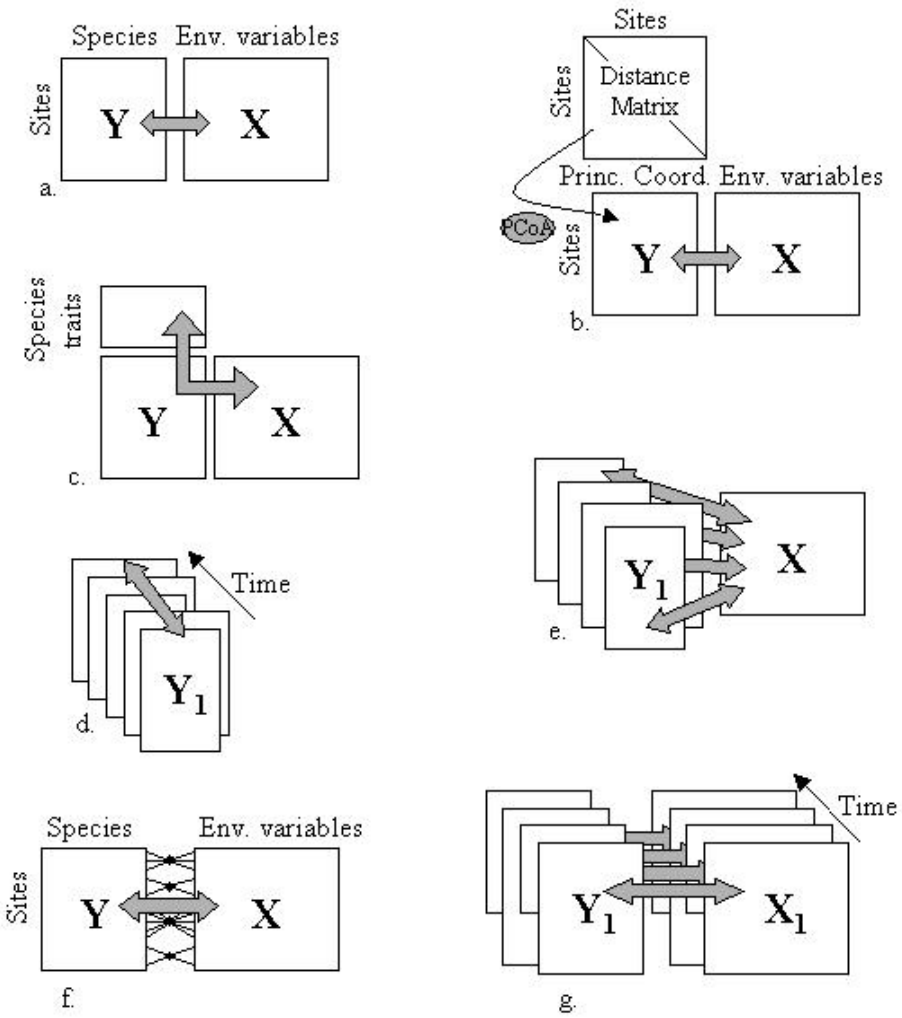


Figure 3: Some extensions based on the co-inertia criterion. **Y** is a sites by species table and **X** is a sites by environmental variables table. (a) classical co-inertia analysis. (b) coupling an environmental table and principal coordinates obtained by the principal coordinate analysis (PCoA) of a distance matrix. (c) RLQ analysis to link environmental variables with species traits through **Y**. (d) Multiple co-inertia analysis to link k tables measuring k sets of variables for the same sites. (e) Concordance analysis to link k tables to a reference table. (f) between-class or within-class co-inertia analysis to study the species-environment relationships according to groups of sites. (g) STATICO to link k couples of tables.

APPLICATION

We propose an ecological example to illustrate the co-inertia approach. The data set concerns the distribution of 91 taxa of macroinvertebrate and the measurements of 9 environmental variables in 16 ponds (Friday 1987). The 91 species are divided in 10 taxonomic groups. Data and taxa names can be found in the original paper. Friday (1987) analyzed these data by simple and multiple linear regressions. Data can be analyzed by a simple coupling analysis between all the species and the environmental variables or by separate coupling analyses for each taxonomic group. Numerical conditions (16 individuals for 91 and 9 variables) imply that methods based on regression (CCA and RDA) are unsuitable and CIA is the only alternative. These data have been previously used to demonstrate the efficiency of the co-inertia approach and its developments: Dolédec and Chessel (1994) performed a simple CIA on these data; Chessel and Hanafi (1996) analyzed the 10 species tables by multiple co-inertia analysis. In this paper, we go further in the co-inertia approach and use concordance analysis (Lafosse and Hanafi 1997) to link the 10 species tables ($\mathbf{Y}_1$, $\mathbf{Y}_2$, …$\mathbf{Y}_{10}$) with environment ($\mathbf{X}$) to identify the common structure of the species – environment relationships and the variations around this structure for the different taxonomic groups. As the aim of the study is the effects of lake acidification, which is a limiting factor for some species, environmental variables are analyzed by a normed PCA and species data by centered by species PCAs. Separate CIAs have been performed and all show that the data are structured by only one axis. According to the results of separate CIAs, concordance analysis was performed between the 10 taxonomic groups tables and the environmental table and the results are presented only for the first axis. When two environmental gradients can be identified in the data, the results of multivariate analyses are often represented on a plane defined by the first two axes of the analysis. In our example, the data correspond to a one-axis structure (lake acidification) and so the results will be presented on one-dimensional graphics corresponding to the first axis of the analysis. Concordance analysis finds a score of sites obtained by linear combination of environmental variables ($\mathbf{Xa}$) and ten sets of sites scores corresponding to each taxonomic group ($\mathbf{Y}_1\mathbf{b}_1$, …, $\mathbf{Y}_{10}\mathbf{b}_{10}$) maximizing the co-inertia defined by $\sum_{i=1}^{10} \mathrm{cov}^2\left(\mathbf{Xa}, \mathbf{Y}_i\mathbf{b}_i\right)$. The vector $\mathbf{a}$ contains the coefficients of environmental variables and the vectors $\mathbf{b}_i$ contains the coefficients of taxa for the group i. The coefficients are contained in vectors and not in matrices because we present the results for the first axis only. So the number of columns of $\mathbf{a}$, $\mathbf{b}_1$, …$\mathbf{b}_{10}$ is equal to one. The first axis of concordance analysis is an acidification factor and provides an ordination of ponds based on environmental variables ($\mathbf{Xa}$), which is linked to variations in species composition (figure 4). Coefficients of environmental variables ($\mathbf{a}$) are represented on figure 4c and this figure shows that pH, Alcalinity and Hardness are the main variables used to construct the lake acidification factor. Representation of the score of ponds based on environmental data shows that basic ponds are often situated in woodland (figure 4b). This confirms the close agreement between land-use and pond water chemistry identified by Friday (1987, table 5). Each taxonomic group provides an ordination of sites ($\mathbf{Y}_1\mathbf{b}_1$, …, $\mathbf{Y}_{10}\mathbf{b}_{10}$), which is linked to the environmental gradients. The concordance between the ordination of the sites in the species viewpoint and in the environmental viewpoint can be measured by the co-inertia criterion (figure 5). The figure 5 shows, for the first axis of concordance analysis, the relationships between the score of ponds obtained by environmental variables (X-axis) and the score of ponds obtained by species data (Y-axis) for each taxonomic group. In order to make the results more legible, we have decided to split the graphic according to the ten taxonomic groups. The tables have been centered and so the co-inertia is simply a sum of square covariances where each square covariance measures the strength of the species-environment relationships for each taxonomic group. Concordance analysis confirmed that

Ephemeroptera taxa are very strongly linked to acidification (highest square covariance) as stated by Friday (1987 p. 97) and that Coleoptera (lowest square covariance) are less sensitive. This indicates the typological value of each group for the study of lake acidification. Hence, sampling can be focused only on groups with high typological value (e.g. Ephemeroptera, Oligochaeta) rather than groups with low value (Coleoptera or Hydracarina) for studying lake acidification. Problems concerning the lack of adequacy, observed by Friday (1987 p. 98), for pond L between chemical characteristics and species composition is also demonstrated by concordance analysis (figure 5): the point corresponding to this pond is often well below the regression line.

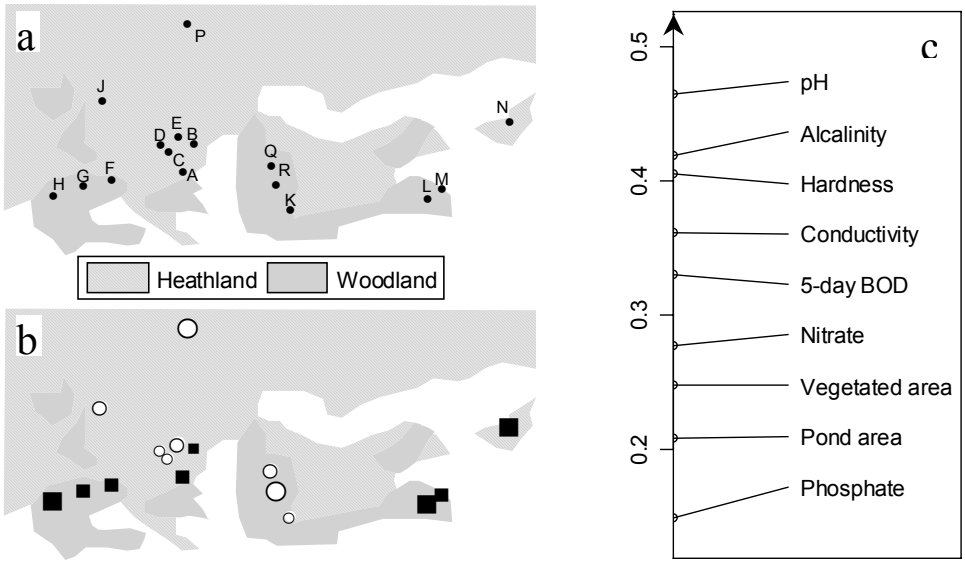


Figure 4: Results of the data of Friday (1987) for the first axis of concordance analysis. (a) localization of the 16 ponds. (b) sites scores from the environmental data (i.e. **Xa**). Black squares indicate positive values, white circles indicate negative values. The size of symbols is proportional to the absolute value of the score. Low values correspond to acidic ponds while high values correspond to basic ponds (high pH). (c) coefficients of environmental variables (i.e. **a**) used to compute the environmental score of sites

Taxas can be represented by their coefficients ($\mathbf{b}_1$, …, $\mathbf{b}_{10}$) used to compute the first axis of concordance analysis on a one-axis graphic. This graphic has also been split by taxonomic group to improve its readability (figure 6). The majority of taxa are in the positive part of the gradient, corresponding to high pH. Very few taxa occur in acid ponds, with a few exceptions: *Hesperocorixa castanea* (A3), *Acamptocladius* (F4), or *Helochares punctatus* (E1). Mollusca, Ephemeroptera and Malacostraca are clearly absent at low pH (Friday 1987 p. 101).

The structure induced by acidification on species composition of ponds is very strong and a simple CIA provides quite similar results than concordance analysis. So, it is legitimate to ask oneself why use a sophisticated method rather than a simple one. Concordance analysis provides a theoretical framework that is really suitable to the structure of the data. The adequacy between the objectives of the study, the structure of the data and the statistical method allows to obtain more results than a classical method that does not take into account the taxonomic information. Co-inertia criterion gives information about the suitability of each taxonomic group to study lake acidification. This information, that is not available in a simple coupling method, is very important for the ecologist and can be used to guide further investigations.

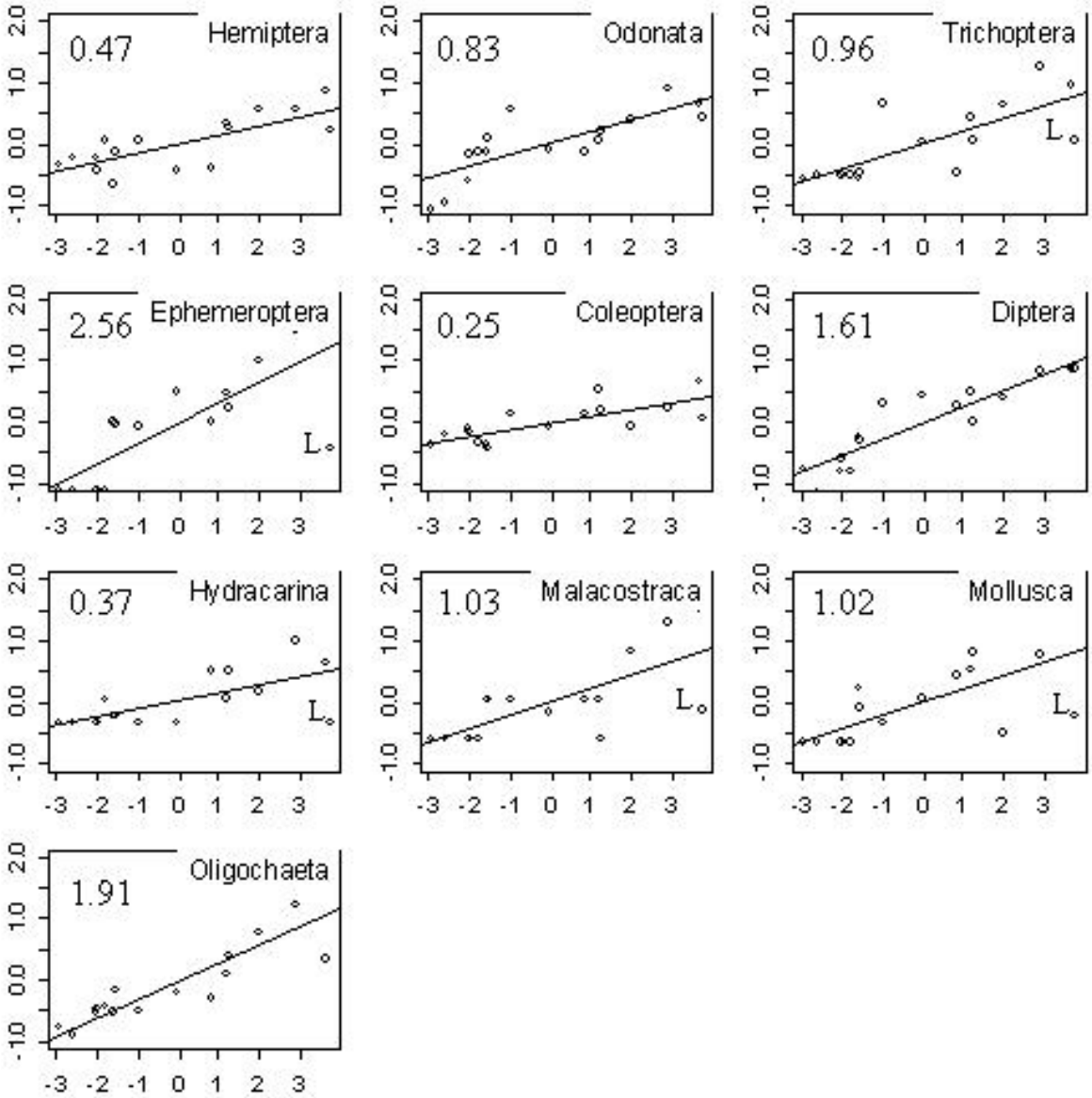


Figure 5: Results of the data of Friday (1987) for the first axis of concordance analysis. Concordance between sites scores obtained by environmental data and species data. Environmental site scores (i.e. **Xa**) are represented on X-axis and species sites scores (i.e. $\mathbf{Y}_i\mathbf{b}_i$) on Y-axis. The co-inertia criterion, which is maximized by the analysis is a sum of square covariances (i.e. $\sum_{i=1}^{10} \mathrm{cov}^2\left(\mathbf{Xa}, \mathbf{Y}_i\mathbf{b}_i\right)$). We decomposed it and indicated in each graph the corresponding value (i.e. $\mathrm{cov}^2\left(\mathbf{Xa}, \mathbf{Y}_i\mathbf{b}_i\right)$). The position of pond L is indicated for some taxonomic groups (L). The regression line is also indicated.

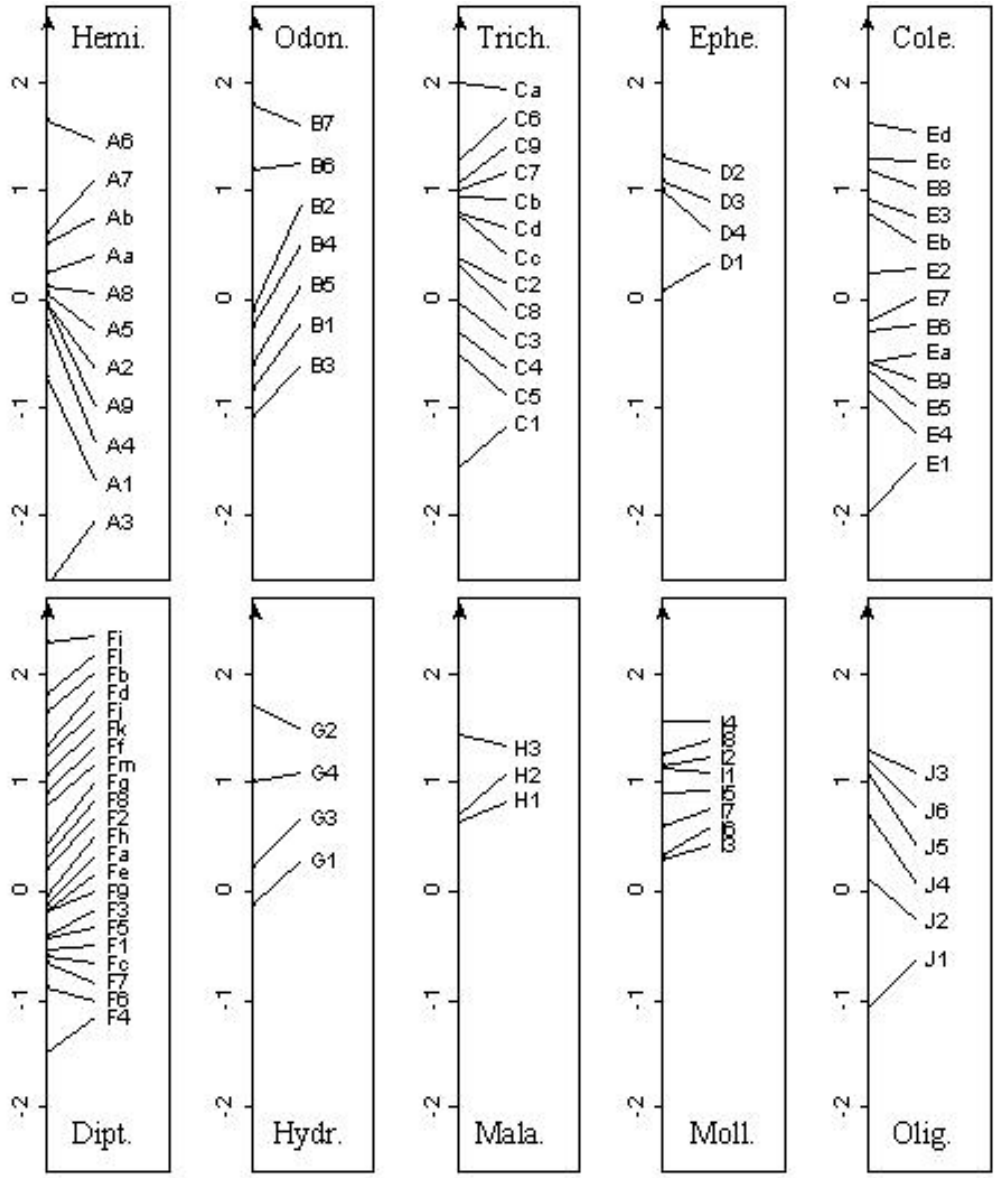


Figure 6: Results of the data of Friday (1987) for the first axis of concordance analysis. Coefficients of taxa used to compute the co-inertia axes for each taxonomic group (i.e. $\mathbf{b}_i$). High values indicate affinities to basic environment and low values correspond to affinities to acidic environment. Species name can be found in the original paper.

DISCUSSION

In this example, we have used one-dimensional graphics instead of the well-known factor maps used in multivariate analysis. This is not particularity linked to CIA. One-dimensional graphics can be used with any multivariate analysis method. They give same information and use less space than factor maps when the number of interesting axes is only one. They can also be used adequately when the number of interesting axes is odd: fro example, to display three axes F1, F2 and F3, on can use one dimensional graphic for F1 and a factor map for F2, F3. Moreover, we have split these graphics according to taxons instead of displaying all the taxa on the same graphic. This technique takes up more space, but gives graphics that are much more clear and easy to interpret. These two graphical techniques (one-dimensional graphic and splitting by categories) can also be used for biplot or triplot.

Co-inertia analysis is the most robust method for the coupling of two tables. Analyses with respect to instrumental variables (e.g. CCA, RDA…) require a small number of environmental variables, and CANCOR requires a small number of species *and* environmental variables compared to the number of sites (figure 2). This comes from the two simultaneous multivariate regressions in CANCOR, and from the multivariate regression in analyses with respect to instrumental variables. Instead, co-inertia analysis is linked to partial least squares regression, a robust alternative to classical regression (Tenenhaus 1998). In the case of qualitative environmental variables, the dimension of the environmental space increases quickly with the number of variables, and co-inertia analysis is often the only alternative.

The previous considerations concerning the number of variables, the species response model, the units of interest, the point of reference of the study and the significance of species absence must be mixed in order to chose the "good" statistical method. The theoretical framework induced by co-inertia will probably introduce new coupling methods. Adding new methods will not make easier the choice for ecologists but we hope that the previous considerations will give practical elements to guide this choice. Experience shows that in many cases, different methods will give similar results, but that in particular situations, the results of a study can greatly depend on the choice of the multivariate method.

Moreover, the generalization of the co-inertia criterion for linking more than two tables provides new efficient tools to study species-environment relationships. The introduction of species traits and the possibility to study simultaneously the spatial and temporal variations would probably greatly improve the quality of results in ecological studies. Even if some answers to these questions are given by classical coupling, the use of more sophisticated methods provides additional results of great interest to ecologists.

All these methods are available in the free ADE-4 software (Thioulouse et al. 1997) which will soon be distributed also under the form of an R package (Ihaka and Gentleman 1996).

ACKNOWLEDGMENTS

We wish to thank P. Legendre, whose suggestions and comments have allowed us to improve the first version of this text.

LITERATURE CITED

Ben-Shahar, R. 1987. Grasses and habitat relationships on a sour bushveld nature reserve. Vegetatio **72**:45-49.

Ben-Shahar, R., and J.-D. Skinner. 1988. Habitat preferences of african ungulates derived by uni-and multivariate analyses. Ecology **69**:1479-1485.

Birks, H. J. B., S. M. Peglar, and H. A. Austin. 1996. An annotated bibliography of canonical correspondence analysis and related constrained ordination methods 1986-1993. Abstracta Botanica **20**:17-36.

Cadet, P., and J. Thioulouse. 1998. Identification of soil factors that relate to plant parasitic nematode communities on tomato and yam in the French West Indies. Applied Soil Ecology **8**:35-49.

Cazes, P. 1980. L'analyse de certains tableaux rectangulaires décomposé en blocs : généralisation des propriétés rencontrées dans l'étude des correspondances multiples. Les cahiers de l'analyse des données **5**:145-161.

Chessel, D., and M. Hanafi. 1996. Analyse de la co-inertie de K nuages de points. Revue de Statistique Appliquée **44**:35-60.

Chevenet, F., S. Dolédec, and D. Chessel. 1994. A fuzzy coding approach for the analysis of long term ecological data. Freshwater Biology **31**:295-309.

De Maesschalck, R., D. Jouan-Rimbaud, and D. L. Massart. 2000. The Mahalanobis distance. Chemometrics and intelligent laboratory systems **50**:1-18.

Del Moral, R. 1982. Control of vegetation on contrasting substrates: herb patterns on serpentinite and sandstone. American Journal of Botany **69**:227-238.

Dolédec, S., and D. Chessel. 1987. Rythmes saisonniers et composantes stationnelles en milieu aquatique I- Description d'un plan d'observations complet par projection de variables. Acta Oecologica - Oecologia Generalis **8**:403-426.

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

Dolédec, S., D. Chessel, and C. Gimaret-Carpentier. 2000. Niche separation in community analysis: a new method. Ecology **81**:2914-2927.

Dolédec, S., D. Chessel, C. J. F. ter Braak, and S. Champely. 1996. Matching species traits to environmental variables: a new three-table ordination method. Environmental and Ecological Statistics **3**:143-166.

Dray, S., D. Chessel, and J. Thioulouse. submitted-a. Procrustean co-inertia analysis for the linking of multivariate data sets. Ecoscience.

Dray, S., N. Pettorelli, and D. Chessel. submitted-b. Matching data sets from two different spatial samplings. Journal of Vegetation Science.

Esposito Vinzi, V. 2001. Explanatory methods for comparative analyses. Chemometrics and intelligent laboratory systems **58**:275-286.

Franquet, E., and D. Chessel. 1994. Approche statistique des composantes spatiales et temporelles de la relation faune-milieu. Comptes Rendus de l'Academie des Sciences Serie III - Sciences de la Vie **317**:202-206.

Franquet, E., S. Dolédec, and D. Chessel. 1995. Using multivariate analyses for separating spatial and temporal effects within species-environment relationships. Hydrobiologia **300/301**:425-431.

Friday, L. E. 1987. The diversity of macroinvertebrate and macrophyte communities in ponds. Freshwater Biology **18**:87-104.

Gauch, H. G. 1982. Multivariate analysis in community ecology. Cambridge University Press, Cambridge.

Gauch, H. G., and T. R. Wentworth. 1976. Canonical correlation analysis as an ordination technique. Vegetatio **33**:17-22.
Gimaret-Carpentier, C., D. Chessel, and J.-P. Pascal. 1998. Non-symmetric correspondence analysis: an alternative for species occurrences data. Plant Ecology **138**:97-112.
Green, R. H. 1971. A multivariate statistical approach to the Hutchinsonian niche: bivalve Molluscs of Central Canada. Ecology **52**:543-556.
Hill, M. O. 1974. Correspondence analysis : A neglected multivariate method. Applied Statistics - Journal of the Royal Statistical Society Series C **23**:340-354.
Hill, M. O., and A. J. E. Smith. 1976. Principal component analysis of taxonomic data with multi-state discrete characters. Taxon **25**:249-255.
Hotelling, H. 1936. Relations between two sets of variates. Biometrika **28**:321-377.
Huang, C. L., and R. Del Moral. 1988. Plant-environment relationships on the Montlake wildlife area, Seattle Washington, USA. Vegetatio **75**:103-113.
Ihaka, R., and R. Gentleman. 1996. R: A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics **5**:299-314.
Jackson, D. A., and K. M. Somers. 1991. Putting things in order: the ups and downs of detrended correspondence analysis. American Naturalist **137**:704-712.
Kroonenberg, P. M., and R. Lombardo. 1999. Nonsymmetric correspondence analysis: a tool for analysing contingency tables with a dependence structure. Multivariate Behavioral Research **34**:367-396.
Lafosse, R., and M. Hanafi. 1997. Concordance d'un tableau avec K tableaux : définition de K+1 uples synthétiques. Revue de Statistique Appliquée **45**:111-126.
Lauro, N., and L. D'ambra. 1984. L'analyse non symétrique des correspondances. Pages 433-446 *in* R. Tomassone, editor. Data Analysis and Informatics III. Elsevier, North-Holland.
Legendre, P., and M. J. Anderson. 1999. Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. Ecological Monographs **69**:1-24.
Legendre, P., and E. D. Gallagher. 2001. Ecologically meaningful transformations for ordination of species data. Oecologia **129**:271-280.
Legendre, P., R. Galzin, and M. L. Harmelin-Vivien. 1997. Relating behavior to habitat: solutions to the fourth-corner problem. Ecology **78**:547-562.
Lods-Crozet, B., E. Castella, D. Cambin, C. Ilg, S. Knispel, and H. Mayor-Simeant. 2001. Macroinvertebrate community structure in relation to environmental variables in a Swiss glacial stream. Freshwater Biology **46**:1641-1661.
McArdle, B. H., and M. J. Anderson. 2001. Fitting multivariate models to community data: a comment on distance-based redundancy analysis. Ecology **82**:290-297.
Mercier, P., D. Chessel, and S. Dolédec. 1992. Complete correspondence analysis of an ecological profile data table: a central ordination. Acta Oecologica - International Journal of Ecology **13**:25-44.
Montana, C., and P. Greig-Smith. 1990. Correspondence analysis of species by environmental variable matrices. Journal of Vegetation Science **1**:453-460.
Noy-Meir, I. 1973. Data transformation in ecological ordination. I. some advantages of non-centring. Journal of Ecology **61**:329-341.
Noy-Meir, I., D. Walker, and W. T. Williams. 1975. Data transformation in ecological ordination. II. On the meaning of data standardization. Journal of Ecology **63**:779-800.
Noy-Meir, I., and R. H. Whittaker. 1977. Continuous multivariate methods in community analysis: some problems and developments. Vegetatio **33**:79-98.

Ojeda, F., J. Arroyo, and T. Marañon. 1998. The phytogeography of European and Mediterranean heath species (Ericoideae, Ericaceae): a quantitative analysis. Journal of Biogeography **25**:165-178.

Palmer, M. W. 1993. Putting things in even better order: the advantages of canonical correspondence analysis. Ecology **74**:2215-2230.

Peet, R. K., R. G. Knox, J. S. Case, and R. B. Allen. 1988. Putting things in order: the advantages of detrended correspondence analysis. American Naturalist **131**:924-934.

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

Reynaud, P. A., and J. Thioulouse. 2000. Identification of birds as biological markers along a neotropical urban-rural gradient (Cayenne, French-Guiana), using co-inertia analysis. Journal of Environmental Management **59**:121-140.

Romane, F. 1972. Utilisation de l'analyse multivariable en phytoécologie. Investigacion Pesquera **36**:131-139.

Simier, M., L. Blanc, F. Pellegrin, and D. Nandris. 1999. Approche simultanée de K couples de tableaux : Application à l'étude des relations pathologie végétale-environment. Revue de Statistique Appliquée **47**:31-46.

Strahler, A. H. 1978. Binary discriminant analysis: a new method for investigating species-environment relationships. Ecology **59**:108-116.

Tenenhaus, M. 1998. La régression PLS. Théorie et Pratique. Editions Technip, Paris.

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

ter Braak, C. J. F., and C. W. N. Looman. 1986. Weighted averaging, logistic regression and the Gaussian response model. Vegetatio **65**:3-11.

Thioulouse, J., D. Chessel, S. Dolédec, and J. M. Olivier. 1997. ADE-4: a multivariate analysis and graphical display software. Statistics and Computing **7**:75-83.

Tucker, L. R. 1958. An inter-battery method of factor analysis. Psychometrika **23**:111-136.

Wartenberg, D., S. Ferson, and F. J. Ohlf. 1987. Putting things in order: a critique of detrended correspondence analysis. American Naturalist **129**:434-448.

Willby, N. G., V. J. Abernethy, and B. O. L. Demars. 2000. Attribute-based classification of European hydrophytes and its relationship to habitat utilization. Freshwater Biology **43**:43-74.

Procrustean co-inertia analysis for the linking of multivariate datasets

Stéphane Dray, Daniel Chessel, Jean Thioulouse

UMR CNRS 5558, Laboratoire de Biométrie et Biologie Evolutive, Université Claude Bernard Lyon 1, 69622 Villeurbanne Cedex, France.

Abstract: Procrustes analysis is a method to fit a set of points to another. These two sets of points are often defined by the measurements of two sets of variables for the same individuals (e.g. measurements of both species abundances and environmental variables at the same sites). We present a solution for graphical representation of the results of procrustes analysis when the number of variables in each dataset exceeds 2. This method is named procrustean co-inertia analysis because it is based on the joint use of procrustes analysis and co-inertia analysis, which is a coupling method finding linear combinations of two sets of variables of maximal covariance. It provides better graphical representation of the concordance between the two datasets than classical co-inertia analysis. Moreover, distance matrices can be introduced in the analysis to improve its ecological meaning. Lastly, a randomization test equivalent to PROTEST is proposed as an alternative to the Mantel test. An ecological example is presented to illustrate the method.

Résumé : L'analyse procrustéenne est une méthode permettant d'ajuster un nuage de points sur un autre. Ces deux nuages de points sont souvent définis par les mesures de deux types de variables sur les mêmes individus (e.g. mesures d'abondances spécifiques et de variables environnementales à des sites). Nous présentons une solution pour la représentation graphique des résultats d'une analyse de procruste quand le nombre de variables dans chacun des deux jeux de données est supérieur à 2. Cette méthode est appelée analyse de co-inertie procrustéenne car elle est basée sur l'utilisation conjointe de l'analyse de procruste et de l'analyse de co-inertie qui est une méthode cherchant des combinaisons linéaires de deux ensemble de variables de covariance maximale. Elle fournit de meilleures représentations graphiques de la concordance des deux jeux de données que l'analyse de co-inertie classique. De plus, des matrices de distances peuvent être introduites dans l'analyse afin d'en améliorer le sens écologique. Enfin, un test de randomisation, équivalent à PROTEST, est proposé comme alternative du test de Mantel. Une illustration écologique est présentée.

Keywords: Procrustes, Co-Inertia, matrix concordance, graphical representation, distance matrices

Introduction

Procrustes was the leader of a band of brigands in Greek mythology. He was in the habit of putting his victims in a bed and to stretch or cut their limbs in such a way that they "fit" in the bed (Digby & Kempton, 1987). By analogy, procrustes analysis is a method based on rotation, reflection, translation, and dilation of a set of points in order to fit it to another fixed set of points (Gower, 1971). Ecologists often deal with the coupling of species data and environmental data measured over several sites but the use of procrustes rotation is unusual for this task and Jackson (1995) claims that "Procrustean methods are used infrequently in ecology. This lack of use likely reflects the previously limited availability of the procedure." Another reason to this disinterest is probably the fact that the procrustes method provides a measurement of the concordance between the two datasets but produces a graphical representation of this concordance only in the case of two variables in each dataset. If there are more than two variables in one dataset, an alternative may be to perform two separate PCAs to summarize the main patterns of variation of each dataset and then to study their concordance with procrustean analysis on the first two principal components (Peres-Neto & Jackson, 2001). Original variables may then be plotted by their correlations with principal components. This approach is interesting but it requires that the main variation of the two datasets are described by neither more nor less than their first two principal components. Graphical representations of species, sites and environmental variables through biplot or triplot (Gabriel, 1971) are actually the main procedure to interpret the results of a multivariate analysis (ter Braak, 1994). Hence, coupling multivariate analyses such as co-inertia analysis (CIA, Dolédec & Chessel, 1994), redundancy analysis (RDA, Rao, 1964) or canonical correspondence analysis (CCA, ter Braak, 1986) which provide graphical representation of the results are preferred. However, procrustes analysis has demonstrated its adequacy for comparing the results from different ordinations methods of the same dataset (Digby & Kempton, 1987; Jackson, 1993) or for studying the concordance of ecological tables (Fasham & Foxton, 1979; Kenkel & Bradfield, 1986; Olden, Jackson & Peres-Neto, 2001; Paszkowski & Tonn, 2000).

