N° d'ordre: 230-2005 Année 2005

Thèse
présentée devant l'Université Claude Bernard - LYON I

pour l'obtention du
Diplôme de doctorat
(*arrêté du 25 avril 2002*)

Soutenue le 6 décembre 2005

par

Sandrine Pavoine

Méthodes Statistiques pour la Mesure de la Biodiversité

Composition du jury

Directeur de thèse :	M. Daniel Chessel, Professeur à l'Université Claude Bernard Lyon I
Rapporteurs :	M. François Houllier, Directeur de recherche INRA-CIRAD, AMAP, Montpellier M. Jean-Dominique Lebreton, Directeur de recherche CNRS, CEFE, Montpellier
Examinateurs :	M. Robert Barbault, Professeur à l'Université Pierre et Marie Curie Paris VI M. Pierre Legendre, Professeur à l'Université de Montréal, Canada Mme Dominique Pontier, Professeur à l'Université Claude Bernard Lyon I

Remerciements

Je tiens à remercier toutes les personnes qui m'ont soutenue, encouragée, accueillie, conseillée durant ces quelques années au laboratoire de biométrie et biologie évolutive, particulièrement :

– Mon directeur de thèse, Daniel Chessel, pour m'avoir laissée libre de découvrir le monde de la recherche par moi-même, pour avoir mis tout en œuvre pour que cette découverte se fasse dans les meilleures conditions, pour avoir été toujours présent pour de longues et riches discussions, et pour avoir toujours répondu à mes questions par de multiples conseils variés et parfois opposés, afin de me guider, sans jamais s'imposer, dans mes décisions et mes choix ;

– Anne-Béatrice Dufour pour tout le soutien qu'elle m'a apporté, pour les nombreux conseils qu'elle m'a donné et pour son agréable compagnie lors de nos nombreux voyages ;

– Les membres du jury, François Houllier, Jean-Dominique Lebreton, Robert Barbault, Pierre Legendre et Dominique Pontier pour avoir pris le temps de lire et juger cette thèse, pour leurs remarques avisées et leurs nombreux conseils ;

– Tous les membres du laboratoire de biométrie et biologie évolutive, tout spécialement ceux qui m'ont aidée et soutenue dans mes recherches, en particulier Dominique Pontier et Christian Biémont et tous ceux qui m'ont aidée ou dépannée dans mes enseignements particulièrement Sandrine Charles, Dominique Allainé, Dominique Mouchiroud, Vincent Lacroix, Samuel Venner ;

– L'institut français de la biodiversité pour avoir permis et provoqué la réunion de jeunes chercheurs tous travaillant sur la biodiversité, que ce soit aux niveaux des plantes, des animaux et/ou des micro-organismes, mais avec la richesse d'approches très variées, écologiques, génétiques, géographiques, statistiques, ethnologiques, philosophiques, juridiques, économiques et politiques.

– Le groupe de phylogénie et écologie des communautés, et surtout les organisateurs de notre premier Workshop, Jérôme Chave et Olivier Hardy ;

– L'équipe de Guyane, Eric Marcon, Christopher Baraloto, Jean-Christophe Roggy, et François Morneau. Un grand merci à Eric pour m'avoir invitée à présenter mes recherches à l'INRA de Kourou, pour m'avoir montré les recherches actuelles faites en Guyane sur la biodiversité fonctionnelle des peuplements végétaux, autant sur les aspects théoriques que sur les relevés et expérimentations de terrain. Merci à Christopher pour m'avoir présenté et expliqué le fonctionnement et l'intérêt de l'herbier de Guyane. Tout a été réuni pour me donner un nouveau regard sur ce magnifique, complexe et impressionnant écosystème qu'est la forêt amazonienne ;

– Tous ceux qui ont fait part de leurs commentaires sur mes recherches et/ou de leur désir de partager leur recherche : Carlo Ricotta, Janos Izsak, Jose Alexandre F. Diniz-Filho,

Cajo ter Braak, Ross Crozier, Owen Petchey, Arne Mooers ;

– Raphaël Pélissier pour avoir accueilli très volontiers mes études et remarques sur la décomposition de la biodiversité et pour m'avoir donné l'occasion de m'exprimer et de m'intégrer dans le groupe de phylogénie et écologie des communautés ;

– Jacques Blondel pour tout le soutien qu'il m'a apporté tout au long de cette thèse et pour tous les encouragements dont il me fait part actuellement pour mes projets en cours ;

– Laurent Excoffier pour m'avoir accueillie quelques jours dans son laboratoire, afin que je comprenne tous les tenants et aboutissants de l'AMOVA ;

– Robert May pour m'avoir ouvert les portes du département de Zoology à l'université d'Oxford et Michael Bonsall pour m'y avoir invitée et accueillie, et pour m'avoir aidée à développer les projets que j'avais en tête ;

– Tous ceux qui m'ont apporté leur soutien et leurs conseils dans la préparation de dossiers pour tout ce qui va suivre maintenant. Un grand merci à Alice, pour nos discussions, et Emmanuelle Porcher, pour tous ses conseils.

– Robb Ogden pour m'avoir enseigné de façon ludique les rudiments de la langue anglaise, pour m'avoir entraînée aux conférences internationales, pour avoir relu et corrigé mes articles et surtout pour tous nos rires et sourires ;

– Clément pour toutes nos discussions plus ou moins philosophiques, pour nos moments de détente, et pour notre long voyage...

A ma famille, mes grands-parents, ma mère, mon père, Olivier, Pierre-Yves, Gabriella, Bernard, Alexandre, William, Sophie, Alice, Isabelle, Maxime, et tous nos chats. Vous avez été là à chaque étape importante de ma vie, vous m'avez soutenu sans réserve même dans mes rêves les plus ambitieux. Merci. Je vous dédie cette thèse.

« Il n’existe que deux espèces de folies contre lesquelles on doit se protéger. L’une est la croyance selon laquelle nous pouvons tout faire. L’autre est celle selon laquelle nous ne pouvons rien faire. » André Brink

R´ ´

Face à l'accumulation des indices développés pour mesurer la biodiversité, la détermination de schémas fondamentaux est devenue nécessaire. Cette thèse démontre que : 1) l'axiomatisation de Rao constitue un schéma statistique pour l'analyse de la variation, en particulier variance et diversité ; 2) au cœur de ce schéma, un indice, l'entropie quadratique, basé sur une matrice de dissimilarités est défini sur l'ensemble des distributions de fréquences ; 3) la décomposition de cet indice généralise des méthodes utilisées pour l'analyse de la variation en statistique (ANOVA), génétique (AMOVA) et écologie, et est égale à la décomposition de l'inertie d'un nuage de points dans un espace euclidien déterminé ; 4) l'entropie quadratique appliquée à des dissimilarités ultramétriques présente trois propriétés qui sont fondamentales pour un indice de biodiversité. Cette thèse analyse l'unité de ce schéma qui réunit les concepts de diversité, inertie, dissimilarité, ordination et originalité.

Mots-clés: Analyse multivariée, ANOVA, CATANOVA, entropie quadratique, espèce rare, indice de dissimilarité, indice de diversité, ordination, test d'hypothèses, statistique.

A

Given the accumulation of indices developed for measuring biodiversity, the determination of fundamental patterns has become necessary. This thesis demonstrates that : 1) Rao's axiomatization constitutes a statistical framework for the analysis of variation, especially variance and diversity ; 2) At the heart of this framework, an index, the quadratic entropy, which is based on a matrix of dissimilarities, is defined on the set of frequency distributions ; 3) the decomposition of this index generalizes methods used to analyze variation in statistics (ANOVA), genetics (AMOVA) and ecology, and it is equal to the decomposition of the inertia of a cloud of points in a specified Euclidean space ; 4) the quadratic entropy applied to ultrametric matrices has three properties which are fundamental for an index of biodiversity. This thesis analyzes the unity of this framework, which assembles the concepts of diversity, inertia, dissimilarity, ordination and originality.

Keywords: Multivariate analysis, ANOVA, CATANOVA, quadratic entropy, rare species, index of dissimilarity, index of diversity, ordination, hypothesis testing, statistics.

Table des matières

Chapitre 1

Introduction

Le mot "biodiversité" est apparu peu à peu quand on a pris conscience de la disparité des espèces de la planète et de la disparition depuis tout temps de certaines d'entre elles. Une multitude de points de vue s'entremêlent, autant dans le language courant que dans et entre disciplines scientifiques, sur ce qu'est cette biodiversité. Si on en croit son étymologie, ce mot désigne toute la variété du monde vivant, sa pluralité, sa complexité, ses multitudes représentations et apparences. C'est cette définition très générale qui a été adoptée pour aborder cette thèse et sur laquelle nous nous baserons pour comprendre et ordonner les mesures de biodiversité.

Pourquoi étudier la biodiversité ?

Je souhaiterais répondre à cette question en cinq points choisis qui me semblent essentiels sans être pour autant exhaustifs.

1) La biodiversité est l'un des moteurs de la vie sur Terre.
L'étude de la biodiversité nous permet d'abord d'améliorer notre connaissance et notre compréhension des origines, de la dynamique et de l'évolution du monde vivant. D'après les théories les plus récentes, les êtres vivants sont nés de la combinaison chimique de seulement quatre molécules. Mais nous savons que deux lettres suffisent pour écrire un language très complexe. Chaque organisme résulte d'une combinaison unique de milliards de nucléotides. Et l'ensemble de ces combinaisons crée la diversité qui est sans doute un des facteurs qui a permis le maintien de la vie dans un environnement changeant. La diversité est donc à l'origine de la persistance de la vie sur Terre.

> "Multiplication et diversité sont les deux faces d'un même phénomène, d'une même nécessité : la perpétuation de la vie dans un monde changeant" (Barbault 1997)

2) La biodiversité est encore mal connue.
Les grands voyages des siècles passés ont montré qu'il existe de par le monde des faunes et des flores très variées. A chaque découverte de nouvelles espèces, des noms leur étaient attribués parfois extrêmement complexes et longs. La description de nouvelles espèces s'est multipliée et les synonymes, c'est-à-dire des noms différents attribués aux mêmes groupes d'organismes, se sont aussi multipliés. Carl von Linné et ses successeurs ont permis d'ordonner ces différentes espèces, de rassembler les connaissances, et de détecter les synonymes. Cependant 99% des espèces aujourd'hui répertoriées ont juste un nom et n'ont pas été analysées (Wilson 1993). Actuellement, de nouvelles espèces sont découvertes chaque année ; par exemple en 1990 un nouveau primate a été découvert sur une île à soixante cinq kilomètres de Sao Paulo (Wilson 1993). Or la nature nous réservera sûrement encore bien des surprises. Le 18 juillet 1895, Joubin découvre, dans le ventre d'un cachalot capturé aux Açores près de Terceira, un calmar (*Lepidoteuthis grimaldii*) portant plusieurs milliers d'écailles (Compte-rendu de l'académie des sciences 1905). Aucun autre céphalopode ne présente une semblable disposition tégumentaire, les écailles étant en général réservées aux poissons et reptiles. Quelle est l'utilité d'un tel céphalopode ? En existe-t-il d'autres ? Personne actuellement n'est capable de donner de réponse précise à ces questions.

3) Le monde actuellement voit une accélération drastique de la perte de biodiversité, accélération non pressentie à cause de sa nature anthropique.
Erwin (1991) commence un article en faisant remarquer qu'aux époques féodales et monarchiques, des parts des royaumes étaient préservées [entre autres du défrichement] pour satisfaire l'aristocratie lors de traditionnelles parties de chasse. Aujourd'hui, les zones préservées dans lesquelles on s'efforce de réduire localement l'impact des sociétés humaines représentent une infime partie de la surface terrestre. Le taux d'extinction des espèces augmente et le processus de spéciation, qui créera une part de la biodiversité future est sévèrement contraint par l'élimination d'habitats contigus. Nous vivons actuellement une phase d'accélération du rythme d'extinction des espèces. C'est cette réalité qui nécessite une augmentation de l'intérêt des hommes à l'étude de la diversité biologique. L'extinction des espèces a toujours existé, mais la vitesse du phénomène a changé. Plus un phénomène est rapide moins les groupes d'organismes ont le temps de s'y adapter. Ce que nous observons actuellement ressemble fort au commencement d'une grande extinction, une sixième période d'extinction massive.
Qu'appelle-t-on une grande extinction ? Le nombre d'espèces nommées et enregistrées sur Terre est entre 1.4 millions et 1.8 millions. Les estimations du nombre total d'espèces présentes sur Terre varient de 3 millions à 30 millions (May 2002), voire même 50 millions (Barbault 1997) et peut être plus. Selon les estimations, entre 1 et 10% des espèces ayant vécu sur Terre sont toujours présentes. Pour les oiseaux et les mammifères, la durée de vie moyenne d'une espèce actuelle est évaluée entre 100 et 1000 années soit 10^3 à 10^5 fois plus courte que celle évaluée à partir de données fossiles. Les estimations pour l'ensemble des organismes sont difficiles puisque dans les enregistrements fossiles 95% des espèces sont des animaux marins alors qu'actuellement seulement 15% des plantes et des animaux vivent en milieu marin (May 1995).
Le taux d'extinction des espèces est donc une mesure complexe à déterminer surtout parce que beaucoup d'espèces vivant actuellement restent inconnues, sont peu étudiées, ou ont une biologie et une écologie peu connue, notamment chez les invertébrés et les microorganismes. Les évaluations dépendent du lieu géographique et de la méthode appliquée. Cependant la plupart des scientifiques s'accordent pour dire que ce taux est indubitablement élevé. Par exemple, au niveau des populations, en supposant que l'extinction d'une population est une fonction linéaire de la perte de l'habitat, environ une population toutes les deux secondes est détruite dans les forêts tropicales (Hugues *et al.* 1997).

4) La perte de biodiversité a des conséquences fortes.
Les facteurs actuels du déclin sont multiples. Les principaux sont (Wilson 2003) :

- destruction de l'habitat ;
- déplacement écologique suite à l'introduction d'espèces exogènes ;
- pollution chimique ;
- hybridation avec d'autres espèces ;
- pêche ou chasse excessive.

Ces facteurs peuvent agir sur une espèce ou sur tout un habitat et donc sur toutes les espèces vivant dans cet habitat. Chaque population d'une espèce joue plusieurs fonctions dans un écosystème et même au delà une population peut avoir un rôle important dans la régulation des cycles biogéochimiques. Les conséquences de la perte de ces populations dépendent de la capacité d'autres populations d'autres espèces à remplacer les fonctions perdues. La rapidité estimée des extinctions contemporaines fait que les conséquences peuvent être sévères. De plus, "chaque

espèce est une bibliothèque d'informations acquises par l'évolution sur des centaines de milliers, voire des millions d'années. Ce sont des bibliothèques entières que nous brûlons. Or si nous avons une idée de ce que la déstabilisation entraînera (moindre productivité, moindre sûreté, changements du climat etc.), nous n'avons aucune idée de la valeur pour l'humanité de ce que nous perdons en termes d'information (Wilson et Postel-Vinay 2000)". Les expériences du passé ont montré qu'après une grande extinction il faut des dizaines de millions d'années pour retrouver une diversité comparable en nombre, et non en contenu, à celle d'avant la période de grande extinction. Ainsi, face à cette accélération des phénomènes d'extinction, il y a eu une prise de conscience que la perte de chaque population ou de chaque espèce est une perte de diversité, de fonctions et à un autre niveau de connaissance biologique.

5) Nous avons besoin de suffisamment de connaissance pour comprendre et stopper la perte de biodiversité.

Face à ces destructions est née la conviction qu'il est nécessaire d'agir. Le problème de la sauvegarde de la biosphère est maintenant concentré dans une discipline récente "la biologie de la conservation" au cœur même de l'écologie.

> "Le lien [...] entre l'Histoire moderne de l'espèce humaine d'un côté et le devenir de la biosphère de l'autre est au cœur de ce qu'on appelle aujourd'hui, faute de mieux, la biologie de la conservation. En fait, il serait plus juste de dire que nous sommes ici tout simplement plongés au cœur même de l'écologie, « science de l'Homme et de la nature », écrivait justement Jean-Paul Deléage en sous-titre de son « Histoire de l'écologie »." (Barbault 1997)

La diversité de la vie sur Terre, et les mécanismes qui la génèrent, se retrouvent dans tout jeu de données en écologie. Les problèmes de la conservation de la biodiversité sont clairement énoncés dans les deux citations ci-dessous :

> "As more and more species face extinction in the wild over the next few decades, how do we go about making choices for the ineluctably limited number of places on the ark ? [...] How do we go about designing a limited set of reserves that will, in some defined sense optimize what is saved ?" (May 1990)
>
> "How can we protect the most species per dollar invested ?" (Myers *et al.* 2000)

Comment étudier la biodiversité ?

Pour analyser et essayer de gérer cette situation, il faut mesurer, chiffrer et donc recueillir sur le terrain des données. Si nous sommes sur le point de conserver autant de biodiversité qu'il est possible avec des ressources limitées, nous devons savoir où se trouve la plus grande part de biodiversité et pour ça nous aurons besoin d'être capable de la mesurer (Williams et Humphries 1994). Pour évaluer l'impact d'une pollution par exemple sur la faune et la flore d'une région, ou pour estimer quelles régions actuellement concentrent le plus de diversité, il faut des moyens de classer, de comparer ces données. Ce rôle est joué par des fonctions statistiques appelées "indices" dont les valeurs numériques sont appelées "mesures".

Le problème émergeant, il y a quelques décennies, est l'absence de mesure de biodiversité qui permette d'avoir un nombre résumant la diversité ou un aspect de la diversité d'une

région pour pouvoir surtout comparer ce nombre avec celui d'une autre région. Dans les années 70 et 80, de nombreux chercheurs se sont penchés sur la mesure de la biodiversité ou plus généralement de la diversité. Le manque de formalisme au niveau du mot diversité a abouti à d'abondantes façons de définir, mesurer et interpréter ce concept. Le problème auquel nous nous confrontons maintenant est qu'il existe une multitude d'indices et peu de schémas fondamentaux reliant, classant ces indices.

A travers ce foisonnement, trois points sont abordés dans cette thèse et structurent chaque chapitre.

1. La mise en évidence de la diversité des approches destinées à décrire et mesurer la biodiversité. Cette recherche inclus à la fois, du point de vue biologique (génétique et écologique), la diversité dans la façon de regarder et définir la biodiversité et aussi, du point de vue des statistiques, la diversité dans la façon de mesurer, évaluer la biodiversité.
2. La recherche de points communs dans ces approches variées afin de mettre en évidence des schémas fondamentaux.
3. La proposition à partir de ces schémas de méthodes statistiques nouvelles permettant de mieux décrire la biodiversité à différentes échelles.

1) Mise en évidence de la diversité des approches.

Du point de vue biologique, la biodiversité est le plus souvent mesurée par le nombre d'espèces. D'autres indices intègrent l'abondance relative des espèces. Ces distributions d'abondance ont été modélisées, et incluses dans des modèles afin de comprendre comment un profil observé de diversité peut apparaître, quels ont été les mécanismes à l'origines de la diversité observés : mécanismes neutres (migration, mort, spéciation) ou adaptatifs (sélection naturelle). Mais même dans ces modèles on ne regarde souvent pour mesurer la biodiversité que le nombre d'espèces. Si la sélection naturelle agit sur les espèces c'est qu'elle agit sur des caractères des espèces, et si elle agit différemment sur chaque espèces, c'est que les espèces diffèrent, selon des caractères variés. La plupart des indices résument, simplifient considérablement la biodiversité en faisant l'hypothèse que les espèces sont équivalentes, interchangeables. Par simplification, ils omettent que les espèces ont une histoire commune, matérialisée par la phylogénie, qu'elles ont des traits phénotypiques et génétiques qui les rendent semblables selon certains critères et uniques selon d'autres critères. Le manque de prise en compte de ces différences et leur absence très fréquente lors des prises de décision destinées à gérer cette biodiversité, a été un des moteurs de cette thèse. Ce manque a été souligné (Cousins 1991), discuté, des méthodes ont été proposées (e.g. Hendrickson et Ehrlich 1971, Vane-Wright *et al.* 1991, Faith 1992a) mais leur application concrète est encore très rare. Ainsi dans le livre de Magurran (2004), seule la diversité taxonomique incorporant les liens taxonomiques entre espèces est abordée et très succinctement, ce qui reflète effectivement bien la très faible part donnée à ces mesures innovantes dans la littérature.

Du point de vue statistique, certains indices sont des moyennes, d'autres des variances, parfois on utilise des coefficients de variation, d'autres enfin sont développés de façon *ad hoc* pour un type de données et de problèmes et inclus plusieurs paramètres définis arbitrairement. La multiplicité des manières de mesurer un même aspect de la biodiversité fait que les indices n'ont pas les mêmes propriétés mathématiques ou statistiques. Nous allons voir que certaines propriétés

mathématiques peuvent permettre de s'attaquer à une description plus complexe de la biodiversité. Leurs absences pour certains indices limitent alors l'apport, l'impact de ces indices dans la mesure détaillée de la biodiversité.

2) Recherche de schémas fondamentaux.
Parmi ces différentes manières de concevoir et mesurer la biodiversité, il est possible de détecter des structures qui se répètent laissant place à des schémas centraux et fondamentaux. Ce que l'on peut raisonnablement demander à un schéma fondamental c'est d'englober un certain nombre d'indices de biodiversité comme cas particulier d'un indice plus général, de montrer les avantages et les désavantages de chaque indice, de définir dans quels cadres chaque indice fonctionne le mieux et où il ne doit pas être utilisé. Lorsqu'on utilise un indice on doit connaître ses propriétés, dont son domaine de définition, on doit pouvoir lui associer un estimateur et connaître son espérance, sa variance, on doit pouvoir expliquer précisément ce qu'il mesure, par exemple est-ce une somme de quelque chose, une moyenne, une variance, une distance, une probabilité ? Il faudrait pouvoir définir un schéma fondamental qui soit cohérent avec les connaissances actuelles sur la biodiversité et qui puisse évoluer en même temps que les connaissances futures. Le fil conducteur de cette thèse a été la recherche d'un tel schéma. Cette recherche a été faite avec la contrainte que ce schéma puisse intégrer, dans la mesure de la biodiversité, les différences entre espèces quelles que soient leurs natures (phylogénétique, phénotypiques, génétiques ou autres).

3) Propositions de méthodes statistiques nouvelles.
Les questions centrales suivantes ont motivées et orientées mes recherches pour développer de nouvelles méthodes :

- A quelles échelles spatiales se trouve la plus grande part de diversité ? Par exemple, est-ce dans les habitats d'une région ? Est-ce entre ces habitats ? Entre régions ? (Chapitre 3 et Annexe 2)
- Peut-on mettre en évidence des patrons, des structures de biodiversité dans un espace hétérogène ? Comment peut-on les expliquer ? (Chapitre 4 et Annexe 1)
- Les profils de biodiversité sont-ils les mêmes selon que l'on considère pour décrire les liens et différences entre espèces, la phylogénie, des traits phénotypiques, des informations génétiques ? (Annexe 1)
- Quelle est la contribution de chaque espèce à la biodiversité d'un ensemble (par exemple une communauté ou un écosystème) ? (Chapitre 5 et Annexes 3 et 4)

Mon travail s'est alors concentré à développer de nouvelles méthodes statistiques qui permettent et permettront, dans leurs applications réalisées et dans leurs applications futures, d'apporter des éléments de réponses à ces questions, et aussi de façon intéressante d'ouvrir sur d'autres problèmes et questionnements.

Pour mener à bien ces trois points, mon travail a commencé par la description des classes de données sur lesquelles la biodiversité est étudiée. La diversité est une propriété intrinsèque à tout jeu de données. La biodiversité existe dans tout ce qui nous entoure et qui est vivant. Tout jeu de données en biologie reflète une partie de ce que l'on appelle biodiversité. Un grand nombre de disciplines historiquement séparées se partagent donc l'étude de la diversité. Chacune avec ses méthodes et ses mots aborde les problèmes et développe des outils avec un regard propre et

unique. Et pourtant des méthodes identiques ont été développées de nombreuses fois indépendamment dans des disciplines différentes, sous des appellations, des notations différentes, mais avec de grandes similitudes dans leurs interprétations. Au contraire, à l'intérieur d'une même discipline, des évaluations radicalement différentes de la diversité, autant d'un point de vue biologique que mathématique coexistent. Chaque mesure part d'un point de vue sur la diversité.

On constate que les informations les plus capitales concernant la mesure de la biodiversité sont dispersées dans la littérature. Pour cette recherche, des revues biologiques, écologiques, génétiques ou moléculaires, mathématiques et économiques, ont été consultées. Le travail a d'abord commencé par déterminer à partir de quels schémas de données la biodiversité est couramment mesurée. Mes recherches ont pris un tournant à la lecture de l'analyse de variance moléculaire (AMOVA, Excoffier *et al.* 1992), méthode devenue extrêmement populaire en génétique. Cette méthode décompose la variabilité génétique contenue dans des populations structurées selon un schéma similaire à l'analyse de variance (ANOVA, Fisher 1925) hiérarchique mais avec des données multivariées. Dans certains cas, l'AMOVA pourrait être considérée comme une alternative non-paramétrique à l'analyse de variance multivariée (MANOVA). Cette méthode s'insère parfaitement dans le schéma fondamental des mesures de diversité qu'a proposé Rao en 1986 autour d'un indice de diversité nommé l'"entropie quadratique". L'entropie quadratique peut nous permettre d'intégrer des mesures de différences entre espèces dans la mesure de la biodiversité. Toute la thèse s'est alors centrée sur cet indice et toutes les questions posées par la suite ont été formulées à partir de propriétés particulières de l'entropie quadratique, pour arriver finalement à une redéfinition de ce qu'est, ou plutôt ce que devrait être, un indice de biodiversité.

Le plan de cette thèse est le suivant :

Le chapitre 2 pose la structure des données utilisées pour mesurer la biodiversité à toutes les échelles du monde vivant. Les indices traditionnellement utilisés pour mesurer la biodiversité sont donnés. L'entropie quadratique est présentée.

Le chapitre 3 rassemble des méthodes développées en génétique et en écologie, pour décomposer la diversité biologique. La biodiversité existe au niveau des régions comme au niveau des individus. La diversité doit alors pouvoir être décomposée sur les différentes échelles.

Le chapitre 4 traite des représentations graphiques de la diversité pour visualiser précisément la diversité intra et inter-communautés par exemple. La biodiversité peut être vue comme l'inertie de points dans un espace euclidien.

Le chapitre 5 débat d'une contradiction. L'entropie est appelée par Rao (1986) "mesure parfaite de diversité" mais est qualifiée d'"indice faible de diversité" par Ricotta (2002). Pourquoi une telle divergence de point de vue ? Quelle définition ont-ils chacun adoptée pour désigner les "indices de diversité" ? Existe-t-il un ensemble d'axiomes ou de propriétés que doivent vérifier l'ensemble des indices de biodiversité ? Il traite également de la mesure de l'originalité d'une espèce dans un ensemble.

Chapitre 2

Indices de biodiversité

Sommaire

Résumé

A travers les différentes échelles (gène→espèce→règne ; espèce→région) et les différents critères (e.g. phénotypique, fonctionnel) de mesure de la biodiversité, la première partie dégage une structure générale de données basée sur trois objets : entité, catégorie, et collection. Une collection (e.g. communauté) est un ensemble d'entités (e.g. organismes) réparties entre catégories (e.g. espèces).

Une façon de mesurer la diversité est d'étudier cette répartition des entités entre catégories dans une collection, c'est-à-dire d'étudier les abondances relatives des catégories dans la collection. Une liste d'indices de biodiversité est donnée autour de trois indices traditionnels : la richesse, l'indice de Gini-Simpson et l'entropie de Shannon. Bien que rassemblés dans deux formules générales, ils diffèrent cependant par leur sensibilité vis-à-vis des catégories rares. Nous faisons alors une revue bibliographique de leurs propriétés (concavité, espérance, variance), et des tests statistiques qui leur ont été associés.

En montrant les limites de ces indices traditionnels, nous soulignons que, pour la mesure de la diversité, il est primordial de considérer les différences qui séparent les catégories (telles que les distances phylogénétiques entre espèces). Aux mesures de ces différences sont associés des termes mathématiques que nous définissons : dissimilarité, semi-distance, distance, métrique, euclidien, circum-euclidien et ultramétrique. Deux types d'indices de diversité sont présentés. Les premiers, basés sur l'attribution d'une valeur à une catégorie, ont été fortement critiqués. Les seconds, basés sur des matrices de dissimilarités entre catégories, ont l'inconvénient de ne pas tenir compte de l'abondance relative des catégories.

Nous arrivons alors à l'entropie quadratique, indice qui tient compte à la fois des fréquences des catégories et de leurs différences. L'histoire de la création de l'entropie quadratique est retranscrite telle que j'ai pu la découvrir dans la littérature écologique, génétique, et statistique. L'entropie quadratique généralise dans une même formule à la fois l'indice de Gini-Simpson et la variance, et permet ainsi de faire apparaître un schéma général dans la foisonnante littérature sur la mesure de la diversité.

2.1 A ´ - - ´ ?

Williams et Humphries (1996) :

> "A measure of biodiversity must solve two problems. First, what ultimately is to be measured ? Second, how realistically, can appropriate data be obtained ?"

Lorsqu'on fait intervenir un indice de diversité, cela implique qu'avant on a décidé de ce qu'on veut mesurer, ou souvent de ce qu'on est capable de mesurer avec un certain budget et un certain temps alloué. Lorsqu'un groupe de chercheurs mesure la biodiversité, il s'intéresse réellement à ce de quoi il est spécialiste. Les différentes disciplines des sciences de la vie englobent des résolutions très petites, telles que les molécules, et des résolutions très grandes, telles que les paysages. Il est bien sur impossible de déterminer toutes les molécules organiques contenues dans un paysage, si bien que les disciplines se sont spécialisées sur un petit bout de diversité. Par exemple, des généticiens ou des biologistes moléculaires s'intéressent aux molécules d'ADN de petits groupes d'organismes.

Le paragraphe 2.1 ci-dessous s'intéresse ainsi à ce que l'on mesure, c'est à dire aux différents aspects de la biodiversité que l'on est amené à étudier. Les paragraphes 2.2 à 2.4 sont alors centrés sur comment on mesure, c'est à dire quels outils statistiques, quels indices peuvent être utilisés pour évaluer, mesurer chacun de ces aspects de la biodiversité.

2.1.1 Différentes échelles

Ainsi la biodiversité est mesurée à différentes échelles ou résolutions. Dans son livre "Biogéographie, approche écologique et évolutive", Jacques Blondel (2000) affirme que la biodiversité doit être appréhendée à des niveaux différents et interdépendants : diversité génétique, diversité spécifique, diversité des assemblages d'espèces, diversité des écosystèmes au sein des paysages, et diversité dans le temps de systèmes biologiques qui se transforment au fil de l'évolution. De cette description, nous pouvons retenir que la biodiversité doit être mesurée à différentes échelles bio-écologiques, spatiales et temporelles.

Pour l'échelle bio-écologique, Blondel retient la diversité des gènes, des espèces, des assemblages d'espèces et des écosystèmes. En rassemblant les termes des écologues et des généticiens, nous pouvons organiser des niveaux biologiques et écologiques de la façon suivante :

$$\begin{array}{c} \text{règne} \\ \cup \\ \text{embranchement} \\ \cup \\ \text{classe} \\ \cup \\ \text{ordre} \\ \cup \\ \text{famille} \\ \cup \\ \text{genre} \\ \cup \end{array}$$

région $\supset$ paysage $\supset$ écosystème $\supset$ communauté $\supset$ guilde $\supset$ **espèce** $\supset$ population $\supset$ dème $\supset$ organisme $\supset$ gène.

Dans les faits, un de ces niveaux hiérarchiques devient prédominant. Le terme espèce a été mis en gras car il est généralement considéré comme la plaque tournante de toute étude de biodiversité. Pour définir les régions qui devront avoir la priorité dans les stratégies de conservation, l'espèce est considérée comme la forme de biodiversité la plus importante et la plus facilement reconnaissable (Myers *et al.* 2000). Entreprendre des actions de conservation demande un financement dépendant du grand public, qui pourra apprécier plus spontanément le niveau de l'espèce qu'un niveau plus fin et moins directement perceptible tel que celui du gène.

> "It is easier to recognize 'biodiversity' immanent in species - especially charismatic vertebrates or colourful plants - than in gene pools or ecosystem." (May 1995)

L'indice de biodiversité le plus courant, et le plus utilisé en biologie de la conservation, est la richesse égale au nombre d'espèces présentes sur un site. Le niveau de l'espèce sera donc central dans notre description des différentes échelles. Le terme 'espèce' se situe dans la liste ci-dessus à la jointure de deux grandes échelles afin de distinguer l'échelle biologique et évolutive qui part des gènes pour arriver aux règnes, de l'échelle écologique qui aboutit à la région. L'espèce est le plus bas niveau d'une taxonomie qui contient espèce (*Homo sapiens*), genre (Homo), famille (Hominidés), ordre (Primates), classe (Mammifères), embranchement (Chordés), règne (Animaux). Quantifier la diversité à ces niveaux taxonomiques distincts, des gènes aux règnes, est une démarche différente de quantifier la diversité le long de la hiérarchie allant des espèces aux régions. En effet, les hiérarchies taxonomiques et phylogénétiques gènes→espèces→règnes soulignent généralement les origines et relations évolutives alors que les hiérarchies espèces→régions tendent à mettre en évidence des différences ou similarités écologiques contemporaines dans différents cadres environnementaux et géographiques (May 1995). La taxonomie a été présentée avec les sept niveaux de la systématique : espèce, genre, famille, ordre, classe, embranchement, règne. D'autres niveaux y ont été définis tels que le sous-ordre, la sous-classe, etc. Un niveau plus bas que l'espèce a aussi été défini : la sous-espèce. Mais si le concept d'espèce est difficile à définir, celui de la sous-espèce, lui, est très souvent controversé. Une sous-espèce est un ensemble d'organismes d'une même espèce, dotés de traits distinctifs et occupant une certaine partie de l'aire de distribution de l'espèce. Ainsi, plus le nombre de traits considérés pour établir la taxonomie est élevé, plus le nombre de sous-espèces peut être élevé (Wilson 1993). A ce niveau, les classifications évoluent beaucoup au fil des recherches et les familles les plus étudiées telles que les félidés ou les ursidés sont plus susceptibles de contenir beaucoup de sous-espèces uniquement parce que beaucoup de chercheurs s'y sont intéressés. Le niveau de l'espèce est donc généralement préféré à celui de sous-espèce pour mesurer la biodiversité. La considération d'un niveau plus haut que l'espèce permettrait de réduire les coûts d'échantillonnage et ainsi d'étudier plus d'espace. Le nombre de familles fournit une bonne estimation du nombre d'espèces pour de nombreux groupes et de nombreuses régions (Williams et Gaston 1994). Le niveau de l'espèce reste cependant l'unité de prédilection pour l'étude de la biodiversité. Et pourtant il présente un inconvénient majeur : l'absence d'une définition universelle du concept d'espèce. Les critères utilisés pour définir une espèce sont souvent différents d'une classe à l'autre (Faith 1995). Etant donné son importance pour la mesure de la biodiversité, il convient donc de préciser ce que l'on entend par 'espèce'.

> "In short, one of the basic conceptual issues in quantifying biological diversity is the extent to which a 'species' does or does not represent the same unit of evolutionary currency for a bacterium, a protozoan, a mite and a bird." (May 1995)

La notion d'espèce est depuis longtemps l'objet d'un ardent débat autour de trois définitions principales toutes trois jugées équivoques ou insuffisantes (Gayon 1996). La première est biologique : une espèce est un groupe de populations naturelles dont les organismes sont réellement ou potentiellement capables de se reproduire entre eux, et sont incapables de se reproduire avec d'autres groupes d'autres espèces (Mayr 1942). Cette définition est difficilement applicable aux organismes asexués. De plus, nous savons que la reproduction entre individus d'espèces différentes est parfois possible mais les hybrides obtenus ne peuvent eux-mêmes se reproduire. On pourrait donc plutôt proposer la définition biologique suivante : les espèces sont des groupes de populations naturelles dont les organismes sont réellement ou potentiellement capables de se reproduire entre eux, et dont la reproduction éventuelle avec d'autres groupes d'autres espèces conduit à des organismes incapables de se reproduire. La deuxième définition est évolutive : une espèce est une lignée évoluant séparément avec son propre rôle et ses propres tendances (Simpson 1961). La troisième définition est écologique et prend en compte les connexions entre l'isolation des lignées évolutives et les aires adaptatives distinctes qu'elles exploitent (Van Valen 1976, Gayon 1996). La littérature sur le concept d'espèce est très abondante. Gayon (1996) observe que :

> "Modern evolutionary theory is therefore close to suggesting that the notion of a species is a verbal fiction, which directly contradicts its own explicitly stated and most central claims, though perhaps not contradicting Darwin's thinking on the same subject."

Il commente une autre signification donnée par Ghiselin, et avant lui Buffon, à la notion d'espèce : celle d'individu. L'espèce serait un "individu" avec des limites spatio-temporelles, avec des parties et des constituants. Gayon ne défend pas ce concept. Les espèces ne sont probablement pas des groupes avec des limites rigides mais des ensembles flous ("fuzzy sets"), leurs frontières estompées par le transfert horizontal de gènes, l'hybridation, et l'isolation récente (Agapow *et al.* 2004, et références associées). Avec les avancées de la biologie moléculaire, une définition phylogénétique de l'espèce connaît une popularité de plus en plus importante. Une espèce serait, selon cette définition, un groupe d'organismes partageant au moins un caractère dérivé unique, peut-être avec un profil partagé d'ancêtres et de descendants ou monophylie (Agapow *et al.* 2004). Cette définition a l'avantage d'être applicable aux organismes asexués et aux populations allopatriques. Elle conduit à un nombre plus important d'espèces que la définition la plus courante qui est la définition biologique. Et cette inflation du nombre d'espèces se fait à un taux qui dépend de la sophistication des marqueurs et des techniques (Crandall *et al.* 2000, Mace *et al.* 2003). Ce concept d'espèce phylogénétique a eu un impact plus important chez certains taxa et dans certaines régions (Mace *et al.* 2003). Les espèces ainsi définies, comme elles sont plus nombreuses, comprennent chacune moins d'individus et occupent des aires plus restreintes. Il peut donc arriver qu'une espèce, relativement abondante, définie selon le concept biologique soit divisée en plusieurs espèces, chacune relativement rare et donc considérée en danger d'extinction, selon le concept phylogénique. Cela signifie qu'en changeant le concept d'espèce, nous changerions les évaluations de biodiversité qui, actuellement, sont faites à partir du nombre d'espèces, et nous risquerions probablement de changer des ordres de priorité de conservation. L'accélération du rythme des extinctions d'espèces a conduit à l'élaboration de listes répertoriant des espèces du monde entier avec leur statut en terme d'abondance : de

communes à éteintes à l'état sauvage en passant par vulnérables. Les études récentes suggèrent que les longueurs des listes d'espèces en danger d'extinction (Agapow *et al.* 2004), surtout lorsqu'on s'intéresse aux modifications de ces longueurs dans le temps et l'espace, sont plus affectées par la définition du concept d'espèce que par les processus réels d'extinction (Mace *et al.* 2003). En passant du concept biologique au concept phylogénétique, les listes seraient probablement considérablement allongées. Agapow *et al.* (2004) concluent que

> "The best response to the terrible ambiguity of species may be for scientists not to work to reduce it, nor to fear making conservation mistakes, but to learn how to work with it."

Wilson (1993) note que

> "Tous les biologistes n'admettent pas forcément que le concept d'espèce soit solide ou qu'il représente la pierre angulaire sur laquelle repose toute description de la diversité biologique. Ils pensent que ce rôle est mieux rempli par l'idée de gène ou d'écosystème, ou alors ils se contentent de travailler dans l'anarchie conceptuelle. Je pense qu'ils ont tort."

Pour Wilson, le concept d'espèce est crucial pour l'étude de la biodiversité parce que, malgré les erreurs probables dues en partie à la confusion illustrée par ces débats, il s'agit d'une unité potentiellement identifiable. De nombreuses équipes de scientifiques peuvent ainsi travailler sur le même groupe d'organismes, comme *Drosophila melanogaster* en génétique.

En dessous de l'espèce, nous rentrons dans le domaine de l'étude écologique ou génétique des populations. Une population est l'ensemble des organismes appartenant à une même espèce et évoluant dans un même espace au même moment. A ce niveau, les généticiens ont défini l'unité ayant une signification évolutive (ESU "Evolutionary Significant Unit") qui est une population d'organismes reproductivement isolés des autres populations de la même espèce, et représentant un important composant dans l'héritage de l'espèce (DeSalle et Amato 2004). Une population peut être fragmentée lorsque les appariements de ses individus ne sont pas aléatoires. Le dème est alors, dans une population, un ensemble d'organismes qui se croisent entre eux au hasard ; tout organisme a une probabilité égale de s'accoupler avec un autre organisme quelle que soit sa localisation. Un dème est inclus dans une population ; il peut correspondre à cette population. L'organisme, c'est-à-dire l'être vivant, résulte d'une combinaison unique de milliards de paires de nucléotides. Le gène est un segment d'ADN participant à la synthèse d'une protéine. En réalité l'ADN tout entier peut être considéré pour la mesure de la biodiversité. La diversité génétique d'une espèce est l'objet brut le plus basique, et en même temps le plus fondamental, sur lequel les processus évolutifs agissent (May 1995).

Au dessus de l'espèce, se situent les études écologiques des communautés et paysages. Le premier niveau au dessus de celui de l'espèce est la guilde. Une guilde est un ensemble d'espèces vivant en un même lieu et collectant la même nourriture par des moyens semblables. Ce terme peut aussi être défini comme un groupe d'espèces occupant des niches similaires, la niche étant l'habitat et la fonction (l'impact dans cet habitat) d'une espèce ou d'une population. Le second niveau est la communauté. Plusieurs définitions de ce niveau coexistent dans la littérature. Wilson (1993) pose qu'une communauté est un ensemble d'espèces liées par la chaîne alimentaire et par toutes les activités qui influencent leurs cycles vitaux. Il précise que les limites de cet ensemble peuvent rarement être repérées avec exactitude. Root (2001) définit

la communauté comme un assemblage de populations qui coexistent dans une aire. Le terme 'assemblage' est ambigu. Il est souvent employé au niveau des espèces : l'assemblage d'espèces est défini soit comme un ensemble d'espèces vivant au même endroit mais dont le profil d'organisation (compétition, prédation, etc.) est inconnu soit comme une communauté dans laquelle les espèces appartiennent à un groupe taxonomique donné. Et en général, les communautés sont définies et étudiées non pas pour l'ensemble des organismes mais pour un taxon donné. Nous parlerons par exemple de communautés d'oiseaux. L'écosystème est formé d'une communauté (ou biocénose) et du milieu abiotique où celle-ci vie (ou biotope). Le paysage est une mosaïque d'écosystèmes dans une aire géographique donnée. La région ou aire biogéographique est une zone qui peut s'étendre sur les territoires de plusieurs états et qui possède une faune et une flore conditionnée par les mêmes facteurs écologiques tel que le climat.

Tous ces niveaux hiérarchiques sont interdépendants (Gaston 1996, Wagner *et al.* 2000). Chaque fois que nous étudions la biodiversité, nous avons besoin de nous situer à un ou plusieurs niveaux donnés. Bien sûr, cette décomposition hiérarchique est arbitraire. Agir sur un niveau provoque une modification des autres. Les niveaux que nous distinguons sont des repères nécessaire dans la hiérarchie totale de la variation ; l'hétérogénéité existe à chaque niveau (Sarkar et Margules 2002, Faith 2003). Ces échelles écologiques discontinues, c'est-à-dire discrètes, ont fait l'objet d'une théorie hiérarchique (O'Neill *et al.* 1991) selon laquelle les profils observés à différentes échelles dans un paysage structuré hiérarchiquement sont indépendants car causés par des processus isolés à des échelles discrètes (Wagner *et al.* 2000). La diversité biologique est donc extrêmement complexe. Les niveaux que nous distinguons sont une représentation théorique de faits réels.

En plus de ces échelles biologiques et écologiques, on observe que la biodiversité varie dans l'espace (Escudero *et al.* 2003) et le temps (selon les saisons, les années (Warwick *et al.* 2002), les décennies, ou sur des millions d'années) (Chakraborty et Rao 1991). L'abondance relative par exemple n'est pas une propriété figée d'une espèce mais varie dans l'espace et le temps.

2.1.2 Différents critères

Une fois que l'on a choisi une échelle, selon quels critères mesure-t-on la biodiversité ? La démarche la plus tentante, et peut-être la moins coûteuse, est de compter, ou d'estimer, le nombre d'entités à chaque niveau de l'échelle. Mais ces nombres rendent-ils compte de la diversité ? En plus des niveaux hiérarchiques définis dans la partie précédente, il existe d'autres repères, des critères, chacun définissant une facette de la biodiversité. L'utilisation d'un critère donné se justifie par l'objectif précis d'une étude. "Choisir un aspect particulier à mesurer est, en pratique, choisir quel aspect de la biodiversité possède la *valeur* qui doit être mesurée. En conséquence, décider quel aspect de la biodiversité doit être mesuré est un problème plus profond que se prononcer en faveur de la préférence des biologistes moléculaires à compter les nombres de substitutions entre séquences nucléotidiques, celle des taxonomistes à compter les espèces, et celle des écologues à compter les formations végétales." (traduit de Williams et Humphries 1996)

Dans son livre "Ecological diversity and its measurement", Magurran (1988) cite des diversités aussi variées que (dans l'ordre d'apparition) : diversité de la hauteur du feuillage, diversité taxonomique, diversité spatiale, diversité architecturale, diversité structurale (distribution verticale et horizontale des plantes), diversité des stades d'âge au sein d'une communauté, diversité de la taille de la niche (diversité des ressources qu'un organisme ou une espèce utilise). Blondel introduisait les différentes échelles de la biodiversité. Avec cette citation du livre de Magurran, nous avons différentes natures de diversité.

On constate que l'évaluation de la diversité est liée à celui qui la regarde. Prenons la diversité des couleurs par exemple. Deux personnes auront du mal à se mettre d'accord sur la diversité des couleurs qu'ils ont devant eux. Leur querelle peut être résolue avec un peu de physique en mesurant la diversité des longueurs d'onde émises par les objets ou les êtres. L'évaluation de la diversité est dépendante des connaissances, des choix voire des préférences de l'observateur. Par exemple, la longueur d'onde peut être prise pour "marqueur", et le spectrophotomètre est alors un des "outils" pour mesurer la diversité chromatique selon ce marqueur.

Différents types de données peuvent permettre de mesurer la diversité génétique. Chaque type de données est obtenu à partir d'un marqueur qui est ici une région du génome, appelée locus, génétiquement variable, *i.e.* qui produit des variants distincts, ou allèles, quand elle est analysée. Les techniques moléculaires utilisées pour obtenir ces marqueurs diffèrent par la façon dont elles résolvent les différences génétiques, par le type de données qu'elles génèrent et par les niveaux taxonomiques auxquels il est le plus approprié de les utiliser (Karp *et al.* 1996). Il est possible de mesurer la diversité à partir de marqueurs biochimiques tels que les protéines ou allozymes, mais ils présentent l'inconvénient de ne porter que sur une infime partie du génome d'un organisme. Des marqueurs moléculaires ont alors été prônés tels que des microsatellites (Slatkin 1995, Michalakis et Excoffier 1996), des séquences nucléotidiques (Nei et Li 1979, Nei et Tajima 1981, Nei et Jin 1989, Crease *et al.* 1990, Lynch et Crease 1990, Nei et Miller 1990, Holsinger et Mason-Gamer 1996), des sites de restriction (Nei et Tajima 1981, Nei et Miller 1990) (étudiés par RFLP Restriction Fragment Length Polymorphism ou AFLP Amplification Fragment Length Polymorphism) (Zhu *et al.* 1998), et des séquences d'ADN amplifiées aléatoirement (par RAPD Random amplified Polymorphism DNA) (Stewart et Excoffier 1996). Les généticiens, en biologie de la conservation, commencent également à regarder avec intérêt les puces à ADN et les loci responsables de la variabilité de caractères quantitatifs (QTL "quantitative-trait locus") (DeSalle et Amato 2004). Pour choisir un de ces marqueurs, les propriétés suivantes sont regardées : neutralité, co-dominance pour les organismes diploïdes, dispersion, variabilité. Un marqueur est neutre s'il n'est pas soumis à la sélection naturelle. Lorsqu'il est co-dominant, chez les organismes diploïdes, nous n'aurons souvent pas accès au génotype c'est-à-dire aux deux allèles. La dispersion est sa répartition dans le génome. Un marqueur peut être dispersé dans tout le génome, dans l'ADN codant ou dans l'ADN non codant. Ce qui nous intéresse tout particulièrement pour la mesure de la diversité, c'est sa variabilité. Un marqueur permettant de mesurer la diversité est un marqueur prenant plusieurs formes, il est donc dit polymorphe. Un locus est souvent considéré comme polymorphe lorsque la fréquence de l'allèle le plus commun est inférieure ou égale à 0.99 ou 0.995. Les microsatellites, se situant surtout dans l'ADN non codant mais aussi dans l'ADN codant, ont l'avantage d'avoir une forte variabilité. Les méthodes telles que la RAPD et l'AFLP fournissent deux allèles par

locus : présence ou absence d'une bande sur un gel d'électrophorèse. Ce polymorphisme bas est compensé par le nombre de bandes (Russell *et al.* 1997).

La diversité génétique est souvent mesurée entre individus au sein d'une population, donc pour une espèce donnée. Lorsqu'on s'intéresse à plusieurs espèces, les caractéristiques moléculaires de chaque espèce sont plus difficiles à évaluer. Swingland (2001) note que dans la stratégie de conservation mondiale (IUCN/UNEP/WWF 1980) la taxonomie, largement basée sur la morphologie, est utilisée comme substitut pour refléter les différences génétiques entre espèces. Il est d'autre part maintenant possible au moins pour certains taxons de mesurer une diversité phylogénétique. Erwin (1991) affirme que la richesse spécifique à un site est un indice aisément observable du nombre d'interactions entre espèces et de la façon dont ces espèces sont groupées et forment une unité vivante sur ce site. Cependant il affirme aussi que pour comprendre la signification de la valeur d'un indice de biodiversité dans plusieurs zones géographiques, on a besoin d'un contexte et que ce contexte est fourni par les relations entre les espèces et la connaissance des lignées auxquelles elles appartiennent.

Comme le montre Magurran (1988), il est possible de s'intéresser à un seul caractère ou une seule propriété. La taille des organismes est l'un des principaux critères utilisés pour mesurer la biodiversité (*e.g.*, Staudhammer et LeMay 2001). Par exemple, en réponse à l'accès à la lumière les plantes peuvent avoir différentes tailles. Les herbivores vont consommer ces différentes plantes et vont évoluer eux-mêmes vers différentes tailles. Ils vont donc constituer des proies de tailles différentes pour leurs prédateurs, etc. Le fait qu'il existe des plantes de tailles différentes permet une diversification des herbivores et de leurs prédateurs. A l'inverse la présence de prédateurs de tailles différentes rend possible la diversification des herbivores et des plantes (Whittaker 1972). Des espèces de tailles différentes sont susceptibles d'occuper des niches différentes, et donc de remplir des fonctions différentes dans la communauté. Elles présentent alors moins de compétition. Une communauté possédant des espèces de tailles variées auraient ainsi une plus forte adaptabilité.

Outre la taille, d'autres mesures du corps peuvent intervenir sur la définition de la niche, ce qui nous amène à la diversité morphométrique. L'hypothèse selon laquelle la morphométrie reflète les différents moyens par lesquels une espèce utilise les ressources disponibles et donc exploite sa niche a été soutenue par de nombreux auteurs. De plus, la morphométrie fournit une information différente de celle des analyses génétiques puisque les variations morphométriques comprennent deux composants : un composant génétique et un composant non-génétique dû aux conditions environnementales sous lesquelles un individu se développe (James 1983, Merilä *et al.* 2001). Parfois aussi, de faibles différences génétiques sont à l'origine de fortes différences phénotypiques.

La recherche de l'existence et de la nature des forces influant sur l'évolution des espèces et des communautés d'espèces est un problème majeur en écologie. Plusieurs types d'évolution peuvent être observés. L'évolution divergente est celle d'une même espèce qui présente différentes adaptations. Un cas particulier de l'évolution divergente peut être provoqué par la compétition (Schluter 2000a). Lorsque deux espèces ont des niches écologiques similaires, elles peuvent présenter un déplacement de caractères lorsqu'elles sont en sympatrie. Lorsque ce processus apparaît un grand nombre de fois, il peut conduire à la formation de nouvelles espèces à

partir d'une seule. On parle alors de radiation adaptative : prolifération relativement rapide de nouvelles espèces à partir d'un même ancêtre, accompagnée par une extension vers de nouvelles ressources et de nouveaux environnements et par la divergence dans les traits phénotypiques utilisés pour exploiter ces environnements (Schluter 2000b). L'évolution convergente correspond à des adaptations similaires, chez des organismes relativement distants, à cause de pressions évolutives communes. L'évolution parallèle correspond à l'évolution convergente chez des organismes relativement proches. Ces trois types d'évolution sont très liés (*e.g.*, Rüber et Adams 2001).

Des cas de radiations adaptatives ou d'évolution divergente au niveau de la morphologie ont été observés. Un exemple très connu d'évolution divergente est celui des fringillidés de Darwin dans l'archipel des Galapagos. Peter et Rosemary Grant ont passé 30 ans, dont deux articles récents font le bilan (Grant et Grant 2002, Hosken et Balloux 2002), à les étudier. D'autres exemples ont été donnés chez les lézards Phynosomalid (*Urosaurus ornatus*) (Herrel *et al.* 2001), chez des scarabées rhinocéros géants lorsqu'ils sont en sympatrie (Kawano 2002), chez des poissons osseux de l'espèce *Galaxias platei* en fonction de la quantité de présence de prédateurs (Milano *et al.* 2002). Schluter (2000a) ouvre une discussion sur les différences morphologiques entre espèces dues à la compétition. Losos et Miles (2002) proposent une méthode pour tester l'existence d'une radiation adaptative.

En ce qui concerne l'évolution parallèle, un exemple en est donné par les lézards Scincid d'Amérique du Nord (Richmond et Reeder 2002). Les ressemblances morphologiques entre la tête d'un insecte et celle d'un myriapode pourraient résulter de contraintes environnementales communes et d'adaptations fonctionnelles parallèles (Hwang *et al.* 2001).

Des cas de convergences morphologiques ont été observés chez des grands scarabées en milieu désertique (Chown *et al.* 1998), chez les lézards Anolis (Irschick et Losos 1998, 1999, Losos 1990a, b, c, 1992), chez les colibris et chez les mites des fleurs (Colwell 2000), chez des espèces de rongeurs présentes dans des régions de dunes d'Israël et dans plusieurs guildes d'Amérique du Nord (Ben-Moshe *et al.* 2001). A l'inverse, certaines espèces ne semblent pas présenter de convergence morphologique. Par exemple Leal *et al.* (2002) proposent une discussion sur l'absence de convergence morphologique chez des espèces de lézards aquatiques Anolis des Antilles, alors que la convergence morphologique chez de nombreuses espèces de lézard Anolis non aquatiques de cette région a souvent été observée. Ils proposent, parmi d'autres, l'hypothèse suivante pour expliquer cette absence de convergence : il n'existerait pas une seule façon de s'adapter à un même habitat.

Ainsi l'information fournie par la morphologie peut être très différente de celle fournie par les analyses génétiques. La morphométrie est aussi étudiée pour son importance fonctionnelle. Certains traits morphologiques des oiseaux peuvent être regroupés en complexes fonctionnels ou complexes d'adaptation reflétant le degré de liaison avec des variables trophiques et des variables de substrat. Ces groupements dépendent de l'espèce étudiée (*e.g.*, Rothstein 1973, James 1982, Leisler et Winkler 1985). Chez les oiseaux, d'année en année, la longueur ou la forme d'appendices tels que l'aile, les tarses ou la queue, peuvent changer selon des modifications de l'environnement liées entre autres à la recherche de nourriture (Hespenheide 1973, Rothstein 1973, Carrascal *et al.* 1990), et au degré de présence de compétiteurs (Willson 1969).

La diversité peut être mesurée sur des données biologiques pour de multiples raisons. Dans le cadre de la biologie de la conservation, que souhaite-t-on sauver ? Pour Williams *et al.* (1994),

il n'y a pas de justification théorique particulière à considérer les attributs génétiques comme la "monnaie" de la conservation. Au contraire, selon eux, une plus grande valeur pour les attributs phénotypiques héritables peut être justifiée pas tellement parce qu'il s'agit d'une monnaie directement perçue mais parce que ces attributs peuvent être directement plus utiles à l'homme.

Ainsi différents aspect de la biodiversité peuvent être mesurés selon que l'on choisit d'étudier les gènes, la morphométrie ou d'autres caractères des êtres vivants. Une fois qu'un critère a été choisi, disons, la morphométrie, il est impossible d'amasser toutes les informations existantes sur la forme et la taille des membres ou parties des corps de beaucoup d'organismes. Il faut toujours choisir et se restreinte à plusieurs mesures spécifiées telles que la taille des tarses, la longueur de la queue, etc. A chaque fois que l'on mesure la biodiversité, on mesure un aspect de cette biodiversité qui dépend étroitement de tous ces choix.

2.1.3 Schéma général

D'une façon générale, avec la plupart des indices actuels, la diversité est mesurée dans une **collection** d'**entités**. Les entités sont souvent réparties en groupes que nous appellerons de façon générale **catégories**. Souvent en écologie, les collections désignent des zones d'études, les catégories représentent des espèces, et les entités des organismes. Dans ce cas, la densité est parfois mesurée en terme de biomasse. A un autre niveau, les collections peuvent être des populations d'une même espèce, les entités des individus et les catégories des profils d'ADN. Les catégories peuvent être recensées dans une collection sous forme de présence/absence, d'abondance ou de fréquence (ou encore de densité). En résumé, en biologie, les entités peuvent être de natures très variées, d'une séquence d'ADN à des paysages. Nous pourrons garder en mémoire que tous les indices présentés peuvent être appliqués à toute nature de données pourvu que leurs formes soient du type entité $\in$ (catégorie $\times$ collection) (Fig. 1). Si toutes les entités sont différentes, c'est-à-dire, si selon le critère utilisé, il est impossible de répartir les entités dans des groupes alors nous considérerons chaque entité comme une catégorie. Dans ce cas, une catégorie aura une abondance de une entité et une fréquence de 1/nombre total d'entités.

2.2 D ´ ,

2.2.1 Présentation des indices

Nous nous intéresserons ici à trois indices très utilisés en écologie et génétique. L'indice le plus simple et le plus utilisé pour mesurer la biodiversité est le nombre S de catégories diminué de 1, afin qu'une collection comprenant une seule catégorie ait une biodiversité nulle. Cet indice s'écrit

$$H_{\mathrm{r}} = S - 1,$$

et est appelé richesse.

Beaucoup ont argumenté contre ce type d'indices en affirmant que les fréquences des catégories doivent être considérées pour mesurer la biodiversité (Fig. 2). Shannon (1948), Gini

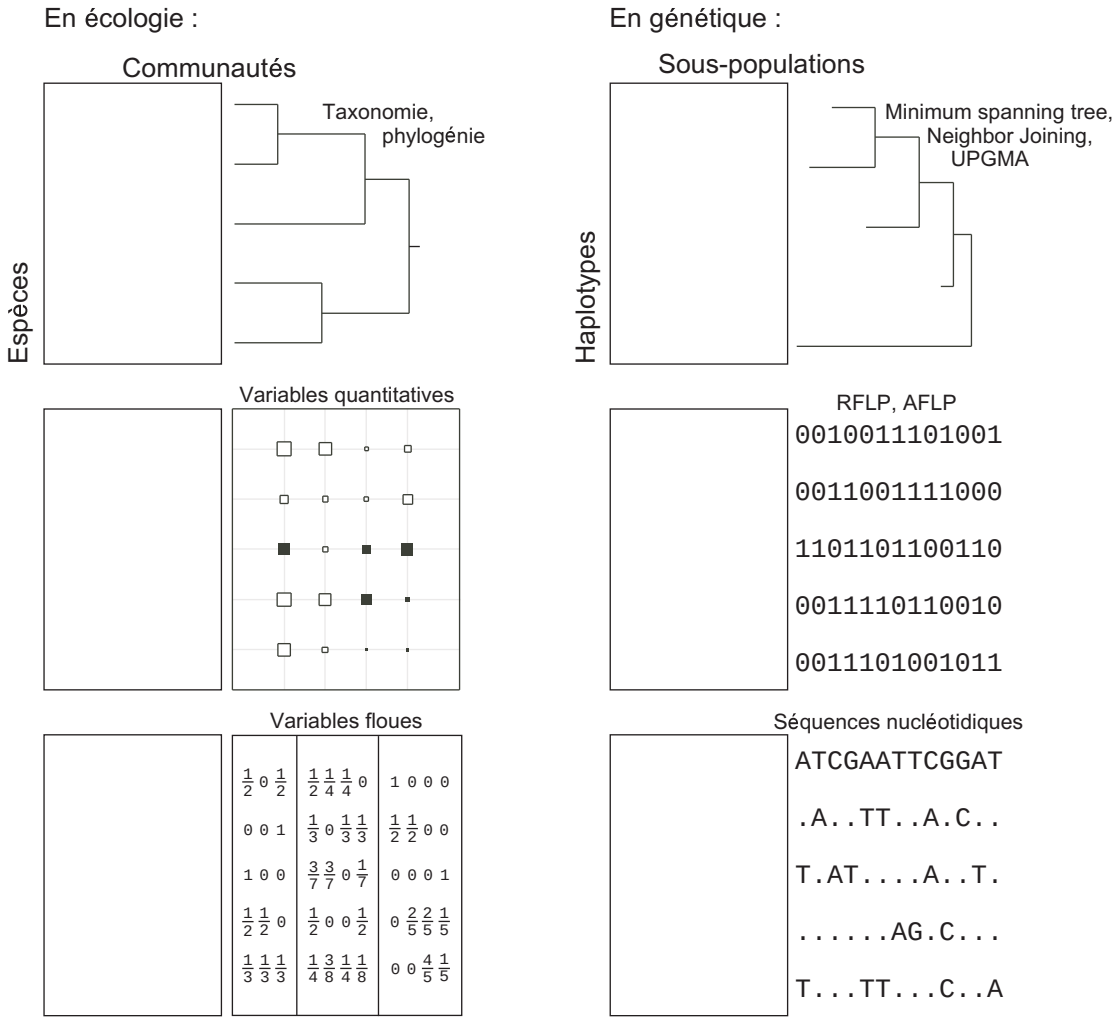


FIG. *1 – Exemples de jeux de données utilisés pour mesurer la biodiversité. Les rectangles vides représentent des tableaux d'abondance ou de présences/absences.*

et Simpson (Gini 1912, Simpson 1949) ont proposé des indices corrigeant la richesse par les fréquences relatives des catégories. Ces indices sont donc tous définis sur l'ensemble

$$\mathcal{P} = \left\{ \mathbf{p} = (p_1, ..., p_k, ..., p_S),\ p_k \geq 0\ \forall k = 1, ..., S,\ \sum_{k=1}^{S} p_k = 1 \right\}.$$

Shannon (1948) a développé son indice H_S dans le cadre de la théorie de l'information qui suppose que la diversité peut être mesurée de la même façon que l'information contenue dans un code ou un message. Les indices développés dans le cadre de cette théorie sont qualifiés de fonctions d'entropie. Soient p_k la fréquence de la catégorie k, et $\mathbf{p} = (p_1, ..., p_k, ..., p_S)$ la distribution de fréquences des catégories, l'indice de Shannon est

$$H_S(\mathbf{p}) = -\sum_{k=1}^{S} p_k \ln(p_k).$$

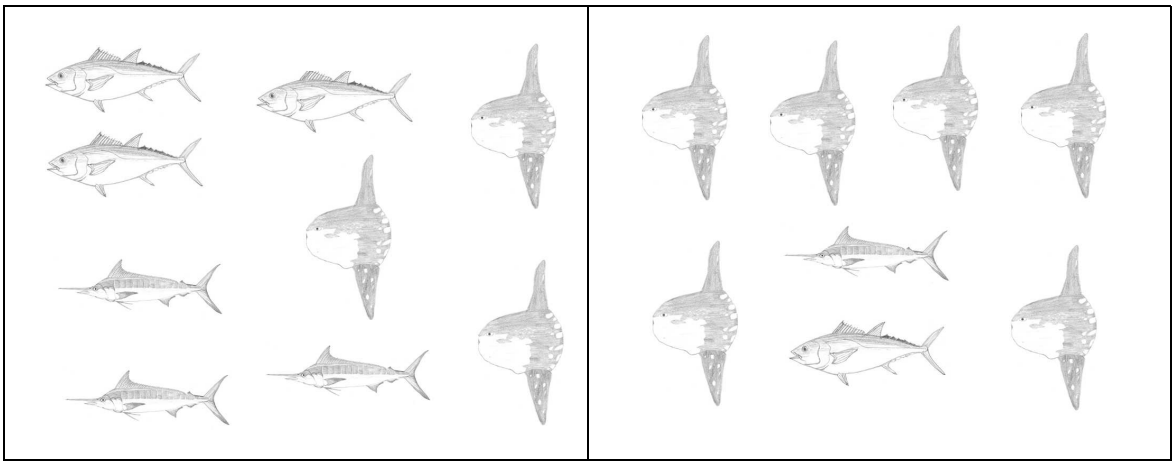


FIG. *2 – La diversité dépend des fréquences des catégories. La diversité de l'assemblage théorique de gauche est plus grande que celle de l'assemblage de droite, ce dernier possédant une catégorie dominante. Actuellement* Mola mola *(poisson lune) est une espèce commune ; les deux autres* Thunnus thynnus *(thon rouge) et* Makaira nigricans *(marlin bleu) sont classés vulnérables dans la liste de l'IUCN ("International union for the conservation of nature").*

Cette mesure peut être interprétée de plusieurs façons. Une première interprétation est que l'indice de Shannon mesure la perte d'information due à la perte d'une entité. Si toutes les entités appartiennent à la même catégorie, c'est-à-dire si la diversité est nulle, alors nous ne perdons aucune information en retirant une de ces entités. A l'opposé, si toutes les entités appartiennent à des catégories différentes, la perte d'information due à la perte d'une entité est grande. Une autre interprétation possible est que l'indice de Shannon est une mesure d'incertitude. Si nous tirons au hasard une seule entité de la collection, l'indice de Shannon mesure l'incertitude que nous avons sur le résultat, c'est-à-dire : quelle catégorie allons-nous tirer ? Par exemple, si toutes les entités appartiennent à la même catégorie, alors nous sommes certains de tirer la catégorie en question. Le degré d'incertitude est nul. La diversité est nulle. A l'opposé, si toutes les entités appartiennent à des catégories différentes, l'incertitude quant au résultat est maximale. La diversité est dans ce cas maximale pour le nombre de catégories présentes dans la collection.

L'indice $H_{\text{G-S}}$ de Gini-Simpson (Gini 1912, Simpson 1949) est égal à la probabilité de tirer, avec remise, dans une collection deux entités appartenant à deux catégories différentes :

$$H_{\text{G-S}}(\mathbf{p}) = 1 - \sum_{k=1}^{S} p_k^2.$$

Gini (1912) l'avait suggéré comme mesure de diversité écologique. Simpson (1949) propose de nouveau cet indice à partir d'une mesure de concentration $\lambda = \sum_{k=1}^{S} p_k^2$ qui donne une valeur à la répartition des entités en catégories. Ainsi λ est égale à $1/S$ pour une communauté présentant la plus petite concentration, c'est-à-dire autant d'entités dans chaque catégorie, et donc la plus grande diversité. A l'opposé, λ est égale à 1 pour une communauté présentant la plus grande concentration, c'est-à-dire lorsque toutes les entités appartiennent à une seule catégorie, ce qui correspond à la plus petite diversité. La diversité est donc inversement proportionnelle

TAB. *1 – Autres indices d'entropie. Nayak (1983)*

$H_{\mathrm{R}} = \log\left(\sum_{k=1}^{S} p_k^{\alpha}\right) / (1-\alpha),\ \alpha > 0,\ \alpha \neq 1$	Indice de Renyi
$H_{\gamma} = \left[1 - \left(\sum_{k=1}^{S} p_k^{\alpha}\right)^{1/\alpha}\right] / \left[1 - 2^{1/\alpha - 1}\right],\ \alpha > 0,\ \alpha \neq 1$	Entropie-γ
$H_{\mathrm{P}} = -\sum_{k=1}^{S} p_k \log p_k - \sum_{k=1}^{S} (1-p_k)\log(1-p_k)$	Entropie appariée
$H_{\mathrm{D}} = 1 - \sum_{k=1}^{S} p_k^2 - \sum_{k=1}^{S} p_k^2\left(1-p_k^2\right)$	Indice de Rao (1982c)

à la concentration. Plusieurs indices de diversité associés à λ ont été proposés : $1/\lambda$, $-\ln(\lambda)$ et $1 - \sum_{k=1}^{S} p_k^2$. Ce dernier indice est principalement connu sous le nom d'indice de Gini-Simpson ($H_{\text{G-S}}(\mathbf{p})$). En écologie, il est beaucoup utilisé pour mesurer la diversité spécifique, les catégories étant des espèces. En génétique, dans une population panmictique, il est équivalent à la proportion d'individus hétérozygotes, c'est pourquoi l'indice de Gini-Simpson est également appelé "hétérozygotie". Les catégories sont alors des allèles et la collection une population. La mesure de concentration est alors appelée homozygotie. L'indice de Gini-Simpson, lorsqu'il est mesuré à l'intérieur d'une population dans laquelle les individus ne s'apparient pas aléatoirement, est appelé diversité de gènes ou diversité allélique. La mesure de concentration est alors appelée identité allélique. L'inverse de l'identité allélique ($1/\lambda$) est appelée nombre efficace d'allèles, il est égal au nombre réel d'allèles seulement si les allèles ont les mêmes fréquences, et il estime le nombre d'allèles de même fréquence qui auraient produit la même diversité allélique que la diversité observée. La notion de diversité allélique a été introduite par Nei (1973), l'indice est donc également connu en génétique sous le nom de diversité de Nei. Il désigne la probabilité que deux gènes tirés aléatoirement dans une population soient différents dans leur type génétique. Nayak (1983) fait également remarquer son utilisation en économie (Agresti et Agresti 1978) et en linguistique (Greenberg 1956). Son usage s'étend maintenant aussi à la microbiologie : diversité en OTU "operational taxonomic units" (Hill *et al.* 2003).

D'autres indices d'entropie ont été étudiés, dont deux basés sur la mesure $\sum_{k=1}^{S} (p_k)^{\alpha}$ (Tab. 1).

Des liens unissent la richesse H_{r}, l'indice de Shannon H_{S} et celui de Gini-Simpson $H_{\text{G-S}}$. Hill (1973) introduit la formule suivante

$$N_a = \left(p_1^a + p_2^a + \ldots + p_S^a\right)^{1/(1-a)}$$

En faisant varier la valeur de a, on retrouve des indices connus. $N_{-\infty}$ est égal à l'inverse de la fréquence de la catégorie la plus rare ; N_0 est une fonction de la richesse : $N_0 = S = H_{\mathrm{r}} + 1$; N_1 est une fonction de l'indice de Shannon : $N_1 = \exp(H_{\mathrm{S}})$; N_2 une fonction de l'indice de Gini-Simpson : $N_2 = 1/(1 - H_{\text{G-S}})$; N_{∞} est égal à l'inverse de la fréquence de la catégorie la plus commune. Lorsque $a = 2$, l'indice de Hill s'écrit simplement $N_2 = 1/\lambda$. Hill (1973) affirme alors que pour unifier les mesures de diversité en les rassemblant dans une même formule, l'in-

dice de Gini-Simpson devrait être abandonné au profil de $1/\lambda$ ou éventuellement $-\ln(\lambda)$.

Mais un autre indice, celui de Havrda et Charvat (1967) généralise lui aussi la richesse, l'indice de Shannon et celui de Gini-Simpson, dans une seule formule :

$$H_{\text{H-C}}(\mathbf{p}) = \frac{1 - \sum_{k=1}^{S} p_k^{\alpha}}{\alpha - 1}, \alpha \geq 0, \alpha \neq 1.$$

Lorsque $\alpha = 0$, $H_{\text{H-C}}(\mathbf{p}) = H_R$; quand $\alpha \to 1$, $H_{\text{H-C}}(\mathbf{p}) \to H_S(\mathbf{p})$; et si $\alpha = 2$, alors $H_{\text{H-C}}(\mathbf{p}) = H_{\text{G-S}}$. L'indice de Havrda et Charvat comme celui de Shannon est un indice d'entropie mais pour d'autres raisons : il est utilisé en physique depuis les travaux de Tsallis (Cho 2002). Le terme d'entropie est utilisé en physique pour mesurer le degré de désordre d'un système et aussi, comme nous l'avons vu, en communication pour mesurer le degré d'incertitude du contenu d'un message. Ainsi dans la formule de l'indice de Havrda et Charvat c'est la forme $1 - \lambda$ et non $1/\lambda$ qui est utilisée.

A partir de $H_{\text{H-C}}$, Patil et Taillie (1982) montrent que les trois indices, H_{r}, H_{S} et $H_{\text{G-S}}$, peuvent être réécrits comme des moyennes d'une fonction de rareté $R(p_k)$, la rareté d'une catégorie diminuant lorsque sa fréquence augmente :

$$H_{\text{H-C}}(\mathbf{p}) = \sum_{k=1}^{S} p_k \left(\frac{1 - p_k^{\alpha-1}}{\alpha - 1} \right).$$

Selon ces indices, la rareté d'une catégorie est donc égale à

$$R(p_k) = \begin{cases} \left(1 - p_k^{\alpha-1}\right) / (\alpha - 1) & \text{si } \alpha \neq 1 \\ -\ln(p_k) & \text{si } \alpha = 1 \end{cases}$$

Les trois indices H_{r}, H_{S} et $H_{\text{G-S}}$ ont les fonctions de rareté respectives :

$$\begin{aligned} R_{\text{r}} &= \frac{1}{p_k} - 1 \\ R_{\text{S}} &= -\ln(p_k) \\ R_{\text{G-S}} &= 1 - p_k. \end{aligned}$$

Ils diffèrent par leur sensibilité vis à vis des espèces rares (Fig. 3). Le plus sensible est H_{r} et le moins sensible $H_{\text{G-S}}$. La caractéristique des indices de Gini-Simpson et de Shannon est qu'ils donnent un faible poids aux catégories rares (surtout l'indice de Gini-Simpson). Leurs définitions s'appuient sur l'inconvénient de la richesse de ne pas considérer les fréquences des catégories et donc de donner, selon un certain point de vue, trop de poids aux catégories rares. Ce point de vue est celui de l'homogénéité des collections. Si une collection contient une catégorie dominante et plusieurs rares, le poids des catégories rares y est plus faible que dans une collection où les catégories sont présentes à la même fréquence. Lorsque les valeurs de ces indices sont estimées, le nombre de catégories observées dans un échantillon est très dépendant de la taille de l'échantillon, alors que les indices de Gini-Simpson et de Shannon atteignent plus

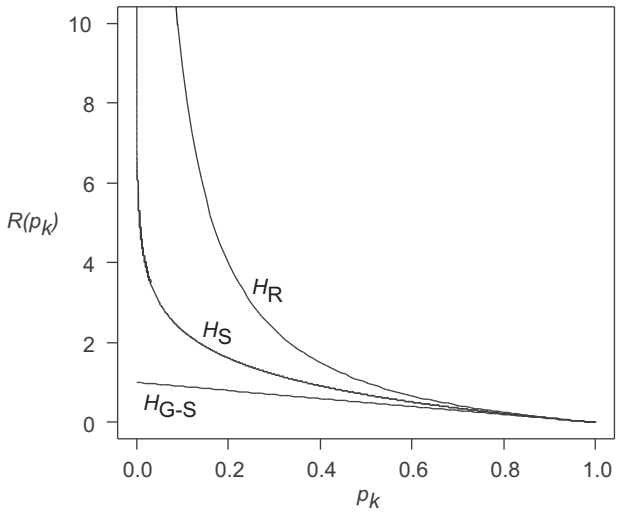


FIG. *3 – La rareté est une fonction décroissante de la fréquence. Les fonctions de rareté associées à trois indices sont considérées : la richesse H_r, l'indice de Shannon H_S et celui de Gini-Simpson $H_{G\text{-}S}$*

vite leur valeur asymptotique.

Les deux indices de diversité en catégories, H_S et $H_{G\text{-}S}$, se contredisent parfois lors du classement des collections (Patil et Taillie 1982). De ces deux indices de diversité en catégories, H_S est considéré comme le plus utilisé actuellement (Magurran 2004). La signification de cet indice a souvent été discutée :

> "This measure is not directly related to any genetic entity" (Nei 1973) ;
>
> "It does not have a clear cut biological meaning" (Chakraborty et Rao 1991).

Whittaker (1972) admet qu'il n'y a aucune raison particulière pour interpréter la diversité comme information ou incertitude, mais avance que l'indice a des qualités distinctives et appropriées. Il affirme que l'indice de Gini-Simpson est fortement affecté par les 1, 2 ou 3 catégories les plus abondantes, et que l'indice de Shannon au contraire est plus affecté par les catégories moyennes. Etant ainsi moins affecté par les catégories abondantes et rares, l'indice de Shannon serait selon Whittaker peu sensible aux biais d'échantillonnage. Whittaker en déduit donc que $H_{G\text{-}S}$ est un indice moins efficace que H_S.

En réalité, comme l'indice de Shannon donne une plus forte valeur de R que l'indice de Simpson, le produit $p_k \times R(p_k)$ est plus équilibré pour l'indice de Shannon, c'est-à-dire que sa valeur pour les espèces rares est plus proche de celle pour des espèces abondantes avec l'indice de Shannon qu'avec celui de Gini-Simpson. Whittaker appelle donc l'indice de Shannon, indice d'"équitabilité" ("equitability"). Contrairement au point de vue de Whittaker, il est donc possible de considérer que l'indice de Shannon est très sensible aux catégories rares et qu'il fournit donc une valeur exagérée de la diversité (Rao 1982a). Pour ces raisons, l'indice de Shannon semble désavantagé par rapport à d'autres indices, en particulier par rapport à l'indice de Gini-Simpson.

2.2.2 Propriétés

H_r, H_S, $H_{G\text{-}S}$ et $H_{H\text{-}C}$ sont concaves (Nayak 1983, Lande 1996) c'est-à-dire qu'ils augmentent dans un mélange.

Une fonction f, définie sur un espace E (par exemple $\mathbb{R}$) peut être concave ou convexe. Elle est concave si et seulement si, soit α et β deux nombres réels tels que $\alpha + \beta = 1$ et a et b deux objets de E,

$$f(\alpha a + \beta b) \geq \alpha f(a) + \beta f(b).$$

La fonction est convexe si et seulement si

$$f(\alpha a + \beta b) \leq \alpha f(a) + \beta f(b).$$

Pour que ces définitions puissent être applicables il faut que E vérifie une propriété particulière : si a et b appartiennent à E, $\alpha a + \beta b$ appartient aussi à E. L'ensemble $\mathcal{P}$ des vecteurs de fréquences vérifie cette propriété.