In this paper, we present a solution for graphical representation of the results of procrustes analysis. We named this approach procrustean co-inertia analysis (PCIA) because it is based on the principles of procrustes analysis and co-inertia analysis. We present the method and demonstrate its efficiency for the coupling of ecological data and distance matrices. This allows to incorporate more ecological meaning in statistical analyses than classical methods such as canonical correspondence analysis (Legendre & Anderson, 1999; Legendre & Gallagher, 2001). An ecological example is presented.

Procrustes analysis

To illustrate this method, we analyzed data concerning the cephalofacial growth of a monkey (Macaca nemestrina) studied at the ages of 0.9 and 5.77 years using the spatial coordinates of 72 fixed points (Olshan, Siegel & Swindler, 1982). Data are represented in figure 1a-1b. Data concerning the 5.77-year-old monkey have been rotated by 90 degrees for the sake of example.

In this paper, Procrustes analysis is based on a least-square Procrustean rotation and other types of rotation are not considered. Procrustes analysis aims to transform a set of points to fit another set of points. Let **X** (n by p) and **Y** (n by q) be two matrices containing the values of respectively p and q variables for the same n individuals. The two configurations of points are given by matrices **X** and **Y** in an r - dimensional space. If the number of variables is the same in the two tables, then $r=p=q$. If the number of variables of **X** is not equal to the number of variables of **Y** then $r=\max(p, q)$ and columns of zeros are added to the smaller

table to match the size of the larger one. The fit between the two configurations of points is measured by:

$$d^2(\mathbf{X},\mathbf{Y}) = \|\mathbf{X}-\mathbf{Y}\|^2 = \sum_{i=1}^{n}\sum_{j=1}^{r}\left(x_{ij}-y_{ij}\right)^2 \qquad [1]$$

The fit is simply measured by the square Euclidean distances between the rows of **X** and those of **Y** (square Euclidean norm). Centering the tables **X** and **Y** in order to make the centroid of **X** coincide with that of **Y** is important for the rotational process (figure 1c-1d). Taking **X** to be fixed, a rotation **R** is applied to the coordinates of **Y** so that the sum of squared distances:

$$\begin{aligned} m_{\mathbf{YX}}^2 &= d^2\left(\mathbf{X},\widehat{\mathbf{Y}}\right) = \left\|\mathbf{X}-\mathbf{Y}\mathbf{R}^t\right\|^2 = \sum_{i=1}^{n}\sum_{j=1}^{r}\left(x_{ij}-\widehat{y}_{ij}\right)^2 \\ &= trace((\mathbf{X}-\mathbf{Y}\mathbf{R}^t)^t(\mathbf{X}-\mathbf{Y}\mathbf{R}^t)) \qquad [2] \\ &= trace(\mathbf{X}^t\mathbf{X})^t + trace(\mathbf{Y}^t\mathbf{Y}) - 2trace(\mathbf{R}\mathbf{Y}^t\mathbf{X}) \end{aligned}$$

is minimum. **R** (r by r) is an orthogonal matrix satisfying $\mathbf{R}^t\mathbf{R} = \mathbf{I}_r = \mathbf{R}\mathbf{R}^t$. If we consider the singular value decomposition of $\mathbf{Y}^t\mathbf{X} = \mathbf{V}\Theta\mathbf{U}^t$, where Θ is a diagonal matrix with the singular values θ_{ii}, then the solution of procrustes analysis is simply expressed as $\mathbf{R} = \mathbf{U}\mathbf{V}^t$ (figure 1e-1f).

In ecological studies, a preliminary rescaling of **X** and **Y** may be necessary if the two tables contain different types of variables. Each column can be standardized to a variance equal to one if the different variables of one table are in different units (e.g. temperature, slope, …). If all the variables of one table are in the same units (e.g. abundances of species), one can wish to keep the relative variance between columns with a global rescaling. This rescaling can be asymmetric if the modifications are defined by only one set of points (Schönemann & Carroll, 1970). In the rest of the paper we have adopted a symmetric rescaling. For this task, Gower (1971) proposes to use a common scale, transforming the matrices by $\frac{\mathbf{X}}{\sqrt{trace(\mathbf{X}^t\mathbf{X})}}$ and $\frac{\mathbf{Y}}{\sqrt{trace(\mathbf{Y}^t\mathbf{Y})}}$ to have unit sum of squares. This rescaling implies that the procrustes analysis focuses only on variations of shape and removes the variations in size (figure 1g-1h). In order to simplify notations, we name **X** and **Y** the centered and scaled tables in the rest of the paper. $\mathbf{X}_{rot} = \mathbf{X}\mathbf{U}\mathbf{V}^t$ is the configuration of the points from **X** that fits on **Y** and $\mathbf{Y}_{rot} = \mathbf{Y}\mathbf{V}\mathbf{U}^t$ is the configuration of the points from **Y** that fits on **X**. The bi-representation of the points is easy when r=2 because the two sets of points (**X** and $\mathbf{Y}_{rot}$ or **Y** and $\mathbf{X}_{rot}$) are contained in a plane. When r>2, Peres-Neto & Jackson (2001) proposed to first modify the data (the first two principal components are used in the place of the original data) and then perform the procrustean rotation between the first two principal components of each table.

In the case where r>2 and no preliminary analyses are performed, it is necessary to project the rows of **X** and $\mathbf{Y}_{rot}$ or those of **Y** and $\mathbf{X}_{rot}$ on a plane to visualize the fitting between the two datasets. Several possibilities have been proposed (Mouttet, 1981) such as the principal component analysis (PCA) of the bound tables (e.g. $\begin{bmatrix}\mathbf{X}_{rot}\\ \mathbf{Y}\end{bmatrix}$) or the PCA of the average table (e.g. $\frac{1}{2}\left(\mathbf{X}_{rot}+\mathbf{Y}\right)$). The problem of this approach is that the graphical representations obtained by the PCA of $\begin{bmatrix}\mathbf{X}_{rot}\\ \mathbf{Y}\end{bmatrix}$ or $\frac{1}{2}\left(\mathbf{X}_{rot}+\mathbf{Y}\right)$ is different from those

obtained by the PCA of $\begin{bmatrix} \mathbf{Y}_{rot} \\ \mathbf{X} \end{bmatrix}$ or $\frac{1}{2}\left(\mathbf{Y}_{rot} + \mathbf{X}\right)$. All these approaches provide quite similar but different representations. Theoretically, as rotations do not change the shape of the configurations of 2n points (n reference points and n rotated points), the representation in a low dimensional space must be unique for the two analyses (i.e. $\mathbf{X}$ and $\mathbf{Y}_{rot}$ or $\mathbf{Y}$ and $\mathbf{X}_{rot}$).

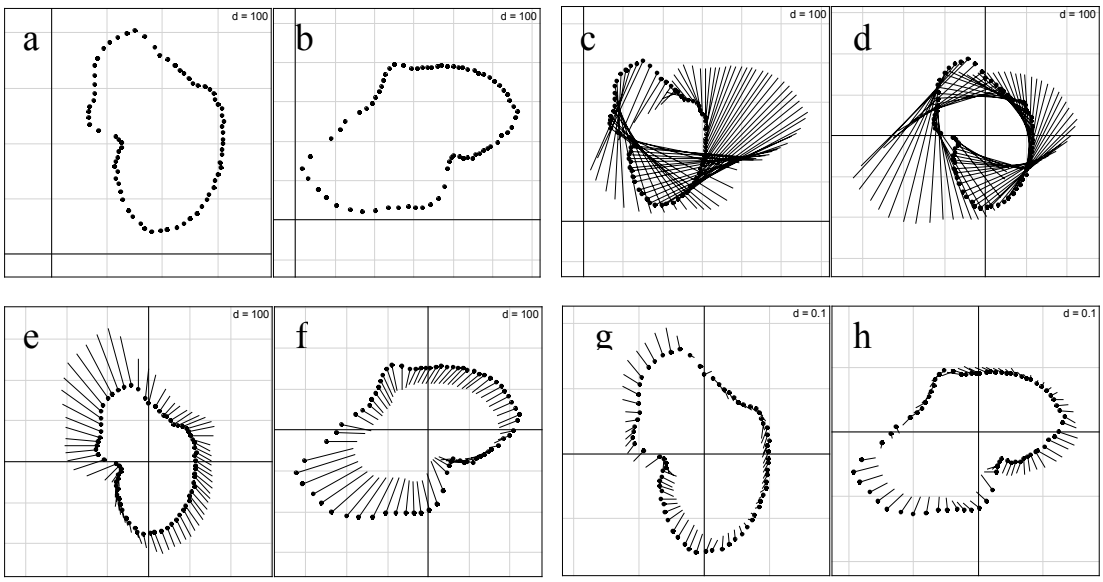


Figure 1: Procrustes analysis of Monkey data. Original data for 0.9-year old (a) and 5.77-years old monkey (b). Graphical representation of the concordance of the two tables for (c) original data and (d) centered data. Lines segments are residuals between scores of 0.9-year old and 5.77-years old monkey. Results from procrustes analysis fitting data of 5.77-years old monkey to data of 0.9-year old monkey (e) and fitting data of 0.9-year old monkey to data of 5.77-year old monkey (f). Results from rescaled procrustes analysis with fitting data of 5.77-years old monkey to data of 0.9-year old monkey (g) and fitting data of 0.9-year old monkey to data of 5.77-year old monkey (h). The value of d indicates the size of squares of the grid.

Co-inertia analysis

Consider now that $\mathbf{X}$ contains the measurements of p environmental variables and $\mathbf{Y}$ the abundances of q species in n sites. There is a cloud of the sites in the species hyperspace and another one in the environmental hyperspace. CIA is based on the diagonalization of $\mathbf{R}^{\frac{1}{2}}\mathbf{Y}^t\mathbf{DXQX}^t\mathbf{DYR}^{\frac{1}{2}}$ where $\mathbf{D}$ is the diagonal matrix of row weights, $\mathbf{Q}$ and $\mathbf{R}$ are respectively the diagonal matrices of column weights of $\mathbf{X}$ and $\mathbf{Y}$. Mathematical details are not described in this paper and can be found in Dolédec & Chessel (1994). The diagonalization of CIA results in a set of site scores in the species hyperspace (i.e. linear combination of species) and in a set of site scores in the environmental hyperspace (i.e. linear combination of environmental variables) with maximal square covariance. This analysis, by maximizing the square covariance, maximizes simultaneously the variance of the sites in the species viewpoint, the variance of the sites in the environmental viewpoint and the square correlation:

$$\mathrm{cov}^2\left(\mathbf{Xu}, \mathbf{Yv}\right) = \mathrm{corr}^2\left(\mathbf{Xu}, \mathbf{Yv}\right) * \mathrm{var}(\mathbf{Xu}) * \mathrm{var}(\mathbf{Yv}) \qquad [3]$$

To represent the concordance between the two datasets, the two sets of site scores can be normalized and their correlations can then be plotted. But this representation is not the best because the scores are normalized after the analysis and so it is the correlations that are plotted, while the analysis maximizes the covariances. Torre & Chessel (1995) present an extension of CIA in the case of fully matched tables (i.e. same individuals and same variables). This analysis can be used for example to study the temporal stability between two

tables containing the measurements of the same variables at the same sites at two dates. For this kind of tables, the sites are represented two times in the same hyperspace and so CIA finds only one co-inertia axis instead of a pair of co-inertia axes (one in each hyperspace) in the usual case. The co-inertia axis maximizes the covariance of the projected coordinates of the two multidimensional clouds in the same hyperspace. If we consider two totally matched tables **A** and **B**, the co-inertia axes are the eigenvectors of $\frac{1}{2}(\mathbf{A}^t\mathbf{B}+\mathbf{B}^t\mathbf{A})$.

Procrustean co-inertia analysis

The CIA of fully matched tables can be applied to **X** and $\mathbf{Y}_{rot}$ because these two datasets are contained in the same hyperspace. It results in finding the eigenvectors of $\frac{1}{2}(\mathbf{Y}_{rot}{}^t\mathbf{X}+\mathbf{X}^t\mathbf{Y}_{rot})=\frac{1}{2}(\mathbf{U}\mathbf{V}^t\mathbf{Y}^t\mathbf{X}+\mathbf{X}^t\mathbf{Y}\mathbf{V}\mathbf{U}^t)=\frac{1}{2}(\mathbf{U}\mathbf{V}^t\mathbf{V}\Theta\mathbf{U}^t+\mathbf{U}\Theta\mathbf{V}^t\mathbf{V}\mathbf{U}^t)=\mathbf{U}\Theta\mathbf{U}^t$. In the same way, if **Y** is fixed, the CIA of totally matched tables **Y** and $\mathbf{X}_{rot}$ finds the eigenvectors of $\mathbf{V}\Theta\mathbf{V}^t$. The environmental rows score data is obtained by projecting the rows of **X** on **U** or those of $\mathbf{X}_{rot}$ on **V** and are contained in $\mathbf{S}_\mathbf{X}=\mathbf{X}_{rot}\mathbf{V}=\mathbf{X}\mathbf{U}\mathbf{V}^t\mathbf{V}=\mathbf{X}\mathbf{U}$ .In the same way, the species rows score is obtained by projecting the rows of **Y** on **V** or those of $\mathbf{Y}_{rot}$ on **U** and are contained in $\mathbf{S}_\mathbf{Y}=\mathbf{Y}_{rot}\mathbf{U}=\mathbf{Y}\mathbf{V}\mathbf{U}^t\mathbf{U}=\mathbf{Y}\mathbf{V}$. Variables can be represented by the coefficients contained in **U** (environmental variables) and **V** (species). Hence, these two analyses (**X** and $\mathbf{Y}_{rot}$ or **Y** and $\mathbf{X}_{rot}$) provide the same representation and the same quality of representation. So, PCIA provides a common solution to the representation in procrustes analysis. PCIA finds co-inertia axes maximizing the covariance between linear combinations of **X** and $\mathbf{Y}_{rot}$ (or **Y** and $\mathbf{X}_{rot}$). The appendix presents a complete example of computation of PCIA.

Randomization test

PROTEST (Jackson, 1995) is a permutation test available for procrustes analysis. If the two tables have been rescaled, $m^2_{\mathbf{YX}}=m^2_{\mathbf{XY}}=m^2=2(1-\sum_{k=1}^{r}\theta_k)$ is a goodness-of-fit statistic (θ_k are the singular values of $\mathbf{Y}^t\mathbf{X}$). PROTEST is based on the statistic $m_{12}=1-\left(\sum_{k=1}^{r}\theta_k\right)^2$ (i.e. based on sum of singular values of $\mathbf{Y}^t\mathbf{X}$) and consists in computing new values of $\underline{m}_{12}$ after permuting entire rows of one table. The observed value is then compared to the set of values obtained by permutation. The hypothesis tested is that there is no link between the two tables against the hypothesis that there is a significant common structure. A permutation test of the same hypothesis is also available in CIA based on the total co-inertia, which is simply, except for a constant, $\alpha^2=\sum_{k=1}^{r}\theta_k{}^2$ (i.e. sum of eigenvalues of $\mathbf{Y}^t\mathbf{X}\mathbf{X}^t\mathbf{Y}$). The co-inertia test is based on $\sum_{k=1}^{r}\mathrm{cov}^2(\mathbf{X}\mathbf{u}_k,\mathbf{Y}\mathbf{v}_k)$, while PROTEST is based on $\sum_{k=1}^{r}\mathrm{cov}(\mathbf{X}\mathbf{u}_k,\mathbf{Y}\mathbf{v}_k)$. The test of CIA is based on the computation of $trace(\mathbf{Y}^t\mathbf{X}\mathbf{X}^t\mathbf{Y})$ and is strictly equivalent to a test based on the RV coefficient (Heo & Gabriel, 1997). PROTEST is based on the residuals of the analysis and so a low value indicates a link whereas the RV-test measures the co-structure of the two matrices and so a high value indicates a link. In the case of two distance matrices, PROTEST is as powerful or more powerful than the usual Mantel test (Peres-Neto & Jackson, 2001). Empirical experiences based on many numerical examples show that the results from PROTEST and RV-test are very similar.

Introduction of ecological distances

The Euclidean distance used in PCA and RDA and the chi-square distance of CA and CCA are not always appropriate for the analysis of ecological data. In the same way, PCIA on original data lacks of ecological consideration for the analysis of the relationships between sites and species. The introduction of distances that are considered better for ecological data, such as Bray-Curtis distance, in ordination methods is then a significant improvement. Distance-based redundancy analysis (Legendre & Anderson, 1999; McArdle & Anderson, 2001) allows to couple a distance matrix based on species data to a table of environmental or experimental data. This approach consists in computing a matrix of distances among sites form table **Y** and the principal coordinates of this distance matrix are computed with principal coordinate analysis (PCoA), possibly including a correction for negative eigenvalues. The data describing the experiment are then coupled to the principal coordinates through RDA. This approach allows introducing ecological distances in ordination methods but the species cannot be plotted on the graphical representation because table **Y** has been substituted by its principal coordinates. Legendre & Gallagher (2001) propose to plot species score through the use of weighted correlation with principal coordinates. Another alternative is available to avoid this problem: table **Y** can be transformed so that the Euclidean distance among sites computed with the transformed table **Y** is ecologically meaningful (Legendre & Gallagher, 2001). For example, if we consider the transformation $y'_{ij} = \frac{y_{ij}}{\sqrt{\sum_{j=1}^{q} y_{ij}^2}}$ then Euclidean distances measured between sites with **Y'** correspond to Chord distances measured with **Y**. Biplot and triplot are then available with ecological distances. However some interesting distances which are not Euclidean based (e.g. Bray-Curtis, Legendre & Legendre, 1998) can not be obtained using a simple transformation of **Y**. One of the advantages of PCIA is that it can also be used to couple transformed data or to couple any distance matrix to a set of environmental variables. Moreover, the robustness of CIA concerning the number of variables relative to the number of individuals allows the coupling of two distance matrices. In this case, the original data are used to compute distances between sites from table **X** and from table **Y**. Two separate PCoAs are then applied to the two distance matrices and two sets of principal coordinates are then obtained. The PCoA of an n by n distance matrix often produces n-1 principal coordinates and so the two tables of principal coordinates have n rows and n-1 columns. The two sets of principal coordinates can be linked by PCIA. Indeed, CIA is very robust concerning the number of variables, as opposed to redundancy analysis that requires few explanatory variables compared to the number of individuals, and is therefore not appropriate in this case. So, PCIA is the only alternative for the coupling of two distance matrices.

Ecological illustration

This example (Verneaux, 1973) is proposed to illustrate the PCIA approach. It concerns the relationships between 11 environmental variables and the distributions of 27 fish species in the Doubs river (France). Environmental variables were measured to describe the morphological aspects of the river (distance to the source, altitude, slope, flow) and the water quality (pH, calcium, phosphate, nitrate, ammonium, oxygen, biological demand in oxygen) for the 29 samples This dataset has been previously analyzed by CCA (Chessel, Lebreton & Yoccoz, 1987). Data have been centered by column and environmental variables have been scaled to unit variance. Barplot of singular values indicates that PCIA based on original data identify a two-axes structure (figure 2a). The first axis is related with the upstream-downstream structure of the river (altitude and slope decreasing, flow increasing with the distance to the source). The second axis is a pollution factor (ammonium, phosphate, nitrate),

which produces a decrease of the species richness in several sites (figure 2b). Species are separated essentially along the upstream-downstream structure (trout, minnow and loach versus the other) and bleak seems to be insensible to pollution (figure 2c). The fit between the two tables is good ($\underline{m}^2$=0.67) and high residuals concern essentially polluted sites. We constructed Bray-Curtis distance matrix from the fish data and linked it to environmental variables. The use of the Bray-Curtis distance (figure 2d-f) improves the fit of the analysis (decrease of the $\underline{m}^2$ statistic from 0.67 to 0.54) and the two axes corresponding to the upstream-downstream structure and to the pollution factor are also identified. The improvement of the fit is especially observed for the polluted sites because the Bray-Curtis distance is more sensible to variations in relative composition and less sensible to variations of absolute abundance than the Euclidean distance. Hence, the effects of pollution on species richness have low importance in this analysis. Note that the use of distance matrix implies that species are replaced by their principal coordinates on the plot (figure 2f). Lastly, we computed the Bray-Curtis distance matrix based on species data and the Euclidean distance matrix computed with environmental data. PCIA was then performed between the two sets of principal coordinates of these two distances matrices. Results are obviously the same as those obtained when coupling environmental variables to the Bray-Curtis distance but environmental variables are replaced by their principal coordinates on the plot (figure 3c). This example is only given to illustrate the use of the different permutation tests because the mantel test can only be applied in the case of two distance matrices. Permutation tests have been performed for this last case (figure 3d-f). The three tests were significant (p-value <0.0001) but the position of the observed values compared to the distribution of the simulated values indicates that the Mantel test seems to have less power than the two others.

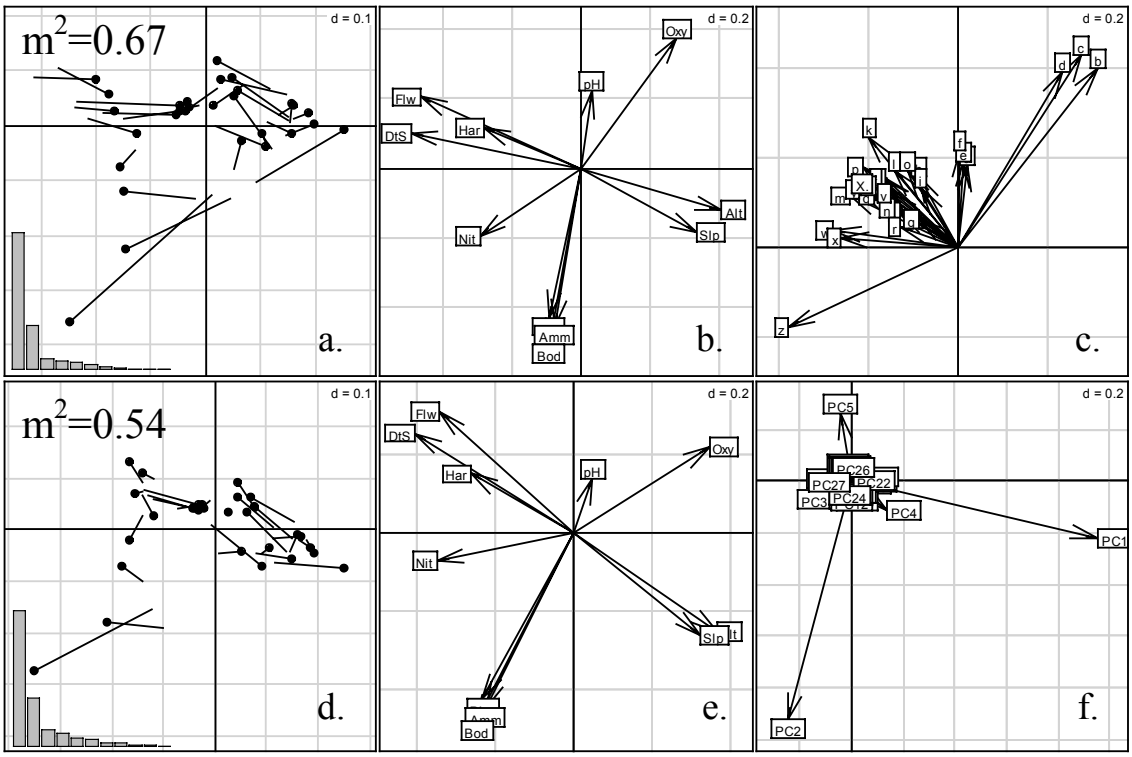


Figure 2: Procrustean co-inertia analysis of fish data. Coupling of original data (a-c) and coupling of Bray-Curtis distance matrix based on species data with environmental variables (d-f). (a, d) concordance between sites. Lines segments are residuals between scores obtained from the environmental matrix and the fish species matrix. Barplot of singular values and m^2 statistic are indicated. (b, e) coefficients of fish species or coefficients of their principal coordinates. (c, f) coefficients of environmental variables. The value of d indicates the size of squares of the grid.

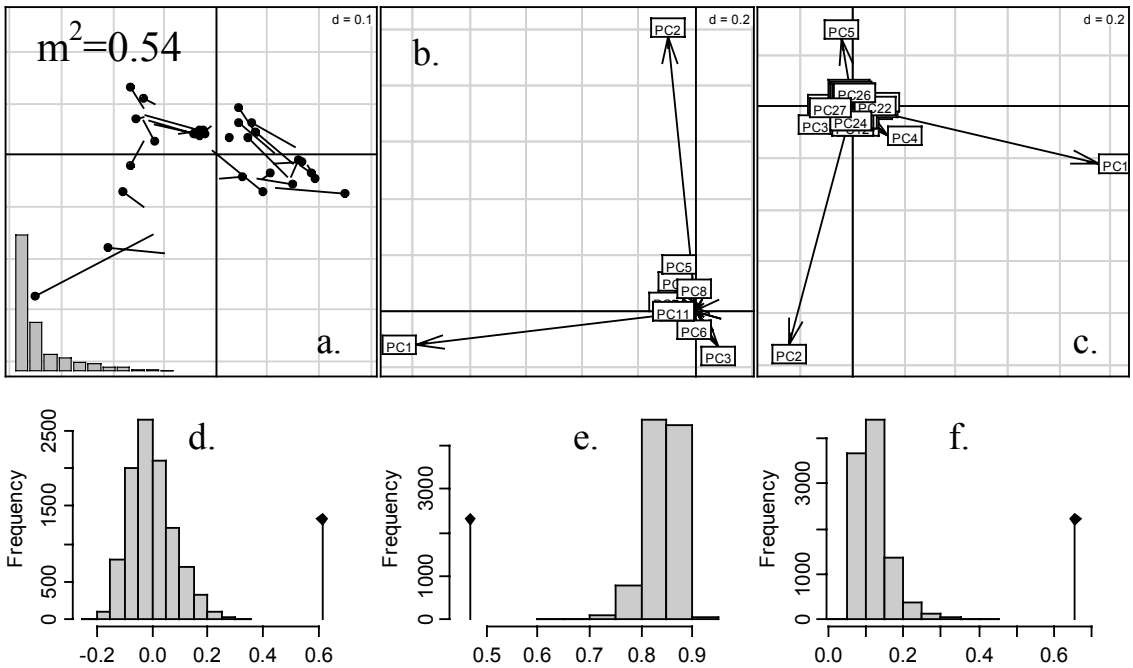


Figure 3: Procrustean co-inertia analysis between two distance matrices (Bray-Curtis distance for species data and Euclidean distance for environmental data). (a) concordance between sites. Lines segments are residuals between scores obtained from the environmental matrix and the fish species matrix. Barplot of singular values and m^2 statistic are indicated. (b) coefficients of principal coordinates of the distance matrix based on species data. (c) coefficients of principal coordinates of the distance matrix based on environmental variables. (d-f) Results of randomization tests (observed value is indicated by the vertical line) with 9999 permutations: (d) mantel test ($p<0.0001$), (e) PROTEST ($p<0.0001$) and (f) RV-test ($p<0.0001$). The value of d indicates the size of squares of the grid.

Conclusion

PCIA demonstrates its efficiency to link ecological data tables. This method accepts various kinds of data such as raw data, modified data or distance matrices (figure 4). This is very promising and allows including more ecological meaning in ordination methods in the same way as the approach introduced by distance-based RDA. PCIA provides a convenient solution to graphical representation of results of procrustes analysis when there are more than two variables in one dataset. Representation of the results obtained from the approach of Peres-Neto & Jackson (2001) are similar to those obtained by PCIA when the structure of each table can be well summarized by the first two principal components. If the structure of one table is more or less complex, our approach is better because it takes into account the global structure of the data and not only the part of the structure that can be summarized on a plane. Moreover, the representation of the concordance between the two datasets in PCIA is better than the representation used in classical CIA because the two systems of sites are in the same hyperspace and no rescaling is needed. Classical CIA is more general and can accept qualitative variables and various weighting of sites and of variables. PCIA is only devoted to the analysis of quantitative variables and cannot include weights but the possibility to weight rows and columns in PCIA is conceivable. In the case of quantitative environmental variables, PCIA appears as a good alternative whereas when table **X** contains the design of an experiment, methods based on RDA (e.g. distance-based RDA) are more suitable because the variables in table **X** are fixed by the experiment and it is then necessary to take into account this dissymmetry. In fact, classical CIA, distance-based RDA and PCIA are three complementary tools. In the context of unimodal response species curves, canonical correspondence analysis has proved its efficiency to separate species niche centroids. PCIA used with the Chi-square distance can also be used in the unimodal context as an alternative to

CCA and avoid the much debated step of CCA consisting in weighting the sites by their species richness (Dolédec, Chessel & Gimaret-Carpentier, 2000). Moreover, CCA is more stringent concerning the number of variables relative to the number of sites. Indeed, CCA is based on multivariate regression and requires a low number of explanatory variables. Lastly, PCIA is more general than CCA and can be used in many other contexts than the study of species-environment relationships. Concerning the randomization procedures, the RV-test of PCIA provides results similar to PROTEST but is less time-consuming because PROTEST performs a singular value decomposition for each step whereas the RV-test is based only on the computation of the trace of a matrix.

Lastly, a weighted version of PCIA can be used to incorporate weights of individuals and variables as in classical CIA. Moreover, CIA has been extended to the case of coupling k tables (k>2) under the name of multiple co-inertia analysis (Chessel & Hanafi, 1996), to the analysis of the concordance of k tables with a reference table (Lafosse & Hanafi, 1997) and to the case of the analysis of k couples of tables (Simier *et al.*, 1999). These different approaches will probably help provide graphical representation for generalized procrustes analysis (Gower, 1975) in the case of more than two tables.

All analyses and graphical representations have been made with ADE-4 software (Thioulouse *et al.*, 1997) freely distributed at http://pbil.univ-lyon1.fr/ADE-4/ADE-4.html. A new version, developed under the form of an R package (Ihaka & Gentleman, 1996) will soon be available.

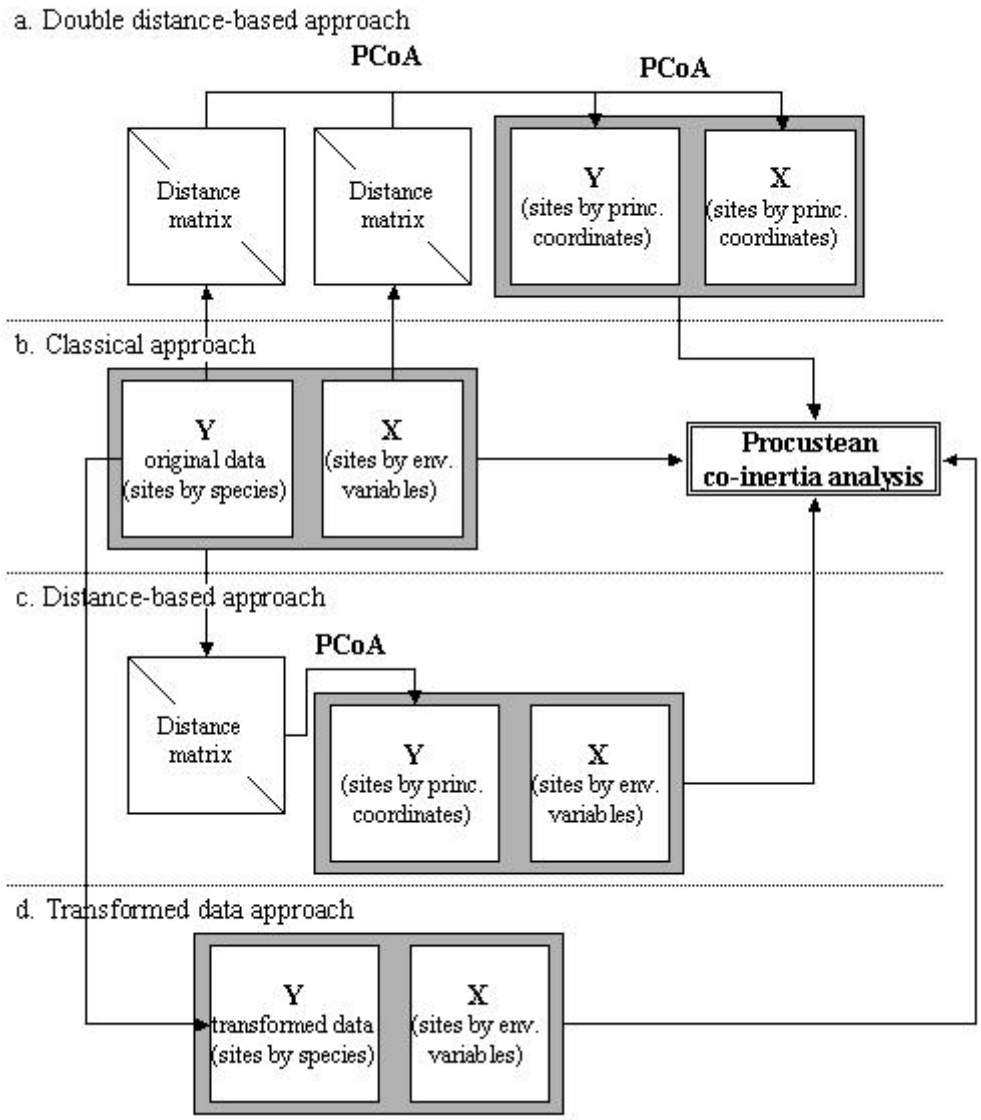


Figure 4: Different approaches available for procrustean co-inertia analysis. (a) Double distance-based approach. The two original tables are used to create two between-sites distance matrices. Principal coordinate analyses are then performed and two new tables (sites by principal coordinates) are created. These two tables can be linked by PCIA. (b) Classical approach. The original data are directly linked by PCIA. (c) Distance-based approach. Species matrix is used to construct a distance matrix. Principal coordinates obtained by PCoA are then coupled to environmental matrix by PCIA. (d) Transformed data approach. Species matrix is modified so that the Euclidean distances measured between sites of the modified table correspond to distances of ecological interest (e.g. Chi2...). The transformed matrix is linked to environmental variables by PCIA.

References

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

Chessel, D. & M. Hanafi, 1996. Analyse de la co-inertie de K nuages de points. Revue de Statistique Appliquée, 44: 35-60

Digby, P. G. N. & R. A. Kempton (ed.), 1987. Multivariate Analysis of Ecological Communities. Chapman and Hall, London.