La concavité est une des propriétés importantes pour mesurer la diversité totale d'une région et la décomposer en un composant de diversité au sein d'habitats de cette région et un composant de diversité entre les habitats, mesurant les différences entre habitats (cf. partie 3).

Nous avons considéré pour l'instant que les fréquences des catégories ainsi que le nombre de catégories dans la collection sont connus. En réalité, il s'agit souvent de valeurs observées obtenues à partir d'un échantillon. A chaque fois que l'on fait intervenir l'inférence, beaucoup de résultats différents sont fournis par la variété des données sur lesquelles la diversité est mesurée.

Notons $\hat{S}$ l'estimateur de la richesse dans un échantillon. Il s'agit simplement du nombre de catégories observées. Notons également π_k l'abondance relative réelle de la catégorie k dans la collection à partir de laquelle l'échantillon a été tiré, et n le nombre d'entités dans l'échantillon. $\hat{S}$ a pour espérance

$$E\left(\hat{S}\right) = S - \sum_{k=1}^{S} (1 - \pi_k)^n$$

(Grassle et Smith 1976) et pour variance

$$V\left(\hat{S}\right) = \sum_{k=1}^{S} (1 - \pi_k)^n \left[1 - (1 - \pi_k)^n\right] + 2 \sum_{k=1}^{n-1} \sum_{l=k+1}^{n} \left[(1 - \pi_k - \pi_l)^n - (1 - \pi_k)^n (1 - \pi_l)^n\right]$$

(Lande 1996). La catégorie k a une forte chance d'être présente dans l'échantillon seulement si $\pi_k n > 1$. Ainsi, l'estimateur $\hat{S}$ a de fortes chances de sous-estimer la richesse d'une collection contenant beaucoup de catégories rares.

En écologie, où les catégories sont des espèces, des modifications de l'indice de la richesse ont été proposées pour tenir compte du nombre d'individus échantillonnés (n) : $(S - 1)/\ln(n)$ (Margalef 1958), $S/\sqrt{n}$ (Menhinick 1964). La division par une fonction du nombre total d'individus a pour but d'éliminer les biais d'échantillonnage. En effet un effort d'échantillonnage plus grand aurait peut-être permis d'observer plus d'espèces.

La mesure la plus simple de l'effort d'échantillonnage, et qui est considérée dans les indices de Margalef et Menhinick, est le nombre d'individus observés. D'autres mesures telles que la surface de la zone étudiée ou le temps d'observation sont également utilisées. Au lieu de diviser la richesse par l'effort d'échantillonnage, il a été proposé d'utiliser différents degrés d'effort d'échantillonnage pour estimer la richesse par extrapolation. On suppose alors que la richesse augmente lorsqu'on augmente l'effort d'échantillonnage jusqu'à atteindre un seuil à partir duquel la richesse n'augmente plus ou très peu par une augmentation supplémentaire de l'effort d'échantillonnage. En écologie, la première méthode d'estimation de la richesse en espèces est l'utilisation de courbes d'accumulation d'espèces. Les courbes d'accumulation des espèces sont des graphiques représentant le nombre d'espèces découvertes dans une zone donnée, en fonction d'une mesure de l'effort d'échantillonnage. Les extrapolations supposent donc que ces courbes atteignent un plateau. La dépendance des mesures de richesse vis à vis de l'effort d'échantillonnage est un des problèmes majeurs dans l'estimation de la diversité particulièrement dans les environnements marins où on atteint rarement l'asymptote de la relation espèce-aire (Warwick et Clarke 2001) et chez les microorganismes où

> "In a sample with 4000 species it would take 25,000 samples (i.e. clones) to reveal 2000 different species and 285,000 samples to be reasonably sure of sampling the 2000 most abundant species." (Curtis et Sloan 2004)

Toujours en écologie, des méthodes paramétriques et non-paramétriques d'estimation de la richesse en espèces ont été développées. Les méthodes paramétriques utilisent la forme de la distribution d'abondance des espèces. Les méthodes non-paramétriques sont basées sur l'hypothèse que les espèces manquantes (non-observées) sont des espèces rares. Des bilans sur l'ensemble de ces méthodes sont donnés par Colwell et Coddington (1994) et Magurran (2004).

Prenons par exemple les méthodes non-paramétriques de Chao (1984, 1987) (voir Colwell et Coddington 1994), le premier estimateur de la richesse d'un assemblage est

$$S_1^* = S_{\text{obs}} + \left(a^2/2b\right),$$

où S_{obs} est le nombre d'espèces observées dans un échantillon, a est le nombre d'espèces observées qui sont représentées par un seul individu et b est le nombre d'espèces observées représentées par exactement 2 individus dans cet échantillon. Cet estimateur suppose que des données d'abondance des espèces ont pu être obtenues. Sa variance est

$$V(S_1^*) = b\left[\left(\frac{a}{4b}\right)^4 + \left(\frac{a}{b}\right)^3 + \left(\frac{a}{2b}\right)^2\right].$$

Si seules des données de présences/absences sont connues, il est alors nécessaire d'analyser plusieurs échantillons de l'assemblage. Un deuxième estimateur

$$S_2^* = S_{\text{obs}} + \left(L^2/2M\right)$$

prend alors en compte le nombre L d'espèces qui n'ont été observées que dans un échantillon et le nombre M d'espèces observées dans exactement 2 échantillons. La variance de cet estimateur est

$$V(S_2^*) = M\left[\left(\frac{L}{4M}\right)^4 + \left(\frac{L}{M}\right)^3 + \left(\frac{L}{2M}\right)^2\right].$$

Avant de parler d'inférence sur des indices prenant en compte des fréquences, regardons d'abord les valeurs maximales. La richesse augmente linéairement avec le nombre de catégories. Pour un nombre de catégories S fixé, les indices traditionnels de diversité prenant leurs valeurs dans l'ensemble $\mathcal{P} = \left\{\mathbf{p} = (p_1, ..., p_k, ..., p_S),\ p_k \geq 0\ \forall k = 1, ..., S,\ \sum_{k=1}^{S} p_k = 1\right\}$ présentent un maximum. Ainsi, les indices H_{S}, $H_{\text{G-S}}$, $H_{\text{H-C}}$, et aussi tous ceux du tableau 1, atteignent leur valeur maximale pour la distribution uniforme $\left(\frac{1}{S}, ..., \frac{1}{S}, ..., \frac{1}{S}\right)^t$. La valeur maximale de $H_{\text{H-C}}$ est donc

$$\max_{\mathbf{p}} \left(H_{\text{H-C}}\left(\mathbf{p}\right)\right) = \begin{cases} \frac{S^{\alpha-1}-1}{(\alpha-1)S^{\alpha-1}} & \text{si } \alpha \geq 0 \text{ et } \alpha \neq 1 \\ \ln\left(S\right) & \text{si } \alpha = 1 \end{cases}$$

Ainsi l'indice H_{S} dont la valeur maximale est $\ln\left(S\right)$ n'est pas borné ; alors que $H_{\text{G-S}}$ est borné par 0 et 1 puisque sa valeur maximale

$$\max_{\mathbf{p}} \left(H_{\text{G-S}}\left(\mathbf{p}\right)\right) = \frac{S-1}{S}$$

tend vers 1 lorsque S tend vers l'infini. La valeur de ces indices, peut être considérée au maximum comme une mesure de richesse parce qu'il s'agit de fonctions monotones croissantes du nombre de catégories

$$H\left(1/S, 1/S, ..., 1/S\right) \leq H\left(1/(S+1), 1/(S+1), ..., 1/(S+1)\right).$$

Il a été proposé de diviser les indices de Shannon et de Gini-Simpson par leur maximum respectif afin qu'ils deviennent indépendants de la richesse. Les indices ainsi obtenus mesurent le degré d'"équirépartition" des entités dans les catégories et sont appelés mesures d'égalité ou mesures d'équitabilité ("evenness measures"). Ainsi un indice de diversité en catégories est le produit d'une mesure d'égalité et d'une fonction de la richesse. Pour H_{S}, l'indice d'égalité est donc $H_{\mathrm{S}}/ln\left(S\right)$ (Pielou 1969). La valeur $\texttt{exp}\left(H_{\mathrm{S}}\right)/S$ a été proposée pour éviter le quotient logarithmique (Lloyd et Ghelardhi 1964, Buzas et Gibson 1969, Whittaker 1972).

Passons aux propriétés de ces indices lorsque leurs valeurs dans une collection doivent être estimées à partir d'un échantillon. Nayak (1983) part sur l'hypothèse que les variables (N_1, ..., N_k, ..., N_S) des nombres d'entités dans chaque catégorie suivent une loi multinomiale de paramètres n et $\pi = (\pi_1, ..., \pi_k, ..., \pi_S)$, avec $\pi_k > 0$, pour tout k, $1 \leq k \leq S$. Si $\pi_k = 0$, pour un k donné, alors les résultats sont valables sur l'ensemble moins cette catégorie. Soient p_1, p_2, ..., p_S les valeurs prises par les variables N_1, N_2,..., N_S dans un échantillon de taille n. $P_k = N_k/n$ est un estimateur non biaisé de π_k. Notons $P = (N_1/n, ..., N_S/n)^t$.

$H_{\text{G-S}}\left(P\right)$ est un estimateur biaisé de $H_{\text{G-S}}\left(\pi\right)$:

$$E\left(H_{\text{G-S}}\left(P\right)\right) = \frac{n-1}{n}\left(1 - \sum_{k=1}^{S} \pi_k^2\right).$$

Le biais dû au terme $(n-1)/n$ diminue avec la valeur de n. L'estimateur non biaisé de $H_{\text{G-S}}\left(\pi\right)$ et dont la variance est minimale est

$$\frac{n}{n-1} H_{\text{G-S}}\left(P\right).$$

L'espérance de $H_S(P)$ est plus complexe. Elle peut néanmoins être approchée par la valeur :

$$E(H_S(P)) \approx -\sum_{k=1}^{S} \pi_k \ln(\pi_k) + \frac{S-1}{n}$$

(Peet 1974, Chakraborty et Rao 1991, page 278). Le biais de H_S dépend de la richesse réelle qui est inconnue. Comme pour l'estimateur de la richesse, il est impossible d'obtenir un estimateur non biaisé de l'indice de Shannon.

Nayak (1983, 1985) démontre les variances asymptotiques suivantes pour les indices $H_{G\text{-}S}$, H_S et $H_{H\text{-}C}$

$$V(H_{G\text{-}S}(P)) \approx \frac{4}{n}\left[\sum_{k=1}^{S} \pi_k^3 - \left(\sum_{k=1}^{S} \pi_k^2\right)^2\right]$$

$$V(H_S(P)) \approx \frac{1}{n}\left[\sum_{k=1}^{S} \pi_k (\ln \pi_k)^2 - \left(\sum_{k=1}^{S} \pi_k \log \pi_k\right)^2\right]$$

$$V(H_{H\text{-}C}(P)) \approx \frac{1}{n}\left(\frac{\alpha}{\alpha-1}\right)^2\left[\sum_{k=1}^{S} \pi_k^{2\alpha-1} - \left(\sum_{k=1}^{S} \pi_k^{\alpha}\right)^2\right]$$

Lorsque n est faible, les formules exactes des variances peuvent être trouvées dans Chakraborty et Rao (1991) par exemple. En génétique, les catégories sont les allèles d'un locus. Actuellement nous ne sommes capables d'examiner qu'une petite fraction du génome. Ainsi souvent plusieurs loci sont étudiés et la valeur de la diversité est moyennée sur l'ensemble des loci. Il y a alors deux échantillonnages (Zhang et Allard 1986) : (1) la sélection des individus dans la population ; (2) la sélection des loci dans le génome. Dans ce cas, la variance de l'indice comprend une variance inter-loci et une variance intra-locus. Les calculs se compliquent si les loci ne sont pas indépendants. Il faut également considérer le cas des polyploïdes et en particulier des diploïdes. Le calcul de la variance de l'indice de Gini-Simpson pour des données multi-loci pourra être trouvé dans Zhang et Allard (1986).

Nous avons vu que l'indice de Shannon est traditionnellement le plus utilisé en écologie. Pourtant l'indice de Gini-Simpson présente plus d'avantages. Il possède une définition simple : probabilité pour que deux entités tirées au hasard dans une collection soient différentes. Les biais dans l'estimation de la richesse et de l'indice de Shannon dépendent de paramètres inconnus : le nombre de catégories et les fréquences des catégories dans la collection pour laquelle on ne connaît qu'un échantillon. C'est seulement pour l'indice de Gini-Simpson qu'il existe un estimateur non biaisé. De plus l'indice de Gini-Simpson est celui qui a la plus petite variance (Lande 1996).

Des questions restent soulevées quant à l'estimation des indices définis sur l'ensemble $\mathcal{P}$ des distributions de fréquences. Les espérances et variances de ces indices sont estimées en supposant une distribution multinomiale des nombres d'entités dans chaque catégorie, ce qui n'est pas toujours le cas. En particulier, toutes les formules d'espérance et de variance de ces indices devraient également prendre en compte la variation spatiale et les différences de difficulté à détecter les espèces (Yoccoz *et al.* 2001). Yoccoz *et al.* (2001) citent plusieurs références qui

pourront être consultées pour prendre en considération ces probabilités de détection. Elles n'ont pas été considérées dans cette thèse faute de temps, et due à la complexification des problèmes statistiques qu'elles peuvent engendrer dans leur intégration aux développements méthodologiques nouveaux présentés dans les chapitres suivants. Elles pourront néanmoins faire l'objet d'études complémentaires à cette thèse.

2.2.3 Tests

Schluter et Ricklefs (1993) proposent d'utiliser l'analyse de variance pour tester l'égalité des richesses de plusieurs assemblages. Ils développent ce test dans le but de détecter ou d'invalider la présence de convergence entre plusieurs régions soumises aux mêmes conditions climatiques. Dans chaque région plusieurs habitats sont étudiés, ces habitats sont choisis de sorte qu'ils se ressemblent le plus possible entre les régions. Le nombre d'espèces présentes dans chaque habitat de chaque région est évalué. Il y aura convergence si le nombre d'espèces varie avec l'habitat de la même façon dans toutes les régions. Région et habitat constituent deux facteurs croisés. La variance totale de la richesse comprend quatre composants :

$$\sigma^2_{\text{total}} = \sigma^2_{\text{H}} + \sigma^2_{\text{R}} + \sigma^2_{\text{H}\times\text{R}} + \sigma^2_{\text{e}},$$

σ^2_{H} est la part de la variance totale due aux effets des habitats, σ^2_{R} la part due aux effets des régions, $\sigma^2_{\text{H}\times\text{R}}$ la part due à des interactions entre effets de l'habitat et effets de la région, et σ^2_{e} est un composant de variance résiduelle. Le composant σ^2_{H} inclut toutes les variations de richesse entre habitats qui ont le même profil dans toutes les régions. Ce composant est donc considéré comme une mesure de la convergence entre régions. L'estimateur de σ^2_{H} est

$$V_{\text{C}} = \frac{(MS_{\text{H}} - MS_{\text{e}})\,(h-1)}{nrh},$$

où C désigne la convergence, h est le nombre d'habitats, r le nombre de régions, n le nombre de sites dans chaque combinaison habitat-région, et MS désigne le carré moyen ("Mean Square") de l'analyse de la variance (ANOVA) (Fisher 1925). V_{C} est le composant typique de l'ANOVA multiplié par $(h-1)/h$, modification effectuée pour tenir compte du fait que l'habitat est un facteur fixé plutôt qu'aléatoire. V_{C} est un estimateur non-biaisé de σ^2_{H}. Toutes les mesures de richesse sont ln-transformées avant l'analyse (avant la transformation logarithmique, toutes les valeurs sont augmentées de une unité si certaines richesses sont nulles). L'ANOVA requiert cette transformation parce que la variation entre sites augmente généralement avec le nombre moyen d'espèces présentes. Remarquons que les problèmes d'estimation du nombre d'espèces vus dans la partie précédente ont pour conséquence que les richesses enregistrées à chaque site sont globalement sous-estimées. Le composant régional de la variance est

$$\begin{aligned} V_{\text{Rtotal}} &= V_{\text{R}} + V_{\text{R}\times\text{H}} \\ &= \frac{(MS_{\text{R}} - MS_{\text{e}})\,(r-1)}{nrh} \\ &+ \frac{(MS_{\text{R}\times\text{H}} - MS_{\text{e}})\,(h-1)\,(r-1)}{nrh} \end{aligned}$$

Schluter et Ricklefs proposent de calculer la fraction de convergence :

$$I_{\text{C}} = \frac{V_{\text{C}}}{V_{\text{total}}},$$

où $V_{\text{total}} = V_{\text{C}} + V_{\text{Rtotal}} + MS_{\text{e}}$.

Schluter et Ricklefs (1993) ont choisi les données de Blondel *et al.* (1984) pour illustrer leur analyse. Au lieu de considérer des habitats, Schluter et Ricklefs s'intéressent ici aux catégories alimentaires d'espèces d'oiseaux dans trois successions : une en Californie, une au Chili et une en Provence (tableau 3 dans Blondel *et al.* (1984)). Leur étude examine les différences entre catégories alimentaires et les différences entre régions en terme de richesse spécifique. Ils trouvent des différences fortes entre les richesses associées aux différentes catégories alimentaires ($I_{\text{C}} = 0.74$, $P < 0.001$), montrant ainsi une convergence de la répartition des richesses entre catégories alimentaires pour les trois régions Californie, Chili et Provence. Blondel *et al.* avaient également analysé la région de Bourgogne en tant que région témoin. Nous l'avons rajoutée l'analyse, et nous obtenons $I_{\text{C}} = 0.79$, $P < 0.001$. Le test proposé par Schluter and Ricklefs suggère l'existence d'une convergence entre les trois régions méditerranéennes et la Bourgogne et très peu de différences entre les régions. Il s'en suit que si une convergence existe, elle n'est pas due au climat. Le test proposé par Schluter et Ricklefs est donc une façon de tester les différences de richesse entre collections.

Prenons maintenant, pour mesurer la diversité, les indices définis sur l'ensemble $\mathcal{P}$ des distributions de fréquences. En choisissant pour variables les nombres d'entités dans chaque catégories $(N_1, ..., N_k, ..., N_S)$, avec pour hypothèse que ces variables suivent une loi multinomiale de paramètres n et $(\pi_1, ..., \pi_k, ..., \pi_S)$, où $\pi_k > 0$, pour tout k, $1 \leq k \leq S$, les estimateurs des fonctions H_{S}, $H_{\text{G-S}}$ et plus généralement de la fonction $H_{\text{H-C}}$ vérifient (Nayak 1985, théorème 2.3.1),

$$\sqrt{n}\left(\hat{H} - H\right) \rightarrow N\left(0, \tau^2\right),$$

où n est le nombre total d'entités observées et τ^2 est égal à n fois la variance asymptotique de la fonction H considérée.

Prenons une fonction de diversité D vérifiant

1. D est maximum si $\pi_1 = ... = \pi_S = 1/S$
2. $\sqrt{n}\left(\hat{D} - D\right) \rightarrow N\left(0, \tau^2\right)$

En particulier, les trois fonctions H_{S}, $H_{\text{G-S}}$ et $H_{\text{H-C}}$ vérifient ces deux propriétés. Les tests suivants ont été développés par Nayak (1983, 1985) pour toute fonction D.

Test 1 : La collection est-elle homogène ? Pour répondre à cette question, il faut tester l'hypothèse $\text{H}_{01} : D = D_{\max}$, contre l'hypothèse $\text{H}_{11} : D \neq D_{\max}$, où $D_{\max}$ est la valeur maximale de D obtenue pour $\pi_1 = ... = \pi_S = 1/S$. Le test diffère selon la fonction de diversité. Pour l'indice H_{S} de Shannon, $\max_{\mathbf{p}}\left[H_{\text{S}}(\mathbf{p})\right] = \ln(S)$. Sous H_{01},

$$2n\left(\ln S - \hat{H}_{\text{S}}\right) \sim \chi^2_{(S-1)},$$

H_{01} est rejetée avec un risque d'erreur α si

$$2n\left(\ln S - \hat{H}_{\text{S}}\right) > \chi^2_{(S-1);\alpha}.$$

Pour les indices $H_{\text{G-S}}$ de Gini-Simpson et $H_{\text{H-C}}$ de Havrda et Charvat, H_{01} est équivalente à $\sum_{k=1}^{S} \pi_k^{\alpha} = S^{1-\alpha}$, avec $\alpha = 2$ pour $H_{\text{G-S}}$, et $\alpha > 0$ et $\neq 1$ pour $H_{\text{H-C}}$. Soit $T = \sum_{k=1}^{S} p_k^{\alpha}$, sous H_{01}

$$\frac{2nS^{\alpha-1}}{\alpha(\alpha-1)}\left(T - S^{1-\alpha}\right) \sim \chi^2_{(S-1)}.$$

H_{01} est rejetée au risque d'erreur δ si

$$\frac{2nS^{\alpha-1}}{\alpha(\alpha-1)}\left(T - S^{1-\alpha}\right) > \chi^2_{(S-1);\delta}.$$

Test 2 : La diversité de la collection est-elle égale à la valeur D_0 ($D_0 \neq D_{\max}$) ? $\text{H}_{02} : D = D_0$. D'après la condition (2) vérifiée par D, et si $\hat{\tau}$ est l'estimation de τ obtenue en remplaçant les π_k par leurs observations p_k,

$$\frac{\sqrt{n}(\hat{D} - D)}{\hat{\tau}} \rightarrow N(0,1)\text{, pourvu que } \tau^2 \neq 0.$$

La variance τ^2 de D s'annule pour la distribution uniforme qui correspond à $D_{\max}$. Donc si D_0 est différent de $D_{\max}$ alors $\tau^2 \neq 0$. H_{02} peut être testée contre trois hypothèses : $\text{H}_{12} : D > D_0$, $\text{H'}_{12} : D < D_0$ et $\text{H"}_{12} : D \neq D_0$. Ainsi, H_{02} est rejetée, avec un risque α, contre

$$\begin{aligned}
\text{H}_{12} : D > D_0 \quad &\text{si} \quad \frac{\sqrt{n}(\hat{D} - D_0)}{\hat{\tau}} > \varepsilon_{\alpha}, \\
\text{H'}_{12} : D < D_0 \quad &\text{si} \quad \frac{\sqrt{n}(\hat{D} - D_0)}{\hat{\tau}} < -\varepsilon_{\alpha}, \\
\text{H"}_{12} : D \neq D_0 \quad &\text{si} \quad \frac{\sqrt{n}|\hat{D} - D_0|}{\hat{\tau}} > \varepsilon_{\alpha/2},
\end{aligned}$$

où ε_{α} est le seuil de la loi normale au niveau α.

Test 3 : Les diversités de plusieurs collections sont-elles égales ? $\text{H}_{03} : D_1 = D_2 = ... = D_r$. Notons n_i le nombre d'entités dans la collection i et

$$\bar{D} = \left[\sum_{i=1}^{r} \frac{n_i \hat{D}_i}{\tau_i^2}\right] / \left[\sum_{i=1}^{r} \frac{n_i}{\tau_i^2}\right].$$

Sous H_{03},

$$\sum_{i=1}^{r} \frac{n_i(\hat{D}_i - \bar{D})^2}{\hat{\tau}_i^2} \sim \chi^2_{(r-1)}.$$

H_{03} est rejetée avec un risque d'erreur α si

$$\sum_{i=1}^{r} \frac{n_i(\hat{D}_i - \bar{D})^2}{\hat{\tau}_i^2} > \chi^2_{(r-1);\alpha}.$$

Un intervalle de confiance de D peut être donné au niveau $100\,(1-\alpha)\,\%$ et pour des grands échantillons :

$$\hat{D} - \frac{\hat{\tau}}{\sqrt{n}}\varepsilon_{\alpha/2} < D < \hat{D} + \frac{\hat{\tau}}{\sqrt{n}}\varepsilon_{\alpha/2}.$$

Les données écologiques ne vont pas toujours satisfaire l'hypothèse selon laquelle les variables $N_1, ..., N_S$ suivent une loi multinomiale de paramètres n et $\pi = (\pi_1, ..., \pi_S)^t$. Des alternatives, utilisant des méthodes de jackknife, et bootstrap, ont alors été proposées (Efron 1982, Magurran 1988, 2004). Le jackknife d'un indice de diversité consiste à diviser les n individus de l'échantillon en m groupes. Sont alors calculés la diversité g^0 de l'échantillon, la diversité de l'échantillon privée du groupe i notée $g_i^{(-i)}$ et la pseudovaleur g_i, effet du groupe i sur la diversité de l'échantillon, calculée par

$$g_i = mg^0 - (m-1)\,g_i^{(-i)},\ i = 1, ..., m.$$

L'estimation jackknife est alors

$$\hat{H} = \frac{1}{m}\sum_{i=1}^{m} g_i.$$

La variance des pseudovaleurs peut être calculée et servir dans des procédure de tests (Zahl 1977). L'estimation par bootstrap de la diversité est obtenue en rééchantillonnant les individus de l'échantillon, avec remise. L'estimation est la moyenne des diversités obtenues sur l'ensemble des rééchantillonnages. La variance de ces diversités peut également être calculée. Grâce au jackknife et bootstrap, il est possible de tester la différence entre les diversités de deux collections (Caron 2000).

2.3 D ´ ´

Deux critiques peuvent alors être faites à tous ces indices traditionnels, définis sur $\mathcal{P}$.

1. La première est que plus une collection possède de catégories, moins les indices de Shannon et de Gini-Simpson sont sensibles aux différences de fréquences entre catégories dans cette collection. La mesure de diversité qu'ils fournissent est alors très proche de la richesse.
2. La deuxième critique concerne une des propriétés fondamentales de ces indices : ils sont invariables par permutation des catégories. D'après les définitions de ces indices, les catégories sont interchangeables : si dans une collection, une catégorie est remplacée par une autre alors qu'elle a des caractéristiques très différentes des autres, les indices précédents ne détecteront pas ce changement. La deuxième critique est donc que ces indices attribueraient la même diversité à une région dans laquelle seraient présents une autruche, un mulot et un lion, qu'à une région dans laquelle se trouveraient un campagnol, un mulot et un rat. Alors que d'un certain point de vue, nous avons l'intuition que la première région est plus diverse.

Pour avoir une mesure exhaustive de la diversité, il nous faudrait connaître les génomes de tous les êtres vivants, leurs phénotypes, leurs comportements donc leur rôle fonctionnel vis à vis des autres et du fonctionnement de la planète. Nous savons qu'une telle connaissance est actuellement impossible. Cette complexité de la diversité nous indique que chaque catégorie que nous considérons en écologie (espèce, allèle, etc) possède un rôle particulier et en règle général, n'est pas interchangeable. Pour tenir compte de ce fait, une suggestion est qu'une mesure générale de biodiversité doit non seulement inclure une mesure du nombre d'espèces, mais aussi une mesure du degré de différences entre ces espèces (Williams et Humphries 1994). L'interchangeabilité des espèces est une propriété traditionnellement prônée pour les mesures de biodiversité. Mais nous verrons pourtant que l'avantage des indices prenant en compte les différences entre les espèces est justement de ne plus vérifier cette propriété.

> "A measure of the biodiversity of a site ought to say something about how different the inhabitants are from each other." (Harper et Hawksworth 1995)

2.3.1 Matrice de dissimilarités : définition

• Dissimilarités, semi-distances et distances :

Dans la démarche de mesurer des différences, les termes suivants sont souvent employés en mathématique, plus précisément en géométrie euclidienne : dissimilarité, distance, métrique. Le terme 'différence', dans le language courant, représente ce qui distingue ou oppose deux choses. Les autres termes ont les significations mathématiques précises suivantes. L'emploi de l'un ou l'autre n'est donc pas anodin.

Soit $\mathbf{D} = \left[d_{ij}\right]$ une matrice symétrique de diagonale nulle :

$$d_{ij} = 0 \ \forall i, j \text{ (C1 : diagonale nulle) et } d_{ij} = d_{ji} \text{ (C2 : symétrie).}$$

Si **D** vérifie C1 et C2 et si de plus

$$d_{ij} \geq 0 \ \forall i, j \neq i \text{ (C3 : positivité).}$$

alors les d_{ij} sont des dissimilarités et **D** une matrice de dissimilarités.
Si **D** vérifie C1, C2 et si de plus

$$d_{ij} + d_{jk} \geq d_{ik} \ \forall i, j, k \text{ (C4 : inégalité triangulaire),}$$

alors les d_{ij} sont des semi-distances et **D** une matrice de semi-distances.
Si **D** vérifie C1, C2, C4 et si de plus

$$d_{ij} = 0 \Rightarrow i = j \ \forall i, j \text{ (C5),}$$

alors les d_{ij} sont des distances et **D** une matrice de distances. Notons si C4 est vérifiée, C3 l'est par implication.

• Matrices euclidiennes, circum-euclidiennes et ultramétriques :

En plus des propriétés C1 à C5, d'autres propriétés peuvent être ajoutées. Nous parlerons notamment de matrices de distances euclidienne, circum-euclidienne et ultramétrique. Contrairement aux termes dissimilarités, distances et ultramétriques, les termes "euclidien" et "circum-euclidien" s'appliquent aux matrices et non aux d_{ij} aux-mêmes. Le terme euclidien est aussi utilisé pour caractériser un espace vectoriel contenant tous les n-tuples de nombres réels $(x_1, x_2, ..., x_n)$. Un espace euclidien est souvent noté $\mathbb{R}^n$ et est aussi appelé espace cartésien. Le terme "distance euclidienne" existe mais il désigne la valeur prise par une formule particulière appelée métrique euclidienne ou métrique canonique. La distance euclidienne entre deux points dans un espace euclidien est calculée par la métrique euclidienne appliquée aux coordonnées des deux points et vaut la longueur du segment de droite qui relie ces deux points.

***Théorème* 2.3.1.1** *Théorème de Gower : Une matrice de dissimilarités* $\mathbf{D} = [d_{kl}]$ *est euclidienne si et seulement si S points* M_k *(*$k = 1, ..., S$*) peuvent être inclus dans un espace euclidien tels que la distance euclidienne (calculée avec la métrique euclidienne) entre* M_k *et* M_l *est* d_{kl}*. (Gower et Legendre 1986)*

Dans le cas où une matrice est euclidienne, nous dirons que ses éléments d_{kl} sont des distances qui ont des propriétés euclidiennes pour éviter la confusion avec le terme "distance euclidienne". Les matrices euclidiennes sont des matrices de distances. La matrice **D** est circum-euclidienne, si elle est euclidienne et si les points M_k sont situés sur le bord d'une même hypersphère. Elle est de plus ultramétrique si

$$d_{kl} \leq \max(d_{ki}, d_{il}) \text{ pour tous } k, l \text{ et } i.$$

Toutes ces définitions sont emboîtées dans l'ordre où elles ont été présentées. Par exemple une matrice euclidienne est une matrice de distances, qui est elle-même une matrice de dissimilarités.

- Genèse des matrices de dissimilarités :

Une matrice de dissimilarités peut être observée, lorsque par exemple la dissimilarité entre deux objets est évaluée directement sans passer par une variable. Elle peut aussi être calculée à partir d'une fonction. La fonction qui permet d'obtenir une distance est appelée métrique. Soit g une telle fonction. g est définie sur un ensemble $E \times E$ (par exemple $\mathbb{R}^n \times \mathbb{R}^n$). Soient x, y, et z trois éléments de E, g vérifie les propriétés C1, C2, C4 et C5 :

$$g(x, x) = 0$$

$$g(x, y) = g(y, x)$$

$$g(x, y) + g(y, z) \geq g(x, z)$$

$$g(x, y) = 0 \Rightarrow x = y$$

(g vérifie C3 par implication : C4 $\Rightarrow$ C3). Si g ne vérifie pas C5 alors la fonction est appelée pseudo-métrique et construit des semi-distances.

Pour mesurer les différences entre catégories, nous utiliserons toujours des dissimilarités. Ce terme "dissimilarité" sera donc employé par la suite pour désigner de façon générale toute matrice possédant au moins les propriétés C1, C2 et C3 définissant les dissimilarités, quelles ques soient les propriétés supplémentaires possédées par la métrique utilisée. Ces propriétés pourront être ensuite étudiées et signalées, mais d'une façon générale, le terme "dissimilarité" sera gardé.

Plusieurs fonctions peuvent être utilisées pour mesurer des dissimilarités entre objets. Elles sont désignées par un terme général : "indice de dissimilarité" ; ou aussi souvent par "coefficient de dissimilarité". Lorsque le nom d'un de ces indices de dissimilarité commence par "métrique", c'est qu'il vérifie effectivement les critères définissant une métrique. La réciproque n'est pas vraie. Dans la pratique, tous les indices qui sont des métriques ne portent pas forcément des noms commençant par "métrique". Il arrive parfois que ce nom commence par "distance" et il apparaît que c'est souvent indépendant des propriétés mathématiques de la fonction. "distance" a dans ce cas un sens commun de différence. Nous garderons donc le nom le plus utilisé pour chaque fonction mais préciserons s'il s'agit effectivement d'une métrique conduisant à des mesures de distances ou simplement un indice de dissimilarité ne vérifiant pas les propriétés C4 et C5 par lesquelles nous avons défini ici la distance.

Nous mesurerons des dissimilarités entre deux types d'objets : les catégories (section suivante) et les collections (chapitre 4). Beaucoup de fonctions ne dépendent pas de la nature biologiques des données, mais de type mathématique. Par exemple une fonction peut être définie sur l'ensemble des vecteurs de fréquences, quelle que soit la nature de ces fréquences. Ainsi beaucoup de fonctions permettant de calculer des dissimilarités entre catégories peuvent à un autre niveau biologique permettre de calculer des dissimilarités entre collections. A l'inverse, nous verrons que certaines façons de calculer des dissimilarités sont étroitement liées à la nature des données biologiques.

2.3.2 Dissimilarités entre catégories

Les catégories peuvent être caractérisées par une ou plusieurs variables quantitatives $\{X_1, ..., X_i, ..., X_I\}$; par exemple des variables morphométriques caractérisant des espèces. Soient $\mathbf{x}_k = (x_{1k}, ..., x_{ik}, ..., x_{Ik})^t$ et $\mathbf{x}_l = (x_{1l}, ..., x_{il}, ..., x_{Il})^t$ deux vecteurs contenant les valeurs prises par les catégories k et l respectivement, pour chaque variable considérée. Les fonctions suivantes peuvent être utilisées pour mesurer la dissimilarité entre les catégories k et l :

- la métrique ou distance euclidienne, citée précédemment, encore appelée métrique canonique

$$d(\mathbf{x}_k - \mathbf{x_l}) = \sqrt{(\mathbf{x}_k - \mathbf{x}_l)^t (\mathbf{x}_k - \mathbf{x}_l)} = \|\mathbf{x}_k - \mathbf{x}_l\|,$$

- la métrique de Joreskog

$$d(\mathbf{x}_k - \mathbf{x_l}) = \sqrt{(\mathbf{x}_k - \mathbf{x}_l)^t \mathbf{V}^{-1} (\mathbf{x}_k - \mathbf{x}_l)} = \|\mathbf{x}_k - \mathbf{x}_l\|_{\mathbf{V}^{-1}},$$

où $\mathbf{V} = diag(V(\mathbf{Y}_1), ..., V(\mathbf{Y}_i), ..., V(\mathbf{Y}_I))$ est la matrice diagonale contenant les variances des variables considérées,

- la métrique de Mahalanobis

$$d(\mathbf{x}_k - \mathbf{x_l}) = \sqrt{(\mathbf{x}_k - \mathbf{x}_l)^t \mathbf{W}^{-1} (\mathbf{x}_k - \mathbf{x}_l)} = \|\mathbf{x}_k - \mathbf{x}_l\|_{\mathbf{W}^{-1}},$$

où **W** est la matrice des variances-covariances des variables X_1, ..., X_i, ..., X_I. Ces trois fonctions sont définies sur $\mathbb{R}^I \times \mathbb{R}^I$. Elles métriques et mesurent des distances conduisant à des matrices euclidiennes.

Les catégories peuvent être caractérisées par une ou plusieurs variables dites floues. Une variable floue est une variable qualitative par exemple les habitudes alimentaires d'espèces d'oiseaux (modalités : granivores, insectivores, etc), ou quantitatives mais découpées arbitrairement en classes, par exemple la taille des ailes des oiseaux (de 50 à 70 mm, de 70 à 90 mm, etc). Considérons une variable à J modalités. Les données sont exprimées sous forme de pourcentages qui représentent les affinités des catégories pour chaque modalité de la variable. Par exemple si l'alimentation des individus d'une espèce d'oiseau comprend en moyenne 90% de graines et 10% d'insectes alors l'affinité de cette espèce pour la modalité "granivores" sera de 0.9, son affinité pour la modalité "insectivores" de 0.1 et ses affinités pour les autres modalités seront de 0. Soit p_{jk} (resp. p_{jl}) la fréquence ($\Leftrightarrow$ affinité) associée à la modalité j (parmi m) pour la catégorie k (resp. l), et $\mathbf{p}_k$ et $\mathbf{p}_l$ les vecteurs de fréquences pour les catégories k et l respectivement sur l'ensemble des modalités. Toutes les fonctions définies sur $\mathcal{P} \times \mathcal{P}$, où $\mathcal{P}$ est l'ensemble des vecteurs de fréquences, peuvent être utilisées pour mesurer la dissimilarité entre catégories avec ce type de données. Nous en verrons 17 dans le chapitre 4. Citons, par exemple, la distance d'Edwards (1971)

$$d(\mathbf{p}_k - \mathbf{p}_l) = \sqrt{1 - \sum_{j=1}^{m} \sqrt{p_{jk} p_{jl}}}.$$

Lorsque plusieurs variables de ce type sont utilisées, la dissimilarité entre deux catégories peut être calculée comme la moyenne des valeurs obtenues pour chaque variable. La distance d'Edwards est une métrique et les matrices de dissimilarités obtenues par cette fonction sont euclidiennes.

Les fonctions que nous venons de voir ont été développées en statistiques, où dérivent de fonctions développées en statistiques. Elles ont donc un champ d'utilisation très large. Les fonctions qui suivent ont, elles, été développées pour répondre à des problèmes soulevés en biologie.

Les catégories peuvent être caractérisées par une ou plusieurs variables binaires. Par exemple, en génétique les profils de bandes observés sur électrophorèse conduisent souvent à des données binaires. Une catégorie est un haplotype. A un haplotype correspond un profil de bandes particulier. Ces profils sont obtenus par des méthodes découpant l'ADN en fragments séparés par électrophorèse. Chaque haplotype est alors caractérisé par un vecteur de 0 et 1 : 0 = absence d'une bande, 1 = présence de cette bande. Soit B l'ensemble des vecteurs binaires. Toutes les fonctions de dissimilarité définies sur $B \times B$ peuvent être utilisées pour calculer une dissimilarité entre deux haplotypes. Ces fonctions ont été pour la plupart développées en écologie, pour calculer la similarité et la dissimilarité entre deux listes, non pas de bandes, mais d'espèces (cf. chapitre 4). La plus courante de ces fonctions est celle qui est basée sur l'indice de Jaccard (1901)

$$d = 1 - \frac{a}{a + b + c},$$

où d désigne ici une mesure de la dissimilarité entre deux haplotype, a désigne le nombre de

bandes communes aux deux haplotypes comparés k et l, b le nombre de bandes présentes dans le premier haplotype mais pas dans le second, et inversement c désigne le nombre de bandes présentes dans le second haplotype mais pas dans le premier. D'autres mesures sont présentées dans la partie 4.1.1. Certains indices peuvent être développés pour un type de données et un type d'organisme. Par exemple, en travaillant sur les profils AFLP chez des bactéries, Mougel *et al.* (2002) ont remarqué la chose suivante. Si on augmente artificiellement les différences entre deux individus, la distance de Jaccard augmente plus vite entre deux individus proches qu'entre deux individus éloignés. Or la vitesse d'évolution est supposée constante. Ils ont alors introduit une autre mesure de dissimilarité, basée sur celle de Jaccard et qui tient compte de cette observation :

$$d = 1 - \left(\frac{a}{a+b+c}\right)^{1/r},$$

où r est le nombre de sites nucléotidiques identiques nécessaires pour le partage d'un fragment AFLP par deux isolats (clones).

En génétique, les catégories peuvent être des séquences nucléotidiques. Soit L la taille des séquences c'est-à-dire le nombre de loci considérés. Les deux dissimilarités les plus simples entre ces deux séquences sont le nombre x_{ij} et la proportion x_{ij}/L de nucléotides qui diffèrent entre les séquences k et l, sur l'ensemble des loci. D'autres mesures de dissimilarités plus complexes ont été développées.

La fonction de Jukes et Cantor (1969) permet de prendre en compte les substitutions multiples qui ont pu avoir lieu entre les deux séquences depuis leur plus récent ancêtre commun :

$$d = \frac{3}{4}\ln\left(1 - \frac{4}{3}\frac{x_{kl}}{L}\right).$$

La fonction de Kimura à 2 paramètres (Kimura 1980) permet de prendre en compte les nombres de transition n_{TS} (A $\rightleftharpoons$ G ou C $\rightleftharpoons$ T) et de transversion n_{TV} (A $\rightleftharpoons$ T, A $\rightleftharpoons$ C, G $\rightleftharpoons$ T ou G $\rightleftharpoons$ C) :

$$d = -\frac{1}{2}\ln\left(1 - 2\frac{n_{\text{TS}}}{L} - \frac{n_{\text{TV}}}{L}\right) - \frac{1}{4}\ln\left(1 - 2\frac{n_{\text{TV}}}{L}\right)$$

Une modification de cet indice a été proposée (Jin et Nei 1990). Elle suppose que les taux de mutations sont distribués selon une loi Gamma, de paramètre a, inverse du coefficient de variation du taux de mutations :

$$d = \frac{a}{2}\left[\left(1 - 2\frac{n_{\text{TS}}}{L} - \frac{n_{\text{TV}}}{L}\right)^{-1/a} + \frac{1}{2}\left(1 - 2\frac{n_{\text{TV}}}{L}\right)^{-1/a} - \frac{3}{2}\right]$$

D'autres indices ont été proposés pour tenir compte du taux de G+C (Tamura 1992), des fréquences de chaque nucléotide, A, G, C et T (Tajima et Nei 1984), et des taux de transitions entre purines (A $\rightleftharpoons$ G) et entre pyrimidines (C $\rightleftharpoons$ T) (Tamura et Nei 1993).

En systématique, on étudie la construction d'arbre en vue de la classification de la diversité biologique. Dans cette discipline, les catégories sont des taxa et sont les feuilles ou les nœuds d'un schéma appelé arbre ou dendrogramme. Les feuilles sont aussi appelées "nœuds terminaux", les autres nœuds sont alors des "nœuds internes". Une branche est un segment qui

relie deux nœuds sans en traverser d'autres. Il existe plusieurs types d'arbres. Le terme dendrogramme désigne n'importe quel diagramme ayant la structure d'un arbre. Différentes méthodes ont été développées pour classer les êtres vivants, en particulier, la cladistique, la phylogénie, la phénétique et la systématique linnéenne. En cladistique, les arbres sont des cladogrammes décrivant les relations historiques entre les taxa et dont les longueurs de branches sont arbitraires. En phylogénie, deux types d'arbres existent : les phylogénies donnant les relations historiques entre des lignées d'organismes ou leurs parties (*e.g.* des gènes) ; les phylogrammes qui décrivent les relations historiques entre organismes et dont les longueurs de branches sont proportionnelles aux nombres de changements estimés de caractères. La phénétique s'intéresse au ressemblance entre taxa en prenant en compte, de manière équivalente, le maximum de caractères. En phénétique, les phénogrammes décrivent les ressemblances entre taxa. Dans un phénogramme deux individus sont proches s'ils se ressemblent, s'ils ont beaucoup de caractères communs, y compris ceux obtenus par évolution convergente en vivant par exemple dans un même habitat. La phénétique s'oppose à la phylogénie et à la cladistique qui cherchent à regrouper les êtres vivants uniquement en fonction de leurs liens de parentés. Phylogénie et cladistique se basent donc sur des caractères homologues c'est-à-dire obtenus par ascendance commune. En systématique linnéenne, la taxonomie des êtres vivants peut aussi être représentée sous la forme d'un arbre. Les feuilles sont les espèces. En partant de ces feuilles, les premiers nœuds rencontrés sont les genres, les seconds les familles, etc. A une branche qui relie deux nœuds successifs (par exemple la longueur de la branche séparant le nœud terminal *Felis catus* du nœud interne *Felis*) on attribue généralement une longueur d'une unité, qui représente un changement de niveau taxonomique.

La distance entre deux catégories sur un arbre peut être soit le nombre de nœuds soit la somme des longueurs de branches qui les séparent. En général, on calcule la somme des longueurs de branches, lorsqu'elles sont connues, plutôt que le nombre de nœuds. Les dissimilarités obtenues à partir d'un phylogramme correspondent alors à la "distance patristique" mesurant la différence entre deux taxa par le nombre de changements de caractères qui les séparent. Celles calculées à partir d'un phénogramme sont des "distances phénétiques". Enfin celles obtenues à partir de taxonomies mesurent le nombre de niveaux hiérarchiques qui séparent deux taxa et sont appelées simplement "distances taxonomiques". La taxonomie la plus courante comprend les niveaux espèce, genre, famille, ordre, classe, embranchement et règne. Mais d'autres niveaux ont été définis pour certains organismes : sous-famille, sous-ordre, sous-classe, super-classe, etc. Supposons que les longueurs de branches entre deux niveaux juxtaposés (par exemple espèce/genre ou genre/famille) soient d'une unité comme affirmé précédemment. Un nouveau problème se pose : si l'on compte la somme des longueurs de branches pour définir la distance entre deux taxa, et si la taxonomie comprend espèce, genre, famille, sous-ordre et ordre, on comptera de la même façon l'écart entre genres et familles qu'entre sous-ordres et ordres. Le problème est double et non résolu. (1) Doit-on pondérer différemment les écarts entre deux niveaux ? (2) Que faire lorsque pour certaines espèces de nombreux niveaux taxonomiques ont été définis et pour d'autres non, simplement parce que plus de spécialistes ont travaillé sur les premières que sur les secondes, ou bien parce que différents spécialistes peuvent avoir des politiques de définition de groupes différentes ?

Le terme "distance" est couramment utilisé pour désigner toutes ces dissimilarités obtenues à partir d'arbres. Cependant leur propriété métrique n'est pas garantie. Signalons que s'il est possible d'obtenir des distances entre catégories, de nombreux arbres sont, eux, obtenus à partir

de matrices de dissimilarités, notamment en génétique. Si la dissimilarité entre deux entités est définie à partir de données issues d'un échantillonnage, alors la variance de l'estimateur de cette dissimilarité peut être étudiée (Nei 1987, page 64 par exemple).

Les données de départ pour mesurer la diversité seront alors, d'une façon générale, une matrice décrivant les abondances, fréquences ou présences/absences d'un ensemble de catégories dans une ou plusieurs collections et une matrice de dissimilarités entre toutes les catégories (Fig. 4).

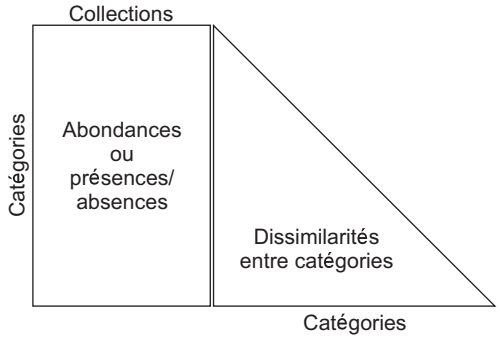


FIG. *4 – Données de départ pour mesurer la biodiversité.*

2.3.3 Mesurer la diversité en attribuant un poids à chaque catégorie

Certains indices prennent en compte la différence entre catégories bien qu'ils ne soient pas basés pour autant sur une matrice de dissimilarités entre ces catégories. Au lieu de travailler sur les dissimilarités entre catégories, ces indices attribuent des poids différents aux catégories. L'utilisation de ces indices concerne essentiellement l'écologie et particulièrement les mesures de diversité taxonomique ou phylogénétique.

Ces indices attribuent des poids à des taxons à partir d'un arbre dont les longueurs de branches sont unitaires, préférentiellement un cladogramme. La mesure de diversité, dite, taxonomique est alors la somme de ces poids. Ce type de mesure est dû à Vane-Wright *et al.* (1991) qui ont proposé de pondérer un taxon par une valeur inversement proportionnelle au nombre de groupes (nœuds) séparant ce taxon de la racine de l'arbre. Cette valeur étant exprimée en pourcentage est appelée "valeur P". Elle est considérée comme la contribution de ce taxon à la diversité taxonomique totale. Les valeurs P sont utilisées pour estimer le pourcentage de la diversité taxonomique totale contenue dans un sous-ensemble d'espèces. La diversité taxonomique relative d'une région est alors calculée comme la somme des valeurs P des espèces qu'elle contient (Fig. 5).

Une première critique de cet indice a été faite par May (1990) : l'indice de Vane-Wright *et al.* (1991) ne tient pas compte des nœuds non-résolus, c'est-à-dire desquels plus de deux branches sont issues. May (1990) a alors proposé une amélioration de l'indice en comptant le nombre de branches à chaque nœud plutôt que le nombre de nœuds lui-même (Fig. 6).

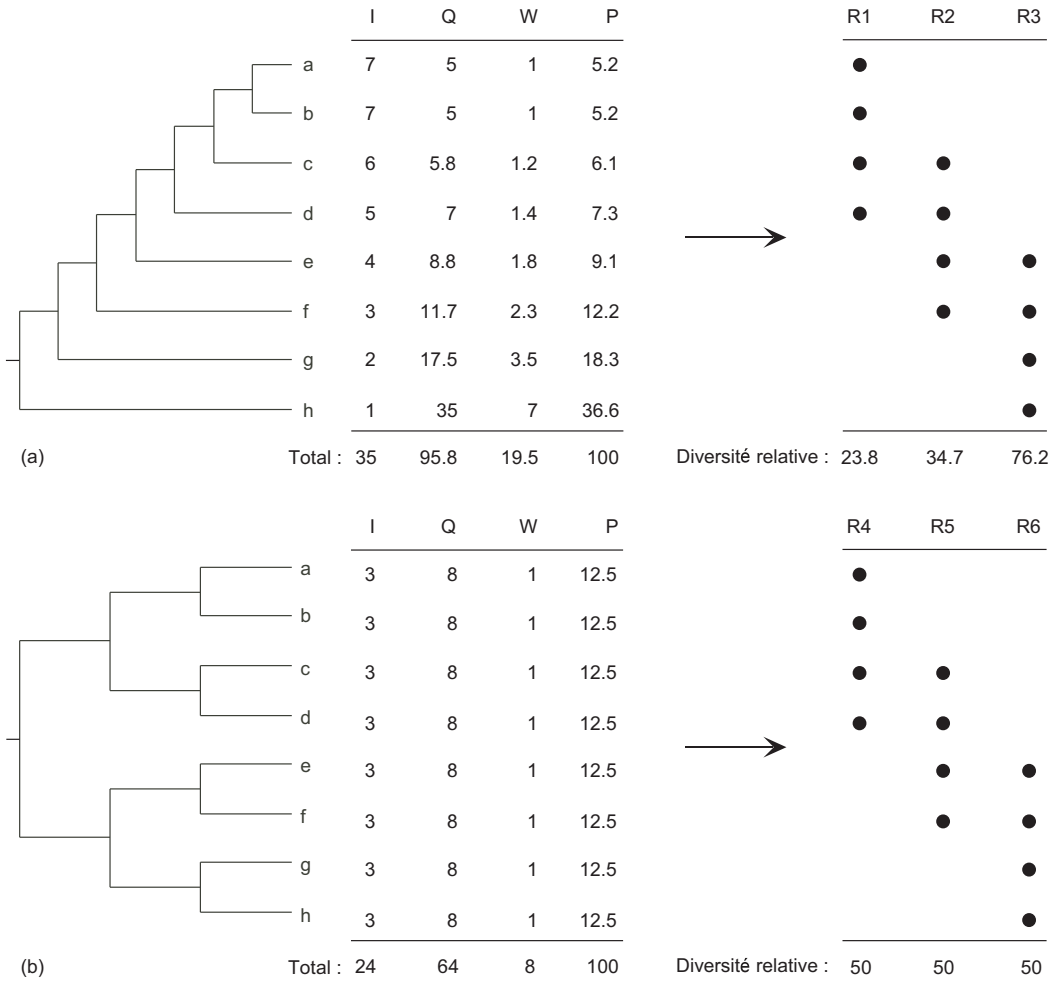


FIG. *5 – Calcul de la diversité taxonomique par l'indice de Vane-Wright* et al. *(1991) pour deux types d'arbres théoriques (a) en forme de "peigne" et (b) en forme de "buisson". Pour chaque figure, le premier tableau contient 4 valeurs. La valeur I indique le nombre de groupes auxquels chaque taxon appartient. La valeur Q est égale à (somme des valeurs I)/I. La valeur W est égale à Q/min(Q). La valeur P, poids de chaque taxon, est égale à W/(somme des valeurs W) et est donc inversement proportionnelle à I. Elle est exprimée en pourcentage. Le deuxième tableau de chaque figure propose le calcul de la diversité taxonomique relative pour trois régions théoriques. Chaque région contient 4 taxa. La diversité taxonomique relative d'une région (en pourcentage) est la somme des valeurs P de ses quatre taxa.*

La deuxième critique faite par Solow *et al.* (1993) et Humphries et Williams (1994) à propos de cet indice porte sur le calcul des diversités taxonomiques relatives des régions. Avec un arbre en forme de peigne, la diversité taxonomique relative la plus grande est attribuée à la région R3 contenant les quatre taxa séparés de la racine par les plus petits nombres de nœuds (Fig. 5a). Avec un arbre en forme de buisson, la diversité taxonomique relative dépend uniquement du nombre de taxa (Fig. 5b). La deuxième critique est donc que l'indice de Vane-Wright *et al.* n'attribue pas la plus grande diversité taxonomique relative à une région contenant les espèces les plus divergentes (Solow *et al.* 1993, Humphries et Williams 1994). Par exemple, dans la figure 5b, Vane-Wright *et al.* (1991) à travers leur indice considèrent que les trois régions ont

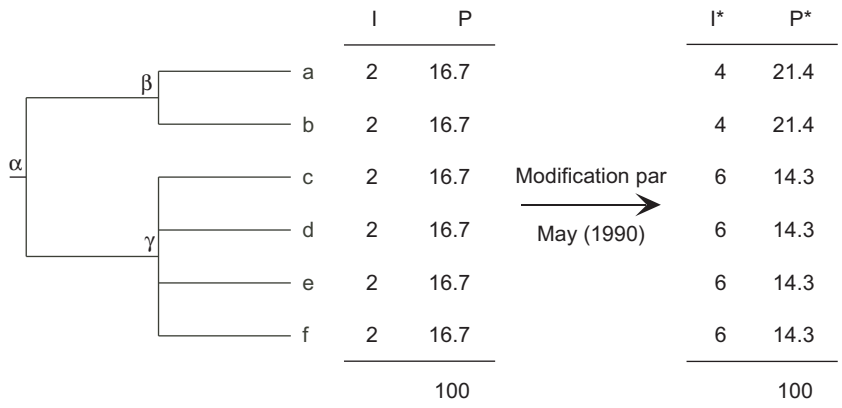


FIG. *6 – Amélioration de l'indice de Vane-Wright* et al. *(1991) par May (1990). Des nœuds α, β et γ descendent respectivement 2, 2 et 4 branches. Les nouvelles valeurs P tiennent compte de ces différences de résolution. Plus un taxon a de taxa proches moins il contribue à la diversité taxonomique. Les valeurs P peuvent être directement calculées à partir des valeurs I : P = (1/I) / (somme des valeurs 1/I).*

la même diversité taxonomique. Pourtant la région R5 qui comprend deux espèces de chacun des grands groupes devrait avoir une plus grande diversité taxonomique. Une région contenant, par exemple, les espèces a, c, e et g, devrait avoir une des plus grandes diversités taxonomiques alors que selon l'indice de Vane-Wright *et al.* elle a la même diversité taxonomique que les autres. Cette critique reste applicable aux autres indices développés ensuite sur le même modèle.

Parmi ces autres indices, ceux de Nixon et Wheeler (1992) pondèrent un taxon par les richesses des clades auxquelles il appartient. Ces indices sont plutôt destinés aux arbres qui ne sont pas déterminés jusqu'au niveau des espèces. Chaque feuille est alors un groupe d'espèces dont on connaît le nombre. Nixon et Wheeler (1992) définissent deux indices : un indice non pondéré et un indice pondéré. Pour l'indice non pondéré, à chaque nœud du cladogramme est attribuée la valeur 1 si les clades descendant de ce nœud possèdent plus d'espèces que les clades de ses nœuds frères, c'est-à-dire issus du même nœud que lui, et la valeur 0 dans le cas contraire. Pour les nœuds non résolus, des niveaux supplémentaires sont rajoutés. Pour une espèce k, les valeurs des nœuds rencontrés pour aller de la racine à l'espèce par le plus court chemin sont mis côte à côte, dans l'ordre de rencontre, formant ainsi un vecteur binaire. Tout vecteur binaire peut être transformé en une valeur décimale par l'utilisation des puissances de 2. Par exemple, le vecteur 100110 correspond à la valeur décimale $1*2^5+0*2^4+0*2^3+1*2^2+1*2^1+1*2^0 = 38$. Le poids u_k d'une espèce est inversement proportionnel à cette valeur décimale que nous noterons v_k. Nous prendrons $u_k = 1/v_k$ (Fig. 7, Tab. 2). Pour éviter une division par 0, il faut augmenter l'indice non pondéré de 1. Pour l'indice pondéré de diversité, le poids w_k d'une espèce est inversement proportionnel à la somme μ_k des nombres d'espèces présentes dans tous les sous-clades auxquels l'espèce appartient. Nous prendrons $w_k = 1/\mu_k$ (Tab. 3).

Comme Vane-Wright *et al.* l'ont fait pour leur indice, nous allons transformer les poids des espèces u_k et w_k calculés par les indices de Nixon et Wheeler en pourcentages : $u_k^* = u_k / (\sum_k u_k)$ et $w_k^* = w_k / (\sum_k w_k)$. Cette transformation en pourcentages permettra de comparer les valeurs

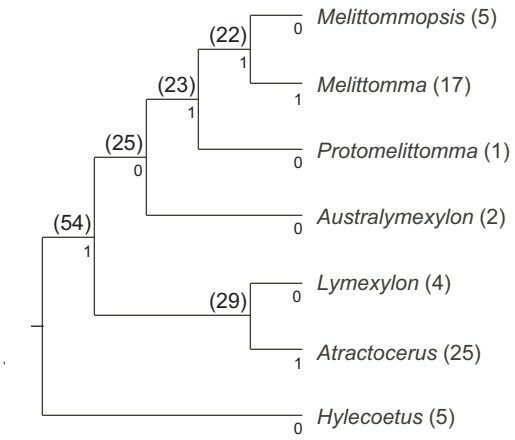


FIG. *7 – Cladogramme pour la famille Lymexydae (Coléoptères) (données de Nixon et Wheeler (1992)). Le nombre d'espèces dans chaque groupe est indiqué entre parenthèses. Les autres chiffres indiquent les valeurs binaires (0 ou 1) attribuées aux groupes.*

TAB. *2 – Indice non pondéré de Nixon et Wheeler (1992)*

Nœud	1	2	3	4	5	6	Nb binaire	Nb décimal (v)	u^*
Melittommopsis	1	0	0	1	1	0	100110	38	2.24
Melittomma	1	0	0	1	1	1	100111	39	2.18
Protomelittomma	1	0	0	1	0	-	100100	36	2.36
Australimexylon	1	0	0	0	-	-	100000	32	2.64
Atractocerus	1	1	1	-	-	-	111000	56	1.53
Lymexylon	1	1	0	-	-	-	110000	48	1.78
Hylecoetus	0	0	-	-	-	-	000000	0	87.27

P de Vane-Wright avec les valeurs u^* et w^* dérivées de Nixon et Wheeler (1992). Ces indices sont appliqués à la mesure des poids des genres de la famille Lymexydae (Coléoptères) dans la diversité de cette famille (Fig. 7). Les résultats sont donnés dans les tableaux 2 et 3. Le genre *Hylecoetus* le plus isolé est rendu très distinct par les deux indices de Nixon et Wheeler. La particularité de ces indices est donc d'attribuer de très forts poids à toutes les espèces appartenant à des clades isolés.

Comme le soulignent Altschul et Lipman (1990) et Faith (1993), les indices de Vane-Wright *et al.*, May et Nixon et Wheeler ne tiennent pas compte des longueurs de branches sur les arbres phylogénétiques. Et certains auteurs ont proposé des solutions.

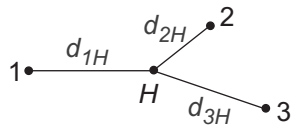
Crozier considère des arbres dans lesquels les longueurs de branches sont comprises entre 0 et 1 et représente le degré de changement phylogénétique. Il pose que l'unicité d'une espèce k est la somme de deux termes :

– la probabilité qu'il y ait eu une mutation sur la branche qui amène à cette espèce ;
– le produit de la probabilité qu'il n'y ait eu aucun changement le long de cette branche par

TAB. 3 – *Indice pondéré de Nixon et Wheeler (1992)*

						Somme (μ)	w^*
Melittommopsis	5	22	23	25	54	129	3.01
Melittomma	17	22	23	25	54	141	2.75
Protomelittomma		1	23	25	54	103	3.77
Australimexylon			2	25	54	81	4.79
Atractocerus			25	29	54	108	3.60
Lymexylon			4	29	54	87	4.46
Hylecoetus					5	5	77.62

les probabilités qu'il y ait eu des changements sur les autres branches.
Par exemple, considérons trois espèces reliées par un nœud H, et notons d_{kH} la longueur de la branche reliant l'espèce k au nœud H :



L'unicité de l'espèce 1 est

$$U_1 = d_{1H} + (1 - d_{1H})\, d_{2H} d_{3H}.$$

Faith (1994a) critique l'absence de symétrie dans cet indice et propose une version corrigée

$$U_1 = d_{1H}\,(1 - d_{2H})\,(1 - d_{3H}) + (1 - d_{1H})\, d_{2H} d_{3H}.$$

Crozier et Kusmierski (1994) répondent qu'ils considèrent un nombre infini d'allèles alors que Faith n'en considère que deux. En effet, sous les hypothèses que la taille de la population est finie et constante au fil des générations et que le nombre total d'allèles est très élevé, toute mutation conduit à un allèle qui n'a pas été vu auparavant dans la population. En d'autres termes, sous ces hypothèses, un changement sur la branche amenant à l'espèce 1 assure que le nouveau caractère issu de ce changement sera différent des caractères des autres espèces.

Contrairement à Vane-Wright *et al.* (1991), Crozier définit sa mesure de diversité génétique indépendamment des poids attribués aux espèces. La mesure qu'il propose est la probabilité qu'il soit advenu au moins une mutation sur l'une des branches de l'arbre :

$$P = 1 - \prod_k (1 - b_k)\,,$$

où b_k est la longueur de la branche k. Selon le modèle de l'arbre, il arrive que cette formule soit corrigée par une constante C de sorte que Cb_k et non b_k soit une mesure de probabilité de changement le long d'une branche.

Soient $T = (s_1, ..., s_i, ..., s_S)$ un ensemble de S espèces et E un sous-ensemble de ces espèces. Soit d_i la dissimilarité moyenne entre l'espèce s_i et les autres espèces de T. Il s'agit du poids de l'espèce s_i dans la diversité de l'ensemble. Avec des données génétiques, Eiswerth et Haney (1992) définissent la diversité de E par

$$H(E) = \sum_{s_i \in E} d_i.$$

De la même façon, Ricotta (2004) appelle "caractère taxonomique distinctif w_k d'une espèce k" la moyenne des distances entre cette espèce et les autres sur un arbre taxonomique. Mais contrairement aux indices précédents, Ricotta tient compte de l'abondance relative des espèces. Soit p_k la fréquence de l'espèce k dans un ensemble, Ricotta propose de mesurer la diversité taxonomique par l'espérance du caractère taxonomique distinctif $T(m)$ quand un échantillon aléatoire de taille m est obtenu par tirage avec remise à partir d'une communauté :

$$T(m) = \frac{\sum_{k=1}^{S} w_k \left(1 - (1 - p_k)^m\right)}{\sum_{k=1}^{S} \left(1 - (1 - p_k)^m\right)}.$$

Il avait auparavant proposé d'autres mesures de diversité taxonomique :

$$H(\mathbf{p}, \mathbf{w}) = \sum_{k=1}^{S} p_k \ln w_k \text{ (Ricotta et Avena 2003),}$$

et

$$\Delta_\beta = \frac{1}{\beta} \sum_{k=1}^{S} w_k p_k \left(1 - p_k^\beta\right) \text{ (Ricotta 2002)}$$

dont la limite lorsque $\beta \to 0$ est

$$\Delta_0 = -\sum_{k=1}^{S} w_k p_k \ln p_k$$

Ricotta a défini ces indices avec des distances taxonomiques mais suggère aussi leurs utilisations avec des distances fonctionnelles entre espèces. En référence à Webb (2000), Ricotta (2004), comme Vane-Wright *et al.* (1991), suggère que, pour pouvoir comparer les diversités dans plusieurs communautés, les poids des espèces soient calculés à partir d'un arbre englobant les espèces de l'ensemble des communautés, ce qui est également cohérent avec la mesure de Eiswerth et Haney (1992).

En tant que mesures de diversité, ces indices sont très peu utilisés dans la pratique, notamment en biologie de la conservation (Warwick et Clarke 2001). Toutes ces mesures échouent à attribuer la plus grande diversité à un ensemble contenant les espèces les plus divergentes. Elles seront étudiées à nouveau dans la partie 5.3 avec un autre champ d'utilisation. Pour notre but actuel qui est de mesurer la diversité en tenant compte des différences entre espèces, d'autres indices doivent donc être cherchés.

2.3.4 Mesurer la diversité à partir d'une matrice de dissimilarités entre catégories

Izsak et Papp (2000) proposent que la diversité soit simplement mesurée par la somme des termes de la matrice de dissimilarités entre catégories :

$$F = \sum_{k=1}^{S} \sum_{l=1}^{S} d_{kl}.$$

Cet indice a l'avantage d'être monotone pour l'ajout d'une ou plusieurs catégorie(s) dans l'ensemble des catégories de départ : soient A un ensemble de S catégories et x une catégorie non présente dans A,

$$F(A \cup \{x\}) = \sum_{k=1}^{S} \sum_{l=1}^{S} d_{kl} + \sum_{k=1}^{S} d_{k,S+1} > \sum_{l=1}^{S} d_{kl} = F(A).$$

L'avantage de F est qu'il n'est ni un indice de richesse ni une fonction croissante de la richesse. Il apporte donc une information différente de celle de la richesse. D'autres indices plus complexes ont été proposés.

Weitzman (1992) a développé sa théorie à partir des travaux de Solow *et al.* (1993) qui, malgré leur date de publication, avaient été écrits avant l'article de Weitzman. Soit A un ensemble d'espèces. La fonction V de diversité est définie par un algorithme de façon inductive comme la solution de :

$$V(A) = \max_{k \in A} \left(V(A \backslash k) + d(k, A \backslash k) \right)$$
$$\text{où } d(k, A \backslash k) = \min_{l \in \{A \backslash k\}} (d_{kl}).$$

$V(A)$ est déterminée de façon récursive. La solution est unique pour d_0 fixé tel que $V(k) = d_0$ pour tout k (la diversité d'un ensemble contenant une seule espèce est une constante). Solow *et al.* (1993) proposent $d_0 = 0$. Sous certaines conditions (Weitzman 1992, pages 398-400), cet indice est équivalent à celui de Shannon. Cet algorithme fait intervenir une méthode de classification hiérarchique qui aboutit à un schéma de type arbre (Weitzman 1992). A chaque nœud de cet arbre est attribuée la valeur égale à la distance qui le sépare des feuilles (ici les espèces). La diversité de l'ensemble des espèces de l'arbre est égale à la somme des valeurs attribuées aux nœuds. Elle est aussi égale à la somme totale des longueurs de branches sur l'arbre moins une fois la hauteur totale de l'arbre.

Prenons par exemple l'arbre de la figure 8a. Notons s l'ensemble {a,b,c,d}. La plus petite distance entre deux espèces est

$$\min_{kl} (d_{kl}) = d_{\text{ac}} = l_1 + l_3 + l_4.$$

Donc $V(s) = d_{\text{ac}} + \max_{k \in \{\text{a,c}\}} V(s \backslash k)$.

$$V(s \backslash a) \quad = \quad d_{\text{cb}} + \max_{k \in \{\text{b,c}\}} V(s \backslash \{a, k\})$$

$$
\begin{aligned}
&= d_{\mathrm{cb}} + d_{\mathrm{bd}} = 2l_2 + 2l_3 + l_4 + l_5 + l_6 \\
V(s\backslash c) &= d_{\mathrm{ab}} + \max_{k\in\{\mathrm{a,b}\}} V(s\backslash\{c,k\}) \\
&= d_{\mathrm{ab}} + d_{\mathrm{bd}} = l_1 + 2l_2 + l_3 + l_5 + l_6
\end{aligned}
$$

comme $V(s\backslash c) < V(s\backslash a)$,

$$
\begin{aligned}
V(s) &= d_{\mathrm{ac}} + V(s\backslash a) \\
&= d_{\mathrm{ac}} + d_{\mathrm{cb}} + d_{\mathrm{bd}} \\
&= l_1 + 2l_2 + 3l_3 + 2l_4 + l_5 + l_6
\end{aligned}
$$

L'arbre théorique associé à cette procédure est donné dans la figure 8b. Il est ultramétrique. Pour construire cet arbre, à chaque étape de l'algorithme, les espèces sont séparées en deux catégories : lien et représentant. Le lien est l'espèce dont la perte provoquerait la plus petite diminution de diversité.

Pour notre exemple, la plus petite distance entre deux espèces est d_{ac}. Les deux premières espèces à comparer sont donc *a* et *c*. La perte de *a* provoquerait une plus petite diminution de diversité ($V(s\backslash c) < V(s\backslash a)$). *a* est donc le lien et *c* le représentant du groupe $\{a, c\}$. La hauteur de la branche séparant les espèces *a* et *c* de leur premier nœud commun est $d_{ac} = l_1 + l_3 + l_4$. La plus petite distance entre deux espèces dans le groupe $s\backslash a$ est d_{cb}. Les deux espèces a comparer sont donc maintenant *b* et *c*. Comme $V(s\backslash\{a, c\}) < V(s\backslash\{a, b\})$, *c* est le lien et *b* le représentant du groupe $\{a, b, c\}$. La distance séparant les espèces *a*, *b* et *c* de leur premier nœud commun est donc $d_{bc} = l_2 + l_3 + l_4$. L'ensemble $s\backslash\{a, c\}$ ne comprend plus que deux espèces (*b* et *d*), la distance entre les espèces *a*, *b*, *c* et *d* et leur premier nœud commun est donc $d_{bd} = l_2 + l_3 + l_5 + l_6$. Le choix de *b* ou *d* comme représentant n'a à ce stade final aucune importance. La construction de l'arbre peut se résumer dans le tableau suivant :

Représentant	Lien	Distance
c	a	$l_1 + l_3 + l_4$
b	c	$l_2 + l_3 + l_4$
d	b	$l_2 + l_3 + l_5 + l_6$

Si la distance est ultramétrique alors $V(A) = V(A\backslash k) + d(k, A\backslash k)$. Une des limites de cette méthode, soulignée par Solow et Polasky (1994) est que, en dehors des distances ultramétriques, l'indice de Weitzman n'est pas strictement monotone par rapport aux distances. Par exemple, dans le cas de trois espèces, la mesure de Weitzman correspond à la somme de la plus grande et de la plus petite distance. Elle est donc insensible à toute modification de la distance intermédiaire.

La question de prendre en compte la taxonomie et la phylogénie dans les indices de diversité a été soulevée il y a plusieurs années (Vane-Wright *et al.* 1991, Humphries *et al.* 1995). Faith (1992a) propose un indice appelé "diversité phylogénétique" (PD, "Phylogenetic Diversity") qui est égal à la somme des longueurs de branches sur un arbre phylogénétique. Le but de cette mesure est de décrire la richesse en caractères, nombre total de caractères (phénotypiques, physiologiques, etc.) présents dans un ensemble d'espèces. Un caractère est ici l'état (*e.g.* bleu)

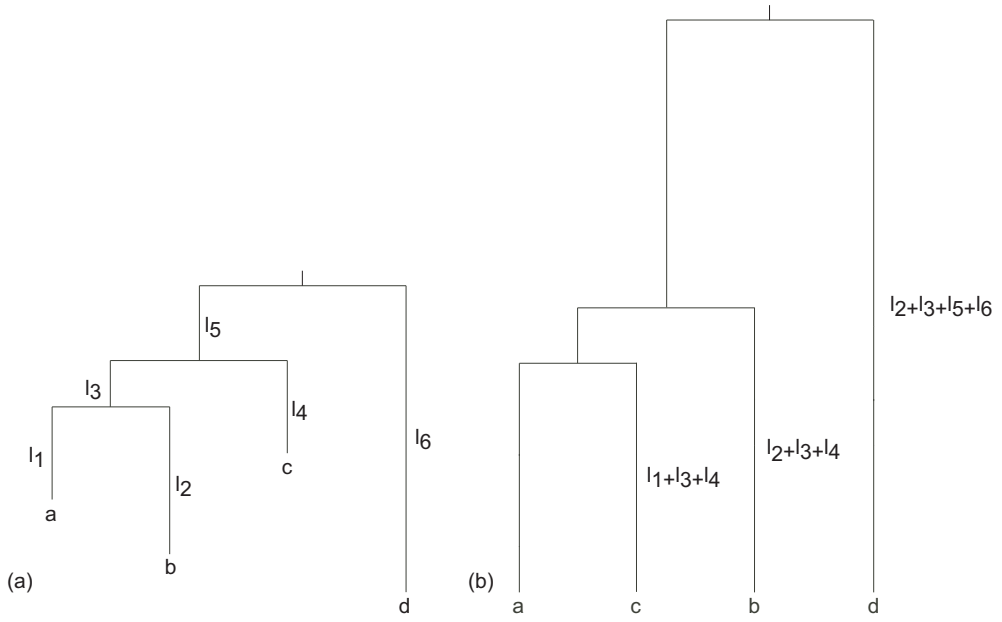


FIG. *8 – Calcul de diversité phylogénétique. (a) Exemple théorique d'arbre phylogénétique. Les valeurs l_i indiquent des longueurs de branches. (b) Arbre induit par la méthode de Weitzman.*

pris par une variable (*e.g.* couleur des yeux). Comme il est impossible de connaître tous les caractères correspondant aux états de toutes les variables biologiques de toutes les espèces, la phylogénie va nous aider à estimer cette richesse. Cela demande d'obtenir une phylogénie dont les longueurs de branches peuvent être raisonnablement exprimées en nombre de caractères. Deux espèces éloignées sur une telle phylogénie ont plus de caractères différents que deux espèces proches.

En partant d'un arbre phylogénétique de S espèces, la mesure PD sur un sous-ensemble s de ces espèces est égale à la somme de toutes les branches qui tracent le plus petit chemin ("minimum spanning path") défini par s. Je montre ici que la somme des longueurs de branches de ce plus petit chemin peut être calculée directement à partir d'une matrice de dissimilarités. Cette démonstration nous permettra ensuite de mettre en évidence des similitudes entre la théorie de Weitzman et celle de Faith.

La distance d_{ab} entre deux espèces a et b est la somme des longueurs de branches sur le chemin qui les relie. Faith (1992a) montre que si un sous-ensemble s est un ensemble d'espèces faisant l'objet d'un programme de conservation alors il est possible de calculer le gain G sur la valeur PD apporté par l'ajout d'une autre espèce :

$$G = \min_{k,l \in s} \left[\frac{1}{2} (d_{xk} + d_{xl} - d_{kl}) \right].$$

Notons cette valeur de gain $G_{/s}(x)$ (gain apporté par x à l'ensemble s ou gain de x sachant s) et $\mathrm{PD}(s)$ la valeur de la diversité phylogénique dans s. De cette définition de G, nous pouvons déduire que

$$\mathrm{PD}(s) = \mathrm{PD}(s \setminus j) + G_{/(s \setminus j)}(j),$$

pour toute espèce j de s. Cela veut aussi dire que l'on peut définir PD de manière itérative en partant de deux espèces seulement et en rajoutant à chaque étape le gain apporté par une nouvelle espèce. Reprenons par exemple l'arbre de la figure 8a. Ce processus itératif peut être décrit de la façon suivante :

1. Sous-ensemble {a,b}

$$\mathrm{PD}(\{a,b\}) = d_{ab}$$

2. Sous-ensemble {a,b,c}

$$\begin{aligned}\mathrm{PD}(\{a,b,c\}) &= \mathrm{PD}(\{a,b\}) + G_{/\{a,b\}}(c)\\ &= \mathrm{PD}(\{a,b\}) + \frac{1}{2}(d_{ca} + d_{cb} - d_{ab})\\ &= \frac{1}{2}(d_{ab} + d_{ca} + d_{cb})\end{aligned}$$

3. Ensemble {a,b,c,d}

$$\begin{aligned}\mathrm{PD}(\{a,b,c,d\}) &= \mathrm{PD}(\{a,b,c\}) + G_{/\{a,b,c\}}(d)\\ &= \mathrm{PD}(\{a,b,c\}) + \min_{kl \text{ dans } \{a,b,c\}} \frac{1}{2}(d_{dk} + d_{dl} - d_{kl})\end{aligned}$$

Gains :

$$\begin{aligned}G_{/\{a,b\}}(d) &= \frac{1}{2}(d_{da} + d_{db} - d_{ab}) = \frac{1}{2}((l_1 + l_3 + l_5 + l_6) + (l_2 + l_3 + l_5 + l_6) - (l_1 + l_2))\\ &= l_3 + l_5 + l_6\\ G_{/\{a,c\}}(d) &= \frac{1}{2}(d_{da} + d_{dc} - d_{ac}) = \frac{1}{2}((l_1 + l_3 + l_5 + l_6) + (l_4 + l_5 + l_6) - (l_1 + l_3 + l_4))\\ &= l_5 + l_6\\ G_{/\{b,c\}}(d) &= \frac{1}{2}(d_{db} + d_{dc} - d_{bc}) = \frac{1}{2}((l_2 + l_3 + l_5 + l_6) + (l_4 + l_5 + l_6) - (l_2 + l_3 + l_4))\\ &= l_5 + l_6\end{aligned}$$

Ainsi,

$$\begin{aligned}\mathrm{PD}(\{a,b,c,d\}) &= \mathrm{PD}(\{a,b,c\}) + \frac{1}{2}(d_{da} + d_{dc} - d_{ac})\\ &= \frac{1}{2}(d_{ab} + d_{ac} + d_{bc}) + \frac{1}{2}(d_{da} + d_{dc} - d_{ac})\\ &= \frac{1}{2}(d_{ab} + d_{bc} + d_{cd} + d_{ad})\end{aligned}$$

C'est-à-dire

$$\begin{aligned}\mathrm{PD}(\{a,b,c,d\}) &= \frac{1}{2}((l_1 + l_2) + (l_2 + l_3 + l_4) + (l_4 + l_5 + l_6) + (l_1 + l_3 + l_5 + l_6))\\ &= l_1 + l_2 + l_3 + l_4 + l_5 + l_6\end{aligned}$$

Nous pouvons conclure que l'indice de diversité de Weitzman et celui de Faith se font sur des schémas qui se ressemblent. Pour un arbre ultramétrique, ces deux indices sont très liés : l'indice de Weitzman est égal à l'indice de Faith (somme totale des longueurs de branches sur l'arbre ultramétrique (PD)) moins la hauteur de l'arbre ultramétrique.

Faith (1992a) a développé l'indice PD pour des arbres phylogénétiques et donc des dissimilarités phylogénétiques entre espèces. D'un autre côté, il suggère aussi que l'indice PD soit utilisé à partir de dissimilarités phénétiques (Faith 1992b). Petchey et Gaston (2002) ont proposé d'utiliser le même indice sur un arbre ultramétrique obtenu à partir de distances fonctionnelles entres espèces. Nee et May (1997) utilisent un indice similaire à partir d'arbres ultramétriques représentant l'histoire évolutive des espèces.

Williams *et al.* (1991) ont proposé une série d'indices basés sur une mesure de divergence entre deux espèces ; mesure obtenue à partir d'un cladogramme. La divergence U_{kl} entre l'espèce k et l'espèce l est le nombre de nœuds possédés par k, c'est-à-dire présents entre la feuille k et la racine de l'arbre, mais pas par l. Cette mesure est asymétrique : $U_{kl} \neq U_{lk}$. Ils définissent aussi S_{kl} qui est le nombre de nœuds partagés par les espèces depuis la racine de l'arbre. Williams *et al.* (1991), en prenant comme référence la richesse (mesure I) et l'indice de Vane-Wright *et al.* (1991) (mesure II, cf. partie 2.3.2) définissent alors les indices de diversité suivants :

$$[S \times \mathrm{m.}(d_{kl})]_{sample} \times 100 \Big/ [S \times \mathrm{m.}(d_{kl})]_{clade},$$

avec

$$d_{kl} = 1/S_{kl} \qquad \text{(mesure III)}$$

$$d_{kl} = U_{kl} + U_{lk} + 1 \qquad \text{(mesure IV)}$$

$$d_{kl} = \left(U_{jk} + U_{kj} + 1\right) \Big/ \left(2S_{jk} + U_{jk} + U_{kj}\right) \qquad \text{(mesure V)}$$

$$d_{kl} = \left[\left(U_{jk} + U_{kj}\right) \Big/ S_{jk}\right] + 1 \qquad \text{(mesure VI)}$$

$$[S \times (\mathrm{m.}(d_{kl}) - \mathrm{e.t.}(d_{kl}))]_{sample} \times 100 \Big/ [S \times (\mathrm{m.}(d_{kl}) - \mathrm{e.t.}(d_{kl}))]_{clade},$$

avec

$$d_{kl} = U_{kl} + U_{lk} + 1 \qquad \text{(mesure VII)}$$

"m." désigne la moyenne, "e.t." l'écart type et S le nombre d'espèces.

La mesure III est appelée "richesse en taxa plus élevés" puisque la valeur S_{ij} sera moins grande pour des espèces ayant divergé près de la racine du cladogramme. La mesure de dispersion VII a pour but de donner une plus grande diversité à un ensemble d'espèces dispersées donc éloignées les unes des autres, dans le cladogramme. Observant que, selon la mesure VII, dans certaines circonstances une espèce peut diminuer la diversité d'un ensemble dans lequel elle est ajoutée, Williams et coll. (Humphries et Williams 1994, Williams *et al.* 1993) introduisent une autre mesure de dispersion :

$$\nu + e^{[-\mathrm{e.t.}(d_{ij})/\mathrm{m.}(d_{ij})]}$$

où ν est le nombre total de nœuds dans le sous-arbre contenant uniquement les espèces de l'ensemble considéré.

May (1990) et Izsak et Papp (2000) affirment que nous avons besoin de combiner les mesures quantitatives de distinction taxonomique avec des considérations d'abondances, ou de fréquences, qui sont plus familières à l'écologie. Or, parmi les indices cités dans cette partie, seul celui de Ricotta rentre dans ce cadre. Mais il présente le même défaut que l'indice de Vane-Wright *et al.* : il n'attribue pas la plus grande diversité à une région dont les espèces sont très distinctes. Des travaux récents permettent maintenant de progresser dans la mesure des diversités taxonomique et phylogénétique (Izsak et Papp 2000, Warwick et Clarke 1995) incluant les fréquences des espèces, et même dans la mesure de la diversité au sens large.

2.4 L'

Tous les indices cités dans la partie 2.3 intègrent des mesures de différences entre espèces, mais omettent les notions d'abondance et de rareté qui sont centrales dans les indices de Shannon et de Gini-Simpson. Gore (1994), dans un commentaire à Solow et Polasky (1994), mentionne que c'est agir comme si l'on "jetait le bébé avec l'eau du bain". Par exemple comme le souligne Warwick et Clarke (1995), dans le cas de l'évaluation de l'impact d'une pollution sur la diversité d'un assemblage, des données d'abondance doivent être prises en compte en plus d'une description des degrés de dissimilarités entre espèces.

En 1982, Rao, généralisant les travaux de Nei et coll. (Nei et Li 1979, Nei et Tajima 1981), introduit un même indice dans deux revues différentes (Rao 1982a, b). Il donne alors deux noms différents à cet indice : coefficient de diversité et entropie quadratique. Son indice sera aussi connu plus tard sous le nom de diversité quadratique. Par la suite, il a choisi le terme "entropie quadratique" qui est gardé ici. Rao présente cet indice comme une mesure fondamentale qui peut être utilisée dans tous les domaines touchant à la diversité. Il cite par exemple, l'anthropologie, la génétique, l'économie, la sociologie, et la biologie. L'histoire de cet indice peut être décrite ainsi.

2.4.1 Quatre développements indépendants

La lecture d'articles provenant de disciplines différentes m'a permis de voir que l'entropie quadratique a en fait été développé au moins quatre fois en biologie : trois fois en écologie (une fois pour mesurer la biodiversité d'une façon générale en prenant les espèces comme référentiel, une fois pour mesurer la diversité des interactions entre espèces, une fois pour mesurer la diversité taxonomique en biologie marine) et une fois en génétique. Dans les quatre paragraphes suivants, ces quatre développements sont donnés par ordre chronologique.

- Hendrickson et Ehrlich (1971) :

En écologie, l'idée d'utiliser à la fois les abondances des espèces et les différences entre ces espèces pour construire un indice de biodiversité semble être apparue pour la première fois dans Hendrickson et Ehrlich (1971). Quand Hendrickson et Ehrlich parlent de différence entre espèces, ils entendent n'importe quel type de différence qu'elle soit écologique, morphologique, évolutive, ou tout autre. Ces auteurs ont suggéré de modifier l'indice de Gini-Simpson pour

prendre en compte les différences entre espèces. Ils proposent alors l'indice suivant :

$$\frac{\sum_{k=1}^{s} n_k \left(\sum_{l=1}^{s} n_l x_{kl}\right)}{N^2 - N}$$

où n_k et n_l sont respectivement les nombres d'individus des espèces k et l, x_{kl} est une estimation de la différence entre les deux espèces k et l, et N est le nombre total d'individus dans toutes les espèces confondues. Pour que la valeur de l'indice soit comprise entre 0 et 1, Hendrickson et Ehrlich proposent que la valeur de la différence entre deux espèces soit toujours comprise entre 0 et 1 : $0 \leq x_{kl} \leq 1$ pour tout k et l. Ils proposent également que cette valeur pourrait être calculée en terme de recouvrement de niche :

$$\frac{\cup(\text{niche 1, niche 2}) - \cap(\text{niche 1, niche 2})}{\cup(\text{niche 1, niche 2})}$$

.

• Nei et Li (1979) :

En génétique, Nei et Li (1979) introduisent un indice similaire à celui de Hendrickson et Ehrlich pour mesurer la diversité d'une population à partir de séquences nucléotidiques. Les catégories sont donc les séquences distinctes et la dissimilarité entre deux séquences est la proportion de nucléotides qui diffèrent entre ces deux séquences. Cette mesure est née de l'observation que, lorsqu'il existe beaucoup de gènes ou d'allèles (par exemple sur l'ADN mitochondrial), la diversité allélique, mesurée par l'indice de Gini-Simpson, est proche de 1 pour beaucoup de populations. Pour que la variété génétique de ces populations soit évaluée par une mesure plus appropriée, une solution est de prendre le nombre de différences nucléotidiques par site entre deux séquences d'ADN tirées aléatoirement.

• Margalef et Gutierrez (1983) :

En écologie, Margalef et Gutierrez (1983) ont développé un indice similaire à celui de Hendrickson et Ehrlich pour mesurer la diversité de processus d'interactions entre espèces telle que la compétition. La dissimilarité entre deux espèces est alors une valeur, comprise entre 0 et 1, quantifiant leur niveau d'interactions.

• Warwick et Clarke (1995) :

En écologie marine, Warwick et Clarke (1995) introduisent une mesure de diversité taxonomique. La formule qu'ils proposent est identique à celle de Hendrickson et Ehrlich. La différence entre deux espèces est mesurée par le nombre de niveaux taxonomiques qui les séparent. Par exemple, une distance de 1 est attribuée à deux espèces d'un même genre, 2 entre deux espèces de genres différents mais de même famille, etc.

2.4.2 La généralisation de Rao : l'entropie quadratique

Rao (1982a) part des travaux de Nei pour introduire l'entropie quadratique.

Considérons une collection d'entités regroupées en catégories. Soient $\mathbf{p} = (p_1, ..., p_k, ..., p_S)^t$ la distribution de fréquences des catégories et d_{kl}^{cat} une mesure de la différence entre deux catégories, l'entropie quadratique est égale à

$$H_{\mathbf{D}^{\text{cat}}}(\mathbf{p}) = \sum_{k=1}^{S} \sum_{l=1}^{S} p_k p_l d_{kl}^{\text{cat}},$$

où $\mathbf{D}^{\text{cat}}$ est la matrice contenant les dissimilarités entre catégories. Avec une écriture matricielle,

$$H_{\mathbf{D}^{\text{cat}}}(\mathbf{p}) = \mathbf{p}^t \mathbf{D}^{\text{cat}} \mathbf{p}.$$

• Un indice très général :

D'un point de vue statistique, cet indice généralise la variance et l'indice de Gini-Simpson que nous avons vu précédemment. En effet, en choisissant $d_{kl}^{\text{cat}} = 1$ pour tout $k \neq l$, nous retrouvons l'indice de Gini-Simpson (Rao 1982a) :

$$d_{kl}^{\text{cat}} = 1 \text{ pour tout } k \neq l \Rightarrow H_{\mathbf{D}^{\text{cat}}}(\mathbf{p}) = 2 \sum_{k=1}^{S-1} \sum_{l=k+1}^{S} p_k p_l = 1 - \sum_{k=1}^{S} p_k^2.$$

L'indice le plus courant qui permet de tenir compte des différences entre catégories pour mesurer la variabilité d'une collection est la variance. La variance d'une variable Y est habituellement vue comme la moyenne des carrés des écarts entre les valeurs $\{y_k\}$ pour chaque catégorie et la valeur moyenne $y_\bullet$:

$$V(Y) = \sum_{k=1}^{S} p_k (y_k - y_\bullet)^2. \tag{2.1}$$

Cette formule peut être réécrite comme le carré de la différence attendue entre deux entités tirées au hasard dans cette collection (Lebart 1969, Light et Margolin 1971) :

$$V(Y) = \frac{1}{2} \sum_{k=1}^{S} \sum_{l=1}^{S} p_k p_l (y_k - y_l)^2. \tag{2.2}$$

Dans cette formule $|y_k - y_l|$ est la mesure de la distance euclidienne, encore appelée distance canonique (cf. partie 2.3.1), entre les valeurs prises par la variable Y pour les catégories k et l. Notons δ_{kl}^{cat} cette distance entre les catégories k et l. La formule de la variance de Y devient

$$V(Y) = \frac{1}{2} \sum_{k=1}^{S} \sum_{l=1}^{S} p_k p_l (\delta_{kl}^{\text{cat}})^2. \tag{2.3}$$

Notons maintenant d_{kl}^{cat} la moitié du carré de la distance euclidienne entre y_k et y_l :

$$d_{kl}^{\text{cat}} = \frac{\left(\delta_{kl}^{\text{cat}}\right)^2}{2} = \frac{(y_k - y_l)^2}{2}$$

Ainsi d'après les égalités 2.1, 2.2 et 2.3,

$$d_{kl}^{\text{cat}} = \frac{(y_k - y_l)^2}{2} \text{ pour tout } k, l \Rightarrow H_{\mathbf{D}^{\text{cat}}}(\mathbf{p}) = \sum_{k=1}^{S} p_k (y_k - y_\bullet)^2.$$

• Nouvelle écriture, dissimilarité d et δ :

Si au lieu de mesurer la différence entre catégories à partir d'une variable qualitative, nous considérons n'importe quel type de dissimilarité δ_{kl}^{cat} entre deux catégories k et l, cette mesure devient

$$H_{\Delta^{\text{cat}}}(\mathbf{p}) = \frac{1}{2} \sum_{k=1}^{S} \sum_{l=1}^{S} p_k p_l (\delta_{kl}^{\text{cat}})^2,$$

où Δ^{cat} désigne la nouvelle matrice contenant l'ensemble des dissimilarités entre catégories. Cette formule correspond à une nouvelle écriture de l'entropie quadratique.

Par la suite, nous nous intéresserons aux deux matrices $\mathbf{D}^{\text{cat}} = [\left(\delta_{kl}^{\text{cat}}\right)^2/2] = [d_{kl}^{\text{cat}}]$ et $\Delta^{\text{cat}} = [\delta_{kl}^{\text{cat}}]$. Si d_{kl}^{cat} est une mesure de dissimilarité alors δ_{kl}^{cat} l'est aussi et inversement. De plus si d_{kl}^{cat} et δ_{kl}^{cat} sont des mesures de dissimilarité alors n'importe quelle puissance positive appliquée à ces mesures fournit une valeur de dissimilarité. Le choix de la dissimilarité se fera donc d'un point de vue biologique selon le type de données analysées. Par la suite, nous parlerons de **"dissimilarité** d**"** (pour d_{kl}^{cat}) et **"dissimilarité** δ**"** (pour δ_{kl}^{cat}), ceci pour indiquer la façon dont elles seront employées dans l'entropie quadratique.

• L'entropie quadratique est concave :

Une des propriétés qui ont été mises en avant pour les fonctions cherchant à mesurer la diversité en catégories, en particulier pour les indices $H_{\mathbf{r}}$, H_{S}, $H_{\text{G-S}}$ et $H_{\text{H-C}}$, est la concavité. Une condition suffisante pour que l'entropie quadratique soit concave est que $\mathbf{D}^{\text{cat}}$ soit conditionnellement définie négative (Rao et Nayak 1985), c'est-à-dire $\mathbf{a}^t\mathbf{D}^{\text{cat}}\mathbf{a} \leq 0$ pour tout vecteur $\mathbf{a}$ de longueur S et vérifiant $\mathbf{a}^t\mathbf{1}_S = 0$, où $\mathbf{1}_S$ est le vecteur de longueur S ne contenant que des 1. $\mathbf{D}^{\text{cat}}$ est conditionnellement définie négative si et seulement si Δ^{cat} est euclidienne. Cette propriété sera utilisée dans la partie 4.

• Espérance et variance :

Regardons la valeur de l'entropie quadratique lorsqu'elle doit être estimée à partir d'un échantillon. Soit $\pi = (\pi_1, \pi_2, ..., \pi_S)^t$ le vecteur de la distribution de fréquences inconnue dans

une collection. Considérons un échantillon de cette collection. Notons n_k, $k = 1, ..., S$ le nombre observé d'entités dans la catégorie k et n le nombre total d'entités échantillonnées. Alors $\hat{\pi}_k = n_k/n = p_k$ est une estimation de π_k. Soit le vecteur $\mathbf{p} = (p_1, p_2, ..., p_S)^t$. Une estimation de $H_{\mathbf{D}^{\text{cat}}}(\pi)$ est

$$H_{\mathbf{D}^{\text{cat}}}(\mathbf{p}) = \mathbf{p}^t \mathbf{D}^{\text{cat}} \mathbf{p}.$$

En supposant que les variables $(N_1, ..., N_S)$ des nombres d'entités dans chaque catégorie suivent une loi multinomiale de paramètres n et $\pi = (\pi_1, ..., \pi_S)^t$, et en notant $\mathbf{P} = (N_1/n, ..., N_S/n)^t$ Nayak (1983, 1985) démontre que

$$E(\mathbf{P}^t \mathbf{D}^{\text{cat}} \mathbf{P}) = \frac{n-1}{n} \pi^t \mathbf{D}^{\text{cat}} \pi$$

et

$$\begin{aligned} V(\mathbf{P}^t \mathbf{D}^{\text{cat}} \mathbf{P}) = \frac{n-1}{n^3} \quad & [(6-4n)(\pi^t \mathbf{D}^{\text{cat}} \pi)^2 + 8(n-2) \\ & \{\sum_{i=1}^{S-2} \sum_{j>i}^{S-1} \sum_{k>j}^{S} \pi_i \pi_j \pi_k \left(d_{ij}^{\text{cat}} d_{ik}^{\text{cat}} + d_{ij}^{\text{cat}} d_{kj}^{\text{cat}}\right)\} \\ & +4 \sum_{i=1}^{S-1} \sum_{j>i}^{S} \left(d_{ij}^{\text{cat}}\right)^2 \pi_i \pi_j \left\{(n-2)\left(\pi_i + \pi_j\right) + 1\right\}]. \end{aligned}$$

Notons Σ la matrice de variances-covariances des variables $N_1, ..., N_S$:

$$\Sigma = \begin{pmatrix} \pi_1(1-\pi_1) & -\pi_1\pi_2 & ... & -\pi_1\pi_S \\ -\pi_1\pi_2 & \pi_2(1-\pi_2) & ... & -\pi_2\pi_S \\ ... & ... & ... & ... \\ -\pi_1\pi_S & -\pi_2\pi_S & ... & \pi_S(1-\pi_S) \end{pmatrix},$$

alors $V(\mathbf{P}^t \mathbf{D}^{\text{cat}} \mathbf{P})$ peut être réécrit plus simplement ainsi :

$$V(\mathbf{P}^t \mathbf{D}^{\text{cat}} \mathbf{P}) = \frac{4\pi^t \mathbf{D}^{\text{cat}} \Sigma \mathbf{D}^{\text{cat}} \pi}{n}.$$

L'estimateur $H_{\mathbf{D}^{\text{cat}}}(\mathbf{P}) = \mathbf{P}^t \mathbf{D}^{\text{cat}} \mathbf{P}$ pour $H_{\mathbf{D}^{\text{cat}}}(\pi)$ est donc biaisé, contrairement à l'estimateur choisi par Hendrickson et Ehrlich (1971)

$$\frac{n}{n-1} \mathbf{P}^t \mathbf{D}^{\text{cat}} \mathbf{P}$$

qui lui n'est pas biaisé. Cependant ce biais s'annule puisque $(n-1)/n$ tend vers 1 lorsque n tend vers l'infini. L'estimateur $H_{\mathbf{D}^{\text{cat}}}(\mathbf{P})$ est donc asymptotiquement non biaisé. De plus $V(H_{\mathbf{D}^{\text{cat}}}(\mathbf{P}))$ tend vers 0 quand n tend vers l'infini. Ainsi $H_{\mathbf{D}^{\text{cat}}}(\mathbf{P})$ est un bon estimateur pour $H_{\mathbf{D}^{\text{cat}}}(\pi)$ (Nayak 1983).

- Tests d'hypothèses :

Sauf dans un cas très particulier (Nayak 1983, propositions 4.4.6 et 4.4.7),

$$\sqrt{n}\left(H_{\mathbf{D}^{\text{cat}}}(\mathbf{P}) - H_{\mathbf{D}^{\text{cat}}}(\pi)\right) \rightarrow N\left(0, 4\pi^t \mathbf{D}^{\text{cat}} \Sigma \mathbf{D}^{\text{cat}} \pi\right).$$

Les tests des hypothèses

– $H_0 : D = D_0$, la diversité d'une collection est égale à la valeur D_0,
– et $H_0 : D_1 = D_2 = ... = D_r$, les diversités de plusieurs collections sont égales,

proposés pour les indices de Shannon H_S, de Gini-Simpson $H_{G\text{-}S}$ et de Havrda et Charvat $H_{H\text{-}C}$ s'appliquent également à $H_{\mathbf{D}^{cat}}$ en prenant $\tau^2 = 4\pi^t\mathbf{D}^{cat}\Sigma\mathbf{D}^{cat}\pi$ et $\hat{\tau}^2 = 4\mathbf{p}^t\mathbf{D}^{cat}\Sigma\mathbf{D}^{cat}\mathbf{p}$ (cf. page 30).

2.4.3 Utilisation actuelle de l'entropie quadratique

"Ecologists often revive old ideas by giving them new purpose." (Veech *et al.* 2002)

L'entropie quadratique, sous sa forme générale, est très utilisée en génétique grâce aux travaux de Nei. En écologie elle a été beaucoup plus rare. Pourtant, depuis le milieu des années 90, de plus en plus d'écologues s'y intéressent.

Izsak et Papp (1995) ont introduit, en même temps que Warwick et Clarke (1995), le même indice de diversité taxonomique, à une constante prêt. Izsák et Papp introduisent cet indice comme la continuité des travaux de Rao. Ils l'utilisent pour comparer la diversité taxonomique de drosophiles dans deux parcs nationaux et deux zones protégées des basses montagnes du nord de la Hongrie. Dans leur exemple, la diversité taxonomique est très corrélée à la diversité spécifique mesurée par l'indice de Gini-Simpson, sans doute en partie du fait de la présence de quelques espèces très dominantes. Izsak et Papp (1995) calculent aussi une diversité des types de ressources utilisées, en regroupant les drosophiles non pas par espèces mais par groupes trophiques. Ils soulignent que l'entropie quadratique pourrait être très utile dans la mesure de la diversité pour n'importe quel type d'habitude de vie et en particulier les habitudes alimentaires.

Warwick et Clarke (1995) notent leur indice de diversité taxonomique Δ. Ils l'utilisent pour mesurer la diversité d'assemblages en milieu marin selon plusieurs degrés de pollution. En plus de cet indice, ils introduisent une mesure de spécificité taxonomique ("taxonomic distinctness") :

$$\Delta^* = \frac{\sum_{i=1}^{S-1}\sum_{j=i+1}^{S} n_i n_j d_{ij}}{\sum_{i=1}^{S-1}\sum_{j=i+1}^{S} n_i n_j}.$$

Cette mesure est le nombre moyen de niveaux taxonomiques distincts entre deux individus provenant de deux espèces différentes. Clarke et Warwick (1998) considèrent également le cas de données de présence/absence pour lesquels les deux indices Δ et Δ^* sont égaux entre eux et égaux à

$$\Delta^+ = \frac{\sum_{i=1}^{S-1}\sum_{j=i+1}^{S} d_{ij}}{S\left(S-1\right)/2}.$$

Cet indice Δ^+ est très proche de la mesure IV de Williams *et al.* (1991). En supposant que m individus sont tirés aléatoirement de l'ensemble des n individus de départ, ils notent Δ_m, Δ^*_m, et Δ^+_m les valeurs prises par les trois indices sur l'échantillon obtenu et démontrent que les indices

Δ_m et Δ_m^+ sont, de par leur définition, non biaisés et que Δ_m^* est asymptotiquement non biaisé. Ils déterminent également la variance de Δ_m^+ :

$$V(\Delta_m^+) = 2(s-m)\frac{(s-m-1)\sigma_d^2 + 2(s-1)(m-2)\sigma_{\bar{d}}^2}{m(m-1)(s-2)(s-3)},$$

où

$$\sigma_d^2 = \frac{\sum_{i=1}^{S-1}\sum_{j=i+1}^{S} d_{ij}^2}{S(S-1)/2} - \bar{d}^2,\ \sigma_{\bar{d}}^2 = \frac{\sum_{i=1}^{S}\bar{d}_i^2}{S} - \bar{d}^2,\ \bar{d}_i = \frac{\sum_{j,\, j\neq i} d_{ij}}{S-1},\ \bar{d} = \frac{\sum_{i=1}^{S}\bar{d}_i}{S} = \Delta^+.$$

σ_d^2 est la variance des dissimilarités d_{ij} entre espèces différentes et $\sigma_{\bar{d}}^2$ est la variance des distances moyennes $\bar{d}_i$ entre une espèce et toutes les autres.

L'avantage de ce type d'indice par rapport aux mesures traditionnelles de biodiversité est qu'il a moins de chance d'être influencé par d'éventuelles erreurs (mauvaises identifications) dans les taxonomies (Izsak et Price 2001).

Si la liste complète des S espèces d'une région est connue, et si dans un certain site de cette région, m espèces ont été observées, il est alors possible de tester si le site en question a une variété taxonomique plus grande (resp. plus petite) que celle de l'ensemble de la région. Pour faire ce test, 1000 échantillons aléatoires de taille m sont prélevés à partir de la liste globale. La valeur de Δ^+ observée est alors comparée à l'ensemble des valeurs Δ^+ obtenues sur les échantillons théoriques. Une valeur-p ("p-value") peut alors être calculée comme la proportion des valeurs théoriques supérieures (resp. inférieures) ou égales à la valeur observée (Clarke et Warwick 1998, Warwick et Clarke 1998).

Un autre indice qui peut mesurer la diversité taxonomique d'un ensemble d'espèces avec des données de type présence/absence est PD (Faith 1992a, cf. partie 2.3.3). Selon Warwick et Clarke, PD est très dépendant de l'effort d'échantillonnage. Ils proposent PD/S mais la valeur de cet indice pour un échantillon aléatoire de m espèces donne une estimation biaisée de sa valeur pour la liste globale.

Δ^+ mesure la dissimilarité taxonomique moyenne entre deux espèces d'un ensemble. Pour un même nombre d'espèces et de niveaux taxonomiques, la même valeur de Δ^+ peut être obtenue avec des arbres taxonomiques différents. Clarke et Warwick (2001) proposent alors de compléter l'étude de Δ^+ par celle d'une mesure de la variance de la dissimilarité entre deux espèces d'un ensemble :

$$\Lambda^+ = \frac{\sum_{i=1}^{S-1}\sum_{j=i+1}^{S}\left(d_{ij} - \bar{d}\right)^2}{S(S-1)/2}.$$

Cette variance permet par exemple de distinguer les arbres de la figure 9.

Nous avons vu dans la partie 2.3.1, que la dissimilarité taxonomique entre deux espèces dépend de la définition des longueurs de branches, sur l'arbre taxonomique, entre deux niveaux taxonomiques successifs (par exemple entre espèce et genre). En donnant des poids différents à ces longueurs de branches, l'importance des niveaux hiérarchiques les plus hauts par exemple pourrait être augmentée. Clarke et Warwick ont étudié les effets de la variation des longueurs

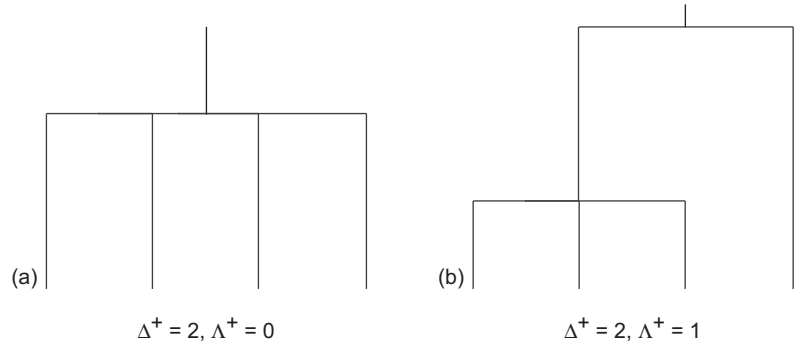


FIG. 9 – *Deux arbres taxonomiques théoriques ayant même spécificité taxonomique moyenne* Δ^+*, mais des variances de ces spécificités taxonomiques* Λ^+ *différentes : une variance nulle pour l'arbre (a) dont la forme est dite "en râteau" et une variance de 1 pour l'arbre (b) dont une espèce se distingue des autres.*

de branches entre deux niveaux successifs sur les valeurs observées prises par Δ^+ (Clarke et Warwick 1999). Dans leur étude, ils comparent la diversité taxonomique de nématodes marins aux îles Sorlingues et le long des côtes de Grande-Bretagne. En pondérant uniformément les longueurs de branches, ils trouvent que la diversité taxonomique des nématodes aux Sorlingues n'est pas différente de celle des eaux côtières de Grande-Bretagne. Par contre, en faisant varier les longueurs de branches, ils observent une plus grande dispersion de la diversité aux niveaux taxonomiques les plus élevés (sous-classes), ce qui correspond à une répartition plus homogène des espèces entre clades aux Sorlingues que dans les stations des eaux côtières de Grande-Bretagne. Après plusieurs modifications des longueurs de branches, Clarke et Warwick (1999) concluent finalement qu'en général des longueurs de branches constantes semblent adéquates pour mesurer la diversité taxonomique dans son ensemble. Dans toute la suite nous utiliserons des longueurs de branches constantes pour toutes les taxonomies.

L'indice Δ^+ était déjà utilisé avant les travaux de Warwick et Clarke et est toujours actuellement utilisé en morphométrie. En se référant à Hendrickson et Ehrlich, Findley (1973, 1976) propose trois indices pour mesurer la diversité phénotypique dans une faune. Ses indices sont basés sur la distance taxonomique de Sneath et Sokal (1973)

$$d_{kl} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(X_{kj} - X_{lj}\right)^2}$$

où X_{kj} et X_{lj} sont les valeurs d'une variable morphométrique pour les espèces k et l respectivement, et n est le nombre de traits évalués pour chaque espèce. La première mesure de diversité proposée $\bar{d}_{\min}$ est la distance moyenne entre une espèce et sa plus proche voisine dans la faune. La seconde mesure $\bar{d}$ est la distance moyenne entre deux espèces dans la faune. La troisième $\bar{d}_{fc}$ est la distance moyenne entre une espèce et une "espèce moyenne" hypothétique ou "centroïde de la faune". Les mesures $\bar{d}$ et Δ^+ sont identiques bien qu'elles n'aient pas été définies pour les mêmes données.

Losos et Miles (2002) ont introduit récemment la notion de divergence morphologique dans

un clade pour tester l'existence de radiations adaptatives, en utilisant une distance moyenne similaire à la mesure $\bar{d}$ de Findley. Contrairement à Findley, ces auteurs utilisent la distance de Mahalanobis pour calculer les dissimilarités entre espèces.

2.5 P

La crise actuelle de la biodiversité a suscité un grand intérêt des scientifiques, et ensuite du large public. Il est apparu une urgence pour l'exploration de la diversité biologique. Pour pouvoir comparer la biodiversité dans plusieurs régions du monde, il a fallu trouver des indicateurs, des indices donc des mesures décrivant et surtout résumant en valeurs numériques cette biodiversité. Les théoriciens se sont alors plongés dans la recherche de tels indices, en économie, en anthropologie, en écologie et en génétique. Des centaines de mesures ont été proposées dont certaines ont eu plus d'impact que d'autres sur le monde scientifique, chaque mesure correspondant à un aspect de la biodiversité et à une façon de résumer cet aspect.

Certains ont essayé de mettre de l'ordre en regroupant plusieurs indices comme cas particuliers d'un indice plus général. Ce foisonnement dans la littérature scientifique est nécessaire et générateur d'idées. Nous avons vu qu'un indice développé à plusieurs reprises afin de répondre à chaque fois de façon adéquate à des problèmes spécifiques peut être replacé dans un schéma mathématique plus général que la somme des problèmes particuliers. Il apparaît aujourd'hui que malgré les barrières entre disciplines, nous creusons parfois les mêmes questions, ou des questions proches. Ce foisonnement de la littérature est donc encore aujourd'hui générateur d'idées nouvelles car il devient possible d'y aller chercher des outils dans une autre discipline.

D'un autre côté, ce foisonnement peut être réducteur car on se noie facilement dans un amoncellement de procédures correspondant à des problèmes très particuliers. Seuls des schémas fondamentaux peuvent alors nous permettre d'élargir notre regard sans nous perdre. Il faut des clés, des structures pour s'y retrouver.

Chapitre 3

Décomposition de la biodiversité

Sommaire

Résumé

Ce chapitre commence par une revue bibliographique des méthodes de décomposition de la variation, diversité ou variance, en génétique et écologie. De cette revue, nous montrons que l'analyse de variance moléculaire (AMOVA) développée en génétique sort du lot parce qu'elle est basée implicitement sur l'entropie quadratique. Cette méthode a eu un fort impact dans l'analyse de la variation en génétique. Or la même structure de données existe en écologie.

Une question se pose alors : existe-il un équivalent de l'AMOVA en écologie ?

Nous démontrons que l'AMOVA correspond en fait à une utilisation particulière de l'APQE, décomposition hiérarchique de l'entropie quadratique développée par Rao. De plus, L'APQE a pour cas particuliers l'ANOVA, analyse de la variance et la CATANOVA analyse de la variance sur variable catégorielle, ou décomposition de l'indice de Gini-Simpson. On obtient alors un cadre général des méthodes de décomposition de la variation qui peut s'appliquer à n'importe quelle discipline et en particulier à la génétique et à l'écologie. Nous établissons alors les liens entre toutes les méthodes présentées à la fois en génétique, en écologie et en statistique.

Ce chapitre se termine par la proposition d'une application possible de la décomposition de l'entropie quadratique permettant l'analyse du profil spatial de la diversité taxonomique dans une zone d'étude.

3.1 D´ ´ ´

En génétique, un grand intérêt est porté aux les structures des populations. Lorsque dans une population les individus s'apparient aléatoirement pour la reproduction, cette population est dite panmictique. Sa composition génétique suit des lois particulières régies par le modèle d'Hardy-Weinberg. Ce modèle suppose de plus que la population est de taille infinie, qu'il n'y a ni sélection, ni mutation, ni migration, et qu'il y a absence de croisement entre générations différentes. Si cette population n'est pas panmictique mais est formée de sous-populations panmictiques, toutes les lois basées sur le principe d'Hardy-Weinberg peuvent s'appliquer dans les sous-populations, mais pas dans la population entière. En effet, dans ce cas, il y a moins d'homozygotes, donc plus d'hétérozygotes, dans la population entière qu'en moyenne dans chaque sous-population, c'est-à-dire l'hétérozygotie de la population entière est plus grande. Cette réalité est appelée principe de Walhund et est défini relativement au modèle d'Hardy-Weinberg (Hattemer 1982). De cette façon, la structure de chaque sous-population est prise en compte dans chaque analyse génétique. L'hétérozygotie est une mesure de diversité. De plus elle est décomposable, concave, puisque l'hétérozygotie de la population totale est plus grande que l'hétérozygotie moyenne dans les sous-populations panmictiques. C'est une des mesures les plus répandues en génétique pour mesurer la diversité.

3.1.1 Les statistiques-F

Les statistiques-F, ou indices de fixation, ont été introduites par Wright (1951). Wright a d'abord défini un coefficient F de consanguinité. F est alors la probabilité que des allèles soient identiques dans une population parce qu'elles ont été héritées d'un même ancêtre commun. Supposons qu'une population est divisée en sous-populations. Wright définit alors trois composants : F_{IT} et F_{IS} sont les corrélations entre les deux gamètes qui s'unissent pour produire les individus relativement à la population totale pour F_{IT} et aux sous-populations pour F_{IS} ; et F_{ST} est la corrélation entre deux gamètes tirés aléatoirement dans chaque sous-population. Ces statistiques-F sont reliées par la formule

$$1 - F_{IT} = (1 - F_{IS})(1 - F_{ST}). \tag{3.1}$$

Nei et Wright ont par la suite modifié les définitions de ces statistiques-F (Nei 1987). Soient h_O le taux d'hétérozygotie moyen observé dans les sous-populations, h_S le taux d'hétérozygotie moyen attendu au sein des populations selon l'équilibre d'Hardy-Weinberg, c'est-à-dire sous panmixie, et h_T le taux d'hétérozygotie total attendu (dans toutes les populations mélangées) sous panmixie. Soient p_{ki} la fréquence de l'allèle k dans la sous-population i et $\bar{p_k}$ la fréquence moyenne de l'allèle k sur la totalité des sous-populations. Les hétérozygoties attendues h_S et h_T sont définies par

$$h_S = 1 - \frac{1}{r} \sum_i \sum_k p_{ki}^2$$

et

$$h_T = 1 - \sum_k \bar{p_k^2}.$$

Les statistiques-F sont alors calculées par les formules suivantes :

$$F_{IT} = (h_T - h_O)/h_T$$

$$\begin{aligned} F_{IS} &= (h_S - h_O)/h_S \\ F_{ST} &= (h_T - h_S)/h_T \end{aligned}$$

Dans le cas où ces valeurs sont estimées à partir de petits échantillons, des estimateurs non-biaisés de ces statistiques peuvent être trouvés dans Nei (1987, pages 164-165). Wright a également défini les indices P_{IT}, P_{IS} et P_{ST}, égaux respectivement à $1 - F_{IT}$, $1 - F_{IS}$ et $1 - F_{ST}$, P désignant la panmixie. Ces indices sont donc unis par la relation

$$P_{IT} = P_{IS} P_{ST}.$$

h_S et h_T sont deux utilisations de l'indice de Gini-Simpson :

$$h_S = 1 - \frac{1}{r} \sum_i \sum_k p_{ki}^2 = \frac{1}{r} \sum_{i=1}^{r} \left(1 - \sum_{k=1}^{S} p_{ki}^2 \right)$$

est la diversité moyenne au sein des sous-populations en pondérant les sous-populations uniformément, et

$$h_T = 1 - \sum_{k=1}^{S} \bar{p}_k^2$$

est la diversité totale dans le mélange des sous-populations. Ainsi, puisque l'indice de Gini-Simpson est concave, si h_T représente la diversité totale, h_S la diversité moyenne dans les sous-populations, alors $(h_T - h_S)$ est un composant de diversité allélique entre sous-populations. F_{ST} représente donc la part de la diversité totale due aux différences entre sous-populations, i.e. la diversité inter-sous-populations. Cependant, F_{IT} et F_{IS} ne sont pas des mesures de diversité totale et intra mais des mesures de l'écart entre la structure de la population observée et celle qu'elle devrait avoir selon le modèle d'Hardy-Weinberg.

F_{ST} est nécessairement positif et de plus compris entre 0 et 1. Au contraire, F_{IS} et F_{IT} peuvent être négatifs (Wright 1951). Les comportements de ces indices dans des structures extrêmes de populations sont assez inattendus. Par exemple, si toutes les sous-populations n'ont qu'un seul génotype et si ce génotype est hétérozygote, alors $h_O = 1$, et $h_S = 1 - \left((1/2)^2 + (1/2)^2 \right) = 1/2$ donc $F_{IS} = -1$.

La statistique F_{ST} a été introduite pour mesurer la consanguinité et estimer des paramètres décrivant la structure et la dynamique des populations tels que les flux de gènes (Neigel 2002). Des définitions de F_{ST} différentes de celle de Wright ont été développées par la suite ; elles portent des noms différents : G_{ST} (Nei 1973), θ (Weir et Cockerham 1984), N_{ST} (Lynch et Crease 1990), R_{ST} (Slatkin 1995), et Φ_{ST} (Excoffier *et al.* 1992). Si elles ne correspondent pas à des coefficients de consanguinité, ces nouvelles statistiques n'auront pas nécessairement, comme la statistique F_{ST}, des relations avec des paramètres tels que les flux de gènes (Neigel 2002). Cependant, dans le cadre de la décomposition de la diversité, ce ne sont pas ces pseudo-statistiques F qui vont directement nous intéresser, mais surtout les méthodes de décompositions de variance ou de diversité génétique qui ont été développées pour les estimer. Nous allons regarder ces méthodes en les regroupant selon l'indice de diversité qu'elles utilisent.

3.1.2 Décomposition de la diversité allélique

Lewontin (1972), Nei (1973) et Weir et Cockerham (1984) ont décomposé la diversité allélique de populations structurées hiérarchiquement, explicitement ou implicitement selon les indices de Shannon et de Gini-Simpson.

Lewontin (1972) propose une décomposition de l'indice de Shannon pour étudier la diversité génétique humaine. Trois niveaux sont considérés : espèce humaine $\supset$ race $\supset$ population. Posons r le nombre de races, m_i le nombre de populations dans la race i, m_+ le nombre total de populations, et $\mathbf{p}_{ji} = \left(p_{1ji}, ..., p_{Sji}\right)^t$, $\mathbf{p}_{\bullet i} = (p_{1\bullet i}, ..., p_{S\bullet i})^t$ et $\mathbf{p}_{\bullet\bullet} = (p_{1\bullet\bullet}, ..., p_{S\bullet\bullet})^t$ les vecteurs contenant les distributions de fréquences des allèles d'un locus respectivement dans la population j de la race i, en moyenne dans toutes les populations de la race i et en moyenne dans toutes les races de l'espèce entière. Lewontin impose de considérer que, dans la mesure de la diversité, les poids des populations d'une race soient uniformes, donc $\mathbf{p}_{\bullet i} = \sum_{j=1}^{m_i} \mathbf{p}_{ji}/m_i$, et que les poids des races soient fonctions du nombre de populations qu'elles contiennent, donc $\mathbf{p}_{\bullet\bullet} = \sum_{i=1}^{r} m_i \mathbf{p}_{\bullet i}/m_+$. Il définit les composants de diversité suivants (H_S désigne l'indice de Shannon) :

- $H_{O_{ji}} = H_S\left(\mathbf{p}_{ji}\right) = -\sum_{k=1}^{S} p_{kji} \ln\left(p_{kji}\right)$, la diversité au sein de la population j de la race i,
- $H_{pop_i} = \sum_{j=1}^{m_i} H_{O_{ji}}/m_i$, la diversité moyenne dans les populations de la race i,
- $H_{race_i} = H_S\left(\mathbf{p}_{\bullet i}\right) = -\sum_{k=1}^{S} p_{k\bullet i} \ln\left(p_{k\bullet i}\right)$, la diversité au sein de la race i.
- $H_{esp} = H_S\left(\mathbf{p}_{\bullet\bullet}\right) = -\sum_{k=1}^{S} p_{k\bullet\bullet} \ln\left(p_{k\bullet\bullet}\right)$, la diversité totale dans l'espèce humaine,
- $\bar{H}_{pop} = \sum_{i=1}^{r} m_i H_{pop_i}/m_+$, la diversité moyenne au sein des populations de l'ensemble des races,
- $\bar{H}_{race} = \sum_{i=1}^{r} m_i H_{race_i}/m_+$, la diversité moyenne au sein des races.

Lewontin introduit également la différence entre $\bar{H}_{race}$ et $\bar{H}_{pop}$ comme composant de diversité inter-populations intra-race et la différence entre H_{esp} et H_{race} comme le composant de diversité inter-races. Il montre ainsi que la diversité totale dans l'espèce peut être divisée additivement en trois composants :

$$\overbrace{H_{esp}}^{\text{Diversité dans l'espèce}} = \overbrace{\bar{H}_{pop}}^{\substack{\text{Diversité moyenne} \\ \text{dans les populations}}} + \overbrace{\left[\bar{H}_{race} - \bar{H}_{pop}\right]}^{\substack{\text{Diversité moyenne} \\ \text{entre les populations d'une race}}} + \overbrace{\left[H_{esp} - \bar{H}_{race}\right]}^{\text{Diversité entre races}}.$$

Lewontin démontre donc que l'indice de Shannon peut être décomposé à différentes échelles liées hiérarchiquement. Ce résultat fut analysé ensuite par un écologue (Allan 1975), puis popularisé en écologie par un généticien (Lande 1996).

Nei (1973) choisit d'utiliser l'indice de Gini-Simpson pour mesurer la diversité allélique ("gene diversity") dans les populations subdivisées. Soit une population divisée en r sous-populations, soient p_{ki} la fréquence de l'allèle k à un locus donné dans la sous-population i et $\mathbf{p}_i = (p_{1i}, ..., p_{Si})^t$ le vecteur contenant la distribution de fréquences des allèles dans la sous-population i. La diversité allélique de la sous-population i est égale à $H_i = H_{G\text{-}S}\left(\mathbf{p}_i\right) = 1 - \sum_k p_{ki}^2$. La diversité allélique entre la sous-population i et la sous-population i' est égale à $D_{ii'} = H_{ii'} + \left(H_i + H_{i'}\right)/2$, où $H_{ii'} = 1 - \sum_k p_{ki} p_{ki'}$, ce que nous pouvons aussi écrire par la

formule de Jensen (1906) appliquée à l'indice de Gini-Simpson :

$$D_{ii'} = 2H_{\text{G-S}}\left(\frac{\mathbf{p}_i + \mathbf{p}_{i'}}{2}\right) - H_{\text{G-S}}(\mathbf{p}_i) - H_{\text{G-S}}(\mathbf{p}_{i'}).$$

Il est intéressant de remarquer que

$$D_{ii'} = \sum_k \frac{(p_{ki} - p_{ki'})^2}{2}.$$

$D_{ii'}$ est appelée "distance minimale" par Nei (Hattemer 1982, Finkeldey 1994).

La diversité allélique dans la population totale est égale à $H_T = H_{\text{G-S}}(\mathbf{p}_\bullet) = 1 - \sum_k \bar{p}_k^2$, où $\bar{p}_k = \sum_i w_i p_{ki}$ avec w_i poids de la sous-population i ($\sum_i w_i = 1$), et $\mathbf{p}_\bullet = (\bar{p}_1, ..., \bar{p}_S)^t$. Nei (1973), comme Lewontin (1972), choisit une pondération uniforme des populations : $w_i = 1/r$. La diversité allélique totale se décompose de façon additive en une diversité allélique moyenne au sein des sous-populations $H_S = \sum_i H_i/r$, et une diversité allélique entre paires de sous-populations $D_{ST} = \sum_i \sum_{i'} D_{ii'}/r^2$. Ainsi

$$H_T = H_S + D_{ST}. \tag{3.2}$$

Pour avoir une bonne image de la différenciation allélique entre les populations, plusieurs loci doivent être étudiés. Dans ce cas, les composants de diversité, H_T, H_S et D_{ST} sont alors définis comme les moyennes de ceux obtenus pour chaque locus. La relation $H_T = H_S + D_{ST}$ reste valable.

D'autres niveaux hiérarchiques peuvent être rajoutés à cette décomposition. Par exemple, si les sous-populations sont divisées en colonies, alors la décomposition suivante de la diversité allélique peut être utilisée :

$$H_T = H_C + D_{CS} + D_{ST}.$$

A partir de la formule 3.2, le composant

$$G_{ST} = D_{ST}/H_T$$

est introduit pour mesurer la différenciation génique entre sous-populations. Soient $J_T = 1 - H_T$ et $J_S = 1 - H_S$,

$$(1 - G_{ST})(1 - J_T) = (1 - J_S)$$

(Nei 1973). La relation entre ces composants est différente de celle qui existe entre les trois composants de Wright (formule 3.1). Le composant G_{ST} est équivalent au F_{ST} de Wright alors que les composants J_S et J_T sont différents des statistiques F_{IS} et F_{IT}. Les valeurs de J_S et J_T sont comprises entre 0 et 1 et mesurent la probabilité que deux allèles tirés au hasard, dans les sous-populations pour J_S et dans la population entière pour J_T, soient identiques. Les valeurs de F_{IS} et F_{IT} peuvent être négatives et représentent la déviation de l'organisation des sous-populations par rapport à l'organisation qu'elles auraient sous le modèle d'Hardy-Weinberg.

Le composant D_{ST} inclut la comparaison des sites avec eux-mêmes. Nei (1973), estimant préférable d'éliminer cette comparaison propose l'indice

$$\tilde{D}_m = \sum_{i \neq i'} D_{ii'}/[r(r-1)].$$

La relation de décomposition de la diversité devient

$$H'_T = H_S + \tilde{D}_m.$$

Nei (1973) définit alors la quantité relative de différenciation allélique entre sous-populations

$$G'_{ST} = \frac{\tilde{D}_m}{H'_T},$$

et le rapport entre la diversité inter-populations et la diversité intra-populations :

$$R_{ST} = \frac{\tilde{D}_m}{H'_S}.$$

Cette méthode s'adresse à des marqueurs génotypiques permettant l'accès aux allèles. Les individus étudiés peuvent être haploïdes ou diploïdes. Dans le cas où le nombre d'individus étudiés est supérieur à 50, les fréquences observées des allèles dans les sous-populations échantillonnées peuvent être utilisées. Dans le cas contraire, il est nécessaire d'utiliser des estimateurs non biaisés (Nei 1987, Hudson *et al.* 1992, Kremer *et al.* 1998).

Weir et Cockerham (1984) proposent d'étudier des populations d'individus diploïdes. Le marqueur concerné est la présence ou l'absence d'un allèle donné. Pour l'instant, un seul allèle est donc considéré. Weir inscrit cette analyse dans le cadre d'une application de l'ANOVA hiérarchique à ce type de données. Quatre niveaux hiérarchiques sont considérés : populations ⊃ dèmes ⊃ individus ⊃ gènes. Les notations sont les suivantes :

- r, le nombre de populations ;
- m_i, le nombre de dèmes dans la population i ;
- $m_+ = \sum_{i=1}^{r} m_i$, le nombre total de dèmes ;
- n_{ji}, le nombre d'individus dans le dème j de la population i ;
- $n_{+i} = \sum_{j=1}^{m_i} n_{ji}$, le nombre d'individus dans la population i ;
- $n_{++} = \sum_{i=1}^{r} n_{+i}$, le nombre total d'individus ;
- $\tilde{p}_{ji}$ et $\tilde{h}_{ji}$, les fréquences observées respectivement de l'allèle considéré et des hétérozygotes pour cet allèle dans le dème j de la population i ;
- $\tilde{p}_{\bullet i} = \sum_{j=1}^{m_i} n_{ji}\tilde{p}_{ji}/n_{+i}$, la fréquence moyenne observée de l'allèle considéré dans la population i ;
- $\tilde{p}_{\bullet\bullet} = \sum_{i=1}^{r}\sum_{j=1}^{m_i} n_{ji}\tilde{p}_{ji}/n_{++}$, la fréquence globale observée de l'allèle, c'est-à-dire la fréquence de cet allèle dans l'ensemble des populations mélangées.

Avec ces notations, par similitude avec l'ANOVA, Weir et Cockerham (1984) définissent les carrés moyens ("MS" "mean squares") suivants :

$$MS\,P = 2\sum_{i=1}^{r} n_{+i}\left(\tilde{p}_{\bullet i} - \tilde{p}_{\bullet\bullet}\right)^2 / (r-1)$$

$$MS\,D = 2\sum_{i=1}^{r}\sum_{j=1}^{m_i} n_{ji}\left(\tilde{p}_{ji} - \tilde{p}_{\bullet i}\right)^2 / (m_+ - r)$$

$$MSI = \left[2 \sum_{i=1}^{r} \sum_{j=1}^{m_i} n_{ji} \tilde{p}_{ji} \left(1 - \tilde{p}_{ji}\right) - \frac{1}{2} \sum_{i=1}^{r} \sum_{j=1}^{m_i} n_{ji} \tilde{h}_{ji} \right] / \left(n_{++} - m_{+}\right)$$

$$MSG = \sum_{i=1}^{r} \sum_{j=1}^{m_i} n_{ji} \tilde{h}_{ji} / 2n_{++}$$

A partir de ces carrés moyens, ils calculent des composants de variances : a pour les populations, b_2 pour les dèmes dans les populations, b_1 pour les individus dans les dèmes, et c pour les gamètes au sein des individus :

$$\begin{aligned} c &= MSG, \\ b_1 &= \frac{1}{2}\left(MSI - MSG\right), \\ b_2 &= \frac{1}{2n_1}\left(MSD - MSI\right), \\ a &= \frac{1}{2n_1 n_3}\left[n_1 MSP - n_2 MSD - \left(n_1 - n_2\right) MSI\right], \end{aligned}$$

où les valeurs

$$\begin{aligned} n_1 &= \frac{1}{m_{+} - r}\left(n_{++} - \sum_{i=1}^{r} \sum_{j=1}^{m_i} \frac{n_{ji}^2}{n_{+i}}\right), \\ n_2 &= \frac{1}{r-1}\left(\sum_{i=1}^{r} \sum_{j=1}^{m_i} \frac{\left(n_{++} - n_{+i}\right) n_{ji}^2}{n_{+i} n_{++}}\right), \\ n_3 &= \frac{1}{r-1}\left(n_{++} - \sum_{i=1}^{r} \frac{n_{+i}^2}{n_{++}}\right). \end{aligned}$$

permettent d'appréhender des populations dont les tailles (en nombre d'individus) diffèrent. Ces composants de variances, a, b_2, b_1, et c, conduisent à l'estimation de trois paramètres apparentés aux statistiques-F de Wright : F corrélation de gènes au sein des individus, θ_1 corrélation de gènes entre individus dans les dèmes, θ_2 corrélation de gènes entre dèmes dans les populations. Les estimations $\hat{F}$, $\hat{\theta_1}$ et $\hat{\theta_2}$ sont obtenues par

$$\begin{aligned} 1 - \hat{F} &= \frac{c}{a + b_2 + b_1 + c}, \\ \hat{\theta_1} &= \frac{a + b_2}{a + b_2 + b_1 + c}, \\ \hat{\theta_2} &= \frac{a}{a + b_2 + b_1 + c}. \end{aligned}$$

Nous avons montré comment retrouver le lien entre cette analyse et l'ANOVA. Pour cela, il nous faut considérer une variable Y_{lkji} telle que

$$Y_{lkji} = \begin{cases} 1 & \text{si l'allèle est présent sur le chromosome } l \text{ de l'individu } k \\ & \text{appartenant au dème } j \text{ de la population } i, \\ 0 & \text{sinon.} \end{cases}$$

Soit y_{lkji} la valeur observée prise par Y_{lkji} pour le chromosome l de l'individu k du dème j du groupe i. Soient $y_{\bullet kji}$, $y_{\bullet\bullet ji}$, $y_{\bullet\bullet\bullet i}$ et $y_{\bullet\bullet\bullet\bullet}$, les valeurs *moyennes* de y respectivement pour l'individu k du dème j du groupe i, dans le dème j du groupe i, dans le groupe i et dans la totalité des groupes mélangés. Les fréquences précédentes peuvent être exprimées en fonction de ces valeurs moyennes : Si l'individu k est homozygote pour l'absence de l'allèle, alors

$$\sum_{l=1}^{2}\left(y_{lkji}-y_{\bullet kji}\right)^2=(0-0)^2+(0-0)^2=0.$$

S'il est homozygote pour la présence de l'allèle,

$$\sum_{l=1}^{2}\left(y_{lkji}-y_{\bullet kji}\right)^2=(1-1)^2+(1-1)^2=0$$

S'il est hétérozygote,

$$\sum_{l=1}^{2}\left(y_{lkji}-y_{\bullet kji}\right)^2=(1-1/2)^2+(0-1/2)^2=1/2$$

Ainsi $\tilde{h}_{ji}=\left(2/n_{ji}\right)\sum_{k=1}^{n_{ji}}\sum_{l=1}^{2}\left(y_{lkji}-y_{\bullet kji}\right)^2$ et de plus

$$\tilde{p}_{ji}=y_{\bullet\bullet ji},\ \tilde{p}_{\bullet i}=y_{\bullet\bullet\bullet i},\ \tilde{p}_{\bullet\bullet}=y_{\bullet\bullet\bullet\bullet}.$$

Avec ces nouvelles notations, les composants de variance de Weir et Cockerham peuvent être écrits sous la forme des carrés moyens de l'ANOVA (Weir 1996) :

$$\begin{aligned}
MSP &= \sum_{i=1}^{r}2n_{+i}\left(y_{\bullet\bullet\bullet i}-y_{\bullet\bullet\bullet\bullet}\right)^2/(r-1)\\
MSD &= \sum_{i=1}^{r}\sum_{j=1}^{m_j}2n_{ji}\left(y_{\bullet\bullet ji}-y_{\bullet\bullet\bullet i}\right)^2/(m_+-r)\\
MSI &= \sum_{i=1}^{r}\sum_{j=1}^{m_j}\sum_{k=1}^{n_{ji}}2\left(y_{\bullet kji}-y_{\bullet\bullet ji}\right)^2/(n_{++}-m_+)\\
MSG &= \sum_{i=1}^{r}\sum_{j=1}^{m_i}\sum_{k=1}^{n_{ji}}\sum_{l=1}^{2}\left(y_{lkji}-y_{\bullet kji}\right)/n_{++}.
\end{aligned}$$

Démonstrations : Les démonstrations pour MSP, MSD et MSG sont immédiates par le changement de notation. Pour MSI, le changement de notation implique que

$$\begin{aligned}
MSI &= \left[2\sum_{ij}n_{ji}y_{\bullet\bullet ji}\left(1-y_{\bullet\bullet ji}\right)-\frac{1}{2}\sum_{ij}n_{ji}\frac{2}{n_{ji}}\sum_{kl}\left(y_{lkji}-y_{\bullet kji}\right)^2\right]/(n_{++}-m_+)\\
&= \left[2\sum_{ij}n_{ji}y_{\bullet\bullet ji}-2\sum_{ij}n_{ji}y_{\bullet\bullet ji}^2-2\sum_{ijkl}y_{lkji}^2+2\sum_{ijk}y_{\bullet kji}^2\right]/(n_{++}-m_+)
\end{aligned}$$

Or $y_{lkji}^2 = y_{lkji}$ et $2\sum_{ij} n_{ji} y_{\bullet\bullet ji} = 2\sum_{ijkl} y_{lkji}$ donc

$$\begin{aligned} MSI &= \left[2\sum_{ijk} y_{\bullet kji}^2 - 2\sum_{ij} n_{ji} y_{\bullet\bullet ji}^2\right] / (n_{++} - m_+) \\ &= \sum_{i=1}^{r}\sum_{j=1}^{m_j}\sum_{k=1}^{n_{ji}} 2\left(y_{\bullet kji} - y_{\bullet\bullet ji}\right)^2 / (n_{++} - m_+) \end{aligned}$$

Nous allons démontrer maintenant que l'analyse de Weir et Cockerham s'inscrit à la fois dans la structure de l'ANOVA et aussi dans le schéma de la décomposition additive de l'indice de Gini-Simpson. Dans le schéma de l'ANOVA, la variable Y_{lkji} est considérée comme une variable quantitative. En fait, comme elle est binaire, elle peut aussi être considérée comme une variable catégorielle, qualitative, où la catégorie 1 est la présence de l'allèle et la catégorie 0 l'absence de cet allèle. Notons A et $\bar{A}$ ces deux catégories, montrons alors que l'analyse de Weir est aussi une décomposition de l'indice de Gini-Simpson. Soit $\mathbf{p}_{kij} = \left(p_{Akji}, p_{\bar{A}kji}\right)^t$ le vecteur des fréquences des deux allèles dans l'individu k du dème j de la population i. D'après l'indice de Gini-Simpson la diversité associée à cette distribution de fréquences est

$$H_{\text{G-S}}\left(\mathbf{p}_{kji}\right) = 1 - p_{Akji}^2 - p_{\bar{A}kji}^2 = 2p_{Akji}p_{\bar{A}kji}.$$

Si l'individu k est homozygote alors $H_{\text{G-S}}\left(\mathbf{p}_{kji}\right) = 0$. S'il est hétérozygote $H_{\text{G-S}}\left(\mathbf{p}_{kji}\right) = 1/2$. Notons $\mathbf{p}_{\bullet ij} = \left(p_{A\bullet ji}, p_{\bar{A}\bullet ji}\right)^t$, $\mathbf{p}_{\bullet\bullet i} = (p_{A\bullet\bullet i}, p_{\bar{A}\bullet\bullet i})^t$ et $\mathbf{p}_{\bullet\bullet\bullet} = (p_{A\bullet\bullet\bullet}, p_{\bar{A}\bullet\bullet\bullet})^t$, les vecteurs des distributions moyennes des deux allèles respectivement dans le dème j de la population i, la population i et dans la totalité des populations mélangées. Les fréquences de l'analyse de Weir et Cockerham peuvent être réécrites :

$$\tilde{h}_{ji} = \left(2/n_{ji}\right)\sum_{k=1}^{n_{ji}} H_{\text{G-S}}\left(\mathbf{p}_{kji}\right),\ \tilde{p}_{ji} = p_{A\bullet ji},\ \tilde{p}_{\bullet i} = p_{A\bullet\bullet i},\ \tilde{p}_{\bullet\bullet} = p_{A\bullet\bullet\bullet}.$$

Les composants de variance de Weir et Cockerham peuvent alors être écrits en fonction des composants de la décomposition hiérarchique de l'indice de Gini-Simpson :

$$\begin{aligned} MSP &= n_{++}\left[H_{\text{G-S}}\left(\mathbf{p}_{\bullet\bullet\bullet}\right) - \sum_{i=1}^{r}\lambda_i H_{\text{G-S}}\left(\mathbf{p}_{\bullet\bullet i}\right)\right] / (r-1) \\ MSD &= n_{++}\left[\sum_{i=1}^{r}\lambda_i H_{\text{G-S}}\left(\mathbf{p}_{\bullet\bullet i}\right) - \sum_{i=1}^{r}\lambda_i\sum_{j=1}^{m_i}\mu_{ji} H_{\text{G-S}}\left(\mathbf{p}_{\bullet ji}\right)\right] / (m_+ - r) \\ MSI &= n_{++}\left[\sum_{i=1}^{r}\lambda_i\sum_{j=1}^{m_i}\mu_{ji} H_{\text{G-S}}\left(\mathbf{p}_{\bullet ji}\right) - \sum_{i=1}^{r}\lambda_i\sum_{j=1}^{m_i}\mu_{ji}\sum_{k=1}^{n_{ji}} w_{kji} H_{\text{G-S}}\left(\mathbf{p}_{kji}\right)\right] / (n_{++} - m_+) \\ MSG &= n_{++}\left[\sum_{i=1}^{r}\lambda_i\sum_{j=1}^{m_i}\mu_{ji}\sum_{k=1}^{n_{ji}} w_{kji} H_{\text{G-S}}\left(\mathbf{p}_{kji}\right)\right] / n_{++}, \end{aligned}$$

où les paramètres $\lambda_i = n_{+i}/n_{++}$, $\mu_{ji} = n_{ji}/n_{+i}$, et $w_{kji} = 1/n_{ji}$ désignent les poids relatifs attribués respectivement aux populations, dèmes et individus. Ces poids sont exprimés en tailles relatives par rapport au nombre de copies d'ADN étudiées. Par exemple, le poids d'un individu dans le dème j de la population i est de deux copies sur $2n_{ji}$ soit $1/n_{ji}$.

Démonstration : Avec les dernières notations,

$$\begin{aligned} MSP &= 2\sum_{i=1}^{r} n_{+i}\left(p_{A\bullet\bullet i} - p_{A\bullet\bullet\bullet}\right)^2/(r-1) \\ &= n_{++}\left[\sum_{i=1}^{r}\frac{n_{+i}}{n_{++}}2p_{A\bullet\bullet i}^2 - 2p_{A\bullet\bullet\bullet}^2\right]/(r-1) \end{aligned}$$

Sachant que $p_{A\bullet\bullet i}^2 = p_{A\bullet\bullet i} - p_{A\bullet\bullet i}p_{\bar{A}\bullet\bullet i}$ et $p_{A\bullet\bullet\bullet}^2 = p_{A\bullet\bullet\bullet} - p_{A\bullet\bullet\bullet}p_{\bar{A}\bullet\bullet\bullet}$ et que $\sum_{i=1}^{r}\frac{n_{+i}}{n_{++}}2p_{A\bullet\bullet i} = 2p_{A\bullet\bullet\bullet}$, l'équation devient

$$\begin{aligned} MSP &= n_{++}\left[\sum_{i=1}^{r} -\lambda_i 2p_{A\bullet\bullet i}p_{\bar{A}\bullet\bullet i} + 2p_{A\bullet\bullet\bullet}p_{\bar{A}\bullet\bullet\bullet}\right]/(r-1) \\ &= n_{++}\left[H_{\text{G-S}}(\mathbf{p}_{\bullet\bullet\bullet}) - \sum_{i=1}^{r}\lambda_i H_{\text{G-S}}(\mathbf{p}_{\bullet\bullet i})\right]/(r-1) \end{aligned}$$

Le même raisonnement est utilisé pour MSD et les démonstrations des réécritures de MSI et MSG sont immédiates après changement de notations.

En conclusion, nous avons démontré que l'analyse de variance de Weir et Cockerham peut être écrite aussi bien dans le schéma de l'ANOVA que dans celui de la décomposition de l'indice de Gini-Simpson. Ce deuxième schéma présente plus d'intérêts parce qu'il nous permet de mettre l'analyse de Weir et Cockerham sur le même plan que celle de Nei et en conséquence de montrer que l'analyse de Weir et Cockerham reste valable pour une étude multi-allélique. Combinons les raisonnements de Nei (1973) (analyse multi-alléliques) et de Weir et Cockerham (1984) (considération pour des individus diploïdes du niveau intra-individu), nous mettons ainsi en évidence que la décomposition de l'indice de Gini-Simpson permet d'obtenir une analyse multiallélique tenant compte du niveau intra-individu.

3.1.3 L'Analyse de Variance Moléculaire, AMOVA

Nei *et al.* (Nei et Li 1979, Nei et Tajima 1981) proposent de mesurer la diversité nucléotidique par la formule qui sera généralisée par la suite et appelée entropie quadratique (Rao 1982a, cf. partie 2.4)

$$\hat{v}_i = \frac{2\sum_{x<y} n_{ix}n_{iy}\hat{d}_{xy}}{n_i(n_i-1)},$$

où n_i est le nombre de séquences analysées dans la population i, n_{ix} le nombre observé d'haplotypes x dans cette même population, et $\hat{d}_{xy}$ une estimation du nombre moyen de substitutions par site nucléotidique entre les haplotypes x et y. Lynch et Crease (1990) définissent à partir de

$\hat{v}_i$ la diversité moyenne dans les populations :

$$\hat{v}_w = \frac{\sum_{i=1}^{r} \hat{v}_i}{r}.$$

Selon Nei et Li (1979), le nombre moyen de substitutions par site nucléotidique entre les populations i et j peut être estimé par

$$\hat{v}_{ij} = \hat{v}'_{ij} - \frac{\hat{v}_i + \hat{v}_j}{2},$$

où

$$\hat{v}'_{ij} = \sum_{x,y} \frac{n_{ix}}{n_i} \frac{n_{jy}}{n_j} \hat{d}_{xy}.$$

Lynch et Crease (1990) définissent ensuite une mesure de variation inter-populations :

$$\hat{v}_b = \frac{2 \sum_{i<j} \hat{v}_{ij}}{r\,(r-1)}.$$

Ils présentent alors

$$N_{ST} = \frac{\hat{v}_b}{\hat{v}_w + \hat{v}_b},$$

où la somme de $\hat{v}_w$ et $\hat{v}_b$ est une estimation de la diversité nucléotidique totale, toutes populations mélangées, comme étant analogue à la statistique F_{ST} de Wright (1951). Lynch et Crease proposent ensuite des estimations des variances des différents composants en supposant qu'à la fois les abondances et les dissimilarités entre haplotypes sont issues d'un échantillonnage. Ils affirment que les procédures de ces calculs peuvent être étendues à une analyse hiérarchique de la structure de populations, par exemple en considérant la variation dans des dèmes, entre dèmes dans des sites et entre sites, et d'autre part que des procédures de rééchantillonnage pourraient être utiles dans le développement de tests statistiques explicites. Ces conclusions ont trouvé écho dans l'analyse de variance moléculaire d'Excoffier *et al.* (1992).

Excoffier *et al.* (1992) introduisent une méthode générale, appelée analyse de variance moléculaire (AMOVA, "Analysis of MOlecular VAriance"), qui est susceptible d'inclure toute sorte de diversité génétique (Excoffier 1994, 2001). Pour présenter l'AMOVA, ils ont considéré trois niveaux : groupes, populations, individus. Prenons les notations suivantes :

- G, le nombre de groupes,
- I_g, le nombre de populations dans le groupe g,
- n_{ig}, le nombre d'individus dans la population i du groupe g,
- n_{+g}, le nombre total d'individus dans le groupe g,
- n_{++}, le nombre total d'individus,
- $\delta_{jj'}$, une mesure de différence entre les haplotypes des individus j et j'.

La somme totale des carrés des déviations (SSD, "sum of squared deviation") est égale à

$$SSD(Total) = \frac{1}{2n_{++}} \sum_{j=1}^{n_{++}} \sum_{j'=1}^{n_{++}} \left(\delta_{jj'}^{\text{ind}}\right)^2$$

Cette mesure est divisée en un composant de variations entre individus au sein des populations ($SSD(WP)$), un composant de variations entre populations au sein des groupes ($SSD(AP/WG)$), et un composant de variations entre groupes ($SSD(AG)$) :

$$SSD(WP) = \sum_{g=1}^{G} \sum_{i=1}^{I_g} \frac{\sum_{j=1}^{n_{ig}} \sum_{j'=1}^{n_{ig}} \left(\delta_{jj'}^{\text{ind}}\right)^2}{2n_{ig}}$$

$$SSD(AP/WG) = \sum_{g=1}^{G} \left(\frac{\sum_{i=1}^{I_g} \sum_{j=1}^{n_{ig}} \sum_{i'=1}^{I_g} \sum_{j'=1}^{n_{i'g}} \left(\delta_{jj'}^{\text{ind}}\right)^2}{2n_{+g}} - \sum_{i=1}^{I_g} \frac{\sum_{j=1}^{n_{ig}} \sum_{j'=1}^{n_{ig}} \left(\delta_{jj'}^{\text{ind}}\right)^2}{2n_{ig}} \right)$$

$$SSD(AG) = \frac{\sum_{j=1}^{n_{++}} \sum_{j'=1}^{n_{++}} \left(\delta_{jj'}^{\text{ind}}\right)^2}{2n_{++}} - \sum_{g=1}^{G} \frac{\sum_{i=1}^{I_g} \sum_{j=1}^{n_{ig}} \sum_{i'=1}^{I_g} \sum_{j'=1}^{n_{i'g}} \left(\delta_{jj'}^{\text{ind}}\right)^2}{2n_{+g}}$$

La décomposition est additive :

$$SSD(Total) = SSD(WP) + SSD(AP/WG) + SSD(AG)$$

Par similitude avec l'ANOVA, Excoffier *et al.* (1992) définissent des degrés de liberté (df, "degrees of freedom"), carrés moyens (MS, "mean squares"), quotient des sommes des carrés des déviations par leurs degrés de liberté, et des composants de variances obtenus à partir de l'espérance des carrés moyens (Tab. 4). Les constantes intervenant dans les expressions des espérances des carrés moyens sont estimées en considérant que les effectifs des populations peuvent être inégaux.

TAB. *4 – Schéma de l'analyse moléculaire de variance.*

Source	df	SSD	MS	E(MS)
Inter-groupes	$r-1$	$SSD(AG)$	$SSD(AG)/(r-1)$	$\sigma^2+n_2^{(*)}\sigma_b^2+n_3^{(*)}\sigma_a^2$
Inter-populations intra-groupe	m_+-r	$SSD(AP/WG)$	$SSD(AP/WG)/(m_+-r)$	$\sigma^2+n_1^{(*)}\sigma_b^2$
Intra-population	$n_{++}-m_+$	$SSD(WP)$	$SSD(WP)/(n_{++}-m_+)$	σ^2
Totale	$n_{++}-1$	$SSD(Totale)$		

(*) $n_1=\frac{1}{m_+-r}\left(n_{++}-\sum_{g=1}^{G}\sum_{i=1}^{I_g}\frac{n_{ig}^2}{n_{+g}}\right)$; $n_2 = \frac{1}{r-1}\left(\sum_{g=1}^{G}\sum_{i=1}^{I_g}\frac{(n_{++}-n_{+g})n_{ig}^2}{n_{+g}n_{++}}\right)$; $n_3 = \frac{1}{r-1}\left(n_{++} - \sum_{g=1}^{G}\frac{n_{+g}^2}{n_{++}}\right)$

Les composants de variance σ_a^2, σ_b^2 et σ^2 sont utilisés pour l'estimation de statistiques Φ analogues aux statistiques F de Wright :

$$\Phi_{ST} = \frac{\sigma_a^2+\sigma_b^2}{\sigma_T^2}\ ,\ \Phi_{SC} = \frac{\sigma_b^2}{\sigma_b^2+\sigma^2},\ \Phi_{CT} = \frac{\sigma_a^2}{\sigma_T^2}$$

Ces trois composants sont liés par la même relation que les statistiques F de Wright :

$$(1-\Phi_{ST}) = (1-\Phi_{SC})\,(1-\Phi_{CT})\,.$$

Φ_{ST} est vu comme le rapport entre la corrélation entre haplotypes tirés aléatoirement dans les populations et la corrélation entre deux haplotypes tirés aléatoirement dans l'ensemble des haplotypes sans considération de la population ni de la région d'appartenance ; Φ_{SC} comme la corrélation entre deux haplotypes tirés aléatoirement au sein des populations relativement à la corrélation entre deux haplotypes tirés aléatoirement dans la région ; Φ_{CT} comme la corrélation entre deux haplotypes tirés aléatoirement au sein des régions relativement à la corrélation entre deux haplotypes tirés aléatoirement dans l'ensemble des haplotypes sans considération de la population ni de la région d'appartenance.

Cette analyse de variance moléculaire permet de tester l'effet de chaque niveau hiérarchique sur la variation. Les hypothèses nulles à tester sont l'absence de variation entre groupes (HoA : $\sigma_a^2 = 0$), l'absence de variation entre populations dans les groupes (HoB : $\sigma_b^2 = 0$), et l'absence de variation au sein des populations (HoC : $\sigma^2 = 0$). Pour obtenir les distributions des composants de variance sous chaque hypothèse nulle, Excoffier *et al.* (1992) utilisent des schémas de permutations : permutations des populations entre les groupes (test de HoA) ; permutations des haplotypes entre les populations, mais au sein des groupes (test de HoB) ; permutations des haplotypes entre les populations, et entre les groupes (test de HoC).

L'AMOVA est disponible dans le logiciel ARLEQUIN (Schneider *et al.* 2000). Dans la "FAQ list" de ce logiciel il peut être lu la question suivante :

> Why do I get negative variance components in AMOVA ?

La réponse donnée par l'équipe d'ARLEQUIN est :

> Negative variance components can sometimes occur, because they are rather covariances. Their associated fixation indices can also be seen as correlation coefficients. Usually, slightly negative variance components can occur in absence of genetic structure, because the true value of the parameter you want to estimate is zero. Thus, if the expectation of the estimator is zero, you can have, by chance, slightly positive or slightly negative variance components. Most of the time, these negative variance components indicate an absence of genetic structure. They can have a biological meaning, though. For instance, in outcrossing organisms, genes from different populations can be more related to each other than genes from the same population.

Essayons d'apporter quelques compléments à cette réponse. Si la vraie valeur de variance est positive et proche de 0, les estimateurs des composants de variance n'étant pas contraint à être positifs, l'estimation peut être négative et proche de 0. Si la vraie valeur est effectivement négative, alors dans ce cas la pertinence des composants de variance est mise en cause d'où l'affirmation "they are rather covariances".

Le modèle classique d'une analyse de variance sur plan hiérarchique est le suivant : $Y_{jig} = \mu + A_g + B_{ig} + \varepsilon_{jig}$. Toutes les variables Y_{jig} suivent la même loi normale de moyenne μ et de variance σ_Y^2. Toutes les variables aléatoires, A_g représentant l'effet groupe, B_{ig} représentant l'effet collection à l'intérieur des groupes, et ε_{jig} variables résiduelles dues à l'échantillonnage des individus dans les localités à l'intérieur des régions, sont supposées normales, indépendantes les unes des autres, de moyenne nulle et de variances respectives σ_A^2, σ_B^2 et σ^2 de sorte que $\sigma_Y^2 = \sigma_A^2 + \sigma_B^2 + \sigma^2$. Notons que dans ce schéma, les facteurs groupes et collections sont

considérés comme aléatoires c'est-à-dire résultant d'un échantillonnage. Excoffier *et al.* (1992) proposent de considérer le même raisonnement sur des vecteurs : $\mathbf{y}_{jig} = \mu + \mathbf{a}_g + \mathbf{b}_{ig} + \varepsilon_{jig}$ où $\mathbf{y}_{jig}$ est un vecteur de longueur N contenant les valeurs prises par N variables pour l'haplotype de l'individu j de la population i du groupe g. La distance entre les haplotypes des individus j et j' est définie par

$$\delta^2_{jj'} = \left(\mathbf{y}_{jig} - \mathbf{y}_{j'i'g'}\right)^t \mathbf{W} \left(\mathbf{y}_{jig} - \mathbf{y}_{j'i'g'}\right). \tag{3.3}$$

Lorsque $\mathbf{W} = \mathbf{I}_N$, où $\mathbf{I}_N$ est la matrice d'identité de dimensions $N \times N$, l'AMOVA correspond à la somme des ANOVA appliquées à chaque variable. Les composants de variance sont donc les sommes des variances des variables indépendantes ; ils sont donc toujours positifs. Dans Excoffier *et al.* (1992), il est simplement signalé que $\mathbf{W}$ peut être quelconque. Signalons néanmoins que pour assurer la positivité de $\delta^2_{jj'}$ il suffit que $\mathbf{W}$ soit définie positive, c'est-à-dire pour tout vecteur $\mathbf{a}$, $\mathbf{a}^t\mathbf{W}\mathbf{a} \geq 0$. Si les variables considérées sont quantitatives, un choix intéressant pour $\mathbf{W}$ est la matrice inverse de variance/covariance, la métrique est alors celle de Mahalanobis. Lorsque $\mathbf{W}$ est une matrice définie positive quelconque, mais constante (définie indépendamment des variables Y), alors les composants de variances sont des combinaisons de variances et covariances. Ils peuvent donc être négatifs.