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

Dolédec, S., D. Chessel & C. Gimaret-Carpentier, 2000. Niche separation in community analysis: a new method. Ecology, 81: 2914-2927

Fasham, M. J. R. & P. Foxton, 1979. Zonal distribution of pelagic decapods (Crustacea) in the eastern North Atlantic and its relation to the physical oceanography. Journal of Experimental Marine Biology and Ecology, 37: 225-253

Gabriel, K. R., 1971. The biplot graphic display of matrices with application to principal component analysis. Biometrika, 58: 453-467

Gower, J. C., 1971. Statistical methods of comparing different multivariate analyses of the same data. 138-149 138-149 *in* Tautu, P. (ed.). Mathematics in the archaeological and historical sciences. Edinburgh University Press, Edinburgh

Gower, J. C., 1975. Generalized procrustes analysis. Psychometrika, 40: 33-51

Heo, M. & K. R. Gabriel, 1997. A permutation test of association between configurations by means of the RV coefficient. Communications in Statistics - Simulation and Computation, 27: 843-856

Ihaka, R. & R. Gentleman, 1996. R: A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics, 5: 299-314

Jackson, D. A., 1993. Multivariate analysis of benthic invertebrate communities: the implication of choosing particular data standardizations, measures of association, and ordination methods. Hydrobiologia, 268: 9-26

Jackson, D. A., 1995. PROTEST: A PROcrustean Randomization TEST of community environment concordance. Ecoscience, 2: 297-303

Kenkel, N. C. & C. E. Bradfield, 1986. Epiphytic vegetation on Acer macrophyllum: A multivariate study of species-habitat relationships. Vegetatio, 68: 43-53

Lafosse, R. & M. Hanafi, 1997. Concordance d'un tableau avec K tableaux : définition de K+1 uples synthétiques. Revue de Statistique Appliquée, 45: 111-126

Legendre, P. & L. Legendre (ed.), 1998. Numerical Ecology, 2nd English edition. Elsevier Science, Amsterdam.

Legendre, P. & M. J. Anderson, 1999. Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. Ecological Monographs, 69: 1-24

Legendre, P. & E. D. Gallagher, 2001. Ecologically meaningful transformations for ordination of species data. Oecologia, 129: 271-280

McArdle, B. H. & M. J. Anderson, 2001. Fitting multivariate models to community data: a comment on distance-based redundancy analysis. Ecology, 82: 290-297

Mouttet, F., 1981. Comparaison de tableaux par la méthode Procuste. Université Paris VI, Paris.

Olden, J. D., D. A. Jackson & P. R. Peres-Neto, 2001. Spatial isolation and fish communities in drainage lakes. Oecologia, 127: 572-585

Olshan, A. F., A. F. Siegel & D. R. Swindler, 1982. Robust and least-squares orthogonal mapping: Methods for the study of cephalofacial form and growth. American Journal of Physical Anthropology, 59: 131-137

Paszkowski, C. A. & W. M. Tonn, 2000. Community concordance between the fish and aquatic birds of lakes in northern Alberta, Canada: the relative importance of environmental and biotic factors. Freshwater Biology, 43: 421-437

Peres-Neto, P. R. & D. A. Jackson, 2001. How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test. Oecologia, 129: 169-178

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

Schönemann, P. H. & R. M. Carroll, 1970. Fitting one matrix to another under choice of a central dilation and a rigid motion. Psychometrika, 35: 245-256

Simier, M., L. Blanc, F. Pellegrin & D. Nandris, 1999. Approche simultanée de K couples de tableaux : Application à l'étude des relations pathologie végétale-environment. Revue de Statistique Appliquée, 47: 31-46

ter Braak, C. J. F., 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology, 67: 1167-1179

ter Braak, C. J. F., 1994. Canonical community ordination. Part I: Basic theory and linear methods. Ecoscience, 1: 127-140

Thioulouse, J., D. Chessel, S. Dolédec & J. M. Olivier, 1997. ADE-4: a multivariate analysis and graphical display software. Statistics and Computing, 7: 75-83

Torre, F. & D. Chessel, 1995. Co-structure de deux tableaux totalement appariés. Revue de Statistique Appliquée, 43: 109-121

Verneaux, J., 1973. Cours d'eau de Franche-Comté (Massif du Jura). Recherches écologiques sur le réseau hydrographique du Doubs. Essai de biotypologie. Université de Besançon, Besançon.

Appendix:
Species code for the 27 fish species

Code	English name	Scientific name
a	chub	*Cottus gobio* L.
b	trout	*Salmo trutta fario* L.
c	minnow	*Phoxinus phoxinus* L.
d	loach	*Nemacheilus barbatulus* L.
e	grayling	*Thymallus thymallus* L.
f	soufie	*Telestes soufia agassizi* C.
g	nase	*Chondrostoma nasus* L.
h	toxostome	*Chondrostoma toxostoma* Vallot
i	dace	*Leuciscus leuciscus* L.
j	chub	*Leuciscus cephalus cephalus* L.
k	barbel	*Barbus barbus* L.
l	spirlin	*Spirlinus bipunctatus* Bloch
m	gudgeon	*Gobio gobio* L.
n	pike	*Esox lucius* L.
o	perch	*Perca fluviatilis* L.
p	bitterling	*Rhodeus amarus* Bloch
q	pumpkinseed	*Lepomis gibbosus* L.
r	red-eye rudd	*Scardinius erythrophtalmus* L.
s	carp	*Cyprinus carpio* L.
t	tench	*Tinca tinca* L.
u	bream	*Abramis brama* L.
v	black bullhead	*Ictalurus melas* Rafinesque
w	ruffe	*Acerina cernua* L.
x	roach	*Rutilus rutilus* L.
y	silver bream	*Blicca bjoerkna* L
z	bleak	*Alburnus alburnus* L.
+	eel	*Anguilla anguilla* Shaw

Environmental variables recorded at 29 sites and codes used as labels in the figures

No.	Code	Environmental variable
1-	DtS	Distance to the source (km x 10)
2-	Alt	Altitude (m)
3-	Slp	Slope (‰ x 10)
4-	Flw	Minimum flow (m^3/s x 100)
5-	pH	pH (x 10)
6-	Har	Total hardness (mg/l of calcium)
7-	Pho	Phosphate (mg/l x 100)
8-	Nit	Nitrate (mg/l x 100)
9-	Amm	Ammonium (mg/l x 100)
10-	Oxy	Dissolved oxygen (mg/l x 10)
11-	Bod	5-days Biological Oxygen Demand (mg/l x 10)

Numerical example of PCIA

Consider the two matrices:

$$\mathbf{X} = \begin{bmatrix} -0.02 & 0.24 \\ -0.74 & 1.84 \\ 0.35 & 1.99 \\ -0.27 & 2.21 \end{bmatrix} \text{ and } \mathbf{Y} = \begin{bmatrix} -0.90 & -0.01 & -0.90 \\ 0.09 & 0.23 & 0.12 \\ 1.48 & 0.46 & 1.74 \\ 1.22 & 0.70 & 1.66 \end{bmatrix}$$

The first step consists in centering and normalizing ($\frac{\mathbf{X}}{\sqrt{trace(\mathbf{X}^t\mathbf{X})}}$) the two matrices:

$$\mathbf{X} = \begin{bmatrix} 0.09 & -0.76 \\ -0.33 & 0.15 \\ 0.30 & 0.24 \\ -0.06 & 0.37 \end{bmatrix} \text{ and } \mathbf{Y} = \begin{bmatrix} -0.46 & -0.12 & -0.52 \\ -0.13 & -0.04 & -0.18 \\ 0.34 & 0.04 & 0.37 \\ 0.25 & 0.12 & 0.34 \end{bmatrix}$$

The singular value decomposition $\mathbf{Y}^t\mathbf{X} = \mathbf{V}\Theta\mathbf{U}^t$ is then performed and results in:

$$\Theta = \begin{bmatrix} 0.80 & 0 \\ 0 & 0.02 \end{bmatrix}, \ \mathbf{U} = \begin{bmatrix} -0.17 & -0.99 \\ -0.99 & 0.17 \end{bmatrix} \text{ and } \mathbf{V} = \begin{bmatrix} -0.65 & -0.09 \\ -0.17 & 0.98 \\ -0.74 & -0.15 \end{bmatrix}$$

Rotated matrices are simply computed by $\mathbf{X}_{rot} = \mathbf{XUV}^t$ and $\mathbf{Y}_{rot} = \mathbf{YVU}^t$.

$$\mathbf{X}_{rot} = \begin{bmatrix} -0.46 & -0.34 & -0.51 \\ 0.03 & 0.36 & 0.02 \\ 0.21 & -0.20 & 0.25 \\ 0.22 & 0.18 & 0.24 \end{bmatrix} \text{ and } \mathbf{Y}_{rot} = \begin{bmatrix} -0.12 & -0.70 \\ -0.04 & -0.22 \\ 0.13 & 0.48 \\ 0.03 & 0.44 \end{bmatrix}$$

The concordance between the rows of the two tables can be represented with the two systems of scores $\mathbf{S_X} = \mathbf{XU}$ and $\mathbf{S_Y} = \mathbf{YV}$.

$$\mathbf{S_X} = \begin{bmatrix} 0.74 & -0.21 \\ -0.10 & 0.35 \\ -0.29 & -0.25 \\ -0.35 & 0.12 \end{bmatrix} \text{ and } \mathbf{S_Y} = \begin{bmatrix} 0.71 & 0.00 \\ 0.22 & 0.00 \\ -0.50 & -0.05 \\ -0.44 & 0.04 \end{bmatrix}$$

Variables of **X** and **Y** are represented respectively by the coefficients contained in **U** and **V**.

Matching data sets from two different spatial samplings

Dray, Stéphane[*] ; Pettorelli, Nathalie & Chessel, Daniel

UMR CNRS 5558, Laboratoire de Biométrie et Biologie Evolutive, Université Claude Bernard Lyon 1, 69622 Villeurbanne Cedex, France;
[*] Corresponding author: Fax : +33 (0)4 78 89 27 19; E-mail : dray@biomserv.univ-lyon1.fr

Abstract: Methods for coupling two data sets (species composition and environmental variables for example) are well known and often used in ecology. All these methods require that variables of the two data sets have been recorded at the same sampling stations. But if the two data sets arise from different sampling schemes, sampling locations can be different. In this case, scientists usually transform one data set to conform with the other one that is chosen as a reference. This inevitably leads to some loss of information. We propose a new ordination method, named spatial-RLQ analysis, for coupling two data sets with different spatial samplings. Spatial-RLQ analysis is an extension of co-inertia analysis and is based on neighbourhood graph theory and classical RLQ analysis. This analysis finds linear combinations of variables of the two data sets which maximise the spatial cross-covariance. This provides a co-ordination of the two data sets according to their spatial relationships. A vegetation study concerning the forest of Chizé (western France) is presented to illustrate the method.

Keywords: Co-inertia analysis; Neighbourhood relationship; Ordination; RLQ analysis; Spatial cross-covariance.
Abbreviations: CA: correspondence analysis; CCA: canonical correspondence analysis; GIS: Geographic Information System ; GSVD: generalised singular value decomposition; PCA: principal component analysis; RDA: redundancy analysis.
Nomenclature: Rameau et al. (1989)

Introduction

Ecology often deals with the coupling of two data sets. In most cases, the two data sets consist in environmental and species data collected in the same sites. When the sampling units are in agreement, a number of ordination methods can be used to link the two tables (ter Braak & Verdonschot 1995). Canonical correspondence analysis (CCA, ter Braak 1986, 1987) is probably the most frequently used method for this purpose in finding linear combinations of environmental variables that maximise the separation of species niches (Lebreton et al. 1988). There are several reasons for the success of CCA, and one of these is the dissymmetry of the approach, which uses environmental variables to model species composition structure by a step of multivariate regression. Redundancy Analysis (RDA, Rao 1964) contains also a multivariate regression step and can be preferable to CCA because the chi-square metric used by CCA over-emphasizes the importance of the rare species in the data set. Note that nonlinear relationships can also be modelled using nonlinear RDA and CCA (Makarenkov & Legendre 2002). The regression step of CCA or RDA requires that the number of environmental variables must be much lower than the number of samples, like in canonical correlation analysis. In this context, co-inertia analysis (Dolédec & Chessel 1994) is a good alternative because it is more robust than CCA regarding the number of variables compared to the number of individuals (ter Braak & Verdonschot 1995). Furthermore, co-inertia analysis has been extended, under the name of RLQ-method, to the case of linking three tables (Dolédec et al. 1996). This method has been named "RLQ" because it finds linear combination of the variables of table **R** (external information about rows) and linear combinations of the variables of table **Q** (external information about columns) of maximal covariance weighted by data contained in table **L** (link table). For example, RLQ has been used to link species traits to environmental variables by way of a species by sites table (Ribera et al. 2001).

For the methods discussed, measurements of species abundances and environmental variables must have been done in the same locations. In some cases, the two sampling schemes are different and so measurements are not done in the same locations. In biogeographic studies, for example, environmental data are available for meteorological stations whereas species abundances are measured at the quadrat level and are available from museum or atlas data. In vegetation science, people interested in different purposes can sample the same area at different locations and scales. Hitherto, there has been no method that reconciles the two sampling schemes, and the simplest way to analyse data is to estimate (e.g. by weighted averaging) the values from one table for the sampling points of the other one in order to have the same sampling units for the two tables (Mourelle & Ezcurra 1996, Hill 1991).

In this paper we propose a new ordination approach based on the RLQ ordination method and neighbourhood graph theory in order to link two data sets corresponding to different spatial sampling plans.

Neighbourhood matrix

The first step of the analysis is to establish a neighbourhood relationship between the sites of the two sampling schemes. Let us consider the situation where the first sampling involves m_1 sites whereas the second involves m_2 sites. A neighbourhood matrix **G** with m_1 rows and m_2 columns must be constructed where:

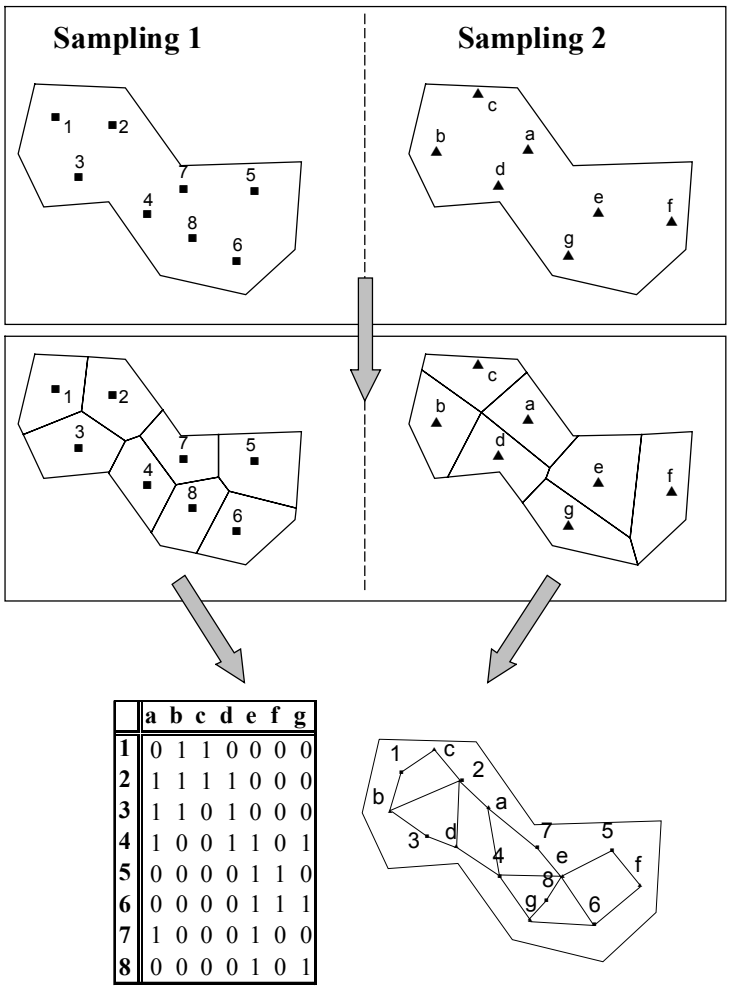
$$\mathbf{G}_{ij} = 1 \text{ if site } i \text{ and site } j \text{ are neighbours}$$
$$\mathbf{G}_{ij} = 0 \text{ otherwise.}$$

This kind of matrices is currently used in spatial ordination (Thioulouse et al. 1995). In the case where sites are 2-dimensional objects (i.e. quadrats, polygons...) we can easily fill this matrix by considering that:

$\mathbf{G}_{ij}$ = 1 if polygon *i* intersects polygon *j*
$\mathbf{G}_{ij}$ = 0 otherwise.

In the case where sites are considered as points, we propose to use tessellation to create neighbourhood relationships (Green & Sibson 1978). We consider the two different spatial samplings. Voronoi polygons can be easily constructed for each system of sampling with tessellation (figure 1). With these two tessellations, sites can be considered as polygons and we can apply the following decision rule:

$\mathbf{G}_{ij}$ = 1 if polygon induced by site *i* intersects polygon induced by site *j*
$\mathbf{G}_{ij}$ = 0 otherwise.



	a	b	c	d	e	f	g
1	0	1	1	0	0	0	0
2	1	1	1	1	0	0	0
3	1	1	0	1	0	0	0
4	1	0	0	1	1	0	1
5	0	0	0	0	1	1	0
6	0	0	0	0	1	1	1
7	1	0	0	0	1	0	0
8	0	0	0	0	1	0	1

Figure 1: Definition of a neighbourhood matrix from two different spatial sampling. From the two data sets, two tessellations are performed. Two voronoi polygons are neighbours if they intersect. Then, a neighbourhood matrix is constructed with 1 if the two individuals are neighbours and 0 otherwise.

Note that the matrix **G** can also be filled in computing the area of the intersection between polygons with GIS. Taking into account the overlapped area will give more weight to isolated points that produce large Voronoi polygons. Moreover, we propose to define the neighbourhood relations using two tessellations but simplest method can be used. For example, methods based on nearest neighbours criteria or on intervals of Euclidean distances that are used to define neighbourhood in the case of one set of points can be extended in the case of two sets of points. However, our choice has been guided by the fact that others methods can produce points with no neighbours (methods based on distance) or are difficult to be applied in the case of two sets of points and must be configured by the user (number of nearest neighbours or distance must be specified by the user). Our method based on

tessellation has the advantages that the whole zone is covered (all the points have neighbours) and that the method is defined without parameters entered by the user. However, it must be keep in mind that neighbourhood relation represents the strength of the potential interaction between two points and so the choice of the neighbourhood matrix can greatly influence the results of the analysis.

Measurements of spatial covariance

A major purpose of spatial statistics is to understand the spatial distribution of the values of an attribute over the whole study region (Bailey & Gatrell 1995). Is the value observed at a particular location correlated to those observed at neighbouring points? To answer this question, spatial covariance and covariograms are well known tools (see Cressie 1991 for example). We consider a single variable **x** measured at *n* locations (x_1, …, x_n) defining a *n* by *n* neighbourhood matrix $\mathbf{G}^*$. The matrix $\mathbf{G}^*$ is symmetric and allows to define a diagonal matrix of neighbouring weights $\mathbf{D}^* = diag(p^*_{1+},...,p^*_{i+},...,p^*_{n+})$ where $p^*_{i+} = \sum_{j=1}^{n} \mathbf{G}^*_{ij} / 2m$ is the neighbouring weight for the point *i* and $m = \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{G}^*_{ij} / 2$ is the number of pairs of neighbours. For a single variable **x**, the spatial covariance is (Thioulouse et al. 1995):

$$Cov_{spat}(\mathbf{x}) = \frac{1}{2m}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{G}^*_{ij}(x_i - \bar{x}_{\mathbf{D}^*})(x_j - \bar{x}_{\mathbf{D}^*})$$

where $\bar{x}_{\mathbf{D}^*} = \sum_{i=1}^{n} p^*_{i+} x_i$ is the mean of the variables **x** given the weights $\mathbf{D}^*$. The word covariance is rather ambiguous here because in the spatial context, it concerns the same variable and not two variables like in the general statistical context. But we can extend the idea of spatial covariance for a single variable (**x**) to the cross-covariance between two variables (**x**, **y**):

$$Cov_{spat}(\mathbf{x},\mathbf{y}) = \frac{1}{2m}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{G}^*_{ij}(x_i - \bar{x}_{\mathbf{D}^*})(y_j - \bar{y}_{\mathbf{D}^*})$$

The use of cross-covariance has been introduced in kriging methods for spatial interpolation (see Bailey & Gatrell 1995). Suppose that **x** has been recorded at *n* sites, and additional information on possible covariate **y** is recorded at *n+a* sites. Then, we can apply co-kriging (based on cross-covariance) by using covariate information to improve the prediction of **x** at a general point in the whole study region. While co-kriging requires that **x** and **y** be measured at the same *n* locations, the notion of cross-covariance is still legitimate when **x** and **y** have been sampled at different locations.

Spatial-RLQ ordination

Let **R** be an m_1 by *p* matrix containing the measurements of *p* variables at m_1 sites. Let **Q** be an m_2 by *q* matrix containing the measurements of *q* variables at m_2 sites. Spatial-RLQ is based exactly on the same principles as classical RLQ ordination. The only difference concerns the central table **L**. The table **L** derives from a species by sites abundance table in classical RLQ whereas it derives from an m_1 by m_2 neighbourhood matrix in spatial-RLQ analysis. The first step of the method consists of separate ordinations of **R**, **L** and **Q**. The second step is the study of the common structure of **R** and **Q** through **L** with spatial-RLQ analysis.

Correspondence analysis of the central table

Let us consider the m_1 by m_2 matrix **G** where $\mathbf{G}_{ij}$=1 if sites *i* and *j* are neighbours and $\mathbf{G}_{ij}$=0 otherwise. The table **P** of neighbourhood relative frequencies has m_1 rows and m_2 columns and contains the relative frequencies $\mathbf{P}_{ij} = \mathbf{G}_{ij}/g_{++}$. Moreover, let $g_{i+} = \sum_{j=1}^{m_2} \mathbf{G}_{ij}$, $g_{+j} = \sum_{i=1}^{m_1} \mathbf{G}_{ij}$ and $g_{++} = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \mathbf{G}_{ij}$ be the row totals, the column totals and the grand total of neighbours, respectively. We define $\mathbf{D}_{m_1}$, the diagonal matrix of row neighbouring weights: $\mathbf{D}_{m_1} = diag(p_{1+},...,p_{i+},...,p_{m_1+})$ where $p_{i+} = g_{i+} / g_{++}$. In the same way, the diagonal matrix of column neighbouring weights is $\mathbf{D}_{m_2} = diag(p_{+1},...,p_{+j},...,p_{+m_2})$ where $p_{+j} = g_{+j} / g_{++}$. Correspondence analysis (CA) of neighbourhood table **G** is the generalised singular value decomposition (GSVD) of the statistical triplet (**L**, $\mathbf{D}_{m_2}$, $\mathbf{D}_{m_1}$) with $\mathbf{L} = \mathbf{D}_{m_1}^{-1}\mathbf{P}\mathbf{D}_{m_2}^{-1} - \mathbf{1}_{m_1 m_2}$ (i.e. $\mathbf{L}_{ij} = \frac{\mathbf{G}_{ij}}{g_{i+}g_{+j}} - 1$, Greenacre 1984). This GSVD finds a $\mathbf{D}_{m_2}$-normed vector **c** and a $\mathbf{D}_{m_1}$-normed vector **d** maximising the quantity (Dolédec et al. 1996)

$$\mathbf{d}^t\mathbf{D}_{m_1}\mathbf{L}\mathbf{D}_{m_2}\mathbf{c} = \frac{1}{g_{++}}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\mathbf{G}_{ij}(d_i - \bar{d}_{\mathbf{D}_{m_1}})(c_j - \bar{c}_{\mathbf{D}_{m_2}}) = Cov_{spat}(\mathbf{c},\mathbf{d})$$

Since **c** and **d** are normed vectors, the above quantity is simply the spatial cross-correlation:

$$Cov_{spat}(\mathbf{c},\mathbf{d}) = Corr_{spat}(\mathbf{c},\mathbf{d})\sqrt{Var(\mathbf{c})}\sqrt{Var(\mathbf{d})} = Corr_{spat}(\mathbf{c},\mathbf{d})$$

Lebart (1984) has applied correspondence analysis to find row and column scores with maximum spatial correlation for the symmetric matrix **G** of a simple graph. In the case of an asymmetric neighbourhood matrix, correspondence analysis finds row and column scores maximising the spatial cross-correlation between neighbours. Eigenvectors corresponding to highest eigenvalues describe global structures whereas eigenvectors corresponding to the lowest eigenvalues describe local structures. These eigenvectors can be used as a good alternative to polynomial functions in trend surface analyses to describe spatial patterns (Thioulouse et al. 1995).

Analyses of R and Q

$\mathbf{D}_q$ and $\mathbf{D}_p$ are, respectively, a *q* by *q* and a *p* by *p* diagonal matrices of column weights associated to tables **Q** (m_2 by *q*) and **R** (m_1 by *p*). Different ordination methods can be used to analyse the tables **R** and **Q**. Different table transformations imply different analyses (see Dolédec et al. 2000). For example, if the variables in **R** are centred then the GSVD of (**R**, $\mathbf{I}_p$, $\mathbf{D}_{m_1}$), ($\mathbf{I}_p$ being the *p* by *p* identity matrix) is a centred PCA whereas if the variables in **R** are centred and scaled to unit variance then the same GSVD is a normed PCA. Note that in the GSVD of (**R**, $\mathbf{I}_p$, $\mathbf{D}_{m_1}$), row weights (p_{i+}) derive from the correspondence analysis of table **L**. Finally, GSVD of (**R**, $\mathbf{D}_p$, $\mathbf{D}_{m_1}$) and (**Q**, $\mathbf{D}_q$, $\mathbf{D}_{m_2}$) can result in different types of analyses (normed PCA, centred PCA, CA, or multiple CA…).

Spatial-RLQ analysis

The purpose of spatial-RLQ analysis is to study the common structure of tables **R** and **Q** through the neighbourhood relationship instead of analysing tables **R** and **Q** separately and trying to find a common spatial structure. Spatial-RLQ is defined as the GSVD of

$\left(\mathbf{R}^t\mathbf{D}_{m_1}\mathbf{L}\mathbf{D}_{m_2}\mathbf{Q}, \mathbf{D}_q, \mathbf{D}_p\right)$. This analysis consists in finding a $\mathbf{D}_p$-normed axis $\mathbf{u}_1$ and a $\mathbf{D}_q$-normed component $\mathbf{v}_1$ so that the quantity:

$$\mathbf{u}_1{}^t\mathbf{D}_p\mathbf{R}^t\mathbf{D}_{m_1}\mathbf{L}\mathbf{D}_{m_2}\mathbf{Q}\mathbf{D}_q\mathbf{v}_1$$

is maximised.
We can rewrite the previous equation with $\mathbf{a} = \mathbf{R}\mathbf{D}_p\mathbf{u}_1$ and $\mathbf{b} = \mathbf{Q}\mathbf{D}_q\mathbf{v}_1$:

$$\mathbf{u}_1{}^t\mathbf{D}_p\mathbf{R}^t\mathbf{D}_{m_1}\mathbf{L}\mathbf{D}_{m_2}\mathbf{Q}\mathbf{D}_q\mathbf{v}_1 = \mathbf{a}^t\mathbf{D}_{m_1}\mathbf{L}\mathbf{D}_{m_2}\mathbf{b}$$

It can be easily demonstrated that:

$$\mathbf{a}^t\mathbf{D}_{m_1}\mathbf{L}\mathbf{D}_{m_2}\mathbf{b} = Cov_{spat}(\mathbf{a},\mathbf{b})$$

So, spatial-RLQ finds linear combinations of variables of **R** ($\mathbf{a} = \mathbf{R}\mathbf{D}_p\mathbf{u}_1$) and linear combinations of variables of **Q** ($\mathbf{b} = \mathbf{Q}\mathbf{D}_q\mathbf{v}_1$) that have maximum spatial cross-covariance. The spatial cross-covariance can be decomposed into three terms, like classical covariance:

$$Cov_{spat}(\mathbf{a},\mathbf{b}) = Corr_{spat}(\mathbf{a},\mathbf{b})\sqrt{Var(\mathbf{a})}\sqrt{Var(\mathbf{b})}$$

This decomposition shows that spatial-RLQ is a compromise between the three separate analyses. The first part ($Corr_{spat}(\mathbf{a},\mathbf{b})$) corresponds to the correspondence analysis of **L**, the second ($\sqrt{Var(\mathbf{a})}$) to the analysis of **R** and the third ($\sqrt{Var(\mathbf{b})}$) to the analysis of **Q**. In RLQ, we maximise a compromise that finds a score induced by the variables of **R** and a score induced by the variables of **Q** which have a maximum spatial cross-correlation. Maximum values are obtained from separate analyses and so we can compare ordination obtained from spatial-RLQ analysis to those obtained from separate analyses. Furthermore, a Monte-Carlo test is available by permuting rows of tables **R** and **Q** in order to test the legitimacy of the spatial-RLQ analysis.

Application

Data were collected from a 2 614 ha managed forest located in Chizé (western France, figure 2a). The first data set was collected at the sub-plot scale, which corresponds to the level of forestry management (4 ha on average). These data were available for the two main vegetation strata, the timber stands and the coppices. This data set was collected by foresters of the Office National des Forêts and is used essentially for forest management. For each sub-plot, foresters determine the three dominant species for the coppices and the four dominant species for the timber stand, and their cover (in %). Subspecies and rare species were not determined and data were pooled resulting in ten categories for timber stand data and five for coppice data (Table 1). The second data set concerns the same area but it contains information about vegetation accessible to roe deer (height < 1.20 m, Duncan et al. 1998). This data set was collected at the scale of one square meter sampling plots. This sampling is part of a population dynamic study aiming to understand relationships between roe deer population and their available food. For this second data set, taxonomic information is recorded at the genera level. For each quadrat, the presence (or absence) of genera was recorded. In total, 613 sub-plots (data set 1) and 578 points (data set 2) were recorded (figure 2). The first data set results in table **R** with 613 rows and 15 columns and the second one in table **Q** with 578 rows and 58 columns. The purpose of our study is an examination of the canopy (timber stand and coppice) – understory (vegetation lower than 1.20 m height) relationships when the two data sets are measured on different spatial scales.

Table 1: Taxonomic names and codes

Data set 1

Timber stand	*Timber stand*
Name	**Code**
Quercus sp.	QueT
Acer sp.	AceT
Pinus sp.	PinT
Other deciduous	DecT
Cedrus sp.	CedT
Carpinus betulus	CarT
Prunus avium	PruT
Fagus sylvatica	FagT
Abies douglasi	AbiT
Other coniferous	ConT
Coppices	
Name	**Code**
Quercus sp.	QueC
Acer sp.	AceC
Other deciduous	DecC
Carpinus betulus	CarC
Fagus sylvatica	FagC

Data set 2

Vegetation lower than 1.20m	
Name	**Code**
Acer sp.	Ace
Ajuga reptans	Ajure
Allium sativum	Allsa
Anemone nemorosa	Anene
Arum sp.	Aru
Calamentha sp.	Cal
Carex sp.	Car
Carpinus betulus	Carbe
Clematis vitalba	Clevi
Convolvulus sp.	Con
Cornus sp.	Cor
Corylus avellana	Corav
Crataegus sp.	Cra
Epilobium sp.	Epi
Eupatorium cannabinum	Eupca
Euphorbia sp.	Eup
Evonymus europeatus	Evoeu
Fagus sylvatica	Fagsy
Ficaria ranunculoides	Ficra
Fragaria sp.	Fra
Fraxinus excelsior	Fraex
Gallium sp.	Gal
Geum sp.	Geu
Glechoma sp.	Gle
graminaceous	gra
Hedera helix	Hedhe
Hieracium sp.	Hie
Hyacinthoides sp.	Hya
Hypericum sp.	Hyp
Ilex aquifolium	Ileaq
Lathyrus sp.	Lat
Ligustrum vulgare	Ligvu
Lithospermum sp.	Lit
Lonicera periclymenum	Lonpe
Melitta sp.	Mel
Ornithogalum sp.	Orn
other labiates	othla
other prunus	Othpr
Pinus sp.	Pin
Potentilla sterilis	Potst
Prunus spinosa	Prusp
Pulmonaria sp.	Pul
Quercus sp.	Que
Ranunculus sp.	Ran
Rhamnus sp.	Rha
Rosa sp.	Ros
Rubia peregrina	Rubpe
Rubus sp.	Rub
Ruscus aculeatus	Rusac
Senecio sp.	Sen
Sorbus domestica	Sordo
Sorbus torminalis	Sorto
Stachys sp.	Sta
Ulmus sp.	Ulm
Veronica sp.	Ver
Viburnum lantana	Vibla
Vicia sp.	Vic
Viola sp.	Vio

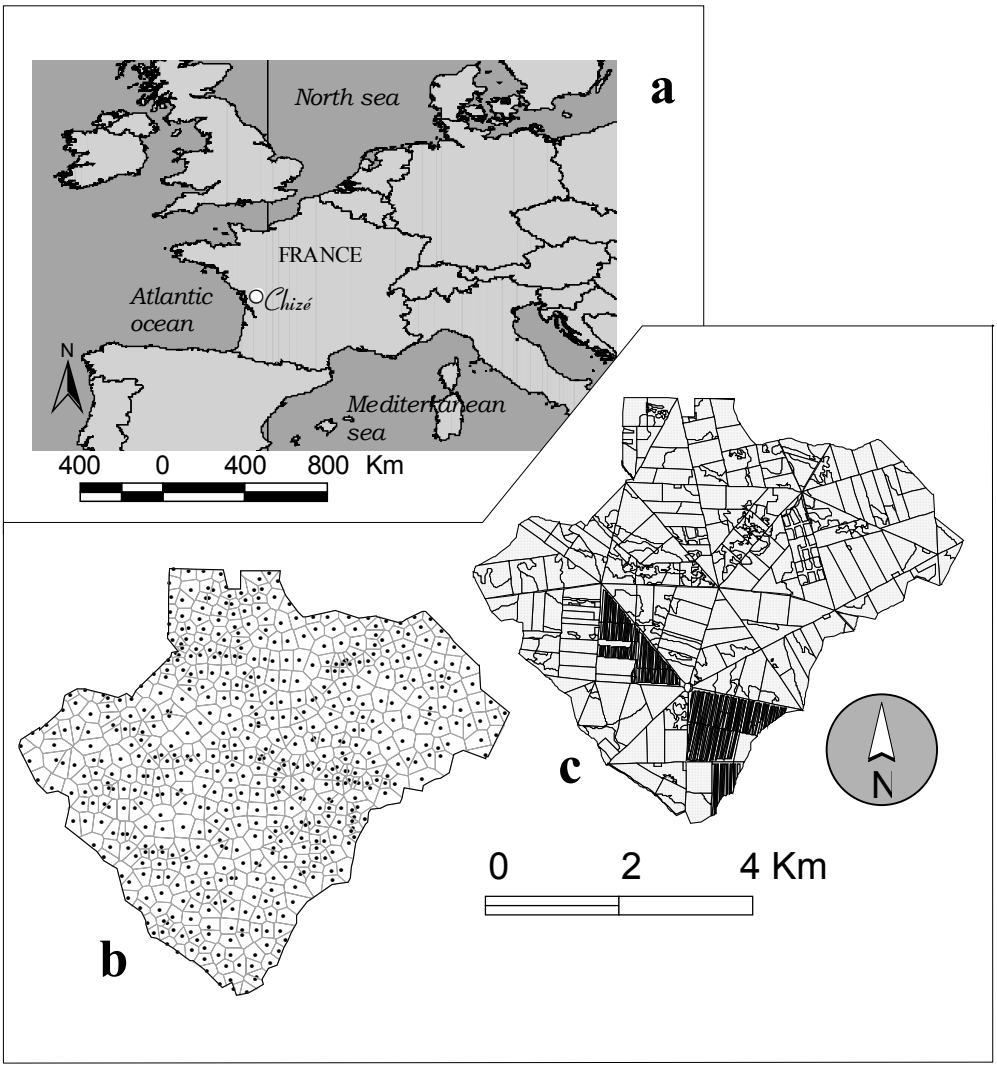


Figure 2: Spatial location (a) and sampling scheme (b, c) of the Chizé forest (Western France). Information have been collected for points (b) and for sub-plots (c). From the point data, a tessellation is applied in order to construct a neighbourhood matrix.

The data have been georeferenced and introduced in a Geographic Information System (GIS). Then, a tessellation on data points has been carried out and we have used the above decision rules to construct a neighbourhood matrix **L** with 613 rows and 578 columns. Separates analyses have been performed: PCAs for tables **R** and **Q** and CA for central table **L**. We performed a randomisation test to check for the statistical significance of the relationship between **R** and **Q**. This test is based on permutation of rows of tables **R** and **Q** and for each permutation the total inertia of the analysis is computed. The total inertia increases with the intensity of the link between **R** and **Q** through **L**. We used a Monte-Carlo version of the test with 1000 permutations, demonstrating a significant relationship ($p<0.001$: all permutations have values smaller than observed total inertia) validating the use of spatial-RLQ analysis. The first axis of the spatial-RLQ analysis takes into account 94% (935/990, table 2) of the total co-structure and we focus on results for this axis only. As seen before, spatial-RLQ analysis maximise the spatial covariance between linear combinations of the variables of **R** and linear combinations of the variables of **Q**. This covariance can be decomposed as the product of two standard deviations by their spatial correlation. Hence, it is possible to measure the proportion of variance attributed to each table and this can be compared to those obtained by separate analyses (table 2). For table **R**, the first axis of

spatial-RLQ analysis takes into account 97% (2988/3075) of the maximal potential inertia obtained by separate analysis and 89% (0.7419/0.8281) for table **Q**. Regarding spatial cross-correlation, the maximum is achieved by CA of table **L** and is equal to the square root of the first eigenvalue of CA ($\sqrt{0.9944}=0.9972$). The decrease of spatial cross-correlation (from 0.9972 to 0.6495) is due to the fact that row and column scores are constrained to be linear combinations of the variables of **R** and **Q** respectively whereas there is no constraint in CA.

Table 2: Inertia decomposition for spatial-RLQ analysis (three tables ordination). Inertia: maximal projected variability; Var: variance of the sets of factorial scores computed for the first axis; Cov: covariance between the two sets of factorial scores projected on the first spatial-RLQ axis; Cor: correlation between the two sets of factorial scores projected on the first spatial-RLQ axis.

	Spatial-RLQ analysis	*Maximal potential values (obtained by separate analysis)*
Total inertia	990	
Inertia projected on F1	935	
Cov (F1-**R**, F1-**Q**)	30.58	
Cor (F1-**R**, F1-**Q**)	0.6495	0.9972 (CA of **L**)
Var (F1-**R**)	2988	3075 (PCA of **R**)
Var (F1-**Q**)	0.7419	0.8281 (PCA of **Q**)

Taxonomic information has been projected onto the first axis of spatial-RLQ (figure 3). On this axis, genera are plotted according to the link of their spatial distribution with global spatial patterns. It is apparent that stands with *Fagus sylvatica* in the canopy (upper side of the first axis) tend to have *Rubus sp.* and *Ruscus aculeatus* in the understory and are mostly distributed in the south of the forest. The north of the forest is mainly occupied by stands with oak in the canopy and *Carpinus betulus* and *Ornithogalum sp.* (lower side of the first axis). So, it seems that the first axis indicates a species turnover from the north to the south of the forest involving different species communities in these two parts of the forest. Representation of scores of sub-plots and sample points by a smoothing by two-dimensional weighted local regression (Cleveland & Devlin, 1988) confirms these trends (figure 4a,b). Sample scores, which are defined by species composition, are structured from the north to the south. However, there are some differences between theses two maps especially in the southwest of the forest where the two scores are quite different. This lack of correspondence between the two scores is probably due to the fact that in this area, trees are young and these sub-plots do not contain mature timber stands.

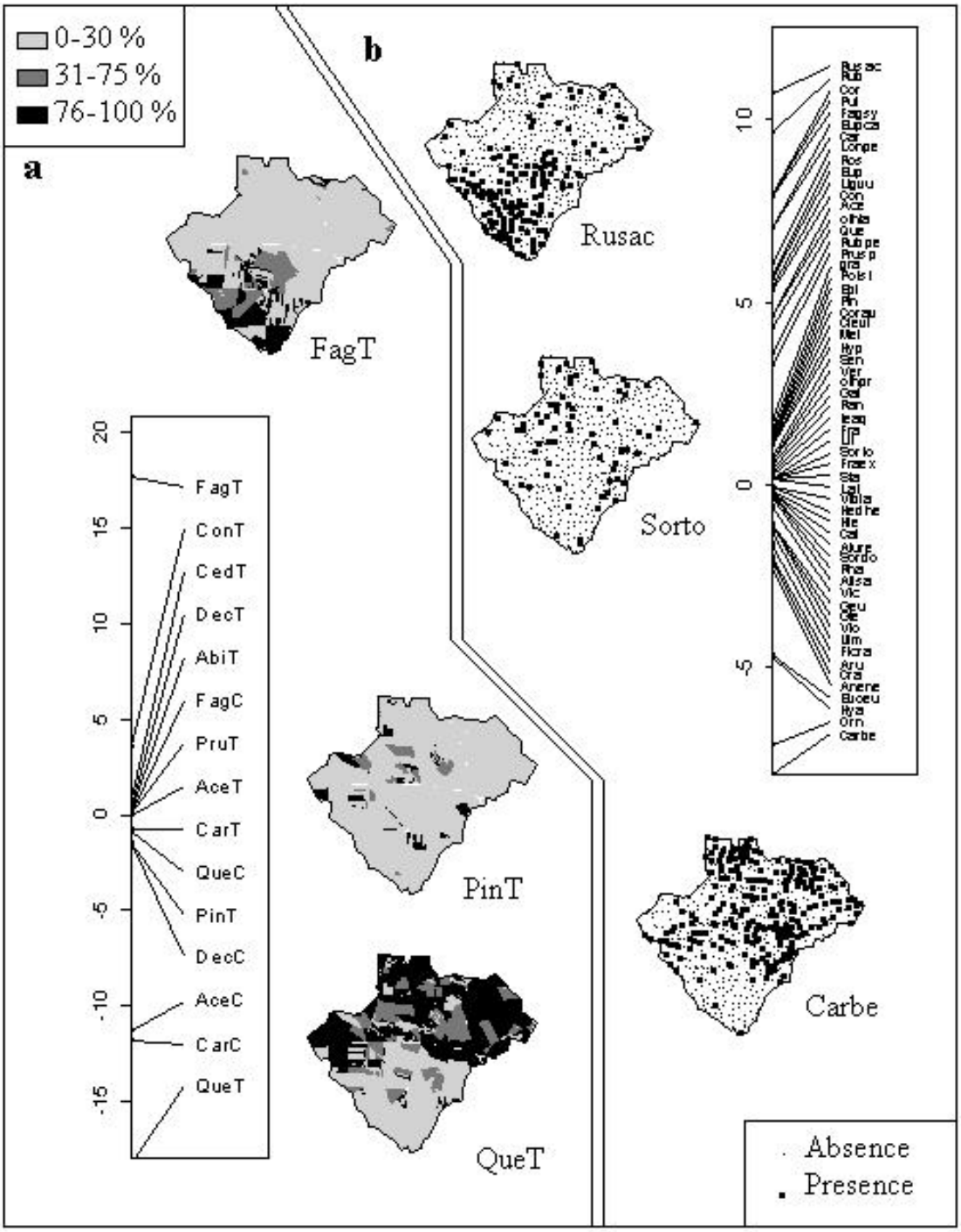


Figure 3: Spatial-RLQ analysis of Chizé data. First axis of species ordination for timber stand and coppices data (a) and for understory vegetation (b). Spatial distributions of some species are represented.

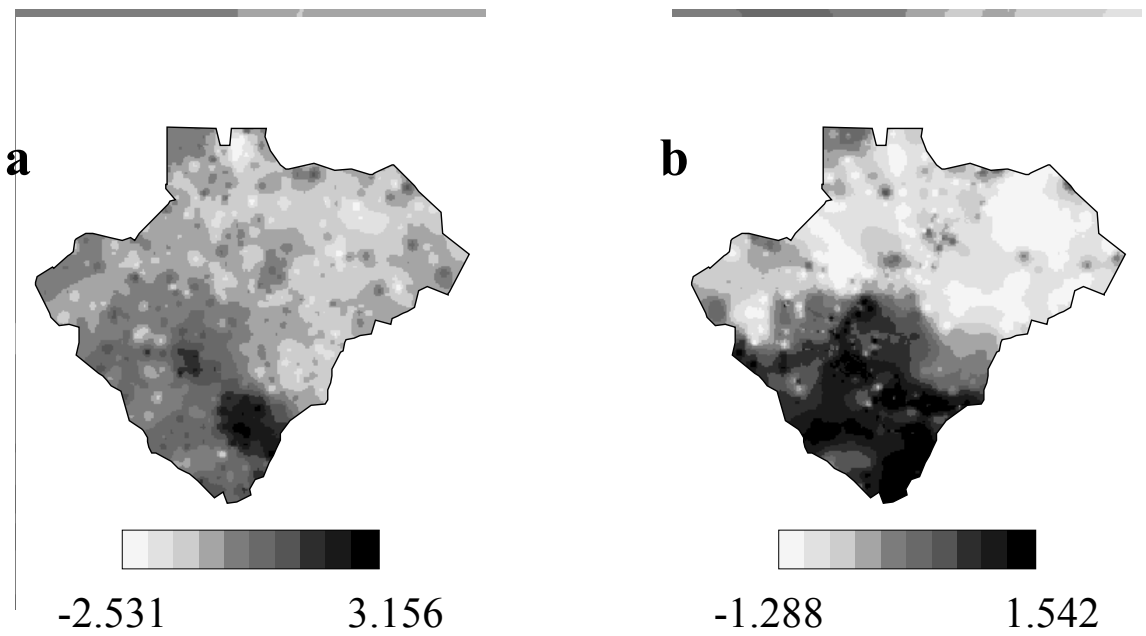


Figure 4: Spatial-RLQ analysis of Chizé data. Smoothing by two-dimensional weighted local regression of (a) points scores of understory vegetation, (b) sub-plots scores of timber stand and coppices.

Conclusion

Spatial-RLQ analysis is a new methodological perspective for coupling two data sets. This method is close to co-inertia analysis, which is used for linking two data sets with the same samples. In our method, the two data sets are considered symmetrically but it could be interesting to introduce an asymmetric part in order to explain one data set by the other. As CCA can be considered as an asymmetric co-inertia analysis, doubly constrained CCA (Böckenholt & Böckenholt 1990, Lavorel et al. 1998,1999) could be a good starting point for an asymmetric form of spatial-RLQ analysis. Moreover, we define in this paper a way to construct the neighbourhood matrix. This matrix represents the strength of the potential interactions between locations. The use of GIS permits the definition of a number of procedures to construct spatial neighbourhood matrices (Anselin & Getis 1992). Obviously, our choice is very simple but spatial-RLQ is flexible and can admit different kinds of neighbourhood matrices. The only constraint is due to the CA of the central table implying that all elements of this table must be non-negative. For example, taking into account the area of one sample crossing another one would probably make the analysis more realistic. Since the analytical results may be sensitive to the specification of the neighbourhood matrix, different spatial neighbourhood matrices may be needed for different purposes of studies. There is no universal type of neighbourhood matrix that can be used in spatial analysis. The choice of neighbourhood matrix and multiple possibilities of analyses of marginal tables **R** and **Q** make spatial-RLQ analysis appears as a flexible and a general method for spatial co-ordination of data.

Acknowledgements

We are grateful to the Office National des Forêts and to all fied assistants and volunteers that spent time collecting data on the study site. We wish to thank J. Oksanen and two anonymous reviewers, whose suggestions and comments have allowed us to improve the first version of this text.

References

Anselin, L. & Getis, A. 1992. Spatial statistical analysis and geographic information systems. *Ann. Reg. Sci.* 26: 19-33.
Bailey, T.C. & Gatrell, A.C. 1995. *Interactive spatial data analysis*. Longman, Harlow.
Böckenholt, U. & Böckenholt, I. 1990. Canonical analysis of a contingency tables with linear constraints. *Psychometrika*. 55: 633-639.
Cleveland, W.S. & Devlin, S.J. 1988. Locally weighted regression: an approach to regression analysis by local fitting. *J. Amer. Statist. Assn.* 83: 596-610.
Cressie, N. 1991. *Statistics for spatial data*. John Wiley and Sons, New-York.
Dolédec, S. & Chessel, D. 1994. Co-inertia analysis: an alternative method for studying species-environment relationships. *Freshwater Biol.* 31: 277-294.
Dolédec, S., Chessel, D. & Gimaret-Carpentier, C. 2000. Niche separation in community analysis: a new method. *Ecology*. 81: 2914-2927.
Dolédec, S., Chessel, D., ter Braak, C.J.F. & Champely, S. 1996. Matching species traits to environmental variables: a new three-table ordination method. *Environ. Ecol. Stat.* 3: 143-166.
Duncan, P., Tixier, H., Hofman, R.R. & Lechner-Doll, M. 1998. Feeding strategies and the physiology of digestion in roe deer. In: Andersen R., Duncan P. & Linnell J.D.C. (eds.), *The european roe deer : the biology of success*, pp. 91-116.Scandinavian University Press, .
Green, P. & Sibson, R. 1978. Computing Dirichlet tessellations in the plane. *Comput. J.* 21: 168-173.
Greenacre, M.J. 1984. *Theory and Applications of Correspondence Analysis*. Academic Press, London.
Hill, M.O. 1991. Patterns of species distribution in Britain elucidated by canonical correspondence analysis. *J. Biogeogr.* 18: 247-255.
Lavorel, S., Rochette, C. & Lebreton, J.D. 1999. Functional groups for response to disturbance in Mediterranean old fields. *Oikos*. 84: 480-498.
Lavorel, S., Touzard, B., Lebreton, J.D. & Clément, B. 1998. Identifying functional groups for response to disturbance in an abandoned pasture. *Acta Oecol.* 19: 227-240.
Lebart, L.1984. Correspondence analysis of a graph structure. *Bull. Tech. CESIA*. 2: 5-19.
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 Oecol. - Oecol. Gener.* 9: 53-67.
Makarenkov, V. & Legendre P. 2002. Nonlinear redundancy analysis and canonical correspondence analysis based on polynomial regression. *Ecology*. 83: 1146-1161.
Mourelle, C. & Ezcurra, E. 1996. Species richness of Argentine cacti: a test of biogeographic hypotheses. *J. Veg. Sci.* 7: 667-680.
Rameau, C., Mansion, D. & Dumé, G. 1989. *Flore forestière française, Plaine et Colline*. Institut pour le développement forestier, Paris.
Rao, C. R. 1964. The use and interpretation of principal component analysis in applied research. *Sankhya A*. 26: 329-359.
Ribera, I., Dolédec, S., Downie, I.S. & Foster, G.N. 2001. Effect of land disturbance and stress on species traits: a three table analysis of ground beetle assemblage. *Ecology*. 82: 1112-1129.
ter Braak, C.J.F. 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology*. 67: 1167-1179.
ter Braak, C.J.F. 1987. The analysis of vegetation-environment relationships by canonical correspondence analysis. *Vegetatio*. 69: 69-77.
ter Braak, C.J.F. & Verdonschot, P.F.M. 1995. Canonical correspondence analysis and related

multivariate methods in aquatic ecology. *Aquat. Sci.* 57: 255-289.
Thioulouse, J., Chessel, D. & Champely, S. 1995. Multivariate analysis of spatial patterns: a unified approach to local and global structures. *Environ. Ecol. Stat.* 2: 1-14.

Chapitre IX : Logiciels

Cette partie est consacrée aux deux utilitaires implémentés pendant cette thèse et distribués gratuitement aux utilisateurs. CoCoAn est une librairie du logiciel R qui est téléchargeable à l'adresse http://cran.r-project.org/src/contrib/PACKAGES.html#CoCoAn. Le développement et la mise à disposition de cette librairie ont été réalisé à la demande de B. D. Ripley, à la suite de l'échange du 16-17 février 2001 sur le forum de R (cf I.3. et sur le site de R-HELP à http://www.r-project.org/nocvs/mail/r-help/2001/index.html). Deux fonctions sont inclues dans cette librairie afin de réaliser des AFC, des ACC et les représentations graphiques associées (biplot, triplot).

AVADE est une extension d'ArcView écrite en Avenue permettant l'interface avec ADE-4. Cette extension utilise la possibilité de lancement automatisé des modules d'ADE-4 au moyen de fichiers batch. Il est ainsi possible de réaliser des analyses multivariées, disponibles dans ADE-4, sans quitter le SIG. Parmi les méthodes disponibles, figurent l'ACP normé ou centré, l'AFC, les ACP spatialisées (Thioulouse *et al.* 1995) et le couplage de deux tableaux par l'analyse canonique des correspondances. Les résultats sont directement importés sous forme de table et il est donc aisé de cartographier les résultats de l'analyse dans ArcView. En l'absence de l'extension AVADE, cette démarche était assez laborieuse puisqu'elle nécessitait de nombreux transferts et transformations des fichiers entre ArcView et ADE-4. AVADE offre d'autres fonctionnalités comme l'import et l'export de fichiers binaires, d'un découpage surfacique (fichiers .area) ou l'import, la création et l'export d'une grille de quadrats (fichiers .lat).

Afin d'illustrer cette partie, nous avons inséré les manuels d'utilisations de ces deux utilitaires.

The CoCoAn Package

March 22, 2002

Title Constrained Correspondence Analysis

Version 1.0-2

Date 2000/02/26

Author Stephane Dray <dray@biomserv.univ-lyon1.fr>

Maintainer Stephane Dray <dray@biomserv.univ-lyon1.fr>

Description Two functions to compute correspondence analysis and constrained correspondence analysis and to make the associated graphical representation.

License GPL2

R topics documented:

CAIV
CAIV.plot
Index

CAIV	Function to perform correspondence analysis with (or without) respect to instrumental variables

Description

Multivariate analysis. This function performs correspondence analysis or constrained correspondence analysis. This latter is better known under the name of canonical correspondence analysis. This analysis finds coefficients of variables to obtain a row score of unit variance. This row score is used to compute by weighted averaging a column score of maximized variance.

Usage

```
CAIV(L, E=diag(1, dim(L)[1], dim(L)[1]), normE=TRUE)
```

Arguments

L	a (i,j) matrix of non-negative number
E	an (i,p) optional matrix of p external variables

normE TRUE to normalize variables in matrix E, FALSE otherwise

Details

This function compute correspondence analysis (enter L) or constrained correspondence analysis (enter L and E). The function return the coefficient (B) to compute a row score of unit variance (R) that maximize the between-column inertia (column score in F obtained by weighting averaging). D contains the intra-set covariance (correlation if normE=TRUE). For correspondence analysis, CAIV(t(L)) gives a column score of unit variance that maximize the between-rows inertia. Note that this function does not use convenient rescaling and so is a little bit different of ter Braak's CCA. (We use the algorithm of Chessel et al.)

Value

A	list with components
ev	a vector containing eigenvalues
B	coefficients of variables of E (only in constrained analysis)
D	covariance matrix between external variables and row scores (only in constrained analysis)
R	row coordinates of unit variance
F	column coordinates of variance ev[i]

Author(s)

Stephane DRAY <dray@biomserv.univ-lyon1.fr>

References

ter Braak (1986): Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology 67(5), 1167–1179.
Chessel, Lebreton and Yoccoz (1987): Propriétés de l'analyse canonique des correspondances; une illustration en hydrobiologie. Revue de Statistique Appliquée 35(4) 55–72.

See Also

CAIV.plot

Examples

```
##correspondence analysis
L <- matrix(c(4,2,0,2,0,5,1,3,2,4,0,2,2,0,3,1),4,4)
CAIV(L)
CAIV(t(L))
##canonical correspondence analysis
E <-matrix(c(1.5,2.3,2,1.6,0.9,0.8,1.2,1.5),4,2)
CAIV(L,E)
```

CAIV.plot Biplot or Triplot after using the CAIV function

Description

CAIV performs correspondence analysis or constrained correspondence analysis. CAIV.plot allows a two-dimensional representation (biplot) of row, column and coefficients of variables.

Usage

```
CAIV.plot(obj, x=1, y=2, add.row=TRUE, add.col=TRUE, add.var=FALSE, row.names="",
col.names="", var.names="")
```

Arguments

obj	an object created by the CAIV function
x	an integer that defines which score corresponds to x axis
y	an integer that defines which score corresponds to y axis
add.row	TRUE to add rows on biplot FALSE otherwise
add.col	TRUE to add columns on biplot FALSE otherwise
add.var	TRUE to add variables on biplot FALSE otherwise (only for constrained analysis)
row.names	a vector of strings containing row names (Ri otherwise)
col.names	a vector of strings containing columns names (Ci otherwise)
var.names	a vector of strings containing variables names (Vari otherwise)

Author(s)

Stephane DRAY <dray@biomserv.univ-lyon1.fr>

See Also

CAIV

Examples

```
## correspondence analysis
L <- matrix(c(4,2,0,2,0,5,1,3,2,4,0,2,2,0,3,1),4,4)
CAIV.plot(CAIV(L))
CAIV.plot(CAIV(L),row.names=c("a","b","c","d"))
## canonical correspondence analysis
E <- matrix(c(1.5,2.3,2,1.6,0.9,0.8,1.2,1.5),4,2)
CAIV.plot(CAIV(L,E))
CAIV.plot(CAIV(L,E),add.var=TRUE)
```

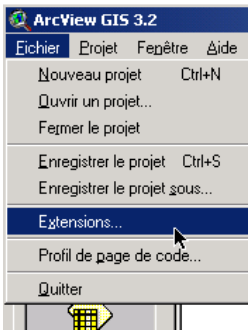
AVADE : Interface entre ArcView et ADE-4 (Windows)

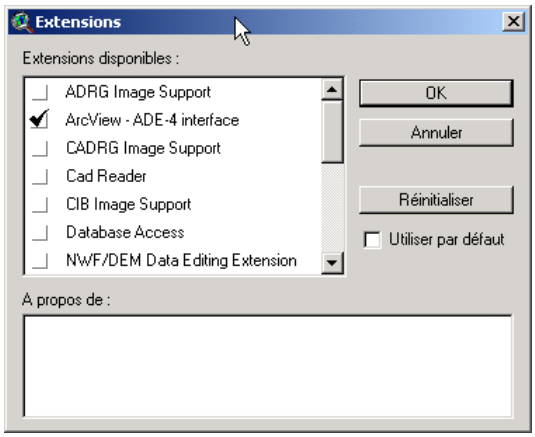
Avade est une extension ArcView permettant d'échanger des données avec ADE-4, de lancer quelques analyses et de récupérer les résultats dans ArcView. Cette interface utilise les nouvelles capacités d'ADE-4 permettant de lancer des modules en mode automatique (batch). Il est donc impératif d'avoir installer la nouvelle version d'ADE-4 (interface Metacard et modules)
N.B. : La mise en place de cette extension a nécessité quelques petites modifications des modules Lattice, CCA, Areas et NGStat. La nouvelle version de ces modules est disponible à http: //pbil.univ-lyon1.fr/ADE-4/ depuis le 05/04/2001.

PLAN

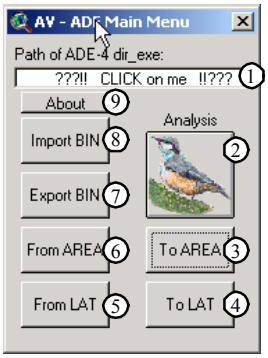
1. Installation

Placer l'extension avade.avx dans le dossier d'extension d'ArcView. Ce dossier est, par exemple, C:\ESRI\AV_GIS30\ARCVIEW\EXT32. Lancer ArcView, ouvrir un projet. Dans le menu **Fichier**, choisir **Extension…** et sélectionner l'extension :





Si on sélectionne une vue, un nouveau bouton contenant l'icône d'ADE-4 apparaît dans la barre de boutons et permet de lancer l'interface :

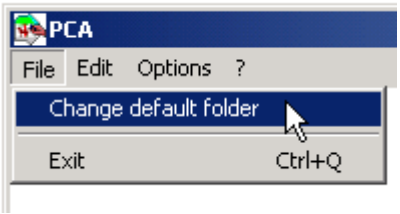


2. Le Menu Principal

A partir de la fenêtre du menu principal on peut :
1 : Indiquer le répertoire des modules exécutables (répertoire dir_exe)
2 : Accéder au menu des analyses
3 : Exporter un thème de polygone en un fichier area
4 : Créer et exporter une grille de quadrats en un fichier lattice
5 : Importer une grille de quadrats décrit par un fichier lattice
6 : Importer un fichier area
7 : Exporter une table en un fichier binaire
8 : Importer un fichier binaire
9 : Obtenir des informations sur l'interface

Avant de lancer une option veiller à indiquer le répertoire dir_exe contenant les modules exécutables d'ADE-4.

Attention : tous les fichiers sur lesquels on travaille doivent se trouver dans le dossier de travail courant défini par l'interface MetaCard.
Attention : il faut sélectionner le dossier de travail courant avec l'option "Change default folder" du menu "File" du module (sinon une erreur se produit systématiquement) : (cf. Thema 15)



Cette documentation s'appuie sur les données des cartes Mafragh. Ces données contiennent de l'information floristique (56 espèces) et mésologique (11 variables) pour 97 relevés. On a créé des fichiers binaires Mafrag_Mil et Mafrag_Flo et un fichier de numéro des sites LabelSite.txt (1…97).

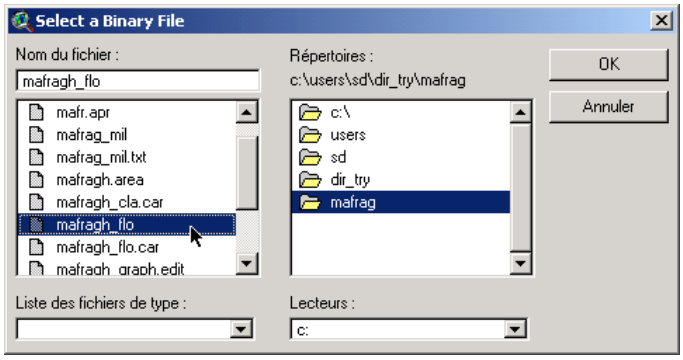
3. Les fonctions d'Import-Export

L'interface permet en premier lieu l'importation et l'exportation de fichiers issus d'ADE-4.

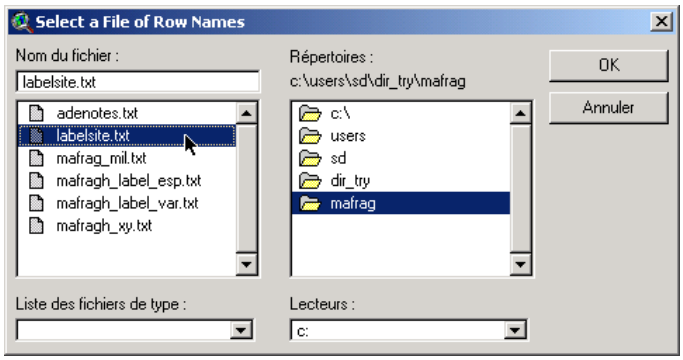
3.1. Fichiers binaires

Import

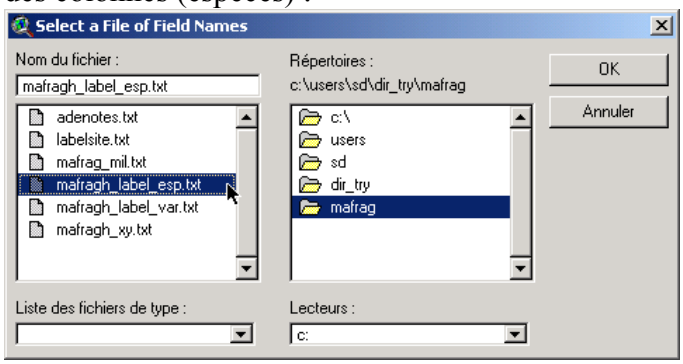
Dans le menu principal, cliquer sur le bouton **Import BIN**. Sélectionner le fichier de données floristiques dans le dossier de travail :



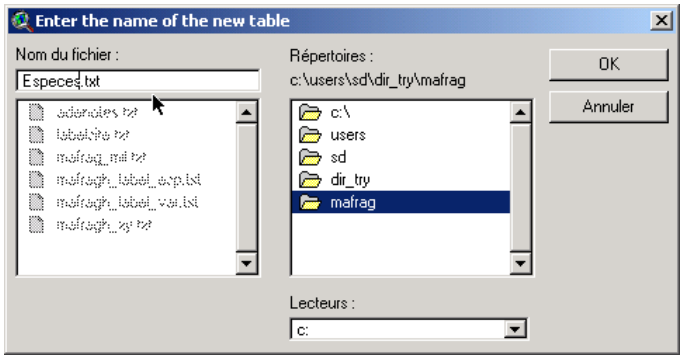
Puis le fichier de label des sites :



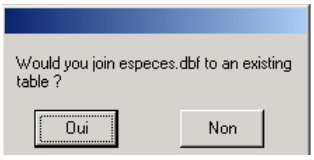
Et le fichier de label des colonnes (espèces) :



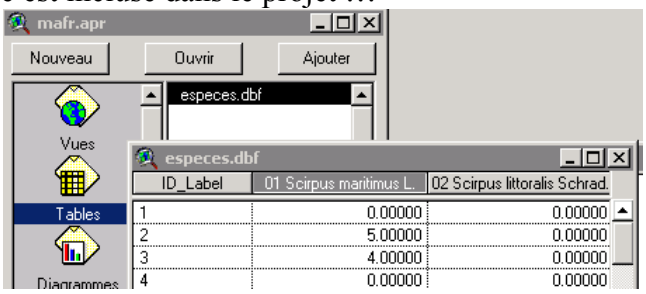
Entrer un nouveau nom de fichier :



On peut éventuellement joindre la table créée à une table existante (choisir non) :

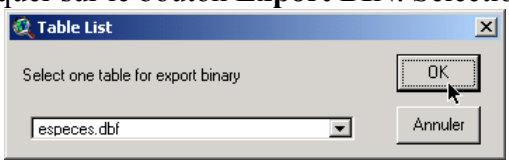


La nouvelle table créée est incluse dans le projet …



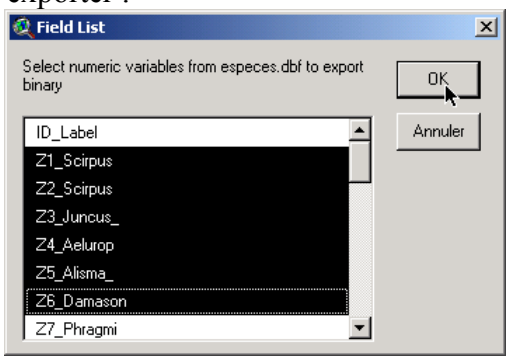
Export

Dans le menu principal, cliquer sur le bouton **Export BIN**. Sélectionner la table à exporter :



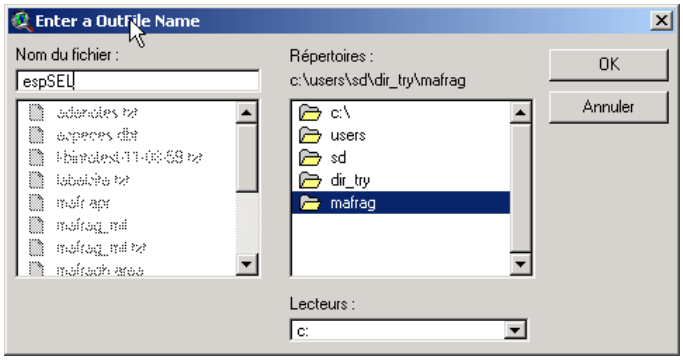
Remarque : L'export ne concerne que les enregistrements sélectionnés dans la table. Si il n'y a pas de sélection, tous les enregistrements sont exportés.

Sélectionner les variables à exporter :

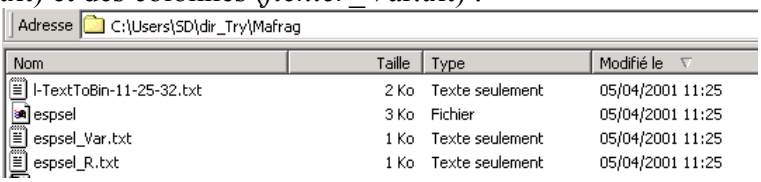


Sélectionner un camp d'identifiant des lignes :

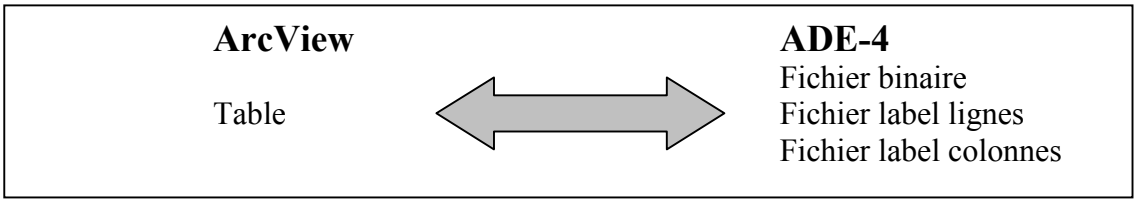
Indiquer un nom de fichier et veiller à bien être dans le dossier de travail :



L'interface a créé le fichier binaire (*fichier*), et deux fichiers texte contenant les labels des lignes (*fichier*_R.txt) et des colonnes (*fichier*_Var.txt) :

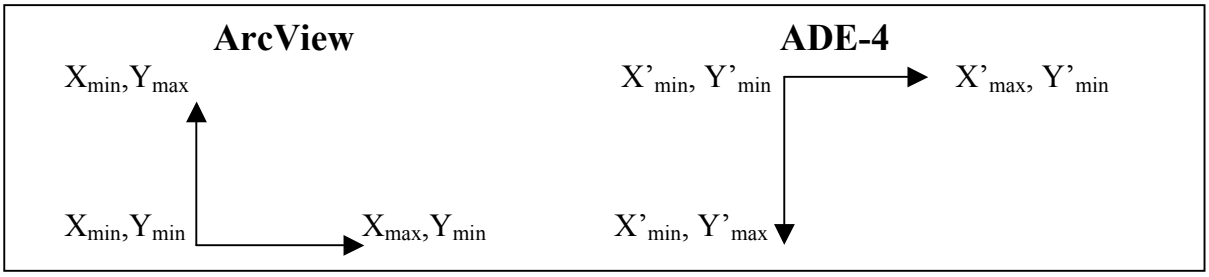


En résumé :



3.2. Fichiers graphiques

ADE-4 permet de travailler sur des fichiers graphiques à partir notamment des modules Areas et Lattices. Dans ADE-4, les coordonnées sont données dans un système pixel. Dans ArcView, un utilisateur peut travailler dans divers systèmes de coordonnées. Les principales différences sont une inversion de l'axe des Y et le fait que, dans le système en pixel, les coordonnées des points doivent être contenues dans l'écran.