Démonstration :

Soit la variable

- $y^{[k]}_{jig} = \mu^{[k]} + a^{[k]}_g + b^{[k]}_{ig} + \varepsilon^{[k]}_{jig}$

et ses moyennes

- $y^{[k]}_{\bullet ig} = \frac{1}{n_{ig}} \sum_{j=1}^{n_{ig}} y^{[k]}_{jig} = \mu^{[k]} + a^{[k]}_g + b^{[k]}_{ig} + \varepsilon^{[k]}_{\bullet ig}$
- $y^{[k]}_{\bullet\bullet g} = \frac{n_{ig}}{n_{+g}} \sum_{i=1}^{I_g} y^{[k]}_{\bullet ig} = \mu^{[k]} + a^{[k]}_g + b^{[k]}_{\bullet g} + \varepsilon^{[k]}_{\bullet\bullet g}$
- $y^{[k]}_{\bullet\bullet\bullet} = \frac{n_{+g}}{n_{++}} \sum_{g=1}^{G} y^{[k]}_{\bullet\bullet g} = \mu^{[k]} + a^{[k]}_\bullet + b^{[k]}_{\bullet\bullet} + \varepsilon^{[k]}_{\bullet\bullet\bullet}$

Tous les effets a, b et ε sont d'espérance nulle. Notons ${\sigma^{[k]}_a}^2$ la variance de $a^{[k]}_g$, et $\sigma^{[k][l]}_a$ la covariance entre $a^{[k]}_g$ et $a^{[l]}_g$. Notons de la même façon, ${\sigma^{[k]}_b}^2$ la variance de $b^{[k]}_{ig}$, et $\sigma^{[k][l]}_b$ la covariance entre $b^{[k]}_{ig}$ et $b^{[l]}_{ig}$; et ${\sigma^{[k]}}^2$ la variance de $\varepsilon^{[k]}_{jig}$, et $\sigma^{[k][l]}$ la covariance entre $\varepsilon^{[k]}_{jig}$ et $\varepsilon^{[l]}_{jig}$.

$$\begin{aligned}
E\left(MS\left(WP\right)\right) &= E\left(\frac{1}{n_{++}-m_+}\sum_{g=1}^{G}\sum_{i=1}^{I_g}\sum_{j=1}^{n_{ig}}\sum_{j'=1}^{n_{ig}}\frac{\left(\mathbf{y}_{jig}-\mathbf{y}_{j'ig}\right)^t\mathbf{W}\left(\mathbf{y}_{jig}-\mathbf{y}_{j'ig}\right)}{2n_{ig}}\right)\\
&= E\left(\frac{1}{n_{++}-m_+}\sum_{g=1}^{G}\sum_{i=1}^{I_g}\sum_{j=1}^{n_{ig}}\sum_{j'=1}^{n_{ig}}\sum_{k=1}^{N}\sum_{l=1}^{N}\frac{w_{kl}\left(y^{[k]}_{jig}-y^{[k]}_{j'ig}\right)\left(y^{[l]}_{jig}-y^{[l]}_{j'ig}\right)}{2n_{ig}}\right)\\
&= E\left(\frac{1}{n_{++}-m_+}\sum_{g=1}^{G}\sum_{i=1}^{I_g}\sum_{j=1}^{n_{ig}}\sum_{j'=1}^{n_{ig}}\sum_{k=1}^{N}\sum_{l=1}^{N}\frac{w_{kl}\left(\varepsilon^{[k]}_{jig}-\varepsilon^{[k]}_{j'ig}\right)\left(\varepsilon^{[l]}_{jig}-\varepsilon^{[l]}_{j'ig}\right)}{2n_{ig}}\right)\\
&= E\left(\frac{1}{n_{++}-m_+}\sum_{g=1}^{G}\sum_{i=1}^{I_g}\frac{1}{2n_{ig}}\sum_{j=1}^{n_{ig}}\sum_{j'=1}^{n_{ig}}\left[\sum_{k=1}^{N}w_{kk}\left(\varepsilon^{[k]}_{jig}-\varepsilon^{[k]}_{j'ig}\right)^2+2\sum_{k=1}^{N-1}\sum_{l>k}^{N}w_{kl}\left(\varepsilon^{[k]}_{jig}-\varepsilon^{[k]}_{j'ig}\right)\left(\varepsilon^{[l]}_{jig}-\varepsilon^{[l]}_{j'ig}\right)\right]\right)\\
&= \left\{\begin{array}{l}
\frac{1}{n_{++}-m_+}\sum_{g=1}^{G}\sum_{i=1}^{I_g}\frac{1}{2n_{ig}}\left[2n_{ig}\sum_{k=1}^{N}w_{kk}\left[\sum_{j=1}^{n_{ig}}E\left(\varepsilon^{[k]2}_{jig}\right)-n_{ig}E\left(\varepsilon^{[k]2}_{\bullet ig}\right)\right]\right]\\
+\frac{1}{n_{++}-m_+}\sum_{g=1}^{G}\sum_{i=1}^{I_g}\frac{1}{2n_{ig}}\left[4\sum_{k=1}^{N-1}\sum_{l>k}^{N}w_{kk}\sum_{j=1}^{n_{ig}}n_{ig}E\left(\varepsilon^{[k]}_{jig}\varepsilon^{[l]}_{jig}\right)\right]\\
-\frac{1}{n_{++}-m_+}\sum_{g=1}^{G}\sum_{i=1}^{I_g}\frac{1}{2n_{ig}}\left[4\sum_{k=1}^{N-1}\sum_{l>k}^{N}w_{kl}\sum_{j=1}^{n_{ig}}\sum_{j'=1}^{n_{ig}}E\left(\varepsilon^{[k]}_{jig}\varepsilon^{[l]}_{j'ig}\right)\right]
\end{array}\right.
\end{aligned}$$

Or $E\left(\varepsilon_{jig}^{[k]^2}\right) = \sigma^{[k]^2}$, $E\left(\varepsilon_{\bullet ig}^{[k]^2}\right) = \frac{\sigma^{[k]^2}}{n_{ig}}$, $E\left(\varepsilon_{jig}^{[k]}\varepsilon_{jig}^{[l]}\right) = \sigma^{[k][l]}$ et $E\left(\varepsilon_{jig}^{[k]}\varepsilon_{j'ig}^{[l]}\right) = 0$ car les individus sont indépendants. Ainsi

$$E\left(MS\left(WP\right)\right) = \sum_{k=1}^{N} w_{kk}\sigma^{[k]^2} + 2\frac{n_{++}}{\left(n_{++} - m_{+}\right)}\sum_{k=1}^{N-1}\sum_{l>k}^{N} w_{kl}\sigma^{[k][l]}.$$

De la même façon, en considérant que les effets a, b et ε et les individus sont indépendants entre eux, on démontre que

$$E\left(MS\left(AP/WG\right)\right) = n_1\sum_{k=1}^{N} w_{kk}\sigma_b^{[k]^2} + \sum_{k=1}^{N} w_{kk}\sigma^{[k]^2} + 2\frac{n_{++}}{\left(m_{+} - G\right)}\sum_{k=1}^{N-1}\sum_{l>k}^{N} w_{kl}\sigma_b^{[k][l]},$$

et

$$E\left(MS\left(AP/WG\right)\right) = n_3\sum_{k=1}^{N} w_{kk}\sigma_a^{[k]^2} + n_2\sum_{k=1}^{N} w_{kk}\sigma_b^{[k]^2} + \sum_{k=1}^{N} w_{kk}\sigma^{[k]^2} + 2\frac{n_{++}}{\left(G-1\right)}\sum_{k=1}^{N-1}\sum_{l>k}^{N} w_{kl}\sigma_a^{[k][l]}.$$

Si $w_{kl} = 0$ pour tout k et $l \neq k$, alors ces espérances sont des combinaisons positives de variances.

D'après Excoffier *et al.*, l'avantage des composants de variance est de permettre de déterminer les groupements de populations fournissant la plus grande variance entre groupes. Cependant, dans le logiciel ARLEQUIN qui implémente l'AMOVA, les fonctions calculant les dissimilarités entre haplotypes ne font pas intervenir la formule 3.3 page 75. Les $\delta_{jj'}$ sont définis indépendamment des vecteurs $\mathbf{y}_{jig}$. Quelle est alors la signification des composants de variance et comment assurer leur positivité ? La question est ouverte. Tant qu'aucune réponse ne sera faite à cette question, il me semble que, pour déterminer les groupements de populations fournissant la plus grande variance entre groupes, il est préférable d'utiliser des statistiques similaires à celle de l'ANOVA (pseudo statistique de Fisher) et d'exécuter des méthodes de classification ou d'ordination. De même, pour l'estimation de statistiques similaires à celles de Wright, les choix faits par Nei (1987) et Weir et Cockerham (1984) semblent préférables :

$$\Phi_{SC}^{*} = \frac{SSD\left(AP/WG\right)}{SSD\left(WG\right)} \text{ et } \Phi_{CT}^{*} = \frac{SSD\left(AG\right)}{SSD\left(Total\right)}$$

Montrons maintenant que l'AMOVA est en fait implicitement basée sur l'entropie quadratique. Plusieurs individus partagent souvent le même haplotype. Ainsi, dans l'AMOVA, chaque population peut être caractérisée par son nombre total d'individus et l'abondance relative, ou la fréquence, de chaque haplotype. Soient n_{kig} l'abondance de l'haplotype k dans la population i du groupe g, S le nombre total d'haplotypes distincts dans l'ensemble des groupes, et δ_{kl}^{hap} la dissimilarité entre les haplotypes k et l. La somme totale des carrés des écarts peut être réécrite de la façon suivante :

$$SSD(Total) = n_{++}\sum_{g=1}^{G}\frac{n_{+g}}{n_{++}}\sum_{i=1}^{I_g}\frac{n_{ig}}{n_{+g}}\sum_{g'=1}^{G}\frac{n_{+g'}}{n_{++}}\sum_{i'=1}^{I_{g'}}\frac{n_{i'g'}}{n_{+g'}}\sum_{k=1}^{S}\sum_{l=1}^{S}\frac{n_{kig}}{n_{ig}}\frac{n_{li'g'}}{n_{i'g'}}\frac{\left(\delta_{kl}^{\text{hap}}\right)^2}{2}$$

Choisissons les notations suivantes :

- μ_{ig}, le poids de la population i dans le groupe g en terme de taille relative (nombre d'individus présents dans la population i relativement au nombre total d'individus dans le groupe g),

- λ_g, le poids, en terme de taille relative, du groupe g par rapport à l'ensemble des groupes,
- $p_{kig} = n_{kig}/n_{ig}$, la fréquence de l'haplotype k dans la population i du groupe g,
- $p_{k\bullet g} = \sum_{i=1}^{I_g} \mu_{ig} p_{kig}$, la fréquence moyenne de l'haplotype k dans le groupe g,
- $p_{k\bullet\bullet} = \sum_{g=1}^{G} \lambda_g p_{k\bullet g}$, la fréquence moyenne totale de l'haplotype k, c'est-à-dire dans l'ensemble des groupes mélangés.

Avec ces notations, $SSD(Total)$ devient :

$$\begin{aligned} SSD(Total) &= n_{++} \sum_{g=1}^{G} \lambda_g \sum_{i=1}^{I_g} \mu_{ig} \sum_{g'=1}^{G} \lambda_{g'} \sum_{i'=1}^{I_{g'}} \mu_{i'g'} \sum_{k=1}^{S} \sum_{l=1}^{S} p_{kig} p_{li'g'} \frac{\left(\delta_{kl}^{\text{hap}}\right)^2}{2} \\ &= n_{++} \sum_{k=1}^{S} \sum_{l=1}^{S} \left(\sum_{g=1}^{G} \lambda_g \sum_{i=1}^{I_g} \mu_{ig} p_{kig} \right) \left(\sum_{g'=1}^{G} \lambda_{g'} \sum_{i'=1}^{I_{g'}} \mu_{i'g'} p_{li'g'} \right) \frac{\left(\delta_{kl}^{\text{hap}}\right)^2}{2} \\ &= n_{++} \sum_{k=1}^{S} \sum_{l=1}^{S} p_{\bullet\bullet k} p_{\bullet\bullet l} \frac{\left(\delta_{kl}^{\text{hap}}\right)^2}{2}, \end{aligned}$$

On reconnaît l'entropie quadratique.

Soient $H_{\Delta^{\text{hap}}}$ l'entropie quadratique appliquée aux dissimilarités entre haplotypes, et $\mathbf{p}_{\bullet\bullet} = (p_{1\bullet\bullet}, ..., p_{S\bullet\bullet})^t$ le vecteur contenant la distribution des fréquences globales des haplotypes. Ainsi

$$SSD(Total) = n_{++} H_{\Delta^{\text{hap}}}(\mathbf{p}_{\bullet\bullet})$$

Posons $\mathbf{p}_{ig} = \left(p_{1ig}, ..., p_{Sig}\right)^t$ le vecteur contenant la distribution des fréquences des haplotypes dans la population i du groupe g, et $\mathbf{p}_{\bullet g} = \left(p_{1\bullet g}, ..., p_{S\bullet g}\right)^t$ celui de la distribution de fréquences moyennes dans le groupe g. De la même façon, les autres composants peuvent être reformulés comme suit :

$$SSD(WP) = n_{++} \sum_{g=1}^{G} \lambda_g \sum_{i=1}^{I_g} \mu_{ig} H_{\Delta^{\text{hap}}}\left(\mathbf{p}_{ig}\right)$$

$$SSD(AP/WG) = n_{++} \left(\sum_{g=1}^{G} \lambda_g H_{\Delta^{\text{hap}}}\left(\mathbf{p}_{\bullet g}\right) - \sum_{g=1}^{G} \lambda_g \sum_{i=1}^{I_g} \mu_{ig} H_{\Delta^{\text{hap}}}\left(\mathbf{p}_{ig}\right) \right)$$

$$SSD(AG) = n_{++} \left(H_{\Delta^{\text{hap}}}(\mathbf{p}_{\bullet\bullet}) - \sum_{g=1}^{G} \lambda_g H_{\Delta^{\text{hap}}}\left(\mathbf{p}_{\bullet g}\right) \right)$$

L'AMOVA est donc basée sur une décomposition de l'entropie quadratique.

Nei et Li (1979) et Lynch et Crease (1990) utilisent également l'entropie quadratique pour la mesure de la diversité nucléotidique en génétique. Ils proposent une mesure de la différence entre deux distributions de fréquences :

$$D(\mathbf{p}, \mathbf{q}) = H_{\Delta^{\text{hap}}}(\mathbf{p}, \mathbf{q}) - \frac{1}{2}\left(H_{\Delta^{\text{hap}}}(\mathbf{p}) + H_{\Delta^{\text{hap}}}(\mathbf{q})\right)$$

où $H_{\Delta^{\text{hap}}}(\mathbf{p},\mathbf{q}) = \sum_{k=1}^{S}\sum_{l=1}^{S} p_k q_l \frac{(\delta_{kl}^{\text{hap}})^2}{2}$. Une autre façon d'écrire le composant $SSD(AP/WG)$ est

$$SSD(AP/WG) = n_{++}\sum_{g=1}^{G}\lambda_g\left(\sum_{i=1}^{I_g}\mu_{ig}\sum_{i'=1}^{I_g}\mu_{i'g}H_{\Delta^{\text{hap}}}\left(\mathbf{p}_{ig},\mathbf{p}_{i'g}\right) - \sum_{i=1}^{I_g}\mu_{ig}H_{\Delta^{\text{hap}}}\left(\mathbf{p}_{ig}\right)\right),$$

ce qui correspond à

$$SSD(AP/WG) = n_{++}\sum_{g=1}^{G}\lambda_g\left(\sum_{i=1}^{I_g}\mu_{ig}\sum_{i'=1}^{I_g}\mu_{i'g}\left[H_{\Delta^{\text{hap}}}\left(\mathbf{p}_{ig},\mathbf{p}_{i'g}\right) - \frac{1}{2}\left(H_{\Delta^{\text{hap}}}\left(\mathbf{p}_{ig}\right) + H_{\Delta^{\text{hap}}}\left(\mathbf{p}_{i'g}\right)\right)\right]\right),$$

donc à

$$SSD(AP/WG) = n_{++}\sum_{g=1}^{G}\lambda_g\left(\sum_{i=1}^{I_g}\sum_{i'=1}^{I_g}\mu_{ig}\mu_{i'g}D\left(\mathbf{p}_{ig},\mathbf{p}_{i'g}\right)\right).$$

De la même façon,

$$SSD(AG) = n_{++}\left(\sum_{g=1}^{G}\sum_{g'}^{G}\lambda_g\lambda_{g'}D\left(\mathbf{p}_{\bullet g},\mathbf{p}_{\bullet g'}\right)\right).$$

L'AMOVA est donc bâtie sur un schéma similaire à celui de la décomposition de la diversité nucléotidique de Lynch et Crease (1990). Les propositions d'Excoffier *et al.* et de Lynch et Crease diffèrent cependant sur la définition d'une statistique analogue au F_{ST} de Wright. Lynch et Crease, comme Nei l'avait proposé avec l'indice de Gini-Simpson, développent une statistique basée sur les sommes des carrés des écarts (SSD) alors qu'Excoffier *et al.* choisissent de baser leur statistique sur les composants de variances.

Pour mieux rendre compte de la décomposition de l'entropie quadratique, une dernière notation peut être utilisée. Notons $\mu_g = \left(\mu_{1g},...,\mu_{I_+g}\right)^t$ le vecteur contenant les poids des populations dans le groupe g. Ce vecteur est de taille I_+, nombre total de populations. Une population n'appartenant pas à g a un poids de zéro par rapport à g. Notons de la même façon, $\lambda = (\lambda_1,...,\lambda_G)^t$ le vecteur de la distribution de poids des groupes. Par similitude avec la matrice Δ^{hap}, nous pouvons définir une matrice de dissimilarités entre toutes les populations $\Delta^{\text{pop}} = \left[\delta_{igi'g'}^{\text{pop}}\right]$, où $\delta_{igi'g'}^{\text{pop}} = \sqrt{2D\left(\mathbf{p}_{ig},\mathbf{p}_{i'g'}\right)}$, et une matrice de dissimilarités entre groupes $\Delta^{\text{gro}} = \left[\delta_{gg'}^{\text{gro}}\right]$, où $\delta_{gg'}^{\text{gro}} = \sqrt{2D\left(\mathbf{p}_{\bullet g},\mathbf{p}_{\bullet g'}\right)}$. Ainsi les termes de l'AMOVA peuvent tous être exprimés sous forme d'entropie quadratique :

$$SSD(Total) = n_{++}H_{\Delta^{\text{hap}}}(\mathbf{p}_{\bullet\bullet})$$

$$SSD(WP) = n_{++}\sum_{g=1}^{G}\lambda_g\sum_{i=1}^{I_g}\mu_{ig}H_{\Delta^{\text{hap}}}\left(\mathbf{p}_{ig}\right)$$

$$SSD(AP/WG) = n_{++}\sum_{g=1}^{G}\lambda_g H_{\Delta^{\text{pop}}}\left(\mu_g\right). \tag{3.4}$$

$$SSD(AG) = n_{++} H_{\Delta^{\text{gro}}}(\lambda).$$

Pour présenter l'AMOVA, Excoffier *et al.* sont partis de données haploïdes. Cette analyse peut être facilement étendue à des données diploïdes lorsque la phase gamétique est connue, c'est-à-dire lorsque les deux haplotypes de tout individu sont connus (Schneider *et al.* 2000). Dans ce cas, le niveau de la diversité inter-haplotypes intra-individu est rajouté à la décomposition de la variation moléculaire. Par contre, le modèle se complexifie lorsque cette phase gamétique est inconnue, donc pour des marqueurs codominants tels que ceux obtenus par la technique moléculaire RAPD (cf. partie 2.1.2) (Stewart et Excoffier 1996).

3.2 D´ ´ ´

Nous avons vu dans la partie 2.1.3 que les structures des jeux de données en génétique et en écologie sont similaires. Les approches pour analyser ces données ont pourtant souvent été différentes.

3.2.1 Le point de vue de Whittaker : diversités α, β, γ

Whittaker (1972) divise la diversité en trois composants : diversité α, diversité ou différenciation β et diversité γ :

La diversité α est la diversité dans une communauté. Elle est mesurée soit par la richesse soit par des indices de diversité spécifique. Elle est liée à la complexité de l'hyperespace contenant les niches de toutes les espèces présentes. Les axes de cet hyperespace sont les gradients de ressources. Plus il y a d'espèces dans une communauté, plus le nombre de gradients utilisés et le nombre de niches différentes augmentent, donc plus la complexité de l'hyperespace augmente. La diversité α est mesurée à petite échelle spatiale. Lorsqu'elle est évaluée sur un échantillon, celui-ci est appelé "échantillon α".

La diversité β est la différenciation des communautés le long d'un gradient d'habitats. Dans la pratique, différents échantillons seront pris le long de ce gradient ; ils sont appelés échantillons alpha.

La diversité γ est la diversité d'un paysage ou d'une région. Elle est mesurée dans l'ensemble des échantillons alpha réunis.

Pour choisir ces dénominations utilisés pour décrire ce schéma de décomposition de la diversité, Whittaker (1960) s'est inspiré de l'article de Fisher *et al.* (1943) qui introduisent une constante α dans une fonction représentant le lien entre nombre d'espèces et nombre d'individus observés. Cette constante α quantifie l'augmentation du nombre d'espèces observées quand la taille de l'échantillon est augmentée de e.

Whittaker définit la relation suivante entre les trois composants de diversité :

$$\beta = \frac{\gamma}{\alpha}.$$

Il propose entre autre de mesurer la diversité β à partir de la richesse par

$$BD = \frac{S_e}{\overline{S} - 1},$$

où S_e est le nombre d'espèces dans l'ensemble composite (combinaison des échantillons α), et $\bar{S}$ est le nombre moyen d'espèces dans les échantillons α. La valeur 1 est retranchée au quotient de sorte que la différenciation β mesurée sur un seul échantillon soit nulle. Dans cette formule S_e est une mesure de diversité γ et $\bar{S} - 1$ une mesure de diversité α. Egalement à partir de la richesse, Koch (1957) avait défini une fonction appelée indice de dispersion biotique (IBD, index of biotal dispersity)

$$\text{IBD} = \frac{T - S}{(r-1)S} \times 100,$$

où T est la somme des richesses de chaque échantillon, S est la richesse totale (nombre d'espèces différentes dans l'ensemble des échantillons) et r est le nombre d'échantillons. Cet indice vaut 0 lorsque les compositions spécifiques des échantillons sont disjointes, $1/100$ lorsque les échantillons ont en commun une seule espèce, et 100 lorsque les échantillons sont identiques. Cet indice mesure la similarité globale entre les échantillons. La mesure 100 − IBD peut être alors proposée comme une mesure de diversité β calculée à partir de la richesse.

Une autre mesure proposée par Whittaker est

$$\frac{\exp\left(H_{S_e}\right)}{\exp\left(\bar{H_S}\right)},$$

où H_{S_e} est la valeur de l'indice de Shannon dans l'ensemble composite, et $\bar{H_S}$ la valeur moyenne de cet indice dans les échantillons α. L'exponentielle est appliqué à H_S pour éviter le quotient de logarithme (cf. section 2.2.3).

Plusieurs indices ont été développés pour mesurer la dissimilarité entre deux échantillons (*e.g.*, Jaccard 1901, cf. partie 4.1). Selon Koch (1957) la principale limite de ces indices, et en particulier celui de Jaccard, est qu'ils ne sont calculés qu'entre deux assemblages seulement. Whittaker propose néanmoins, comme une alternative possible, que la diversité β soit calculée comme la moyenne des dissimilarités entre paires d'assemblages (Whittaker 1972).

Au départ, la diversité β était destinée à l'évaluation des différences entre communautés le long d'un gradient environnemental. Whittaker souligne néanmoins que ces trois diversités peuvent être étudiées à d'autres échelles que communauté, gradient et paysage.

Les composants de Whittaker sont très différents à cette échelle de ceux proposés par Wright à une autre échelle. Ils ont cependant en commun leur liaison par une relation multiplicative :

$$P_{IT} = P_{IS} P_{ST} \text{ et } (1 - F_{IT}) = (1 - F_{IS})(1 - F_{ST}).$$

$$\gamma = \alpha\beta$$

Le composant F_{ST} s'apparente néanmoins à une mesure de diversité β.

3.2.2 Décomposition additive de la richesse et de la diversité spécifique

Lande (1996) propose une décomposition additive de la richesse spécifique. Soient S_T la richesse totale, i.e. le nombre total d'espèces différentes dans un ensemble d'assemblages, S_i la

richesse de l'assemblage i et μ_i un poids attribué à l'assemblage i en fonction par exemple de sa taille ou de son importance. La diversité moyenne dans les assemblages est $S_w = \sum_i q_i S_i$. Le composant inter-assemblages de la richesse totale est $S_a = S_T - S_w$ de sorte que

$$S_T = S_w + S_a.$$

Il a été suggéré dans la partie précédente que l'indice de Koch (1957) sous la forme 100-IBD soit utilisé comme mesure de diversité β. Cet indice s'écrit

$$\text{100-IBD} = 100\left[1 - \frac{T - S}{(r-1)S}\right]$$

soit

$$\text{100-IBD} = 100\left[\frac{r}{r-1}\left(\frac{S - \frac{T}{r}}{S}\right)\right].$$

En réécrivant cet indice avec les notations de Lande, nous obtenons

$$\text{100-IBD} = 100\left[\frac{r}{r-1}\left(\frac{S_a}{S_T}\right)\right],$$

où S_a/S_T est le rapport de la diversité inter-assemblages selon Lande (1996) à la diversité totale, en donnant des poids uniformes aux assemblages. La multiplication de S_a/S_T par $100r/(r-1)$ permet que la valeur maximale de l'indice soit 100. Avec cette multiplication, l'indice S_a/S_T prend une forme qui peut être exprimée en pourcentage. En effet, en supposant qu'aucun assemblage n'est vide, la plus grande valeur de S_a/S_T est obtenue lorsque les assemblages possèdent tous une, et une seule, espèce distincte. Ainsi $S_w = 1$, $S_T = r$ et $S_a = r - 1$. Donc la plus grande valeur prise par S_a/S_T est $(r-1)/r$, et en multipliant par $100r/(r-1)$ on obtient un indice dont les valeurs sont comprises entre 0 et 100 inclus.

Nous avons vu que Lewontin (1972) a proposé, en génétique, une décomposition additive de l'indice de Shannon. Levins (1968) avait déjà utilisé la décomposition de l'indice de Shannon selon le même schéma mais pour des données écologiques. Pour décrire la décomposition de l'indice de Shannon, Levins (1968, page 49) affirme simplement que, lorsque plusieurs collections sont déterminées dans une communauté, la diversité totale de la communauté est égale à la somme de la diversité moyenne au sein des collections et de la diversité due à des différences entre ces collections. Il fait remarquer qu'en utilisant différentes saisons ou sites comme "collections", nous pouvons mesurer l'importance relative des composants de niche le long de différentes dimensions.

Pielou (1975) introduit une décomposition croisée de l'indice de Shannon. Contrairement au schéma de Lewontin, le schéma de décomposition de Pielou ne se fait pas sur les collections mais sur les catégories. Les entités sont réparties selon deux classifications A et B. Soit n_{ij} le nombre d'entités associées à la catégorie i de la classification A et à la catégorie j de la classification B. Soient n_{i+} et n_{+j} les effectifs marginaux et n_{++} l'effectif total. La diversité totale est la diversité des combinaisons des catégories des deux classifications :

$$H(AB) = -\sum_i \sum_j \frac{n_{ij}}{n_{++}} \ln\left(\frac{n_{ij}}{n_{++}}\right).$$

Les diversités marginales en catégories selon les classifications A et B sont égales à

$$H(A) = -\sum_i \frac{n_{i+}}{n_{++}} \ln\left(\frac{n_{i+}}{n_{++}}\right), H(B) = -\sum_j \frac{n_{+j}}{n_{++}} \ln\left(\frac{n_{+j}}{n_{++}}\right).$$

Pielou définit également les diversités conditionnelles suivantes :

$$H_A(B) = \sum_i \frac{n_{i+}}{n_{++}} \left[-\sum_j \frac{n_{ij}}{n_{i+}} \ln\left(\frac{n_{ij}}{n_{i+}}\right)\right], H_B(A) = \sum_j \frac{n_{+j}}{n_{++}} \left[-\sum_i \frac{n_{ij}}{n_{+j}} \ln\left(\frac{n_{ij}}{n_{+j}}\right)\right].$$

Tous ces termes sont connectés par la relation

$$H(AB) = H(A) + H_A(B) = H(B) + H_B(A).$$

De plus si les classifications A et B sont indépendantes alors $H(AB) = H(A) + H(B)$.

Soit une fonction H définie sur u et vérifiant les trois propriétés suivantes :
- H prend sa valeur maximale pour distribution uniforme.
- Entre deux distributions uniformes, la première de longueur S et la deuxième de longueur $S + 1$, H attribue une valeur plus grande à la deuxième distribution.
- H peut être décomposée selon deux classifications croisées par le procédé indiqué ci-dessus.

La seule fonction définie sur $\mathcal{P}$ vérifiant ces propriétés, est de la forme $H(\mathbf{p}) = -C \sum_k p_k \ln(p_k)$, $C > 0$ (Pielou 1975). Ce type de décomposition serait donc réservé à des indices positifs égaux ou proportionnels à l'indice de Shannon.

Pielou (1975) propose également une décomposition hiérarchique. Le raisonnement se fait toujours au niveau des catégories. Elle propose, par exemple, de décomposer l'indice de Shannon selon les niveaux taxonomiques espèces (E), genres (G) et familles (F). Ces trois niveaux constituent trois classifications des individus d'une communauté. Ces trois classifications sont emboîtées hiérarchiquement puisque $E \subset G \subset F$. Soient n_{kji}, n_{+ji}, n_{++i} les nombres d'individus appartenant respectivement à l'espèce k du genre j de la famille i, au genre j de la famille i, à la famille i, et n_{+++} le nombre total d'individus dans la communauté considérée. La décomposition proposée comporte quatre termes :

$$\begin{aligned}
H(EGF) &= -\sum_{i=1}^{r} \sum_{j=1}^{m_i} \sum_{k=1}^{S_{ji}} \frac{n_{kji}}{n_{+++}} \ln\left(\frac{n_{kji}}{n_{+++}}\right), \\
H(F) &= -\sum_{i=1}^{r} \frac{n_{++i}}{n_{+++}} \ln\left(\frac{n_{++i}}{n_{+++}}\right), \\
H_F(G) &= \sum_{i=1}^{r} \frac{n_{++i}}{n_{+++}} \left[-\sum_{j=1}^{m_i} \frac{n_{+ji}}{n_{++i}} \ln\left(\frac{n_{+ji}}{n_{++i}}\right)\right], \\
H_{GF}(E) &= \sum_{i=1}^{r} \frac{n_{++i}}{n_{+++}} \sum_{j=1}^{m_i} \frac{n_{+ji}}{n_{++i}} \left[-\sum_{k=1}^{S_{ji}} \frac{n_{kji}}{n_{+ji}} \ln\left(\frac{n_{kji}}{n_{+ji}}\right)\right].
\end{aligned}$$

La relation entre ces termes est

$$H(EGF) = H(F) + H_F(G) + H_{GF}(E).$$

Pour présenter la décomposition hiérarchique de Pielou, Allan (1975) choisit les facteurs de classification suivants : microhabitats (M), sites (S) et espèces (E). La décomposition proposée est alors

$$H(MSE) = H(E) + H_E(S) + H_{SE}(M). \tag{3.5}$$

La relation hiérarchique suppose que $M \subset S \subset E$. Or une même espèce peut être présente dans plusieurs microhabitats et dans plusieurs sites. Nous pouvons donc affirmer que les relations qui relient E à M d'une part et à S d'autres part ne sont pas hiérarchiques mais croisées. Dans la recherche d'Allan pour présenter une décomposition hiérarchique se trouve sous-jacente une décomposition semi-hiérarchisée ($E \times (M \subset S)$). Regroupons alors les deux derniers termes de l'équation 3.5, nous obtenons $H(MSE) = H(E) + H_E(MS)$. Puisque nous avons déterminé que les facteurs E et MS sont en fait croisés, $H(MSE)$ est aussi égal à $H(MS) + H_{MS}(E)$ et M étant inclus dans S,

$$H(MS) = H(S) + H_S(M).$$

Ainsi,

$$H(MSE) = H(S) + H_S(M) + H_{MS}(E).$$

Levins (1968) avait eu le même raisonnement que Lewontin pour décomposer l'indice de Shannon. Sa proposition a été ensuite modifiée pour que chaque terme de diversité corresponde à une estimation de taille de niche pour une espèce donnée (Colwell et Futuyma 1971, Allan 1975). Les catégories ne sont plus des espèces, elles sont déterminées par deux classifications emboîtées hiérarchiquement : microhabitat (M) et sites (S) ($M \subset S$). La nouvelle version correspond à une décomposition de la diversité en microhabitats et en sites pour une espèce donnée. La diversité est mesurée par la répartition des individus d'une espèce entre microhabitats et entre sites. Pour l'espèce k, la décomposition est $H(MS)^{[k]} = H(S)^{[k]} + H_S(M)^{[k]}$. $H(MS)^{[k]}$ mesure la taille totale de la niche de l'espèce k, $H(S)^{[k]}$ la taille de la niche au niveau des sites, et $H_S(M)^{[k]}$ la taille moyenne de la niche au niveau des microhabitats au sein d'un site.

Lande (1996) compare les décompositions, selon un schéma hiérarchique sur les collections, des trois indices traditionnels de mesure de diversité : richesse, indice de Shannon et indice de Gini-Simpson. Une des propriétés étudiées est leur concavité qui permet la décomposition. Nous avons vu dans la partie 2.2.1 que ces trois indices sont des cas particulier de l'indice de Havrda et Charvat. Il peut donc être noté que, l'indice de Havrda et Charvat étant concave, les décompositions de la richesse, de l'indice de Gini-Simpson et celle de l'indice de Shannon correspondent à trois cas particuliers de la décomposition de l'indice de Havrda et Charvat. De son étude, Lande conclut que l'indice de Gini-Simpson présente un avantage important sur les deux autres : un estimateur non-biaisé et de faible variance peut être facilement calculé pour cet indice. Lande, généticien, met ainsi en avant, dans le domaine de l'écologie, la décomposition de l'indice de Gini-Simpson, très utilisé en génétique depuis Nei (1973). Il a eu un grand impact en écologie car Lande y faisait référence aux diversités α, β et γ de Whittaker, faisant ainsi le

lien entre la décomposition multiplicative de Whittaker et les décompositions additives de la richesse, de l'indice de Gini-Simpson et de l'indice de Shannon (Veech *et al.* 2002). Cet article fait ainsi sans doute partie des travaux à l'origine de la prolifération récente des recherches sur la décomposition de la diversité en écologie (Fournier et Loreau 2001, Gering *et al.* 2003, pour des exemples d'applications).

Toutes les méthodes d'analyse de la diversité aux diverses échelles citées dans cette partie font appel à des décompositions additives. Suite aux travaux de Whittaker, en écologie la décomposition de la diversité est traditionnellement devenue multiplicative. Les deux types de décompositions, multiplicative et additive, ont été développés à la fin des années 60. La théorie multiplicative de Whittaker a eu plus d'impact que les décompositions additives de Levins et de Pielou. Pourtant depuis quelques années, suite à l'article de Lande (1996), un débat a été lancé pour promulguer la décomposition additive. L'article de Veech *et al.* (2002) en fait un résumé. Les composants α et β de diversité sont généralement interprétés séparément. Contrairement à la décomposition multiplicative, la décomposition additive permet de mesurer tous les composants de diversité de la même façon et de les exprimer avec la même unité de sorte qu'ils sont directement comparables (Wagner *et al.* 2000, Fournier et Loreau 2001, Ricotta 2003). Ce concept de décomposition de la diversité reconnaît ainsi explicitement que la diversité β peut être mesurée et définie relativement à la diversité α (Veech *et al.* 2002). Cette nouvelle approche offre la possibilité de réunifier les trois dimensions : diversités α, β et γ (Fournier et Loreau 2001).

La décomposition additive de la diversité est utilisée pour analyser les changements de diversité selon des dimensions spatiales horizontales (Wagner *et al.* 2000, Gering *et al.* 2003) et verticales (DeVries *et al.* 1997), et selon des dimensions temporelles (DeVries *et al.* 1999). Elle permet l'étude comparative des diversités locales (intra-patch par exemple) et globales (γ). De plus, l'analyse du composant inter (β) peut permettre de déterminer si les communautés écologiques locales sont saturées à une certaine diversité constante, ou si la diversité locale (α) varie proportionnellement à la diversité régionale (γ). Cette étude dépendra de l'échelle à laquelle la communauté locale est définie, à travers l'organisation hiérarchique des multiples échelles allant du petit voisinage à la biosphère (Loreau 2000).

Les méthodes de décomposition de la diversité spécifique présentées dans cette partie sont descriptives. Notons que Gering *et al.* (2003) propose de tester les différences entre échantillons en permutant, sans remise, les individus entre les échantillons tout en conservant la taille des échantillons. Pour chaque permutation, il calcule les diversité intra et inter-échantillons, avec la richesse et les indices de Shannon et de Gini-Simpson. Pour chaque indice de diversité, il compare ensuite les composants de diversité théoriques intra et inter-échantillons avec les valeurs observées de ces composants dans l'échantillon de départ.

Toutes les méthodes présentées dans cette partie permettent de décomposer la diversité mais supposent que les catégories (*e.g.* des espèces) sont interchangeables. Ce que nous devons trouver maintenant, c'est comment prendre en compte les degrés de différence entre catégories avec des données écologiques ?

3.2.3 Existe-t-il un équivalent de l'AMOVA en écologie ?

Dans cette partie, nous allons montrer qu'il existe effectivement une méthode développée récemment en écologie qui est construite sur un schéma se rapprochant de celui de l'AMOVA.

Cette méthode est décalée par rapport aux méthodes que nous avons vu précédemment pour deux raisons. D'abord elle ne s'intéresse qu'au niveau inter-collections, et non à la diversité dans ces collections. Et ensuite elle a un objectif plus inférentiel que descriptif. Le but de cette méthode est de comparer des groupes de relevés (ou plus généralement des collections) selon plusieurs facteurs hiérarchiques ou croisés. Malgré deux différences importantes, cette méthode s'emboîte parfaitement dans le schéma de l'AMOVA.

La méthode en question a été indépendamment proposée par Pillar et Orlóci (1996) puis par Anderson (2001). Ce sont les notations d'Anderson que nous utiliserons ci-dessous. Anderson appelle "l'analyse de variance multiple et non paramétrique" (npMANOVA, "non-parametric Multiple ANalysis Of VAvriance") cette méthode qui est une réécriture de l'analyse de la redondance basée sur des distances (dbRDA, "distance-based ReDundancy Analysis") (Legendre et Anderson 1999, McArdle et Anderson 2001). Anderson fait remarquer que la décomposition de la variation est particulièrement importante pour tester des hypothèses dans des systèmes écologiques complexes avec une variabilité naturelle temporelle et spatiale. Son analyse a donc pour but de tester des différences entre communautés ou échantillons qui contiennent des individus appartenant à différentes espèces. Pour ramener cette analyse à un point de vue plus général, appelons collections ces communautés ou échantillons. Ces collections se répartissent dans G groupes à raison de I collections par groupes. La première étape est la définition par l'utilisateur de dissimilarités entre collections selon leurs compositions spécifiques (cf. partie 4.1). Notons δ_{ij} la dissimilarité entre les collections i et j. Par similarité avec l'analyse de variance (ANOVA), Anderson définit

• une somme totale des carrés des écarts

$$SS_T = \frac{1}{IG} \sum_{i=1}^{IG-1} \sum_{j=i+1}^{IG} \delta_{ij}^2,$$

• une somme des carrés des écarts à l'intérieur des groupes

$$SS_W = \frac{1}{I} \sum_{i=1}^{IG-1} \sum_{j=i+1}^{IG} \delta_{ij}^2 \varepsilon_{ij},$$

où ε_{ij} est égal à 1 si les deux collections i et j sont dans le même groupe et à 0 sinon,

• et un dernier composant : la somme des carrés des écarts entre les groupes

$$SS_B = SS_T - SS_W.$$

En quoi cette méthode ressemble-t-elle à l'AMOVA ? Intéressons nous au « composant inter-populations au sein des groupes » de l'AMOVA. Dans la dernière réécriture de l'AMOVA, nous

avons vu que ce composant est égal à (cf. formule 3.4 partie 3.1.3)

$$SSD(AP/WG) = n_{++} \sum_{g=1}^{G} \lambda_g \left(\sum_{i=1}^{I_g} \sum_{j=1}^{I_g} \mu_{ig}\mu_{jg} \frac{\left(\delta_{ig,jg}^{\text{pop}}\right)^2}{2} \right).$$

Or le composant SS_W de la npMANOVA peut être réécrit ainsi :

$$SS_W = IG \sum_{g=1}^{G} \frac{1}{G} \left(\sum_{i=1}^{I} \sum_{j=1}^{I} \frac{1}{I}\frac{1}{I} \frac{\delta_{ig,jg}^2}{2} \right).$$

Dans la npMANOVA, contrairement à l'AMOVA, la diversité intra-collection n'est pas considérée, d'où la multiplication par IG, nombre total de collections, au lieu de n_{++} nombre total d'individus. De plus, dans la npMANOVA, les poids des collections et ceux des groupes sont uniformes. Cette uniformité est due à la prise en compte uniquement de jeux de données équilibrés (même nombre de collections dans chaque groupe). A partir des deux écritures suivantes de SS_T et SS_W

$$\begin{aligned} SS_T &= IG \sum_{g=1}^{G} \sum_{g'=1}^{G} \frac{1}{G}\frac{1}{G} \sum_{i=1}^{I} \sum_{j=1}^{I} \frac{1}{I}\frac{1}{I} \frac{\delta_{ig,jg'}^2}{2}, \\ SS_W &= IG \sum_{g=1}^{G} \sum_{g'=1}^{G} \frac{1}{G}\frac{1}{G} \left(\frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{I} \frac{1}{I}\frac{1}{I} \frac{\delta_{ig,jg}^2}{2} + \frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{I} \frac{1}{I}\frac{1}{I} \frac{\delta_{ig',jg'}^2}{2} \right), \end{aligned}$$

il peut être déduit que

$$SS_B = IG \sum_{g=1}^{G} \sum_{g'=1}^{G} \frac{1}{G}\frac{1}{G} \left[\sum_{i=1}^{I} \sum_{j=1}^{I} \frac{1}{I}\frac{1}{I} \frac{\delta_{ig,jg'}^2}{2} - \frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{I} \frac{1}{I}\frac{1}{I} \frac{\delta_{ig,jg}^2}{2} - \frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{I} \frac{1}{I}\frac{1}{I} \frac{\delta_{ig',jg'}^2}{2} \right].$$

Entre crochets apparaît la différence de Jensen appliquée à la formule de l'entropie quadratique au niveau inter-collections. En posant que la dissimilarité entre deux collections est la racine carrée de deux fois cette différence de Jensen alors :

$$\delta_{gg'}^{\text{gro}} = \sqrt{2 \left[\sum_{i=1}^{I} \sum_{j=1}^{I} \frac{1}{I}\frac{1}{I} \frac{\delta_{ig,jg'}^2}{2} - \frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{I} \frac{1}{I}\frac{1}{I} \frac{\delta_{ig,jg}^2}{2} - \frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{I} \frac{1}{I}\frac{1}{I} \frac{\delta_{ig',jg'}^2}{2} \right]},$$

et

$$SS_B = IG \left(\sum_{g=1}^{G} \sum_{g'=1}^{G} \frac{1}{G}\frac{1}{G} \frac{\left(\delta_{gg'}^{\text{gro}}\right)^2}{2} \right).$$

Pour tester l'absence de différences entre groupes (hypothèse $\mathtt{H}_0$), Anderson propose d'utiliser des statistiques similaires aux statistiques-F de Fisher utilisées dans l'ANOVA. Elle les appelle pseudo-quotients F. Ici,

$$F = \frac{SS_B/(G-1)}{SS_W/(I-1)}.$$

Anderson propose le même schéma de permutations que Excoffier *et al.* (1992) : permuter les collections entre les groupes, en considérant que sous $\mathtt{H}_0$ les collections sont interchangeables. Elle suggère de regarder toutes les permutations possibles pour obtenir la distribution exacte du pseudo-quotient F. Soit F^{π} désignant une valeur théorique prise par le pseudo-quotient F après permutation, alors une valeur-p est définie par

$$\mathrm{p} = \frac{(\text{Nombre de } \mathrm{F}^{\pi} \geq F)}{(\text{Nombre total de } \mathrm{F}^{\pi})}$$

(la valeur F est incluse dans les valeurs F^{π} puisqu'elle correspond à une des permutations possibles). Si le nombre total de permutations possibles ($\frac{(GI)!}{G!(I!)^G}$) est trop grand, Anderson suggère de prendre 1000 permutations pour une erreur de première espèce α de 0.05, et 5000 permutations pour un α de 0.01. Pillar et Orlóci (1996) proposent la même procédure de permutations mais pas la même statistique. Ils choisissent comme statistique directement SS_B.

Anderson insiste sur le fait que n'importe quelle mesure de dissimilarités entre collections peut être utilisée. Pourtant, avec certaines mesures, des composants de variation négatifs peuvent être obtenus, par exemple si le composant SS_W devient plus grand que le composant SS_T. L'écriture de la dbRDA montrait que ce problème de composants de variation négatifs n'apparaît pas lorsque des matrices de dissimilarité $\mathbf{\Delta} = \left[\delta_{ij}\right]$ euclidiennes sont utilisées. Et Pillar et Orlóci (1996) restreignent leur analyse à ces matrices.

Pillar et Orlóci (1996) et Anderson (2001) proposent également des modèles plus complexes : des modèles croisés avec interaction entre facteurs et des modèles hiérarchiques. Anderson (2001) se restreint à des plans d'expérience équilibrés. Pillar et Orlóci (1996) signalent seulement que dans le cas de facteurs croisés, l'effet de l'interaction entre les deux facteurs ne peut être calculé que si le jeu de données contient des enregistrements pour toutes les combinaisons des modalités des facteurs. Dans le cas d'un modèle plus complexe, Anderson insiste sur le fait que la détermination de schémas de permutations est difficile et dépend de deux questions :

1. Quelles unités doivent être permutées ?
2. Quelles restrictions doivent être imposées aux permutations pour tenir compte des autres facteurs du plan d'expérience ?

Plusieurs alternatives existent, telles que la permutation des résidus et celle des données brutes entre tous les termes de l'analyse (Legendre et Anderson 1999).

Smith *et al.* (1990) proposent une analyse similaire tout en permettant de pondérer différemment des groupes de tailles différentes. Soient $\bar{W}_g$ la moyenne des dissimilarités entre les populations dans le groupe g, $\bar{B}_{gg'}$ la moyenne des dissimilarités entre les collections du groupe g et celles du groupe g', m_g le nombre de collections dans le groupe g, et m_+ le nombre total de collections. Soient le composant de variation intra-groupe

$$\bar{W} = \sum_{g=1}^{G} \frac{m_g}{m_+} \bar{W}_g,$$

et un composant inter-groupe

$$\overline{B} = \frac{\sum_{g=1}^{G-1} \sum_{g'=g+1}^{G} m_g m_{g'} \overline{B}_{gg'}}{\sum_{g=1}^{G-1} \sum_{g'=g+1}^{G} m_g m_{g'}}.$$

Pour tester s'il existe effectivement des différences entre groupes, Smith *et al.* considèrent la statistique

$$M = \frac{\overline{B}}{\overline{W}}$$

et permutent les collections entre groupes, en gardant la taille des groupes. Il existe en général $N_p = m_+!/\left(m_1!m_2!...m_G!\right)$ permutations possibles, et moins si deux ou plus des groupes ont des tailles égales : sur l'ensemble de ces permutations, seulement $m_+!/\prod\left[(c!)^{\nu_c}\,\nu_c!\right]$ permutations, où ν_c est le nombre de groupes contenant c collections, sont uniques. Smith *et al.* soulignent que ce test peut être entrepris à partir de matrices de dissimilarités ou à partir de matrices de similarités entre collections. Dans le premier cas, l'hypothèse d'égalité des groupes est rejetée si $\overline{B}$ est grand ou si $\overline{W}$ est petit. Dans le second cas, l'hypothèse est rejetée si $\overline{B}$ est petit ou si $\overline{W}$ est grand. Smith *et al.* proposent également d'étudier la contribution de chaque espèce aux différences observées entre groupes. Ainsi, en notant B_{-k} la moyenne des dissimilarités (resp. similarités) inter-groupes lorsque l'espèce k est éliminée, alors

$$INF_k = 100\left(\hat{B}_{-k} - \hat{B}\right)/\hat{B}$$

mesure l'influence relative, en pourcentage, de l'espèce k sur la moyenne des dissimilarités (resp. similarités) inter-groupes. Une grande valeur positive indique une espèce dont les abondances varient peu (resp. beaucoup) entre groupes et une forte valeur négative montre une espèce dont les abondances changent beaucoup (resp. peu) entre les groupes.

Toutes ces méthodes ont des liens avec l'AMOVA. Cependant l'AMOVA présente l'avantage de pointer du doigt l'importance de la prise en compte des dissimilarités entre catégories dans l'étude des variations dans et entre collections. Au contraire, en écologie, les méthodes de Smith *et al.* (1990), Pillar et Orlóci (1996) et Anderson (2001) partent de matrices de dissimilarités entre collections qui, traditionnellement sont basées soit sur les présences/absences soit sur les abondances des espèces, et ne prennent donc pas en compte des mesures de dissimilarités entre espèces ou plus généralement entre catégories. L'autre inconvénient des méthodes de Smith *et al.* (1990), Pillar et Orlóci (1996) et Anderson (2001), comparativement à l'AMOVA, est qu'elles ne contiennent pas d'information sur la diversité dans les collections.

Pourrait-on trouver alors une méthode qui prennent en compte les dissimilarités entre espèces dans l'étude de la décomposition de la diversité ?

La solution est donnée par Smith (1995) dans un commentaire à l'article de Izsak et Papp (1995). Cette solution est un point de vue statistique et elle se trouve dans l'axiomatisation de Rao (1986) sur les mesures de diversité (Pavoine et Dolédec 2005, cf. annexe 2).

3.3 L ´ ,

La conceptualisation de la diversité intra et de la diversité inter-collections doit être faite dans un schéma unifié (Chakraborty et Rao 1991).

3.3.1 L'axiomatisation de Rao

Rao (1986) a proposé un ensemble d'axiomes définissant un indice de diversité, qu'il a lui-même appelé "l'axiomatisation de Rao des mesures de diversité". Il considère un ensemble convexe de distributions de probabilités. Notons Π un tel ensemble. Qu'il soit convexe signifie que pour tout P et Q appartenant à Π et pour tout réel α appartenant à l'intervale $[0, 1]$ on a :

$$\alpha P + (1 - \alpha)Q \in \Pi.$$

Soit H une fonction à valeurs réelles définie sur Π. Pour caractériser H comme une mesure de diversité d'une distribution, Rao considère les axiomes suivants :

C0 : $H(P) = -J_0(P) \geq 0 \; \forall P \in \Pi$, et $H(P) = 0$ si P est dégénéré. [Dans notre cas, la dégénérescence de P correspondra à la convergence de la distribution sur une seule entité, c'est-à-dire à une distribution avec une probabilité de 1 pour une seule entité]

C1 : $J_0 = -H(P)$ est convexe sur Π.

L'axiome C1 est équivalent à "H est concave sur Π" (cf. définition de la convexité et de la concavité d'une fonction partie 2.2.2 page 25) et signifie que la diversité dans un mélange de deux collections ne doit pas être inférieure à la moyenne des diversités au sein de chaque collection, *i.e.*, que la diversité augmente dans des mélanges. Pour $P_1, P_2, ... \in \Pi$ et $\mu_1, \mu_2, ... \in \mathbb{R}^+$, $\mu_1 + \mu_2 + ... = 1$, J_1 la différence de Jensen du premier ordre s'écrit

$$J_1\left(\{P_i\} : \{\lambda_i\}\right) = \sum \lambda_i J_0\left(P_i\right) - J_0\left(\sum \lambda_i P_i\right).$$

C2 : J_1 est convexe sur Π^2.

L'axiome C2 signifie que la dissimilarité entre deux groupes issus d'un mélange de deux collections ($(\mu_1 P_{11} + \mu_2 P_{21})$ et $(\mu_1 P_{12} + \mu_2 P_{22})$, où $\mu_1 + \mu_2 = 1$) ne doit pas être supérieure à la moyenne des dissimilarités entre collections au sein de chaque groupe (i.e. entre P_{11} et P_{12} et entre P_{21} et P_{22}) ; la dissimilarité diminue par des mélanges. Soient P_{11}, P_{12}, ..., P_{1m}, P_{21}, P_{22}, ..., P_{2m}, ..., $P_{rm} \in \Pi$, et λ_1, λ_2, ..., $\lambda_r \in \mathbb{R}^+$, $\sum_{i=1}^{r} \lambda_i = 1$; μ_1, μ_2, ..., $\mu_m \in \mathbb{R}^+$, $\sum_{j=1}^{m} \mu_j = 1$. Soient de plus, $P_{i\bullet} = \sum_{j=1}^{m} \mu_j P_{ij}$, et $P_{\bullet j} = \sum_{i=1}^{r} \lambda_i P_{ij}$. La différence de Jensen du second ordre s'écrit :

$$J_2\left(\left\{P_{ij}\right\} : \left\{\lambda_i \mu_j\right\}\right) = \sum_{i=1}^{r} \lambda_i J_1\left(\left\{P_{ij}\right\} : \left\{\mu_j\right\}\right) - J_1\left(\left\{P_{\bullet j}\right\} : \left\{\mu_j\right\}\right).$$

C3 : J_2 est convexe sur Π^3.

Les différences de Jensen d'ordre supérieur sont définies de façon récursive.

Ci : J_{i-1} est convexe sur $\Pi^{2^{i-1}}$.

Rao appelle "mesure de diversité d'ordre *i*" une fonction H qui vérifie les axiomes C0 à Ci, et "mesure de diversité parfaite" une fonction H qui vérifie ces axiomes pour tout i. Pour qu'une fonction puisse être décomposée, il faut qu'elle vérifie au moins les axiomes C0 et C1.

Les axiomes C0 et C1 sont souvent admis et même exigés (Lande 1996). Si une fonction H vérifie ces deux axiomes ($\Leftrightarrow$ si elle est d'ordre 1) alors H est décomposable selon des facteurs emboîtés hiérarchiquement, quels que soient le nombre de facteurs et le nombre de modalités pour chaque facteur (Rao 1982b, Rao et Boudreau 1984). Une telle décomposition est appelée APDIV (APportionment of DIVersity). Les indices de Shannon et de Gini-Simpson vérifient ces deux conditions (Lewontin 1972, Lande 1996). D'une façon plus générale, l'indice de Havrda et Charvat vérifie ces deux équations pour tout $\alpha > 0$. L'entropie quadratique ($H_{\mathbf{D}}(\mathbf{p}) = \mathbf{p}^t\mathbf{D}\mathbf{p}$) vérifie cette condition si la matrice $\mathbf{D}$ contenant les dissimilarités d_{kl} entre catégories est définie conditionnellement négative,

D est définie conditionnellement négative si et seulement si, pour tout vecteur **a** de longueur S et tel que $\mathbf{a}^t\mathbf{1}_S = 0$, $\mathbf{a}^t\mathbf{D}\mathbf{a} \leq 0$ ($\Leftrightarrow \mathbf{\Lambda}$ euclidienne) (Rao 1986).

Prenons deux niveaux hiérarchiques : collections et groupes. Le raisonnement est exactement le même que celui de Lewontin pour l'indice de Shannon. Considérons r le nombre de groupes, m_i le nombre de collections dans le groupe i, λ_i le poids du groupe i, et μ_{ji} le poids de la population j du groupe i. Soient P_{ji}, $P_{\bullet i} = \sum_{j=1}^{m_i} \mu_{ji} P_{ji}$, $P_{\bullet\bullet} = \sum_{i=1}^{r} \lambda_i P_{\bullet i}$ les distributions de fréquences des catégories respectivement dans la collection j du groupe i, dans le groupe i, et au total. Soient $H_{tot} = H(P_{\bullet\bullet})$ la diversité totale sur l'ensemble des groupes, $\bar{H}_{gro} = \sum_{i=1}^{r} H(P_{\bullet i})$ la diversité moyenne au sein des groupes et $\bar{H}_{col} = \sum_{i=1}^{r} \lambda_i \sum_{j=1}^{m_i} \mu_{ji} H\left(P_{ji}\right)$ la diversité moyenne au sein des collections, alors la décomposition de la diversité est la suivante

$$\overbrace{H_{tot}}^{\text{Diversité totale}} = \overbrace{\bar{H}_{col}}^{\substack{\text{Diversité moyenne} \\ \text{intra-collection}}} + \overbrace{\left[\bar{H}_{gro} - \bar{H}_{col}\right]}^{\substack{\text{Diversité moyenne} \\ \text{inter-collections intra-groupe}}} + \overbrace{\left[H_{tot} - \bar{H}_{gro}\right]}^{\text{Diversité inter-groupes}} .$$

Tous les calculs de diversité inter correspondent à des différences de Jensen du premier ordre. Pour que ces mesures de diversité inter-collections soient positives la condition C1 est nécessaire et suffisante. La condition C0 est nécessaire et suffisante pour que les diversités totale et intra-collection soient positives.

Des fonctions d'ordres supérieurs permettent des décompositions sur facteurs croisés. Une fonction d'ordre i pourra être décomposée selon i facteurs croisés. Une telle décomposition se fait selon le même schéma que l'ANOVA et est appelée ANODIV (ANalysis Of DIVersity). Considérons par exemple deux facteurs croisés A et B. Soit une collection possédant la distribution de probabilité P_{ij} indexée par la modalité i parmi r d'un facteur A et la modalité j parmi m d'un facteur B. Soient λ_i le poids de la modalité i du facteur A et μ_j le poids de la modalité j du facteur B. Le poids de la collection ij est $\lambda_i\mu_j$. Cela signifie que les poids $\{\lambda_i\}$

et $\{\mu_j\}$ doivent être indépendants. Soient $P_{i\bullet} = \sum_{j=1}^{m} \mu_j P_{ij}$ la distribution marginale pour la modalité i du facteur A, $P_{\bullet j} = \sum_{i=1}^{r} \lambda_i P_{ij}$ la distribution marginale pour la modalité j du facteur B, et $P_{\bullet\bullet} = \sum_{i=1}^{r}\sum_{j=1}^{m} \lambda_i \mu_j P_{ij}$ la distribution totale, alors la décomposition de la diversité est la suivante

$$\overbrace{H(P_{\bullet\bullet})}^{\text{Diversité totale}} = \overbrace{\sum_{i=1}^{r}\sum_{j=1}^{m} \lambda_i \mu_j H\left(P_{ij}\right)}^{\substack{\text{Diversité moyenne} \\ \text{intra-collection}}} + \overbrace{J_1\left(\{P_{i\bullet}\} : \{\lambda_i\}\right)}^{\substack{\text{Effet du} \\ \text{facteur A}}} + \overbrace{J_1\left(\left\{P_{\bullet j}\right\} : \left\{\mu_j\right\}\right)}^{\substack{\text{Effet du} \\ \text{facteur B}}} + \overbrace{J_2\left(\left\{P_{ij}\right\} : \left\{\lambda_i\mu_j\right\}\right)}^{\substack{\text{Effet de} \\ \text{l'interaction entre A et B}}}.$$

Un des problèmes principaux de cette décomposition est que les poids $\{\lambda_i\}$ attribués aux modalités d'un facteur doivent être indépendants des poids $\{\mu_j\}$ attribués aux modalités d'un autre facteur. Les poids choisis par Excoffier *et al.* (1992) dans l'AMOVA, pour l'analyse hiérarchique, sont exprimés en pourcentage d'individus représentés par une modalité : par exemple λ_i serait le pourcentage d'individus associés à la modalités i du facteur A. En analyse croisée, notons n_{kij} pour désigner le nombre d'entités associées à la catégorie k et à la collection ij, n_{+ij} le nombre d'entités respectivement dans la collection ij, n_{+i+} le nombre d'entités associées à la modalité i du facteur A, et n_{++j} le nombre d'entités associées à la modalité j du facteur B. Avec les notations précédentes, le choix d'Excoffier *et al.* (1992) dans l'AMOVA correspondrait, sur ces deux facteurs croisés, à $\lambda_i = n_{i+}/n_{++}$ poids de la modalité i du facteur A, $\mu_j = n_{+j}/n_{++}$ poids de la modalité j du facteur B et $w_{ij} = n_{ij}/n_{++}$ poids de la collection associée à la modalité i du facteur A et à la modalité j du facteur B. Les pondérations naturelles sont $\lambda_i = \frac{n_{i+}}{n_{+++}}$, $\mu_j = \frac{n_{++j}}{n_{+++}}$ et $w_{ij} = \frac{n_{+ij}}{n_{+++}}$. Cependant, $\frac{n_{+ij}}{n_{+++}} = \lambda_i\mu_j$ seulement dans le cas de jeux de données équilibrés, i.e. si toutes les collections ont le même nombre d'entités et si chaque combinaison ij des deux facteurs est représentée. En effet, dans ce cas, $\lambda_i = 1/m$, $\mu_j = 1/r$ et $w_{ij} = 1/(rm)$. Dans les autres cas, Nayak (1983) suggère de conserver $\lambda_i = \frac{n_{i+}}{n_{+++}}$ et $\mu_j = \frac{n_{++j}}{n_{+++}}$ et de prendre $w_{ij} = \frac{n_{i+}}{n_{+++}}\frac{n_{++j}}{n_{+++}}$.

Tous les indices de diversité que nous avons vus dans la partie 2.2 (indice de Shannon, Gini-Simpson, Havrda et Charvat, Renyi, entropie-γ, entropie appariée, modification de l'indice de Gini-Simpson par Rao) vérifient les axiomes C0 et C1. Parmi ces indices, celui de Havrda et Charvat $H_{\text{H-C}}$

$$H_{\text{H-C}}(\mathbf{p}) = \frac{1 - \sum_{k=1}^{S} p_k^{\alpha}}{\alpha - 1},\ \alpha > 0,\ \alpha \neq 1$$

vérifie C2 lorsque $S > 2$ et $\alpha \to 1$ ou $1 < \alpha \leq 2$, et quand $S = 2$ et $\alpha \to 1$ ou $\alpha \in]1, 2] \cup [3, 11/3]$. Cet indice est égal à celui de Shannon pour $\alpha \to 1$ et à celui de Gini-Simpson pour $\alpha = 2$. Ces deux derniers indices vérifient donc les axiomes C0 à C2. L'indice de Shannon est d'ordre 2 (Rao 1982b), c'est-à-dire qu'il vérifie uniquement les axiomes C0 à C2. Il peut donc être décomposé sur un nombre quelconque de facteurs hiérarchiques, mais sur deux facteurs croisés seulement. Comme nous allons le voir ci-dessous, l'indice de Gini vérifie les propriétés Ci pour tout i, c'est donc une mesure parfaite de diversité selon Rao.

L'entropie quadratique, notée

$$H_{\mathbf{D}^{\text{cat}}}(\mathbf{p}) = \sum_{k=1}^{S}\sum_{l=1}^{S} p_k p_l d_{kl}^{\text{cat}},$$

ou

$$H_{\mathbf{\Delta}^{\text{cat}}}(\mathbf{p}) = \frac{1}{2}\sum_{k=1}^{S}\sum_{l=1}^{S} p_k p_l (\delta_{kl}^{\text{cat}})^2,$$

(cf. partie 2.4.2) vérifie les propriétés Ci pour tout i, c'est-à-dire, est une mesure parfaite de diversité (mesure parfaite signifiant ici complètement concave) dans le cas où $\mathbf{D}^{\text{cat}}$ est conditionnellement définie négative (Rao 1986), c'est à dire si $\mathbf{\Delta}^{\text{cat}}$ est euclidienne (cf. partie 2.4.2). Lorsque

$$\delta_{kl} = |y_k - y_l|,$$

où y_k et y_l sont les valeurs prises par une variable quantitative Y pour les catégories k et l, l'entropie quadratique est égale à la variance de Y. La matrice $\mathbf{\Delta}$ ainsi obtenue est euclidienne. La variance d'une variable quantitative vérifie donc les propriétés Ci pour tout i et est une mesure parfaite de diversité selon Rao. La variance d'une variable qualitative, qui a été nommée indice de Gini-Simpson depuis la partie 2.2 est aussi un cas particulier de l'entropie quadratique. L'indice de Gini-Simpson considère des dissimilarités uniformes entre catégories : $d_{kl} = 1$, *i.e.*, $\delta_{kl} = \sqrt{2}$ pour tout k et $l \neq k$. Toute matrice de dissimilarités uniformes est euclidienne. L'indice de Gini-Simpson vérifie donc bien les propriétés Ci pour tout i et est une mesure parfaite de diversité selon Rao.

L'ensemble APDIV ("apportionment of diversity", schéma hiérarchique) + ANODIV ("analysis of diversity", schéma croisé) est appelé DEDIV ("decomposition of diversity") (Rao 1982b) et constitue un schéma statistique fondamental pour la mesure de la diversité quel que soit le champ d'application de cette mesure.

3.3.2 Décomposition hiérarchique de l'entropie quadratique, l'APQE

La décomposition hiérarchique de l'entropie quadratique ou APQE (APportionment of Quadratic Entropy), correspond à l'application de l'APDIV sur l'entropie quadratique. L'APQE correspond exactement à la décomposition utilisée dans l'AMOVA (cf. partie 3.1.3). Considérons m_+ collections comprenant des entités regroupées en S catégories ; et r groupes, le groupe i comprenant m_i collections. Les fréquences relatives des catégories sont données par trois vecteurs :

- $\mathbf{p}_{ji}$ la distribution de fréquences des catégories dans la collection j du groupe i,
- $\mathbf{p}_{\bullet i}$ la distribution de fréquences des catégories dans le groupe i,
- $\mathbf{p}_{\bullet\bullet}$ la distribution globale de fréquences des catégories.

Des poids sont attribués aux collections et groupes :

- $\mu_i = \left(\mu_{1i}, ..., \mu_{ji}, ...\mu_{(m_+)i}\right)^t$, le vecteur de poids relatifs des collections dans le groupe i,
- $\lambda = (\lambda_1, ..., \lambda_i, ...\lambda_r)^t$, le vecteur de poids des groupes.

Dans l'AMOVA, ces poids correspondent aux tailles relatives des collections et groupes (en termes de proportions des entités présentes dans chaque collection et chaque groupe). Dans

l'APQE, ils peuvent être quelconques, c'est-à-dire que l'utilisateur de l'APQE est libre de les choisir, pourvu que $\sum_{j=1}^{m_i} \mu_{ji} = 1$ et $\sum_{i=1}^{r} \lambda_i = 1$.

L'APQE a la particularité d'unifier dans une même théorie les concepts de diversité et de dissimilarité (Rao 1982a). Cette unification est montrée grâce à la différence de Jensen d'ordre 1 :

$$D_\Delta(\mathbf{p},\mathbf{q}) = H_\Delta(\mathbf{p},\mathbf{q}) - \frac{1}{2}\left(H_\Delta(\mathbf{p}) + H_\Delta(\mathbf{q})\right),$$

où $\mathbf{\Delta} = [\delta_{kl}]$ est une matrice euclidienne de dissimilarités associée à $\mathbf{D} = [\delta_{kl}^2/2]$, $\mathbf{p}$ et $\mathbf{q}$ sont deux vecteurs appartenant à $\mathcal{P}$, $H_\Delta(\mathbf{p},\mathbf{q}) = \mathbf{p}^t\mathbf{D}\mathbf{q}$, $H_\Delta(\mathbf{p}) = \mathbf{p}^t\mathbf{D}\mathbf{p}$, et $H_\Delta(\mathbf{q}) = \mathbf{q}^t\mathbf{D}\mathbf{q}$. Notons $\mathbf{\Delta}^{\text{cat}} = \left[\delta_{kl}^{\text{cat}}\right]$ la matrice euclidienne des dissimilarités entre catégories. Cette différence de Jensen de premier ordre permet de calculer des dissimilarités entre collections et entre groupes :

$$\begin{aligned}
d_{ji,j'i'}^{\text{col}} &= D_{\Delta^{\text{cat}}}\left(\mathbf{p}_{ji},\mathbf{p}_{j'i'}\right) = H_{\Delta^{\text{cat}}}\left(\mathbf{p}_{ji},\mathbf{p}_{j'i'}\right) - \frac{1}{2}\left(H_{\Delta^{\text{cat}}}\left(\mathbf{p}_{ji}\right) + H_{\Delta^{\text{cat}}}\left(\mathbf{p}_{j'i'}\right)\right),\\
d_{ii'}^{\text{gro}} &= D_{\Delta^{\text{cat}}}(\mathbf{p}_{\bullet i},\mathbf{p}_{\bullet i'}) = H_{\Delta^{\text{cat}}}(\mathbf{p}_{\bullet i},\mathbf{p}_{\bullet i'}) - \frac{1}{2}\left(H_{\Delta^{\text{cat}}}(\mathbf{p}_{\bullet i}) + H_{\Delta^{\text{cat}}}(\mathbf{p}_{\bullet i'})\right),
\end{aligned}$$

Il est également intéressant de remarquer que

$$D_{\Delta^{\text{cat}}}(\mathbf{p}_{\bullet i},\mathbf{p}_{\bullet i'}) = D_{\Delta^{\text{col}}}(\mu_i,\mu_{i'})$$

Démonstration : Notons $\mathbf{D}^{\text{cat}} = \left[d_{kl}^{\text{cat}}\right] = \left[\left(\delta_{kl}^{\text{cat}}\right)^2/2\right]$,

$$\begin{aligned}
D_{\Delta^{\text{col}}}(\mu_i,\mu_{i'}) &= \sum_{jj'}\mu_{ji}\mu_{j'i'}\left(\mathbf{p}_{ji}^t\mathbf{D}^{\text{cat}}\mathbf{p}_{j'i'} - \frac{1}{2}\mathbf{p}_{ji}^t\mathbf{D}^{\text{cat}}\mathbf{p}_{ji} - \frac{1}{2}\mathbf{p}_{j'i'}^t\mathbf{D}^{\text{cat}}\mathbf{p}_{j'i'}\right)\\
&\quad -\frac{1}{2}\sum_{jj'}\mu_{ji}\mu_{j'i}\left(\mathbf{p}_{ji}^t\mathbf{D}^{\text{cat}}\mathbf{p}_{j'i} - \frac{1}{2}\mathbf{p}_{ji}^t\mathbf{D}^{\text{cat}}\mathbf{p}_{ji} - \frac{1}{2}\mathbf{p}_{j'i}^t\mathbf{D}^{\text{cat}}\mathbf{p}_{j'i}\right)\\
&\quad -\frac{1}{2}\sum_{jj'}\mu_{ji'}\mu_{j'i'}\left(\mathbf{p}_{ji'}^t\mathbf{D}^{\text{cat}}\mathbf{p}_{j'i'} - \frac{1}{2}\mathbf{p}_{ji'}^t\mathbf{D}^{\text{cat}}\mathbf{p}_{ji'} - \frac{1}{2}\mathbf{p}_{j'i'}^t\mathbf{D}^{\text{cat}}\mathbf{p}_{j'i'}\right)\\
&= \mathbf{p}_{\bullet i}^t\mathbf{D}^{\text{cat}}\mathbf{p}_{\bullet i'} - \frac{1}{2}\sum_j\mu_{ji}\mathbf{p}_{ji}^t\mathbf{D}^{\text{cat}}\mathbf{p}_{ji} - \frac{1}{2}\sum_{j'}\mu_{j'i'}\mathbf{p}_{j'i'}^t\mathbf{D}^{\text{cat}}\mathbf{p}_{j'i'}\\
&\quad -\frac{1}{2}\mathbf{p}_{\bullet i}^t\mathbf{D}^{\text{cat}}\mathbf{p}_{\bullet i} + \frac{1}{2}\sum_j\mu_{ji}\mathbf{p}_{ji}^t\mathbf{D}^{\text{cat}}\mathbf{p}_{ji}\\
&\quad -\frac{1}{2}\mathbf{p}_{\bullet i'}^t\mathbf{D}^{\text{cat}}\mathbf{p}_{\bullet i'} + \frac{1}{2}\sum_j\mu_{ji'}\mathbf{p}_{ji'}^t\mathbf{D}^{\text{cat}}\mathbf{p}_{ji'}\\
&= \mathbf{p}_{\bullet i}^t\mathbf{D}^{\text{cat}}\mathbf{p}_{\bullet i'} - \frac{1}{2}\mathbf{p}_{\bullet i}^t\mathbf{D}^{\text{cat}}\mathbf{p}_{\bullet i} - \frac{1}{2}\mathbf{p}_{\bullet i'}^t\mathbf{D}^{\text{cat}}\mathbf{p}_{\bullet i'}\\
&= D_{\Delta^{\text{cat}}}(\mathbf{p}_{\bullet i},\mathbf{p}_{\bullet i'})
\end{aligned}$$

Les dissimilarités entre groupes peuvent donc être calculées soit à partir des catégories soit à partir des collections, le résultat étant le même par les deux méthodes.

Notons $\delta_{ji,j'i'}^{\text{col}} = \sqrt{2d_{ji,j'i'}^{\text{col}}}$, $\delta_{ii'}^{\text{gro}} = \sqrt{2d_{ii'}^{\text{gro}}}$, $\mathbf{\Delta}^{\text{col}} = \left[\delta_{ji,j'i'}^{\text{col}}\right]$ est la matrice des dissimilarités δ^{col} entre collections et $\mathbf{\Delta}^{\text{gro}} = \left[\delta_{ii'}^{\text{gro}}\right]$ la matrice des dissimilarités δ^{gro} entre groupes.

Les composants de l'APQE sont

$$H_{\Delta^{\text{cat}}}(\mathbf{p}_{\bullet\bullet}) = \sum_{i=1}^{r} \lambda_i \sum_{j=1}^{m_i} \mu_{ji} H_{\Delta^{\text{cat}}}\left(\mathbf{p}_{ji}\right) + \sum_{i=1}^{r} \lambda_i H_{\Delta^{\text{col}}}(\mu_i) + H_{\Delta^{\text{gro}}}(\lambda).$$

Si $\mathbf{\Delta}^{\text{cat}}$ est euclidienne alors $\mathbf{\Delta}^{\text{col}}$ et $\mathbf{\Delta}^{\text{gro}}$ sont également euclidiennes (Rao et Nayak 1985, Champely et Chessel 2002). Donc si $\mathbf{\Delta}^{\text{cat}}$ est euclidienne $H_{\Delta^{\text{cat}}}$, $H_{\Delta^{\text{col}}}$ et $H_{\Delta^{\text{gro}}}$ sont toutes concaves, c'est pourquoi la décomposition peut ainsi être continuée à un nombre quelconque de niveaux de groupement.

L'APQE généralise l'analyse de variance (ANOVA) (Fisher 1925) et l'analyse de variance sur variable catégorielle (CATANOVA) (Light et Margolin 1971). La CATANOVA correspond à la décomposition de l'indice de Gini-Simpson. L'APQE est égale à la CATANOVA lorsque la différence d_{kl} entre deux catégories k et l ($l \neq k$) est toujours égale à 1. Elle est égale à l'ANOVA lorsque la différence δ_{kl} entre deux catégories k et l est mesurée comme la valeur absolue de la différence entre les valeurs y_k et y_l prises par une variable quantitative Y pour les deux catégories (cf. partie 2.4.2) $|y_k - y_l|$ (Nayak 1983). Dans le cas de la CATANOVA, la dissimilarité $\delta_{ii'}$ entre deux collections i et i' de distributions de fréquences $\mathbf{p}_i$ et $\mathbf{p}_{i'}$ est égale à

$$\delta_{ii'}^{\text{col}} = \sqrt{\sum_{k=1}^{S} (p_{ki} - p_{ki'})^2}.$$

Dans le cas de l'ANOVA, la dissimilarité $\delta_{ii'}^{\text{col}}$ entre deux collections i et i' de distributions de fréquences $\mathbf{p}_i$ et $\mathbf{p}_{i'}$ est égale à

$$\delta_{ii'}^{\text{col}} = |\sum_{k=1}^{S} p_{ki} y_k - \sum_{k=1}^{S} p_{ki'} y_k|.$$

Il s'agit de la valeur absolue de la différence entre les valeurs moyennes prises par la variable Y dans les deux collections.

La décomposition de l'indice de Gini-Simpson correspond à la décomposition de variance utilisée dans la CATANOVA. Ainsi la décomposition de la diversité allélique de Nei et l'analyse de variance de Weir et Cockerham (cf. partie 3.1.2) sont des cas particuliers de la CATANOVA et donc de l'APQE. L'analyse de diversité en microsatellites de Michalakis et Excoffier (1996), qui dérive de l'analyse de Slatkin (1995) est une utilisation de l'ANOVA sur des données moléculaires ; la variable quantitative étudiée est la taille de microsatellites. L'APQE dans sa version la plus générale est de plus très utilisée en génétique à travers l'AMOVA. Son utilisation en écologie est beaucoup plus rare. Nous proposons, dans l'article Pavoine et Dolédec (2005, annexe2), une illustration écologique montrant les changements de diversité, en termes d'habitudes alimentaires, de tailles de corps et de positions dans la taxonomie, d'espèces de trichoptères et coléoptères le long de la Loire.

Ces décompositions de diversité sont descriptives. Pour tester les différences entre collections, Excoffier *et al.* (1992) proposent des schémas de permutations (cf. partie 3.1.3). Nayak

(1986a, b) propose une solution paramétrique asymptotique (i.e. pour un grand nombre d'entités) dans le cas où les vecteurs $\mathbf{v}_i = (n_{1i}, ..., n_{ki}, ..., n_{Si})^t$ (où n_{ki} est l'abondance de la catégorie k dans la collection i pour $i = 1, ..., r$) sont indépendants et distribués selon des lois multinomiales de paramètres n_{+i} et $\Pi_i = (\pi_{1i}, ..., \pi_{ki}, ..., \pi_{Si})^t$. Soient $\hat{W}$, $\hat{B}$, et $\hat{T}$ les composants de diversité respectivement intra-collection, inter-collections et totale, mesurés sur des échantillons à l'aide d'une fonction H. Considérons l'hypothèse $H_0 : \Pi_1 = ... = \Pi_r$. Nayak (1986b) note que si $\Pi_1 = ... = \Pi_r$ alors $B = 0$, mais que l'inverse est vrai si et seulement si H est concave. Il démontre que sous H_0, $\hat{T}$ et $\hat{B}$ sont asymptotiquement indépendamment distribués. Il affirme donc que, contrairement à l'ANOVA, la statistique $\hat{B}/\hat{T}$, plutôt que $\hat{B}/\hat{W}$, devrait être utilisée en inférence statistique. Il démontre également que, pour l'entropie quadratique, la distribution de

$$(S - 1)(n_{++} - 1)\frac{\hat{B}}{\hat{T}}$$

peut être approchée par la loi de probabilité $\chi^2_{(r-1)(S-1)}$. Ces résultats peuvent être étendus à la décomposition hiérarchique de l'entropie quadratique.