Pour permettre de transférer des fichiers entre les deux logiciels, on a choisi d'opérer les transformations suivantes pour l'import de fichier :

ADE-4	→	**ArcView**
x'		x=x',
y'		y=Y'max-y'

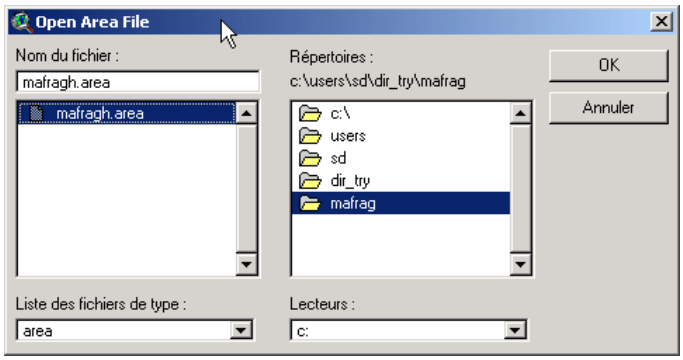
Pour l'export de fichier, on calcule l'étendue maximale et_{max}= *max* (|Ymax-Ymin|,|Xmax-Xmin|) et on la ramène à 500 pixels :

ArcView	→	**ADE-4**
x		$x'=(500/et_{max})*x$
y		$y'=(500/et_{max})*(Ymax-y)$

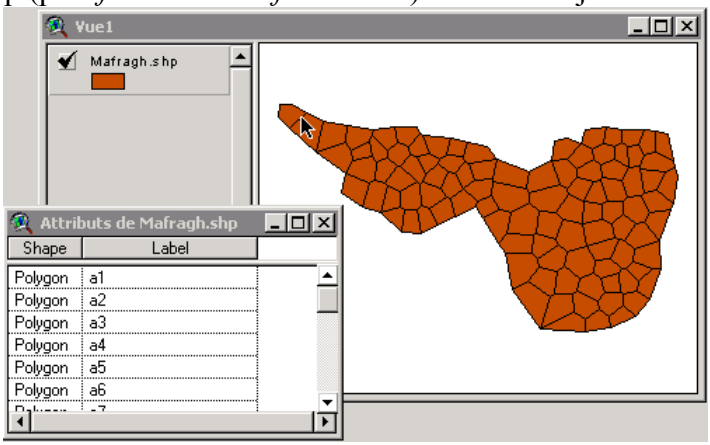
3.2.1. Fichiers area

Import

Cliquer sur une vue dans le projet ArcView pour la rendre active. Dans le menu principal, cliquer sur le bouton **From Area**. Sélectionner un fichier *fichier*.area dans le dossier de travail :

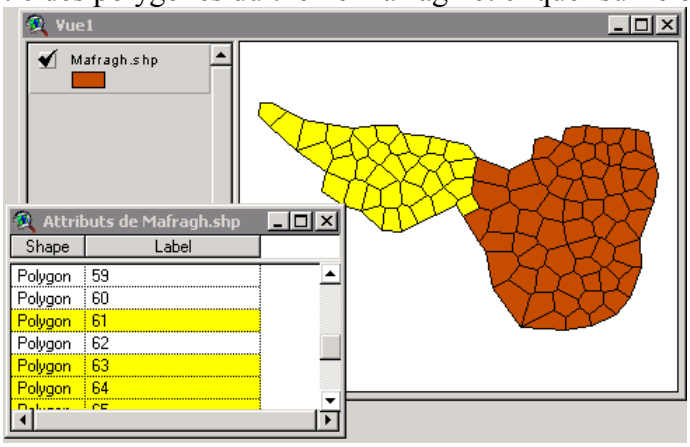


Un thème *fichier*.shp (plus *fichier*.shx et *fichier*.dbf) est crée et ajouté à la vue :



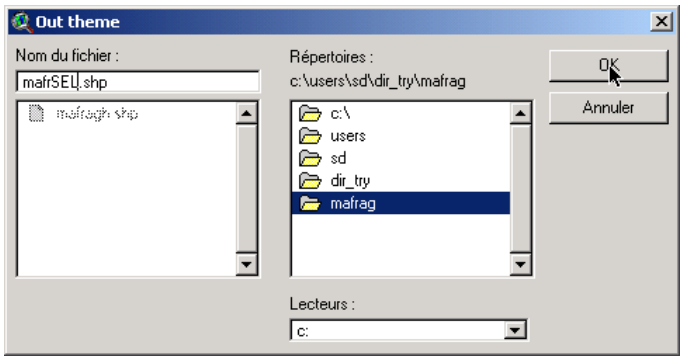
Export

Sélectionner une partie des polygones du thème mafragh et cliquer sur le bouton **To Area** :

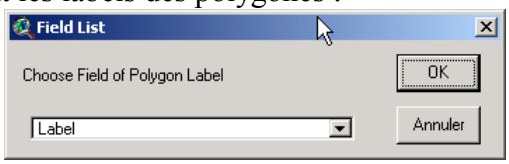


Entrer le nom du thème à exporter :

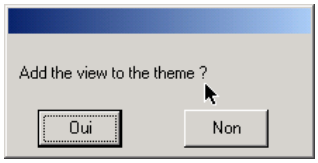
Puis entrer un nom de fichier avec une extension shp :



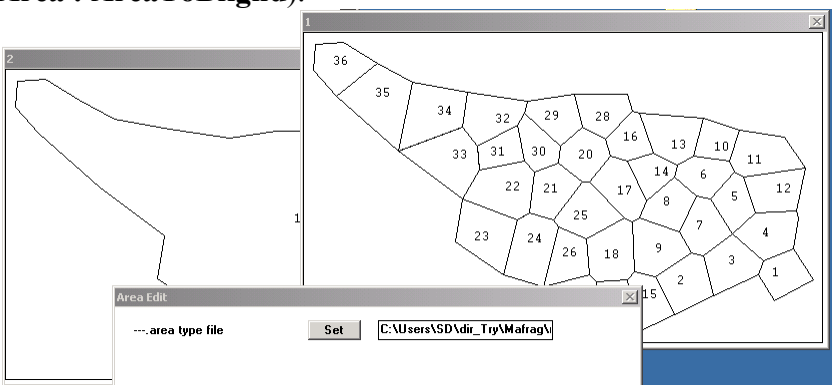
Choisir un champ contenant les labels des polygones :



L'interface a créé un thème de points ArcView reprenant le contour des polygones. On peut ajouter ce thème à la vue sinon il sera effacé :



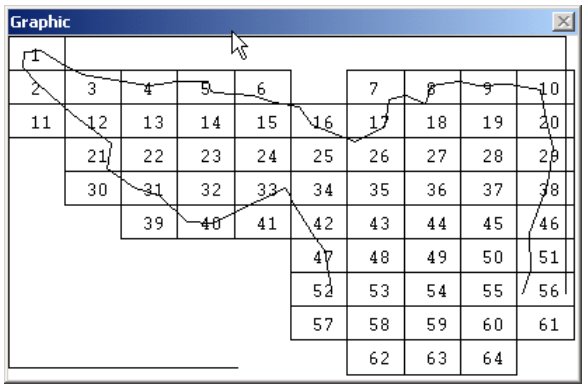
Dans ADE-4, utiliser **Areas : AreaEdit**. L'interface a créé deux fichiers *fichier*.area et *fichier*_c.area. Le second contient un seul polygone de contour et pourra servir pour créer un fond de carte (**Area : AreaToBkgnd**).



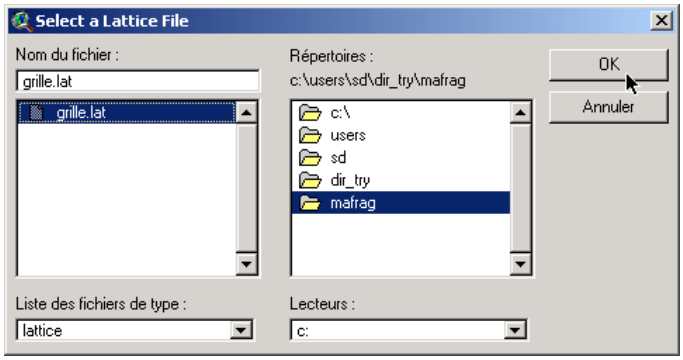
3.2.2. Fichiers lattices

Import

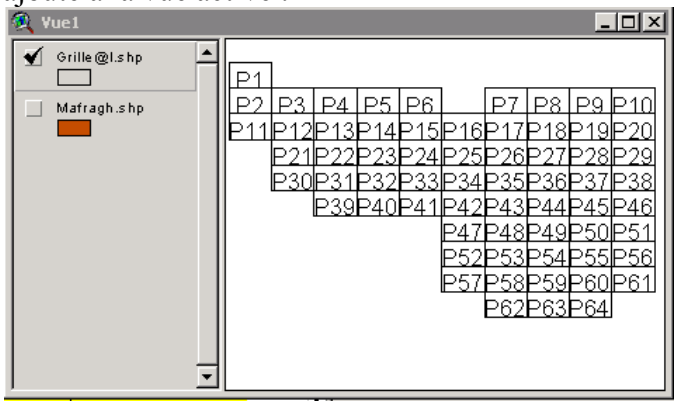
Dans ADE-4, créer un fichier .BMP à partir du fichier mafragh.area (**Area : AreaToBkgnd**) puis implanter une grille de quadrats sur ce fond (**Lattices : Create_Bkgnd**) :



Dans ArcView, sélectionner une Vue et Cliquer sur le bouton **From LAT** et sélectionner la grille préalablement créée.

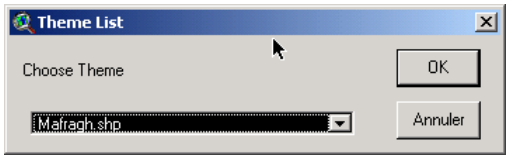


Un thème est créé et ajouté à la vue active :



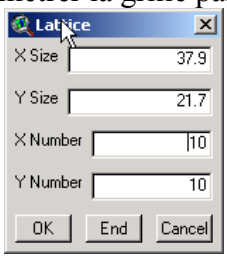
Export

Sélectionner une vue, cliquer sur le bouton **To LAT** et choisir un thème à partir duquel on va créer la grille de quadrats :

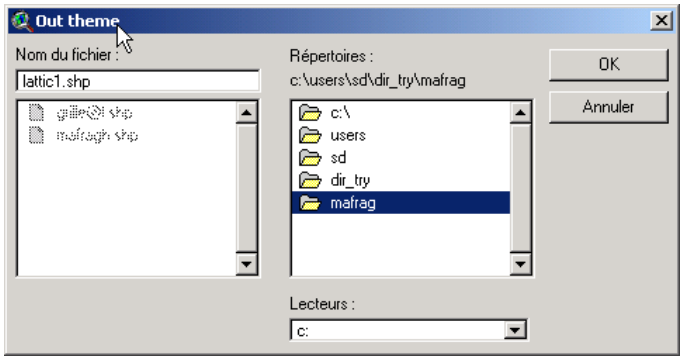


Remarque : L'export ne concerne que les enregistrements sélectionnés dans le thème. S'il n'y a pas de sélection, tous les enregistrements sont exportés.

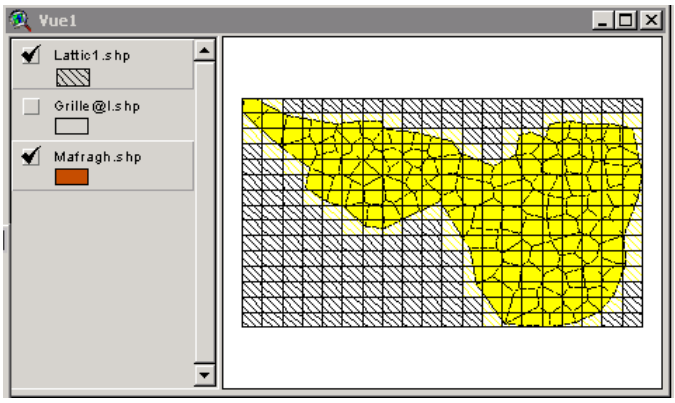
Une fenêtre apparaît et permet de paramétrer la grille par le nombre de quadrats ou leur taille :



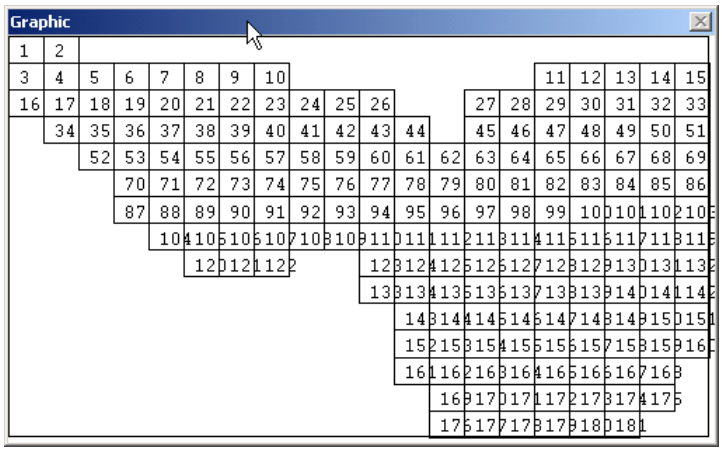
La fenêtre est interactive : si on change un paramètre et on **appuie sur la touche Entrée**, le paramètre dépendant est changé automatiquement.
Quand le paramétrage est choisi, appuyer sur le bouton OK et enregistrer le fichier dans le dossier de travail :



Un thème contenant la grille est créé et ajouté à la vue. Seuls les quadrats coupant des entités du thème d'origine sont sélectionnés :

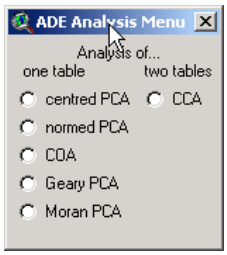


Modifier éventuellement cette sélection (cela peut être nécessaire pour des thèmes de lignes ou de points). Appuyer sur le bouton **End** et le fichier lattice est créé à partir des entités sélectionnées :



4. Le Menu Analysis

Le menu **analysis** permet de réaliser des analyses multivariées à partir d'ArcView. Dans son état actuel, le nombre d'analyses est réduit mais il peut facilement être augmenté (Il suffit de demander…) :

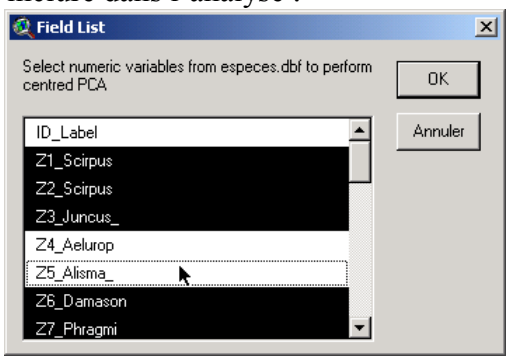


Remarque : Une analyse ne concerne que les enregistrements sélectionnés dans la table. S'il n'y a pas de sélection, tous les enregistrements sont utilisés.

4.1. Analyse en Composantes Principales centrée

Sélectionner centred PCA et choisir une table :

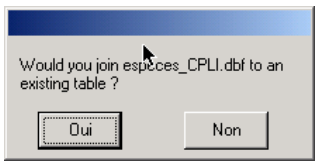
Sélectionner les variables à inclure dans l'analyse :



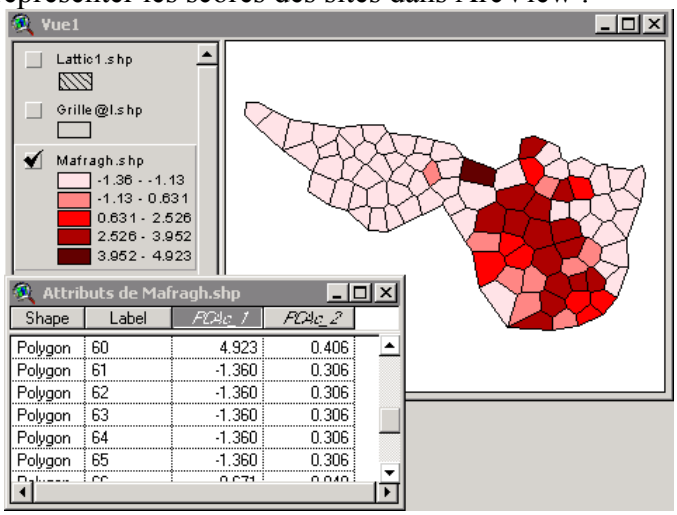
Puis sélectionner un champ de label des lignes :

Et enfin le nombre d'axes à conserver :

L'analyse est lancée, l'ensemble des fichiers d'ADE-4 est créé, une table des scores des lignes *fichier*_CPLI.dbf est créée et introduite dans ArcView. L'interface propose de faire une jointure :



Répondre **Oui** et faire la jointure avec la table Attributs de Mafragh.shp par les champs de label. On peut alors représenter les scores des sites dans ArcView :



4.2. Analyse en Composantes Principales normée

Le principe est le même que pour 4.1.

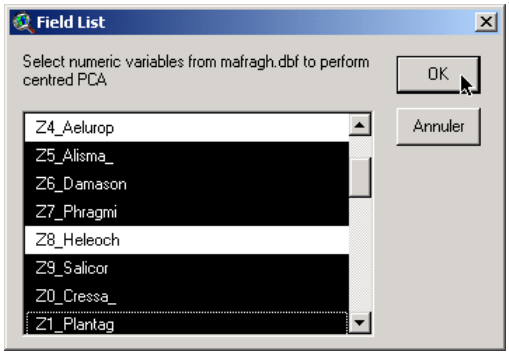
4.3. Analyse Factorielle des Correspondances

Le principe est le même que pour 4.1. Cependant, il faut se souvenir que l'AFC n'admet que des valeurs positives ou nulles.

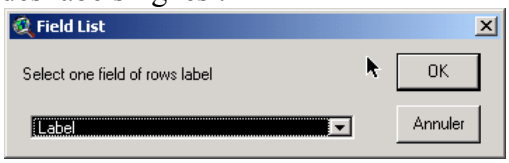
4.4. ACP de Moran

Pour cette option, l'analyse de base est une ACP normée. Faire une jointure entre la table especes.dbf et Attributs de Mafragh.shp par les champs de labels des lignes. Sélectionner une vue, choisir l'option **Moran PCA** et choisir un thème :

Sélectionner les variables de la table du thème à inclure dans l'analyse :

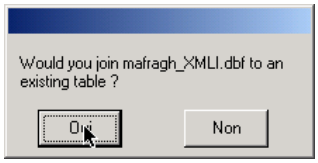


Puis sélectionner le champ des labels lignes :

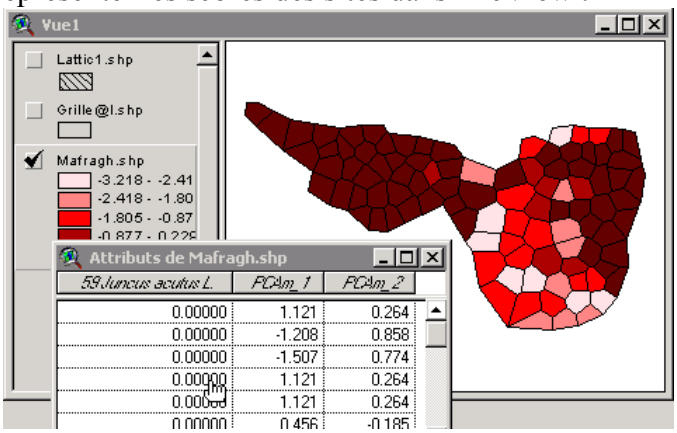


Et enfin le nombre d'axes à conserver :

L'analyse est lancée, l'ensemble des fichiers d'ADE-4 est créé, une table des scores des lignes *fichier*_XMLI.dbf est créée et introduite dans ArcView. L'interface propose de faire une jointure :



Répondre **Oui** et faire la jointure avec la table Attributs de Mafragh.shp par les champs de label. On peut alors représenter les scores des sites dans ArcView :

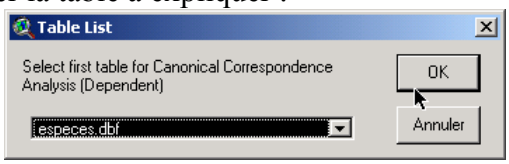


4.5. ACP de Geary

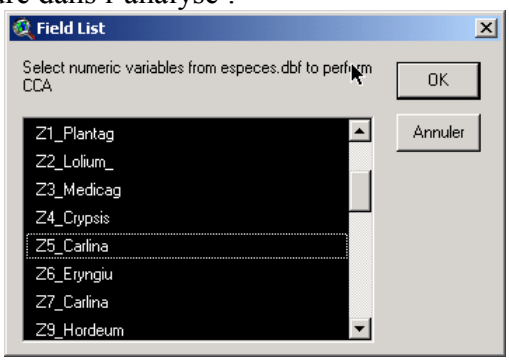
Le principe est le même que pour 4.4.

4.6. Analyse Canonique des Correspondances

L'ACC proposée réalise au préalable une ACP normée sur les variables de milieu. Choisir l'option CCA et sélectionner la table à expliquer :

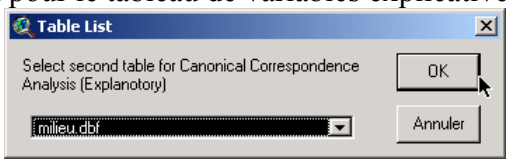


Choisir les variables à inclure dans l'analyse :

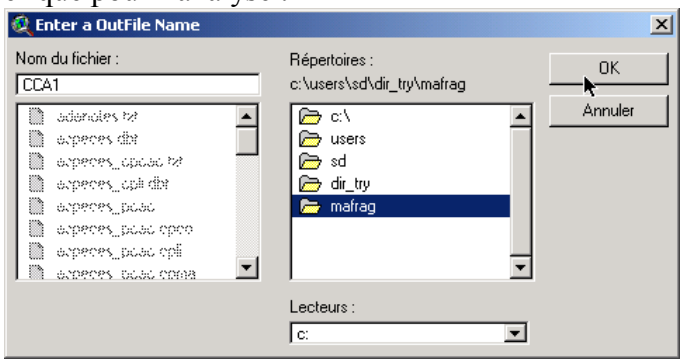


Sélectionner le champ contenant le label des lignes :

Réaliser la même opération pour le tableau de variables explicatives :

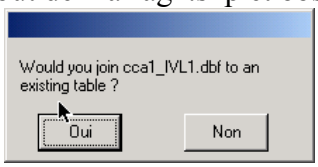


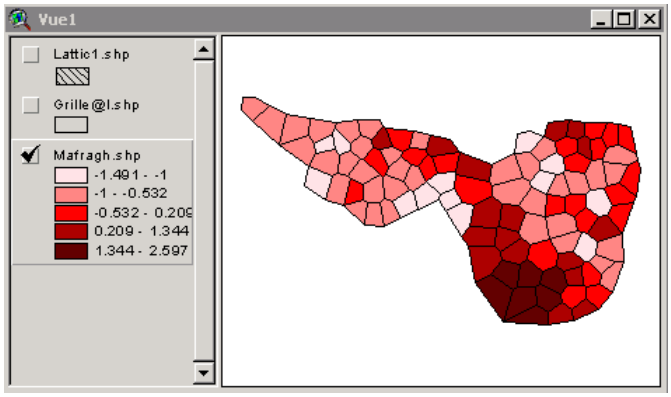
Indiquer un nom générique pour l'analyse :



Le nombre d'axes à conserver :

Faire la jointure avec la table Attribut de mafragh.shp et observer le résultat :





Conclusions

Au terme de ce travail de thèse, nous souhaitons avoir fourni certaines propositions qui faciliteront le traitement de données spatialisées. L'utilisation de la technologie SIG et son couplage avec les méthodes d'analyse multivariée procure de nombreuses perspectives pour améliorer l'analyse de données écologiques. La méthode « *spatial-RLQ* » offre la possibilité du couplage de deux jeux de données échantillonnés à différents sites. Cette méthode, basée sur les principes de l'analyse RLQ (Dolédec *et al.* 1996), nécessite le calcul d'une matrice de voisinage entre les deux plans d'échantillonnage. Il est évident que l'automatisation de cette étape à l'aide d'un SIG permettra désormais de faciliter la mise en œuvre de cette méthode et ainsi de favoriser son utilisation. De la même façon, la méthode proposée pour l'analyse de données incomplètes à l'aide des méthodes PLS requiert l'utilisation d'un SIG pour la création du tableau de données. L'outil SIG est alors fondamental pour la mise en œuvre de la méthode.

L'analyse des listes d'occurrences a mis en exergue la nécessité d'une adéquation entre méthodes, données et objectifs de l'étude. L'analyse canonique des corrélations (AC) est une méthode fondamentale en analyse de données car sa démarche (rechercher des couples de variables de corrélation maximale) se retrouve dans d'autres méthodes comme l'analyse discriminante ou l'analyse factorielle des correspondances. Cependant, les applications directes de l'AC en écologie sont peu nombreuses et elle s'est avérée inefficace pour l'analyse des tableaux espèces-relevés (Gauch & Wentworth 1976). Dans le cadre des listes d'occurrences, l'utilisation de l'analyse canonique ne pose plus de contrainte et apparaît même tout à fait adaptée. La diversité de données et des problématiques biologiques a conduit naturellement au développement de nombreuses méthodes. Chaque méthode ayant ses propriétés d'optimalité, ses conditions d'application, il est légitime que « *no single method has emerged as a solution to all problems of describing and explaining patterns of compositional variation in natural communities* » (Noy-Meir & Whittaker 1977). On peut dès lors s'interroger sur l'usage intensif de l'analyse canonique des correspondances dont l'objectif (maximiser la séparation des niches) ne peut satisfaire l'ensemble des études en écologie. Dans ce cadre, l'usage de l'analyse de co-inertie et de ses variantes dont l'analyse de co-inertie procustéenne mériterait certainement d'être encouragé. Il est évident que les développements méthodologiques ajoutent de nouvelles possibilités d'analyse aux utilisateurs et permettent d'améliorer sensiblement la qualité des résultats. Cependant, pour l'écologue, cette diversité de méthodes complexes augmente sensiblement la difficulté de choisir la « bonne » méthode pour une situation donnée d'autant plus que l'utilisation de plusieurs méthodes sur un même jeu de données peut conduire à différents résultats. Dès 1977, Noy-Meir & Whittaker affirmaient que « *The multiplicity of possibilities, among which the user had to choose subjectively, caused some ecologists to wonder how objective the objective methods really were* ». Le rôle du biométricien et l'importance de la consultation statistique sont alors au centre du débat afin d'aider l'écologue à traiter ses données avec des méthodes adaptées à ses besoins.

Un grand nombre de méthodes d'analyses de données a été développé dans différentes disciplines. Certaines méthodes ont donc été découvertes, puis redécouvertes et une même procédure mathématique peut donc avoir différentes appellations selon le domaine d'application. Ainsi l'analyse inter-batterie de Tucker (1958) qui a été étendue en analyse de co-inertie (Dolédec & Chessel 1994) est connue en morphométrie sous le nom de « *two-blocks partial least-squares analysis* » (2B-PLS, Rohlf & Corti 2000). Les travaux de

synthèse méthodologique interdisciplinaire sont alors très utiles afin de faciliter le transfert de connaissances entre domaines d'application et d'identifier d'éventuelles analogies. Dans le domaine des statistiques spatiales, Dale *et al.* (2002) identifient ainsi des liens entre de nombreuses méthodes utilisées en géographie, géologie ou en écologie notamment. Pour réaliser ce type de synthèse, il est nécessaire de posséder un cadre théorique universel dans lequel chaque méthode apparaît comme un cas particulier d'un modèle général. En analyse de données, la théorie du schéma de dualité peut fournir un cadre commun à l'ensemble des méthodes définies par un triplet (**X**, **Q**, **D**). Il est alors possible de comparer les méthodes et identifier certaines similitudes. Ainsi, dans ce travail, nous avons tenté de replacer l'ensemble des méthodes citées dans la théorie du schéma de dualité. Pour se faire, il est impératif que chaque méthode soit définie rigoureusement d'un point de vue mathématique.

Par exemple, les analyses détendancées telles que la « *detrended correspondence analysis* » (DCA, Hill & Gauch 1980) n'ont pas été mentionnées dans ce manuscrit. La DCA est un ajustement *ad hoc* de l'analyse des correspondances qui permet d'éliminer l'effet Guttman (fer à cheval) observé quand on réalise une représentation dans un plan des résultats d'une AFC, le tableau étant structuré uniquement sur un gradient simple. Dans ce cas, le score observé sur le deuxième axe n'est qu'une fonction quadratique du score sur le premier axe. Le principe de la DCA consiste à diviser le premier axe en segments et, pour chaque segment, la moyenne du score sur le deuxième axe est ramenée à 0. Une controverse est rapidement apparue sur l'utilisation de cette méthode en écologie (Jackson & Somers 1991, Peet *et al.* 1988, Wartenberg *et al.* 1987). Les problèmes relatifs à la DCA concernent notamment le nombre de segments à choisir (les résultats peuvent varier fortement selon ce nombre) et le fait que cette méthode soit dénuée de tout fondement théorique bien qu'elle puisse faciliter l'interprétation des résultats écologiques. La DCA entraîne une perte des propriétés fondamentales de l'AFC et ne peut être définie par des critères optimisés. De plus, l'implémentation de cette méthode dans les logiciels CANOCO et DECORANA contenait des bugs induisant une forte instabilité des résultats lorsque les ordres d'entrée des variables et des individus changeait dans le tableau de données (Oksanen & Minchin 1997). Même si les résultats de la DCA peuvent être de meilleure qualité, une méthode doit être basée sur des principes théoriques clairement explicités et reproductibles et non pas sur des « bricolages » dissimulés dans les codes de logiciels. Il nous semble donc impératif qu'une méthode soit clairement et rigoureusement définie d'un point de vue mathématique et il serait souhaitable d'encourager sa distribution sous la forme de logiciels donnant accès au code source afin de pouvoir vérifier son implémentation.

La modélisation des trypanosomoses et l'étude des relations entre la dynamique de population de chevreuils et structure de l'habitat ont été réalisées avec des méthodes statistiques usuelles. La première a nécessité l'usage du modèle linéaire généralisé, de statistiques spatiales simples. Dans la seconde, classification, modèle linéaire et ACP globale (Thioulouse *et al.* 1995) ont été employés. Le SIG a également été utilisé pour la représentation des résultats. Ces deux travaux n'ont pas engendré de développement méthodologique mais l'objectif était d'utiliser des méthodes en adéquation avec les objectifs de l'étude. Par exemple, l'ACP globale a été préférée à une ACP classique pour identifier les principales structures spatiales des communautés végétales à Chizé. L'ACP globale contient une contrainte d'autocorrélation spatiale et la prise en compte de la structure spatiale des données dans cette méthode était parfaitement adaptée à la situation. Dans ces deux cas, il était important d'utiliser des méthodes conformes aux données et aux objectifs afin de se mettre au service de la problématique biologique. Des méthodes déjà existantes apparaissaient en parfaite adéquation avec les situations et il n'y avait donc aucune raison de faire de la surenchère méthodologique. « *Quand on veut enfoncer un clou, on se sert d'un marteau. Le*

marteau est l'outil, le clou à enfoncer l'objectif. Il n'y a aucune ambiguïté. » (Legay 1973). On peut également utiliser une pince pour enfoncer un clou, ce n'est pas très adapté mais, cela peut marcher. On peut aussi acheter un marteau électrique, c'est exagéré pour un simple clou mais cela marchera aussi. Le biométricien essaiera de choisir le marteau.

Ce travail est basé sur une approche pluridisciplinaire. Les interactions entre analyses multivariées, systèmes d'information géographique et observations écologiques ont été très fortes. Par exemple, lorsque les données concernant la distribution des poids des chevreuils à Chizé ont été entrées dans le SIG, la première manipulation a été simplement d'éditer l'ensemble des cartes représentant la distribution annuelle des poids des chevrillards. La visualisation des données à l'aide d'un outil adapté a alors provoqué naturellement la question : « Est-ce que la structure spatiale de la distribution des poids chevrillards est stable dans le temps ? ». Pour répondre à cette question et compte tenu de la structure des données, le développement d'une nouvelle méthode a été entrepris. Par l'intermédiaire du SIG, une nouvelle problématique biologique s'est présentée et sa résolution a entraîné la mise au point d'une nouvelle méthode. L'intérêt d'une approche pluridisciplinaire est alors évident et favorise l'acquisition de connaissances dans chacune des disciplines concernées. Ainsi, ce travail a permis de répondre à des problématiques biologiques, de mettre au point de nouveaux outils et de nouvelles méthodes.

Le choix de se positionner à l'interface de plusieurs disciplines n'est pas anodin. C'est à cet endroit précis que le dialogue peut être le plus fertile. L'écologue sait que dans sa discipline, l'habitude veut qu'on traite les données avec une méthode X. Il ne connaît pas forcément les fondements théoriques. Le mathématicien connaît le cadre théorique de cette méthode mais s'intéresse peu aux applications. Entre les deux, le dialogue est difficile voire impossible. Le rôle du biométricien est alors de favoriser le transfert de connaissances afin d'offrir aux écologues un ensemble d'outils performants mais aussi de vérifier la bonne utilisation qui est faite de ces méthodes et notamment leur adéquation avec les problématiques biologiques. Pour se faire, il est nécessaire de se positionner à l'interface et de comprendre et parler les langages des différentes disciplines. Ce rôle semble primordiale afin d'établir de véritables collaborations scientifiques pluridisciplinaires.

Bibliographie

ANSELIN L. & GETIS A. (1992) Spatial statistical analysis and geographic information systems. Annals of Regional Science 26: 19-33.

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

BAILEY T. C. & GATRELL A. C. (1995) Interactive spatial data analysis. Longman, Harlow. 413 p.

BARBAULT R. (1995) Ecologie générale: structure et fonctionnement de la biosphère. Masson, Paris. 275 p.

BATEK M. J., REBERTUS A. J., SCHROEDER W. A., HAITHCOAT T. L., COMPAS E. & GUYETTE R. P. (1999) Reconstruction of early nineteenth-century vegetation and fire regimes in the Missouri Ozarks. Journal of Biogeography 26: 397-412.

BENZÉCRI J. P. (1969) Statistical analysis as a tool to make patterns emerge from data. Pages 35-60 *dans* **WATANABE S.**, ed. Methodologies of pattern recognition. Academic Press, NewYork.

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

BLANC L. (2000) Données spatio-temporelles en écologie et analyses multitableaux : examen d'une relation. Thèse de doctorat. Université Lyon I, Lyon. 266 p.

BÖCKENHOLT U. & BÖCKENHOLT I. (1990) Canonical analysis of a contingency tables with linear constraints. Psychometrika 55: 633-639.

BOHNING-GAESE K. (1997) Determinants of avian species richness at different spatial scales. Journal of Biogeography 24: 49-60.

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

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

BURROUGH P. A. (2001) GIS and geostatistics: Essential partners for spatial analysis. Environmental and Ecological Statistics 8: 361-377.