Lorsque les distributions des vecteurs $\mathbf{v}_i$ sont inconnues, Liu et Rao (1995) proposent d'utiliser le bootstrap. Soit $X_i = \{X_{li}, l = 1, ..., n_{+i}\}$ l'échantillon de la collection i. La distribution empirique de cette collection est $\mathbf{p}_i = (n_{1i}/n_{++}, ..., n_{ki}/n_{++}, ..., n_{Si}/n_{++})$. Des échantillons de bootstrap $X_i^* = \{X_{li}^*, l = 1, ..., n_{+i}\}$ sont tirés selon cette distribution. Soit $X = (X_1, ..., X_r)$ l'ensemble des échantillons observés sur l'ensemble des collections, et $X^* = \left(X_1^*, ..., X_r^*\right)$ les échantillons de bootstrap également sur l'ensemble des collections. Notons $d\left(X_{li}, Y_{mj}\right)$ la dissimilarité entre l'entité l de l'échantillon X_i et l'entité m de l'échantillon Y_j mesurée par la dissimilarité entre les catégories auxquelles appartiennent ces entités. Liu et Rao définissent

$$B(X, X) = \frac{1}{n_{++}^2}\sum_{i=1}^{r}\sum_{i'=1}^{r}\sum_{l=1}^{n_i}\sum_{l'=1}^{n_i'} d(X_{li}X_{l'i'}) - \frac{1}{n_{++}}\sum_{i=1}^{r}\frac{1}{n_{+i}}\sum_{l=1}^{n_i}\sum_{l'=1}^{n_i} d(X_{li}X_{l'i}).$$

Ils démontrent que la distribution de $n_{++}B(X, X) = n_{++}\hat{B}$ peut être approchée par la distribution de bootstrap de la statistique

$$n_{++}\left[B(X^*, X^*) - 2B(X, X^*) + B(X, X)\right].$$

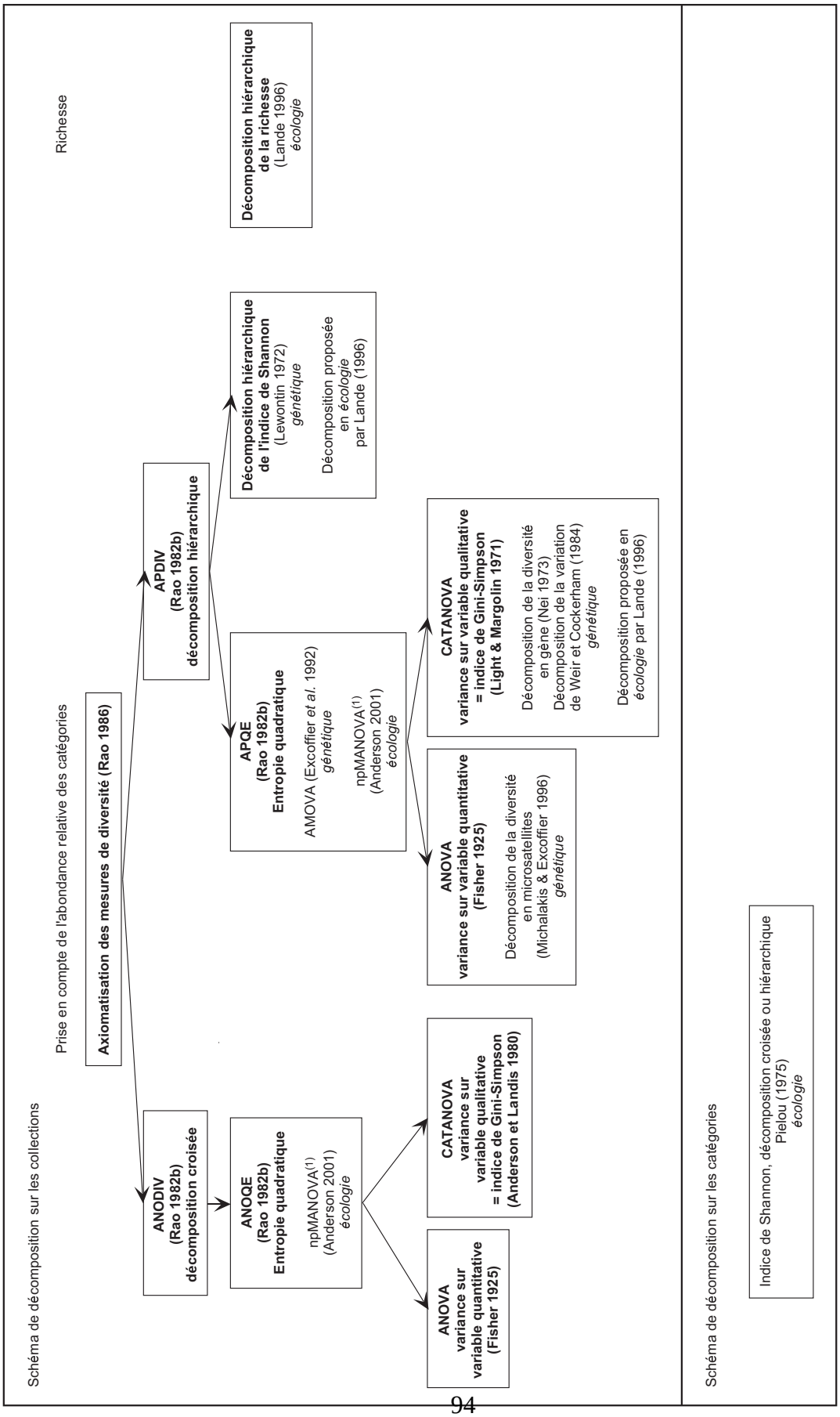
Cette méthode est conceptuellement différente de celle des schémas de permutation d'Excoffier *et al.* (1992) puisque dans l'AMOVA les individus sont permutés entre collections (populations) alors qu'ici les rééchantillonnages se font dans chaque collection.

3.3.3 Bilan sur les liens

Un bilan sur les liens entre les méthodes de décomposition de variation incluant diversité et variance est proposé dans la figure 10.

Les méthodes sont d'abord divisées en deux grands groupes :

– le premier groupe contient les méthodes basées sur un schéma de décomposition sur les *collections* ;



(1) pas de considération de la diversité intra-collections

FIG. 10 -- Bilan sur les liens entre méthodes de décomposition de la variation. Chaque flèche part d'une méthode générale pour arriver à un cas particulier. Tous ces liens concernent uniquement la décomposition de la diversité ou plus généralement de la variation. Ils ne concernent en aucun cas les procédures de tests qui dépendent étroitement à la fois de la méthode et des données.

FIG. *10 –*

– le deuxième groupe contient les méthodes basées sur un schéma de décomposition sur les *catégories*.

Ensuite, le premier groupe comprend les décompositions des indices basés entre autres sur des distributions de fréquences d'une part et la décomposition de la richesse d'autre part. La génétique est très représentée dans l'ensemble de ces méthodes. Rao (1982a) s'est basé sur les mesures de diversité génétique de Nei pour développer sa théorie.

Tous ces liens concernent uniquement les schémas de décomposition de la diversité ou de la variance. Les méthodes diffèrent cependant par les procédures de tests qui leur ont été associés et par la définition des poids des collections, groupes, etc.

Pour les tests, d'autres études pourront être faites pour comparer les méthodes de permutations d'Excoffier *et al.* (1992) et Anderson (2001) aux méthodes de bootstrap de Liu et Rao (1995) comme cela a été fait par exemple pour l'ANOVA (ter Braak 1992). Anderson (2001) affirme qu'une des qualités de la npMANOVA est que les pseudo-quotients F sont égaux aux statistiques F de Fisher lorsqu'une seule variable est considérée et que la distance euclidienne est utilisée. Cependant, la différence entre les statistiques de la CATANOVA ("Inter"/"Total") et celle de l'ANOVA ("Inter"/"Intra") pour obtenir des quotients de deux quantités non corrélées, prouve que les statistiques doivent dépendre des données et que, sur le choix de la statistique, il sera difficile de généraliser.

Pour l'ensemble de ces méthodes, la partie inférentielle, quand elle existe, a pour but de tester l'existence de différences entre collections ou groupes de collections. Dans l'ANOVA, les tests sont basés sur trois hypothèses : (1) indépendance des entités, (2) variable étudiée de loi normale, (3) homoscédasticité entre collections. Regardons cette troisième hypothèse. Pour l'ensemble des méthodes de la figure 10, lorsque des différences entre collections ou groupes sont mises en évidence, ces différences peuvent être dues soit à des différences de diversité ou de variance, soit à des différences de compositions (inertie *versus* position dans un espace pour les matrices de dissimilarités euclidiennes, cf. partie 4 ; variance *versus* moyenne pour l'ANOVA) (Anderson 2001). Faisant cette observation pour l'AMOVA, Stewart et Excoffier (1996) proposent alors un test d'homogénéité de variance moléculaire : HOMOVA ("homogeneity of molecular variance"). Ce test a été développé à partir du test d'homogénéité de variance de Bartlett (1937). Il est formulé en termes de somme des déviations au carré :

$$B = \frac{(N-P)\ln\left(\frac{SSD(T)}{N-P}\right) - \sum_{i=1}^{P}(N_i-1)\ln\left(\frac{SSD(WP)_i}{N_i-1}\right)}{1+\frac{1}{3(P-1)}\left(\sum_{i=1}^{P}\frac{1}{N_i-1}-\frac{1}{N-P}\right)},$$

où P est le nombre de populations, N_i le nombre d'individus dans la population i, N le nombre total d'individus, $SSD(T)$ la somme des déviations au carré sur l'ensemble des populations mélangées, et $SSD(WP)_i$ la somme des déviations au carré dans la population i. Si des distances euclidiennes entre variables aléatoires Gaussiennes étaient considérées pour définir des dissimilarités entre individus, alors B suivrait une distribution du chi-deux avec $P-1$ degrés de liberté. Dans l'AMOVA, ce n'est pas le cas, Stewart et Excoffier proposent donc d'utiliser les mêmes approches de permutations que pour les composants de variances, en l'occurrence une

permutation des individus entre populations.

A propos des poids des collections, Nei (1973, 1987) affirme que des poids égaux doivent être donnés aux différentes populations parce que les généticiens ne s'intéressent pas à la taille des populations quand il s'agit de comparer les compositions génétiques de ces populations. Cependant d'autres suggèrent que les tailles relatives des populations doivent constituer les poids des populations dans la décomposition de la diversité, car la mesure de diversité inter-populations peut être considérablement sur-estimée si les populations sont de tailles très différentes et malgré tout uniformément pondérées (Finkeldey 1994).

3.4 P ´ -

Nous allons juste effleurer le domaine de l'analyse spatiale de la biodiversité à travers une méthode faisant intervenir la décomposition de la diversité à deux niveaux : entre sites géographiques et entre niveaux taxonomiques. Pour cette étude, faite en collaboration avec Raphaël Pélissier, nous sommes partis de l'article de Couteron et Pélissier (2004) sur la décomposition additive de la diversité spécifique.

La problématique est la suivante. On s'intéresse à r sites répartis dans l'espace. Il peut s'agir par exemple d'une région étudiée par l'intermédiaire de r sites échantillonnés. Pour chaque site, on connaît la liste des espèces présentes pour un certain niveau taxonomique, par exemple les Trichoptères, et aussi la fréquence relative de chaque espèce. Par les indices de diversité vus précédemment nous pouvons, avec ces données, estimer la diversité totale de la région, la diversité dans chaque site, ainsi qu'une diversité inter-sites qui est, avec certains indices de diversité seulement, calculée à partir de mesures de dissimilarités entre sites. Nous allons justement analyser ici ces dissimilarités entre sites pour répondre aux questions suivantes. Ces dissimilarités dépendent-elles de l'éloignement géographique ? Des sites très proches se ressemblent-ils ? et des sites plus éloignés sont-ils très différents ?

3.4.1 Dissimilarité spécifique et distance spatiale entre sites

Conformément aux notations précédentes, prenons

- $\mathbf{p}_i = (p_{1i}, ..., p_{ki}, ..., p_{Si})$ la distribution de fréquences des espèces dans le site i ;
- λ_i un poids attribué au site i ;
- $\mathbf{p}_\bullet = \sum_{i=1}^{r} \lambda_i \mathbf{p}_i$ la distribution globale de fréquences des espèces sur l'ensemble des sites.

Dans l'article de Couteron et Pélissier (2004), la diversité totale d'une région est définie par

$$\mathrm{TD} = \sum_{k=1}^{S} w_k p_{k\bullet} (1 - p_{k\bullet}) \tag{3.6}$$

où w_k est le poids que nous souhaitons donner à l'espèce k dans la quantification de la diversité régionale. Cet indice correspond à la richesse lorsque $w_k = 1/p_{k\bullet}$, à l'indice de Shannon quand $w_k = \ln(1/p_{k\bullet})/(1 - p_{k\bullet})$ et à celui de Gini-Simpson si $w_k = 1$ (Pélissier *et al.* 2003).

Le composant de variance contenu dans TD, $\mathrm{SV}(k) = p_{k\bullet}(1 - p_{k\bullet})$, peut être décomposé en variance intra-site et variance inter-sites : $\mathrm{SVwc}(k) = \sum_{i=1}^{r} \lambda_i p_{ki}(1 - p_{ki})$ et $\mathrm{SVac}(k) =$

$\sum_{i=1}^{r} \lambda_i (p_{ki} - p_{k\bullet})^2$. La diversité totale TD peut ainsi être décomposée en diversité intra-site et diversité inter-sites (les lettres "wc" et "ac" signifient "within- and among-classes portions") :

$$\text{TDwc} = \sum_{k=1}^{S} w_k \sum_{i=1}^{r} \lambda_i p_{ki} (1 - p_{ki}), \tag{3.7}$$

$$\text{TDac} = \sum_{k=1}^{S} w_k \sum_{i=1}^{r} \lambda_i (p_{ki} - p_{k\bullet})^2. \tag{3.8}$$

Ce dernier composant peut être réécrit de la façon suivante

$$\text{TDac} = \sum_{k=1}^{S} w_k \frac{1}{2} \sum_{i=1}^{r} \sum_{i'=1}^{r} \lambda_i \lambda_{i'} (p_{ki} - p_{ki'})^2 = \frac{1}{2} \sum_{i=1}^{r} \sum_{i'=1}^{r} \lambda_i \lambda_{i'} \sum_{k=1}^{S} w_k (p_{ki} - p_{ki'})^2$$

Couteron et Pélissier (2004) proposent de prendre

$$d_{ii'}^{\text{sit}} = \sum_{k=1}^{S} w_k (p_{ki} - p_{ki'})^2$$

comme mesure du contraste entre les sites i et i'.

Soient $d_{(i,i')}$ la distance spatiale entre les sites i et i', et C_c une classe de distance spatiale centrée sur c par exemple $C_{40\text{km}} =]0\text{km}, 80\text{km}]$. Pour chaque classe de distance spatiale, Couteron et Pélissier (2004) calculent la dissimilarité moyenne entre sites

$$D(c) = \frac{\sum\limits_{d(i,i') \in C_c} \lambda_i \lambda_{i'} d_{ii'}^{\text{sit}}}{\sum\limits_{d(i,i') \in C_c} \lambda_i \lambda_{i'}}. \tag{3.9}$$

Cette formule a été inspirée par le principe du variogramme (Ver Hoef *et al.* 1993).

Elle dépend des données mais aussi du choix des poids attribués aux espèces. Quelles significations peut-on donner aux composants de diversité obtenus selon les différentes valeurs de w_k ?

Prenons $w_k = 1$. Avec ce choix, les calculs font appel à l'indice de Gini-Simpson. Les composants de diversité TD (eq. 3.6), TDwc (eq. 3.7), et TDac (eq. 3.8) prennent les valeurs suivantes :

$$\text{TD} = H_{\text{G-S}}(p_\bullet).$$

La diversité intra-site est égale à la moyenne des diversités au sein des sites :

$$\text{TDwc} = \sum_{i=1}^{r} \lambda_i H_{\text{G-S}}(\mathbf{p}_i),$$

où $\mathbf{p}_i$ est la distribution de fréquences des espèces dans le site i. Et la diversité inter-sites est égale à la dissimilarité moyenne entre les sites :

$$\text{TDac} = \sum_{i=1}^{r}\sum_{i'=1}^{r} \lambda_i \lambda_{i'} d_{ii'}^{\text{sit}}.$$

Il s'agit donc d'une décomposition additive de l'indice de Gini-Simpson telle qu'elle apparaît dans la CATANOVA (cf. partie 3.3.2).

Couteron et Pélissier (2004) ont introduit les deux autres choix de w_k, $w_k = \ln(1/p_{k\bullet})/(1-p_{k\bullet})$ et $w_k = 1/p_{k\bullet}$, pour intégrer la richesse et l'indice de Shannon dans un schéma similaire à celui de Gini-Simpson. Malheureusement, la méthode utilisée nécessite un calcul de la diversité inter-sites sous une forme quadratique à partir des dissimilarités entre sites, et ce calcul est possible seulement dans le cas où l'indice choisi pour mesurer la diversité peut être mis sous la forme de l'entropie quadratique. C'est le cas pour l'indice de Gini-Simpson (cf. partie 2.4.2 page 50) mais ce n'est pas le cas pour la richesse ni pour l'indice de Shannon. Avec les trois pondérations choisies par Couteron et Pélissier (2004) la diversité totale TD est égale à la richesse ou à l'indice de Shannon ou à l'indice de Gini-Simpson. Mais que l'on prenne $w_k = \ln(1/p_{k\bullet})/(1-p_{k\bullet})$ (de sorte que TD corresponde à l'indice de Shannon cf. page 98) ou $w_k = 1/p_{k\bullet}$ (de sorte que TD corresponde à la richesse cf. page 98), la décomposition reste celle d'une variance sur variable catégorielle (indice de Gini-Simpson), c'est-à-dire que les composants TDwc et TDac sont basés sur l'indice de Gini-Simpson pondéré et non sur l'indice de Shannon ou sur la richesse. Les décompositions de la richesse et de l'indice de Shannon proposées par l'intermédiaire de ces pondérations ne sont pas équivalentes à celles développées par Lande (1996) et Rao (1986). Une propriété essentielle leur manque : la diversité intra-site doit être la moyenne des valeurs prises par l'indice choisi à l'intérieur des sites (Rao 1986, Lande 1996).

Peut-on maintenant proposer un autre type de décomposition pour l'indice de Shannon en suivant la théorie de Rao ? Du point de vue de Rao et Nayak (1985), tout indice de diversité concave H peut être décomposé sous la forme

$$H(\mathbf{p}_\bullet) = \sum_{i=1}^{r} \lambda_i H(\mathbf{p}_i) + \sum_{i=1}^{r} \lambda_i C(\mathbf{p}_i, \mathbf{p}_\bullet),$$

où $C(\mathbf{p}_i, \mathbf{p}_\bullet)$ est une mesure de la différence entre i et un site théorique moyen. Or la richesse, et les indices de Shannon et de Gini-Simpson sont concaves. Lorsque C est une fonction symétrique ($C(\mathbf{p},\mathbf{q}) = C(\mathbf{q},\mathbf{p})$), cette formule peut être réécrite sous la forme

$$H(\mathbf{p}_\bullet) = \sum_{i=1}^{r} \lambda_i H(\mathbf{p}_i) + \sum_{i=1}^{r}\sum_{i'=1}^{r} \lambda_i \lambda_{i'} D(\mathbf{p}_i, \mathbf{p}_{i'}),$$

où $D(\mathbf{p}_i, \mathbf{p}_{i'}) = H\left(\frac{1}{2}\mathbf{p}_i + \frac{1}{2}\mathbf{p}_{i'}\right) - \frac{1}{2}\left(H(\mathbf{p}_i) + H(\mathbf{p}_{i'})\right)$. Rao et Nayak (1985) ont démontré que C est une fonction symétrique si et seulement si H peut être mise sous la forme de l'entropie quadratique. Parmi les indices précédents, seul l'indice de Gini-Simpson est un cas particulier de l'entropie quadratique.

Pour l'indice de Shannon,

$$C_S(\mathbf{p}_i, \mathbf{p}_\bullet) = \sum_{i=1}^{r} p_{ki} \log\left(\frac{p_{ki}}{p_{k\bullet}}\right).$$

Cette fonction, appelée "mesure d'information de Kullback-Leibler" ou "mesure de divergence de Kullback-Leibler" , n'est pas symétrique. Lorsque C n'est pas une fonction symétrique, il est toujours possible de la "symétriser" en remarquant que

$$\sum_{i=1}^{r} \lambda_i C(\mathbf{p}_i, \mathbf{p}_\bullet) = \sum_{i=1}^{r} \sum_{i'=1}^{r} \lambda_i \lambda_{i'} \left[\frac{1}{2} C(\mathbf{p}_i, \mathbf{p}_\bullet) + \frac{1}{2} C(\mathbf{p}_{i'}, \mathbf{p}_\bullet)\right].$$

On obtient donc

$$D^*(\mathbf{p}_i, \mathbf{p}_{i'}) = \frac{1}{2} C(\mathbf{p}_i, \mathbf{p}_\bullet) + \frac{1}{2} C(\mathbf{p}_{i'}, \mathbf{p}_\bullet).$$

Cependant cette mesure présente des inconvénients. Une mesure de divergence nécessite un objet de référence. Elle évalue l'éloignement d'une distribution par rapport à une distribution de référence. Ici cette référence est $\mathbf{p}_\bullet$, la distribution moyenne. Il en découle que si $\mathbf{p}_i$ et $\mathbf{p}'_i$ sont très différentes de la distribution théorique moyenne $\mathbf{p}_\bullet$ alors la valeur de $D^*(\mathbf{p}_i, \mathbf{p}_{i'})$ sera élevée même si ces deux distributions sont très proches l'une de l'autre. En particulier si $\mathbf{p}_i \neq \mathbf{p}_\bullet$ alors $D^*(\mathbf{p}_i, \mathbf{p}_i) \neq 0$. D^* n'est donc pas une mesure de dissimilarité entre deux distributions.

D'un autre côté pour l'établissement d'une formule similaire à la formule 3.9 mais adaptée à d'autres indices que celui de Gini-Simpson, il est possible de calculer l'expression

$$\sum_{i=1}^{r} \sum_{i'=1}^{r} \lambda_i \lambda_{i'} D(\mathbf{p}_i, \mathbf{p}_{i'}) \tag{3.10}$$

pour tout indice de diversité pourvu qu'il soit concave et ceci indépendamment de la décomposition additive de la diversité. Cependant il faut garder en mémoire que la valeur obtenue est égale à la diversité inter selon la décomposition de Rao uniquement pour l'entropie quadratique (indice de Gini-Simpson inclus). En particulier elle n'est pas égale à la diversité inter pour la richesse et pour l'indice de Shannon.

Les formules 3.9 et 3.10 servent à tracer une courbe (voir par exemple la figure 11 page 104) montrant les changements de la dissimilarité moyenne entre sites en fonction de la distance spatiale entre ces mêmes sites. Pour construire cette courbe, des classes de distances spatiales sont définies et la valeur moyenne de dissimilarité entre paires de sites éloignés par une distance comprise dans une classe est associée au centre de cette classe. Couteron et Pélissier (2004) proposent alors de calculer une "enveloppe de confiance" autour de cette courbe. Cette enveloppe contient 95% des valeurs obtenues après permutations aléatoires des compositions faunistiques ou floristiques des sites. Pour les classes de distances spatiales correspondant aux endroits où la courbe dépasse l'enveloppe au dessus ou en dessous, Couteron et Pélissier (2004) concluent que les dissimilarités moyennes observées entre les sites en question sont significativement plus grandes ou plus petites que dans des répartitions aléatoires. Les étapes de la définition de cette enveloppe sont les suivantes :

- Etape 1 : Permutation des compositions des sites ;
- Etape 2 : Calcul des dissimilarités moyennes pour chaque classe de distance (formules 3.9 page 99) ;
- Etape 3 : Recommencer les étapes 1 et 2 pour obtenir 1000 permutations en comptant la composition observée des sites.
- Etape 4 : Pour chaque classe de distance, calculer et ordonner les 1000 valeurs théoriques obtenues pour $D(c)$. Les bornes de l'enveloppe pour chaque classe de distance sont alors les $25^{\text{ième}}$ et $975^{\text{ième}}$ valeur de ces ordinations.

3.4.2 Décomposition spatiale de l'entropie quadratique

La relation, proposée par P. Couteron et R. Pélissier, entre la répartition spatiale des sites et les mesures de dissimilarités entre ces sites calculées à partir d'indices de diversité, est étroitement liée à une décomposition de variance sur variable catégorielle, CATANOVA. Nous allons l'étendre à la décomposition de l'entropie quadratique. Cette extension permettra de prendre en compte, dans l'analyse du profil spatial de la diversité, des estimations de dissimilarités (phylogénétiques ou fonctionnelles) entre espèces.

La diversité totale devient

$$\text{TD} = \sum_{k=1}^{S}\sum_{l=1}^{S} p_{k\bullet}\, p_{l\bullet}\, d_{kl}^{\text{esp}} = H_{\mathbf{D}^{\text{esp}}}(\mathbf{p}_\bullet),$$

où d_{kl}^{esp} est une mesure de la dissimilarité entre deux espèces. Elle est égale à la somme de la diversité intra-site

$$\text{TDwc} = \sum_{i=1}^{r} \lambda_i H_{\mathbf{D}^{\text{esp}}}(\mathbf{p}_i)$$

et de la diversité inter-sites

$$\text{TDac} = \sum_{i=1}^{r}\sum_{i'=1}^{r} \lambda_i \lambda_{i'} D(\mathbf{p}_i, \mathbf{p}_{i'}),$$

où

$$\begin{aligned} D(\mathbf{p}_i, \mathbf{p}_{i'}) &= H_{\mathbf{D}^{\text{esp}}}\left(\frac{1}{2}\mathbf{p}_i + \frac{1}{2}\mathbf{p}_{i'}\right) - \frac{1}{2}\left(H_{\mathbf{D}^{\text{esp}}}(\mathbf{p}_i) + H_{\mathbf{D}^{\text{esp}}}(\mathbf{p}_{i'})\right) \\ &= -\frac{1}{4}(\mathbf{p}_i - \mathbf{p}_{i'})^t\, \mathbf{D}^{\text{esp}}(\mathbf{p}_i - \mathbf{p}_{i'}) \end{aligned}$$

La dissimilarité moyenne entre sites pour chaque classe de distances devient

$$D(c) = \frac{\sum\limits_{d(i,i')\in C_c} \lambda_i \lambda_{i'} D(\mathbf{p}_i, \mathbf{p}_{i'})}{\sum\limits_{d(i,i')\in C_c} \lambda_i \lambda_{i'}}.$$

Le calcul de l'enveloppe de confiance peut se faire de la même façon que précédemment.

Appliquons cette formule aux assemblages de Trichoptères et Coléoptères le long de la Loire. Les données ont été publiées par Ivol *et al.* (1997) et nous les avons réanalysées dans le cadre de la décomposition de la diversité dans l'article Pavoine et Dolédec (2005, cf. annexe 2). Reprenons ici ces données pour illustrer les concepts précédents avec la répartition spatiale des communautés de macroinvertébrés le long de la Loire. Trente-huit sites ont été échantillonnés d'amont en aval le long des 1012km de la Loire. Deux stations situées immédiatement après un barrage ont été éliminées pour l'analyse. Quarante espèces de Trichoptères et Coléoptères ont été observées. Quatre matrices de dissimilarités $\mathbf{D}$ entre espèces sont considérées : $\mathbf{D}^{Uni}$ dissimilarités uniformes, $\mathbf{D}^{Tai}$ dissimilarités entre espèces selon la taille moyenne de leurs corps, $\mathbf{D}^{Ali}$ dissimilarités entre espèces selon leurs habitudes alimentaires et $\mathbf{D}^{Tax}$ dissimilarités taxonomiques entre espèces. Pour le calcul de ces matrices, $\mathbf{D}^{Uni} = \mathbf{1}\mathbf{1}^t - \mathbf{I}$. La dissimilarité entre deux espèces dans $\mathbf{D}^{Tax}$ est le nombre de niveaux taxonomiques qui les séparent (1 pour deux espèces du même genre, 2 pour deux espèces de genres différents mais de même famille, etc.). Pour les deux autres matrices $\mathbf{D}^{Tai}$ et $\mathbf{D}^{Ali}$, les données sont des variables floues. La variable "taille" est divisée en cinq classes allant de $\leq$ 5mm à $\geq$ 40mm. Pour une espèce donnée, des pourcentages appelés pourcentages d'affinité, attribués à ces classes indiquent la distribution de tailles des individus de cette espèce. De la même façon, la variable "habitude alimentaire" est divisée en sept classes : avaleurs, déchiqueteurs, gratteurs, mangeurs de dépôts, filtreurs actifs, filtreurs passifs, perceurs. Pour une espèce donnée, les pourcentages d'affinité attribués à ces classes indiquent les habitudes alimentaires des individus de cette espèce. La métrique de distance utilisée pour ces deux variables floues est celle de Manly (1994, fomule 5.8 p. 68) :

$$d_{kl}^{\text{esp}} = 1 - \frac{\sum_{m=1}^{M} q_{km} q_{lm}}{\sqrt{\sum_{m=1}^{M} q_{km} \sum_{m=1}^{M} q_{lm}}}$$

où M est le nombre de modalités (5 pour la variable "taille" et 7 pour la variable "habitude alimentaire"), et q_{km} et q_{lm} sont les pourcentages d'affinité respectivement de l'espèce k et de l'espèce l pour la modalité m de la variable étudiée ("taille" ou "habitude alimentaire").

Les résultats sont donnés dans la figure 11. Les courbes montrant les changements de la dissimilarité moyenne entre sites $D(c)$ en fonction de la classe de distance c, calculée avec les mesures de diversité de tailles, d'habitudes alimentaires, et selon la taxonomie, sont assez semblables. D'après l'enveloppe de confiance, selon ces trois indices de diversité, seule la dissimilarité moyenne entre sites distants de plus de 805.5km serait plus grande qu'attendue aléatoirement. Inversement, pour l'ensemble des indices de diversité, seule la dissimilarité moyenne entre sites distants de moins de 80.5km serait significativement plus petite qu'on pourrait si attendre aléatoirement.

Ces développements sont une étape pour montrer qu'il est possible d'étendre des méthodes d'analyse spatiale actuellement basées sur la variance à l'utilisation de mesures de diversité.

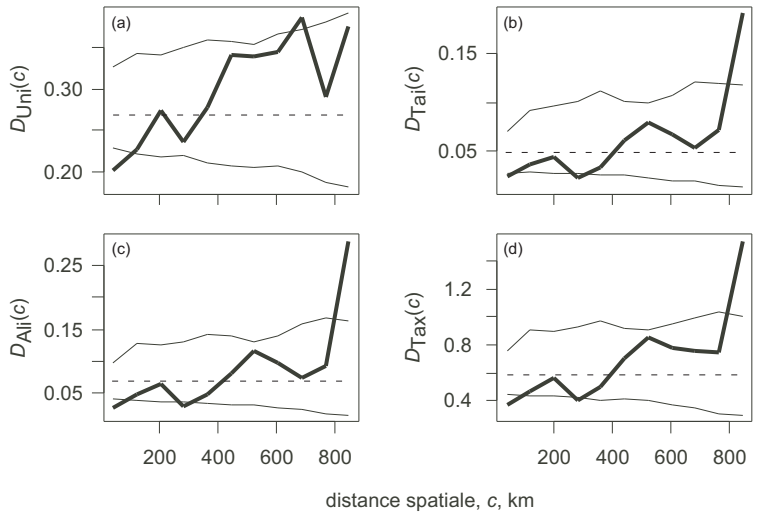


FIG. *11 – Dissimilarité moyenne entre sites D*(*c*) *en fonction de c, distance spatiale entre les sites (lignes épaisses) selon l'entropie quadratique mesurée avec des dissimilarités (a) uniformes ; (b) basées sur les tailles moyennes des espèces ; (c) calculées selon les habitudes alimentaires des espèces ; (d) taxonomiques. Les lignes continues fines délimitent les enveloppes de confiance. La droite discontinue indique la dissimilarité moyenne entre tous les sites sans considération des distances spatiales.*

3.4.3 Dissimilarité taxonomique et distance spatiale

La taxonomie est une classification des êtres vivants qui comprend plusieurs niveaux : espèce, genre, famille, etc. La diversité taxonomique peut être décomposée additivement selon ces différents niveaux.

Appelons x_{ki} la fréquence de l'espèce k dans le site i. Soit $\mathbf{x}_i = (x_{1i}, ..., x_{ki}, ..., x_{Si})^t$. Nous pouvons calculer l'indice de Gini-Simpson, $H_{\text{G-S}}(\mathbf{x}_i) = 2\sum_{k>l} x_{ki}x_{li}$, et la décomposition

$$H_{\text{G-S}}(\mathbf{x}_\bullet) = \sum_{i=1}^{r} \lambda_i H_{\text{G-S}}(\mathbf{x}_i) + \sum_{i=1}^{r}\sum_{i'=1}^{r} \lambda_i \lambda_{i'} D_{\text{G-S}}(\mathbf{x}_i, \mathbf{x}_{i'})$$

Mais ces calculs ne tiennent pas compte de la diversité taxonomique. Supposons que les espèces considérées pour une analyse font partie d'une même classe divisée en trois niveaux taxonomiques : sous-classe, famille et genre. La diversité taxonomique peut être mesurée par l'entropie quadratique de Rao :

$$H_{\text{Tax}}(\mathbf{x}_i) = 2\sum\nolimits_{k>l} x_{ki}x_{li}d_{kl}$$

où d_{kl} est la distance taxonomique entre k et l (c'est-à-dire, le nombre de niveaux taxonomiques qui les séparent). Nous avons alors la décomposition associée suivante :

$$H_{\text{Tax}}(\mathbf{x}_\bullet) = \sum_{i=1}^{r} \lambda_i H_{\text{Tax}}(\mathbf{x}_i) + \sum_{i=1}^{r}\sum_{i'=1}^{r} \lambda_i \lambda_{i'} D_{\text{Tax}}(\mathbf{x}_i, \mathbf{x}_{i'})$$

Soient $\mathbf{g}_i$, $\mathbf{f}_i$ et $\mathbf{s}_i$ les distributions de fréquences dans le site i respectivement des genres, familles et sous-classes ; et soient $\mathbf{g}_\bullet$, $\mathbf{f}_\bullet$ et $\mathbf{s}_\bullet$ les distributions moyennes de fréquences dans la région toujours pour, respectivement, les genres, familles et sous-classes. D'après Shimatani (2001),

$$H_{\text{Tax}}(\mathbf{x}_\bullet) = H_{\text{G-S}}(\mathbf{x}_\bullet) + H_{\text{G-S}}(\mathbf{g}_\bullet) + H_{\text{G-S}}(\mathbf{f}_\bullet) + H_{\text{G-S}}(\mathbf{s}_\bullet)$$

$$H_{\text{Tax}}(\mathbf{x}_i) = H_{\text{G-S}}(\mathbf{x}_i) + H_{\text{G-S}}(\mathbf{g}_i) + H_{\text{G-S}}(\mathbf{f}_i) + H_{\text{G-S}}(\mathbf{s}_i)$$

et ce qui nous intéresse plus particulièrement,

$$D_{\text{Tax}}(\mathbf{x}_i, \mathbf{x}_{i'}) = D_{\text{G-S}}(\mathbf{x}_i, \mathbf{x}_{i'}) + D_{\text{G-S}}(\mathbf{g}_i, \mathbf{g}_{i'}) + D_{\text{G-S}}(\mathbf{f}_i, \mathbf{f}_{i'}) + D_{\text{G-S}}(\mathbf{s}_i, \mathbf{s}_{i'})$$

Ainsi, dans la formule du pseudo-variogramme appliquée à l'entropie quadratique et à des différences taxonomiques entre espèces

$$D_{\text{Tax}}(c) = \frac{\overbrace{\sum_{d(i,i')\in C_h} \lambda_i \lambda_{i'} D_{\text{Tax}}(\mathbf{x}_i, \mathbf{x}_{i'})}^{\text{diversité taxonomique inter-sites}}}{\sum_{d(i,i')\in C_h} \lambda_i \lambda_{i'}},$$

quatre composants peuvent être distingués :

$$D_{\text{Tax}}(c) = \frac{\overbrace{\sum_{d(i,i')\in C_h} \lambda_i \lambda_{i'} D_{\text{G-S}}(\mathbf{x}_i, \mathbf{x}_{i'})}^{\text{diversité en espèces inter-sites}}}{\sum_{d(i,i')\in C_h} \lambda_i \lambda_{i'}} + \frac{\overbrace{\sum_{d(i,i')\in C_h} \lambda_i \lambda_{i'} D_{\text{G-S}}(\mathbf{g}_i, \mathbf{g}_{i'})}^{\text{diversité en genres inter-sites}}}{\sum_{d(i,i')\in C_h} \lambda_i \lambda_{i'}}$$
$$+ \frac{\overbrace{\sum_{d(i,i')\in C_h} \lambda_i \lambda_{i'} D_{\text{G-S}}(\mathbf{f}_i, \mathbf{f}_{i'})}^{\text{diversité en familles inter-sites}}}{\sum_{d(i,i')\in C_h} \lambda_i \lambda_{i'}} + \frac{\overbrace{\sum_{d(i,i')\in C_h} \lambda_i \lambda_{i'} D_{\text{G-S}}(\mathbf{s}_i, \mathbf{s}_{i'})}^{\text{diversité en sous-classes inter-sites}}}{\sum_{d(i,i')\in C_h} \lambda_i \lambda_{i'}}.$$

Cette méthode d'analyse du profil spatial de la diversité à différents niveaux taxonomiques est appliquée à l'étude de la diversité taxonomique de la flore d'une région de Guyane française : Counami. Deux classifications ont été étudiées : celle de Cronquist (1981) (Fig. 12a) et celle de la dernière mise à jour de la phylogénie des angiospermes ("Angiosperm Phylogeny Group", APG II) (Bremer *et al.* 2003) (Fig. 12b). Lors de l'échantillonnage, 291 taxa avec des noms vernaculaires ont été observés. Parmi ces 291 taxa, seuls les 59 correspondant à des espèces botaniques connues ont été retenus.

Les changements de dissimilarité taxonomique moyenne entre sites en fonction des distances spatiales séparant ces sites sont assez semblables aux changements de dissimilarité spécifique moyenne (Fig. 13). Selon la distribution de fréquences des espèces et leurs liens taxonomiques, la dissimilarité moyenne entre sites distants de moins de 4.8km est significativement

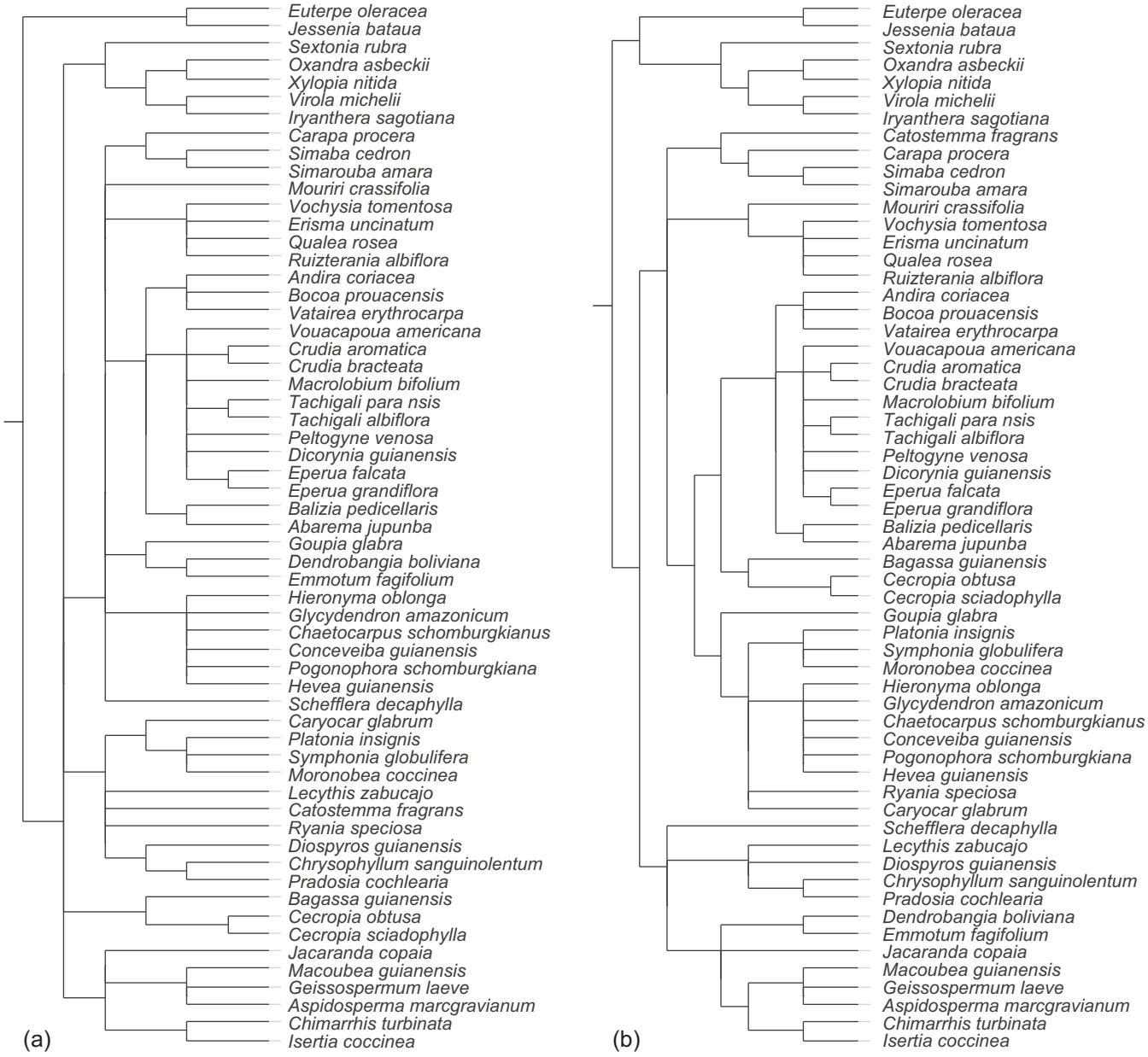


FIG. *12 – Taxonomie des 59 espèces selon (a) Cronquist (1981) et (b) la dernière mise à jour de la classification des angiosperms (APG II, "Angiosperm Phylogeny Group II") (Bremer* et al. *2003).*

plus petite que celle que l'on attendrait aléatoirement. Inversement la dissimilarité moyenne entre sites séparés par une distance comprise entre 6.2 et 10.4km serait plus grande qu'attendue aléatoirement. Des analyses complémentaires ont montré qu'une grande part des différences entre sites est due à l'ordre Fabales qui contient l'espèce la plus commune *Eperua falcata* (Couteron *et al.* 2003). *Eperua falcata* a été rencontrée 2179 fois lors de l'échantillonnage alors que les autres espèces ont, en moyenne, été rencontrées 86 fois chacune.

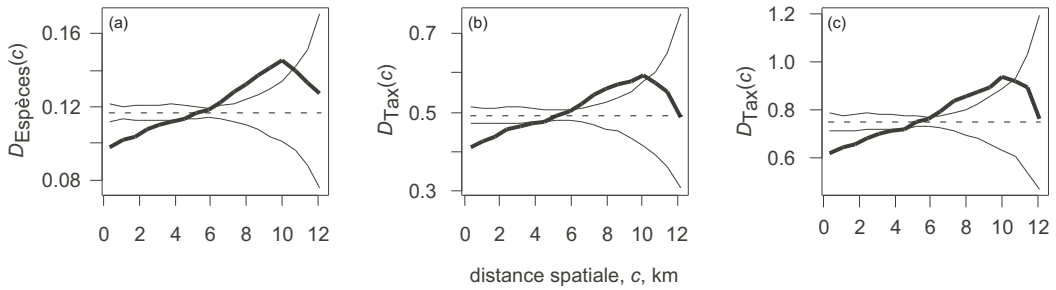


FIG. *13 – Dissimilarité moyenne entre sites, D*(*c*) *en fonction de c, distance spatiale entre les sites (lignes épaisses) (a) avec l'indice de Gini-Simpson ; (b) avec l'entropie quadratique mesurant la diversité taxonomique selon Cronquist (1981) ; (c) avec l'entropie quadratique mesurant la diversité taxonomique selon l'APG II (Bremer* et al. *2003). Les lignes continues fines délimitent les enveloppes de confiance. La droite discontinue indique la dissimilarité moyenne entre tous les sites sans considération des distances spatiales.*

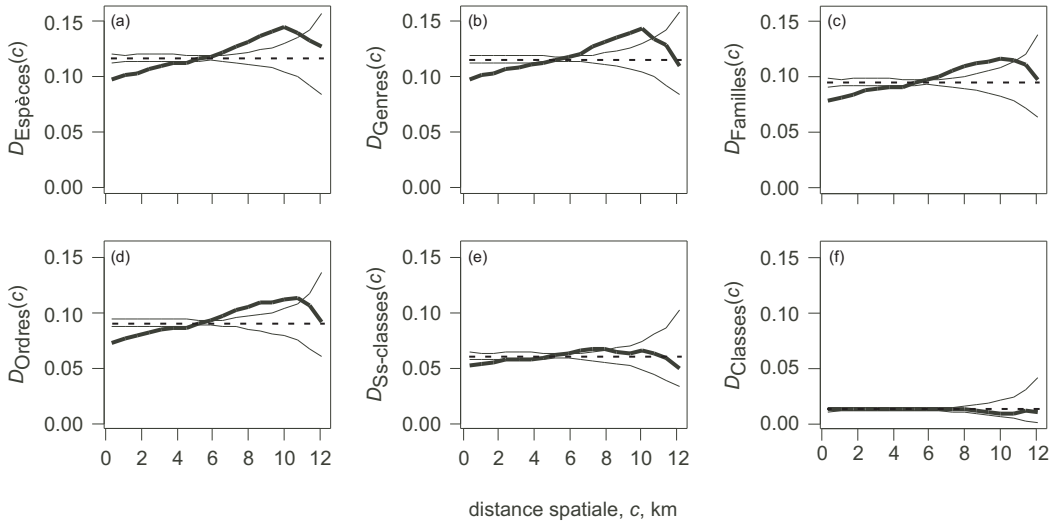


FIG. *14 – Dissimilarité moyenne entre sites, D*(*c*) *en fonction des distances spatiales entre les sites, c (lignes épaisses). D*(*c*) *est mesuré par l'indice de Gini-Simpson à partir de la liste des (a) espèces ; (b) genres ; (c) familles ; (d) ordres ; (e) sous-classes ; (f) classes (Selon Cronquist, 1981) (Pour chaque classe de distance, la somme des ordonnées (courbes aux lignes épaisses) de ces six figures est égale à la celle de la figure 13b). Les lignes continues fines délimitent les enveloppes de confiance. La droite discontinue indique la dissimilarité moyenne entre tous les sites sans considération des distances spatiales.*

Les pseudo-variogrammes obtenus avec la diversité taxonomique mesurée par l'entropie quadratique peuvent être décomposés par niveaux de classification (Fig. 14 et 15). Prenons d'abord la taxonomie de Cronquist. Les pseudo-variogrammes basés sur l'indice de Gini-Simpson à différents niveaux taxonomiques sont très semblables depuis le niveau de l'espèce jusqu'à celui de l'ordre. Le profil spatial de la diversité inter-sites est donc ici peu changé, que l'on échantillonne au niveau des espèces, des genres, des familles ou des ordres. Pour les sous-classes et classes, les formes des pseudo-variogrammes changent légèrement, mais elles influent peu

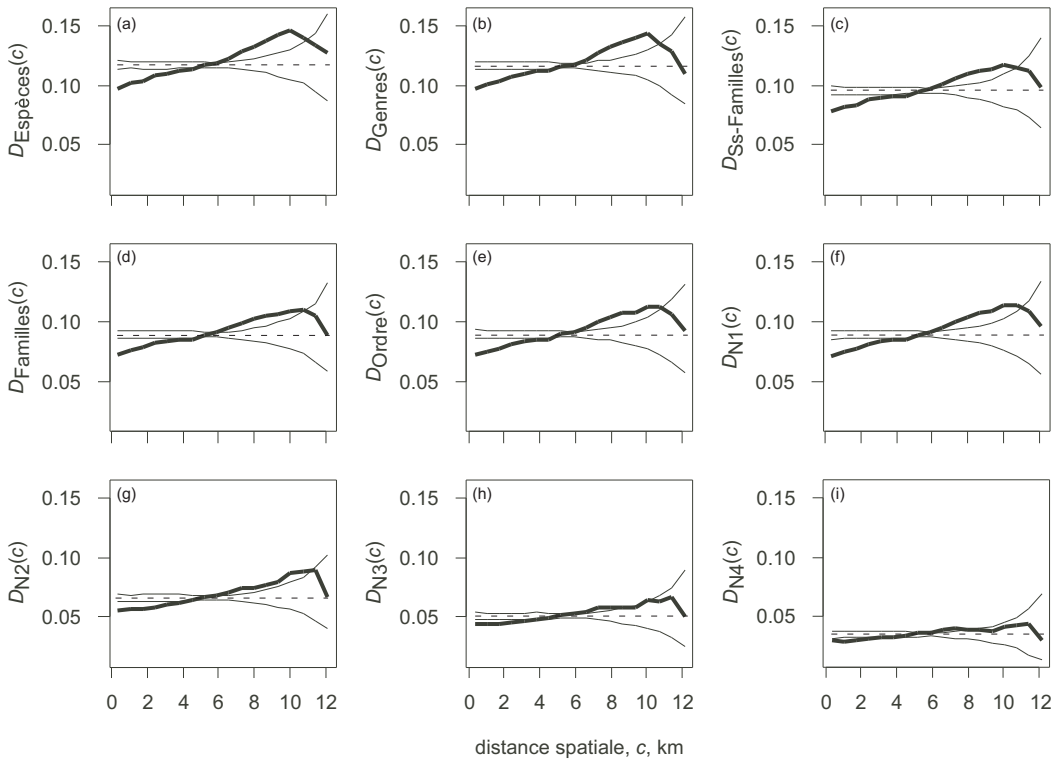


FIG. *15 – Dissimilarité moyenne entre sites, D*(*c*) *en fonction des distances spatiales entre les sites, c (lignes épaisses). D*(*c*) *est mesuré par l'indice de Gini-Simpson à partir de la liste des (a) espèces ; (b) genres ; (c) sous-familles ; (d) familles ; (e) ordres ; (f) (g) (h) (i) taxa de quatre autres niveaux taxonomiques (Selon l'APG II) (Pour chaque classe de distance, la somme des ordonnées (courbes aux lignes épaisses) de ces neuf figures est égale à la celle de la figure 13c). Les lignes continues fines délimitent les enveloppes de confiance. La droite discontinue indique la dissimilarité moyenne entre tous les sites sans considération des distances spatiales.*

sur le résultat final (Fig. 13b) car les dissimilarités entre sites selon ces deux niveaux taxonomiques s'amenuisent. En partie parce qu'il n'y a que deux classes, les dissimilarités entre sites selon leurs compositions en classes sont beaucoup plus faibles que celles calculées à partir des espèces, genres, familles ou ordres. Aucune dissimilarité moyenne observée entre sites selon leurs compositions en classes n'est significativement éloignée d'un modèle aléatoire. Cette décomposition taxonomique nous apprend que le profil de diversité taxonomique inter-sites obtenu dans la figure 13b n'est pas un compromis entre des profils différents pour chaque niveau taxonomique, mais bien un profil supporté par les niveaux taxonomiques, au moins jusqu'à l'ordre.

Les résultats obtenus avec l'APGII sont assez semblables, les pseudo-variogrammes basés sur l'indice de Gini-Simpson à différents niveaux taxonomiques se ressemblent jusqu'au niveau N1 juste après l'ordre. Ensuite les dissimilarités entre sites s'amenuisent. Le profil spatial des dissimilarités entre sites qui existe au niveau des ordres se répercute aux niveaux taxonomiques inférieures (espèce, genre, sous-famille et famille).

3.5 P

En conclusion, dans ce chapitre nous avons montré que l'axiomatisation de Rao permet de rassembler dans un schéma mathématique six décompositions additives de variation (diversité et variance) développées en génétique ou en écologie,

- la npMANOVA (Anderson 2001, écologie), correspondant aux analyses de Pillar et Orlóci (1996, écologie) et de Smith *et al.* (1990, écologie) ;
- l'AMOVA (Excoffier *et al.* 1992, génétique) ;
- la décomposition de la diversité en microsatellites (Michalakis et Excoffier 1996, génétique) ;
- la décomposition de la diversité en gènes (Nei 1973, génétique) ;
- la décomposition de la variation de Weir et Cockerham (Weir et Cockerham 1984, génétique) ;
- la décomposition additive de l'indice de Gini-Simpson (Lande 1996, écologie) ;

ainsi que deux analyses développées en statistiques :

- l'analyse de la variance (ANOVA) (Fisher 1925) ;
- l'analyse de la variance d'une variable catégorielle (CATANOVA)(Light et Margolin 1971).

La décomposition additive de la variation a sur la décomposition multiplicative l'avantage de fournir des composants de variation tous mesurés avec la même unité, donc directement comparables. Une telle décomposition peut avoir un objectif descriptif (décrire, par son évaluation, la variation à plusieurs échelles : décrire par exemple les changements spatiaux ou temporels d'une diversité spécifique) ou plus souvent un objectif inférentiel (tester les différences entre collections ou groupes de collections, ou encore tester les effets de facteurs (expérimentaux ou environnementaux par exemple) sur la variation).

Enfin, les développements que nous avons réalisés à partir du variogramme montrent que des méthodes d'analyse spatiale actuelles prenant, pour indice de variation, la variance d'une variable quantitative peuvent être étendues à d'autres indices pour l'étude spatiale de la diversité.

Chapitre 4

Description des différences entre collections

Sommaire

Résumé

Ce chapitre fait une revue des différentes fonctions permettant de calculer des dissimilarités entre deux collections (e.g. deux populations ou deux sites). Ces mesures vont dépendre des données. Cependant, nous constatons que bien des mesures sont communes entre génétique et écologie même si leurs noms diffèrent. Trois types de fonctions sont considérés : celles qui sont basées sur les présences/absences des catégories, celles qui sont basées sur les abondances relatives des catégories et celles qui tiennent compte des différences entre les catégories.

Les qualités et les limites des représentations traditionnelles des dissimilarités entre collections sont discutées. Alors apparaît une limite importante dans l'étude de la diversité : ces représentations traditionnelles ne se concentrent que sur les collections, puisque, sur ces représentation, on ne retrouve pas les catégories. Or ce sont pourtant ces catégories qui ont permis de calculer les dissimilarités entre collections.

Nous introduisons alors une nouvelle méthode d'ordination que nous appelons Double Analyse en Coordonnées Principale (DPCoA). La DPCoA permet une superposition de la typologie des collections et de celle des catégories dans un même espace euclidien. Nous montrons que la représentation obtenue est en lien direct avec la décomposition de l'entropie quadratique, qui se révèle être une mesure d'inertie dans un espace euclidien. Nous montrons également que cette méthode généralise cinq méthodes d'ordination très utilisées en analyse statistique descriptive et multivariée.

A partir de la DPCoA, nous développons trois extensions, permettant d'introduire une structuration des collections selon des facteurs croisés et selon des facteurs hiérarchiques, et permettant aussi de considérer l'analyse simultanée de plusieurs schémas de DPCoA. Ces extensions nous permettent d'affirmer que nous avons un schéma statistique, dépendant de l'axiomatisation de Rao et du schéma de dualité, qui réunit les concepts de diversité, inertie, dissimilarité, ordination et typologie.

Nous venons de voir un ensemble de méthodes permettant de décomposer la diversité et aussi de tester l'existence de différences significatives entre collections. Une fois que l'on a déterminer l'existence effective de différences entre collections, les analyses précédentes ne nous permettent pas de décrire la typologie de ces différences. Nous allons, dans ce chapitre, étudier l'analyse graphique des différences entre collections d'entités (e.g. communautés en écologie et populations en génétique) en liaison directe avec la mesure de la diversité.

> "The problem of measuring diversity can be viewed as characterizing an aspect of the distribution of points in space." (Solow et Polasky 1994)

4.1 M ´

Le but de cette partie n'est pas de faire une longue liste de tous les indices développés pour mesurer la quantité de différences entre collections, mais plutôt de montrer les liens entre les développements écologiques et génétiques, d'étudier les propriétés de ces indices, et aussi de comprendre pourquoi ils ont été développés. Des listes plus complètes pourront être trouvées dans Legendre et Legendre (1998) et Magurran (2004). Rappelons qu'en écologie chaque collection correspond généralement à une communauté ou à un assemblage d'espèces ; et qu'en génétique chaque collection peut être une population, un dème ou une colonie, les catégories étant alors des allèles.

4.1.1 Indices basés sur les présences/absences des catégories

Soient deux collections i et j prises dans un ensemble de collections à comparer. Notons a le nombre total de catégories présentes à la fois dans les deux collections, b le nombre de catégories présentes dans i mais pas dans j, c le nombre de catégories présentes dans j mais pas dans i et d le nombre de catégories absentes des deux collections. Ce nombre d est calculé par rapport à l'ensemble des collections entre lesquelles des dissimilarités sont calculées. Ces mesures peuvent être résumées dans le tableau suivant :

		j = 1	j = 0	
i	1	a	b	$a + b$
	0	c	d	$c + d$
		$a + c$	$b + d$	

Les indices symétriques suivants ont été développés en *écologie* pour mesurer la similarité entre i et j.

L'indice de Jaccard (1901)

$$S_1 = \frac{a}{a + b + c}$$

est le plus connu et le plus utilisé.

L'indice de Sokal et Michener (1957)

$$S_2 = \frac{a + d}{a + b + c + d}$$

est une modification de l'indice de Jaccard, pour tenir compte des catégories non présentes dans les deux collections comparées mais présentes dans d'autres collections comparables.

Les indices de Sokal et Sneath (1963)

$$S_3 = \frac{a}{a + 2(b + c)}$$

et de Rogers et Tanimoto (1960) (voir Jackson *et al.* 1989)

$$S_4 = \frac{a + d}{a + 2(b + c) + d}$$

sont des modifications respectivement des indices S_1 et S_2, pour donner deux fois plus de poids aux différences (mesures b et c) qu'aux similitudes (mesures a et d).

A l'inverse, l'indice développé par Dice (1945) et Sørensen (1948)

$$S_5 = \frac{2a}{2a + b + c}$$

est une modification de l'indice S_1 pour donner deux fois plus de poids aux similitudes qu'aux différences.

Le coefficient de Hamman, Gower et Legendre (1986)

$$S_6 = \frac{a - (b + c) + d}{a + b + c + d}$$

est égal à $2S_4 - 1$.

Les indices d'Ochiai (1957)

$$S_7 = \frac{a}{\sqrt{(a + b)(a + c)}},$$

de Sokal et Sneath (1963)

$$S_8 = \frac{ad}{\sqrt{(a + b)(a + c)(d + b)(d + c)}},$$

et le Phi de Pearson (Yule 1912, Jackson *et al.* 1989)

$$S_9 = \frac{ad - bc}{\sqrt{(a + b)(a + c)(d + b)(d + c)}}$$

comparent les ressemblances entre les deux collections aux marges du tableau.

Russell et Rao (1940) (voir Jackson *et al.* 1989) ont proposé une modification de S_1 l'indice de Jaccard

$$S_{10} = \frac{a}{a+b+c+d}$$

pour tenir compte du nombre d de catégories absentes des deux collections.

La mesure de dissimilarité D la plus utilisée correspondant à chacune de ces mesures S de similarité est égale à $1-S$. Elle est métrique mais non-euclidienne pour les indices S_1 à S_4, S_6 et S_{10} (Gower et Legendre 1986). $D_1 = 1 - S_1$ est connu comme indice de Marczewski-Steinhaus (Magurran 2004). Pour les autres indices, la mesure D n'est pas métrique. Par contre, pour tous ces 10 indices, une matrice de dissimilarité construite avec la fonction $\sqrt{1-S}$ est euclidienne (Gower et Legendre 1986).

Nous avons vu que l'indice de Sørensen (1948) S_5 est une modification de l'indice S_1 pour donner deux fois plus de poids aux similitudes qu'aux différences. Il existe la même modification pour l'indice S_2 (Legendre et Legendre 1998) :

$$S_{11} = \frac{2a+2d}{2a+(b+c)+2d}.$$

La mesure de dissimilarité $1 - S_{11}$ n'est pas métrique, alors que $\sqrt{1-S_{11}}$ l'est (Gower et Legendre 1986). Par contre une matrice de dissimilarité construite avec la fonction $\sqrt{1-S_{11}}$ n'est pas euclidienne (Gower et Legendre 1986).

Sokal et Sneath (1963) mentionnent d'autres indices tels que

$$S_{12} = \frac{1}{2}\left(\frac{a}{a+b} + \frac{a}{a+c}\right),$$

$$S_{13} = \frac{1}{4}\left(\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{c+d} + \frac{d}{b+d}\right),$$

$$S_{14} = \frac{ad-bc}{ad+bc}.$$

Pour ces trois indices ni $1 - S$, ni $\sqrt{1-S}$ ne sont métriques.

Beaucoup d'autres indices ont été développés (voir liste dans Legendre et Legendre 1998). Il s'agit en général de modifications des indices précédents. Par exemple, Lennon *et al.* (2001) utilisent la mesure de dissimilarité

$$D_{15} = 1 - \frac{a}{a + min\,(b,c)},$$

dans le but de diminuer l'impact d'une forte différence de richesse en catégories entre les deux collections comparées.

La dissimilarité entre deux collections est estimée à partir d'échantillons. Un des problèmes posés par ces échantillonnages est la présence de catégories rares dans les collections, c'est-à-dire de ces catégories qui ont le moins de chance d'être observées dans les échantillons. Toutes les mesures de dissimilarités basées sur des données de présence/absence ont de fortes chances de surestimer la dissimilarité entre deux collections si ces deux collections partagent beaucoup de catégories rares, parce que les catégories rares qui apparaissent dans un échantillon ont une forte probabilité d'être différentes des catégories rares observées dans l'autre échantillon. Au contraire, la dissimilarité peut être sous-estimée si deux collections se ressemblent par quelques catégories très abondantes qu'elles partagent et diffèrent par des catégories rares, chacune présente dans l'une mais pas dans l'autre des collections. Se basant sur des observations de données écologiques réelles et sur des simulations montrant que la situation amenant à une surestimation de la dissimilarité est la plus fréquente, Chao *et al.* (2005) proposent des estimateurs des indices de Jaccard et Sørensen incorporant l'effet des catégories partagées par les deux collections mais non observées. Les données qu'ils ont considérées sont des assemblages (collections) d'espèces (catégories).

Soient S_{12} la probabilité qu'une espèce soit présente dans deux assemblages 1 et 2, S_1 la probabilité qu'elle existe dans l'assemblage 1 et S_2 la probabilité qu'elle existe dans l'assemblage 2. S_{12}/S_2 est alors la probabilité pour qu'une espèce soit présente dans l'assemblage 1 sachant qu'elle est présente dans l'assemblage 2 et inversement S_{12}/S_1 est la probabilité pour qu'une espèce soit présente dans l'assemblage 2 sachant qu'elle est présente dans l'assemblage 1. Parallèlement aux nombres *a*, *b*, *c* et *d* des indices ci-dessus, Chao *et al.* (2005) proposent des nombres notés *A*, *B*, *C* et *D* basés sur les probabilités conditionnelles S_{12}/S_1 et S_{12}/S_2. On tire une espèce à partir de l'assemblage 1 et une autre à partir de l'assemblage 2. La valeur du paramètre *A* est la probabilité que les deux espèces tirées soient chacune présente dans les deux assemblages ; *B* est la probabilité que l'espèce tirée de l'assemblage 2 soit partagée par les deux assemblages, mais pas celle tirée de l'assemblage 1 ; *C* est la probabilité que l'espèce tirée de l'assemblage 1 soit partagée par les deux assemblages, mais pas celle tirée de l'assemblage 2 ; et *D* est la probabilité qu'aucune de ces deux espèces ne soit partagée par les deux assemblages. Ces valeurs sont résumées dans le tableau suivant :

	Espèce tirée de l'assemblage 2	
	Partagée	Non partagée
Espèce tirée de l'assemblage 1		
Partagée	$A=\frac{S_{12}}{S_1}\frac{S_{12}}{S_2}$	$B=\frac{S_{12}}{S_1}\left(1-\frac{S_{12}}{S_2}\right)$
Non partagée	$C=\left(1-\frac{S_{12}}{S_1}\right)\frac{S_{12}}{S_2}$	$D=\left(1-\frac{S_{12}}{S_1}\right)\left(1-\frac{S_{12}}{S_2}\right)$

Les nouvelles écritures des indices de Jaccard et Sørensen sont alors

$$S_1^* = \frac{A}{A+B+C} = \frac{S_{12}}{S_1+S_2-S_{12}}$$
$$S_5^* = \frac{2A}{2A+B+C} = \frac{2S_{12}}{S_1+S_2}$$

Une espèce est en général considérée comme rare si 1, 2 voire 10 au plus individus de cette espèce sont observés dans un échantillon, ou si cette espèce n'est trouvée que dans 1, 2 voire 10 au plus échantillons d'un assemblage. Pour prendre en compte les espèces rares partagées non observées, il est nécessaire de déterminer quelles espèces sont rares et donc d'avoir des données indiquant l'abondance de ces espèces. Dans les formules S_1^* et S_5^*, les espèces sont considérées de façon égale. Dans le cas de données d'abondance, ce sont les individus qui sont considérés de façon égale. Les paramètres *A*, *B*, *C* et *D* sont alors réécrits au niveau des individus. Par exemple si un individu est tiré de l'assemblage 1 et un autre de l'assemblage 2, *A* devient la probabilité que ces deux individus appartiennent à des espèces partagées par les deux assemblages. Soient *U* l'abondance relative (fréquence) totale, dans l'assemblage 1, des espèces partagées, et *V* l'abondance relative (fréquence) totale, dans l'assemblage 2, des espèces partagées. Avec ces notations les nouveaux paramètres *A*, *B*, *C* et *D* sont résumés dans le tableau suivant :

	Individu tiré de l'assemblage 2	
	Espèce partagée	Espèce non partagée
Individu tiré de l'assemblage 1		
Espèce partagée	$A=UV$	$B=U(1-V)$
Espèce non partagée	$C=(1-U)V$	$D=(1-U)(1-V)$

Les nouvelles écritures des indices de Jaccard et Sørensen sont alors

$$S_1^{**} = \frac{A}{A+B+C} = \frac{UV}{U+V-UV}$$
$$S_5^{**} = \frac{2A}{2A+B+C} = \frac{2UV}{U+V}$$

Les valeurs de ces deux indices sont comprises entre 0 et 1.

Pour pouvoir estimer les paramètres *U* et *V* avec des données de présence/absence, il est nécessaire d'obtenir plusieurs échantillons d'un même assemblage. Considérons *w* échantillons d'un assemblage *X* et *z* échantillons d'un assemblage *Y*. L'abondance relative d'une espèce *k* dans les assemblages *X* et *Y* est estimée par le nombre relatif d'échantillons dans lesquels elle est observée :

$$X_k = \sum_{i=1}^{w} x_{ki} \text{ et } Y_k = \sum_{i=1}^{z} y_{ki},$$

où x_{ki} et y_{ki} sont égaux à 1 lorsque l'espèce est présente dans l'échantillon *i* des assemblages respectivement *X* et *Y* et à 0 sinon. Soient les notations suivantes :

- *S* le nombre total d'espèces observées,
- D_{12} le nombre d'espèces observées qui sont partagées entre les deux assemblages,
- f_{1+} le nombre d'espèces partagées qui apparaissent dans un seul échantillon de *X*, et f_{+1} le nombre d'espèces partagées qui apparaissent dans un seul échantillon de *Y*,
- f_{2+} le nombre d'espèces partagées qui apparaissent dans exactement deux échantillons de *X*, et f_{+2} le nombre d'espèces partagées qui apparaissent dans exactement deux échantillons de *Y*,

– $n = \sum_{k=1}^{S} X_k$ et $m = \sum_{k=1}^{S} Y_k$
– I (*condition*) = 1 si "*condition*" est vraie et 0 sinon.

Les estimateurs de U et V sont alors (Chao *et al.* 2005) :

$$\hat{U} = \sum_{k=1}^{D_{12}} \left[\frac{X_k}{n} + \frac{z-1}{z} \frac{f_{+1}}{2 f_{+2}} \left(\frac{X_k}{n} I(Y_k = 1) \right) \right]$$

$$\hat{V} = \sum_{k=1}^{D_{12}} \left[\frac{Y_k}{m} + \frac{w-1}{w} \frac{f_{+1}}{2 f_{+2}} \left(\frac{Y_k}{m} I(X_k = 1) \right) \right]$$

Chao *et al.* (2005) proposent également d'autres estimateurs pour les cas où des données d'abondance seraient disponibles.

Tous ces indices basés sur des données de présence/absence sont beaucoup plus utilisés en écologie où les abondances des espèces sont souvent difficilement estimables, qu'en génétique où, à l'inverse, sont préférés des indices basés sur les fréquences des allèles.

4.1.2 Indices basés sur les abondances des catégories

Considérons deux collections possédant respectivement les distributions de fréquences $\mathbf{p} = (p_1, ..., p_k, ..., p_S)^t$ et $\mathbf{q} = (q_1, ..., q_k, ..., q_S)^t$. Les fonctions suivantes ont été développées en écologie et en génétique pour mesurer la dissimilarité entre deux collections :

La mesure de dissimilarité

$$D_1(\mathbf{p}, \mathbf{q}) = \frac{1}{2} \sum_{k=1}^{S} |p_k - q_k|$$

est utilisée par de nombreux auteurs à la fois en écologie et génétique. Son introduction en écologie semble être due à Gleason (1920) (cf. Whittaker 1960). En génétique, elle est connue comme "distance génétique absolue" de Gregorius (*e.g.*, Laval *et al.* 2002), et a été également proposée par Hattemer (1982) et Prevosti *et al.* (1975) (Nei 1987). D_1 prend ses valeurs entre 0 et 1 et est métrique (Hattemer 1982). Elle est proportionnelle à la fonction connue en mathématique sous le nom de "métrique de Manhattan"

$$D_2(\mathbf{p}, \mathbf{q}) = \sum_{k=1}^{S} |p_k - q_k|.$$

L'utilisation de la métrique de Manhattan a aussi été proposée en taxonomie par Sokal et Michener (1957) et en écologie par Faith *et al.* (1987). Notons que Southwood et Henderson (2000) (Magurran 2004) proposent d'associer à cette métrique la mesure de similarité suivante

$$S(\mathbf{p}, \mathbf{q}) = 100 \left(1 - \frac{1}{2} \sum_{k=1}^{S} |p_k - q_k| \right).$$

La distance euclidienne

$$D_3(\mathbf{p},\mathbf{q}) = \sqrt{\sum_{k=1}^{S}(p_k - q_k)^2}$$

a été utilisée sur des données génétiques par Gower en 1972 et par Goodman en 1973 (Laval *et al.* 2002). Elle a aussi été proposée pour des données écologiques et taxonomiques (Sokal et Sneath 1963, Faith *et al.* 1987). Puisque cette fonction prend ses valeurs entre 0 et $\sqrt{2}$, Rogers (1972) la propose sous la forme :

$$D_4(\mathbf{p},\mathbf{q}) = \sqrt{\frac{1}{2}\sum_{k=1}^{S}(p_k - q_k)^2}$$

dont les valeurs sont comprises entre 0 et 1.

La distance minimale (Nei 1987), proposée en génétique,

$$D_5(\mathbf{p},\mathbf{q}) = \frac{1}{2}\sum_{k=1}^{S}(p_k - q_k)^2$$

est le carré de la métrique D_4, et n'est pas, elle-même, métrique.

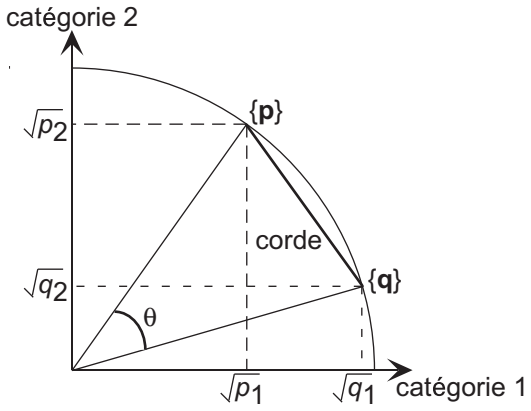


FIG. *16 – Représentation selon Bhattacharyya de deux distributions observées* ***p*** *et* ***q*** *avec deux catégories.*

En statistique, Bhattacharyya (1946) introduit la notion de distance angulaire en considérant deux collections multinomiales caractérisées par deux vecteurs de fréquences $(\pi_1, ..., \pi_S)^t$ et $\left(\pi'_1, ..., \pi'_S\right)^t$. Les deux vecteurs $\left(\sqrt{\pi_1}, ..., \sqrt{\pi_S}\right)^t$ et $\left(\sqrt{\pi'_1}, ..., \sqrt{\pi'_S}\right)^t$ peuvent être considérés comme les directions de deux lignes partant de l'origine d'un espace multidimensionnel et séparées par un angle θ dont le cosinus est (Fig. 16) :

$$\cos\theta = \sum_{k=1}^{S}\sqrt{\pi_k \pi'_k}.$$

Si au lieu de calculer cette valeur sur les collections, on la calcule sur deux échantillons de ces collections avec respectivement les vecteurs de fréquences observées $\mathbf{p} = (p_1, ..., p_S)^t$ et $\mathbf{q} = (q_1, ..., q_S)^t$, la valeur devient :

$$\cos \hat{\theta} = \sum_{k=1}^{S} \sqrt{p_k q_k}.$$

Une mesure de la dissimilarité entre deux collections, utilisée en génétique (Nei 1987), est alors la valeur observée de cet angle

$$D_6 (\mathbf{p}, \mathbf{q}) = [\arccos (\sqrt{p_k q_k})] .$$

Bhattacharyya suggérait de prendre la valeur de θ élevée au carré. Dans la continuité des travaux de Bhattacharyya, Edwards (1971), en génétique, propose la mesure

$$D_7 (\mathbf{p}, \mathbf{q}) = \sqrt{1 - \sum_{k=1}^{S} \sqrt{p_k q_k}},$$

qui est la longueur de la corde entre les deux collections. Nei (1987) considère le carré de l'indice d'Edwards

$$D_8 (\mathbf{p}, \mathbf{q}) = 1 - \sum_{k=1}^{S} \sqrt{p_k q_k}.$$

Une autre série de fonctions de dissimilarités part de la probabilité de tirer deux catégories identiques à partir des deux collections de distributions **p** et **q** qui est

$$\sum_{k=1}^{S} p_k q_k.$$

Nei (1972) propose de normaliser cette probabilité pour obtenir l'indice de similarité suivant entre **p** et **q**

$$S (\mathbf{p}, \mathbf{q}) = \frac{\sum_{k=1}^{S} p_k q_k}{\sqrt{\sum_{k=1}^{S} p_k^2} \sqrt{\sum_{k=1}^{S} q_k^2}}. \qquad (4.1)$$

Smith *et al.* (1990) attribuent l'origine de cette mesure de similarité en écologie à la thèse de Stander (1970). A partir de cet indice, trois mesures de dissimilarité ont été proposées. En génétique, Nei (1972, 1978) introduit une expression logarithmique : $D_9 (\mathbf{p}, \mathbf{q}) = - \ln [S (\mathbf{p}, \mathbf{q})]$

$$D_9 (\mathbf{p}, \mathbf{q}) = - \ln \left(\frac{\sum_{k=1}^{S} p_k q_k}{\sqrt{\sum_{k=1}^{S} p_k^2} \sqrt{\sum_{k=1}^{S} q_k^2}} \right)$$

En écologie, Manly (1994) suggère d'utiliser la relation traditionnelle entre un indice de dissimilarité et un indice de similarité ($D = 1 - S$) : $D_{10}(\mathbf{p}, \mathbf{q}) = 1 - S(\mathbf{p}, \mathbf{q})$

$$D_{10}(\mathbf{p}, \mathbf{q}) = 1 - \frac{\sum_{k=1}^{S} p_k q_k}{\sqrt{\sum_{k=1}^{S} p_k^2} \sqrt{\sum_{k=1}^{S} q_k^2}}$$

Il appelle cet indice, "indice de recouvrement". Orlóci (1967) (voir Faith *et al.* 1987) définit $D_{11}(\mathbf{p}, \mathbf{q}) = \sqrt{2\,[1 - S(\mathbf{p}, \mathbf{q})]}$:

$$D_{11}(\mathbf{p}, \mathbf{q}) = \sqrt{2\left(1 - \frac{\sum_{k=1}^{S} p_k q_k}{\sqrt{\sum_{k=1}^{S} p_k^2} \sqrt{\sum_{k=1}^{S} q_k^2}}\right)}.$$

Par ailleurs, d'autres indices ont été développés tels que celui d'Hedricks (1971) (voir Kuhnlein *et al.* 1989) en génétique

$$D_{12}(\mathbf{p}, \mathbf{q}) = -\ln\left(\frac{1}{S} \sum_{k=1}^{S} \frac{2 p_k q_k}{p_k^2 + q_k^2}\right).$$

Toujours en génétique, Reynolds *et al.* (1983) introduisent un coefficient de parenté ou distance de Reynolds :

$$D_{13}(\mathbf{p}, \mathbf{q}) = \sqrt{\frac{\sum_{k=1}^{S} (p_k - q_k)^2}{2(1 - \sum_{k=1}^{S} p_k q_k)}}.$$

Balakrishnan et Sanghvi (1968) proposent d'utiliser la distance du χ^2 qui augmente l'importance des catégories de faibles fréquences :

$$D_{14}(\mathbf{p}, \mathbf{q}) = \sum_{k=1}^{S} \frac{2(p_k - q_k)^2}{(p_k + q_k)}.$$

$\frac{1}{2} D_{14}$ est une approximation de D_6^2 pour des petites valeurs de D_6 (Bhattacharyya 1946, Chakraborty et Rao 1991). Cet indice $\frac{1}{2} D_{14}$ a été proposé par Mahalanobis dans les années 1930 (voir Bhattacharyya 1946) comme une mesure observée de divergence entre deux collections multinomiales.

Soient $\mathbf{n} = (n_1, ..., n_k, ..., n_S)^t$ et $\mathbf{m} = (m_1, ..., m_k, ..., m_S)^t$ deux vecteurs d'abondance des catégories, Bray et Curtis (1957) proposent de mesurer la dissimilarité entre ces deux vecteurs par

$$D_{15}(\mathbf{n}, \mathbf{m}) = \frac{\sum_{k=1}^{S} |n_k - m_k|}{\sum_{k=1}^{S} n_k + m_k} = 1 - \frac{2W}{A + B},$$

où W est la somme des abondances minimales de toutes les espèces, A le nombre total d'entités dans la première collection, et B le nombre total d'entités dans la seconde collection. Il s'agit d'une modification de la métrique de Manhattan. La racine carrée de cette fonction est métrique (Legendre et Legendre 1998) et a même des propriétés euclidiennes (Legendre et Anderson 1999).

D'un autre côté, MacArthur *et al.* (1966) (Whittaker 1972) en écologie et Nei (1973) en génétique utilisent des différences de Jensen appliquées respectivement à l'indice de Shannon

$$D_{16}(\mathbf{p}, \mathbf{q}) = \frac{1}{2}\left(\sum_{k=1}^{S} p_k \ln(p_k) + \sum_{k=1}^{S} q_k \ln(q_k)\right) - \sum_{k=1}^{S}\left(\frac{p_k + q_k}{2}\right)\ln\left(\frac{p_k + q_k}{2}\right)$$

et à celui de Gini-Simpson

$$D_{17}(\mathbf{p}, \mathbf{q}) = \frac{1}{2}\left(\sum_{k=1}^{S} p_k^2 + \sum_{k=1}^{S} q_k^2\right) - \sum_{k=1}^{S} p_k q_k.$$

Le choix des dissimilarités diffère selon le type de données et selon les traditions différentes notamment entre écologie et génétique.