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

CARMEL Y. & KADMON R. (1999) Effects of grazing and topography on long-term vegetation changes in a Mediterranean ecosystem in Israel. Plant Ecology 145: 243-254.

CAZES P., CHESSEL D. & DOLÉDEC S. (1988) L'analyse des correspondances internes d'un tableau partitionné: son usage en hydrobiologie. Revue de Statistique Appliquée 36: 39-54.

CHESSEL D. (1992) Echanges interdisciplinaires en analyse des données écologiques. Mémoire d'Habilitation à Diriger des Recherches. Université Lyon I, Lyon. 107 p.

CHESSEL D. & HANAFI M. (1996) Analyse de la co-inertie de K nuages de points. Revue de Statistique Appliquée 44: 35-60.

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. Pages 15-43 *dans* **ASSELAIN B.**, ed. Biométrie et Environnement. Masson, Paris.

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

CLEVELAND W. S. (1993) Visualizing data. Hobart Press, Summit, New Jersey. 360 p.

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.

CLIFF A. D. & ORD J. K. (1973) Spatial autocorrelation. Pion, London. 178 p.

COLE R. G. & SYMS C. (1999) Using spatial pattern analysis to distinguish causes of mortality: an example from kelp in north-eastern New Zealand. Journal of Ecology 87: 936-972.

CORMACK R. M. & ORD J. K. (1979) Spatial and temporal analysis in ecology. International Co-operative Publishing House, Fairland. 356 p.

COULON J., MARCHAL P., PUPIER R., RICHOUX P., ALLEMAND R., GENEST L. C. & CLARY J. (2001) Coléoptères de Rhône-Alpes : carabiques et cicindèles. Muséum d'Histoire naturelle de Lyon et Société linnéenne de Lyon, Lyon. 383 p.

COUTERON P. & KOKOU K. (1997) Woody vegetation spatial patterns in a semi-arid savanna of Burkina Faso, West Africa. Plant Ecology 132: 211-227.

DALE M. R. T. (2000) Spatial pattern analysis in plant ecology. Cambridge University Press, Cambridge. 326 p.

DALE M. R. T., DIXON P., FORTIN M.-J., LEGENDRE P., MYERS D. E. & ROSENBERG M. S. (2002) Conceptual and mathematical relationships among methods for spatial analysis. Ecography 25: 558-577.

DALE M. R. T. & ZBIGNIEWICZ M. W. (1995) The evaluation of multi-species pattern. Journal of Vegetation Science 6: 391-398.

DE LA ROCQUE S., MICHEL J. F., CUISANCE D., DE WISPELAERE G., SOLANO P., AUGUSSEAU X., ARNAUD M. & GUILLOBEZ S. (2001) Le risque trypanosomien : une approche globale pour une décision locale. CIRAD, Montpellier. 151 p.

DI BELLA G. & JONA-LASINIO G. (1996) Including spatial contiguity information in the analysis of multispecific patterns. Environmental and Ecological Statistics 3: 269-280.

DIDIER M. (1991) Utilité et valeur de l'information géographique. Economica, Paris. 255 p.

DING Y. & FOTHERINGHAM A. S. (1992) The integration of spatial analysis and GIS. Computers, Environment and Urban Systems 16: 3-19.

DOLÉDEC S. & CHESSEL D. (1987) Rythmes saisonniers et composantes stationnelles en milieu aquatique I- Description d'un plan d'observations complet par projection de variables. Acta Oecologica - Oecologia Generalis 8: 403-426.

DOLÉDEC S. & CHESSEL D. (1989) Rythmes saisonniers et composantes stationnelles en milieu aquatique II- Prise en compte et élimination d'effets dans un tableau faunistique. Acta Oecologica - Oecologia Generalis 10: 207-232.

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

DOLÉDEC S., CHESSEL D., TER BRAAK C. J. F. & CHAMPELY S. (1996) Matching species traits to environmental variables: a new three-table ordination method. Environmental and Ecological Statistics 3: 143-166.

DRAY S. (1999a) L'ordination des listes d'occurrences : quand l'analyse canonique des correspondances est une analyse canonique. Rapport technique, DEA Analyse et Modélisation des Systèmes Biologiques. Université Lyon I, Lyon. 30 p.

DRAY S. (1999b) Utilisation des listes d'occurrences spécifiques spatialisées en écologie et biogéographie. Rapport bibliographique, DEA Analyse et Modélisation des Systèmes Biologiques. Université Lyon I, Lyon. 30 p.

DRAY S., CHESSEL D. & THIOULOUSE J. (sous presse) Procrustean co-inertia analysis for the linking of multivariate data sets. Ecoscience.

DUNGAN J. L., PERRY J. N., DALE M. R. T., LEGENDRE P., CITRON-POUSTY S., FORTIN M.-J., JAKOMULSKA A., MIRITI M. & ROSENBERG M. S. (2002) A balanced view of scale in spatial statistical analysis. Ecography 25: 626-640.

ECCLES N. S., ESLER K. J. & COWLING R. M. (1999) Spatial pattern analysis in Namaqualand desert plant communities: evidence for general positive interactions. Plant Ecology 142: 71-85.

ESCOUFIER Y. (1973) Le traitement des variables vectorielles. Biometrics 29: 750-760.

ESCOUFIER Y. (1983) Réflexions sur les activités du statisticien universitaire. Statistique et Analyse des Données 8: 76-82.

ESCOUFIER Y. (1987) The duality diagramm : a means of better practical applications. Pages 139-156 *dans* **LEGENDRE L.**, ed. Developments in numerical ecology. Springer Verlag, Berlin.

FISHER R. A. (1940) The precision of discriminant functions. Annals of Eugenics 10: 422-429.

FRANQUET E. & CHESSEL D. (1994) Approche statistique des composantes spatiales et temporelles de la relation faune-milieu. Comptes Rendus de l'Academie des Sciences Serie III - Sciences de la Vie 317: 202-206.

FRANQUET E., DOLÉDEC S. & CHESSEL D. (1995) Using multivariate analyses for separating spatial and temporal effects within species-environment relationships. Hydrobiologia 300/301: 425-431.

GAILLARD J. M., DELORME D., BOUTIN J. M., VAN LAERE G., BOISAUBERT B. & PRADEL R. (1993) Roe deer survival patterns: a comparative analysis of contrasting populations. Journal of Animal Ecology 62: 778-791.

GAUCH H. G. (1982) Multivariate analysis in community ecology. Cambridge University Press, Cambridge. 298 p.

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

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 I, Lyon. 235 p.

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

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. SpringerVerlag, Berlin. 351 p.

GOOD I. J. (1969) Some applications of the singular decomposition of a matrix. Technometrics 11: 823-831.

GOODALL D. W. (1954) Objective methods for the classification of vegetation III. An essay on the use of factor analysis. Australian Journal of Botany 2: 304-324.

GOODALL D. W. (1974) A new method for the analysis of spatial pattern by random pairing of quadrats. Vegetatio 29: 135-146.

GOODCHILD M., HAINING R. & WISE S. (1992) Integrating GIS and spatial data analysis: problems and possibilities. International Journal of Geographical Information Systems 6: 407-423.

GOWER J. C. (1971) Statistical methods of comparing different multivariate analyses of the same data. Pages 138-149 *dans* **TAUTU P.**, ed. Mathematics in the archaeological and historical sciences. Edinburgh University Press, Edinburgh.

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.

GREENACRE M. J. (1984) Theory and Applications of Correspondence Analysis. Academic Press, London. 364 p.

GREIG-SMITH P. (1952) The use of random and contiguous quadrats in the study of the structure of plant communities. Annals of Botany 16: 293-316.

GUISAN A., WEISS S. B. & WEISS A. D. (1999) GLM versus CCA spatial modeling of plant species distribution. Plant Ecology 143: 107-122.

GUISAN A. & ZIMMERMANN N. E. (2000) Predictive habitat distribution models in ecology. Ecological Modelling 135: 147-186.

GUTTMAN L. (1941) The quantification of a class of attributes: A theory and method of scale construction. Pages 319-348 *dans* **HORST P.**, ed. The prediction of personal Adjustment. Social Science Research Council, New-York.

HARVILLE D. A. (1997) Matrix algebra from a statistician's perspective. Springer-Verlag, New-York. 630 p.

HAWKSWORTH D. L. (1995) The Resource Base for Biodiversity Assessments. Pages 545-605 *dans* **HEYWOOD V. H.**, ed. Global Biodiversity Assessment. Cambridge University Press, Cambridge.

HAYASHI C. (1950) On the quantification of qualitative data from the mathematico-statistical point of view. Annals of the Institute of Statistical Mathematics 2: 35-47.

HILL M. O. (1973a) The intensity of spatial pattern in plant communities. Journal of Ecology 61: 225-235.

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

HILL M. O. (1974) Correspondence analysis : A neglected multivariate method. Applied Statistics - Journal of the Royal Statistical Society Series C 23: 340-354.

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

HIRSCHFELD H. O. (1935) A connection between correlation and contingency. Proceedings of the Cambridge Philosophical Society. Mathematical and Physical Sciences 31: 520-524.

HIRZEL A. (2001) When GIS come to life. Linking landscape-and population ecology for large population management modelling: the case of Ibex (Capra Ibex) in Switzerland. Thèse de doctorat. Faculté des Sciences de l'Université de Lausanne, Lausanne. 106 p.

HORNIK K. (2002) The R FAQ. http://www.ci.tuwien.ac.at/~hornik/R/.

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

IHAKA R. & GENTLEMAN R. (1996) R: A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics 5: 299-314.

JACKSON D. A. (1995) PROTEST: A PROcrustean Randomization TEST of community environment concordance. Ecoscience 2: 297-303.

JACKSON D. A. & SOMERS K. M. (1991) Putting things in order: the ups and downs of detrended correspondence analysis. American Naturalist 137: 704-712.

JENKS G. F. & CASPALL F. C. (1971) Error on choroplethic maps: definition, measurement, reduction. Annals of the Association of American Geographers 61: 217-244.

JOHNSON L. B. (1990) Analyzing spatial and temporal phenomena using geographical information system. Landscape Ecology 4: 31-43.

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

KADMON R. & DANIN A. (1999) Distribution of plant species in Israel in relation to spatial variation in rainfall. Journal of Vegetation Science 10: 421-432.

KADMON R. & HELLER J. (1998) Modelling faunal responses to climatic gradients with GIS: land snails as a case study. Journal of Biogeography 25: 527-539.

KROONENBERG P. M. & LOMBARDO R. (1999) Nonsymmetric correspondence analysis: a tool for analysing contingency tables with a dependence structure. Multivariate Behavioral Research 34: 367-396.

LAFOSSE R. & HANAFI M. (1997) Concordance d'un tableau avec K tableaux : définition de K+1 uples synthétiques. Revue de Statistique Appliquée 45: 111-126.

LAURO N. & D'AMBRA L. (1984) L'analyse non symétrique des correspondances. Pages 433-446 *dans* **TOMASSONE R.**, ed. Data Analysis and Informatics III. Elsevier, North-Holland.

LAVOREL S., ROCHETTE C. & LEBRETON J. D. (1999) Functional groups for response to disturbance in Mediterranean old fields. Oikos 84: 480-498.

LAVOREL S., TOUZARD B., LEBRETON J. D. & CLÉMENT B. (1998) Identifying functional groups for response to disturbance in an abandoned pasture. Acta Oecologica - International Journal of Ecology 19: 227-240.

LEBART L., MORINEAU A. & TABART N. (1977) Techniques de la description statistique, méthodes et logiciels pour la description des grands tableaux. Dunod, Paris. 351 p.

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

LEBRETON J. D., CHESSEL D., RICHARDOT-COULET M. & YOCCOZ N. (1988b) L'analyse des relations espèces-milieu par l'analyse canonique des correspondances. II. Variables de milieu qualitatives. Acta Oecologica - Oecologia Generalis 9: 137-151.

LEBRETON J. D., SABATIER R., BANCO G. & BACOU A. M. (1991) Principal component and correspondence analyses with respect to instrumental variables : an overview of their role in studies of structure-activity and species-environment relationships. Pages 85-114 *dans* **KARCHER W.**, ed. Applied Multivariate Analysis in SAR and Environmental Studies. Kluwer Academic Publishers.

LEGAY J. M. (1973) La méthode des modèles, état actuel de la méthode expérimentale. Informatique et biosphère, Paris. 69 p.

LEGAY J. M. (1976) Pour une Biométrie. Statistique et Analyse des Donnees 1: 5-11.

LEGAY J. M. (1980) La bio-informatique. Pages 288-291 *dans* Encyclopaedia Universalis.

LEGAY J. M. (1984) Sur les relations Biométrie-Ecologie. Bulletin d'Ecologie 15: 117-119.

LEGENDRE P. & FORTIN M.-J. (1989) Spatial pattern and ecological analysis. Vegetatio 80: 107-138.

MARAVELIAS C. D., REID D. G., SIMMONDS E. J. & HARALABOUS J. (1996) Spatial analysis and mapping of acoustic surey data in the presence of high local

variability: geostatistical application to North Sea herring (Clupea harengus). Canadian Journal of Fisheries and Aquatic Sciences 53: 1497-1505.

MÉOT A. (1992) Explicitation de contraintes de voisinage en analyse multivariée. Thèse de doctorat. Université Lyon I, Lyon. 192 p.

MÉOT A., LEGENDRE P. & BORCARD D. (1998) Partialling out the spatial component of ecological variation: questions and propositions in the linear modelling framework. Environmental and Ecological Statistics 5: 1-27.

MICHEL J. F., MICHEL V., DE LA ROCQUE S., TOURÉ I. & RICHARD D. (1999) Modélisation de l'occupation de l'espace par les bovins. Applications à l'épidémiologie des trypanosomoses animales. Revue d'élevage et de médecine vétérinaire des pays tropicaux 52: 297-301.

MYERS N. (1990) The biodiversity challenge: expanded hot-spots analysis. The Environmentalist 10: 243-256.

NISHISATO S. (1980) Analysis of Categorical Data: Dual Scaling and its Applications. University of Toronto Press, London. 276 p.

NOY-MEIR I. (1973) Data transformation in ecological ordination. I. some advantages of non-centring. Journal of Ecology 61: 329-341.

NOY-MEIR I. & ANDERSON D. J. (1971) Multivariate pattern analysis, or multiscale ordination: towards a vegetation hologram ? Pages 207-232 *dans* **WATERS W. E.**, ed. Statistical Ecology III. Populations, ecosystems, and systems analysis. Pennsylvania State University Press.

NOY-MEIR I., WALKER D. & WILLIAMS W. T. (1975) Data transformation in ecological ordination. II. On the meaning of data standardization. Journal of Ecology 63: 779-800.

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

OBADIA J. (1978) L'analyse en composantes explicatives. Revue de Statistique Appliquée 26: 5-28.

OHMANN J. L. & SPIES T. A. (1998) Regional gradient analysis and spatial pattern of woody plant communities of Oregon forests. Ecological Monographs 68: 151-182.

OJEDA F., ARROYO J. & MARAÑON T. (1998) The phytogeography of European and Mediterranean heath species (Ericoideae, Ericaceae): a quantitative analysis. Journal of Biogeography 25: 165-178.

OKSANEN J. & MINCHIN P. R. (1997) Instability of ordination results under changes in input data order: explanations and remedies. Journal of Vegetation Science 8: 447-454.

PALMER M. W. (1993) Putting things in even better order: the advantages of canonical correspondence analysis. Ecology 74: 2215-2230.

PALMER M. W. (2002) Computer program reviews, a new feature in the Journal of Vegetation Science. Journal of Vegetation Science 13: 291.

PEET R. K., KNOX R. G., CASE J. S. & ALLEN R. B. (1988) Putting things in order: the advantages of detrended correspondence analysis. American Naturalist 131: 924-934.

PÉLISSIER R. (1995) Relations entre l'hétérogénéité spatiale et la dynamique de renouvellement d'une forêt dense humide sempervirente (Foret d'Uppangala - Ghâts Occidentaux de l'Inde). Thèse de doctorat. Université Lyon I, Lyon. 236 p.

PETTORELLI N., GAILLARD J. M., DUNCAN P., OUELLET J. P. & VAN LAERE G. (2001) Population density and small-scale variation in habitat quality affect phenotypic quality in roe deer. Oecologia 128: 400-405.

POIDEVIN D. (1999) La carte moyen d'action : conception - réalisation. Ellipse, Paris. 200 p.

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.

RANJIT DANIELS R. J. (1992) Geographical distribution patterns of amphibians in the Western Ghats, India. Journal of Biogeography 19: 521-529.

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

RICKLEFS R. E. (1990) Ecology. W.H. Freeman and Company, New York. 896 p.

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 Series B-Methodological 39: 172-212.

ROA R. & TAPIA F. (2000) Cohorts in space: geostatistical mapping of the age structure of the squat lobster Pleuroncodes monodon population off central Chile. Marine Ecology - Progress Series 196: 239-251.

ROHLF F. J. & CORTI M. (2000) Use of two-block partial least-squares to study covariation in shape. Systematic Biology 49: 740-753.

RUEDA M. & DEFEO O. (2001) Survey abundance indices in a tropical estuarine lagoon and their management implications: a spatially explicit approach. ICES Journal of Marine Science 58: 1219-1231.

SABATIER D., GRIMALDI M., PRÉVOST M. F., GUILLAUME J., GODRON M., DOSSO M. & CURMI P. (1997) The influence of soil cover organization on the floristic and structural heterogeneity of a Guianan rain forest. Plant Ecology 131: 81-108.

SABATIER R., LEBRETON J. D. & CHESSEL D. (1989) Principal component analysis with instrumental variables as a tool for modelling composition data. Pages 341-352 *dans* **BOLASCO S.**, ed. Multiway data analysis. Elsevier Science Publishers B.V., North-Holland.

SCHAEFER J. A. & MESSIER F. (1994) A paired-quadrat method for use in multiscale ordination. Vegetatio 113: 9-11.

SIMIER M., BLANC L., PELLEGRIN F. & NANDRIS D. (1999) Approche simultanée de K couples de tableaux : Application à l'étude des relations pathologie végétale-environment. Revue de Statistique Appliquée 47: 31-46.

SKALOVA H., KRAHULEC F., DURING H. J., HADINCOVA V., PECHACKOVA S. & HERBEN T. (1999) Grassland canopy composition and spatial heterogeneity in the light quality. Plant Ecology 143: 129-139.

SMITH E. P. (2002) Ecological statistics. Pages 589-602 *dans* **PIERGORSCH W. W.**, ed. Encyclopedia of Environmetrics. John Wiley and Sons, Chichester.

STEWART D. K. & LOVE W. A. (1968) A general canonical correlation index. Psychological Bulletin 70: 160-163.

TAKANE Y., YANAI H. & MAYEKAWA S. (1991) Relationships among several methods of linearly constrained correspondence analysis. Psychometrika 56: 667-684.

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. 458 p.

TENENHAUS M. (1998) La régression PLS. Théorie et Pratique. Editions Technip, Paris. 252 p.

TENENHAUS M. & YOUNG F. W. (1985) An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. Psychometrika 50: 91-119.

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

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

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

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

THIOULOUSE J. (1996) Outils logiciels, méthodes statistiques et implications biologiques : une approche de la biométrie. Mémoire d'Habilitation à Diriger des Recherches. Université Lyon I, Lyon. 106 p.

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.

THIOULOUSE J., CHESSEL D., DOLÉDEC S. & OLIVIER J. M. (1997) ADE-4: a multivariate analysis and graphical display software. Statistics and Computing 7: 75-83.

TUCKER L. R. (1958) An inter-battery method of factor analysis. Psychometrika 23: 111-136.

VER HOEF J. M. & GLENN-LEWIN D. C. (1989) Multiscale ordination: a method for detecting pattern at several scale. Vegetatio 82: 59-67.

VER HOEF J. M., GLENN-LEWIN D. C. & WERGER M. J. A. (1989) Relationship between pattern and vertical structure in a chalk grassland. Vegetatio 83: 147-155.

WARTENBERG D. (1985a) Canonical trend surface analysis: a method for descibing geographic pattern. Systematic Zoology 34: 259-279.

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

WARTENBERG D., FERSON S. & OHLF F. J. (1987) Putting things in order: a critique of detrended correspondence analysis. American Naturalist 129: 434-448.

WATT A. S. (1947) Pattern and process in the plant community. Journal of Ecology 35: 1-22.

WHITTAKER R. H. (1972) Evolution and measurement of species diversity. Taxon 21: 213-251.

WIENS J. A. (1989) Spatial scaling in ecology. Functional Ecology 3: 385-397.

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: 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 I, Lyon. 254 p.

YOM-TOV Y. & KADMON R. (1998) Analysis of the distribution of insectivorous bats in Israel. Diversity and Distributions 4: 63-70.

ANNEXE 1

User's manual to CA-richness and NSCA-Simpson strategies

Summary

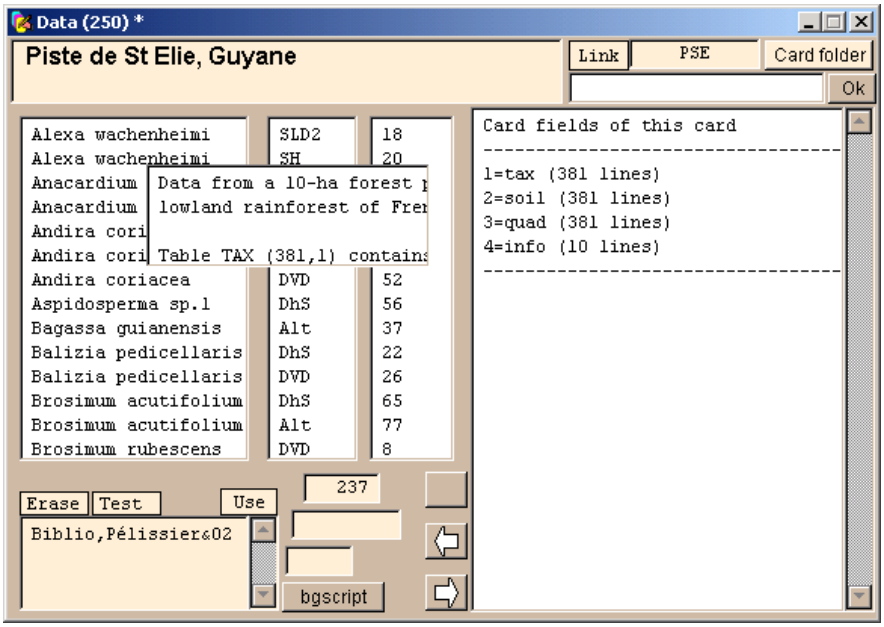
This volume is a *user's manual* describing how the CA-richness and NSCA-Simpson strategies of Pélissier et al. (2002, Consistency between ordination techniques and diversity measurements: two alternative strategies for species occurrences data. Ecology X, ppp-ppp) can be performed using ADE-4 modules. The treated data set is the same as in the original paper and comes from a 10-ha rainforest plot at Piste de St-Elie station in French Guiana.

Contents

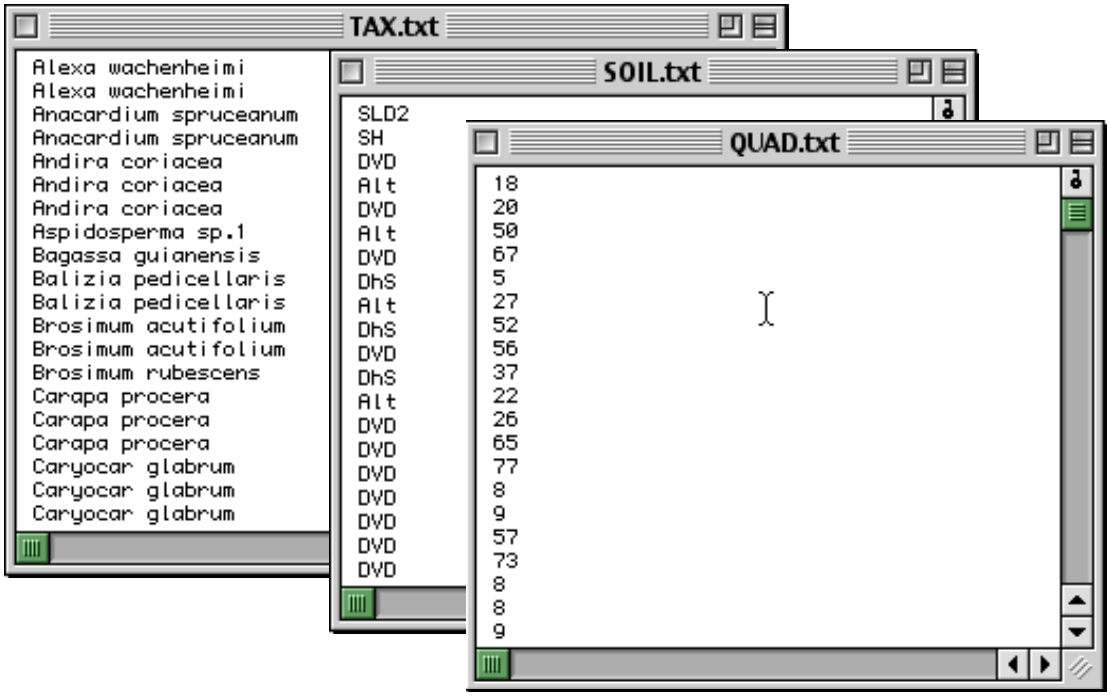
R. Pélissier, S. Dray & P. Couteron

1 — Data input

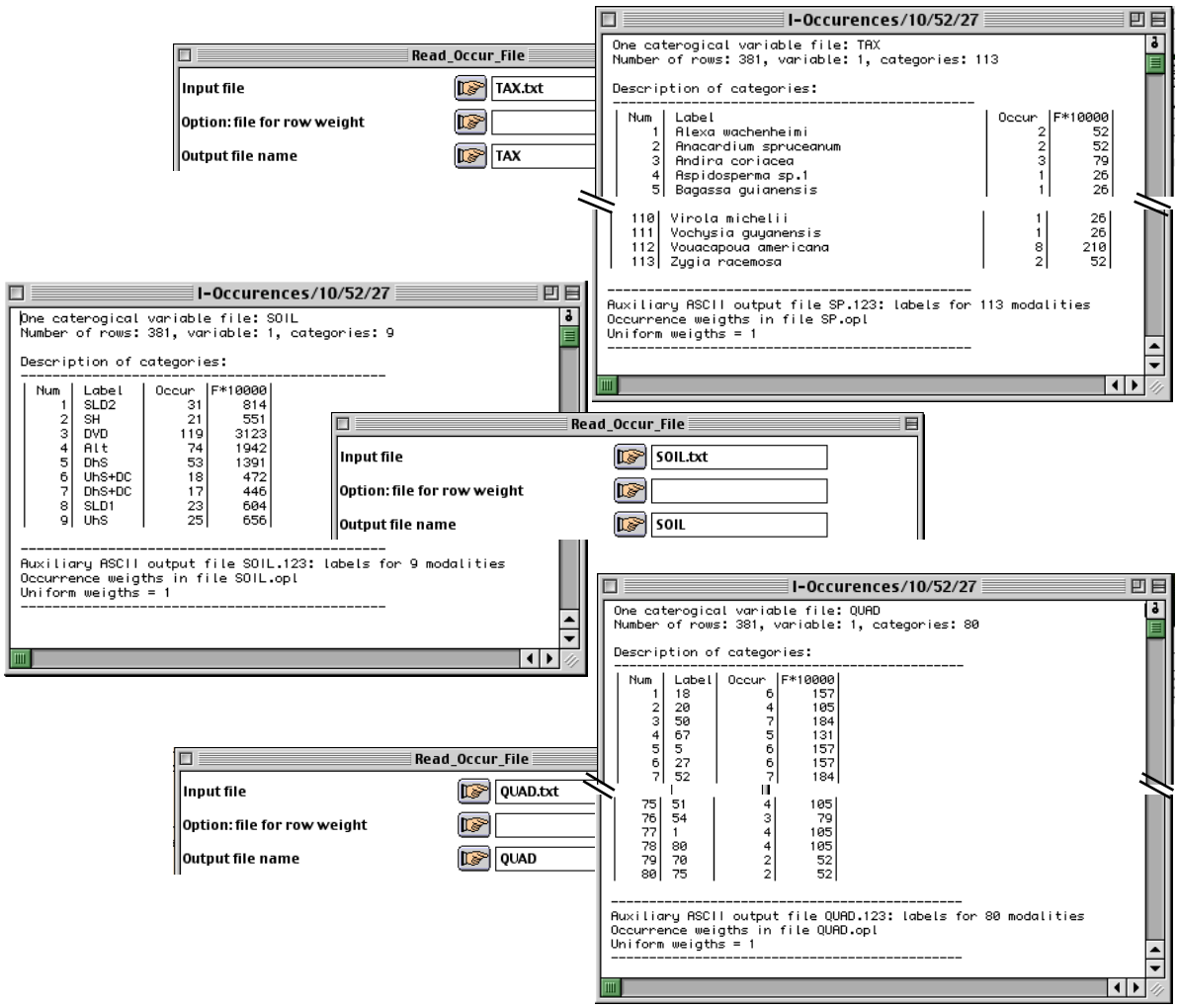
Data are stored in card PSE of ADE-4 data stack. Create a data folder from card PSE:



It contains three text files with 381 rows each, corresponding to the trees with d.b.h. ≥ 50 cm recorded within the 10-ha forest plot at Piste de St-Elie station (taken from transect B in Sabatier et al. 1997). File **TAX.txt** contains the species names; files **SOIL.txt** and **QUAD.txt** contain respectively the codes for the soil classes and for the 50 m x 25 m quadrats to which the trees were allocated.

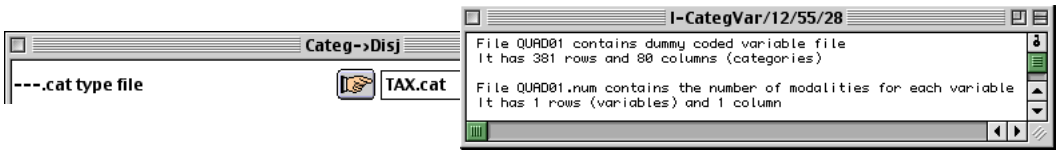


Use module OccurData : Read_Occur_File to create from **TAX.txt**, **SOIL.txt** and **QUAD.txt** specific ADE-4 files for qualitative variables:



It can be checked from the listings that the data set encompass 113 species, 9 soil classes and 80 quadrats.

Use module <u>CategVar : Categ->Disj</u> to create a complete binary species occurrence table from ADE-4 qualitative file **TAX.cat**:



Files **TAX01** corresponds to table **T** (381 occurrences x 113 species) of Pélissier et al. (2002).

2 — Creation of the statistical triplets

ADE-4 programs are designed to deal with statistical triplets (Escoufier 1987). In practice, this means that all analyses are launched from a data matrix identified through a specific file extension (**--.--ta**), automatically associated with two vectors of row (**--.--pl**) and column (**--.--pc**) weights.

Use module COA : NSCA_Row_Profiles to perform Non-Symmetric Correspondence Analysis on the rows of **T**:

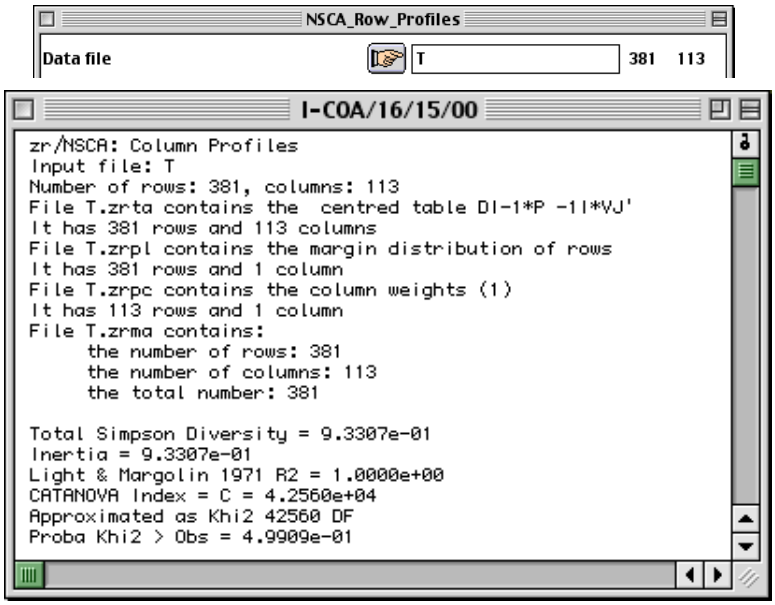


Table **T.zrta** is exactly the centred-by-species occurrence table **Tc** of Pélissier et al. (2002), which is associated with a vector of row weights, **T.zrpl**, and a vector of column weights, **T.zrpc**. Total inertia of this analysis corresponds to Simpson index of diversity ($D = 0.9331$). As table **T** is a species occurrence table (i.e. a complete binary table), **T.zrta**, **T.zrpl** and **T.zrpc** define the basic statistical triplet of the NSCA-Simpson strategy: (**Tc**, $\mathbf{I}_p$, $\mathbf{D}_n$) in Pélissier et al. (2002).

Use module COA : Correspondence Analysis to perform classical CA of **T**:

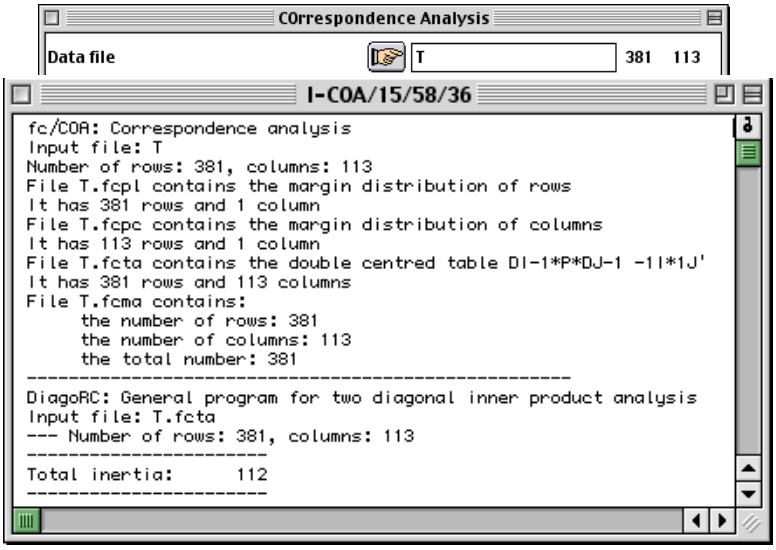
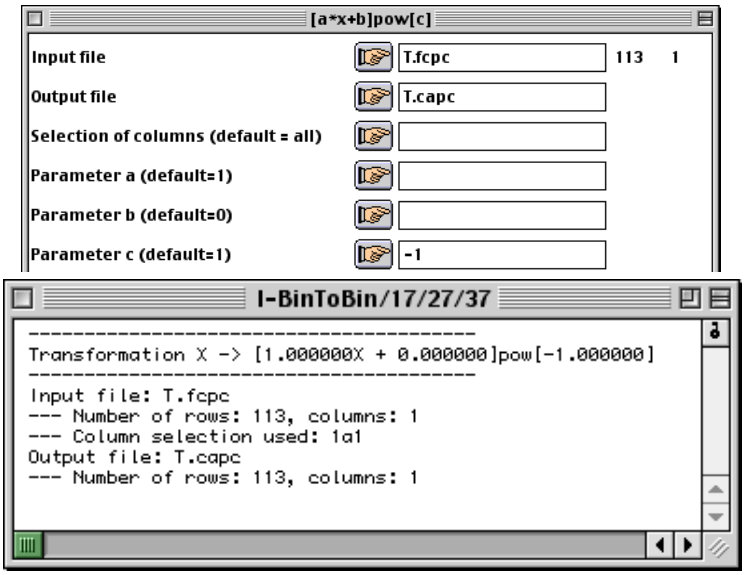
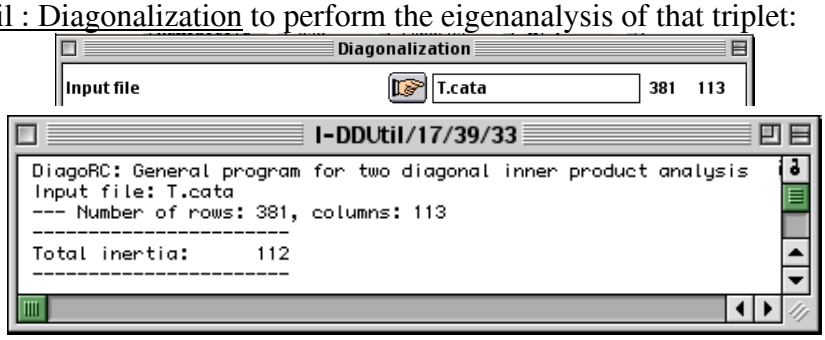


Table **T.fcta** is the classical double centred CA matrix associated with the two weighting vectors **T.fcpl** and **T.fcpc**, which contain the marginal row and column frequencies of **T**. Notice that total inertia of this analysis corresponds to $p - 1 = 112$, where p is the species richness (113 species).

By analogy with the NSCA-Simpson strategy, defined above as the eigenanalysis of (**Tc**, $\mathbf{I}_p$, $\mathbf{D}_n$), the CA-richness strategy can be defined in a non-standard way, from the eigenanalysis of the basic statistical triplet (**Tc**, $\mathbf{D}_p^{-1}$, $\mathbf{D}_n$) (cf. Pélissier et al. 2002). The centred-by-species occurrence table and its associated vector of row weights are thus simply obtained by duplicating the NSCA-files, **T.zrta** and **T.zrpl**, subsequently renamed as **T.cata** and **T.capl**. The corresponding vector of column weights is obtained by computing the inverse of the marginal column frequencies of **T** from file **T.fcpc** and module Bin-Bin : [a*x+b]pow[c], with a = 1, b = 0 and c = -1:



One can check that the CA-richness strategy computed from table **T.cata** associated to the weighting vectors **T.capl** and **T.capc**, has the same total inertia as the classical CA of **T**. Use module DDUtil : Diagonalization to perform the eigenanalysis of that triplet:



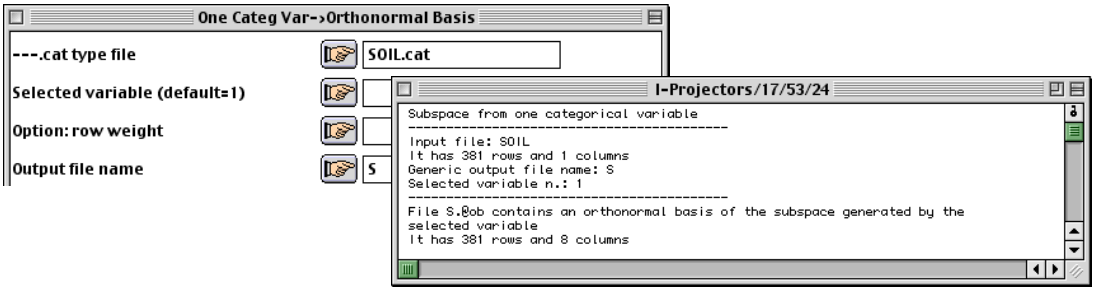
In the following, the NSCA-Simpson strategy will be considered from the basic triplet **T.zrta**, **T.zrpl**, **T.zrpc**, and the CA-richness strategy from the basic triplet **T.cata**, **T.capl**, **T.capc**, keeping in mind that **T.zrta** = **T.cata** and **T.zrpl** = **T.capl**.

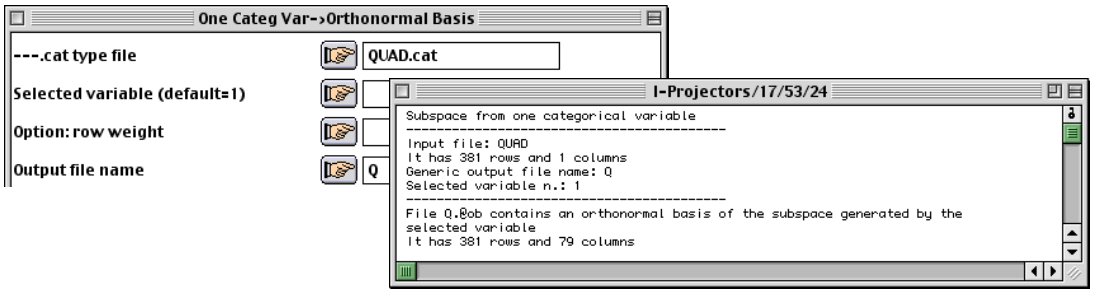
3 — Inertia decomposition

Inertia decomposition of table **Tc** with respect to external variables, representing either the soil classes or the 50 m x 25 m quadrats, is performed using the ADE-4 module Projectors : Triplet Inertia Decomposition. This procedure realizes i) a multivariate linear regression seen as the projection of the dependent variables (in table **Tc**) onto an orthonormal coordinate system defined by the explanatory variables, and ii) computes explained and residual

variances from weighted totals of column norms. Prior to the regression, however, the initial explanatory table have to be appropriately transformed in order to exhibit column variances equal to 1 and covariances equal to 0 in pairs (to avoid multicollinearity).

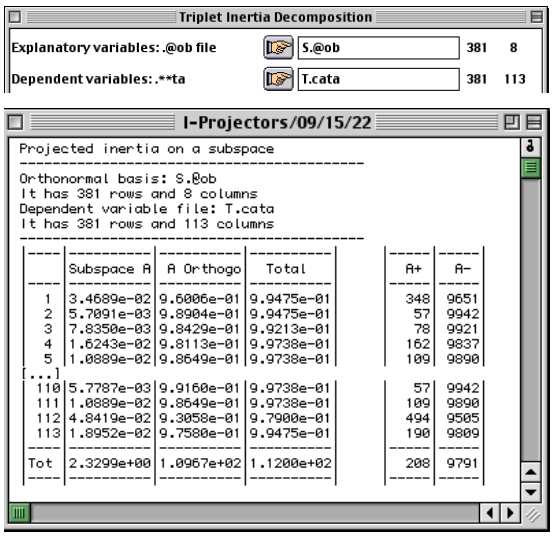
In ADE-4, this transformation is realized by Gram-Schmidt's orthonormalisation (Harville 1997, pp 63-66): the projection of the initial explanatory table onto an orthonormal coordinate system is launch from module Projectors : Table->Orthonormal Basis when the intial table contains quantitative variables (if the variables are uncorrelated the procedure simply normalizes each column) or module Projectors : One Categ Var->Orthonormal Basis when the initial table contains qualitative (dummy) variables. In this last case, ADE-4 qualitative files (**SOIL.cat** and **QUAD.cat** in the present example) can be used directely as input files:

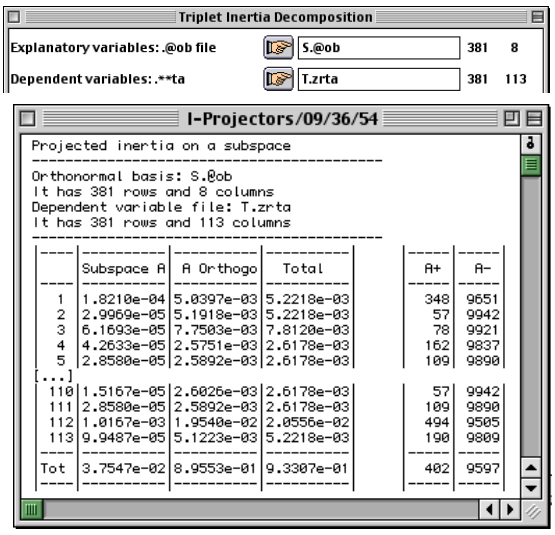




Use Projectors : Triplet Inertia Decomposition to compute explained and residual column norms and their weighted totals according to CA-richness (on left) and NSCA-Simpson (on right) strategies:

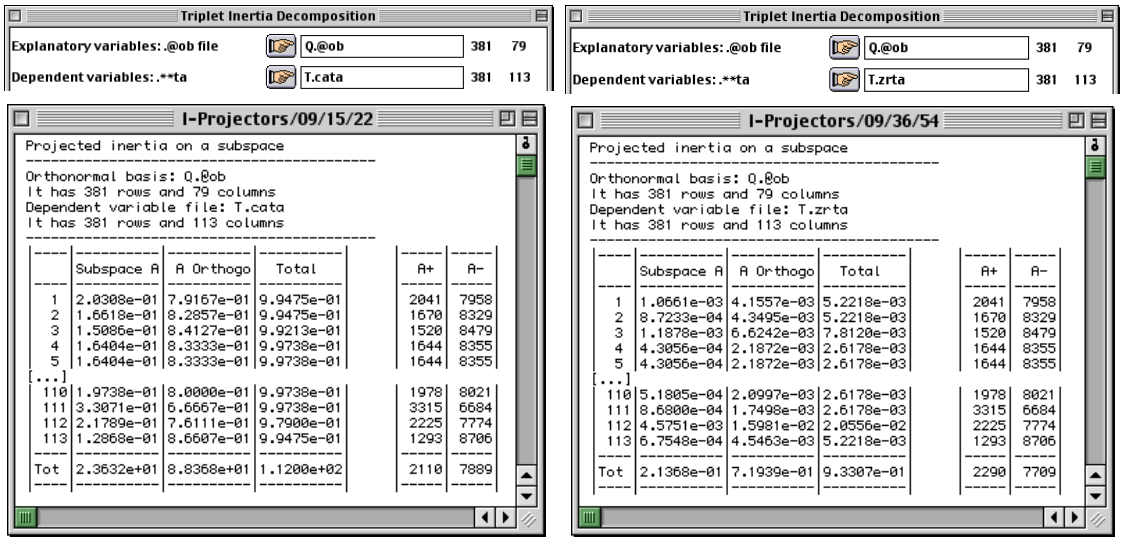
a) with respect to the soil classes (**S.@ob**):





These results correspond to the ones given in Table 2 of Pélissier et al. (2002).

b) with respect to the partition into quadrats of 50 m x 25 m (**Q.@ob**):

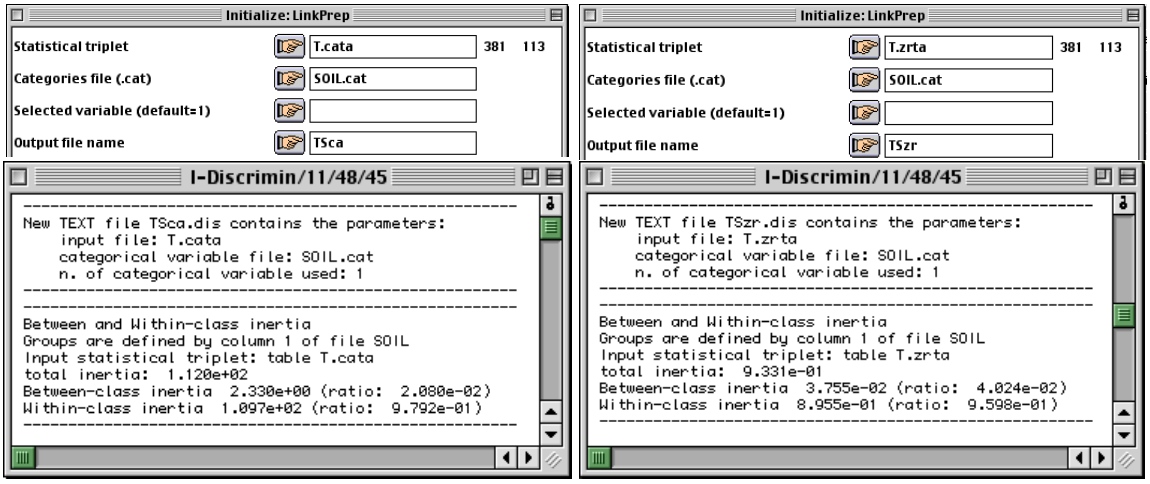


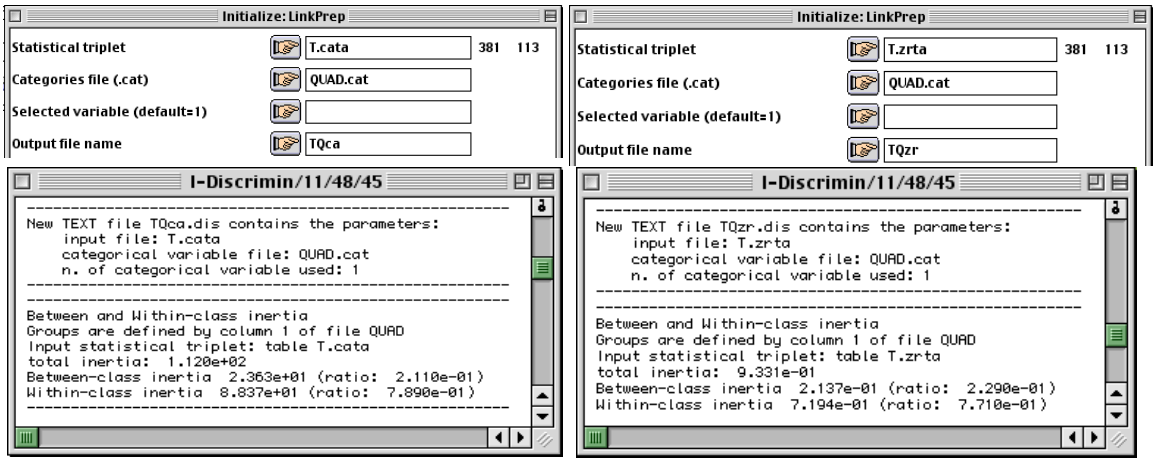
These results correspond to the ones given in Table 3 of Pélissier et al. (2002).

Notice that the proportion of explained inertia for a given species (A+) is exactly the same from the CA-richness and NSCA-Simpson strategies, and that only their relative importance, and thus the weighted totals (or total inertia) differ between the two strategies.

Statistical significance of inertia explained by the explanatory variables can be tested by row permutation tests (Manly 1991). Unfortunately, module Projectors : Subspace/Test cannot run from non-standard file extensions defining the CA-richness and NSCA-Simpson triplets, but module Discrimin : Between analysis/Test can be used instead.

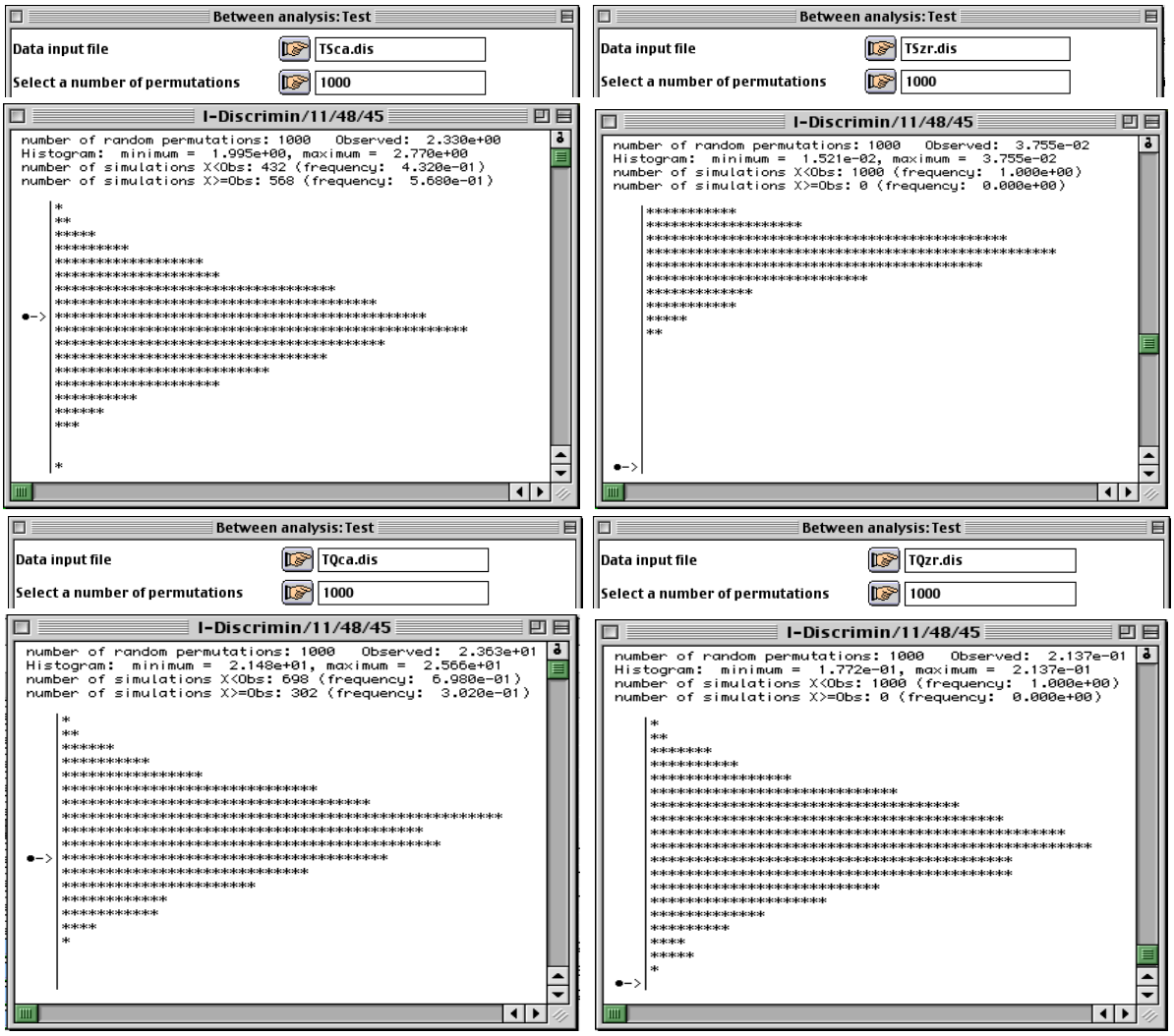
Use Discrimin : Initialize/LinkPrep to prepare the files:





One can check from the listings that between- and within-class analyses (as defined by Dolédec and Chessel 1989) give the same inertia decomposition than CA-richness and NSCA-Simpson strategies viewed as analyses on instrumental variables run either from **T.cata** or **T.zrta**.

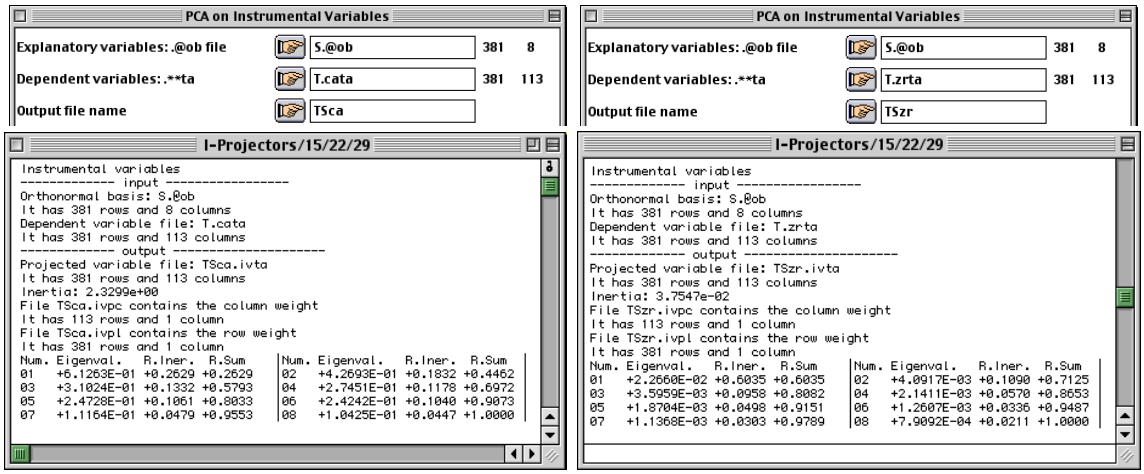
Then, run row permutation tests using Discrimin : Between analysis/Test:



These results are given in Tables 2 and 3 of Pélissier et al. (2002): diversity explained by the soil classes as well as diversity explained by the partition into quadrats are not statistically significant using the CA-richness strategy ($P > 0.3$), but highly statistically significant using the NSCA-Simpson strategy ($P < 0.001$).

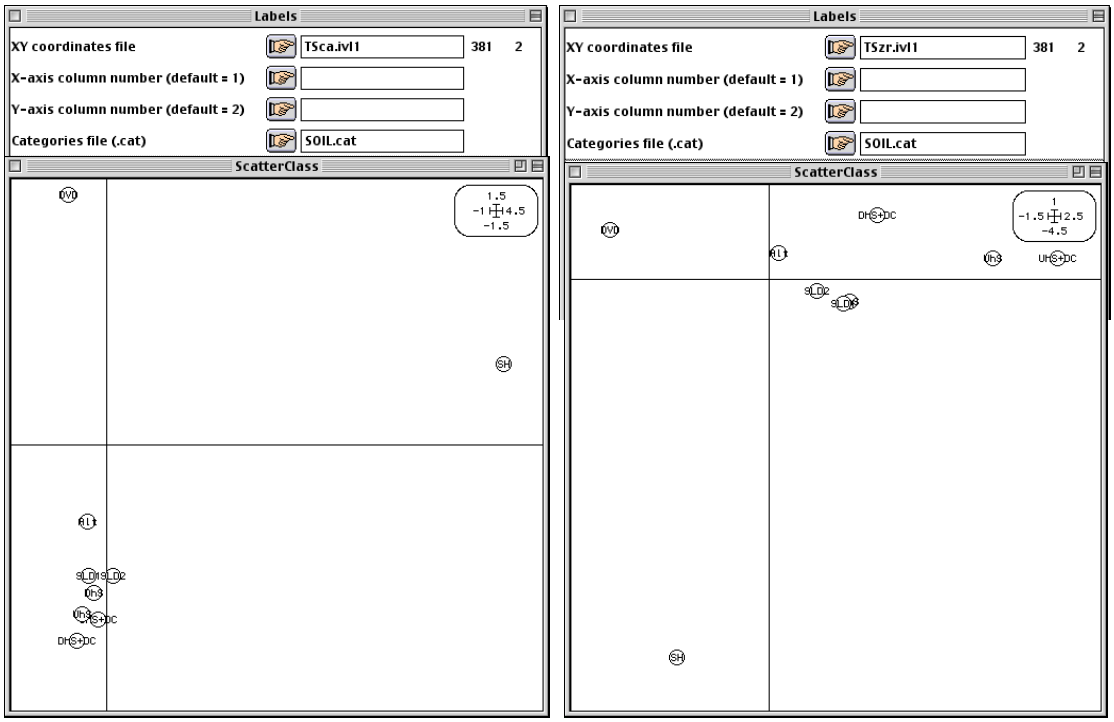
4 — CA *vs.* NSCA on Instrumental Variables

Use module Projectors : PCA on Instrumental Variables to perform CAIV (on left) and NSCAIV (on right) with respect to the soil classes:

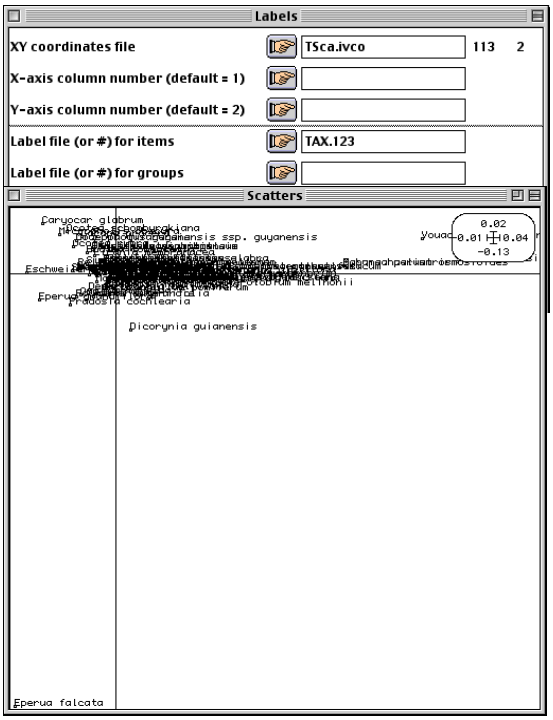


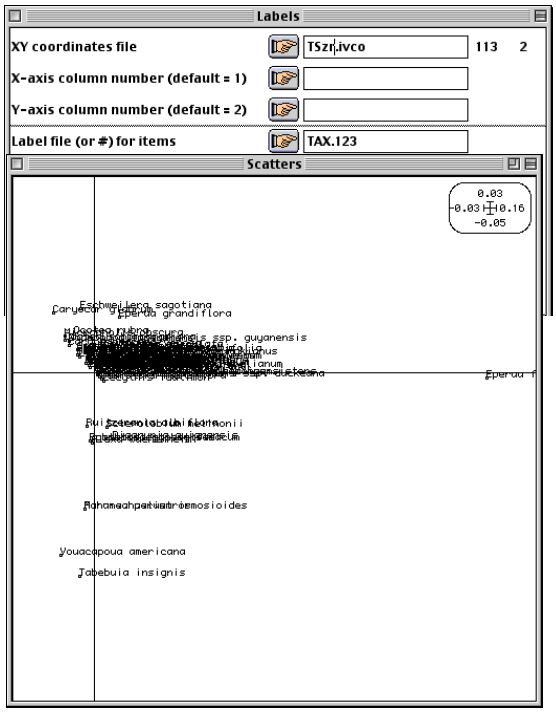
The two first axis account for 26.3% and 18.3% of total inertia of the projected table from the CA-richness strategy, and for 60.3% and 10.9% from the NSCA-Simpson strategy.

Use module ScatterClass : Labels to display the soil classes at the weighted mean position of their occurrences:



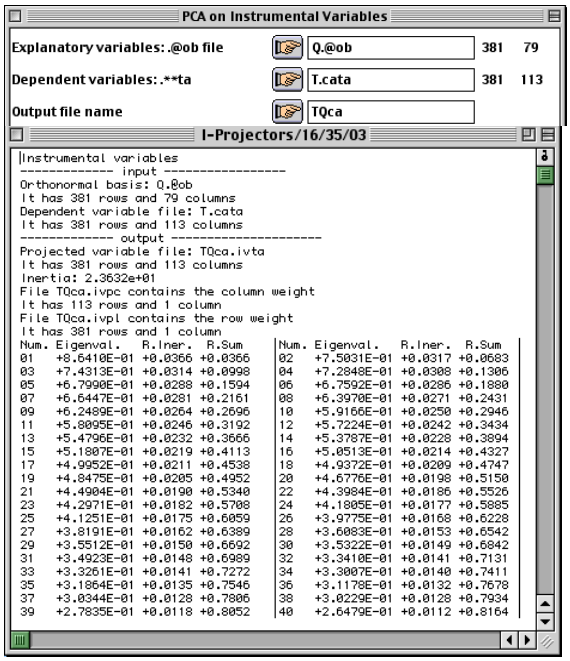
Use module Scatters : Labels to display the position of species:

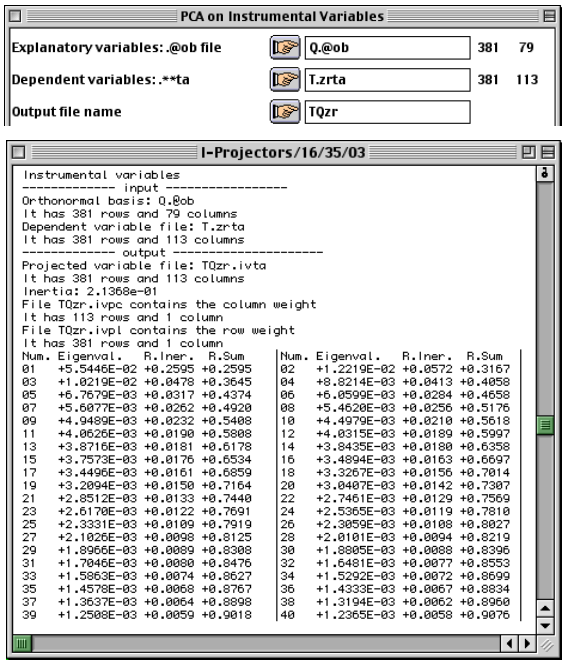




These figures correspond to Figures 2a-b in Pélissier et al. (2002), showing the reversed hierarchy of the main axes obtained from CA-richness strategy and NSCA-Simpson strategy.

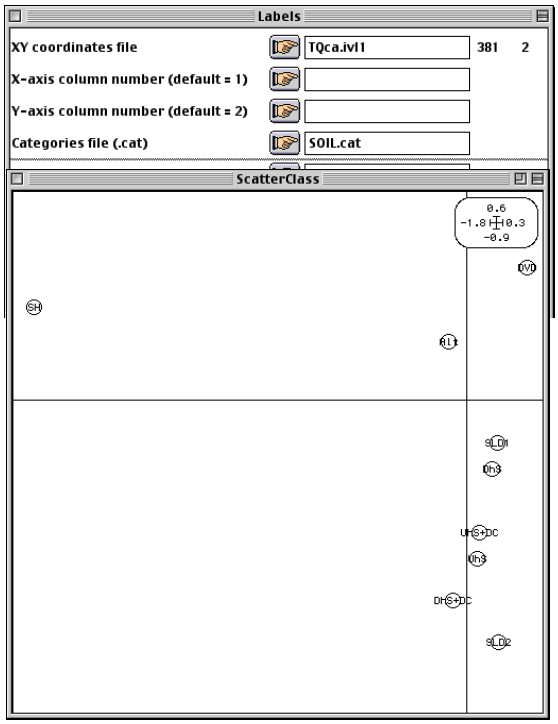
Use similarly module Projectors : PCA on Instrumental Variables to perform CAIV (on left) and NSCAIV (on right) with respect to the partition into 50 m x 25 m quadrats:

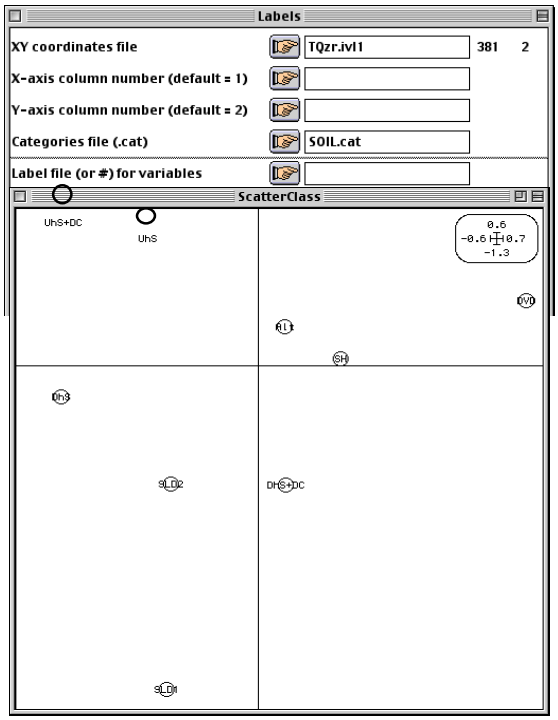




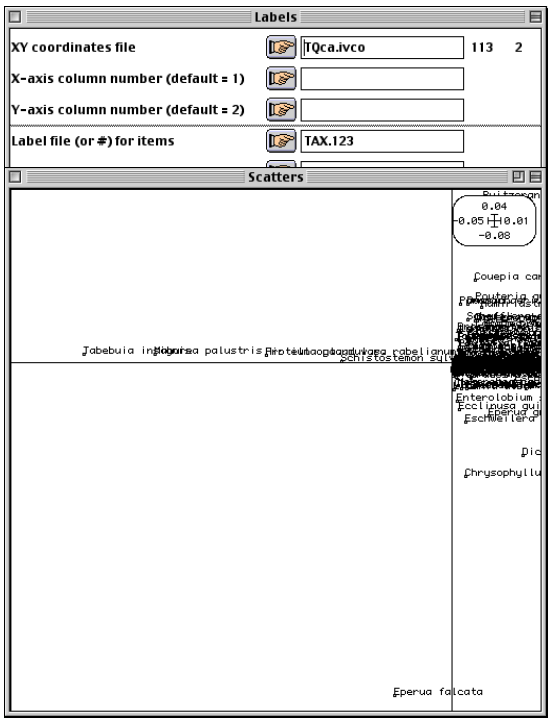
The two first axis account for 3.7% and 3.2% of total inertia of the projected table from the CA-richness strategy, and for 25.9% and 5.7% from the NSCA-Simpson strategy.

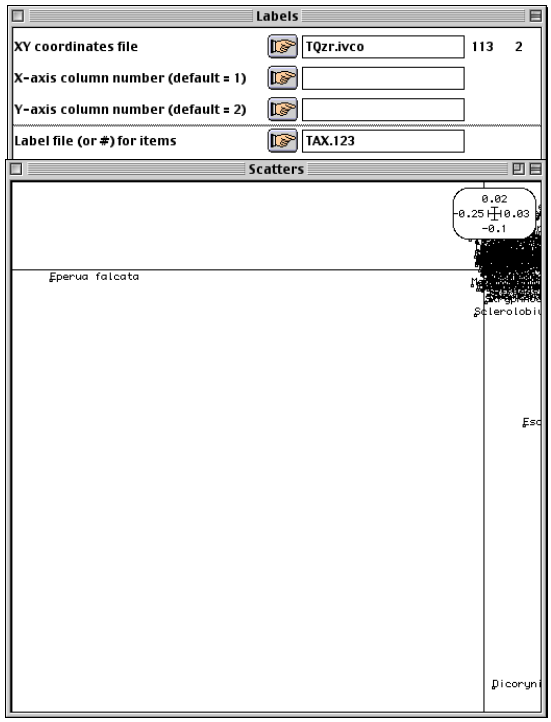
Use module <u>ScatterClass : Labels</u> to display the soil classes at the weighted mean position of their occurrences:





Use module <u>Scatters : Labels</u> to display the position of species:





These figures correspond to Figures 4a-b in Pélissier et al. (2002).

Literature cited

Dolédec, S., and D. Chessel. 1989. Rythmes saisonniers et composantes stationnelles en milieu aquatique. II-Prise en compte et élimination d'effets dans un tableau faunistique. Acta Oecologica **10**: 207–232.Escoufier 1987

Escoufier, Y. 1987. The duality diagram: a means of better practical applications. Pages 139–156 *in* P. Legendre. and L. Legendre, editors. Development in numerical ecology, Springer-Verlag, Berlin, Germany.