Une des particularités des données génétiques, par rapport aux données écologiques, est la présence de plusieurs marqueurs. Pour les données génétiques, les catégories sont les allèles d'un locus (marqueur). La caractérisation d'une collection, par exemple une population, nécessite souvent l'analyse de plusieurs loci. Une matrice de dissimilarités entre populations peut être obtenue pour chaque locus. Ces matrices peuvent alors être comparées. Le plus souvent, une dissimilarité globale est calculée comme une moyenne des dissimilarités sur l'ensemble des loci. Parmi les indices utilisés en génétique, la dissimilarité moyenne entre deux populations est calculée pour les indices D_1, D_2 et D_5 par

$$\bar{D} = \frac{1}{L}\sum_{\ell=1}^{L} D^{[\ell]},$$

où $D^{[\ell]}$ est la mesure de dissimilarité selon le locus ℓ, et L est le nombre total de loci considérés. La dissimilarité moyenne entre deux populations pour les indices D_3 et D_4 est calculée par

$$\bar{D} = \sqrt{\frac{1}{L}\sum_{\ell=1}^{L} (D^{[\ell]})^2}.$$

Pour les indices D_9, D_{10} et D_{11}, elle est calculée en remplaçant la mesure de similarité de la formule 4.1 page 120 par

$$\bar{S}(\mathbf{p},\mathbf{q}) = \frac{\sum_{\ell=1}^{L}\sum_{k=1}^{S} p_k^{[\ell]} q_k^{[\ell]}}{\sqrt{\sum_{\ell=1}^{L}\sum_{k=1}^{S}\left(p_k^{[\ell]}\right)^2}\sqrt{\sum_{\ell=1}^{L}\sum_{k=1}^{S}\left(q_k^{[\ell]}\right)^2}},$$

où $p_k^{[\ell]}$ et $q_k^{[\ell]}$ sont les fréquences de l'allèle k du locus ℓ dans les deux populations comparées. Avec l'indice de Reynolds D_{13}, la dissimilarité moyenne est calculée par

$$\bar{D}_{13} = \sqrt{\frac{\sum_{\ell=1}^{L}\sum_{k=1}^{S}\left(p_k^{[\ell]} - q_k^{[\ell]}\right)^2}{2\sum_{\ell=1}^{L}\left(1 - \sum_{k=1}^{S} p_k^{[\ell]} q_k^{[\ell]}\right)}}.$$

Une des particularités étonnantes de cette liste d'indices de dissimilarités entre collections est la présence simultanée d'un indice D et de l'indice associé $D^2/2$. Par exemple

$$\begin{aligned} D_5 &= D_3^2/2, \\ D_{10} &= D_{11}^2/2. \end{aligned}$$

On y trouve aussi la présence d'un indice D_7 et de son carré D_8 :

$$D_8 = D_7^2.$$

N'importe quelle puissance positive d'une mesure de dissimilarité est une mesure de dissimilarité. Les choix sont souvent faits plus pour rechercher une signification génétique, biologique ou écologique aux indices que pour obtenir des indices aux propriétés mathématiques intéressantes telles que les propriétés métriques et euclidiennes, l'idéal étant bien sûr de satisfaire aux deux critères de choix.

Nous avions déjà observé ce problème de choix pour la définition des dissimilarités entre catégories. En effet, nous avons vu deux écritures de l'entropie quadratique :

$$H_{\mathbf{D}^{\text{cat}}}(\mathbf{p}) = \sum_{k=1}^{S}\sum_{l=1}^{S} p_k p_l d_{kl}^{\text{cat}}, \tag{4.2}$$

$$H_{\Delta^{\text{cat}}}(\mathbf{p}) = \sum_{k=1}^{S}\sum_{l=1}^{S} p_k p_l \frac{\left(\delta_{kl}^{\text{cat}}\right)^2}{2}, \tag{4.3}$$

ce qui nous a conduit à parler de "dissimilarités δ" et "dissimilarités d" où $d = \delta^2/2$. Dans sa présentation de l'entropie quadratique qui correspond à la formule 4.2, Rao affirme que la matrice $\mathbf{D}$ doit être choisie selon des critères qui dépendent de la discipline d'étude de la diversité, dans chaque cas précis. Dans l'ANOVA, où une variable Y est étudiée, ce n'est pas $d_{ij} = \left(\bar{y}_i - \bar{y}_j\right)^2/2$

qui est exprimée dans la même unité que Y mais bien $\delta_{ij} = |\overline{y}_i - \overline{y}_j|$. Dans la dbRDA (cf. page 85), Legendre et Anderson (1999) développent la méthode qui est à l'origine de la npMANOVA (Anderson 2001, cf. page 85) et qui se fondent sur deux méthodes d'ordination. Les résultats sont identiques à ceux de la npMANOVA, par exemple :

$$\begin{aligned} SS_T &= \frac{1}{IG}\sum_{i=1}^{IG-1}\sum_{j=i+1}^{IG}\delta_{ij}^2 \\ &= IG\sum_{g=1}^{G}\sum_{g'=1}^{G}\frac{1}{G}\frac{1}{G}\sum_{i=1}^{I}\sum_{j=1}^{I}\frac{1}{I}\frac{1}{I}\frac{\delta_{ig,jg'}^2}{2}. \end{aligned}$$

où $\delta_{ig,jg'}$ est la dissimilarité entre deux communautés. Legendre et Anderson (1999) affirment que $\delta_{ig,jg'}$, et non $\delta_{ig,jg'}^2/2$, doit être choisi d'un point de vue biologique et avoir des propriétés euclidiennes. Ils choisissent la fonction D_{15} de Bray-Curtis. Notons $D_{15}(ig, jg')$ la valeur prise par D_{15} entre les communauté ig et jg'. La fonction D_{15} ne conduit pas à des matrices euclidiennes de dissimilarités alors que sa racine y conduit. Le plus naturel serait donc de choisir $\delta_{ig,jg'} = \sqrt{D_{15}(ig, jg')}$, ou $\delta_{ig,jg'} = \sqrt{2D_{15}(ig, jg')}$, de sorte que soit $\delta_{ig,jg'}^2$ soit $\delta_{ig,jg'}^2/2$ soit égale à la valeur $D_{15}(ig, jg')$ de l'indice de Bray-Curtis. Le composant SST serait alors la dissimilarité moyenne entre sites selon la fonction D_{15} de Bray-Curtis. Cependant, Legendre et Anderson imposent que $\delta_{ig,jg'}$ soit mesurée par la fonction de Bray-Curtis ce qui les obligent à modifier la matrice Δ obtenue pour qu'elle devienne euclidienne. Legendre et Anderson étudient l'influence de chaque type de transformation pouvant permettre de rendre D_{15} euclidienne : la racine carré qui conduirait à $\delta_{ig,jg'} = \sqrt{D_{15}(ig, jg')}$, la transformation de Cailliez (1983) et celle de Lingoes (1971). Observant que la transformation de Lingoes (1971) a moins d'impact sur la valeur de D_{15}, ils préconisent cette transformation qui consiste à trouver la plus petite constante c telle que

$$\delta_{ig,jg'} = \sqrt{D_{15}(ig, jg')^2 + c}$$

ait des propriétés euclidiennes. Ce raisonnement va dans le sens de notre commentaire sur la variance : la dissimilarité qui a l'unité biologique est δ et non d. Mais d'un autre point de vue, si d et non δ a l'unité biologique alors $H_{\mathbf{D}^{\text{cat}}}(\mathbf{p})$ est une mesure simple et facilement interprétable comme moyenne de distances biologiques. Le débat est donc ouvert.

Notons cependant que, s'ils avaient choisi de baser leur analyse sur la formule 4.1 page 120, le problème aurait été plus complexe étant donné que les dissimilarités $D_{10} = D_{11}^2/2$ et D_{11}, basées sur cette formule, ont été toutes deux proposées pour une utilisation en écologie.

Ce problème du choix des dissimilarités, qu'il soit au niveau des catégories ou des collections, n'est absolument pas résolu, et il est récurrent dans la littérature de biologie théorique en génétique et en écologie.

4.1.3 Indices tenant compte des dissimilarités entre catégories

Le problème de la mesure de la diversité taxonomique a fait l'objet de nombreuses recherches et plusieurs indices ont été proposés (cf. partie 2). Cependant, lorsque plusieurs assemblages d'espèces sont comparés dans l'espace ou le temps, la taxonomie est rarement prise

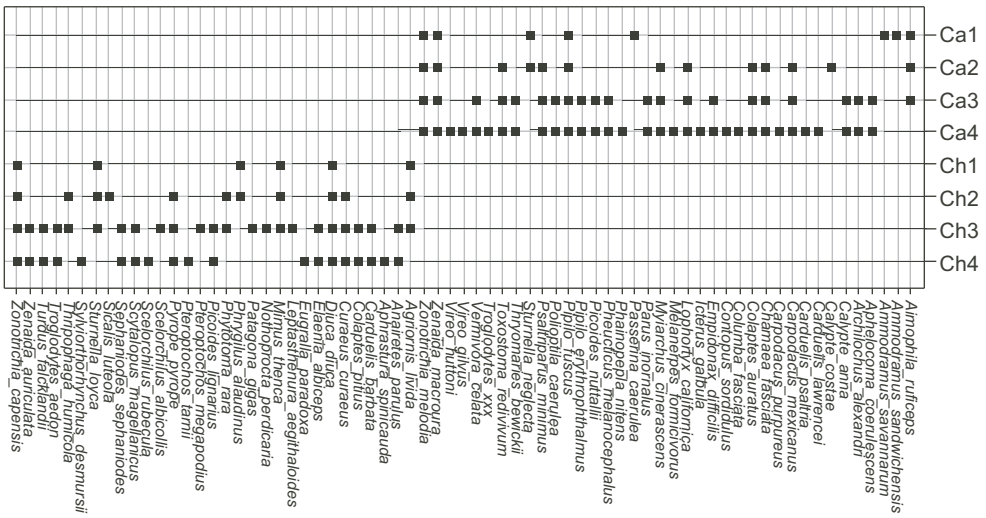


FIG. *17 – Données de Blondel* et al. *(1984) pour la Californie "Ca" et le Chili "Ch". Dans les étiquettes des communautés, les numéros indiquent des stades de végétations le long de chaque succession : de 1 habitat ouvert, à 4 habitat fermé végétation dense. Cette figure a été réalisée avec la fonction 'table.value' du package ade4 de R (Ihaka et Gentleman 1996, Chessel* et al. *2004).*

en compte. Par exemple, Price *et al.* (1999) mesurent la diversité au sein de zones géographiques en utilisant l'indice de spécificité taxonomique de Warwick et Clarke (1995). Mais ne parvenant pas à trouver dans la littérature un indice de dissimilarité taxonomique entre deux zones, ils utilisent l'indice de Sørensen basé uniquement sur les présences/absences des différentes espèces (cf. partie 4.1.2).

Récemment Izsak et Price (2001) proposèrent l'indice suivant :

$$\mathrm{TD} = \frac{\sum_{k=1}^{S_A} w_{kB} + \sum_{l=1}^{S_B} w_{lA}}{S_A + S_B},$$

où S_A et S_B sont les nombres d'espèces dans les sites A et B respectivement, w_{kB} est la distance taxonomique minimale entre l'espèce k du site A et toutes les espèces du site B, et w_{lA} est la distance taxonomique minimale entre l'espèce l du site B et toutes les espèces du site A. Cet indice ne prend pas en compte les fréquences des espèces mais permet de considérer des distances taxonomiques. Il peut être divisé par $L - 1$, où L est le nombre de niveaux dans la taxonomie, pour que ses valeurs soient comprises entre 0 et 1. Il est moins influencé par la richesse spécifique et l'effort d'échantillonnage que les indices classiques (Izsak et Price 2001).

Prenons dans les données de Blondel *et al.* (1984), uniquement les successions de Californie et Chili. Nous avons parlé de ces données dans la partie 2.2.3 dans le cadre d'une méthode de test liée à la richesse spécifique. Il était démontré une convergence dans la répartition des richesses entre catégories alimentaires pour les quatre régions Californie, Chili, Provence et Bourgogne. Nous allons nous intéresser ici non plus aux catégories alimentaires mais à la composition spécifique le long des successions. Chaque succession est divisée en quatre habitat correspondant à

(a)

	Ca1	Ca2	Ca3	Ca4	Ch1	Ch2	Ch3
Ca2	0.69						
Ca3	0.85	0.54					
Ca4	0.92	0.71	0.36				
Ch1	**1.00**	**1.00**	**1.00**	**1.00**			
Ch2	**1.00**	**1.00**	**1.00**	**1.00**	0.46		
Ch3	**1.00**	**1.00**	**1.00**	**1.00**	0.80	0.65	
Ch4	**1.00**	**1.00**	**1.00**	**1.00**	0.91	0.85	0.52

(b)

	Ca1	Ca2	Ca3	Ca4	Ch1	Ch2	Ch3
Ca2	1.62						
Ca3	2.17	0.86					
Ca4	2.45	1.24	0.50				
Ch1	**2.07**	**2.37**	**2.71**	**2.79**			
Ch2	**2.21**	**2.38**	**2.70**	**2.74**	0.71		
Ch3	**2.59**	**2.16**	**2.26**	**2.16**	2.03	1.49	
Ch4	**2.56**	**2.22**	**2.24**	**2.16**	2.40	2.03	0.70

(c)

	Ca1	Ca2	Ca3	Ca4	Ch1	Ch2	Ch3
Ca2	0.72						
Ca3	0.77	0.33					
Ca4	0.83	0.40	0.25				
Ch1	**0.82**	**0.83**	**0.83**	**0.85**			
Ch2	**0.78**	**0.72**	**0.70**	**0.70**	0.44		
Ch3	**0.87**	**0.55**	**0.48**	**0.44**	0.78	0.58	
Ch4	**0.89**	**0.60**	**0.53**	**0.49**	0.84	0.66	0.28

FIG. *18 – Dissimilarités entre les communautés de Californie "Ca" et Chili "Ch" calculées par (a) l'indice de Jaccard (*$1 - S_1$*, partie 4.1.1), (b) l'indice de Izsak et Price et (c) racine de la différence de Jensen appliquée à l'entropie quadratique. Les numéros indiquent des stades de végétations le long de chaque succession : de 1 habitat ouvert, à 4 habitat fermé végétation dense.*

quatre stade de végétations. Ces habitats se ressemblent d'une succession à l'autre. Ces données sont en présence/absence d'espèce d'oiseaux nichant dans ces régions. Dans chaque stade de végétation de chaque succession les espèces d'oiseaux présentes constituent une communauté. Nous calculons ici des dissimilarités entre ces communautés pour une succession et aussi entre les deux successions considérées, Californie et Chili.

Les indices de dissimilarité sur données binaires attribuent une valeur strictement comprise entre 0 et 1 pour les dissimilarités entre communautés évoluant à différents stades d'une même succession (Fig. 18a). Par contre comme aucune espèce commune aux deux successions n'a été observée (Fig. 17), tous les indices de dissimilarité sur données binaires attribuent des dissimilarités égales à 1 entre les communautés de Californie et celles de Chili, c'est-à-dire la valeur maximale possible de dissimilarité qui apparaît pour deux communautés complètement disjointes. Il est donc impossible avec ces seules informations de dire si deux habitats de même stade de végétation mais issus de deux successions différentes se ressemblent selon un point de vue autre que l'identité de l'espèce. Par contre en rajoutant les liens taxonomiques entre espèces, il est possible de calculer des dissimilarités différentes entre les habitats des deux successions. Ainsi, en réorganisant les espèces selon leur classification taxonomique, on observe que les deux successions, bien que sans espèce commune, partagent des mêmes genres, familles et ordres (Fig. 19). L'indice d'Izsak et Price permet alors de calculer des dissimilarités entre les communautés de Californie et celles de Chili (Fig. 18b).

La grande particularité des fonctions de dissimilarité entre collections prenant en compte des mesures de dissimilarité entre catégories, est donc de permettre le calcul de dissimilarités

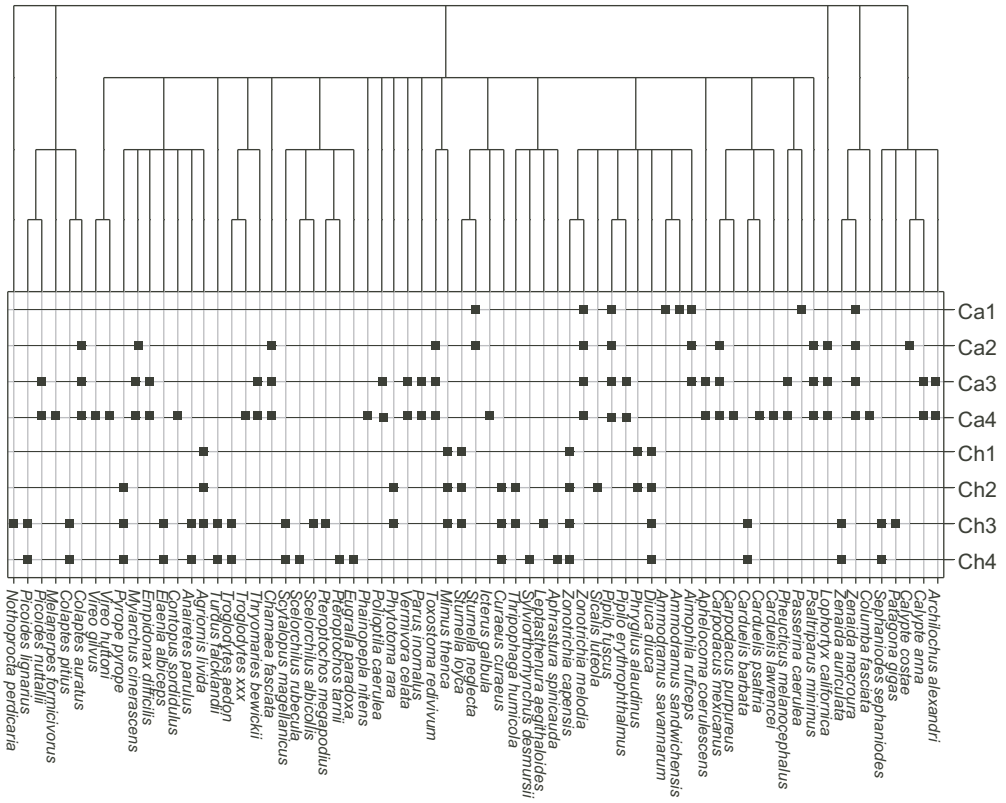


FIG. *19 – Réorganisation des données de Blondel* et al. *(1984) pour la Californie "Ca" et le Chili "Ch", les espèces étant rangées selon leur classification taxonomique. Dans les étiquettes des communautés, les numéros indiquent des stades de végétations le long de chaque succession : de 1 habitat ouvert, à 4 habitat fermé végétation dense. Les niveaux de la taxonomie sont espèce, genre, famille et ordre. Cette figure a été réalisée avec les fonctions 'as.taxo', 'taxo2phylog', 'phylog' et 'table.phylog' du package ade4 de R (Ihaka et Gentleman 1996, Chessel* et al. *2004)*

entre collections ne partageant aucune catégorie.

Un indice présente des avantages sur celui d'Iszak et Price. C'est la différence de Jensen appliquée à l'entropie quadratique, puisqu'elle permet de prendre en compte à la fois les abondances des espèces et leurs différences taxonomiques. Depuis les travaux de Nei en génétique, surtout pour des séquences d'ADN, la différence de Jensen sur entropie quadratique est utilisée (Comas *et al.* 1996). Cependant aucune contrainte n'est placée sur le choix de la matrice de dissimilarité entre catégories sur laquelle l'entropie quadratique est appliquée. Or il faut s'assurer que les valeurs obtenues de dissimilarités entre collections seront toujours positives. Ainsi, Comas *et al.* (1998) utilisent cette fonction à partir de distances nucléotidiques entre séquences d'ADN et obtiennent une matrice de dissimilarités entre populations comprenant plusieurs valeurs négatives. Pour éliminer ces valeurs négatives, ils rajoutent une petite constante à la matrice entière affirmant que cette procédure n'altère pas les relations génétiques et nous rend capable d'utiliser des algorithmes de construction d'arbres. La concavité de l'entropie quadratique, et donc la positivité de la différence de Jensen appliquée à l'entropie quadratique, est

toujours vérifiée si la matrice de dissimilarités entre catégories est euclidienne. Par la suite, il a été choisi de se restreindre à ce type de matrice. La racine carrée de la métrique de Jensen appliquée à l'entropie quadratique devient alors une mesure de distance entre collections, aux propriétés euclidiennes (Rao et Nayak 1985, Champely et Chessel 2002), permettant de prendre en compte non seulement la liste des catégories appartenant à chaque collection, mais aussi une matrice caractérisant un aspect des différences entre ces catégories. La différence de Jensen appliquées à l'entropie quadratique a été aussi utilisée pour mesurer la différence taxonomique entre cortèges avifaunistiques des habitats de Chili et Californie (Fig. 18c). Cette métrique mesure des écarts plus grands entre ces différences, mettant ainsi en évidence une ressemblance plus importante entre les deux régions, Chili et Californie, dans les habitats fermés. Nous avons montré que l'utilisation de cette métrique en écologie associée à des méthodes de représentation graphique se révèle prometteuse (Pavoine *et al.* 2004, cf. annexe 1).

Pierre Legendre m'a fait remarquer un comportement étonnant de l'indice d'Iszak et Price dans le cas particulier illustré par la figure 20. Cette figure montre deux exemples contenant deux sites. Dans le premier exemple, les sites A et B possèdent des espèces majoritairement de familles distinctes, alors que dans l'exemple 2, ils possèdent des espèces de même familles mais de genres différents. L'indice d'Iszak et Price considère que la dissimilarité entre les deux sites est la même dans les deux exemples (TD = 8). La racine de la différence de Jensen appliquée à l'entropie quadratique, suggèrent au contraire que les sites A et B sont beaucoup plus différents dans l'exemple 1 ($\delta_{AB}^{\text{site}} = 1.173$) que dans l'exemple 2 ($\delta_{AB}^{\text{site}} = 0.500$), ce qui est plus conforme à ce que l'on pressent intuitivement en regardant la figure 20. Nous conserverons ces deux indices dans les deux paragraphes suivant pour mesurer la dissimilarité taxonomique, en gardant cette remarque en mémoire.

4.2 R ´ ´

4.2.1 Méthodes d'arbres et de classifications hiérarchiques

Rao (1982c) propose que les coefficients de dissimilarité de Jensen obtenus avec l'entropie quadratique soient utilisés pour construire des dendrogrammes afin d'étudier des classifications des collections. Ainsi, pour obtenir une représentation graphique des différences entre collections, beaucoup d'auteurs utilisent des représentations en arbres. De nombreuses méthodes de classifications hiérarchiques ont été développées. Dans R (Ihaka et Gentleman 1996), la fonction "hclust" propose les méthodes "ward", "single", "complete", "average", "mcquitty", "median" et "centroid".

Pour chaque méthode, au premier pas, chaque collection constitue un groupe et nous disposons d'une matrice de distance entre ces groupes. Ensuite, les deux groupes les plus proches sont fusionnés. Les groupes sont ainsi associés progressivement. Les ordres de regroupement forment une classification hiérarchique. Lors d'un regroupement, deux groupes fusionnent, il faut donc redéfinir les distances entre ce nouveau groupe et les autres groupes. C'est lors de cette étape que les méthodes de classification hiérarchique utilisées par "hclust" diffèrent. Chaque méthode se distinguent par une manière particulière de définir la distance entre deux groupes **a** et **b** nouvellement fusionnés et n'importe quel autre groupe **g**. Lance et Williams (1966, 1967) (voir Legendre et Legendre 1998) ont proposé un schéma général pour définir les mesures uti-

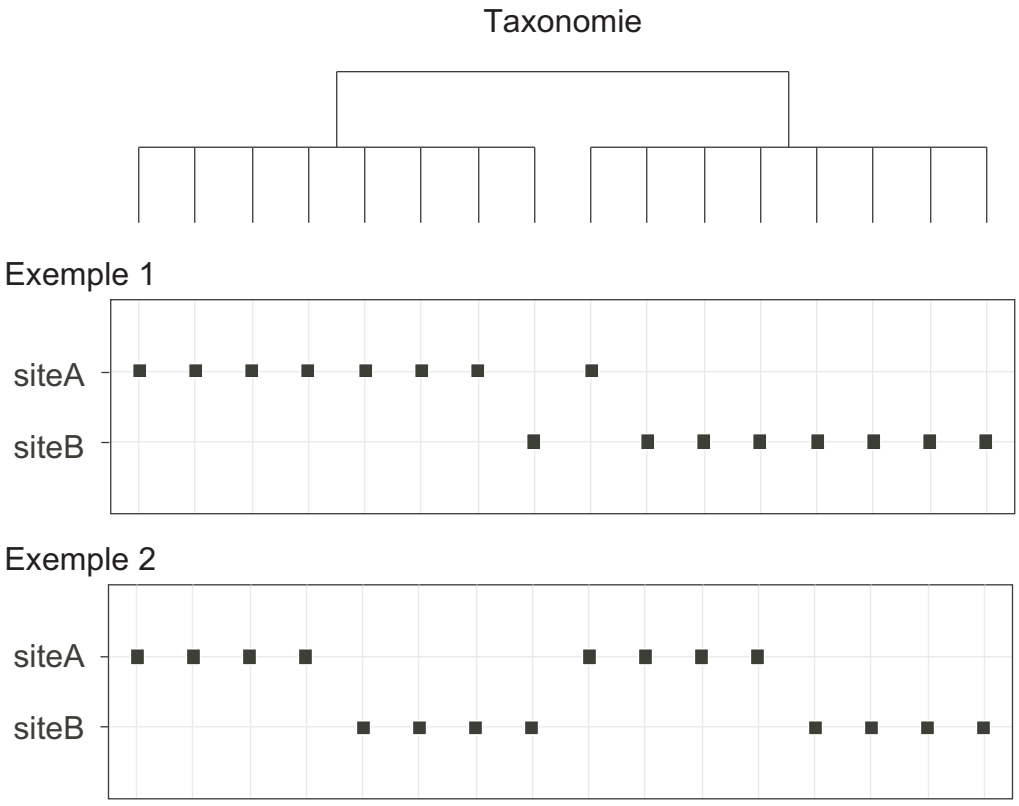


FIG. *20 – Exemple théorique trouvé par Pierre Legendre mettant ainsi en évidence un cas où l'indice d'Iszak et Price donne des résultats inattendus. La dissimilarité entre les sites A et B calculée par l'indice de Iszak et Price est la même que l'on soit dans l'exemple A ou dans l'exemple B.*

lisées. Notons $D(\mathbf{ab}, \mathbf{g})$ la mesure de la distance entre le nouveau groupe **ab** et le groupe **g**, et $D(\mathbf{a}, \mathbf{b})$, $D(\mathbf{a}, \mathbf{g})$ et $D(\mathbf{b}, \mathbf{g})$ les distances entre les groupes avant la fusion. L'équation de Lance et Williams est

$$D(\mathbf{ab}, \mathbf{g}) = \alpha_a D(\mathbf{a}, \mathbf{g}) + \alpha_b D(\mathbf{b}, \mathbf{g}) + \beta D(\mathbf{a}, \mathbf{b}) + \gamma |D(\mathbf{a}, \mathbf{g}) - D(\mathbf{b}, \mathbf{g})|,$$

où les valeurs prises par les paramètres α_a, α_b, β et γ pour chaque méthode sont données dans le tableau 5.

La classification hiérarchique obtenue est schématisée par un arbre. Les collections sont aux feuilles, aussi appelées nœuds terminaux, et le groupe contenant toutes les collections est à la racine de l'arbre. Chaque groupement est représenté par un nœud. Appelons u le nœud correspondant à la fusion de deux groupes a et b. La somme des longueurs des branches reliant les collections du groupe **a** au nœud u est égale à $D(\mathbf{a}, \mathbf{b})$; de même celle des longueurs des branches reliant les collections de **b** à u est $D(\mathbf{a}, \mathbf{b})$ (cf. 21). Ces arbres fournis par la classification hiérarchique sont donc racinés par le groupe englobant toutes les collections et possèdent des feuilles équidistantes à la racines. On qualifie ces arbres d'ultramétriques. Les méthodes "average", "mcquitty", "median", et "centroid" sont également connues sous les noms respectifs "UPGMA" (Unweighted Pair-Group Method using Arithmetic averages), "WPGMA" (Weigh-

TAB. *5 – Paramètres de l'équation de Lance et Williams.*

Méthode	α_a*	α_b*	β*	γ
Ward	$\frac{n_a+n_g}{n_a+n_b+n_g}$	$\frac{n_b+n_g}{n_a+n_b+n_g}$	$\frac{-n_g}{n_a+n_b+n_g}$	0
Single	$1/2$	$1/2$	0	$-1/2$
Complete	$1/2$	$1/2$	0	$1/2$
Average	$\frac{n_a}{n_a+n_b}$	$\frac{n_b}{n_a+n_b}$	0	0
Mcquitty	$1/2$	$1/2$	0	0
Median	$1/2$	$1/2$	$-1/4$	0
Centroid	$\frac{n_a}{n_a+n_b}$	$\frac{n_b}{n_a+n_b}$	$\frac{-n_a n_g}{(n_a+n_b)^2}$	0

* n_a, n_b, et n_g sont les nombres de collections dans chaque groupe *a*, *b* et *g*.

ted Pair-Group Method using Arithmetic averages), "WPGMC" (Weighted Pair-Group Method using Centroids), et "UPGMC" (Unweighted Pair-Group Method using Centroids). Les méthodes "single" et "complete" sont aussi appelées respectivement "méthode du saut minimal" et "méthode du saut maximal". Si les distances de départ sont d_{ij}, la méthode de Ward a la particularité de travailler sur d_{ij}^2 au lieu de d_{ij}. Il peut être observé que les méthodes "ward", "centroid" et "median", lorsqu'elles sont appliquées à des distances ultramétriques fournissent un arbre différent de celui qui est associé à la matrice de départ. A partir de l'arbre de la classification, il est possible de calculer les distances entre collections par la somme des longueurs de branches qui les séparent sur l'arbre. La matrice de distances ainsi fournie est ultramétrique.

Pour obtenir un arbre à partir de distances entre collections, la méthode "average" est la plus utilisée. Cependant, en génétique, une autre méthode, celle du Neighbor-Joining (NJ) (Saitou et Nei 1987, Comas *et al.* 1996, 1998, Bosch *et al.* 1999, Comas *et al.* 2000, Vona *et al.* 2001) est également très utilisée. Il ne s'agit pas d'une méthode de classification hiérarchique puisque l'arbre obtenu n'est ni ultramétrique ni raciné. Cette méthode a été développée par Saitou et Nei (1987) pour des distances entre unités taxonomiques (*e.g.* espèces) mais est utilisée également pour des distances entre collections. Chaque unité taxonomique ou, pour ce qui nous intéresse ici, chaque collection constitue un nœud terminal (feuille) sur l'arbre. La méthode du Neighbor-Joining est résumée par Swofford *et al.* (Weir 1996) en 5 étapes :

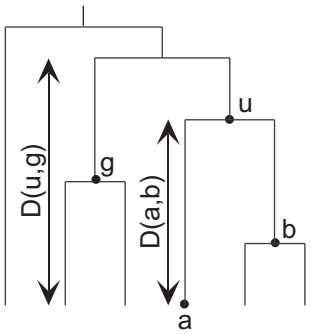


FIG. *21 – La distance entre deux groupes est la distance entre les feuilles (collections) de ces groupes sur l'arbre de classification hiérarchique.*

1. Calcul de la somme des dissimilarités entre un nœud terminal i et tous les autres

$$r_i = \sum_{k=1}^{N} d_{ki}$$

   N est le nombre total de nœuds terminaux.

2. Détermination de la paire de nœuds i, j qui minimise la distance corrigée d^*_{ij} définie par

$$d^*_{ij} = d_{ij} - \frac{r_i + r_j}{N-2}.$$

3. Définition d'un nouveau nœud u qui représente le rassemblement au sein d'un nouveau groupe dans la classification des collections descendant des nœuds i et j. Les longueurs de branches reliant les nœuds i et j au nœud u sont

$$s_{iu} = \frac{d_{ij}}{2} + \frac{r_i - r_j}{2\,(N-2)}, \tag{4.4}$$

$$s_{ju} = \frac{d_{ij}}{2} + \frac{r_j - r_i}{2\,(N-2)}. \tag{4.5}$$

   Les longueurs de branches allant de u à chaque nœud terminal k du nouveau groupe sont

$$d_{ku} = \frac{d_{ik} + d_{jk} - d_{ij}}{2}. \tag{4.6}$$

4. Remplacement dans la matrice de dissimilarités des nœuds i et j par le nœud u et diminution de N d'une unité.

5. S'il reste plus de deux nœuds alors on recommence à l'étape 1. Sinon, la longueur de branches joignant les deux nœuds restants k et l est $s_{kl} = d_{kl}$ et l'arbre est complètement déterminé.

L'équation 4.6 correspond bien à un cas particulier de l'équation de Lance et Williams, où $\alpha_a = 1/2$, $\alpha_b = 1/2$ et $\beta = -1/2$. Mais la méthode diffère des classifications hiérarchiques

par la manière asymétrique de calculer les longueurs de branches (équations 4.4 et 4.5) et par l'absence de racine.

En utilisant par exemple la méthode "average", il est possible d'obtenir une classification (Fig. 22) des communautés de Californie et du Chili pour chacune des matrices de dissimilarités considérées dans la figure 18, page 126, de la partie 4.1.3. Les deux premières matrices (distance de Jaccard et distance d'Izsák et Price aboutissent à des classifications assez similaires 22a et b. Les communautés sont d'abord séparées entre les deux successions (Californie versus Chili) puis au sein des successions entre les stades de végétation (pour la Californie, classification allant de l'habitat 1 (habitat le plus distinct) à l'habitat 4 (habitat partageant beaucoup d'espèces avec les autres habitats de Californie) et pour le Chili, opposition entre les habitats ouverts (1 et 2) et les habitats fermés (3 et 4)). Les informations données par les figures 22a et 22b sont en fait différentes. La première décrit simplement les différences entre communautés dans chaque succession, puisqu'il n'y a pas de calcul de dissimilarités entre successions. En revanche la deuxième classification montre qu'il existe plus de différences taxonomiques entre successions qu'entre stades de végétations dans les successions. Par contre cette représentation graphique est décevante puisque l'indice d'Izsák et Price décrit des dissimilarités variées entre communautés de régions différentes, ce qui n'apparaît pas dans les classifications hiérarchiques. La troisième matrice de dissimilarités (racine des différences de Jensen appliquées à l'entropie quadratique) donne en revanche des résultats très différents 22c. Les deux régions ne sont plus séparées, les résultats sont difficilement interprétables. Nous verrons par la suite que la classification hiérarchique ne parvient, pas dans cette utilisation qu'on veut lui donner, à décrire efficacement les dissimilarités entre habitats. Il nous faut explorer d'autres méthodes de représentations graphiques.

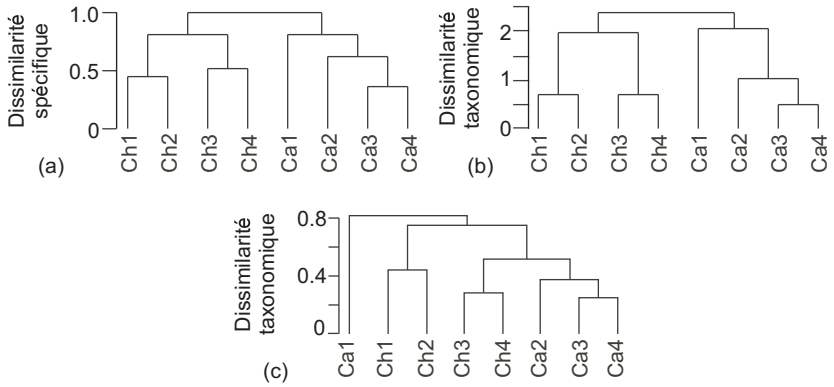
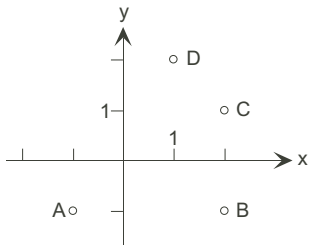


FIG. *22 – Représentation des dissimilarités entre Communautés de Californie et du Chili par la méthode de classification "average" (données de Blondel* et al. *(1984), cf. partie 4.1.3.) : (a) dissimilarité de Jaccard, (b) dissimilarité d'Izsak et Price et (c) racine de la différence de Jensen appliquée à l'entropie quadratique. Les deux successions sont notés 'Ca' pour la Californie et 'Ch' pour le Chili. Les chiffres indiquent le stade de végétation : de 1 habitat ouvert à 4 habitat fermé. Cette figure a été réalisée avec la fonction 'hclust' de la base de R (Ihaka et Gentleman 1996).*

4.2.2 Positionnement multidimensionnel

Deux autres méthodes possibles pour représenter les dissimilarités entre collections sont le positionnement multidimensionnel (MDS, metric multidimensional scaling), aussi appelé analyse en coordonnées principales (PCoA, principal coordinate analysis) (Gower 1984, Gower et Legendre 1986), et le positionnement multidimensionnel non-métrique (NMDS, nonmetric multidimensional scaling) (Kruskal 1964a, b).



	x	y
A	-1	-1
B	2	-1
C	2	1
D	1	2

	A	B	C
B	3		
C	$\sqrt{13}$	2	
D	$\sqrt{13}$	$\sqrt{10}$	$\sqrt{2}$

Représentation graphique Coordonnées Distances

FIG. *23 – Passage d'une représentation graphique à une matrice de distances.*

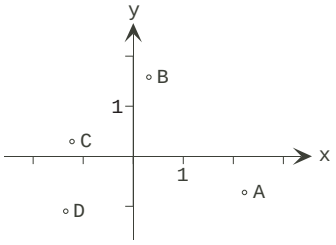
A partir d'une représentation graphique, un tableau de coordonnées peut être obtenu ainsi qu'une matrice de distances entre points (Fig. 23). Si x_{il} (resp. x_{jl}) est la coordonnée du point i (resp. j) sur l'axe l, la distance entre les points i et j est calculée par la métrique euclidienne :

$$\sqrt{\sum_l \left(x_{il} - x_{jl}\right)^2}.$$

A l'inverse, à partir d'une matrice de distances, les méthodes d'ordination PCoA et NMDS permettent de trouver un espace euclidien et, dans cet espace, un système de coordonnées pour représenter les points (Fig. 24).

	A	B	C
B	3		
C	$\sqrt{13}$	2	
D	$\sqrt{13}$	$\sqrt{10}$	$\sqrt{2}$

	x	y
A	-2.24	-0.74
B	0.32	1.57
C	-1.22	0.29
D	-1.35	-1.12



Distances Coordonnées Carte factorielle

FIG. *24 – Passage d'une matrice de distances ou dissimilarités à une carte factorielle : la PCoA.*

Notons $\mathbf{\Delta} = \left[\delta_{ij}\right]$ une matrice de dissimilarités entre r collections. La PCoA recherche un repère orthonormé où chaque axe maximise la variance des coordonnées des points. Soit $\mathbf{D} =$

$\left[\delta_{ij}^2/2\right]$, et soit la matrice de centrage $\mathbf{Q} = \mathbf{I}_r - \frac{1}{r}\mathbf{1}\mathbf{1}^t$. La PCoA correspond à la décomposition en valeurs singulières de la matrice **D** doublement centrée (par lignes et colonnes)

$$-\mathbf{QDQ} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^t.$$

Les matrices **U** et $\mathbf{\Lambda}$ contiennent respectivement vecteurs et valeurs propres. Les coordonnées des collections sont données par les lignes de la matrice $\mathbf{X} = \mathbf{U}\mathbf{\Lambda}^{1/2}$. Si $\mathbf{\Lambda}$ est euclidienne, les valeurs propres sont toutes positives et les distances euclidiennes entre les points dans l'espace multidimensionnel défini par la PCoA sont égales aux dissimilarités δ_{kl}.

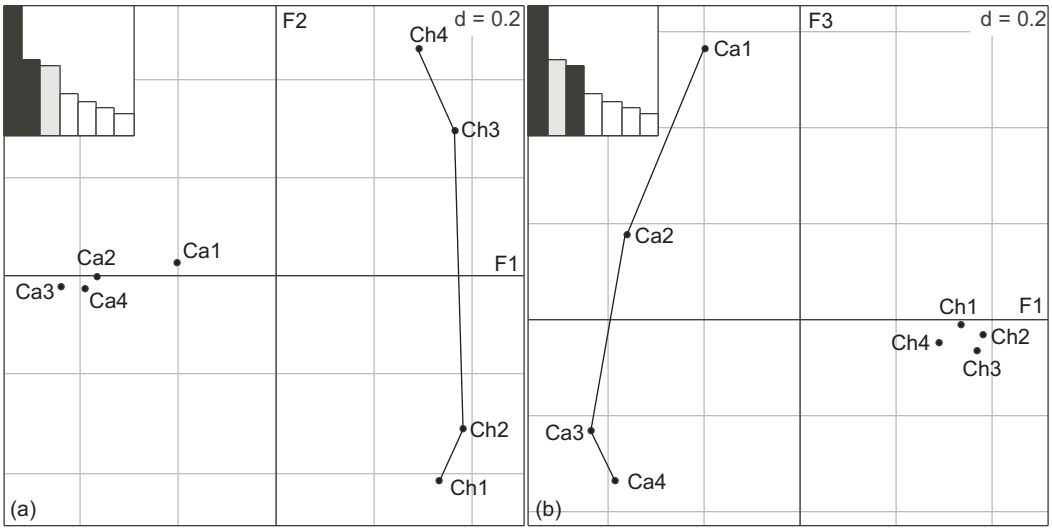


FIG. *25 – Résultat de la PCoA appliquée aux dissimilarités de Jaccard entre les communautés de Californie et du Chili : a) axes factoriels (principaux) 1 et 2 ; b) axes factoriels 1 et 3. Les encadrés donnent le graphe des valeurs propres avec en noir celles correspondant aux axes représentés. Cette figure a été réalisée avec la fonction 'dudi.pco' du package ade4 de R (Ihaka et Gentleman 1996, Chessel* et al. *2004).*

Appliquons la PCoA à la matrice de dissimilarités entre communautés de Californie et du Chili obtenues par l'indice de Jaccard (Fig. 25). Le premier axe principal sépare les communautés de Californie des communautés du Chili. Ensuite les dissimilarités de Jaccard entre les communautés de successions différentes étant toutes égales à 1, le deuxième axe se situe au niveau de l'axe principal des communautés du Chili montrant que les espèces s'organisent le long de la succession, et le troisième axe au niveau de l'axe principal des communautés de Californie montrant que les espèces de Californie s'organisent aussi le long d'une succession. Lorsque la PCoA est appliquée à la matrice de dissimilarités entre communautés de Californie et du Chili obtenues par l'indice d'Izsak et Price (Fig. 26a), on obtient directement en axe 1 la séparation des deux successions et en axe 2 le gradient commun de végétation. Au contraire, avec la racine de la différence de Jensen appliquée à l'entropie quadratique, l'axe 1 suit le gradient commun de végétation et l'axe deux sépare les deux successions (Fig. 26b). Cette dernière analyse montre aussi clairement une convergence des communautés d'oiseaux, sachant les liens taxonomiques entre les espèces d'oiseaux, dans les milieux fermés. Ces deux applications de la

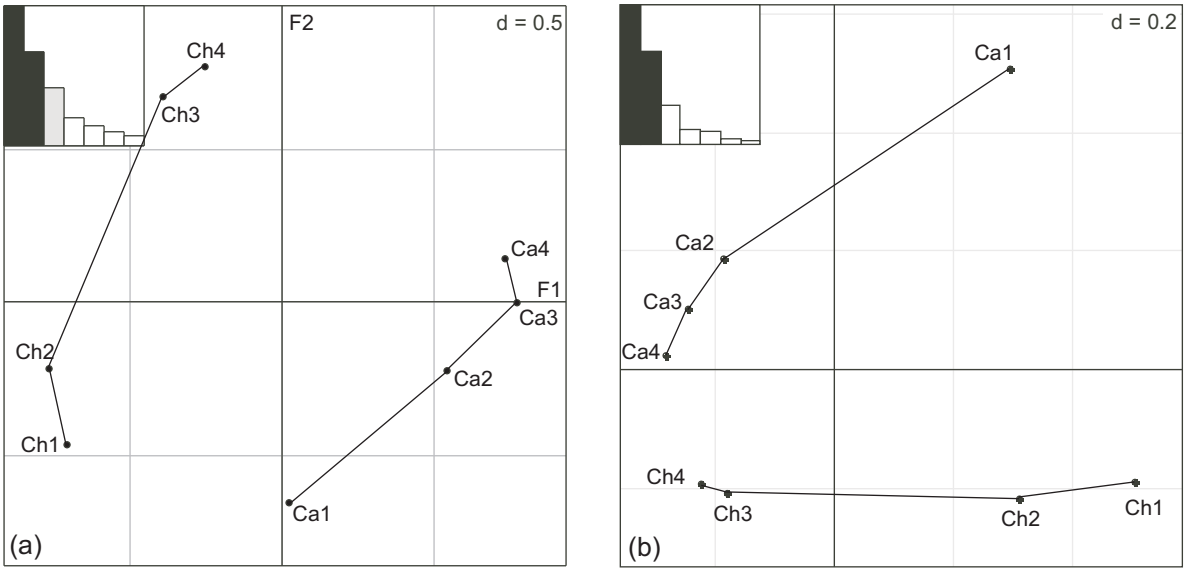


FIG. *26 – Résultat de la PCoA appliquée aux entre les communautés de Californie et du Chili calculées par (a) les dissimilarités de Izsak et Price (b) la racine de la différence de Jensen appliquée à l'entropie quadratique. Les encadrés donnent le graphe des valeurs propres avec en noir celles correspondant aux axes représentés. Cette figure a été réalisée avec la fonction 'dudi.pco' du package ade4 de R (Ihaka et Gentleman 1996, Chessel* et al. *2004).*

PCoA révèlent ainsi que les deux successions, bien que contenant des espèces différentes, ont la même organisation des genres, familles, ordres le long des gradients de végétation. La figure obtenue par la PCoA (Fig. 26b) est donc nettement plus claire que celle fournie par la classification hiérarchique (Fig. 22c). L'utilisation de l'indice de Jaccard ne permet pas de montrer ce gradient commun.

Pour la NMDS, le nombre de dimensions de l'espace de représentation est défini par l'utilisateur. A partir de ce nombre, la NMDS recherche par des processus itératifs des coordonnées des points pour arriver à ce que les distances euclidiennes entre les coordonnées de ces points ($\hat{\delta}_{ij}$) minimisent l'une des fonctions suivantes (Legendre et Legendre 1998) :

$$Stress_1 = \sqrt{\sum_{i=1}^{r}\sum_{j=1}^{r}\left(\delta_{ij} - \hat{\delta}_{ij}\right)^2 / \sum_{i=1}^{r}\sum_{j=1}^{r}\delta_{ij}^2},$$

$$Stress_2 = \sqrt{\sum_{i=1}^{r}\sum_{j=1}^{r}\left(\delta_{ij} - \hat{\delta}_{ij}\right)^2 / \sum_{i=1}^{r}\sum_{j=1}^{r}\left(\delta_{ij} - \bar{\delta}\right)^2},$$

$$Stress_3 = \sqrt{\sum_{i=1}^{r}\sum_{j=1}^{r}\left(\delta_{ij} - \hat{\delta}_{ij}\right)^2}.$$

Dans sa définition du positionnement multidimensionnel, Kruskal (1964a, b) choisit la fonction $Stress_1$.

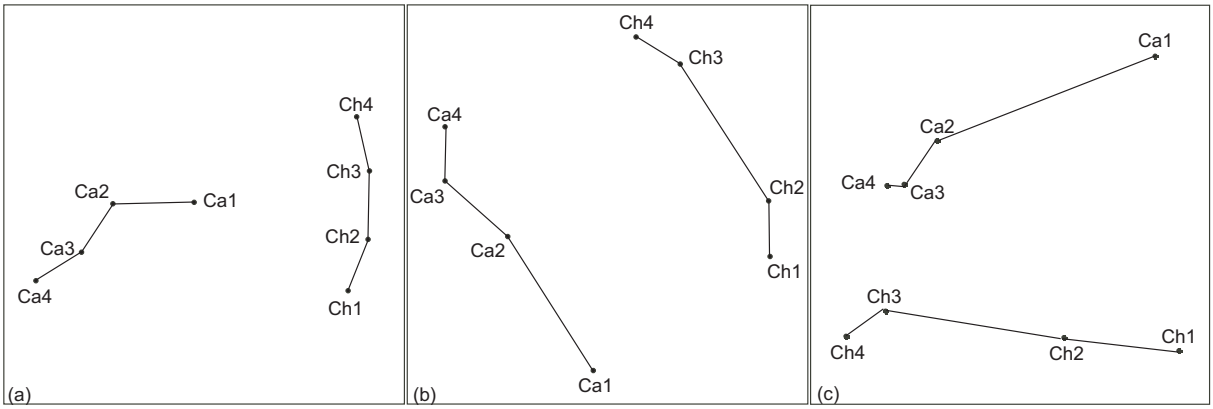


FIG. *27 – Résultat de la NMDS appliquée aux dissimilarités de (a) Jaccard et (b) Izsak et Price entre les communautés de Californie et du Chili et (c) aux racines des différences de Jensen appliquées à l'entropie quadratique. Le nombre de dimensions choisi est 2. La formule "$Stress_1$" est utilisée, les valeurs obtenues sont (a) 0.002 et (b) 1.202 respectivement.*

Appliquée au problème des communautés d'oiseaux de Chili et de Californie, la NMDS montre aussi l'intérêt d'intégrer la taxonomie dans la mesure des dissimilarités entre communautés (Fig. 27). Avec l'indice de Jaccard, les ordinations des communautés des deux successions se trouvent sur le même graphique, mais perpendiculaires puisqu'elles n'ont aucun lien (Fig. 27a). Au contraire, avec l'indice d'Izsak et Price, les deux successions se trouvent parallèles sur le même graphique et cette ordination des communautés étudiées dans ces successions révèle ainsi une même structuration taxonomique des communautés des deux successions le long de leur gradient commun de végétation (Fig. 27b).

4.2.3 Limites

Pour l'ensemble des méthodes, le choix de la fonction définissant les dissimilarités influencera les résultats.

Jackson *et al.* (1989) font le point sur tous les problèmes mis en évidence lors de l'utilisation de méthodes de classification : (1) la nature objective de la classification est compromise par les choix de la fonction de dissimilarité et de la méthode de classification ; (2) les techniques de classification fournissent des groupes même lorsqu'ils n'existent pas ; (3) les valeurs *ex aequo* dans la matrice de dissimilarités ont pour résultat un certain nombre de dendrogrammes différents. Jackson *et al.* (1989) notent que le résultat d'un classement dépend plus du choix des dissimilarités que du choix de la méthode de classification.

Il nous est apparu que toutes les méthodes de groupements ou d'ordination présentent quelques limites pour représenter, de la façon la plus informative possible, les différences entre collections d'entités. Par exemple, Comas *et al.* (1998, 2000) s'intéressent aux origines des populations humaines. Ils évaluent les différences entre populations à partir de leurs diversités nucléotidiques. Dans leurs analyses, ils ont utilisé à la fois des arbres NJ et des PCoA pour représenter graphiquement ces différences. Ils justifient cette double utilisation par le fait qu'une

représentation en arbre des dissimilarités entre populations peut être lue à tort comme une succession de séparations historiques de sous-populations. En effet, l'arbre peut être interprété à tort comme si les populations étaient connectées entre elles seulement par leurs liens historiques avec une population ancestrale commune, en supposant l'absence de migration entre ces populations. Dans Comas *et al.* (1998, 2000), l'arbre NJ ne fournit pas d'informations supplémentaires par rapport à la PCoA. Dans l'étude de Comas *et al.* (1998), l'arbre présente plusieurs valeurs de bootstrap très fortes et est similaire à la carte factorielle F1×F2 (F1 et F2 désignant les deux premiers axes) de la PCoA. Dans l'analyse de Comas *et al.* (2000), l'arbre a des valeurs de bootstrap basses (80% des valeurs en dessous de 50). Comas *et al.* (2000) concluent alors que cet arbre ne semble pas fournir une structure de populations qui puisse être interprétée. Les deux premiers axes de la PCoA, exprimant 69% de la variation inter-populations, permettent en revanche une meilleure description de la typologie des populations.

Dans toutes ces représentations, les catégories sont absentes. En effet, ni la PCoA, ni la NMDS, ni les méthodes de classification ne permettent une connection entre l'ordination des collections et l'ordination des catégories. Cette connection a été développée dans la structure de l'axiomatisation de Rao (Pavoine *et al.* 2004, cf. annexe 1).

4.3 D ´ (DPC A)

4.3.1 Procédure

Dans la décomposition hiérarchique de l'entropie quadratique (APQE), plusieurs matrices de dissimilarités apparaissent, en particulier la matrice $(S \times S)$ $\Delta^{\text{cat}} = [\delta_{kl}^{\text{cat}}]$ contenant les dissimilarités entre les catégories et la matrice $r \times r$ associée $\Delta^{\text{col}} = [\delta_{ij}^{\text{col}}]$ contenant, parmi les dissimilarités entre les collections, celles qui sont égales à la racine de deux fois la formule de Jensen appliquée à l'entropie quadratique. Si Δ^{cat} est euclidienne alors Δ^{col} l'est aussi (Rao et Nayak 1985, Champely et Chessel 2002). La présence de deux matrices euclidiennes dans l'APQE suggère la possibilité de représentations graphiques par une PCoA (Gower 1984, Gower et Legendre 1986). Deux représentations séparées peuvent être envisagées : une représentation des catégories et une représentation des collections. Les deux représentations se situent dans deux espaces différents et ne peuvent donc pas être directement reliées. La double analyse en coordonnées principales (DPCoA) (Pavoine *et al.* 2004, cf. annexe 1) permet de représenter les deux matrices de façon combinées, de sorte qu'il devient possible non seulement de visualiser des typologies de collections, mais aussi éventuellement d'expliquer ces typologies par des catégories spécifiques, et surtout d'avoir une représentation de l'inertie des points catégories autour des collections.

La première étape de la DPCoA est de placer les catégories et les collections dans un même espace. Considérons la matrice $\mathbf{F} = [f_{ki}]$ de dimensions $S \times r$, contenant en lignes les catégories, en colonnes les collections. Dans cette matrice on trouve les fréquences globales des catégories dans les collections. Soient n_{ki} le nombre d'entités de la collection i appartenant à la catégorie k, n_{+i} le nombre d'entités dans la collection i, n_{k+} le nombre total d'entités appartenant à la catégorie k et n_{++} le nombre total d'entités. La fréquence globale de la catégorie k dans la

collection i est $f_{ki} = n_{ki}/n_{++}$. Soient $\lambda_i = n_{+i}/n_{++}$ le poids de la collection i et $\nu_k = n_{k+}/n_{++}$ le poids global de la catégorie k. La fréquence relative de la catégorie k dans la collection i est $p_{ki} = n_{ki}/n_{+i} = \lambda_i f_{ki}$. Considérons la matrice $\mathbf{D}^{\text{cat}} = [d_{ki} = \delta^2_{ki}/2]$ de dimensions $(S \times S)$. Soient $\mathbf{B}_S = diag(\nu_1, ..., \nu_k, ..., \nu_S)$ la matrice diagonale contenant les poids des catégories et $\mathbf{B}_r = diag(\lambda_1, ..., \lambda_i, ..., \lambda_r)$ la matrice diagonale contenant les poids des collections. Soit $\mathbf{Q} = \mathbf{I}_S - \mathbf{1}_S\mathbf{1}_S^t\mathbf{B}_S$

La première étape de la DPCoA (Pavoine *et al.* 2004, cf. annexe 1) est une PCoA pondérée (Gower 1984, Gower et Legendre 1986, Drouet d'Aubigny 1989). Il s'agit de la PCoA de $\mathbf{\Delta}^{\text{cat}}$ pondérée par $\mathbf{B}_S$:

$$-\mathbf{B}_S^{1/2}\mathbf{Q}\mathbf{D}^{\text{cat}}\mathbf{Q}^t\mathbf{B}_S^{1/2} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^t$$
$$\Rightarrow -\mathbf{Q}\mathbf{D}^{\text{cat}}\mathbf{Q}^t = \mathbf{B}_S^{-1/2}\mathbf{U}\mathbf{\Lambda}\mathbf{U}^t\mathbf{B}_S^{-1/2} = \mathbf{B}_S^{-1/2}\mathbf{U}\mathbf{\Lambda}^{1/2}\left(\mathbf{B}_S^{-1/2}\mathbf{U}\mathbf{\Lambda}^{1/2}\right)^t = \mathbf{X}\mathbf{X}^t,$$

où $\mathbf{U}$ est une matrice $S \times K$ contenant les vecteurs propres ou axes principaux, $\mathbf{\Lambda}$ est une matrice diagonale $S \times S$ contenant les valeurs propres, et $\mathbf{X}$ est une matrice $S \times K$ contenant en ligne les coordonnées de chaque haplotype sur les axes principaux. L'utilisation de cette pondération particulière par rapport à la pondération uniforme généralement utilisée à pour but d'obtenir un nuage de points qui soit, selon les pondérations habituelles utilisées dans la décomposition de l'entropie quadratique, centré à la fois pour les catégories et pour les collections. Seul ce double centrage nous permettra de passer d'une représentation des collections dans l'espace de la typologie des catégories à une projections des catégories dans l'espace de la typologie des collections. Les axes principaux orthogonaux du nuage des catégories expriment alors séquentiellement autant de diversité entre les catégories que possible. Chaque collection est positionnée au barycentre pondéré de ses catégories : $\mathbf{Y} = \mathbf{B}_r^{-1}\mathbf{F}^t\mathbf{X}$. De cette façon, les points collections comme les points catégories sont dans le même espace et sont centrés pour leurs pondérations respectives. De plus la distance euclidienne entre le point de la catégorie k et celui de la catégorie l est exactement égale à δ^{cat}_{kl}, et de la même façon la distance euclidienne entre le point de la collection i et celui de la collection j est précisément δ^{col}_{ij}, égale à la racine de deux fois la différence de Jensen appliquée à l'entropie quadratique. Nous avons établis toutes les démonstrations dans Pavoine *et al.* (2004, cf. annexe 1).

Nous avons donc un espace commun à la fois aux catégories et aux collections. Cet espace représente au mieux les dissimilarités entre catégories. Mais si l'objet est l'étude des dissimilarités entre collections, il reste à rechercher, dans cet espace, les axes principaux des collections :

$$\mathbf{Y}^t\mathbf{B}_r\mathbf{Y} = \mathbf{V}\mathbf{\Phi}\mathbf{V}^t$$

où $\mathbf{V}$ est la matrice $K \times G$ contenant les vecteurs propres ou axes principaux (G est le nombre de valeurs propres strictement positives), et $\mathbf{\Phi}$ est la matrice diagonale de dimension $G \times G$ contenant les valeurs propres. Dans ce nouvel espace défini par les axes principaux des collections, les coordonnées des collections sont données par les lignes de la matrice $\mathbf{YV}$ de dimension $r \times G$, et celles des catégories par les lignes de la matrice $\mathbf{XV}$ de dimension $S \times G$. Les coordonnées des anciens axes (axes principaux des catégories) sur les nouveaux (axes principaux des collections) se trouvent dans les lignes de la matrice $\mathbf{V}$. Ces dernières coordonnées peuvent rendre compte de la qualité de la représentation des dissimilarités entre catégories dans l'espace des collections.

4.3.2 Liens avec d'autres méthodes d'ordination

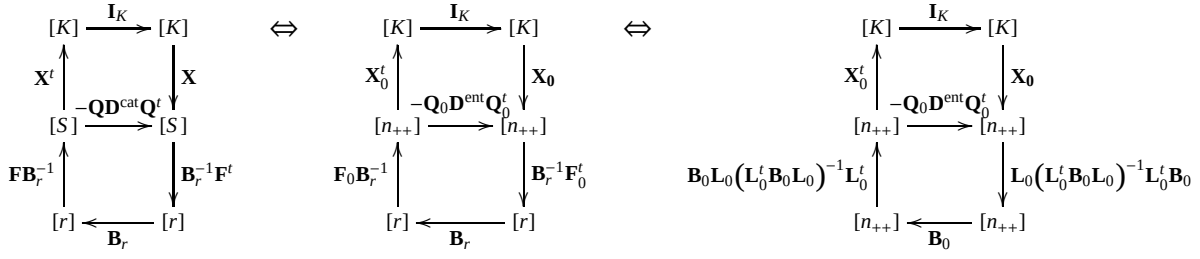
La double analyse en coordonnées principales débute dans l'espace défini par les axes principaux des catégories pondérées par leurs abondances relatives sur l'ensemble du jeu de données. Elle aboutit à l'espace défini par les axes principaux des collections pondérées par leurs tailles relatives. Elle prend pour données de départ une matrice euclidienne de dissimilarités entre catégories et un tableau donnant les abondances, ou les présences, des catégories dans les collections. D'autres analyses sont explicitement ou implicitement basées sur le même schéma. J'ai démontré tous les liens entre méthodes d'ordination dans l'article Pavoine *et al.* (2004, cf. annexe 1). Je propose ici de réécrire ces démonstrations en terme de schéma de dualité.

Pour présenter ces analyses, nous utiliserons les notations suivantes :

- $\mathbf{X}_0$ ($n_{++} \times K$) et $\mathbf{X}$ ($S \times K$) : respectivement la matrice des coordonnées des entités obtenue par PCoA simple et la matrice des coordonnées des catégories obtenue par PCoA pondérée ;
- $\mathbf{B}_0 = diag\left(\frac{1}{n_{++}}, ..., \frac{1}{n_{++}}, ..., \frac{1}{n_{++}}\right) = \frac{1}{n_{++}}\mathbf{I}_{n_{++}}$: matrice ($n_{++} \times n_{++}$) diagonale contenant les poids uniformes des entités ;
- $\mathbf{L}_0$: matrice ($n_{++} \times r$) d'indicatrices donnant l'appartenance de chaque entité à une collection ;
- $\mathbf{F}_0 = \frac{1}{n_{++}}\mathbf{L}_0 = \mathbf{B}_0\mathbf{L}_0$: matrice des fréquences absolues associées ;
- $\mathbf{B}_S$ ($S \times S$) et $\mathbf{B}_r$ ($r \times r$) (cf. partie 4.3.1) : matrices diagonales respectivement des poids des catégories et des poids des collections ; Rq : $\mathbf{L}_0^t\mathbf{B}_0\mathbf{L}_0 = \mathbf{B}_r$;
- $\mathbf{F}$: matrice ($S \times r$) des fréquences absolues des catégories ;
- $\mathbf{Z}_0$ ($n_{++} \times h$) et $\mathbf{Z}$ ($S \times h$) : matrices quantitatives centrées ayant en colonnes des variables, et en lignes respectivement les entités et les catégories ;
- $\mathbf{\Lambda}$ ($K \times K$) la matrice des valeurs propres du nuage des catégories ;
- $\mathbf{\Psi}$ ($G \times G$) la matrice des valeurs propres du nuage des collections ;
- trois matrices de vecteurs propres : $\mathbf{U}_S$ ($S \times K$) et $\mathbf{U}_h$ ($h \times K$) avec en colonnes les coordonnées des axes principaux des catégories respectivement dans l'espace des catégories et dans l'espace de h variables quantitatives centrées, et $\mathbf{V}_h$ ($h \times G$) qui a en colonnes les coordonnées des axes principaux des collections dans un espace défini par h variables quantitatives centrées ;
- $\mathbf{Q}_0 = \mathbf{I}_0 - \mathbf{B}_0\mathbf{1}_0\mathbf{1}_0^t$ et $\mathbf{Q} = \mathbf{I}_S - \mathbf{B}_S\mathbf{1}_S\mathbf{1}_S^t$ deux matrices de centrage.

Nous garderons les mêmes notations pour toutes les méthodes. Néanmoins, deux objets portant le même nom dans deux méthodes différentes sont du même type (par exemple matrice de variables quantitatives centrées) et de mêmes dimensions, mais peuvent représenter des jeux de données différents.

La DPCoA peut être écrite sous la forme de trois schémas équivalents :



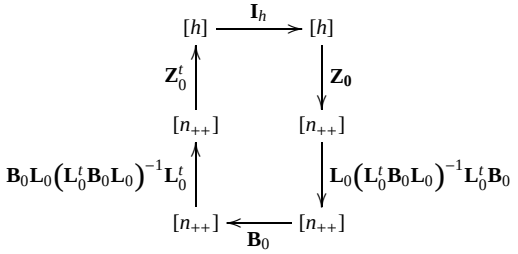
écriture 1 écriture 2 écriture 3

Pour passer de l'écriture 1 à l'écriture 2, il suffit de dérouler les matrices **F** et **X** en recopiant les lignes (ou colonnes) des matrices pour toutes les entités d'une même catégorie. Pour passer de l'écriture 2 à l'écriture 3, il faut se rappeler que $\mathbf{B}_r$ est égal à $\mathbf{L}_0^t\mathbf{B}_0\mathbf{L}_0$. Cette troisième écriture fait apparaître le projecteur $\mathbf{B}_0\mathbf{L}_0\left(\mathbf{L}_0^t\mathbf{B}_0\mathbf{L}_0\right)^{-1}\mathbf{L}_0^t$.

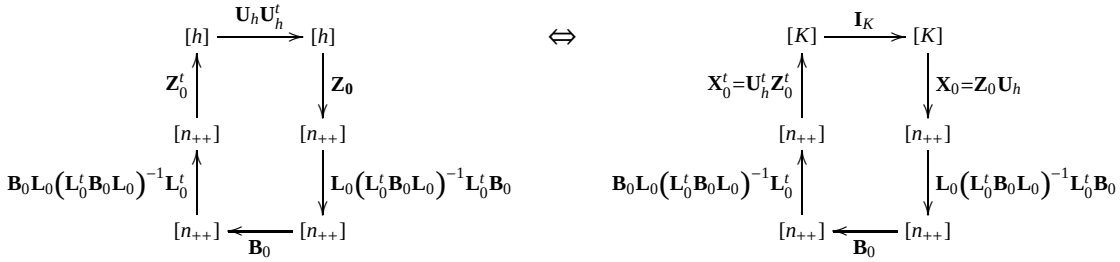
La DPCoA généralise cinq méthodes connues :

- deux analyses basées sur la stratégie des variables instrumentales (variables explicatives) (Rao 1964) :
  - l'analyse en composantes principales inter-classes (BPCA : "Between Principal Component Analysis") (Dolédec et Chessel 1987) ;
  - et l'analyse non symétrique des correspondances (NSCA "Non-Symmetrical Correspondence analysis") (Lauro et D'Ambra 1984).
- deux analyses canoniques :
  - l'analyse discriminante (DA "Discriminant Analysis" = CVA "Canonical Variate Analysis") (Fisher 1936, Rao 1952) ;
  - l'analyse canonique sur coordonnées principales (CAP "Canonical Analysis of Principal coordinates") (Anderson et Willis 2003) ;
- et une analyse qui peut être considérée à la fois comme une analyse canonique et comme une analyse basée sur la stratégie des variables instrumentales :
  - elle est appelée "analyse canonique des correspondances" (CCA "Canonical Correspondence Analysis") (ter Braak 1986, 1987) et "analyse factorielle des correspondances sur variables instrumentales" (AFCvi) (Chessel *et al.* 1987, Lebreton *et al.* 1988a, b).

La BPCA peut être écrite sous la forme du schéma suivant :

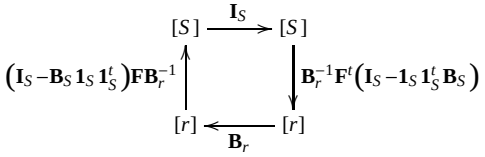


Le calcul des distances euclidiennes entre les lignes de $\mathbf{Z}_0$ conduit à une matrice de distances $\mathbf{\Delta}_0$ entre entités. La PCoA de $\mathbf{\Delta}_0$ pondérée par $\mathbf{B}_0$ correspond à l'analyse en composantes principales (PCA "Principal Component Analysis) de $\mathbf{Z}_0$ centrée pour la pondération donnée par $\mathbf{B}_0$, c'est-à-dire pour la pondération uniforme. Soit $\mathbf{Z}_0\mathbf{U}_h$ la matrice des coordonnées des entités, où $\mathbf{U}_h$ est la matrice des axes principaux associée à $\mathbf{Z}_0^t\mathbf{B}_0\mathbf{Z}_0$. Pour retrouver le lien avec la DPCoA il nous faut passer par un schéma différent du précédent :



Soit $\hat{\mathbf{Z}}_0 = \mathbf{L}_0\left(\mathbf{L}_0^t\mathbf{B}_0\mathbf{L}_0\right)^{-1}\mathbf{L}_0^t\mathbf{B}_0\mathbf{Z}_0$. La décomposition en valeurs singulières du schéma de la BPCA ($\hat{\mathbf{Z}}_0$, $\mathbf{I}_0$, $\mathbf{B}_0$) pourra donner $\hat{\mathbf{Z}}_0{}^t\mathbf{B}_0\hat{\mathbf{Z}}_0 = \mathbf{V}_h\mathbf{\Psi}\mathbf{V}_h^t$. $\mathbf{V}_h$ est la matrice de passage de l'espace défini par les variables quantitatives, colonnes de $\mathbf{Z}_0$, à l'espace défini par les axes principaux des collections. Ainsi pour le deuxième schéma, $\mathbf{U}_h^t\hat{\mathbf{Z}}_0{}^t\mathbf{B}_0\hat{\mathbf{Z}}_0\mathbf{U}_h = \mathbf{U}_h^t\mathbf{V}_h\mathbf{\Psi}\mathbf{V}_h^t\mathbf{U}_h$. En observant que $\mathbf{U}_h^t\mathbf{V}_h$ est la matrice de passage de l'espace des axes principaux des entités à l'espace des axes principaux des collections, on en déduit que, par les deux schémas, les valeurs propres sont identiques, les axes principaux sont identiques, les analyses des deux schémas sont équivalentes.

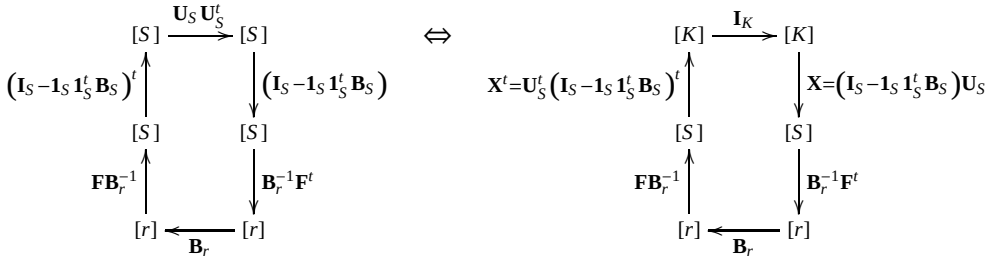
La deuxième analyse basée sur la stratégie des variables instrumentales et généralisée par la DPCoA est la NSCA qui correspond au schéma suivant



Prenons la matrice $\Delta^{\text{cat}} = \sqrt{2}\left(\mathbf{1}_S\mathbf{1}_S^t - \mathbf{I}_S\right)$. Le facteur $\sqrt{2}$ permet que l'inertie d'un nuage de points obtenu selon Δ^{cat} corresponde à l'indice de Gini-Simpson. La PCoA de Δ^{cat} pondérée par $\mathbf{B}_S$ est égale à la PCA centrée de $\mathbf{I}_S$ pondérée par $\mathbf{B}_S$, c'est-à-dire à la décomposition en valeurs singulières du schéma de dualité ($\left(\mathbf{I}_S - \mathbf{1}_S\mathbf{1}_S^t\mathbf{B}_S\right)$, $\mathbf{I}_S$, $\mathbf{B}_S$) :

$$\left(\mathbf{I}_S - \mathbf{1}_S\mathbf{1}_S^t\mathbf{B}_S\right)^t\mathbf{B}_S\left(\mathbf{I}_S - \mathbf{1}_S\mathbf{1}_S^t\mathbf{B}_S\right) = \mathbf{U}_S\mathbf{\Lambda}\mathbf{U}_S^t.$$

Les coordonnées sont dans la matrice $\mathbf{X} = \left(\mathbf{I}_S - \mathbf{1}_S\mathbf{1}_S^t\mathbf{B}_S\right)\mathbf{U}_S$. Comme pour la BPCA, pour retrouver le lien avec la DPCoA, il nous faut passer par un schéma légèrement différent qui correspond à la première écriture de la DPCoA.



Soit $\bar{\mathbf{F}} = \left(\mathbf{I}_S - \mathbf{B}_S\mathbf{1}_S\mathbf{1}_S^t\right)\mathbf{F}\mathbf{B}_r^{-1}$. La décomposition en valeurs singulières du schéma de la NSCA ($\bar{\mathbf{F}}$, $\mathbf{I}_S$, $\mathbf{B}_r$) pourra donner $\bar{\mathbf{F}}\mathbf{B}_r\bar{\mathbf{F}}^t = \mathbf{V}_S\mathbf{\Psi}\mathbf{V}_S^t$. De même pour le deuxième schéma,

$\mathbf{U}_S^t \bar{\mathbf{F}} \mathbf{B}_S \bar{\mathbf{F}}^t \mathbf{U}_S = \mathbf{U}_S^t \mathbf{V}_S \mathbf{\Psi} \mathbf{V}_S^t \mathbf{U}_S$. La matrice $\mathbf{U}_S^t \mathbf{V}_S$ peut être considérée comme une matrice de passage de l'espace défini par les axes principaux des catégories à l'espace défini par les axes principaux des collections. Les valeurs propres sont identiques, les deux analyses sont équivalentes.

Deux autres méthodes généralisées par la DPCoA sont des analyses canoniques :
La première est l'analyse discriminante CVA dont le schéma est

$$\begin{array}{ccc} & \left(\mathbf{Z}_0^t \mathbf{B}_0 \mathbf{Z}_0\right)^{-1} & \\ [h] & \longrightarrow & [h] \\ {\scriptstyle \mathbf{Z}_0^t \mathbf{B}_0 \mathbf{L}_0}\uparrow & & \downarrow{\scriptstyle \mathbf{L}_0^t \mathbf{B}_0 \mathbf{Z}_0} \\ [r] & \longleftarrow & [r] \\ & \left(\mathbf{L}_0^t \mathbf{B}_0 \mathbf{L}_0\right)^{-1} & \end{array}$$

Le calcul des distances de Mahalanobis entre les lignes de $\mathbf{Z}_0$ conduit à une matrice de distances entre entités. Ce calcul revient à l'application d'une PCA centrée de $\mathbf{Z}_0$ suivie du calcul des distances euclidiennes entre les lignes de la matrice $\mathbf{X}_0 = \mathbf{Z}_0 \mathbf{U}_h \mathbf{\Lambda}^{-1/2}$ contenant les composantes principales. De plus $\left(\mathbf{Z}_0^t \mathbf{B}_0 \mathbf{Z}_0\right)^{-1} = \mathbf{U}_h \mathbf{\Lambda}^{-1} \mathbf{U}_h^t$. Une PCoA pondérée par $\mathbf{B}_0$ de la matrice $\mathbf{\Delta_0}$ des distances de Mahalanobis restitue comme matrice de coordonnées des entités exactement la matrice $\mathbf{X}_0$. Ainsi le schéma de la CVA est égal au schéma

$$\begin{array}{ccc} & \mathbf{U}_h \mathbf{\Lambda}^{-1} \mathbf{U}_h^t & \\ [h] & \longrightarrow & [h] \\ {\scriptstyle \mathbf{Z}_0^t}\uparrow & & \downarrow{\scriptstyle \mathbf{Z_0}} \\ [n_{++}] & & [n_{++}] \\ {\scriptstyle \mathbf{B}_0 \mathbf{L}_0 \left(\mathbf{L}_0^t \mathbf{B}_0 \mathbf{L}_0\right)^{-1} \mathbf{L}_0^t}\uparrow & & \downarrow{\scriptstyle \mathbf{L}_0 \left(\mathbf{L}_0^t \mathbf{B}_0 \mathbf{L}_0\right)^{-1} \mathbf{L}_0^t \mathbf{B}_0} \\ [n_{++}] & \underset{\mathbf{B}_0}{\longleftarrow} & [n_{++}] \end{array} \quad \Leftrightarrow \quad \begin{array}{ccc} & \mathbf{I}_K & \\ [K] & \longrightarrow & [K] \\ {\scriptstyle \mathbf{X}_0^t = \mathbf{\Lambda}^{-1/2} \mathbf{U}_h^t \mathbf{Z}_0^t}\uparrow & & \downarrow{\scriptstyle \mathbf{X}_0 = \mathbf{Z}_0 \mathbf{U}_h \mathbf{\Lambda}^{-1/2}} \\ [n_{++}] & & [n_{++}] \\ {\scriptstyle \mathbf{B}_0 \mathbf{L}_0 \left(\mathbf{L}_0^t \mathbf{B}_0 \mathbf{L}_0\right)^{-1} \mathbf{L}_0^t}\uparrow & & \downarrow{\scriptstyle \mathbf{L}_0 \left(\mathbf{L}_0^t \mathbf{B}_0 \mathbf{L}_0\right)^{-1} \mathbf{L}_0^t \mathbf{B}_0} \\ [n_{++}] & \underset{\mathbf{B}_0}{\longleftarrow} & [n_{++}] \end{array}$$

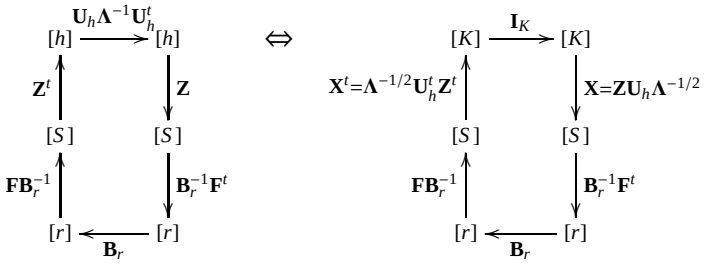
qui correspond à la troisième écriture de la DPCoA.

La deuxième analyse canonique généralisée par la DPCoA est l'analyse canonique sur coordonnées principales CAP qui est en fait l'application de la CVA aux coordonnées principales d'une matrice de dissimilarités. Ces coordonnées principales sont obtenues par PCoA non pondérée. Le lien entre CAP et DPCoA se déduit donc directement de celui démontré ci-dessus entre CVA et DPCoA.