Harville, D. A. 1997. Matrix Algebra From a Statistician's Perspective. New York: Springer-Verlag, New York.

Manly, B. J. F. 1991. Randomization and Monte Carlo methods in biology. Chapman and Hall, London, UK.

Pélissier, R., P. Couteron, S. Dray, and S. Sabatier. 2002. Consistency between ordination techniques and diversity measurements: two alternative strategies for species occurrence data. Ecology: in press.

Sabatier, D., M. Grimaldi, M.-F. Prévost, J. Guillaume, M. Godron, M. Dosso, and P. Curmi. 1997. The influence of soil cover organization on the floristic and structural heterogeneity of a Guianan rain forest. Plant Ecology **131**: 81–108.

ANNEXE 2

Fiche Technique : Mise en œuvre dans R de la modélisation d'une maladie à transmission vectorielle (cf chapitre V).

Chargement des données (216 troupeaux et 13 variables) :

```
> tab_read.table("Donnees.txt",sep="\t",h=T)
> dim(tab)
[1] 216  13
> tab[1:5,]
  Num posi essa tatr tysp tyss dlpp dlps camp hydr espa   Long     lat
1   1   18   20   t3  for  for   d1   d1  oui  oui    7 375455 1213610
2   2    1    2   t1  for  for   d1   d1  oui  non    7 379085 1209301
3   3   15   17   t3  for  for   d1   d1  oui  non    7 378405 1214078
4   4    3    3   t1  for  for   d1   d1  oui  non    1 343773 1205456
5   5    2    2   t1  for  for   d1   d1  oui  non    1 343744 1204543
```

Mise en place du modèle :

```
> f1_as.formula("cbind(posi, essa - posi) ~ (tyss + tatr + camp +
hydr)")
> glm1_glm(formula = f1, family = quasibinomial(), data = tab,start =
c(0.1, 0.4, -0.3, 0.7, 1, 1))
```

Analyse du modèle :

```
> anova(glm1,test="F")
Analysis of Deviance Table

Model: quasibinomial, link: logit

Response: cbind(posi, essa - posi)

Terms added sequentially (first to last)


      Df Deviance Resid. Df Resid. Dev       F    Pr(>F)
NULL                    215     666.51
tyss   1    17.66       214     648.86  7.1207 0.0082147 **
tatr   2    37.37       212     611.49  7.5365 0.0006903 ***
camp   1    59.07       211     552.41 23.8259 2.085e-06 ***
hydr   1    54.17       210     498.24 21.8486 5.271e-06 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

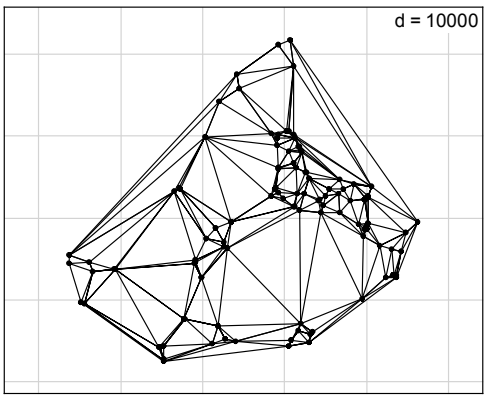
Calcul des risques relatifs :

```
oddratio_function(model){
coeffs_coef(summary(model))
lci_exp(coeffs[,1]-1.96*coeffs[,2])
or_exp(coeffs[,1])
uci_exp(coeffs[,1]+1.96*coeffs[,2])
odds_cbind(lci,or,uci)
odds
}

> oddratio(glm1)
                  lci        or      uci
(Intercept) 0.5387067 1.0872319 2.194280
```

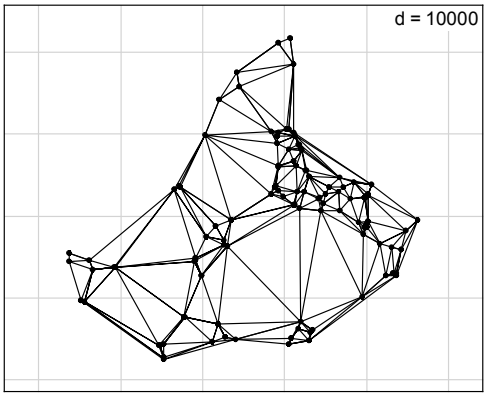
```
tysssou     0.8789475 1.3422377 2.049727
tatrt2      0.3870867 0.6715213 1.164961
tatrt3      1.0349882 1.9939569 3.841458
campoui     1.6943816 2.7035371 4.313735
hydroui     1.9395029 3.4322755 6.073987
```

Tests de l'autocorrélation :

```
> xy_read.table("xy.txt",sep="\t",h=T)
> dim(xy)
[1] 216   2

> library(tripack)
> provi_neighbours(tri.mesh(xy))
> library(ade4)
> provi.neig_neig(list=provi)
> scatter.label(xy,neig=provi.neig,inc=F,addax=F,clab=0,cnei=1)
```




```
> distances_apply(provi.neig,1,function(x) sqrt(sum((xy[x[1],]-
xy[x[2],])^2)))
> myneig_neig(edges=provi.neig[distances<15000,])
> scatter.label(xy,neig=myneig,inc=F,addax=F,clab=0,cnei=1)
```




```
> mynb_neig2nb(myneig)
> library(spdep)
> mylist_nb2listw(mynb)
> moran.mc(residuals(glm1),mylist,999)
> geary.mc(predict(glm1,type="response"),mylist,999)

        Monte-Carlo simulation of Geary's C

data:  predict(glm1, type = "response")
weights: mylist
```

```
number of simulations + 1: 1000

Geary C statistic = 0.4649, rank of observed C = 1
```

Prédictions pour l'ensemble de la population :

```
> tab2_read.table("Apredire.txt",sep="\t",h=T)
> dim(tab2)
[1] 801   9
> tab2[1:5,]
  NUMENQ LATITUDE LONGITUDE NOMBBOV tatr camp tyss dlps hydr
1      1  1213542    377184     110   t3  oui  sou   d3  non
2      2  1213610    375455      90   t3  oui  for   d1  oui
3      3  1214141    377861      40   t3  oui  sou   d3  non
4      4  1214191    377900      50   t3  oui  sou   d3  non
5      5  1208330    384860     300   t3  non  sou   d3  non

> predictions_predict(glm1, type="response", data=tab2)
> predictions[1:5]
[1] 0.9526437 0.7461524 0.8542482 0.7461524 0.7461524
```

Par la suite, le tableau des prédictions a été exporté dans le SIG afin de le coupler au modèle spatial et ainsi d'obtenir une cartographie du risque trypanosomien pour l'ensemble de la zone d'étude.

ANNEXE 3

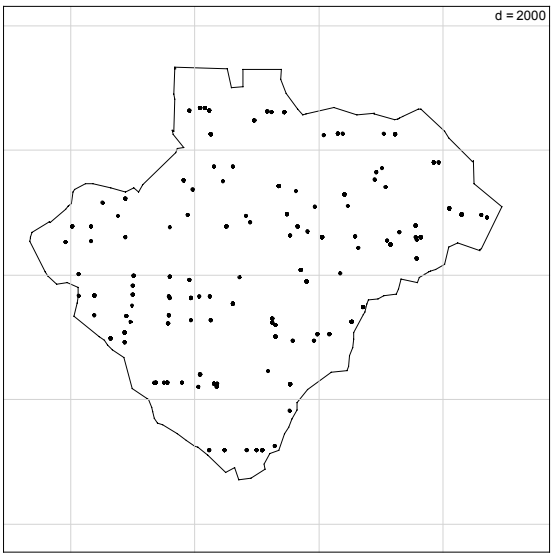
Fiche Technique : Mise en œuvre dans R de la modélisation de la dynamique d'une population de chevreuils (cf chapitre VI).

Chargement des données (1235 faons et 5 variables) :

```
> tab_read.table("natha.txt",sep="\t",h=T)
> dim(tab)
[1] 1235    5
> tab[1:5,]
  an      x       y poids sexe
1  1 385601 2127978    16    m
2  1 385601 2127978    15    m
3  1 385601 2127978    15    m
4  1 385574 2127219    14    m
5  1 385574 2127219    16    m
```

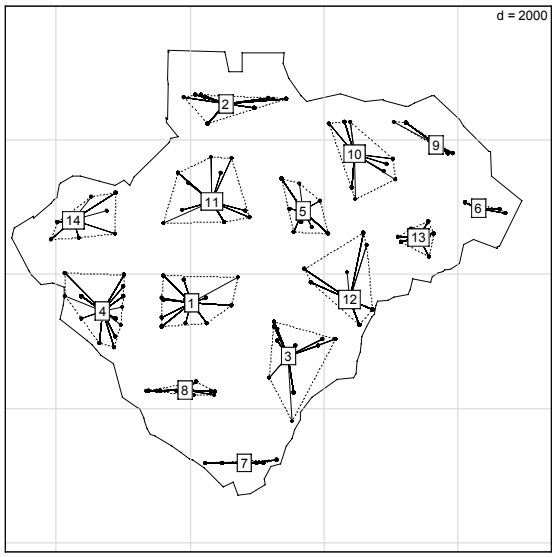
Représentation des données :

```
> contour_read.table("contour.txt",sep="\t",h=T)
> library(ade4)
> limitesx_range(tab$x)+c(-1000,1000)
> limitesy_range(tab$y)+c(-1000,1000)
> xy_cbind(tab$x,tab$y)
> scatter.label(xy,area=contour,clab=0,xlim=limitesx,ylim=limitesy)
```



L'espace est discrétisé :

```
> library(mva)
> c1600_cutree(hclust(dist(xy)),h=1600)
> scatter.class(xy,factor(c1600),inc=F,area=contour,cell=0)
> scatter.chull(xy,factor(c1600),inc=F,optchull=1,add.p=T)
```



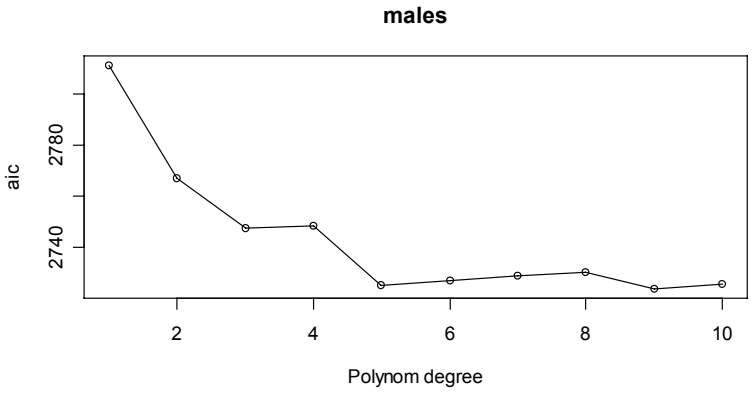

```
> c1600_as.factor(c1600)
> tab_cbind(tab,c1600)
> tab[1:5,]
  an      x       y poids sexe c1600
1  1 385601 2127978    16    m     1
2  1 385601 2127978    15    m     1
3  1 385601 2127978    15    m     1
4  1 385574 2127219    14    m     1
5  1 385574 2127219    16    m     1
```

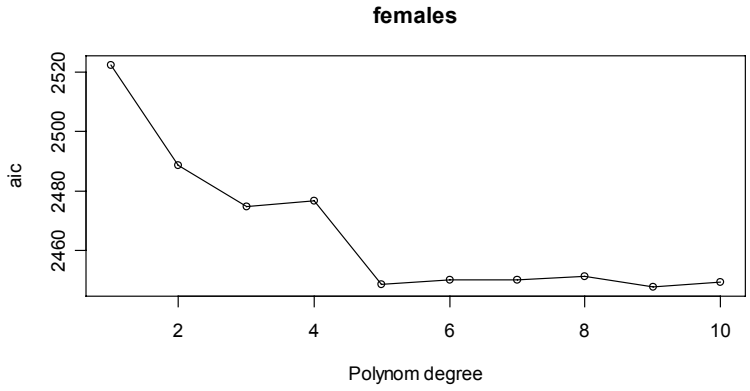
Le tableau de données est partitionné par sexe afin de pouvoir réaliser deux modèles :

```
> tab2_split(tab,tab$sexe)
> names(tab2)
[1] "f" "m"
> tab2$f[1:5,]
    an      x       y poids sexe c1600
640  1 385601 2127978    16    f     1
641  1 385601 2127978    13    f     1
642  1 385574 2127219    16    f     1
643  1 386257 2130252    16    f     2
644  1 387304 2127010    12    f     3
```

Les variations temporelles sont appréhendées par un polynôme de degré 5. Le choix du degré du polynôme a été réalisé en visualisant les variations de l'AIC :

```
> aicglm_vector()
> for (i in 1:10){
aicglm _cbind(aicglm,glm(poids~poly(an,i)+c1600,family="gaussian",
data=tab2$m)$aic)}
> plot(1:10,t(aicglm),ty='l',xlab="Polynom degree",ylab="aic",main="males")
> points(1:10,t(aicglm))
```





Construction des modèles multiplicatif et additif chez les mâles :

```
> lm1M_lm(poids~poly(an,5)*c1600,data=tab2$m)
> lm2M_lm(poids~poly(an,5)+c1600,data=tab2$m)

> anova(lm1M)
Analysis of Variance Table

Response: poids
                  Df  Sum Sq Mean Sq F value    Pr(>F)
poly(an, 5)        5  700.47  140.09 35.7268 < 2.2e-16 ***
c1600             13  238.42   18.34  4.6771 1.072e-07 ***
poly(an, 5):c1600 65  325.59    5.01  1.2774   0.07913 .
Residuals        555 2176.28    3.92
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

> cor(predict(lm1M),tab2$m$poids)^2 # Coef. de corrélation multiple
[1] 0.3675

> anova(lm2M)
Analysis of Variance Table

Response: poids
            Df  Sum Sq Mean Sq F value    Pr(>F)
poly(an, 5)  5  700.47  140.09 34.7171 < 2.2e-16 ***
c1600       13  238.42   18.34  4.5449 1.846e-07 ***
Residuals  620 2501.88    4.04
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

L'interaction n'est pas significative, le modèle additif est donc conservé.

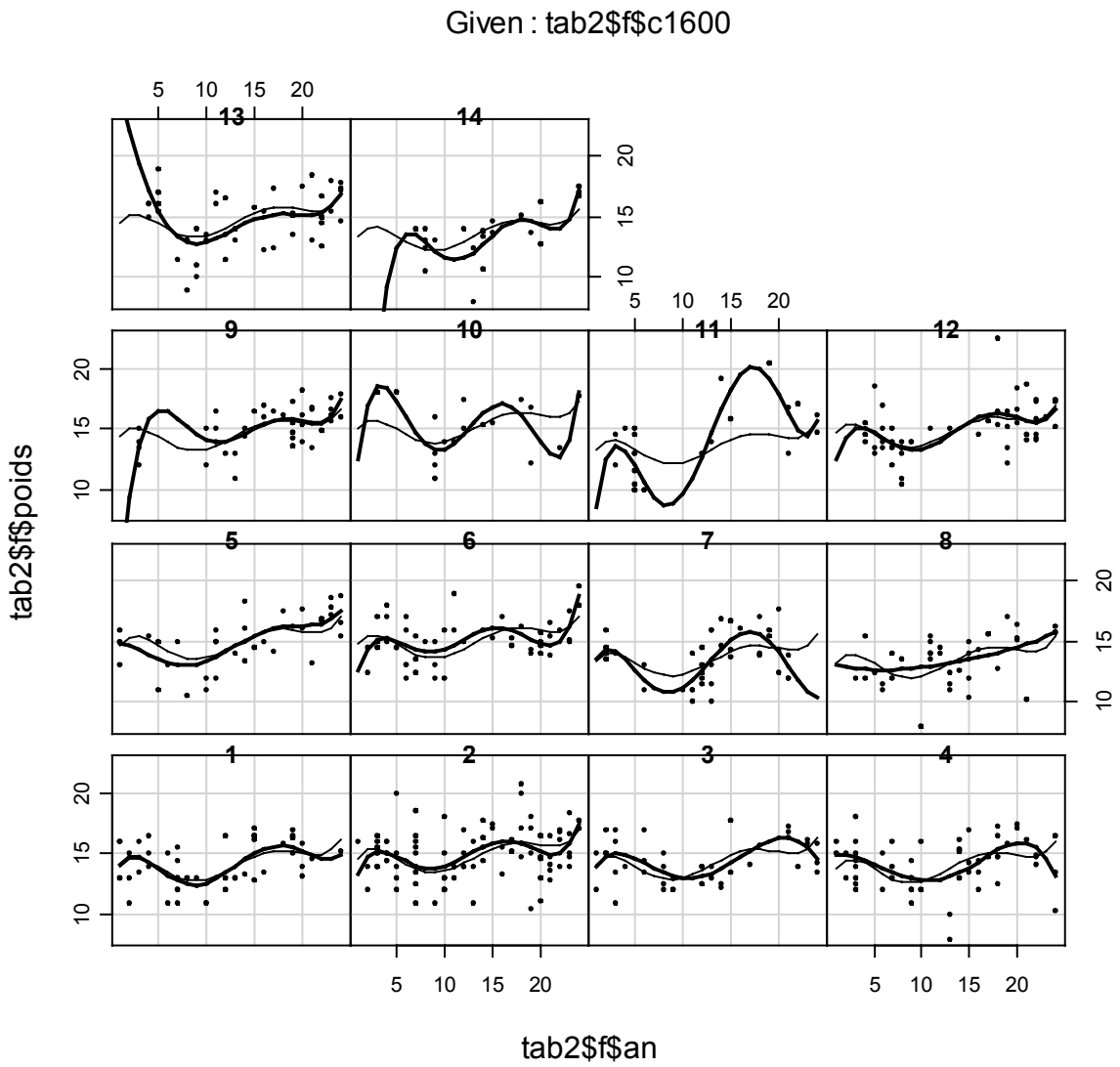
Construction des modèles multiplicatif et additif chez les femelles :

```
> lm1F_lm(poids~poly(an,5)*c1600,data=tab2$f)
> lm2F_lm(poids~poly(an,5)+c1600,data=tab2$f)

> anova(lm1F)
Analysis of Variance Table

Response: poids
                    Df  Sum Sq Mean Sq F value    Pr(>F)
poly(an, 5)          5  540.45  108.09 33.6156 < 2.2e-16 ***
c1600               13  182.33   14.03  4.3618 5.256e-07 ***
poly(an, 5):c1600   65  339.19    5.22  1.6229  0.002468 **
Residuals          512 1646.32    3.22
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

> cor(predict(lm1F),tab2$f$poids)^2 # Coef. de corrélation multiple
[1] 0.3921174

> anova(lm2F)
Analysis of Variance Table

Response: poids
             Df  Sum Sq Mean Sq F value    Pr(>F)
poly(an, 5)   5  540.45  108.09 31.4115 < 2.2e-16 ***
c1600        13  182.33   14.03  4.0758 1.875e-06 ***
Residuals   577 1985.51    3.44
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

L'interaction est significative chez les femelles. Ce résultat est étonnant d'un point de vue biologique car les mâles ont un comportement différent. On peut donc se demander si cette interaction à un véritable sens biologique.

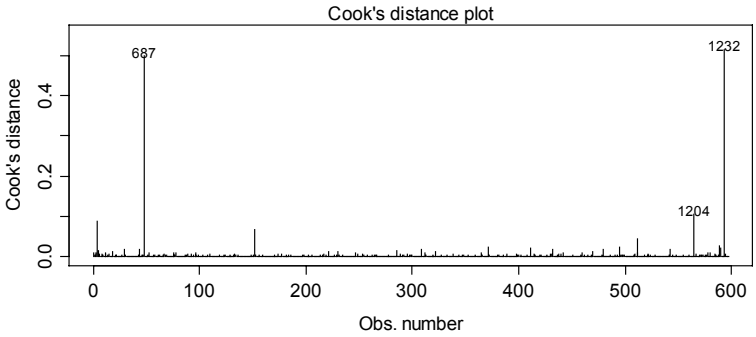
```
f1_function(x,y,col,pch,subscripts) {
      points(x,y,pch=20)
      w0_unique(tab2$f$c1600[subscripts])
      lm3F_lm(poids~an+I(an^2)+I(an^3)+I(an^4)+I(an^5)+c1600,data=tab2$f)
      lm4F_lm(poids~(an+I(an^2)+I(an^3)+I(an^4)+I(an^5))*c1600,data=tab2$f)
      model1_predict(lm3F,newdata=list(an=1:24,c1600=factor(rep(w0,24),
            levels=levels(tab2$f$c1600))))
      model2_predict(lm4F,newdata=list(an=1:24,c1600=factor(rep(w0,24),
            levels=levels(tab2$f$c1600))))
      lines(1:24, model1)
      lines(1:24, model2,lwd=2)
      title(main=as.character(unique(tab2$f$c1600[subscripts])))
}
coplot(tab2$f$poids~tab2$f$an| tab2$f$c1600,show=F,panel=f1,subscripts=T)
```

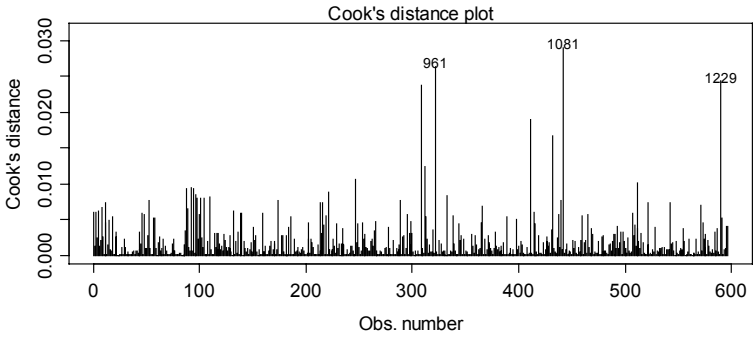
Il y a une grande similitude entre les deux modèles (additif en trait simple et multiplicatif en gras). La zone 11 est celle pour laquelle les deux modèles sont les plus différents. Il semble donc que l'interaction observée chez les femelles soit dû essentiellement à ce cluster.



Afin de comprendre ce phénomène d'interaction, on représente les distances de Cook qui évaluent l'influence de chaque observation dans l'estimation des paramètres du modèle. La distance de Cook mesure la différence entre l'estimation d'une observation faite avec le modèle complet et l'estimation obtenue avec un modèle construit sans cette observation.

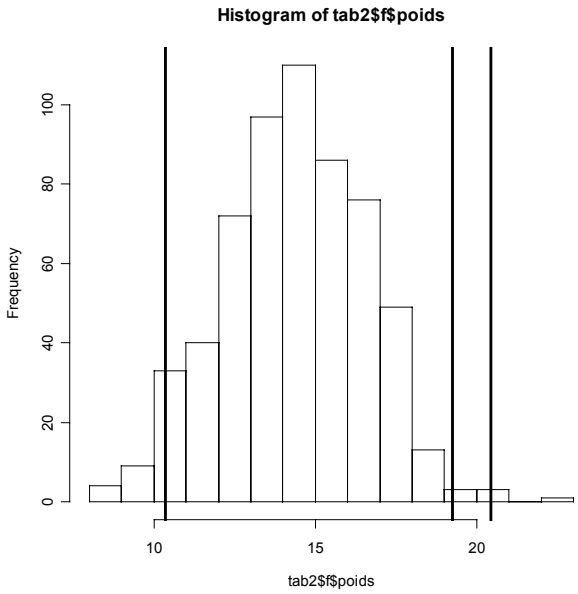
```
> par(mfrow=c(2,1))
> plot(lm1F,which=4)
> plot(lm2F,which=4)
```





Les individus 961, 1081 et 1229 ont une grand influence sur le modèle additif et peu sur le modèle multiplicatif. Ces trois observations correspondent à des valeurs extrêmes et deux d'entres elles appartiennent au cluster 11 :

```
> tab2$f[c(322,442,590),]
     an      x       y poids sexe c1600
961  14 386504 2128773 19.25    f    11
1081 19 386504 2128773 20.45    f    11
1229 24 384876 2127077 10.35    f     4

> hist(tab2$f$poids)
> abline(v=tab2$f$poids[c(322,442,590)],lwd=3)
```

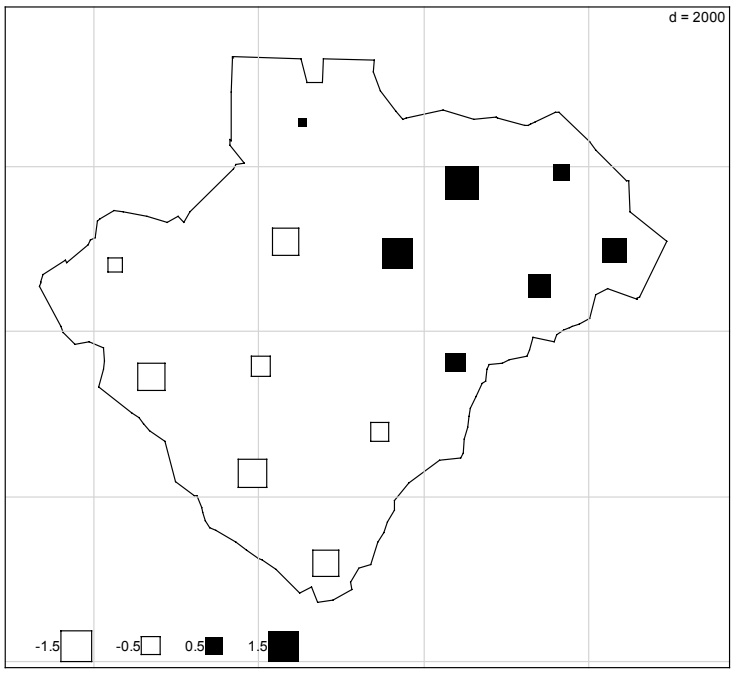


Sans ces trois individus, l'interaction n'est plus significative :

```
> anova(lm(poids ~ poly(an, 5) * c1600,data=tab2$f[-c(322,442,590),]))
Analysis of Variance Table

Response: poids
                   Df  Sum Sq Mean Sq F value    Pr(>F)
poly(an, 5)         5  553.31  110.66 34.7818 < 2.2e-16 ***
c1600              13  201.87   15.53  4.8807 4.393e-08 ***
poly(an, 5):c1600  65  259.23    3.99  1.2535   0.09765 .
Residuals         509 1619.45    3.18
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

La variabilité spatiale des poids s'exprime simplement sur une carte :

```
> xmoy_tapply(tab$x,tab$c1600,mean)
> ymoy_tapply(tab$y,tab$c1600,mean)
> predM_predict(lm2M,newdata=list(an=rep(1,14),c1600=factor(1:14)))
>
scatter.value(cbind(xmoy,ymoy),scale(predM),area=contour,xlim=limitesx
,ylim=limitesy)
```



ANNEXE 4

Programme R pour réaliser une ACP par NIPALS (cf analyse NIPALS spatialisée, chapitre VII)

```
nipals_function(X, nf=2, rec=F){
# X un data frame contenant éventuellement des valeurs manquantes (NA)
# nf nombre de facteurs à conserver
# rec, si rec=T, la reconstitution des données sur les nf premiers axes est
réalisée

    seuil_1e-9 # seuil pour la convergence
    nc <- ncol(X)
    nr <- nrow(X)
    if (rec)
        x_list(li=matrix(0,nr,nf),c1=matrix(0,nc,nf),co=matrix(0,nc,nf),
        eig=rep(0,nf),nb=rep(0,nf),rec=matrix(0,nr,nc))
    else
        x_list(li=matrix(0,nr,nf),c1=matrix(0,nc,nf),co=matrix(0,nc,nf),
        eig=rep(0,nf),nb=rep(0,nf))
    row.names(x$c1)_names(X)
    row.names(x$co)_names(X)
    row.names(x$li)_row.names(X)
    X_scale(X, center=T, scale=T)
    for (h in 1:nf) {
        th_X[,1]
        ph1_rep(1/sqrt(nc),nc)
        ph2_rep(1/sqrt(nc),nc)
        diff_rep(1,nc)
        nb_0
        while (sum(diff^2, na.rm=T)>seuil) {
            for (i in 1:nc) {
                the_th[X[,i]!="NA"]
                ph2[i]_sum(X[,i]*th, na.rm=T)/sum(the*the,na.rm=T)
            }
            ph2_ph2/sqrt(sum(ph2*ph2,na.rm=T))
            for (i in 1:nr) {
                ph2e_ph2[X[i,]!="NA"]
                th[i]_sum(X[i,]*ph2, na.rm=T)/sum(ph2e*ph2e,na.rm=T)
            }
            diff_ph2-ph1
            print(sum(diff^2,na.rm=T))
            ph1_ph2
        nb_nb+1
        }
        X_X-th%*%t(ph1)
        x$nb[h]_nb # nombre d'itérations
        x$li[,h]_th # coordonnées des lignes
        x$c1[,h]_ph1 # coordonnées des colonnes de variance unité
        x$eig[h]_sum(th*th,na.rm=T)/(nr-1) # valeurs propres
        x$co[,h]_x$c1[,h]*sqrt(x$eig[h]) # coord. col. (variance lambda)
    }
    if (rec) {
        for (h in 1:nf) {
            x$rec_x$rec+x$li[,h]%*%t(x$c1[,h]) # tableau reconstitué
        }
    }

    return(x)
}
```

ANNEXE 5

L'analyse de co-inertie procustéenne dans R (cf chapitre VIII)

```
Procuste_function (df1, df2, scale = T, nf = 4, tol = 1e-07)
{
    df1 <- data.frame(df1)
    df2 <- data.frame(df2)
    if (!is.data.frame(df1))
        stop("data.frame expected")
    if (!is.data.frame(df2))
        stop("data.frame expected")
    l1 <- nrow(df1)
    if (nrow(df2) != l1)
        stop("Row numbers are different")
    if (any(row.names(df2) != row.names(df1)))
        stop("row names are different")
    c1 <- ncol(df1)
    c2 <- ncol(df2)
    X <- scale(df1, scale = F)
    Y <- scale(df2, scale = F)
    var1 <- apply(X, 2, function(x) sum(x^2))
    var2 <- apply(Y, 2, function(x) sum(x^2))
    tra1 <- sum(var1)
    tra2 <- sum(var2)
    if (scale) {
        X <- X/sqrt(tra1)
        Y <- Y/sqrt(tra2)
    }
    PS <- t(X) %*% Y
    svd1 <- svd(PS)
    rank <- sum((svd1$d/svd1$d[1]) > tol)
    if (nf > rank)
        nf <- rank
    u <- svd1$u[, 1:nf]
    v <- svd1$v[, 1:nf]
    scor1 <- X %*% u
    scor2 <- Y %*% v
    rot1 <- X %*% u %*% t(v)
    rot2 <- Y %*% v %*% t(u)
    res <- list()
    X <- data.frame(X)
    row.names(X) <- row.names(df1)
    names(X) <- names(df1)
    Y <- data.frame(Y)
    row.names(Y) <- row.names(df2)
    names(Y) <- names(df2)
    res$d <- svd1$d
    res$rank <- rank
    res$nfact <- nf
    u <- data.frame(u)
    row.names(u) <- names(df1)
    names(u) <- paste("ax", 1:nf, sep = "")
    v <- data.frame(v)
    row.names(v) <- names(df2)
    names(v) <- paste("ax", 1:nf, sep = "")
    scor1 <- data.frame(scor1)
    row.names(scor1) <- row.names(df1)
    names(scor1) <- paste("ax", 1:nf, sep = "")
    scor2 <- data.frame(scor2)
    row.names(scor2) <- row.names(df1)
    names(scor2) <- paste("ax", 1:nf, sep = "")
    if ((nf == c1) & (nf == c2)) {
```

```
        rot1 <- data.frame(rot1)
        row.names(rot1) <- row.names(df1)
        names(rot1) <- names(df2)
        rot2 <- data.frame(rot2)
        row.names(rot2) <- row.names(df1)
        names(rot2) <- names(df1)
        res$rot1 <- rot1
        res$rot2 <- rot2
    }
    res$tab1 <- X
    res$tab2 <- Y
    res$load1 <- u
    res$load2 <- v
    res$scor1 <- scor1
    res$scor2 <- scor2
    res$call <- match.call()
    class(res) <- "procuste"
    return(res)
}
```

ANNEXE 6

Script Avenue permettant d'obtenir la matrice de voisinage par le croisement de deux thèmes (cf analyse RLQ spatialisée, chapitre VIII)

```
theClass= av.GetActiveDoc.GetClass.GetClassName
if (theClass<>"View") then
   MsgBox.Error("The active document must be a view for this option","Error")
   Exit
end
theView = av.GetActiveDoc

'********************** Sélection du thème 1***********************************
themeList = theView.getThemes
theTheme1=MsgBox.Choice (themeList, "Choose Theme" , "Theme List")
if (theTheme1=nil) then
  exit
end

theVTab1 = theTheme1.GetFTab
shapeFld1 = theVTab1.FindField("Shape")
theField1=MsgBox.List (theVTab1.GetFields, "Selectionner un champs de "+theVTab1.GetName+" (identifiant
parcelle)", "Field List")
if (theField1=nil) then
  exit
end
'********************** Sélection du thème 2***********************************

theTheme2=MsgBox.Choice (themeList, "Choose Theme" , "Theme List")
if (theTheme2=nil) then
  exit
end

theVTab2 = theTheme2.GetFTab
shapeFld2 = theVTab2.FindField("Shape")
theField2=MsgBox.List (theVTab2.GetFields, "Selectionner un champs de "+theVTab2.GetName+" (identifiant
parcelle)", "Field List")
if (theField2=nil) then
  exit
end
'**********************
workDir = av.GetProject.GetWorkDir
fName = FileName.Make(workDir.AsString).MakeTmp("Cross","")
repeat=0
while (repeat = 0)
  outFileName = FileDialog.Put(fname,"*","Enter a OutFile Name")
  if (outFileName = NIL) then
      exit
  elseif (outFileName.GetBaseName.Contains(" ")) then

     MsgBox.Warning("OutFile name not specified, re-enter","Crossing two fields")
  else
     repeat = 1
  end
end


'*************** Création de la table (matrice de voisinage) ***********************************

theNewVTab=VTab.MakeNew(outFileName,DBase)
theNewFieldR=theField1.Clone
theNewVTab.AddFields({theNewFieldR})
```

```
theNewVTab.SetEditable(TRUE)
for each rec in theVTab1
  theNewVTab.SetValue(theNewFieldR,theNewVTab.AddRecord,theVTab1.ReturnValue(theField1,rec))
end
for each rec in theVTab2
  thefieldName=theVTab2.ReturnValue(theField2,rec).AsString
  theNewVTab.AddFields({Field.Make(thefieldName.AsString,#FIELD_LONG,20,0)})
  for each newrec in theNewVTab
    if ((theVTab1.ReturnValue(shapeFld1,newrec)).Intersects(theVTab2.ReturnValue(shapeFld2,rec))=TRUE)
    then
      theNewVTab.SetValue(thenewVTab.FindField(thefieldName),newrec,1)
    else
      theNewVTab.SetValue(thenewVTab.FindField(thefieldName),newrec,0)
    end
  end
end
theNewVTab.SetEditable(FALSE)
```