La dernière analyse généralisée par la DPCoA est la CCA ou AFCvi qui est à la fois une analyse canonique et une analyse basée sur la stratégie des variables instrumentales. Elle correspond au schéma suivant

$$\begin{array}{ccc} & \left(\mathbf{Z}^t \mathbf{B}_S \mathbf{Z}\right)^{-1} & \\ [h] & \longrightarrow & [h] \\ {\scriptstyle \mathbf{Z}^t \mathbf{B}_S \left(\mathbf{B}_S^{-1} \mathbf{F} \mathbf{B}_r^{-1}\right)}\uparrow & & \downarrow{\scriptstyle \left(\mathbf{B}_r^{-1} \mathbf{F}^t \mathbf{B}_S^{-1}\right) \mathbf{B}_S \mathbf{Z}} \\ [r] & \underset{\mathbf{B}_r}{\longleftarrow} & [r] \end{array}$$

Comme pour l'analyse discriminante, calculer les distances de Mahalanobis entre les lignes de $\mathbf{Z}$ correspond à faire une PCA centrée de $\mathbf{Z}$ pondérée par $\mathbf{B}_S$, et à calculer les distances euclidiennes entre les lignes de la matrice $\mathbf{X} = \mathbf{Z}\mathbf{U}_h\mathbf{\Lambda}^{-1/2}$ contenant les composantes principales. De plus $(\mathbf{Z}^t\mathbf{B}_S\mathbf{Z})^{-1} = \mathbf{U}_h\mathbf{\Lambda}^{-1}\mathbf{U}_h^t$. Une PCoA pondérée par $\mathbf{B}_S$ de la matrice des distances de Mahalanobis restitue, comme matrice de coordonnées des entités, exactement la matrice $\mathbf{X}$. Ainsi le schéma de la CCA est égal au schéma



ce qui correspond à la première écriture de la DPCoA.

Chaque méthode correspond à une fonction particulière pour calculer les dissimilarités entre entités regroupées ou non en catégories, et est plutôt destinée à des matrices d'indicatrices ou à des tableaux d'abondances. La DPCoA généralise ces méthodes en permettant de choisir n'importe quelle matrice de dissimilarités entre entités ou catégories pourvu que cette matrice soit euclidienne, et de l'appliquer à tout type de données, présences ou abondances.

4.3.3 Lien avec l'APQE, illustration

En écologie, Carnes et Slade (1982), Chessel *et al.* (1982), Gimaret-Carpentier *et al.* (1998), et Pélissier *et al.* (2003) ont montré que des techniques d'ordination et des mesures de diversité peuvent être connectées. Carnes et Slade (1982) proposent de mesurer la diversité, au sens de la taille de niche d'une espèce, par une mesure d'inertie dans l'analyse en composantes principales ou l'analyse discriminante. Chessel *et al.* (1982) montrent que l'analyse des correspondances mesure dans ses marges une diversité en espèces des habitats et une amplitude d'habitat pour chaque espèce, c'est à dire la diversité des habitats que cette espèce utilise. Ils proposent de mesurer la diversité écologique d'un relevé comme la diversité des optimums écologiques des espèces présentes dans ce relevé. On est ici dans la logique de l'entropie quadratique qui est de considérer des informations supplémentaires caractérisant les différences entre les espèces dans la mesure de la diversité. Grâce à la double ordination symétrique des espèces et des relevés par l'analyse des correspondances, ils introduisent des mesures optimales d'amplitude d'habitat et de diversité écologique. Il s'agit respectivement de la variance conditionnelle des points relevés pour une espèce (tenant compte de la distribution de l'espèce entre les relevés) et de la variance conditionnelle des points espèces pour un relevé (tenant compte des abondances relatives des espèces dans les relevés). Gimaret-Carpentier *et al.* (1998), et Pélissier *et al.* (2003) démontrent que la décomposition de l'indice de Gini-Simpson peut être traduite en termes d'inertie de points dans l'analyse non-symétrique des correspondances. En attribuant des poids aux espèces, Pélissier *et al.* (2003) établissent des liens entre l'analyse des correspondances et la richesse, et suggèrent une ordination qui correspondrait à l'indice de Shannon.

Néanmoins, la décomposition de diversité qu'ils obtiennent reste celle d'une décomposition de variance qualitative (indice de Gini-Simpson) pondérée et non une décomposition de l'indice de Shannon ou de la richesse, telles que nous les avons étudiées précédemment (cf. partie 3, discussion partie 3.4.1).

De la même façon, la DPCoA est connectée à la décomposition de l'entropie quadratique. L'entropie quadratique est une mesure d'inertie. En faisant une PCoA pondérée de la matrice des dissimilarités entre catégories et en positionnant les collections au barycentre de leur composition en catégories, nous obtenons un espace dans lequel le nuage des points catégories et celui des points collections sont centrés pour leurs pondérations respectives. Dans cet espace commun aux catégories et aux collections, la distance entre les deux points des catégories k et l est δ_{kl}^{cat}, c'est-à-dire précisément la distance de départ qui a servi à construire l'analyse. De même, la distance entre deux collections i et j est δ_{ij}^{col}, racine carrée de la différence de Jensen appliquée à l'entropie quadratique. Ainsi, toujours dans cet espace commun aux catégories et aux collections, l'inertie des points des catégories pondérées par leurs fréquences relatives au sein d'une collection i est égale à la diversité au sein de cette collection ($H_{\mathbf{\Delta}^{\text{cat}}}(\mathbf{p}_i)$). L'inertie des points des collections pondérées par leurs tailles relatives est égale à la diversité inter-collections ($H_{\mathbf{\Delta}^{\text{col}}}(\lambda)$). L'inertie des points des catégories pondérées par leurs fréquences globales dans l'ensemble des collections mélangées est égale à la diversité totale ($H_{\mathbf{\Delta}^{\text{cat}}}(\mathbf{p}_\bullet)$). Nous avons établi toutes les démonstrations dans Pavoine *et al.* (2004). Lorsque les distances entre catégories sont uniformes, l'entropie quadratique est égale à l'indice de Gini-Simpson et la DPCoA est égale à la NSCA. Nous retrouvons donc le résultat de Pélissier *et al.* (2003), à savoir que la NSCA est liée à la décomposition de l'indice de Gini-Simpson.

Prenons le jeu de données qui a été utilisé dans le premier article de l'AMOVA (Excoffier *et al.* 1992). Les données sont disponibles dans le package 'ade4' de R (objet 'humDNAm'), leurs origines sont décrites dans Excoffier *et al.* (1992) et Schneider *et al.* (2000). Des haplotypes d'ADNmt humain ont été échantillonnés dans dix populations représentant cinq groupes régionaux de deux populations chacun : "Asia" (populations "Tharu" et "Oriental"), "West Africa" (population "Wolof" et "Peul"), "America" (populations "Pima" et "Maya"), "Europe" (populations "Finnish" et "Sicilian"), et "Middle-East" (populations "Israeli Jews" et "Israeli Arabs"). Les dissimilarités entre haplotypes sont calculées en terme de nombre de sites de restrictions différents entre séquences. La matrice Δ^{hap} contenant les racines carrées de ces dissimilarités est euclidienne. Les différences entre les haplotypes peuvent être visualisées par un réseau de longueur minimale (Fig. 28). Les fréquences relatives des haplotypes dans chaque populations sont données dans la figure 29.

Nous étudierons trois questions : (1) Existe-t-il des différences entre populations ou entre groupes ? (2) Quels sont les principaux haplotypes impliqués dans ces différences ? (3) Quelle est la typologie des différences entre populations et entre groupes, en particulier quelles sont les populations et groupes qui se distinguent le plus des autres ? L'AMOVA permet de répondre à la première question. La DPCoA va apporter des éléments de réponses aux deux autres questions. Les calculs qui suivent ont été réalisés dans R avec les fonctions développées spécialement "amova", "randtest.amova", "plot.randtest.amova", "DPCoA" et "plot.DPCoA" et intégrées au package "ade4".

Les résultats de l'AMOVA sont donnés dans la figure 30. En moyenne, 60% des sites de restrictions diffèrent entre deux haplotypes : $SSD(Total)/n_{++} = 403.72/672 = 60\%$. Les com-

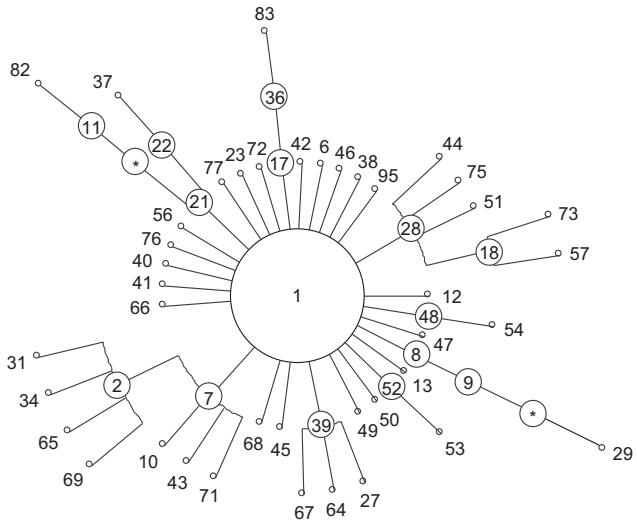


FIG. *28 – Réseau de longueur minimale des 56 haplotypes trouvés dans les 10 populations étudiées. Chaque lien représente une mutation. Les astérisques indiquent deux haplotypes non trouvés dans les populations échantillonnées. L'haplotype 1 central a été trouvé dans toutes les populations. Les feuilles (haplotypes situés en extrémité du réseau) sont indiquées par des petits cercles, les nœuds intermédiaires par des cercles moyens et l'haplotype central par un grand cercle.*

posants de diversité associés sont $SSD(WP)/n_{++} = 47.05\%$, $SSD(AP/WG)/n_{++} = 1.38\%$, et $SSD(AG)/n_{++} = 11.64\%$. La diversité des groupes apparaît bien plus importante que celle des populations au sein des groupes. Les tests de permutation proposés par l'AMOVA sont significatifs pour la variance entre groupes comme pour la variance entre populations au sein des groupes. Ainsi, nous savons qu'il existe des différences entre populations et entre groupes. Quelles sont ces différences ?

Dans l'AMOVA, trois matrices de distances interviennent : Δ^{hap} la matrice de dissimilarités entre haplotypes, Δ^{pop} la matrice de dissimilarités entre populations calculées par la racine de deux fois la différence de Jensen appliquée à l'entropie quadratique, et Δ^{gro} la matrice de dissimilarités entre populations obtenues aussi par différence de Jensen (cf. partie 3.1.3). Appliquons la DPCoA à (1) la matrice d'abondance des haplotypes dans les populations associée à la matrice de distances entre haplotypes (Fig. 31) et à (2) la matrice de poids des populations au sein de chaque groupe associée à la matrice de dissimilarités entre populations (Fig. 32).

Les deux premiers axes principaux des points populations (Fig. 31) expriment respectivement 78% et 8% de la diversité inter-populations. Le premier axe contient donc la majeur partie de l'information : les deux populations africaines diffèrent des autres. Ce premier axe est cohérent avec le premier axe principal du nuage des haplotypes qui lui représente 35% des distances évolutives entre haplotypes. Ainsi, dans ce jeu de données, les populations africaines ont en commun plusieurs haplotypes relativement éloignés des autres d'un point de vue évolutif. Le deuxième axe au contraire ne positionne pas particulièrement les haplotypes selon leurs différences évolutives. Les populations non-Africaines réparties sur cet axe varient en possédant soit différentes proportions de même haplotypes soit différents haplotypes évolutivement proches, soit différents haplotypes évolutivement éloignés en proportion faible. La forme des ellipses

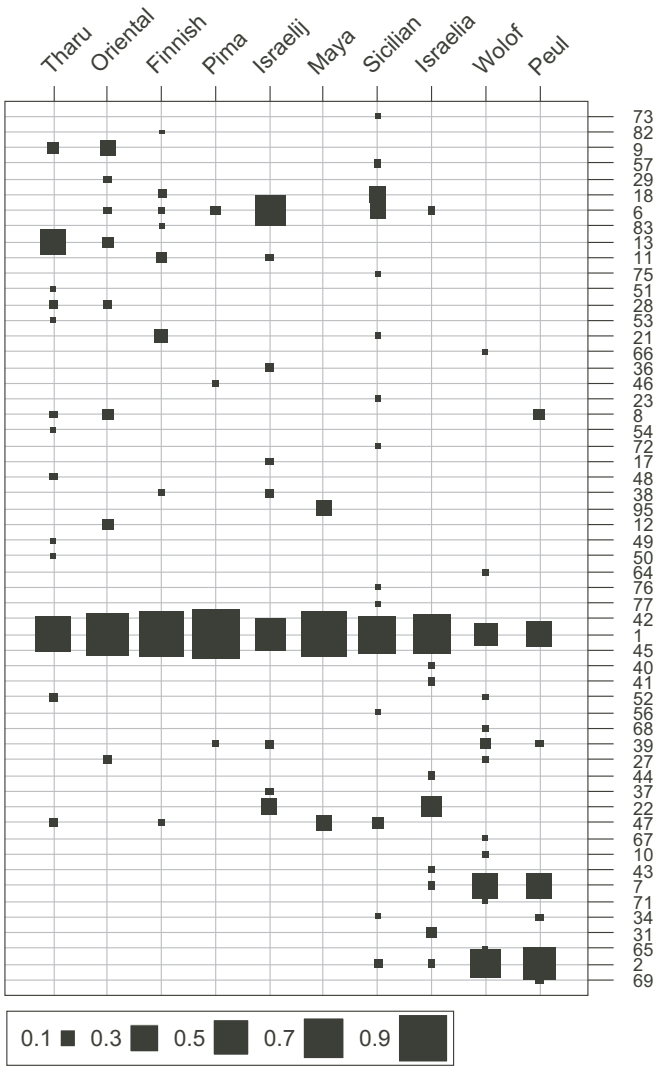


FIG. *29 – Fréquences relatives des haplotypes dans chaque population. Les haplotypes et les populations sont ordonnées selon le premier axe de la dpcoa fournissant une typologie des populations basée sur les haplotypes (Fig. 31). La taille d'un carré est proportionnelle a la fréquence relative d'un haplotype dans une population. L'échelle est donnée en bas de la figure. Cette figure a été réalisée à partir de la fonction 'table.value' du package ade4 (Chessel* et al. *2004) de R (Ihaka et Gentleman 1996).*

indique que les populations Peul, Wolof et Israeli Arab ont les plus grandes inerties internes sur les deux axes conservés.

Les deux premiers axes principaux des points groupes (Fig. 32) expriment 86% et 8% de la diversité inter-groupes. Encore une fois, le premier axe contient la grande majorité de l'information. La typologie des populations est très proche de celles des groupes (Encadré Fig. 32). Les dissimilarités entre populations et entre groupes sont donc toutes bien représentées par la figure 32. La plus grande partie de la diversité entre populations au sein des groupes provient des groupes "Asia" et "Middle East".

Notre méthode d'ordination suppose, comme l'AMOVA, la connaissance a priori de la struc-

```
data(humDNAm)
amovahum <- amova(humDNAm$samples, sqrt(humDNAm$distances), humDNAm$structures)
amovahum # Décomposition de la variation moléculaire
$call
amova(samples = humDNAm$samples, distances = sqrt(humDNAm$distances),
    structures = humDNAm$structures)

$results
                               Df  Sum Sq Mean Sq
Between regions                 4  78.238 19.5595
Between samples Within regions  5   9.285  1.8569
Within samples                662 316.197  0.4776
Total                         671 403.720  0.6017

$componentsofcovariance
                                                  Sigma       %
Variations  Between regions                 0.13381  21.119
Variations  Between samples Within regions 0.02213   3.493
Variations  Within samples                  0.47764  75.387
Total variations                            0.63358 100.000

$statphi
                        Phi
Phi-samples-total   0.24613
Phi-samples-regions 0.04429
Phi-regions-total   0.21119

ramovahum<-randtest(amovahum, 999) # Tests de permutations
ramovahum
class: krandtest
test number:   3
permutation number:   999
  test                         obs   P(X<=obs) P(X>=obs)
1 Variations within samples  0.478 0.001     1
2 Variations between samples 0.022 1         0.001
3 Variations between regions 0.134 1         0.002

plot(ramovahum) # La valeur observée est indiquée par la position du signe
```

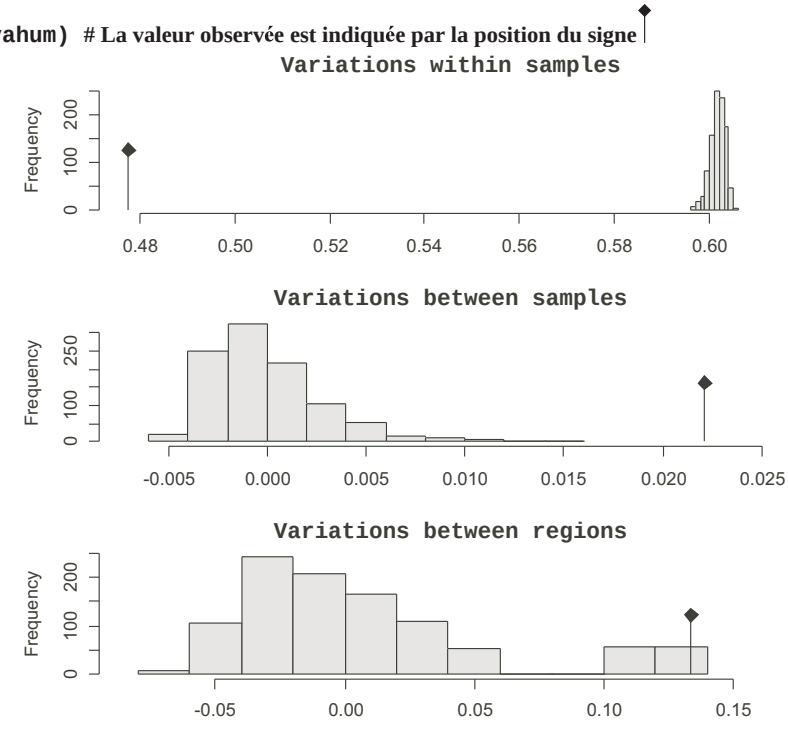


FIG. *30 – Listing dans R (Ihaka et Gentleman 1996) de l'amova appliquée aux données d'Excoffier* et al. *(1992). Les fonctions utilisées ont été développées personnellement pour le package ade4 de R (Chessel* et al. *2004).*

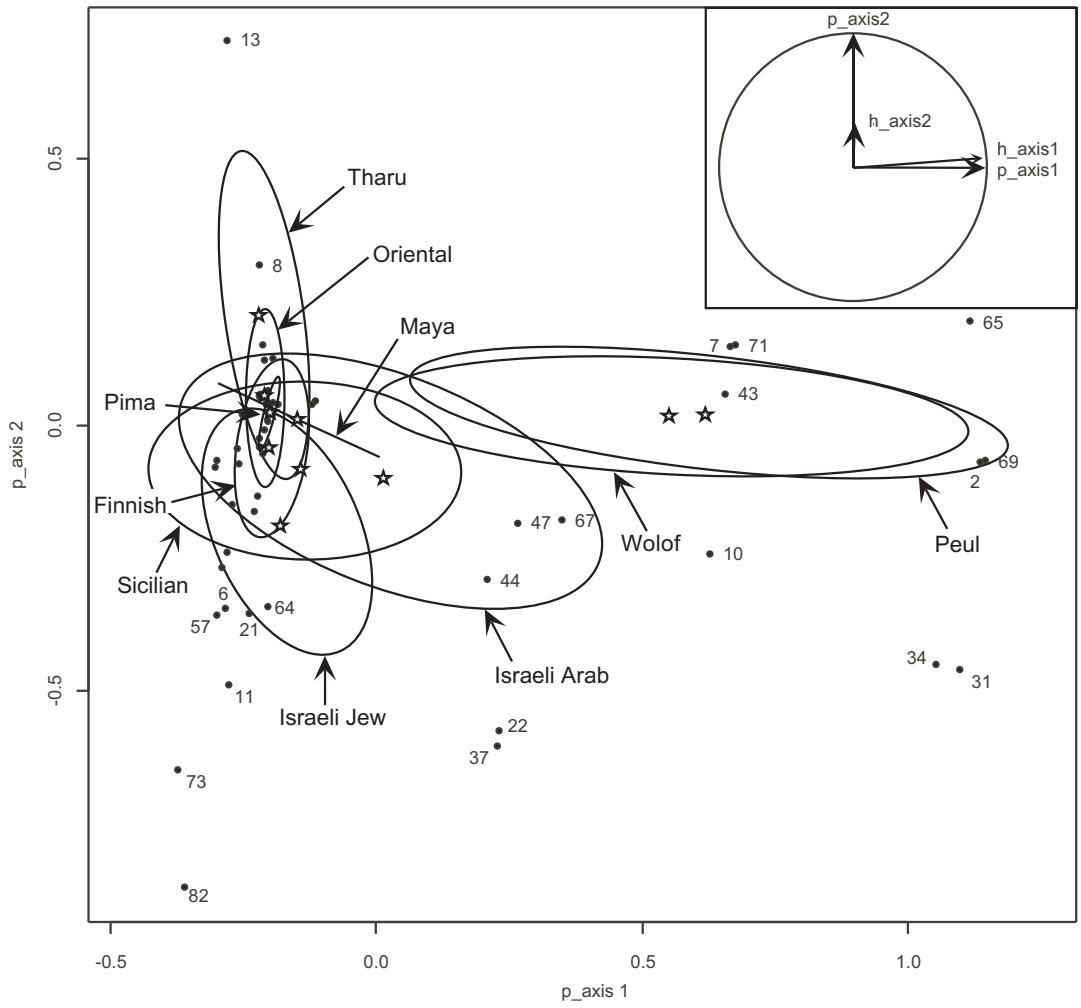


FIG. *31 – Typologie des populations basée sur les haplotypes. Cette représentation est obtenue avec les deux premiers axes principaux du nuage des points populations. Les populations sont indiquées par des étoiles et les haplotypes par des points associés aux numéros correspondant aux notations de l'article Excoffier* et al. *(1992). La diversité de chaque population sur ce plan est représentée par une ellipse centrée sur le point population correspondant. L'encadré contient la projection des deux premiers axes principaux des points haplotypes ("h_axis1" and "h_axis2") sur les deux premiers axes principaux des populations ("p_axis1" and "p_axis2"). La corrélation entre les axes dépend de la direction et de la taille des flèches. Cette figure a été réalisée à partir de la fonction 'DPCoA' développée personnellement pour le package ade4 (Chessel* et al. *2004) de R (Ihaka et Gentleman 1996).*

ture des groupes et populations. Elle ne remplace pas des méthodes telles que celle de Templeton *et al.* (1995), qui identifie des événements de fragmentation des populations. Selon Turner *et al.* (2000), l'AMOVA caractérise les flux de gènes dans des régions fragmentées, et dont on connaît la fragmentation. En complément à l'AMOVA, la DPCoA fournit des informations supplémentaires sur la typologie de groupes. Elle permet de trouver les groupes influençant le plus les différents composants de diversité. Cette influence est dépendante de la composition en haplotypes et des distances évolutives entre ces haplotypes, et est aussi liée à la composition

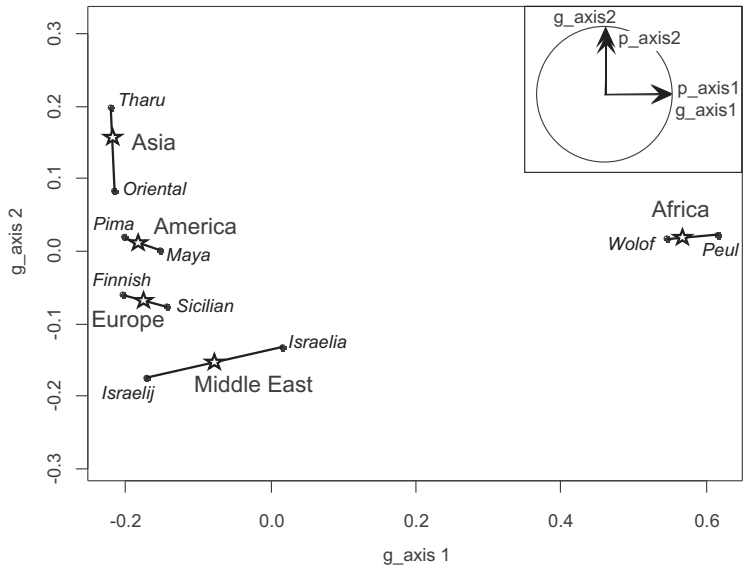


FIG. *32 – Typologie de groupes basée sur les populations. Cette représentation euclidienne est obtenue avec les deux premiers axes principaux des points groupes. Les groupes sont représentés par des étoiles, et les populations par des points. Une ligne joint les populations appartenant à un même groupe. L'encadré donne la projection des deux premiers axes principaux des points populations ("p_axis1" and "p_axis2") sur les deux premiers axes principaux des groupes ("g_axis1" and "g_axis2"). La corrélation entre les axes dépend de la direction et de la taille des flèches. Cette figure a été réalisée à partir de la fonction 'DPCoA' développée personnellement pour le package ade4 (Chessel* et al. *2004) de R (Ihaka et Gentleman 1996).*

en populations et aux distances génétiques entre ces populations. Ainsi les différences entre groupes peuvent être expliquées en termes de compositions et de différences entre populations et entre haplotypes. En étudiant les 945 façons possibles de répartir dix populations dans cinq groupes à raison de deux populations par groupes, Excoffier *et al.* (1992) déterminent que le groupe africain diffère largement des autres. Ce résultat, que le groupe africain se distingue, est obtenu directement par le premier axe de la DPCoA. Pour obtenir le rôle des haplotypes dans les différences entre populations, Excoffier *et al.* (1992) fournissent une visualisation de la diversité intra-population sur le réseau de longueur minimale en indiquant simplement la position des haplotypes de chaque population. Ils concluent que les homoplasies dues à des mutations récurrentes affectent principalement les fréquences des haplotypes localisés aux feuilles du réseau et que à la fois leur faible fréquence et leur position externe minimiseront leur effet sur la partition de la variation. Les résultats de la DPCoA contraste légèrement avec cette conclusion. Globalement, seules les différences entre l'Afrique et les autres groupes sont dues à des haplotypes évolutivement distincts. Les haplotypes n° 2, 65, 69, 31 et 34 sont les principaux responsables de la position du groupe "Africa". Ces cinq haplotypes sont liés dans le réseau, l'haplotype 2 se situe sur un nœud externe et les haplotypes 65, 69, 31 et 34 sont sur des feuilles

reliées à ce nœud. L'haplotype 2, dont la fréquence est forte dans le groupe "Africa", est sans doute la principale source de la position particulière de ce groupe. Au contraire, les différences entre les groupes non-africains ne sont pas dues aux différences évolutives entre haplotypes. La population "Tharu" se particularise par l'haplotype 13, distinct de l'haplotype le plus commun (haplotype 1) par une seule mutation, et fortement présent dans cette population tout en étant absent des autres. A l'opposé sur l'axe 2, les populations d'Europe et d'Israël se distinguent par des haplotypes se situant aux feuilles du réseau et ayant de faibles fréquences. Cette ordination souligne également, grâce aux ellipses, la relation entre la population "Israeli Arab" et celle du groupe "West African".

L'entropie quadratique associée à la DPCoA permet d'avoir une cohérence entre mesure et description des diversités intra et inter. Par exemple, même lorsque des chercheurs utilisent une décomposition additive telle que celle de Gini-Simpson, l'indice de Jaccard est utilisé avec une méthode de classification hiérarchique (*e.g.*, Gering *et al.* 2003) pour décrire les différences entre collections, alors que dans ce cas précis l'analyse non-symétrique des correspondances, cas particulier de la DPCoA, est susceptible de fournir une bien meilleure description des données.

Tout récemment Eckburg *et al.* (2005) montrent que la DPCoA associée à d'autres méthodes statistiques révèle des irrégularités précédemment non reconnues dans l'architecture des communautés microbiennes complexes.

4.4 E DPC A

Nous proposons donc avec la DPCoA une solution au problème de l'ordination d'une matrice d'abondance associée à une matrice de dissimilarités. Mais d'autres problèmes ont été rencontrés.

En particulier, les données d'Excoffier montrent que les collections peuvent être réparties dans des groupes. Collections et groupes forment alors deux niveaux d'une hiérarchie. Ce schéma correspond à la décomposition hiérarchique de l'entropie quadratique (Fig. 33a).

Puis, les données peuvent être organisées sous la forme d'un cube de données d'abondance défini par trois modes. Un des modes correspond à la liste des catégories et est associé à une matrice de dissimilarités. Les deux autres modes sont deux facteurs croisés. Cela correspond en fait à la décomposition croisée de l'entropie quadratique (Fig. 33b).

Enfin, les entités peuvent être réparties en catégories selon plusieurs critères ou marqueurs. A chaque critère correspond une DPCoA. La question qui peut être posée est : les critères aboutissent-ils à la même typologie des collections (Fig. 33c) ?

4.4.1 DPCoA hiérarchique

Considérons deux niveaux hiérarchiques : celui des collections et celui de groupes de collections. Comme dans l'APQE nous considérerons r groupes, le groupe i contenant m_i collections. Les entités de chaque collection se répartissent dans S catégories. Soient n_{kji}, n_{k+i} et n_{k++} le nombre d'entités associées à la catégorie k respectivement dans la collection j du groupe i, dans

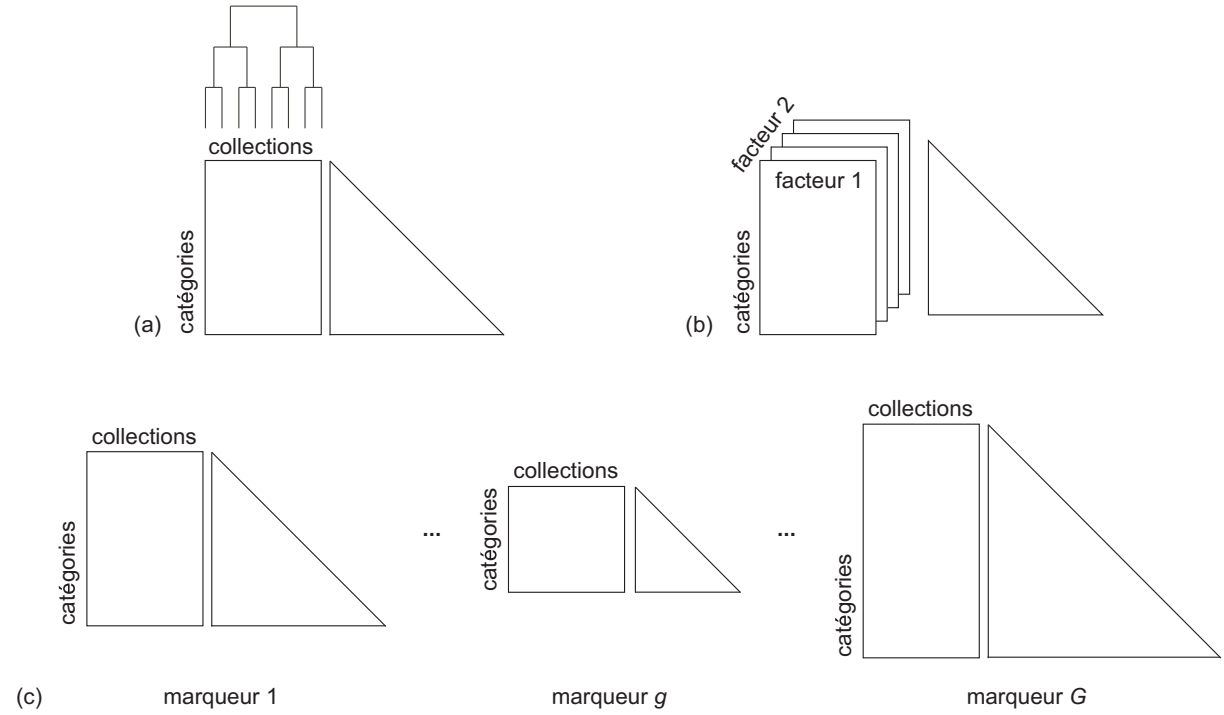


FIG. *33 – Extensions de la DPCoA, types de données : (a) DPCoA hiérarchique, hiérarchie sur les collections ; (b) DPCoA croisée, deux facteurs croisés (collections = facteur 1 × facteur 2) ; (c) DPCoA multiple, plusieurs marqueurs pour définir des catégories. Chaque rectangle représente un tableau d'abondance ou de présence des catégories dans des collections ; et chaque triangle désigne une matrice de dissimilarités entre ces catégories.*

le groupe i, et au total. Notons n_{+ji}, n_{++i} et n_{+++} les nombres d'entités dans la collection j du groupe i, dans le groupe i, et au total. Les différences entre catégories sont mesurées dans la matrice $\mathbf{\Delta}^{\text{cat}}$ qui est euclidienne. Les poids absolus attribués aux catégories, collections et groupes sont respectivement

$$\varepsilon = (\varepsilon_1, ..., \varepsilon_k, ..., \varepsilon_S)\text{, où } \varepsilon_k = n_{k++}/n_{+++}$$
$$\mu^*_+ = \left(\mu^*_{11}, ..., \mu^*_{ji}, ..., \mu^*_{m_r r}\right)\text{, où } \mu^*_{ji} = n_{+ji}/n_{+++}$$
$$\lambda = (\lambda_1, ..., \lambda_i, ..., \lambda_r)\text{, où } \lambda_i = n_{++i}/n_{+++}.$$

De plus les poids relatifs des collections dans le groupe i sont donnés par le vecteur

$$\mu_i = \left(0, ..., 0, \mu_{1i}, ..., \mu_{ji}, ..., \mu_{m_i i}, 0, ..., 0\right)\text{, où } \mu_{ji} = n_{+ji}/n_{++i};$$

dans ce vecteur, un poids nul est donné aux collections n'appartenant pas au groupe i. Définissons les matrices suivantes

$$\mathbf{P}_{/G} = \left[p_{k\bullet i} = \frac{n_{k+i}}{n_{++i}}\right]$$
$$\mathbf{P}_{/C} = \left[p_{kji} = \frac{n_{kji}}{n_{+ji}}\right],$$

où $\mathbf{P}_{/G}$ (dimensions $S \times r$) contient les distributions de fréquences des catégories dans les groupes, et $\mathbf{P}_{/C}$ ($S \times m_+$) contient les distributions de fréquences dans chaque collection.

La première étape est une PCoA de $\mathbf{\Delta}^{\text{cat}}$ pondérée par les $\{\varepsilon_k\}$. Les coordonnées des catégories sont dans une matrice que nous appellerons $\mathbf{X}$. Les groupes et collections sont positionnés aux barycentres de leurs catégories :

$$\begin{aligned}\mathbf{Y}_G &= \mathbf{P}_{/G}^t\mathbf{X}\\ \mathbf{Y}_C &= \mathbf{P}_{/C}^t\mathbf{X}\end{aligned}$$

Dans cet espace, l'inertie des points des groupes, pondérée par $\{\lambda_i\}$, est égale à la diversité inter-groupes ($H_{\mathbf{\Delta}^{\text{gro}}}(\lambda)$) dans l'APQE. L'inertie des points des collections, pondérée par $\{\mu_{ji}\}$, est égale à la diversité inter-collections dans le groupe i de l'APQE ($H_{\mathbf{\Delta}^{\text{col}}}(\mu_i)$). La moyenne de ces diversités inter-collections pondérées par λ_i est égale à la diversité moyenne entre les collections à l'intérieur des groupes. L'inertie pondérée par $\{p_{kji}\}$ des points catégories est égale à la diversité au sein de la collection j du groupe i ($H_{\mathbf{\Delta}^{\text{cat}}}\left(\mathbf{p}_{ji}\right)$). La moyenne, pondérée par $\mu_{ji}\lambda_i$, des diversités à l'intérieur de plusieurs collections est égale à la diversité moyenne au sein des collections. L'inertie totale pondérée par $\{\varepsilon_k\}$ des points catégories est égale à la diversité totale (H_T). Nous avons donc un espace euclidien correspondant à l'APQE.

Démonstration : Soient M_k, C_{ji}, G_i, et G les points correspondant respectivement à l'espèce k, à la collection j du groupe i, au groupe i, et au centre de l'espace. Soit $\mathbf{x}_k$ la ligne k de la matrice $\mathbf{X}$ contenant les coordonnées du point M_k. Notons $\mathbf{c}_{ji}$, et $\mathbf{g}_i$ les vecteurs qui désignent respectivement lignes ji, et i des matrices $\mathbf{Y}_C$, et $\mathbf{Y}_G$. Ces vecteurs contiennent respectivement les coordonnées des points C_{ji} et G_i. Notons $\mathbf{g}$ le vecteur nul contenant les coordonnées du point G. D'après les définitions données ci-dessus, les relations suivantes existent entre ces vecteurs :

$$\begin{aligned}\mathbf{c}_{ji} &= \sum_{k=1}^{S} p_{kji}\mathbf{x}_k\\ \mathbf{a}_i &= \sum_{j=1}^{m_i}\mu_{ji}\mathbf{c}_{ji}\\ \mathbf{g} &= \sum_{i=1}^{r}\lambda_i\mathbf{a}_i = \sum_{i=1}^{r}\sum_{j=1}^{m_i}\mu_{ji}\mathbf{c}_{ji} = \sum_{k=1}^{S}\varepsilon_k\mathbf{x}_k = \mathbf{0}.\end{aligned}$$

Nous pouvons ainsi démontrer que les composants de diversité sont des mesures d'inertie :

$$\begin{aligned}H_{\mathbf{\Delta}^{\text{gro}}}(\lambda) &= \frac{1}{2}\sum_{i=1}^{r}\sum_{i'=1}^{r}\lambda_i\lambda_{i'}\delta_{ii'}^{\text{gro}\,2} = \frac{1}{2}\sum_{i=1}^{r}\sum_{i'=1}^{r}\lambda_i\lambda_{i'}\|\mathbf{a}_i-\mathbf{a}_{i'}\|^2\\ &= \sum_{i=1}^{r}\lambda_i\|\mathbf{a}_i-\mathbf{g}\|^2,\\ H_{\mathbf{\Delta}^{\text{col}}}\left(\mu_{ji}\right) &= \frac{1}{2}\sum_{j=1}^{m_i}\sum_{j'=1}^{m_i}\mu_{ji}\mu_{ji'}\delta_{jj'}^{\text{col}\,2} = \frac{1}{2}\sum_{j=1}^{m_i}\sum_{j'=1}^{m_i}\mu_{ji}\mu_{j'i}\|\mathbf{c}_{ji}-\mathbf{c}_{j'i}\|^2\\ &= \sum_{j=1}^{m_i}\mu_{ji}\|\mathbf{c}_{ji}-\mathbf{g}_i\|^2,\\ H_{\mathbf{\Delta}^{\text{cat}}}\left(\mathbf{p}_{ji}\right) &= \frac{1}{2}\sum_{k=1}^{S}\sum_{l=1}^{S}p_{kji}p_{lji}\delta_{kl}^{\text{cat}\,2} = \frac{1}{2}\sum_{k=1}^{S}\sum_{l=1}^{S}p_{kji}p_{lji}\|\mathbf{x}_k-\mathbf{x}_l\|^2\end{aligned}$$

$$
\begin{aligned}
&= \sum_{k=1}^{S} p_{ki\,j} \|\mathbf{x}_k - \mathbf{c}_{ji}\|^2, \\
H_{\Delta^{\text{cat}}}(\varepsilon) &= \frac{1}{2} \sum_{k=1}^{S} \sum_{l=1}^{S} \varepsilon_k \varepsilon_l \delta_{kl}^{\text{cat}\,2} = \frac{1}{2} \sum_{k=1}^{S} \sum_{l=1}^{S} \varepsilon_k \varepsilon_l \|\mathbf{x}_k - \mathbf{x}_l\|^2 \\
&= \sum_{k=1}^{S} \varepsilon_k \|\mathbf{x}_k - \mathbf{g}\|^2.
\end{aligned}
$$

Dans cet espace, nous pouvons rechercher les axes principaux des points groupes pour analyser la diversité inter-groupes. Les axes principaux des points collections nous informent de la diversité inter-collections. Une représentation de la diversité inter-collections intra-groupe peut être obtenue en couplant la méthode de la DPCoA avec une analyse intra-classe. Cette méthode consiste à déplacer les nuages de points collections en plaçant leur barycentre pour chaque groupe à l'origine.

Avec les données d'Excoffier, les espaces des deux DPCoA (haplotypes × populations et haplotypes × régions) sont si proches (cf. encadré Fig. 32) qu'ils sont presque superposables sans avoir besoin de passer par la DPCoA hiérarchique. Dans les autres cas, la méthode se révélera informative.

4.4.2 DPCoA croisée

Parfois, les collections sont définies par le croisement de deux facteurs comme le temps et l'espace, ou deux traitements différents sur des dispositifs expérimentaux. Les deux facteurs A et B ajoutés à la liste d'espèces forment un cube de données. Facteur A, facteur B et liste d'espèces sont appelés "modes" du cube de données.

La relation entre le facteur A et la liste d'espèces est révélée séparément pour chaque modalité du facteur B par une analyse des correspondances. Pour comparer ces analyses séparées, l'analyse canonique de Foucart propose d'exécuter l'analyse des correspondances d'une matrice moyenne pour obtenir la structure de l'organisation des espèces selon les modalités du facteur A - structure résultant d'un compromis entre les diverses modalités du facteurs B - et de projeter les analyses séparées dans l'espace du "compromis" (Foucart 1978). Nous avons exploré cette analyse dans un article soumis : Pavoine, S., J. Blondel, and D. Chessel. 2006. A new technique for ordering three-dimensional data sets in convergence studies. (cf. annexe 6)

Le problème qui se pose ici est de rajouter en plus une matrice de dissimilarités entre les espèces, et donc de rentrer dans la logique de la DPCoA avec cette fois deux facteurs dont le croisement définit les collections.

Dans la DPCoA, un espace euclidien correspondant à la décomposition de l'entropie quadratique sur un facteur a été défini. Dans la DPCoA croisée, il va être recherché un espace euclidien correspondant à la décomposition croisée de l'entropie quadratique (appelée ANOQE, ANalysis Of Quadratic Entropy) avec deux facteurs croisés.

Soient deux facteurs croisés A et B. Le premier comprend r modalités et le second m. Considérons de plus S catégories. Soit n_{kij} l'abondance de la catégorie k dans la collection ij associée à la modalité i du facteur A et à la modalité j du facteur B. Soient n_{+ij} le nombre total d'entités dans cette collection, n_{+i+} le nombre d'entités associées à la modalité i du facteur A, n_{++j} le nombre d'entités associées à la modalité j du facteur B, et n_{+++} le nombre total d'entités. Soit $\mathbf{\Delta}^{\text{cat}}$ une matrice euclidienne de dissimilarités entre les catégories. Dans l'ANOQE à deux facteurs croisés, chaque catégorie, modalité, et collection est associée à un poids. Les modalités du facteur A sont associées aux poids $\{\lambda_i = n_{+i+}/n_{+++}\}$. De la même façon, les poids des modalités du facteur B sont $\{\mu_j = n_{++j}/n_{+++}\}$. A la collection ij est attribué le poids $w_{ij} = \lambda_i\mu_j$. Ainsi la catégorie k possède un poids total égal à $\varepsilon_k = \sum_{ij} w_{ij}(n_{kij}/n_{+ij})$. Définissons les matrices suivantes

$$\begin{aligned}
\mathbf{P}_{/A} &= \left[p_{ki\bullet} = \sum_{j=1}^{m} \mu_j \frac{n_{kij}}{n_{+ij}}\right] \\
\mathbf{P}_{/B} &= \left[p_{k\bullet j} = \sum_{i=1}^{r} \lambda_i \frac{n_{kij}}{n_{+ij}}\right] \\
\mathbf{P}_{/C} &= \left[p_{kij} = \frac{n_{kij}}{n_{+ij}}\right],
\end{aligned}$$

$\mathbf{P}_{/A}$ (dimensions $S \times r$) et $\mathbf{P}_{/B}$ ($S \times m$) contiennent les distributions de fréquences marginales respectivement pour les facteurs A et B, et $\mathbf{P}_{/C}$ ($S \times rm$) contient les distributions de fréquences dans chaque collection.

La première étape de la DPCoA croisée est une PCoA de $\mathbf{\Delta}^{\text{cat}}$ pondérée par $\{\varepsilon_k\}$. Les coordonnées des catégories sur les axes principaux obtenus par la PCoA sont données dans les lignes d'une matrice notée $\mathbf{X}$. Elle sont centrées pour la pondération $\{\varepsilon_k\}$: notons $\mathbf{B}_{cat} = diag\{\varepsilon_k\}$ la matrice diagonale contenant les poids des catégories, $\mathbf{X}\mathbf{B}_{cat}\mathbf{1}_S = (0, ..., 0)^t$. Les collections et les modalités sont positionnées aux barycentres de leurs catégories. Leurs coordonnées sont données dans les lignes des matrices suivantes :

$$\begin{aligned}
\mathbf{Y}_A &= \mathbf{P}_{/A}^t\mathbf{X} \\
\mathbf{Y}_B &= \mathbf{P}_{/B}^t\mathbf{X} \\
\mathbf{Y}_C &= \mathbf{P}_{/C}^t\mathbf{X}
\end{aligned}$$

Elles sont toutes centrées pour leurs pondérations respectives : soit $\mathbf{B}_A = diag(\lambda_1, ..., \lambda_i, ..., \lambda_r)$, $\mathbf{B}_B = diag(\mu_1, ..., \mu_j, ..., \mu_m)$, et $\mathbf{B}_C = diag(w_{11}, ..., w_{ij}, ..., w_{rm})$ les matrices diagonales contenant respectivement les poids des modalités du facteur A, les poids des modalités du facteur B et les poids des collection, $\mathbf{Y}_A\mathbf{B}_A\mathbf{1}_r = (0, ..., 0)^t$, $\mathbf{Y}_B\mathbf{B}_B\mathbf{1}_m = (0, ..., 0)^t$ et $\mathbf{Y}_C\mathbf{B}_C\mathbf{1}_{rm} = (0, ..., 0)^t$.

Pour l'instant nous sommes dans un espace euclidien dont les axes représentent aux mieux les différences entre catégories. Dans cet espace, l'inertie pondérée par $\{\lambda_i\}$ des points modalités du facteur A est égale à l'effet du facteur A (H_A) dans la décomposition croisée de l'entropie quadratique. L'inertie pondérée par $\{\mu_j\}$ des points modalités du facteur B est égale à l'effet du facteur B (H_B). L'inertie pondérée par $\{p_{kij}\}$ des points catégories est égale à la diversité au sein de la collection ij ($H_{C_{ij}}$). La moyenne, pondérée par w_{ij}, des diversités au sein des collections est

égale à la diversité intra-collection ($\overline{H_C}$). L'inertie totale pondérée par $\{\varepsilon_k\}$ des points catégories est égale à la diversité totale (H_T). De plus l'effet de l'interaction entre les facteurs A et B est égale à $H_{AB} = H_T - \left(H_A + H_B + \overline{H_C}\right)$. Nous avons donc un espace euclidien correspondant à la décomposition croisée de l'entropie quadratique.

Démonstration : Soient M_k, C_{ij}, A_i, B_j et G les points correspondant respectivement à l'espèce k, à la collection ij, à la modalité i du facteur A, à la modalité j du facteur B, et au centre de l'espace. Soit $\mathbf{x}_k$ la ligne k de la matrice $\mathbf{X}$ contenant les coordonnées du point M_k. Les vecteurs $\mathbf{c}_{ij}$, $\mathbf{a}_i$, et $\mathbf{b}_j$ sont formés respectivement de la ligne ij de la matrice $\mathbf{Y}_C$, de la ligne i de la matrice $\mathbf{Y}_A$ et de la ligne j de la matrice $\mathbf{Y}_B$. Ces vecteurs contiennent respectivement les coordonnées des points C_{ij}, A_i, B_j. Notons $\mathbf{g}$ le vecteur nul $(0, ..., 0)^t$ correspondant aux coordonnées du point G. D'après les définitions données ci-dessus, les relations suivantes existent entre ces différents vecteurs :

$$\begin{aligned}
\mathbf{c}_{ij} &= \sum_{k=1}^{S} p_{kij}\mathbf{x}_k \\
\mathbf{a}_i &= \sum_{k=1}^{S}\sum_{j=1}^{m} \mu_j p_{kij}\mathbf{x}_k \\
\mathbf{b}_j &= \sum_{k=1}^{S}\sum_{i=1}^{r} \lambda_i p_{kij}\mathbf{x}_k \\
\mathbf{g} &= \sum_{i=1}^{r} \lambda_i \mathbf{a}_i = \sum_{j=1}^{m} \mu_j \mathbf{b}_j = \sum_{i=1}^{r}\sum_{j=1}^{m} w_{ij}\mathbf{c}_{ij} = \sum_{k=1}^{S} \varepsilon_k \mathbf{x}_k = (0, ..., 0)^t .
\end{aligned}$$

On démontre ainsi que les composants de diversité sont des mesures d'inertie :

$$\begin{aligned}
H_A &= \frac{1}{2}\sum_{i=1}^{r}\sum_{i'=1}^{r} \lambda_i\lambda_{i'} {\delta_{ii'}^{\mathrm{A}}}^2 = \frac{1}{2}\sum_{i=1}^{r}\sum_{i'=1}^{r} \lambda_i\lambda_{i'} \|\mathbf{a}_i - \mathbf{a}_{i'}\|^2 \\
&= \sum_{i=1}^{r} \lambda_i \|\mathbf{a}_i - \mathbf{g}\|^2, \\
H_B &= \frac{1}{2}\sum_{j=1}^{m}\sum_{j'=1}^{m} \mu_j\mu_{j'} {\delta_{jj'}^{\mathrm{B}}}^2 = \frac{1}{2}\sum_{j=1}^{m}\sum_{j'=1}^{m} \mu_j\mu_{j'} \|\mathbf{b}_j - \mathbf{b}_{j'}\|^2 \\
&= \sum_{j=1}^{m} \mu_j \|\mathbf{b}_j - \mathbf{g}\|^2, \\
H_{C_{ij}} &= \frac{1}{2}\sum_{k=1}^{S}\sum_{l=1}^{S} p_{kij}p_{lij}\delta_{kl}^{\mathrm{cat}2} = \frac{1}{2}\sum_{k=1}^{S}\sum_{l=1}^{S} p_{kij}p_{lij} \|\mathbf{x}_k - \mathbf{x}_l\|^2 \\
&= \sum_{k=1}^{S} p_{kij} \|\mathbf{x}_k - \mathbf{c}_{ij}\|^2, \\
H_T &= \frac{1}{2}\sum_{k=1}^{S}\sum_{l=1}^{S} \varepsilon_k\varepsilon_l\delta_{kl}^{\mathrm{cat}2} = \frac{1}{2}\sum_{k=1}^{S}\sum_{l=1}^{S} \varepsilon_k\varepsilon_l \|\mathbf{x}_k - \mathbf{x}_l\|^2 \\
&= \sum_{k=1}^{S} \varepsilon_k \|\mathbf{x}_k - \mathbf{g}\|^2.
\end{aligned}$$

Pour le terme d'interaction entre A et B,

$$H_{AB} = \frac{1}{2}\sum_{i=1}^{r} \lambda_i \sum_{j=1}^{m}\sum_{j'=1}^{m} \mu_j\mu_{j'} {\delta_{iji'j'}^{\mathrm{C}}}^2 - H_B = \frac{1}{2}\sum_{i=1}^{r} \lambda_i \sum_{j=1}^{m}\sum_{j'=1}^{m} \mu_j\mu_{j'} \|\mathbf{c}_{ij} - \mathbf{c}_{ij'}\|^2 - H_B$$

$$
\begin{aligned}
&= \sum_{i=1}^{r}\sum_{j=1}^{m} w_{ij}\|\mathbf{c}_{ij}-\mathbf{b}_j\|^2 - H_B,\\
H_{AB} &= \frac{1}{2}\sum_{j=1}^{m}\mu_j \sum_{i=1}^{r}\sum_{i'=1}^{r}\lambda_i\lambda_{i'}{\delta^{\mathrm{C}}_{iji'j}}^2 - H_A = \frac{1}{2}\sum_{j=1}^{m}\mu_j \sum_{i=1}^{r}\sum_{i'=1}^{r}\lambda_i\lambda_{i'}\|\mathbf{c}_{ij}-\mathbf{c}_{i'j}\|^2 - H_A\\
&= \sum_{i=1}^{r}\sum_{j'=1}^{m} w_{ij}\|\mathbf{c}_{ij}-\mathbf{a}_i\|^2 - H_A.
\end{aligned}
$$

Il s'agit maintenant, connaissant l'un des deux facteurs A et B, d'analyser l'autre. Considérons par exemple l'analyse du facteur A, sachant B. L'analyse de B, sachant A, pourra être déduite par symétrie. Cette analyse doit nous donner les changements, en fonction de chaque modalité du facteur B, dans la typologie des modalités du facteur A fournie par les compositions en catégories. Les étapes sont les suivantes :

1. projection de l'ensemble des points dans un espace orthogonal au facteur B,
2. projection de l'ensemble des points sur les axes principaux des modalités du facteur A.

Etape 1 : Le projecteur orthogonal dans l'espace des modalités de B s'écrit $\mathbf{Y}_B^t\left(\mathbf{Y}_B\mathbf{Y}_B^t\right)^{-1}\mathbf{Y}_B$.

En effet, les axes principaux et les composantes principales du nuage de points pour les modalités du facteur B sont donnés par la décomposition en valeurs singulières du triplet ($\mathbf{Y}_B$, $\mathbf{I}$, $\mathbf{B}_B$), où $\mathbf{B}_B = diag(\{\mu_j\})$

$$
\begin{aligned}
\mathbf{Y}_B &= \mathbf{V}\mathbf{\Lambda}^{1/2}\mathbf{U}^t\\
\mathbf{Y}_B^t\mathbf{B}_B\mathbf{Y}_B &= \mathbf{U}\mathbf{\Lambda}\mathbf{U}^t\\
\mathbf{Y}_B\mathbf{Y}_B^t &= \mathbf{V}\mathbf{\Lambda}\mathbf{V}^t
\end{aligned}
$$

Les matrices $\mathbf{U}$ et $\mathbf{V}$ vérifient $\mathbf{U}^t\mathbf{U} = \mathbf{I}$ et $\mathbf{V}^t\mathbf{B}_B\mathbf{V} = \mathbf{I}$. Le projecteur orthogonal dans l'espace des modalités de B est $\mathbf{U}\left(\mathbf{U}^t\mathbf{U}\right)^{-1}\mathbf{U}^t = \mathbf{U}\mathbf{U}^t$. Les relations suivantes relient les axes principaux et les composantes principales : $\mathbf{V} = \mathbf{Y}_B\mathbf{U}\mathbf{\Lambda}^{-1/2}$ et $\mathbf{U} = \mathbf{Y}_B^t\mathbf{B}_B\mathbf{V}\mathbf{\Lambda}^{-1/2}$. Ainsi $\mathbf{U}\mathbf{U}^t = \mathbf{Y}_B^t\mathbf{B}_B\mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{V}^t\mathbf{B}_B\mathbf{Y}_B^t$. Or

$$\mathbf{V}\mathbf{\Lambda}\mathbf{V}^t\left(\mathbf{B}_B\mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{V}^t\mathbf{B}_B\right)\mathbf{V}\mathbf{\Lambda}\mathbf{V}^t = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^t,$$

donc

$$\mathbf{B}_B\mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{V}^t\mathbf{B}_B = \left(\mathbf{V}\mathbf{\Lambda}\mathbf{V}^t\right)^{-1} = \left(\mathbf{Y}_B\mathbf{Y}_B^t\right)^{-1},$$

et

$$\mathbf{U}\mathbf{U}^t = \mathbf{Y}_B^t\left(\mathbf{Y}_B\mathbf{Y}_B^t\right)^{-1}\mathbf{Y}_B$$

Les coordonnées des catégories, des modalités du facteur A, et des collections sont données par

$$
\begin{aligned}
\mathbf{X}_{/\bar{B}} &= \mathbf{X}\left(\mathbf{I}-\mathbf{Y}_B^t\left(\mathbf{Y}_B\mathbf{Y}_B^t\right)^{-1}\mathbf{Y}_B\right)\\
\mathbf{Y}_{A_{/\bar{B}}} &= \mathbf{Y}_A\left(\mathbf{I}-\mathbf{Y}_B^t\left(\mathbf{Y}_B\mathbf{Y}_B^t\right)^{-1}\mathbf{Y}_B^t\right)\\
\mathbf{Y}_{C_{/\bar{B}}} &= \mathbf{Y}_C\left(\mathbf{I}-\mathbf{Y}_B^t\left(\mathbf{Y}_B\mathbf{Y}_B^t\right)^{-1}\mathbf{Y}_B\right)
\end{aligned}
$$

Etape 2 : Dans ce nouvel espace, les axes principaux des modalités du facteur A sont donnés par la décomposition en valeurs singulières de ($\mathbf{Y}_{A_{/\bar{B}}}$, $\mathbf{I}$, $\mathbf{B}_A$), où $\mathbf{B}_A = diag(\{\lambda_i\})$

$$\mathbf{Y}^t_{A_{/\bar{B}}}\mathbf{B}_A\mathbf{Y}_{A_{/\bar{B}}} = \mathbf{U}_A\mathbf{\Psi}\mathbf{U}^t_A.$$

Les coordonnées finales des catégories, des modalités du facteur A et des collections sont alors données par

$$\begin{aligned}
\mathbf{X}_{/A\cap\bar{B}} &= \mathbf{X}_{/\bar{B}}\mathbf{U}_A \\
\mathbf{Y}_{A_{/A\cap\bar{B}}} &= \mathbf{Y}_{A_{/\bar{B}}}\mathbf{U}_A \\
\mathbf{Y}_{C_{/A\cap\bar{B}}} &= \mathbf{Y}_{C_{/\bar{B}}}\mathbf{U}_A
\end{aligned}$$

Toutes sont centrées pour leurs pondérations respectives ($\{\varepsilon_k\}$, $\{\lambda_i\}$, et $\{w_{ij}\}$). La matrice $\mathbf{X}_{/A\cap\bar{B}}$ contient les coordonnées des catégories. Des coordonnées moyennes des modalités du facteur A sont données par $\mathbf{Y}_{A_{/A\cap\bar{B}}}$. La matrice $\mathbf{Y}_{C_{/A\cap\bar{B}}}$ fournit les coordonnées des modalités du facteur A pour chaque modalité du facteur B.

Prenons les données de Blondel *et al.* (1984). Ils considèrent quatre successions dont trois sont localisées sous un bioclimat méditerranéen : une en Californie, une au Chili, et une en Provence. La quatrième succession se situe en Bourgogne sous un climat tempéré. La séparation en quatre successions constitue un premier facteur. Pour chaque succession, quatre stades d'ouverture de la végétation sont choisis. Ces quatre stades constituent le deuxième facteur. Les quatre habitats des successions sont choisis de sorte qu'ils se ressemblent le plus possible entre successions. "succession" et "stade" sont donc deux facteurs croisés. Dans chaque habitat de chaque succession, les espèces d'oiseaux sont recensées. "succession", "stade" et "espèce" sont trois modes d'un cube de données dont les entrées sont les présences/absences des espèces dans chaque stade de chaque succession. A ce cube de données sont rajoutées les distances taxonomiques entre espèces.

La question posée par Blondel *et al.* était : existe-t-il une convergence morphologique des communautés d'oiseaux nichant sous climat méditerranéen ? Ils proposent les critères suivants pour définir si oui ou non il existe une convergence à l'échelle des communautés.

Si il existe une convergence, les similitudes morphologiques entre les espèces des communautés méditerranéennes devraient être plus grandes que celles entre espèces méditerranéennes et espèces de Bourgogne. Blondel *et al.* illustrent cette hypothèse à travers la figure 34a. S'il n'existe pas de convergence, alors soit les quatre régions sont très similaires (Fig. 34b) soit elles sont très différentes et chacune unique (Fig. 34c).

Or les espèces de Bourgogne sont, d'un point de vue phylogénétique, plus proches de celles de Provence que de celles du Chili ou de Californie. Blondel *et al.* proposent alors une petite modification des modèles précédents. Si la convergence n'est pas importante, et si les similarités taxonomiques sont plus grandes que les similarités morphologiques, alors les similitudes entre successions auront l'allure de celles de la figure 34d correspondant à un modèle "phylogénétique". S'il existe effectivement une convergence notable, alors on devrait se rapprocher du profil de la figure 34e.

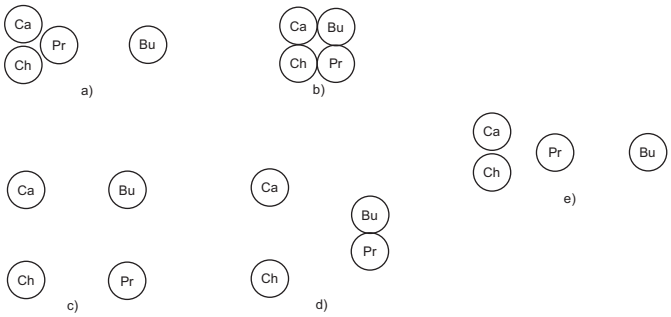


FIG. *34 – Cinq modèles possibles pour expliquer les degrés de similitude entre successions (d'après Blondel* et al. *(1984)) : (a) convergence bioclimatique ; (b,c) absence de convergence ; (d) modèle "phylogénétique" ; (e) modèle de convergence bioclimatique ajusté pour tenir compte des liens phylogénétiques entre les espèces des quatre successions.*

Nous n'allons pas nous intéresser ici aux données morphologiques, mais simplement à la description de la structure taxonomique des espèces avec deux facteurs : stades de végétation et successions. Provence et Bourgogne ont des espèces en commun. Si les deux successions américaines ne partagent aucune espèce avec les deux successions européennes, elles partagent avec chacune de ces successions des genres, des familles et des ordres. La DPCoA croisée peut nous aider à décrire plus précisément les relations taxonomiques entre successions et entre stades de végétations.

Deux DPCoA peuvent être réalisées. La première a pour but de comparer la répartition des espèces, genres, familles et ordres le long de chaque succession. La deuxième, à l'inverse, a pour objet de regarder si la typologie des différences entre successions, en terme de composition taxonomique d'oiseaux, dépend du stade de végétation considéré. Le résultat de cette deuxième analyse pourra être comparé au modèle "phylogénétique" (Fig. 34d).

Les résultats sont que les espèces, genres, familles, ordres s'organisent le long du gradient commun de végétation (Fig. 35). Plus de 80%, en moyenne, des différences entre communautés d'oiseaux intra-succession se reflètent sur le premier axe de l'analyse de l'effet des stades de végétation dans chaque succession (effet "stade sachant succession"). Autour de ce gradient commun, on repère néanmoins des différences entre successions. Nous retrouvons le résultat de la figure 26 page 135 : au Chili, les milieux ouverts 1 et 2 s'opposent aux milieux fermés 3 et 4, alors qu'en Californie, l'habitat 1 se détache des autres. Ces disparités peuvent être dues en partie à des difficultés à déterminer des stades de végétations très similaires dans des régions différentes. Dans la succession de Bourgogne et dans celle de Provence, les compositions avifaunistiques sont beaucoup plus semblables le long des successions qu'en Californie et au Chili. Dans ce jeu de données, autour d'un gradient commun, les différences taxonomiques inter-stades intra-succession diffèrent donc selon la succession considérée.

Par contre les profils de différence entre successions sont très similaires d'un stade de végétation à l'autre (Fig. 36, 37). La typologie des successions obtenue est cohérente avec celle proposée par Blondel *et al.* (1984) (Fig. 34d). En moyenne dans les stades de végétation, 61% des différences taxonomiques entre successions sont dues à des disparités entre continents, 21%

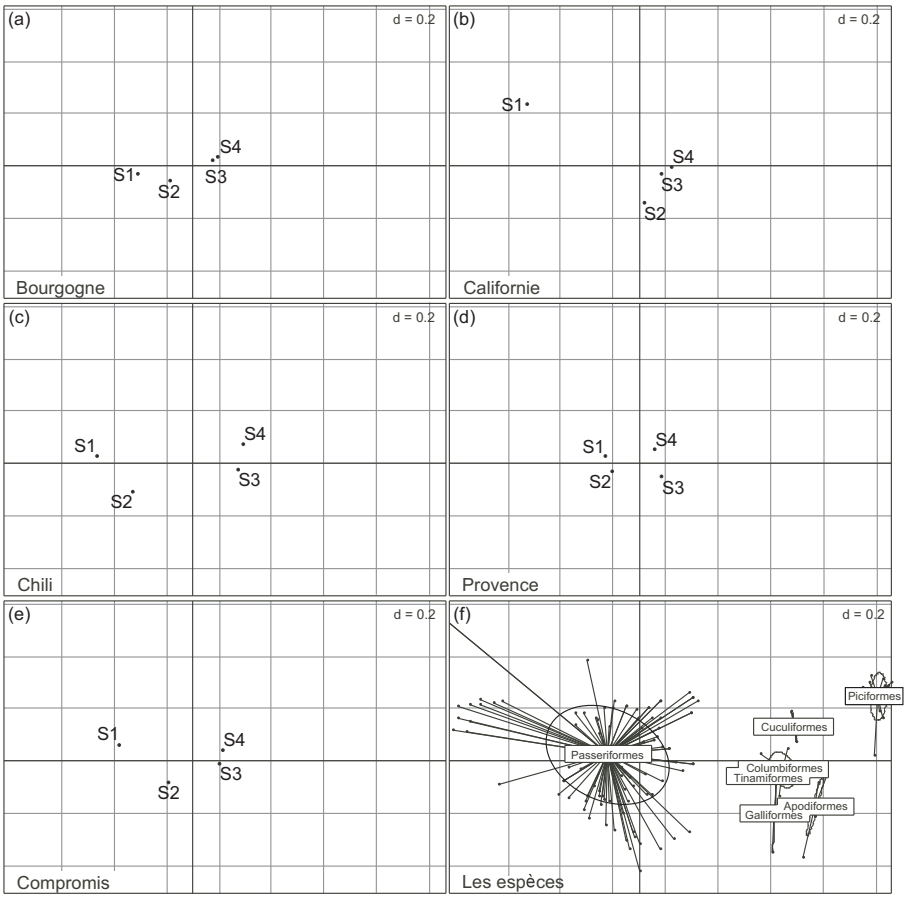


FIG. *35 – Carte factorielle F1 × F2 de l'analyse "stade sachant succession" : (a) typologie des stades de végétation en Bourgogne, (b) typologie des stades de végétation en Californie, (c) typologie des stades de végétation au Chili, (d) typologie des stades de végétation en Provence, (e) typologie moyenne des stades de végétation selon le compromis, et (f) typologie des espèces regroupées par ordre. Ces deux premiers axes principaux représentent 84 et 10% respectivement de la variation entre les points des stades de végétation dans le compromis. La valeur 'd' donne la taille du côté d'un carré dans la grille. Cette figure a été réalisée à partir des fonctions 'DPCoA', développée personnellement, 's.label' et 's.distri' du package ade4 (Chessel* et al. *2004) de R (Ihaka et Gentleman 1996).*

à des différences entre les deux successions d'Amérique et 18% à des disparités entre les deux successions d'Europe.

La DPCoA croisée permet de séparer les effets de deux facteurs dont le mélange dans une analyse globale pourrait conduire à des cartes factorielles difficilement interprétables. L'ordination globale prend en effet en compte en même temps les effets du facteur A, ceux du facteurs B et aussi ceux de l'interaction entre ces deux facteurs. Il est alors impossible de dissocier, sur les cartes factorielles, la part des différents effets. Nous montrons dans l'article Pavoine *et al.* (en révision, cf. annexe 6) que l'interprétation directe des cartes a de fortes chances d'engendrer des erreurs.

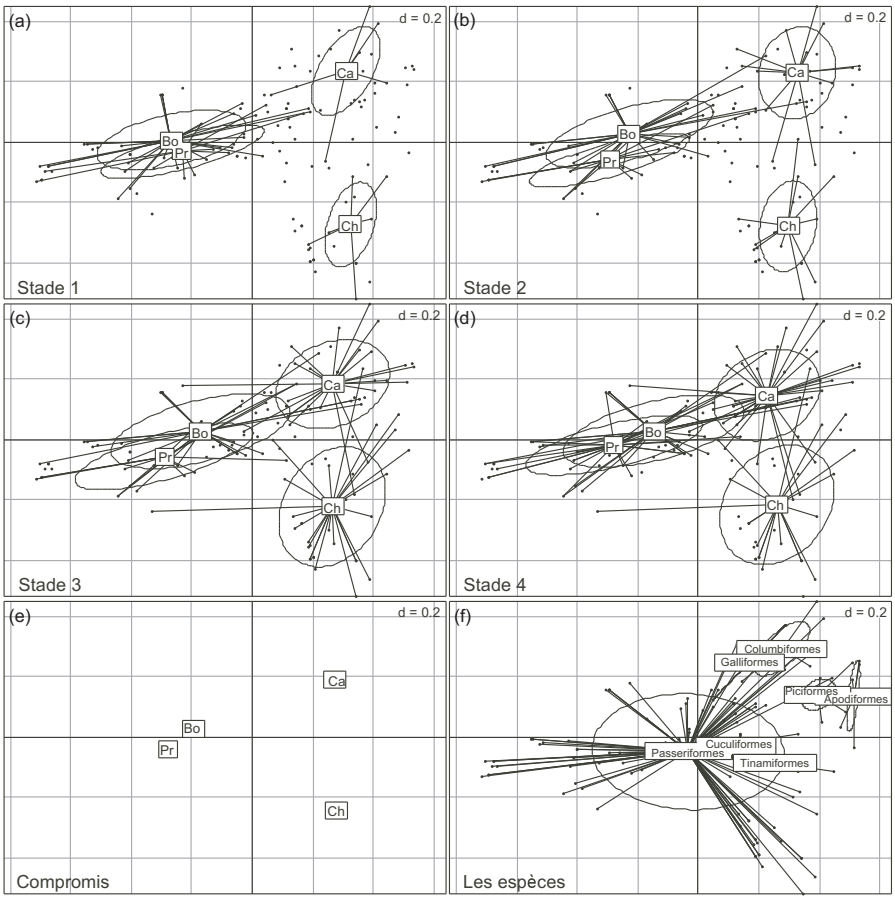


FIG. *36 – Carte factorielle F1 × F2 de l'analyse "succession sachant stade" : (a) à (d) typologie des successions selon les stades 1 à 4 de végétation, (e) typologie moyenne des successions selon le compromis, et (f) typologie des espèces regroupées par ordre. Des ellipses d'inertie des points espèces autour des points successions sont représentées. Ces deux premiers axes principaux représentent 61 et 21% respectivement de la variation entre les points successions dans le compromis. La valeur 'd' donne la taille du côté d'un carré dans la grille. Les notations sont : "Bo" Bourgogne, "Ca" Californie, "Ch" Chili, "Pr" Provence. Cette figure a été réalisée à partir des fonctions 'DPCoA', développée personnellement, 's.label' et 's.distri' du package ade4 (Chessel* et al. *2004) de R (Ihaka et Gentleman 1996).*

4.4.3 DPCoA multiple

L'idée de faire une DPCoA multiple est venue d'un problème actuelle en génétique qui est d'évaluer la cohérence de marqueurs.

Il est impossible, avec les moyens actuels, de connaître tous les génomes d'un grand nombre d'individus pour plusieurs populations. Pour comparer les compositions génétiques de plusieurs populations, les analyses génétiques se focalisent alors sur plusieurs parties du génome appelées marqueurs ou loci. Chaque individu possède une composition nucléotidique particulière ou allèle pour chaque marqueur. Tous les allèles observés sont des catégories qui peuvent servir à mesurer la diversité nucléotidique d'une ou plusieurs populations et aussi la diversité nucléotidique entre populations. Des relations entre ces allèles peuvent être connues et représentées

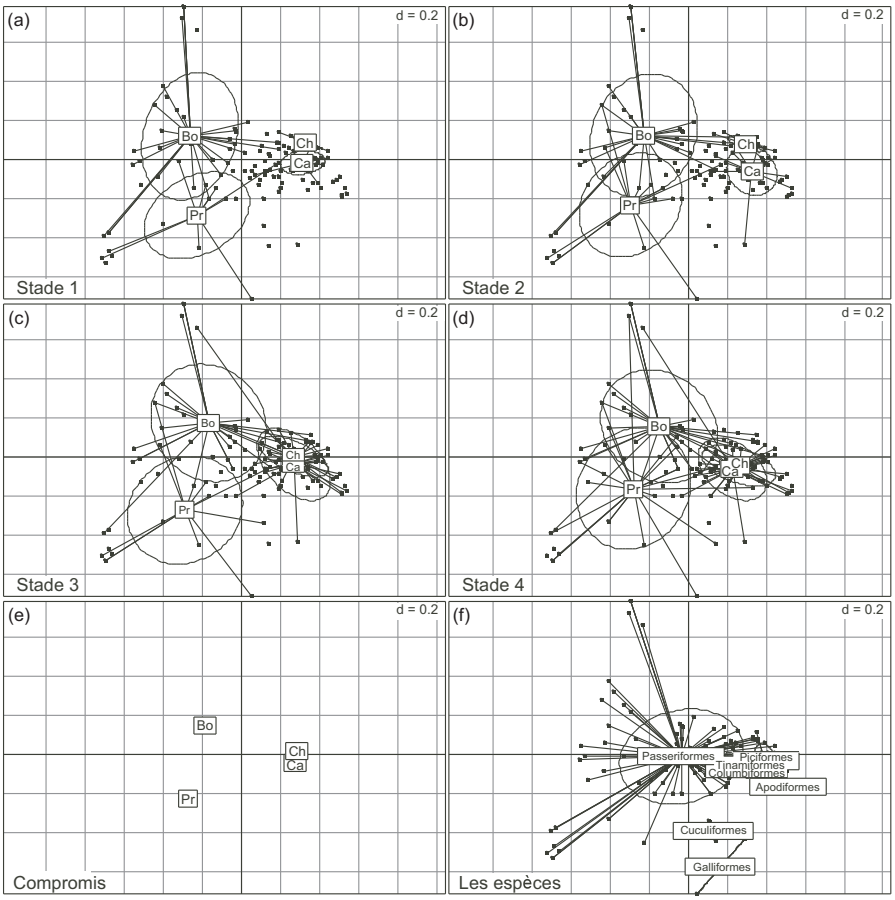


FIG. *37 – Carte factorielle F1 × F3 de l'analyse "succession sachant stade" : (a) à (d) typologie des successions selon les stades 1 à 4 de végétation, (e) typologie moyenne des successions selon le compromis, et (f) typologie des espèces regroupées par ordre. Des ellipses d'inertie des points espèces autour des points successions sont représentées. Ces premier et troisième axes principaux représentent 61 et 18% respectivement de la variation entre les points successions dans le compromis. La valeur 'd' donne la taille du côté d'un carré dans la grille. Les notations sont : "Bo" Bourgogne, "Ca" Californie, "Ch" Chili, "Pr" Provence. Cette figure a été réalisée à partir des fonctions 'DPCoA', développée personnellement, 's.label' et 's.distri' du package ade4 (Chessel* et al. *2004) de R (Ihaka et Gentleman 1996).*

sous forme d'arbres.

Nous avons alors G matrices de fréquences absolues, de type allèles × populations et de dimensions (ρ_g, r), où $g = 1, ..., G$, correspondant aux G marqueurs $\{\mathbf{F}_1, ..., \mathbf{F}_g, ..., \mathbf{F}_G\}$. Chacune de ces matrices est associée à une matrice de dissimilarités $\mathbf{D}^{\text{all}}$ entre allèles $\{\mathbf{D}_1^{\text{all}}, ..., \mathbf{D}_g^{\text{all}}, ..., \mathbf{D}_G^{\text{all}}\}$. Les populations sont les mêmes pour tous les marqueurs mais les allèles sont différents (chaque groupe d'allèles correspond à un marqueur parmi G). La question posée par l'étude de la cohérence des marqueurs est la suivante : la typologie des populations (représentant la diversité inter-populations) est-elle la même selon les différents marqueurs ?

L'analyse de la diversité inter-populations pour un marqueur donné correspond à une DP-

CoA :

- $-\mathbf{Q}_g\mathbf{D}_g^{\text{all}}\mathbf{Q}_g^t = \mathbf{X}_g\mathbf{X}_g^t$,
- $\mathbf{Y}_g = B_r^{-1}\mathbf{F}_g^t\mathbf{X}_g$,
- Décomposition en valeurs singulières de $\left(\mathbf{Y}_g, \mathbf{I}_{\rho_g}, \mathbf{B}_r\right)$.

Considérer l'ensemble des marqueurs fait donc apparaître G triplets

$$(\left(\mathbf{Y}_1, \mathbf{I}_{\rho_1}, \mathbf{B}_r\right), ..., \left(\mathbf{Y}_g, \mathbf{I}_{\rho_g}, \mathbf{B}_r\right), ..., \left(\mathbf{Y}_G, \mathbf{I}_{\rho_G}, \mathbf{B}_r\right)).$$

Des analyses séparées de ces triplets peuvent être exécutées. Pour évaluer l'influence de chaque marqueur sur la typologie des populations, il nous faut trouver un espace dans lequel toutes les analyses séparées peuvent être mises en relation. Dans la suite de cette partie, il va être montré que l'analyse de co-inertie multiple (Chessel et Hanafi 1996) est une solution possible.

L'analyse de co-inertie appliquée aux triplets

$$(\left(\mathbf{Y}_1, \mathbf{I}_{\rho_1}, \mathbf{B}_r\right), ..., \left(\mathbf{Y}_g, \mathbf{I}_{\rho_g}, \mathbf{B}_r\right), ..., \left(\mathbf{Y}_G, \mathbf{I}_{\rho_G}, \mathbf{B}_r\right))$$

fournit, dans une première étape, G vecteurs $\mathbf{u}_g^{[1]}$ normés dans chaque espace $\mathbb{R}^{\rho_g}$, et une variable synthétique $\mathbf{v}^{[1]}$ D_r-normée dans $\mathbb{R}^r$ qui maximisent la quantité

$$\sum_{g=1}^{G} \pi_g \langle \mathbf{Y}_g\mathbf{u}_g | \mathbf{v} \rangle_{\mathbf{B}_r}^2.$$

La valeur π_g est un poids attribué au triplet $\left(\mathbf{Y}_g, \mathbf{I}_{\rho_g}, \mathbf{B}_r\right)$ afin d'homogénéiser les influences des triplets dans l'analyse multiple. On pourra prendre par exemple $\pi_g = 1/\lambda_{g_1}$, où λ_{g_1} est la première valeur propre de l'analyse du triplet individuel $\left(\mathbf{Y}_g, \mathbf{I}_{\rho_g}, \mathbf{B}_r\right)$. Le vecteur $\mathbf{Y}_g\mathbf{u}_g$ contient un système de coordonnées des populations selon le marqueur g. L'opération fournit donc deux vecteurs $\mathbf{u}_g^{[1]}$ et $\mathbf{v}^{[1]}$ maximisant la somme pondérée des carrés des covariances entre un système de coordonnées ($\mathbf{Y}_g\mathbf{u}_g$) représentant le marqueur G et un système de coordonnées ($\mathbf{v}$) synthétisant l'ensemble des marqueurs. $\mathbf{v}^{[1]}$ est la première composante $\mathbf{B}_r$-normée de la PCA correspondant au triplet $\left(\tilde{\mathbf{Y}} = \left[\sqrt{\pi_1}\mathbf{Y}_1 \mid ... \mid \sqrt{\pi_g}\mathbf{Y}_g \mid ... \mid \sqrt{\pi_G}\mathbf{Y}_G\right], \mathbf{I}_{\rho_\bullet}, \mathbf{B}_r\right)$, où $\rho_\bullet = \sum_{g=1}^{G}\rho_g$ est le nombre total d'allèles, tous marqueurs confondus. Et

$$\mathbf{u}_g^{[1]} = \frac{\mathbf{Y}_g^t\mathbf{B}_r\mathbf{v}^{[1]}}{||\mathbf{Y}_g^t\mathbf{B}_r\mathbf{v}^{[1]}||}.$$

Les axes suivants sont déterminés de sorte que

$$\langle \mathbf{v}^{[k]} | \mathbf{v}^{[l]} \rangle_{\mathbf{B}_r} = 0 \text{ et } \langle \mathbf{u}_g^{[k]} | \mathbf{u}_g^{[l]} \rangle = 0 \text{ pour tout } k, l\ (1 \leq k < l) \text{ et pour tout } g\ (1 \leq g \leq G).$$

Pour obtenir les $k^{\text{ièmes}}$ axes, le raisonnement précédent est effectué en remplaçant chaque matrice $\mathbf{Y}_g$ par $\mathbf{Y}_g\left(\mathbf{I}_{\rho_g} - \mathbf{P}_g^{k-1^t}\right)$, où $\mathbf{P}_g^{k-1}$ est un projecteur $\mathbf{I}_{\rho_g}$-orthogonal sur le sous-espace vectoriel de $\mathbb{R}^{\rho_g}$ engendré par les systèmes orthonormés $\{\mathbf{u}_g^{[1]}, ..., \mathbf{u}_g^{[k-1]}\}$

$$\mathbf{P}_g^{k-1} = \mathbf{U}_g^{k-1}\left(\mathbf{U}_g^{k-1^t}\mathbf{I}_{\rho_g}\mathbf{U}_g^{k-1}\right)^{-1}\mathbf{U}_g^{k-1^t} = \mathbf{U}_g^{k-1}\mathbf{U}_g^{k-1^t},$$

où $\mathbf{U}_g^{k-1} = \left[\mathbf{u}_g^{[1]} \mid ... \mid \mathbf{u}_g^{[k-1]}\right]$.

Soient $\mathbf{U}_g$ la matrice $\left[\mathbf{u}_g^{[1]} \mid ... \mid \mathbf{u}_g^{[k]} \mid ... \mid \mathbf{u}_g^{[K]}\right]$ et $\mathbf{V}$ la matrice $\left[\mathbf{v}^{[1]} \mid ... \mid \mathbf{v}^{[k]} \mid ... \mid \mathbf{v}^{[K]}\right]$. Les coordonnées des populations selon le marqueur g sont dans

$$\mathbf{L}_{Y_g} = \sqrt{\pi_g}\mathbf{Y}_g\mathbf{U}_g.$$

Comme $\mathbf{Y}_g = \mathbf{B}_r^{-1}\mathbf{F}_g\mathbf{X}_g$, les coordonnées suivantes peuvent être attribuées aux allèles pour que chaque population soit placée au barycentre de sa composition allélique

$$\mathbf{L}_{X_g} = \sqrt{\pi_g}\mathbf{X}_g\mathbf{U}_g.$$

Les coordonnées synthétiques des populations sont trouvées directement dans $\mathbf{V}$. Ce sont les coordonnées "moyennes" sur l'ensemble des marqueurs. Les allèles étant différents d'un marqueur à l'autre, la notion de coordonnées synthétique d'allèles n'a aucun sens.

Dans le cas de données manquantes, c'est à dire si il manque la composition allélique de certains individus pour certains marqueurs, l'analyse de co-inertie ne peut pas prendre en compte des matrices de poids des populations différentes pour chaque marqueur. Il faudra dans ce cas définir un poids commun à tous les triplets.

Cette analyse a été développée pour analyser les données de Xavier Bailly (doctorant, Laboratoire des symbioses tropicales et méditerranéennes, UMR 1063, Montpellier) en fin d'échantillonnage. Ces données permettront l'étude de populations bactériennes associées à deux espèces de plantes et se développant dans deux types de sols.

4.5 P

Nous constatons donc que de nombreux indices ont été développés pour mesurer la dissimilarité entre deux collections soit à travers deux vecteurs de présences/absences donnant la liste des catégories contenues dans chaque collection, soit à travers deux vecteurs de fréquences décrivant l'abondance relative des catégories.

Parmi ces derniers, la différence de Jensen appliquée à l'entropie quadratique a l'avantage de mesurer la dissimilarité entre deux collections à partir de la liste des catégories dans chaque collection, des fréquences de ces catégories et aussi d'une matrice de dissimilarités résumant numériquement, selon un point de vue choisi, les différences qui peuvent exister entre les catégories. Grâce à l'application de la différence de Jensen à l'entropie quadratique, constituant ainsi une fonction de dissimilarité, nous avons montré qu'il est possible de comparer efficacement des collections qui ne partagent aucune catégorie. Cet avantage devient très intéressant lorsqu'au cours d'une étude, on souhaite comparer deux régions dont les cortèges faunistiques ou floristiques peuvent être considérés comme parfaitement disjoints si on regarde simplement les noms des espèces, mais néanmoins comparables si on s'intéresse à d'autres critères tels que la taxonomie ou la morphométrie.

Nous avons montré que l'entropie quadratique est une mesure d'inertie et que les décompositions additives hiérarchiques et croisées (APQE et ANOQE, cf. partie 3.3) sont des décompositions d'inertie dans des espaces euclidiens que nous avons déterminés. Nous appuyant sur

la proposition de Rao d'une approche unifiée des concepts de diversité et dissimilarité, nous sommes maintenant allés un pas plus loin en rassemblant dans un schéma statistique commun, grâce à l'axiomatisation de Rao et au schéma de dualité, les concepts de diversité, inertie, dissimilarité, ordination et typologie.

Chapitre 5

Quand l'entropie quadratique est une bonne mesure de biodiversité

Sommaire

Résumé

C'est à la suite de cette remarque de Shimatani, Izsak et Szeidl, que j'ai entrepris les recherches décrites dans ce chapitre : l'entropie quadratique peut être maximisée en réduisant la richesse. Quelques exemples permettent d'illustrer les différents comportements de l'entropie quadratique à son maximum. Nous mettons alors en évidence deux comportements extrêmes :

- dans le cas particulier où l'entropie quadratique correspond à l'indice de Gini-Simpson, la valeur maximale est atteinte pour une répartition égale des effectifs entre l'ensemble des catégories possibles ;
- dans le cas particulier où l'entropie quadratique est égale à la variance d'une variable Y, sa valeur maximale est obtenue lorsque les deux catégories correspondant aux valeurs observées extrêmes de Y ont chacune une fréquence de 0.5. Les fréquences de toutes les autres catégories possibles sont dans ce cas nulles.

L'entropie quadratique n'est donc pas toujours maximisée en même temps que la richesse, et nous arrivons à la question suivante : quelle(s) signification(s) biologique(s) pour l'entropie quadratique ?

Ce chapitre démontre que la présence ou non d'une réduction du nombre de catégories pour maximiser l'entropie quadratique dépend de propriétés mathématiques de la matrice de dissimilarités entre catégories qui a été choisie. Une classe de matrice de dissimilarités appelée "SEH-circum-euclidienne" est introduite. Lorsque l'entropie quadratique est appliquée à des matrices ultramétriques, toutes les catégories sont gardées pour atteindre la valeur maximale. J'ai démontré tous les résultats décrits dans l'article de l'annexe 3. Ces démonstrations font appel à l'algèbre linéaire et à la géométrie euclidienne.

Les matrices de distances phylogénétiques sont ultramétriques. Nous montrons que la distribution de fréquences maximisant la valeur de l'entropie quadratique appliquée à des distances phylogénétiques est une mesure de l'originalité relative des espèces les unes par rapport aux autres. Nous discutons de ce résultat dans le débat actuel pour l'attribution d'une valeur à une espèce dans le cadre de la biologie de la conservation. Ce chapitre montre que les indices de diversité attribuant des poids aux espèces, fortement critiqués dans le cadre de la mesure de la biodiversité, sont en fait des indices d'originalité.

Finalement, nous concluons que l'entropie quadratique appliquée à des dissimilarités ultramétriques est bien un indice de diversité possédant trois propriétés qui apparaissent fondamentales pour un tel indice :

- l'unique distribution de fréquences des catégories qui conduit à la valeur maximale de l'indice décrit l'originalité relative des espèces ;
- la valeur maximale de l'indice augmente par l'ajout d'une catégorie dans l'ensemble des possibles ;
- l'indice est complètement concave, il peut être décomposé selon un nombre quelconque de facteurs hiérarchiques et croisés.

5.1 M ' ´, ` ' -

5.1.1 Pourquoi étudier cette maximisation ? Mise en évidence du problème

L'entropie quadratique n'est pas bornée. Ses valeurs numériques dépendent de l'échelle des dissimilarités. Pour normaliser cet indice, une solution est de diviser sa valeur observée dans une collection par sa valeur théorique maximale :

$$H^*_{\mathbf{\Delta}}(\mathbf{p}_{\text{obs}}) = \frac{H_{\mathbf{\Delta}}(\mathbf{p}_{\text{obs}})}{\max\limits_{\mathbf{p}} H_{\mathbf{\Delta}}(\mathbf{p})}.$$

Dans cette formule, $H_{\mathbf{\Delta}}(\mathbf{p}_{\text{obs}})$ est la valeur observée de l'entropie quadratique dans la collection et $H^*_{\mathbf{\Delta}}(\mathbf{p}_{\text{obs}})$ la valeur observée normalisée. La démarche utilisée pour maximiser l'entropie quadratique comprend trois étapes :

1. choisir des catégories dont le nombre est fini ;
2. choisir une matrice de dissimilarités entre ces catégories ;
3. rechercher une distribution de fréquences qui maximise la valeur de l'entropie quadratique pour ces catégories et cette matrice de dissimilarités.

La recherche de cette distribution de fréquences a conduit à un résultat qui est plutôt inattendu pour un indice de diversité : Shimatani (2001) et Izsak et Szeidl (2002) observent que la maximisation de l'entropie quadratique peut réduire considérablement la richesse en espèces ou plus généralement en catégories.

Rao (1986) note que, pour un ensemble de S catégories, toutes les mesures traditionnelles de diversité (les indices de Shannon, d'Havrda et Charvat, de Rényi, l'indice γ-entropie, et l'entropie appariée) ont la propriété d'atteindre leur maximum lorsque la distribution de fréquences est uniforme

$$\mathbf{p}_{\text{uni}} = \left(\frac{1}{S}, \frac{1}{S}, ..., \frac{1}{S}\right),$$

et elles valent 0 si une seule catégorie est représentée, ce qui paraît, au premier abord, être des conditions logiques pour des mesures de diversité. Rao (1986) a recherché l'existence d'autres conditions logiques associées au concept de diversité. Selon lui, les deux conditions minimales pour qu'un indice puisse être qualifié d'"indice de diversité" sont la positivité et la concavité, auxquelles est rajoutée éventuellement la possibilité d'être décomposé selon des facteurs croisés. Les conditions de Rao ne font donc aucune référence à d'éventuelles contraintes sur la valeur maximale d'un indice de diversité. L'entropie quadratique qu'il propose comme "indice parfait de diversité" ne prend pas sa valeur maximale pour une distribution de fréquences uniformes. En écologie, Ricotta (2002) le qualifie alors d'"indice de diversité faible". Quelle est cette valeur maximale de l'entropie quadratique et quelle distribution de fréquences le conduit à cette valeur ?

Shimatani (2001) puis Izsak et Szeidl (2002) observent qu'avec certaines matrices de dissimilarités entre catégories, une distribution de fréquences qui maximise l'entropie quadratique peut contenir un ou plusieurs zéros. Reformulées ces observations prouvent que parfois plusieurs catégories doivent être éliminées pour maximiser l'entropie quadratique. Shimatani, comme Izsák et Sveidl ont bien sûr trouvé que ce résultat est plutôt inattendu pour un indice de diversité.

Champely et Chessel (2002) proposent une méthode itérative pour obtenir une approximation d'un vecteur de fréquences qui maximise l'entropie quadratique. Nous allons utiliser cette procédure sur plusieurs exemples pour observer concrètement le comportement de l'entropie quadratique à son maximum en fonction du choix des dissimilarités.

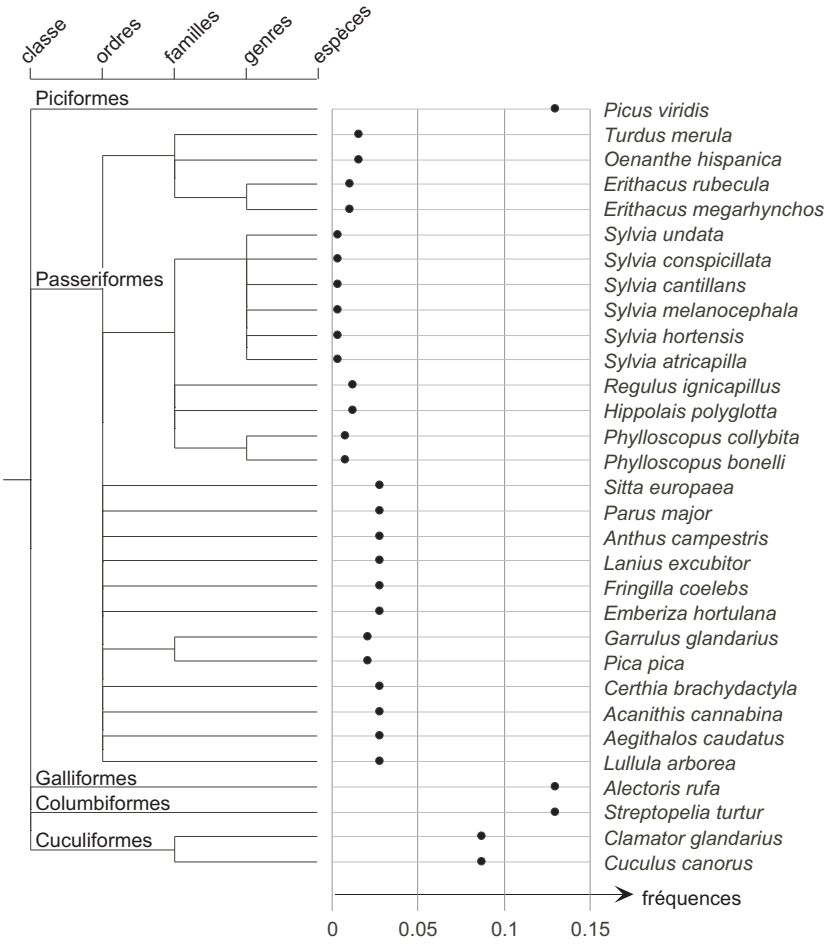


FIG. *38 – Distribution de fréquences maximisant la diversité taxonomique entre 31 espèces d'oiseaux observées en Provence (données de Blondel* et al. *(1984)).*

Commençons par considérer que les catégories sont des espèces. A partir d'un exemple, Shimatani (2001) a remarqué que, si des dissimilarités taxonomiques sont utilisées pour carac-

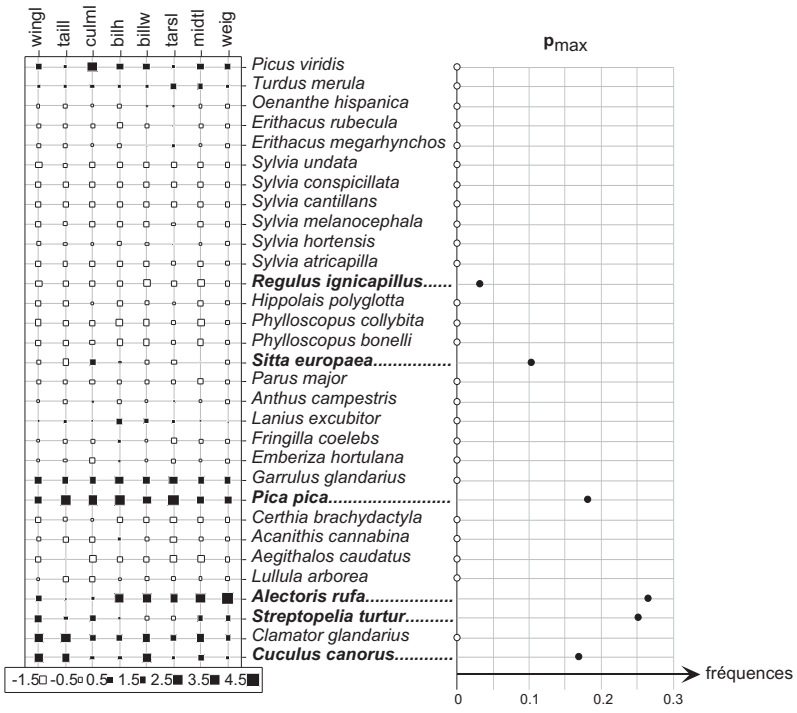


FIG. *39 – Distribution de fréquences maximisant la diversité morphométrique de 31 espèces d'oiseaux observées en Provence (données de Blondel* et al. *(1984)). Huit variables morphométriques sont considérées : "wingl" (wing length, longueur de l'aile (mm)), "taill" (tail length, longueur de la queue (mm)), "culml" (culmen length, longueur du culmen (mm)), "billh" (bill height, hauteur du bec (mm)), "billw" (bill width, largeur du bec (mm)), "tarsl" (tarsus length, longueur des tarses (mm)), "midtl" (middle-toe length, longueur de l'orteil médian (mm)), "weig" (weight, poids (g)). A gauche, ces variables normées sont représentées dans un tableau. L'échelle est donnée par la taille des carrés : plus un carré est grand plus la valeur absolue est grande, et un carré blanc indique une valeur négative alors qu'un carré noir indique une valeur positive. Pour l'analyse de la diversité, les sept premières variables sont divisées par la racine cubique du poids pour éliminer l'effet taille : un oiseau au poids élevé a plus de chances d'avoir de grandes ailes et une grande queue qu'un oiseau de poids faible. Des dissimilarités morphologiques δ sont calculées avec la distance de Mahalanobis appliquée aux logarithmes de ces sept variables. A droite, un diagramme de Cleveland fournit le vecteur maximisant l'entropie quadratique appliquée aux dissimilarités morphométriques entre ces 31 espèces d'oiseaux. Les ronds pleins représentent des valeurs de fréquences strictement positives et les ronds vides des valeurs nulles. Six espèces ont une fréquence strictement positive. Cette figure a été réalisée avec les fonctions 'table.value' et 'dotchart' d'ade4 et de la base de R (Ihaka et Gentleman 1996, Chessel* et al. *2004).*

tériser les différences entre des espèces, toutes ces espèces sont retenues avec des fréquences inégales pour maximiser l'entropie quadratique. En appliquant ce raisonnement à 31 espèces d'oiseaux de Provence, nous retrouvons le même résultat (Fig. 38). Pour compléter la remarque de Shimatani, il peut être rajouté que, pour maximiser la diversité taxonomique mesurée par l'entropie quadratique, les espèces appartenant aux ordres les moins représentés ont une fréquence élevée ; et inversement que, pour maximiser la diversité taxonomique mesurée par l'en-

tropie quadratique, les six espèces du genre *Sylvia* ayant beaucoup d'espèces-sœurs ont les plus faibles fréquences.

Si ces mêmes espèces sont caractérisées par des variables morphologiques, et que la distance de Mahalanobis est utilisée pour calculer des dissimilarités δ entre espèces, alors la distribution de fréquences au maximum est très différente. Seules quelques espèces sont retenues (Fig. 39).

Les comportements de l'entropie quadratique au maximum dépendent donc bien de la matrice de dissimilarités.

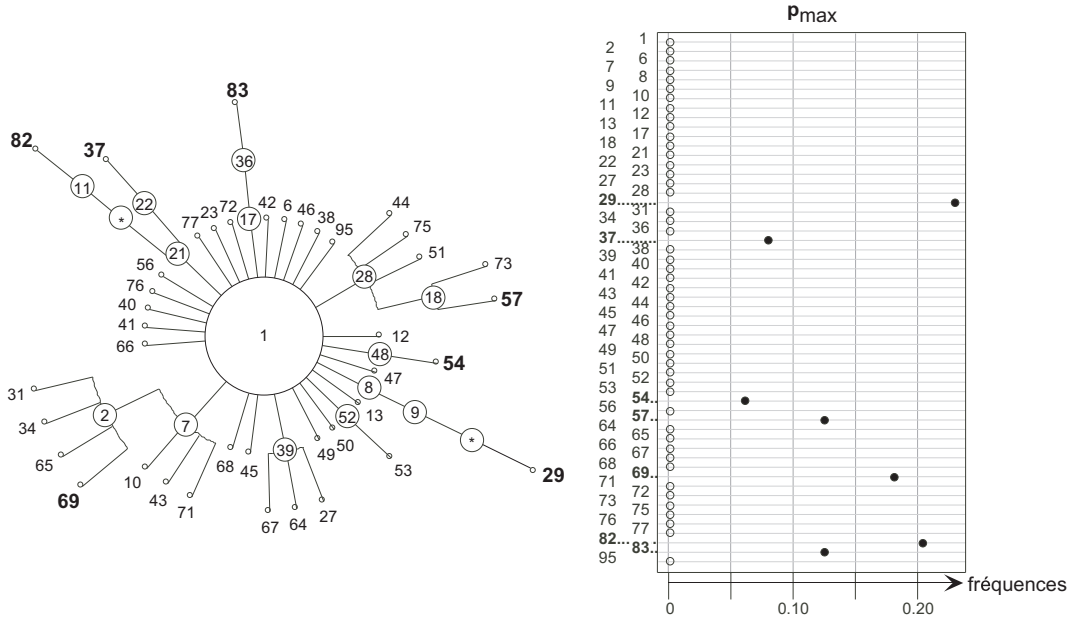


FIG. *40 – Distribution de fréquences maximisant la diversité nucléotidique d'ADN mitochondrial humain selon le jeu de données d'Excoffier* et al. *(1992) comportant 56 haplotypes issus de 10 populations du monde (cf. partie 4.3.3). Des dissimilarités nucléotidiques entre haplotypes sont calculées en termes de nombre de sites de restrictions différents. La figure de gauche est le réseau de longueur minimale reliant les 56 haplotypes. Le segment entre deux nœuds représente une mutation. A droite, un diagramme de Cleveland fournit le vecteur maximisant l'entropie quadratique appliquée aux dissimilarités nucléotidiques entre les haplotypes. Les fréquences nulles sont représentées par des cercles vides et les fréquences strictement positives par des cercles pleins. Sept haplotypes ont une fréquence non-nulle. Ces sept haplotypes sont indiqués en gras dans le réseau.*

Prenons maintenant des données génétiques, celles d'Excoffier *et al.* (1992) (cf. partie 4.3.3). Les catégories ne sont plus des espèces mais des haplotypes. Ces données contiennent 56 haplotypes d'ADN mitochondrial humain observés dans 10 populations du monde. Des dissimilarités nucléotidiques ont été calculées entre les haplotypes et représentées graphiquement par un réseau de longueur minimale (Fig. 40). Les sept haplotypes retenus pour maximiser l'entropie quadratique portent les numéros 29, 37, 54, 57, 69, 82 et 83. Ils se positionnent tous aux feuilles du réseau. Ces haplotypes sont donc particulièrement distincts. Parmi eux, ceux portant les nu-

méros 29, 37, 82 et 83 présentent une position unique dans le réseau. En revanche, les positions des haplotypes 54, 57 et 69 sont comparables à celles d'haplotypes éliminés. Mais un critère les distingue de leurs homologues : la distance moyenne entre chacun d'eux et tous les autres. L'haplotype 69 a une position dans le réseau similaire à celle des haplotypes 31, 34 et 65. Ces quatre haplotypes diffèrent de l'haplotype 2 par une mutation ; mais cette mutation est différente d'un haplotype à l'autre. La mutation de l'haplotype 69 fait que celui-ci possède en moyenne une plus grande dissimilarité avec les autres haplotypes du réseau (4.05 mutations en moyenne contre 3.84 pour les haplotypes 31 et 37 et 3.95 pour l'haplotype 65). De la même façon, l'haplotype 57 est en moyenne plus distant des autres haplotypes que son homologue (4.15 mutations contre 4.05 (haplotype 73)), et l'haplotype 54 est plus distinct en moyenne que son homologue (3.65 mutations contre 3.36 (haplotype 53)). Les haplotypes retenus sont donc les plus extrêmes.

Ces trois exemples sont trois comportements intermédiaires entre deux comportements extrêmes de l'entropie quadratique.

5.1.2 Deux comportements extrêmes

Nous avons vu que l'entropie quadratique possède deux cas particuliers très utilisés en statistique : l'indice de Gini-Simpson, et la variance d'une variable quantitative. Ces deux cas particuliers représentent les deux comportements extrêmes de l'entropie quadratique à son maximum.

Nous avons vu que l'indice de Gini-Simpson est maximum lorsque toutes les catégories possibles sont présentes avec des fréquences égales. C'est la propriété commune aux indices traditionnels de diversité.

Pour la variance, cette distribution est très différente. La variance est au cœur de l'analyse de variance ou ANOVA. Dans la figure 10, partie 3.3.3, page 96, l'ANOVA est vue comme une méthode descriptive de décomposition de la variance. Habituellement, elle est plutôt utilisée pour tester l'existence de différences moyennes entre collections. Une des hypothèses restreignant son utilisation, dans le cadre des tests, est la normalité de la variable étudiée. Si la variable est effectivement distribuée selon une loi normale propre à chaque collection, la variance de cette variable est bien une indication de la diversité de chaque collection, puisque d'une part plus la variance est grande plus la palette des valeurs représentées dans cette population est grande, et d'autre part plus la variance augmente moins certaines valeurs (proches de la moyenne) dominent dans la population et donc plus ces valeurs sont réparties de façon homogène (Fig. 41). Qu'en est-il dans un cadre plus général où toute distribution est possible ?

Prenons tout d'abord uniquement 3 catégories. Attribuons à chacune une valeur tirée au hasard à partir d'une loi normale centrée réduite (Fig. 42a). Les dissimilarités δ entre catégories sont simplement les distances euclidiennes entre les valeurs qui leur ont été attribuées (Fig. 42b). Une représentation triangulaire permet d'observer les différentes valeurs prises par la variance selon les fréquences relatives des trois catégories (Fig. 42c, d). Sur le bord du triangle, lorsque seulement 2 catégories sont présentes, la valeur maximale est obtenue pour une distribution uniforme (0.5 pour chaque catégorie). Lorsqu'une seule catégorie est présente la variance est nulle. La valeur maximale est égale à 1.58 et elle est obtenue pour 50% de catégorie 1 et

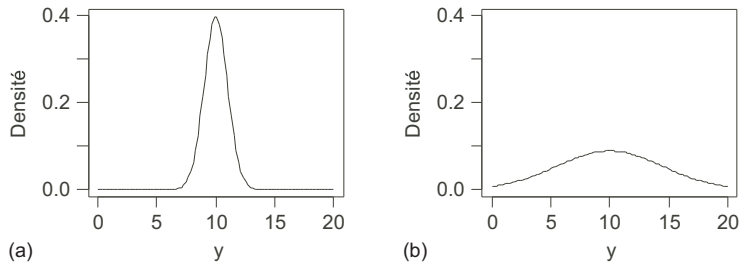


FIG. *41 – Densité d'une loi normale (a)* $\mathcal{N}(10, 1)$*, (b)* $\mathcal{N}(10, 20)$.

50% de catégorie 2. Il s'agit des deux catégories aux valeurs les plus extrêmes. Nous obtenons un résultat bien connu : la variance sur variable quantitative est maximisée si seules les deux catégories les plus extrêmes sont gardées en proportions égales. Par exemple, considérons les espèces d'oiseaux de Provence et la variable morphométrique que nous avions mise de côté précédemment : le poids. L'entropie quadratique, variance du poids des oiseaux de Provence, est maximisée lorsque seules les espèces *Alectoris rufa* (Perdrix rouge, 450 g) et *Regulus ignicapillus* (Roitelet huppé, 5.3 g) sont conservées.

Staudhammer et LeMay (2001) aimeraient mesurer la diversité structurale d'un peuplement végétal. Ils considèrent trois critères : espèce, hauteur et diamètre. Prenons une seule variable : la hauteur des arbres d'une espèce k. Pour mesurer la diversité de ces hauteurs, la première solution proposée est de définir des classes de tailles qui deviendront les catégories, d'évaluer la fréquence de chaque catégorie dans le peuplement et d'utiliser des indices traditionnels tels que celui de Shannon ou celui de Gini-Simpson. Cependant, une partie des informations sur la distribution de tailles est perdue par un regroupement en classes, et la définition des limites des classes est plutôt arbitraire. Staudhammer et LeMay proposent alors de considérer la variance totale de cette hauteur, toutes espèces confondues,

$$S^2 = \frac{\sum_{i=1}^{n} \left[w_i \, (x_i - \bar{x})^2 \right]}{\sum_{i=1}^{n} w_i},$$

où x_i est la hauteur de l'arbre i, $\bar{x}$ la hauteur moyenne, n le nombre total d'arbres et w_i l'aire basale par hectare occupée par l'arbre i. La variance maximale possible est bimodale : soient a et b respectivement la plus petite et la plus grande hauteur d'arbre,

$$S^2_{\max} = \frac{(a-b)^2}{4}.$$

Staudhammer et LeMay affirment que la diversité maximale doit être celle d'une distribution uniforme, comme pour les indices de Gini-Simpson et de Shannon. Cette diversité maximale serait obtenue pour une infinité de valeurs régulièrement distribuées entre a et b :

$$S^2_{\mathrm{U}} = \frac{(a-b)^2}{12}.$$

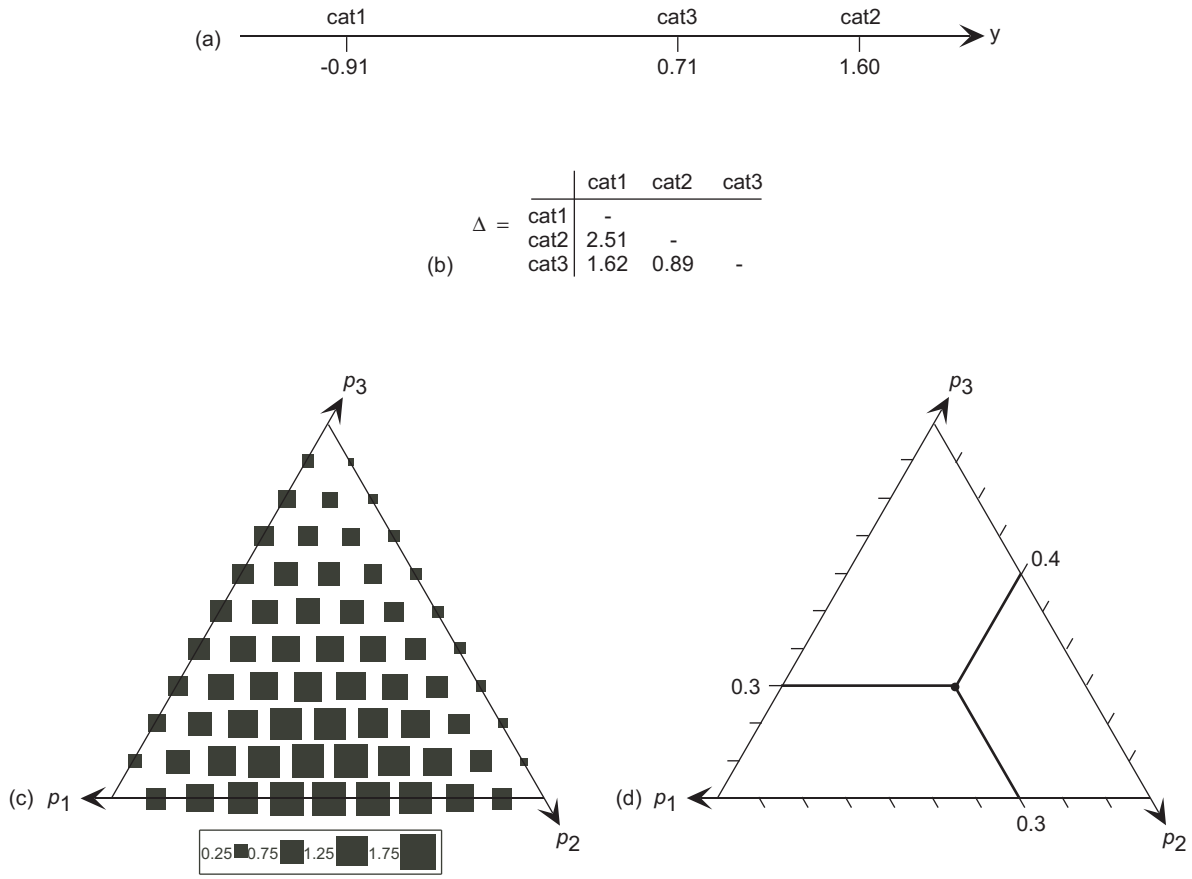


FIG. *42 – Evolution de l'entropie quadratique selon la distribution de fréquences de trois catégories caractérisées par une valeur numérique. La figure (a) donne les valeurs tirées de la loi normale centrée réduite pour les trois catégories (cat1 -0.91, cat2 1.60, cat3 0.71). La métrique euclidienne est utilisée pour définir une matrice* $\mathbf{\Delta}$ *de dissimilarités entre les catégories (b). Appliquée à ces dissimilarités, l'entropie quadratique correspond à la formule de la variance. La figure (c) donne, sous la forme d'une représentation triangulaire, les valeurs prises par l'entropie quadratique pour 66 distributions de fréquences différentes. La taille d'un carré correspond à une valeur prise par l'entropie quadratique. Sa position indique la composition en catégories qui correspond à cette valeur. La figure (d) montre, à partir d'un point particulier, comment obtenir cette composition. Le point représenté correspond à un vecteur de fréquences pour les catégories 1, 2 et 3 respectivement de (0.3, 0.4, 0.3). La valeur de l'entropie quadratique correspondante est de 1.09.*

Staudhammer et LeMay proposent alors un indice de structure basé sur la variance (STVI). Sa valeur pour l'espèce k est

$$\mathrm{STVI}_k = \begin{cases} 1 - \left(\frac{S_\mathrm{U}^2 - S_k^2}{S_\mathrm{U}^2}\right)^{p_1} & \text{si } S_k^2 \leq S_\mathrm{U}^2 \\ 1 - \left(\frac{S_k^2 - S_\mathrm{U}^2}{mS_k^2 - S_\mathrm{U}^2}\right)^{p_2} & \text{si } S_k^2 > S_\mathrm{U}^2 \end{cases},$$

où S_k^2 est la variance observée de la taille des arbres pour l'espèce k, p_1 et p_2 sont des constantes définissant la forme de la courbe qui relie la valeur de l'indice à la variance de l'échantillon, et m

est une constante supérieure ou égale à 1 qui contrôle la valeur de l'indice lorsque la distribution observée est la distribution bimodale maximisant la variance. Si $m = 1$, l'indice s'annule pour cette distribution bimodale.

Un des problèmes que pose leur indice est le choix des constantes. Staudhammer et LeMay proposent de poser trois contraintes sur les valeurs prises par l'indice :

1. l'indice est égal à 0.5 lorsque la variance de l'échantillon est égale à celle d'une distribution uniforme sur la moitié de l'intervalle $[a, b]$;
2. il est aussi égal à 0.5 lorsque la variance de l'échantillon est égale à celle d'une distribution où la moitié des valeurs sont uniformément distribuées au dessous du premier quartile et l'autre moitié au dessus du troisième quartile ;
3. l'indice vaut 0.1 pour la distribution bimodale maximisant la variance.

Avec ces trois contraintes les valeurs suivantes de p_1, p_2 et m sont obtenues :

$$p_1 \simeq 2.4094,\ p_2 \simeq 0.5993 \text{ et } m = 1.1281.$$

Ce qu'il est essentiel de retenir de l'élaboration de cet indice STVI, c'est que Staudhammer et LeMay rejettent fortement l'utilisation directe de la variance pour évaluer la diversité. Il leur a donc fallu rechercher une alternative pour pouvoir malgré tout mesurer efficacement la diversité structurale du peuplement. Reconnaissant que la variance présente l'avantage de prendre en compte une variable quantitative dans la mesure de la variabilité, ils proposent de partir de cet indice pour développer des mesures de diversité. Selon une des étapes de la création de leur indice, la population "idéale", de diversité maximale, serait, du point de vue de la hauteur du feuillage, une population infinie dans laquelle toutes les tailles d'arbres seraient représentées à la même fréquence.

Deux des avantages de l'entropie quadratique sont de pouvoir prendre en compte plusieurs variables dans la caractérisation des catégories ou entités et de pouvoir choisir une autre métrique que la distance euclidienne utilisée par la variance.

La variance étant un cas particulier de l'entropie quadratique, on pourrait envisager d'étendre la démarche de Staudhammer et LeMay à la forme générale de l'entropie quadratique. Mais cette extension peut être complexe. La distribution uniforme, choisie par Staudhammer et LeMay comme la distribution qui doit maximiser leur indice de diversité, constitue un ensemble des tailles d'arbres théoriquement possibles, ensemble arbitrairement placé entre la plus petite et la plus grande hauteur d'arbre observées dans l'échantillon. L'entropie quadratique pouvant être mesurée directement sur des dissimilarités sans passer par des variables, définir dans ce cas un ensemble des possibles pourra être difficile. Et, d'un autre côté, la difficulté reste grande, si des variables quantitatives, binaires et qualitatives sont utilisées ensemble pour calculer les dissimilarités et surtout si ces variables ont des degrés plus ou moins élevés de dépendance. De plus le comportement de l'entropie quadratique n'est pas toujours aussi tranché que celui de la variance puisque, par exemple, la distribution maximale de l'indice de Gini-Simpson, autre cas particulier de l'entropie quadratique, est uniforme.

Tab. 6 – *Compositions des couleurs primaires, des trois couleurs secondaires et des six couleurs intermédiaires. La dernière colonne donne les proportions envisagées d'utilisation de chacune des couleurs.*

	Rouge	Jaune	Bleu	Proportion
Rouge	1	0	0	1/30
Jaune	0	1	0	1/30
Bleu	0	0	1	1/30
Orange	0.5	0.5	0	1/10
Vert	0	0.5	0.5	1/10
Violet	0.5	0	0.5	1/10
Rouge-Orange	0.75	0.25	0	1/10
Jaune-Vert	0	0.75	0.25	1/10
Bleu-Violet	0.25	0	0.75	1/10
Rouge-Violet	0.75	0	0.25	1/10
Jaune-Orange	0.25	0.75	0	1/10
Bleu-vert	0	0.25	0.75	1/10

Il existe donc deux comportements extrêmes de l'entropie quadratique : celui de l'indice de Gini-Simpson et celui de la variance. Entre ces deux comportements extrêmes tous les comportements intermédiaires existent. Le problème est alors dans chaque cas particulier d'être capable d'interpréter les résultats numériques issus de l'application de l'entropie quadratique aux données d'une collection.

5.1.3 Quelle(s) signification(s) biologique(s) pour l'entropie quadratique ?

Shimatani (2001), observant que certaines espèces peuvent être éliminées pour conduire à la valeur maximale de l'entropie quadratique, qualifie ces espèces éliminées de redondantes. Essayons d'expliquer ce que ce terme peut signifier. Lorsque Shimatani l'emploie, il travaille sur des dissimilarités génétiques entre espèces. En général, le terme de redondance est plus fréquemment utilisé à propos des fonctions des espèces dans un écosystème. Pour illustrer ce concept, j'aimerais partir d'une histoire :

Un jeune peintre entreprend de réaliser un tableau représentant un paysage. Il souhaite que ce tableau reflète toutes les couleurs que l'homme perçoit dans son environnement. Pour atteindre son but, il décide de créer une grande palette de couleurs en mélangeant les trois couleurs primaires Rouge, Bleu et Jaune. Il commence par reproduire les couleurs secondaires et intermédiaires et évalue à peu près en quelles proportions il les utilisera dans son tableau (Tab. 6).

Souhaitant savoir avec ces données de base quelle sera la diversité chromatique de son tableau, il utilise l'entropie quadratique. Les catégories sont les couleurs, il a défini une distribution de fréquences de ces couleurs, que nous noterons $\mathbf{q}$, dans une collection : le tableau. Il choisit de calculer la dissimilarité δ entre deux couleurs par la distance euclidienne entre leurs compositions en couleurs primaires. Par exemple, la dissimilarité entre le Rouge et le Rouge-violet est de $\sqrt{(1-0.75)^2+(0-0.25)^2} = 0.25\sqrt{2}$. Il obtient ainsi une matrice $\mathbf{\Delta}^{\text{cou}}$ des

dissimilarités entre couleurs. Avec toutes ces données, il calcule la valeur normée de l'entropie quadratique :

$$H^{*}_{\mathbf{\Delta}^{\text{cou}}}(\mathbf{p}_{\text{obs}}) = \frac{H_{\mathbf{\Delta}^{\text{cou}}}(\mathbf{p}_{\text{obs}})}{\max_{\mathbf{p}} H_{\mathbf{\Delta}^{\text{cou}}}(\mathbf{p})}.$$

Il découvre alors que ses choix ne conduisent qu'à 44% de la diversité maximale possible qu'il pourrait obtenir. Étonné, il recherche en quelles proportions il doit utiliser les couleurs choisies dans son tableau pour maximiser la diversité chromatique. Le résultat est surprenant : les couleurs primaires seules, utilisées en proportions égales, non-mélangées mais simplement juxtaposées, maximiseront la diversité de son tableau. Toute couleur secondaire ou intermédiaire est en effet redondante puisqu'obtenue à partir des seules couleurs primaires. Il n'y avait en fait que trois catégories rouge, bleu et jaune. Déçu mais confiant, le peintre n'utilise que les trois couleurs primaires, sans les mélanger, en les positionnant côte à côte, pour représenter son paysage. Le résultat est une vision très contrastée et évocatrice mais qui ne reflète qu'une partie seulement de la réalité du paysage. Qu'en est-il de cette réalité ?

Revenons à la diversité biologique. Les données définies par ce peintre sont à mettre en relation avec les tableaux dont les lignes sont des espèces et les colonnes des modalités de comportements. Considérons les habitudes alimentaires de macroinvertébrés benthiques (40 espèces de trichoptères et coléoptères) le long de la Loire (Fig. 43) (données de Ivol *et al.* 1997, Pavoine et Dolédec 2005, cf. annexe 2). Sur les 40 espèces, le vecteur maximisant la diversité en retient 18. Toutes les espèces spécialistes sont retenues, dans le sens qu'elles ont des fréquences non nulles dans le vecteur $\mathbf{p}_{\max}$, vecteur optimisant l'entropie quadratique. Treize espèces spécialistes ont été observées : 6 pour la classe des gratteurs (*Limnius perrisi, L. volckmari, L. opacus, Glossosoma conformis, Agapetus delicatulus*), et 7 pour la classe des filtreurs passifs (espèces du genres *Hydropsyche*). Les autres classes sont retenues par 5 non-spécialistes sauf la classe des mangeurs de dépôt qui n'est représentée par aucune des espèces sélectionnées par $\mathbf{p}_{\max}$. Ce vecteur ne maximise ni la richesse en espèce ni la richesse en classes d'habitudes alimentaires. La distribution maximisant l'entropie quadratique conserve donc toutes les espèces spécialistes plus quelques autres représentant des habitudes alimentaires non utilisées par les spécialistes. Avec ce type de données, l'entropie quadratique conserve uniquement les espèces spécialistes s'il en existe au moins une pour chaque modalité (Fig. 44). Et une seule espèce par modalité suffirait. Un tel assemblage ne présenterait effectivement aucune redondance et aucune compétition entre espèces du moins pour le comportement étudié. Cependant chacune de ces espèces aurait un faible pouvoir d'adaptation et à ne conserver qu'elles, nous risquerions de les perdre en cas de changement environnemental. Ce raisonnement est particulier à ce type de données où les espèces sont elles-mêmes des collections d'entités (leurs comportements) regroupées en catégories (types d'habitudes alimentaires, etc.).

Mais qu'en est-il pour les autres cas particuliers de l'entropie quadratique, correspondant à d'autres types de données ? Pour la variance en particulier, existe-t-il des caractères fondamentaux liés au poids, caractères que posséderaient la perdrix rouge et le roitelet huppé et qui nous permettrait d'affirmer que toutes les espèces de poids intermédiaires sont redondantes ?

Deux conclusions peuvent être faites ici. La première est que l'interprétation des valeurs

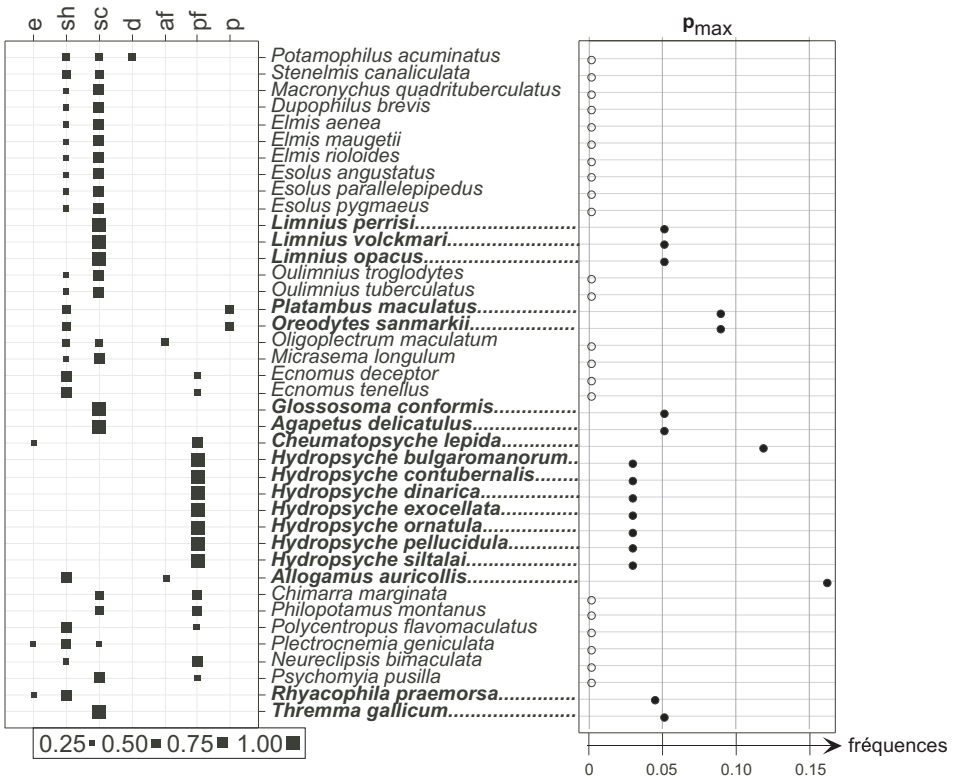


FIG. *43 – Distribution de fréquences maximisant la diversité des habitudes alimentaires de 40 espèces de Trichoptères et Coléoptères de la Loire (fleuve). Sept classes sont considérées : avaleurs ("e", "engulfers"), déchiqueteurs ("sh", "shredders"), gratteurs ("sc", "scrapers"), mangeurs de dépôts ("d", "deposit feeders"), filtreurs actifs ("af", "active filter-feeders"), filtreurs passifs ("pf", "passive filter-feeders"), perceurs ("p", "piercers"). Des dissimilarités δ sont calculées avec la distance d'Edwards (1971). La matrice de dissimilarités ainsi obtenue est euclidienne. Le tableau de gauche donne les affinités (en pourcentages) des espèces pour chaque classe. A droite, un diagramme de Cleveland fournit le vecteur maximisant l'entropie quadratique appliquée aux dissimilarités dans les habitudes alimentaires entre ces 40 espèces de macroinvertébrés. Dix-huit espèces ont une fréquence non-nulle. Cette figure a été réalisée à partir des fonctions 'divcmax', développée personnellement, et 'table2phylog' du package ade4 (Chessel* et al. *2004) de R (Ihaka et Gentleman 1996).*

nulles comme étant de la redondance est insuffisante. Et la seconde est que, si l'entropie quadratique était utilisée sans réflexion préalable sur ce qui est réellement mesuré dans le jeu de données, les résultats obtenus mèneraient parfois à créer une sorte de "règne des extrêmes".

5.2 P

L'indice de Rao est par définition la dissimilarité moyenne entre deux entités tirées au hasard dans une collection, quelle que soit la mesure de dissimilarité choisie. Néanmoins ce choix d'une mesure de dissimilarité a effectivement un impact sur le comportement de l'entropie quadratique ; et dans le cadre de la mesure de la biodiversité, ces différences de comportement

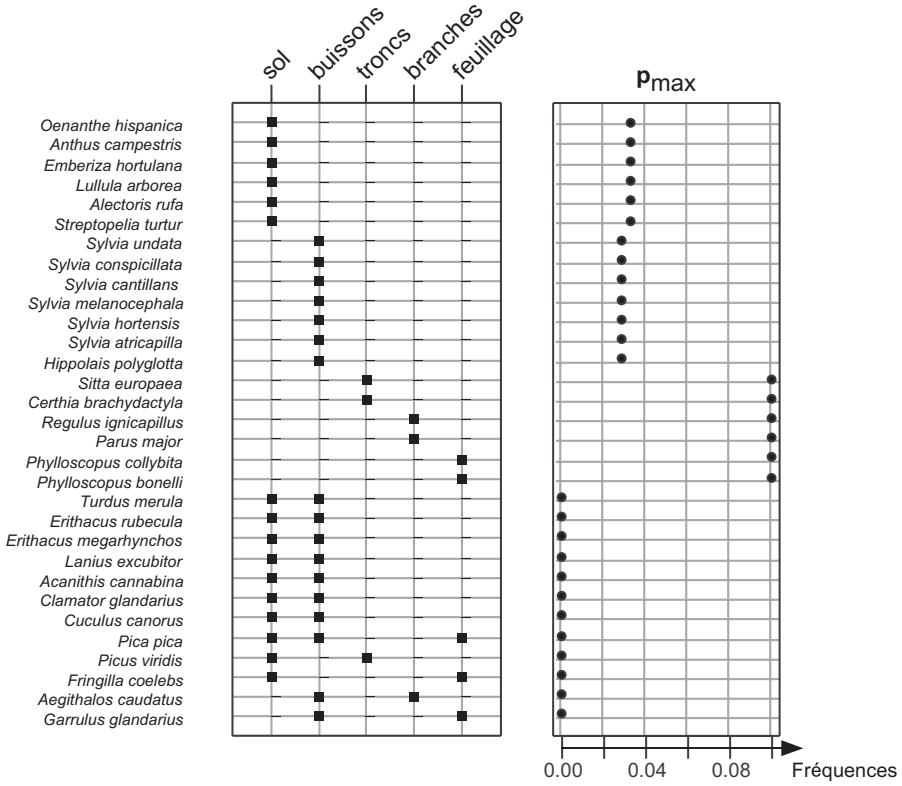


FIG. *44 – Distribution de fréquences maximisant la diversité des sites d'alimentation de 31 espèces d'oiseaux en Provence. Cinq sites d'alimentation ont été observés (Blondel* et al. *1984) : sols, buissons, troncs, branches, feuillage. Des dissimilarités δ sont calculées par la racine de la métrique de Jaccard (Jaccard 1901, cf. partie 4.1.1). La matrice de dissimilarités ainsi obtenue est euclidienne. Dans le tableau de gauche, la présence d'un carré à la ligne k et à la colonne i indique que le site i est utilisé par l'espèce k. A droite, un diagramme de Cleveland fournit le vecteur maximisant l'entropie quadratique appliquée à ces données. Toutes les espèces spécialistes, n'utilisant qu'un seul site, ont une fréquence strictement positive, alors que les autres espèces ont des fréquences nulles. Cette figure a été réalisée à partir des fonctions 'divcmax', développée personnellement, et 'table2phylog' du package ade4 (Chessel* et al. *2004) de R (Ihaka et Gentleman 1996).*

ne doivent pas être éludées.

Staudhammer et LeMay (2001) sous-entendaient que le maximum de diversité pourrait être obtenu dans une collection particulière de taille infinie. Pour toute la suite, nous nous restreindrons à un ensemble fini de catégories qui représentera dans chaque cas "l'ensemble des possibles", par exemple, l'ensemble des espèces d'un taxon donné dans un habitat donné.

5.2.1 Obtention de la valeur maximale exacte de l'entropie quadratique

Avec une écriture matricielle, l'entropie quadratique est égale à

$$H_{\mathbf{D}}(\mathbf{p}) = \mathbf{p}^t\mathbf{D}\mathbf{p}$$

où $\mathbf{p} \in \mathcal{P} = \left\{\mathbf{p} = (p_1, ..., p_k, ..., p_S), p_k \geq 0 \; \forall k, \sum_{k=1}^{S} p_k = 1\right\}$.

Avec un domaine de définition égal à $\mathbb{R}^S$, la fonction s'écrit simplement $\mathbf{x}^t\mathbf{D}\mathbf{x}$. La maximisation de cette fonction a été étudiée dans plusieurs cas particuliers. Izsak et Papp (1995) affirment que, du moment que $\mathbf{D}$ est une matrice symétrique et à valeurs réelles, alors la valeur maximale de l'entropie quadratique est la plus grande valeur propre de $\mathbf{D}$. Cependant cette propriété n'est vraie que si $\mathbf{p}$ appartient à $\mathbf{u}_2$, où $\mathbf{u}_2 = \left\{\mathbf{x} = (x_1, ..., x_k, ..., x_S), \sum_{k=1}^{S} x_k^2 = 1\right\}$. Ainsi, lorsque $||\mathbf{x}||_2 = 1$, c'est-à-dire $\sum_{k=1}^{S} x_k^2 = 1$, alors

$$\sup_{||\mathbf{x}||_2=1} (\mathbf{x}^t\mathbf{D}\mathbf{x}) = \lambda_{\max}(\mathbf{D}),$$

où $\lambda_{\max}(\mathbf{D})$ est la plus grande valeur propre de $\mathbf{D}$. Or l'entropie quadratique n'est pas définie sur $\mathbf{u}_2$ mais sur $\mathcal{P}$. La valeur maximale de l'entropie quadratique n'est donc pas $\lambda_{\max}(\mathbf{D})$.

Etudiant aussi la valeur maximale de l'entropie quadratique, Shimatani (2001) démontre que

$$\sup_{||\mathbf{x}||=1} (\mathbf{x}^t\mathbf{D}\mathbf{x}) = \left(\mathbf{1}_S^t\mathbf{D}^{-1}\mathbf{1}_S\right)^{-1}, \tag{5.1}$$

et

$$\mathbf{x}_{\max} = \mathbf{D}^{-1}\mathbf{1}_S / \left(\mathbf{1}_S^t\mathbf{D}^{-1}\mathbf{1}_S\right). \tag{5.2}$$

Il applique ce résultat à deux matrices de dissimilarités. Pour la première, une matrice de dissimilarité taxonomique, il obtient des résultats cohérents. En revanche, pour la seconde, une matrice de distances basées sur des séquences d'acides aminés, il obtient plusieurs nombres négatifs dans $\mathbf{x}_{\max}$. Shimatani conclut alors que sa formule n'est pas applicable sur ce deuxième exemple et doit se réorienter vers un processus itératif. Donnons quelques explications au problème de Shimatani. Les résultats des équations 5.1 et 5.2 sous-entendent en fait trois hypothèses de départ : (1) $\mathbf{D}$ est inversible ; (2) $\mathbf{1}^t\mathbf{D}^{-1}\mathbf{1} \neq 0$; (3) $\mathbf{x}^t\mathbf{1} = 1$. La deuxième hypothèse équivaut à dire que $\mathbf{D}^{1/2}$ est circum-euclidienne, c'est-à-dire que dans sa représentation euclidienne, les points sont situés sur la bordure d'une hypersphère (Gower 1984). Donc si $\mathbf{D}^{1/2}$ n'est pas circum-euclidienne, $\sup_{||\mathbf{x}||=1}(\mathbf{x}^t\mathbf{D}\mathbf{x}) = \infty$. Autrement dit, $\mathbf{D}$ est une matrice circum-euclidienne si et seulement si $\sup_{||\mathbf{x}||=1}(\mathbf{x}^t\mathbf{D}\mathbf{x}) < \infty$ (Critchley et Fichet 1997). La troisième hypothèse montre que, alors que Shimatani travaillait sur la biodiversité, il ne s'est pas restreint dans cette démonstration à l'ensemble des $\mathbf{x}$ vérifiant $\mathbf{x}^t\mathbf{1} = 1$ mais aussi $x_k \geq 0$ pour tout $k = 1, ..., S$. Or dans le cadre de l'entropie quadratique, nous ne nous intéressons pas à l'ensemble

$$\mathbf{u}_1 = \left\{\mathbf{x} = (x_1, ..., x_k, ..., x_S), \sum_{k=1}^{S} x_k = 1\right\}$$

mais bien à l'ensemble des distributions de fréquences

$$\mathcal{P} = \left\{\mathbf{p} = (p_1, ..., p_k, ..., p_S), p_k \geq 0, \sum_{k=1}^{S} p_k = 1\right\}, \mathcal{P} \subset \mathbf{u}_1.$$

Nous avons vu que la matrice $\mathbf{D}$ est associée à une matrice $\mathbf{\Delta}$ par la relation $\mathbf{\Delta} = [\delta_{kl}]$ et $\mathbf{D} = \left[\delta_{kl}^2/2\right]$. Nous avons également signalé précédemment que l'entropie quadratique est

concave lorsque $\mathbf{\Delta}$ est euclidienne. Les résultats présentés dans la suite de ce chapitre sont limités à ce type de matrices. J'ai établi toutes leurs démonstrations en faisant appel à l'algèbre linéaire et à la géométrie euclidienne. Elles pourront être consultées à la fin de ce mémoire dans Pavoine *et al.* (2005b, cf. annexe 3). Résumons les résultats démontrés dans cet article. Pour trouver la valeur maximale exacte de $H_{\mathbf{D}}(\mathbf{p}) = \mathbf{p}^t\mathbf{D}\mathbf{p}$, où $\mathbf{p} \in \mathcal{P}$ les étapes sont les suivantes :

1. Obtenir une représentation euclidienne des S points séparés par les distances $\{\delta_{kl}\}$;
2. Obtenir le plus petit cercle (en dimension 2), la plus petite sphère (en dimension 3) ou la plus petite hypersphère (pour plus de 3 dimensions) contenant tous les S points (le terme d'hypersphère sera utilisé par la suite plus généralement pour désigner un cercle, une sphère ou une boule à plus de 3 dimensions) ;
3. La valeur maximale de $H_{\mathbf{D}}$ est le carré du rayon de cette hypersphère (proposition 1 dans Pavoine *et al.* 2005b, cf. annexe 3).

Pour trouver la (ou une) distribution de fréquences conduisant à la valeur maximale de $H_{\mathbf{D}}$, les étapes sont :

1. Sélectionner les K points situés sur la bordure de la plus petite hypersphère. Il s'agit de l'ensemble des points supportant l'hypersphère ("ensemble support", noté T).
2. Soit $\mathbf{D}_K$ la matrice $(K \times K)$ contenant les valeurs d_{kl} entre les K points supportant l'hypersphère, alors $\max_{\mathbf{p}}(H_{\mathbf{D}}(\mathbf{p})) = \left(\mathbf{1}_K^t\mathbf{D}_K^{-1}\mathbf{1}_K\right)^{-1}$ (proposition 2 dans Pavoine *et al.* 2005b, cf. annexe 3).
3. Soit $\mathbf{p}_{\max}$ une distribution de fréquences maximisant $H_{\mathbf{D}}$, les fréquences des $S - K$ points situés strictement à l'intérieur de l'hypersphère sont nulles, et les fréquences des K points supportant l'hypersphère sont données par $\mathbf{D}_K^{-1}\mathbf{1}_K / \left(\mathbf{1}_K^t\mathbf{D}_K^{-1}\mathbf{1}_K\right)$ (proposition 3 dans Pavoine *et al.* 2005b, cf. annexe 3).

La matrice $\mathbf{\Delta}_K$ contenant les distances δ_{kl} entre les K points supportant l'hypersphère, est par définition circum-euclidienne, puisqu'elle définit dans un espace euclidien K points situés sur la bordure d'une hypersphère. Le centre de l'hypersphère se situe au barycentre des K points pondérés par les valeurs de $\mathbf{D}_K^{-1}\mathbf{1}_K / \left(\mathbf{1}_K^t\mathbf{D}_K^{-1}\mathbf{1}_K\right)$. Ces valeurs sont positives ou nulles si et seulement si l'hypersphère est la plus petite hypersphère contenant les K points, et c'est le cas ici. En effet, la position des points situés strictement à l'intérieur de l'hypersphère n'a aucune influence sur la définition de cette hypersphère. Une fois qu'un ensemble support de points a été trouvé, l'hypersphère supportée par ces points est inchangée par l'ajout de points dans et sur la bordure de l'hypersphère. Ainsi la plus petite hypersphère contenant les S points est aussi la plus petite hypersphère contenant les K points supports. Donc les valeurs de $\mathbf{D}_K^{-1}\mathbf{1}_K / \left(\mathbf{1}_K^t\mathbf{D}_K^{-1}\mathbf{1}_K\right)$ sont positives. Soient O le centre de l'hypersphère et T l'ensemble support. Soient $\mathbf{x}_O$ et $\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_K$ les vecteurs de coordonnées du centre O et des K points du support dans l'espace euclidien. Notons $(\lambda_1, \lambda_2, ..., \lambda_K)$ les valeurs du vecteur $\mathbf{D}_K^{-1}\mathbf{1}_K / \left(\mathbf{1}_K^t\mathbf{D}_K^{-1}\mathbf{1}_K\right)$. L'égalité suivante relie $\mathbf{x}_O$ aux coordonnées $\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_K$:

$$\mathbf{x}_O = \sum_{k=1}^{K} \lambda_k \mathbf{x}_k,$$

Démonstration : La démonstration se déduit immédiatement des articles de Gower (1982, 1984) qui prouve que, dans la représentation euclidienne d'une matrice $\mathbf{D}_n$ circum-euclidienne et de dimensions $n \times n$, le centre de

l'hypersphère sur laquelle se situent les n points a pour coordonnées $\mathbf{s}^t\mathbf{X}$ où $\mathbf{s}$ est le vecteur $\mathbf{D}_n^{-1}\mathbf{1}_n / \left(\mathbf{1}_n^t\mathbf{D}_n^{-1}\mathbf{1}_n\right)$, et $\mathbf{X}$ est la matrice des coordonnées des n points. Notons que dans la démonstration de Gower le vecteur $\mathbf{D}_n^{-1}\mathbf{1}_n / \left(\mathbf{1}_n^t\mathbf{D}_n^{-1}\mathbf{1}_n\right)$ contient des valeurs positives ou négatives, puisque, pour cette démonstration, l'hypersphère en question n'a pas besoin d'être la plus petite hypersphère qui englobe les n points.

On dit que $O \in conv(T)$. Il peut être ajouté que, quelle que soit la dimension de l'espace euclidien considéré, $K \geq 2$. Ainsi, même si le nuage des S points est en dimension 100, deux points peuvent parfois suffire pour l'ensemble support.

Prenons quelques exemples à une et deux dimensions. Lorsque la représentation euclidienne se situe dans un espace à une seule dimension, l'entropie quadratique correspond à une variance. Il n'est plus possible de parler de cercle, ni de sphère, ni d'hypersphère. Les S points correspondent dans ce cas à S valeurs numériques ordonnées. Les deux points les plus extrêmes jouent le rôle de la frontière entourant tous les autres points. Soit δ la distance entre ces deux points et $d = \delta^2/2$.

$$\mathbf{D}_2^{-1} = \begin{pmatrix} 0 & d \\ d & 0 \end{pmatrix}^{-1} = \begin{pmatrix} 0 & 1/d \\ 1/d & 0 \end{pmatrix},$$

$$\max_{\mathbf{p}}(H_{\mathbf{D}}(\mathbf{p})) = \left(\mathbf{1}_2^t\mathbf{D}_2^{-1}\mathbf{1}_2\right)^{-1} = d/2 = (\delta/2)^2 = H_{\mathbf{D}}(\mathbf{p}_{\max})$$

où

$$\mathbf{p}_{\max} = (\mathbf{D}_2^{-1}\mathbf{1}_2 \Big/ \left(\mathbf{1}_2^t\mathbf{D}_2^{-1}\mathbf{1}_2\right), \overbrace{0, ..., 0}^{(S-2)\times 0}) = (1/2, 1/2, \overbrace{0, ..., 0}^{(S-2)\times 0}).$$

Les $S - 2$ valeurs nulles dans $\mathbf{p}_{\max}$ sont les fréquences des valeurs intermédiaires et 0.5 est la fréquence de chaque valeur extrême.

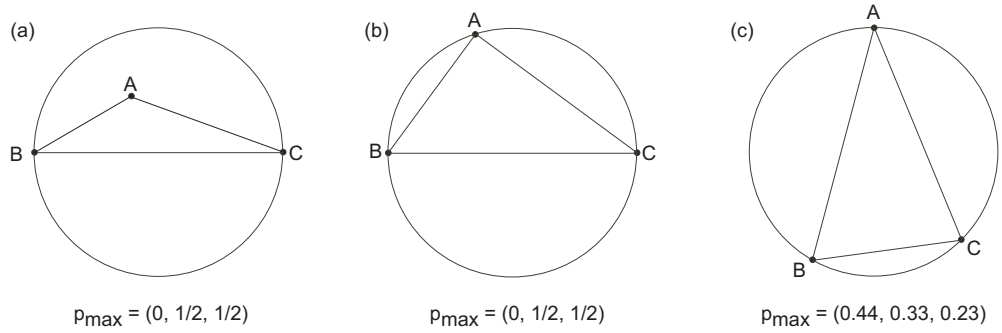


FIG. *45 – Trois cas possibles avec trois catégories en dimension 2 : a) un angle obtus, b) un angle droit, c) trois angles aigus. Le vecteur* $\mathbf{p}_{max}$ *contient les fréquences des points A, B et C dans cet ordre.*

L'exemple le plus simple à deux dimensions contient trois points. Ces trois points vont former un triangle. Il existe alors trois cas possibles (Fig. 45) :

1. un des angles du triangle est obtus ;
2. un des angles du triangle est droit ;

3. les trois angles du triangle sont aigus.

Dans le premier cas, le plus petit cercle est porté par deux points (Fig. 45a), les points B et C. La catégorie A située au sommet de l'angle obtus est dans le cercle, elle a donc une fréquence nulle pour maximiser l'entropie quadratique. Dans le troisième cas, les trois catégories ont des fréquences différentes et non nulles pour maximiser l'entropie quadratique (Fig. 45c). Le deuxième cas représente la situation de transition entre les cas 1 et 3. La catégorie A située au sommet de l'angle droit est sur le cercle, mais la position du centre du cercle est entièrement déterminée par les deux autres points, B et C, puisque le centre se situe au milieu du segment [BC] (Fig. 45b), diamètre du cercle.

Pour illustrer avec plus de catégories ces différentes étapes du processus de maximisation, sortons du domaine de la biologie pour considérer la dispersion spatiale d'une population. La collection est la France métropolitaine continentale ; les entités sont les habitants et les catégories sont les départements. Chaque département est caractérisé par son centre géographique. La dissimilarité entre deux départements est calculée par la métrique euclidienne (distance spatiale) entre les coordonnées de leurs centres géographiques. Les distances spatiales sont des dissimilarités δ dans les notations de l'entropie quadratique. Nous disposons donc d'une collection (France métropolitaine continentale) contenant des entités (habitants) regroupées en catégories (départements) et d'une matrice $\mathbf{\Delta}$ de distances entre ces catégories. L'unité de l'entropie quadratique appliquée à ces données sera donc celle de $d = \delta^2/2$, c'est-à-dire le carré d'une distance spatiale divisée par deux. Concrètement, l'entropie quadratique appliquée à ces données mesure la diversité spatiale des français en France métropolitaine continentale par l'espérance du carré de la distance entre deux français tirés au hasard avec remise. Si les habitants étaient répartis uniformément entre les départements, la diversité spatiale serait seulement à 38% du maximum (Fig. 46a). Avec un gradient Sud-Nord de répartition des habitants, cette diversité serait encore plus faible : 32% du maximum (Fig. 46b). Si le peuplement d'un département était une fonction directe de la somme des distances qui le séparent des autres départements, la diversité spatiale serait à 46% du maximum seulement (Fig. 46c). La diversité maximale théorique ne retient que trois départements : le Finistère, les Alpes-Maritimes et le Bas-Rhin. Les peuplements relatifs de ces départements devraient être respectivement 48.7%, 46.4% et 4.9%. Les centres géographiques de ces trois départements sont en effet les seuls situés sur le plus petit cercle circonscrit à l'ensemble des centres des départements (Fig. 47). La distance entre le Finistère et les Alpes-maritimes est presque égale au diamètre du cercle, c'est pourquoi le Bas-Rhin est sous-représenté dans cette distribution théorique. Cet exemple et celui de la variance montrent que plus le nombre de points excède le nombre de dimensions de l'espace euclidien, plus la richesse (nombre de catégories) risque d'être considérablement diminuée pour maximiser l'entropie quadratique. Nous donnons un exemple plus biologique et des exemples théoriques choisis en dimension 2 dans l'article Pavoine *et al.* (2005b, cf. annexe 3). L'exemple de la diversité chromatique (partie 5.1.3) correspond également à une représentation graphique à deux dimensions (cf. Fig. 48).

5.2.2 Matrices de distances SEH-circum-euclidiennes

J'ai introduit le qualificatif "SEH-circum-euclidienne" pour une matrice de dissimilarités (Pavoine *et al.* 2005b, cf. annexe 3). Une matrice SEH-circum-euclidienne est d'abord eu-

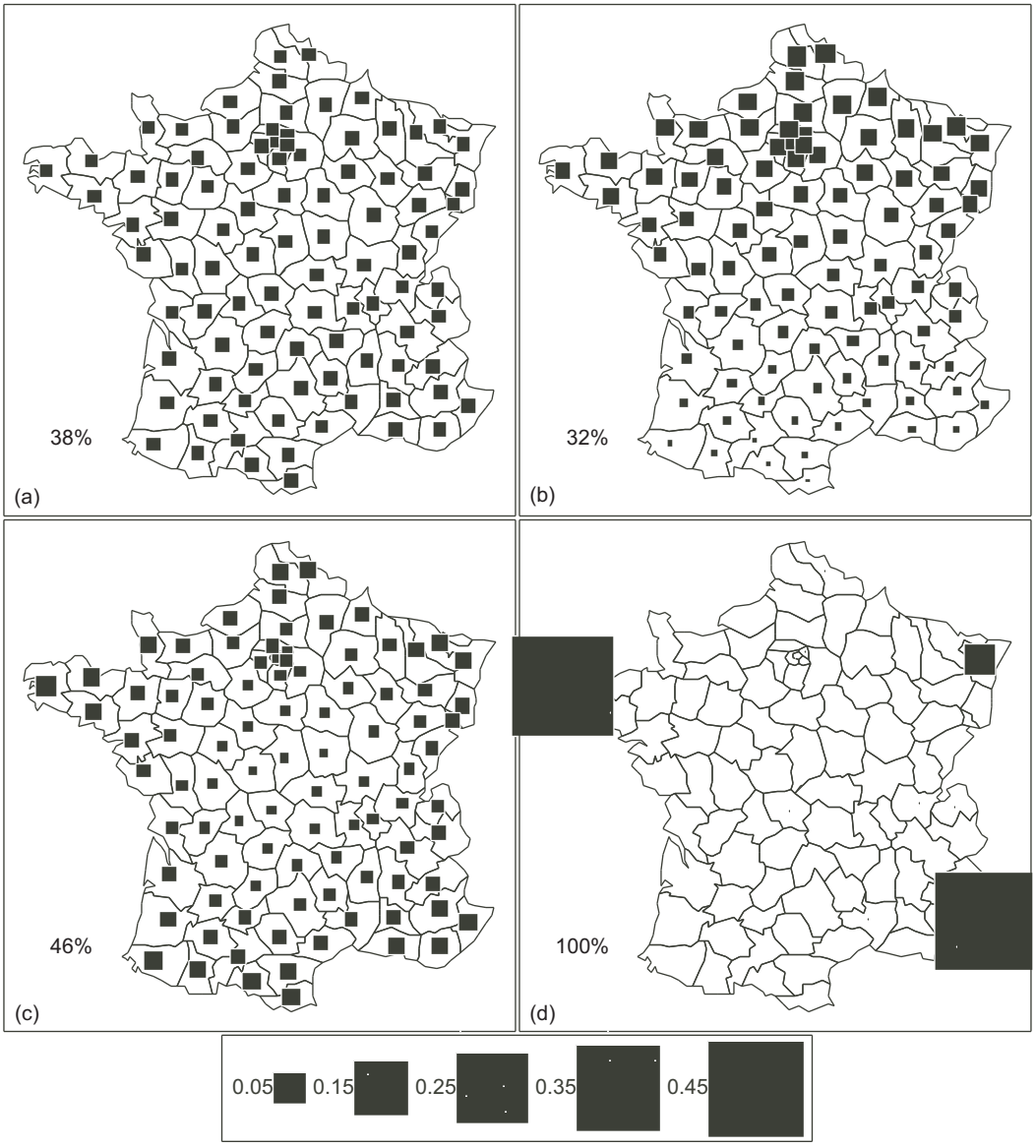


FIG. *46 – Diversité spatiale* $\mathbf{p}^t\mathbf{D}\mathbf{p}$, $\mathbf{D} = \left[\delta_{kl}^2/2\right]$ *où* δ_{kl} *est la distance spatiale entre les centres géographiques de deux départements k et l : (a) distribution uniforme* $\mathbf{p} = (1/94, ..., 1/94)^t$, *(b) gradient Nord-Sud* $\mathbf{p} = \mathbf{y}/(\mathbf{y}^t\mathbf{1})$ *(***y** *: vecteur des ordonnées), (c)* $\mathbf{p} = \mathbf{D1}/(\mathbf{1}^t\mathbf{D1})$, *(d)* $\mathbf{p} = \mathbf{p}_{max}$. *Les carrés indiquent les distributions de fréquences : répartitions théoriques des habitants entre les départements. Les valeurs de diversité sont indiquées en pourcentages du maximum.*

clidienne donc associable à une représentation graphique dans un espace euclidien. Elle est ensuite circum-euclidienne, c'est-à-dire que dans cette représentation graphique, les points se situent sur la bordure d'une hypersphère. Elle est enfin SEH-circum-euclidienne, c'est-à-dire que cette hypersphère est la plus petite qui contient tous les points. Le terme "SEH" désigne "Smallest-Enclosing-Hypersphere".

Pour un ensemble de S catégories, $\max_{\mathbf{p}}(H_{\mathbf{D}}(\mathbf{p})) = H_{\mathbf{D}}(\mathbf{p}_{\max}) = \left(\mathbf{1}_S^t\mathbf{D}^{-1}\mathbf{1}_S\right)^{-1}$ et $\mathbf{p}_{\max} =$

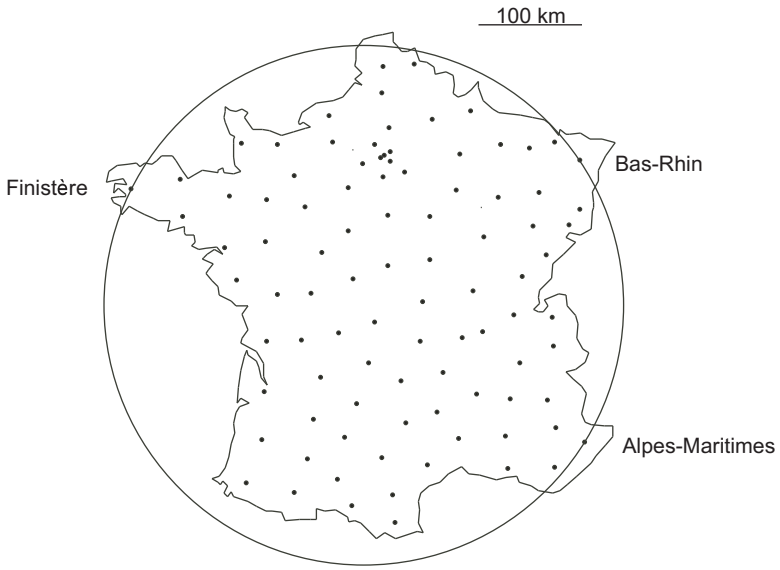


FIG. *47 – Diversité spatiale maximale, départements retenus pour maximiser l'entropie quadratique.*

$\mathbf{D}^{-1}\mathbf{1}_S / \left(\mathbf{1}_S^t \mathbf{D}^{-1}\mathbf{1}_S\right)$ si et seulement si $\mathbf{\Delta}$ est SEH-circum-euclidienne (Pavoine *et al.* 2005b, cf. annexe 3). Ce vecteur $\mathbf{p}_{\max}$ peut contenir des valeurs positives ou nulles mais c'est parmi ces matrices SEH-circum-euclidiennes que se trouvent celles qui lorsqu'elles sont appliquées à l'entropie quadratique conduisent à des vecteurs $\mathbf{p}_{\max}$ ne contenant pas de 0.

On peut définir alors deux sous-classes de matrices SEH-circum-euclidiennes (Pavoine *et al.* 2005b, cf. annexe 3). La première est dite "faible" parce que le vecteur $\mathbf{p}_{\max}$ peut contenir plusieurs valeurs nulles. La deuxième est dite "forte" parce que le vecteur $\mathbf{p}_{\max}$ ne contient que des valeurs strictement positives.

D'après des simulations, il semble possible de transformer une matrice de dissimilarités quelconque en une matrice SEH-circum-euclidienne. Deux méthodes ont été étudiées. La première consiste à ajouter une même constante aux termes non diagonaux de $\mathbf{\Delta}$. La seconde méthode est l'ajout d'une même constante aux termes non diagonaux de $\mathbf{D}$. Ces deux méthodes s'apparentent, respectivement, à un autre niveau, aux procédés de Cailliez (1983) et Lingoes (1971) utilisés pour transformer une matrice de dissimilarités quelconque en une matrice euclidienne. Pourquoi utiliser ces procédés ? La réponse est assez intuitive. Prenons la première méthode. Rajouter une même constante aux termes non diagonaux de $\mathbf{\Delta}$ correspond à l'opération suivante :

$$\mathbf{\Delta}' = \mathbf{\Delta} + c\,(\mathbf{1}\mathbf{1}^t - \mathbf{I})$$

où $c > 0$. La matrice $c\,(\mathbf{1}\mathbf{1}^t - \mathbf{I})$ est elle-même une matrice de dissimilarités, et plus précisément une matrice euclidienne. Sa représentation graphique est, pour trois catégories, un triangle équilatéral, pour quatre catégories un tétraèdre régulier, etc. Le cercle circonscrit à un triangle équilatéral est le plus petit cercle entourant ces trois points (cf. cas de la figure 45c et partie 5.2.1). Cette propriété se généralise à quatre points et plus (cf. partie 5.2.3). La représentation graphique d'une matrice de dissimilarité uniforme de dimensions $S \times S$ est une figure régulière à

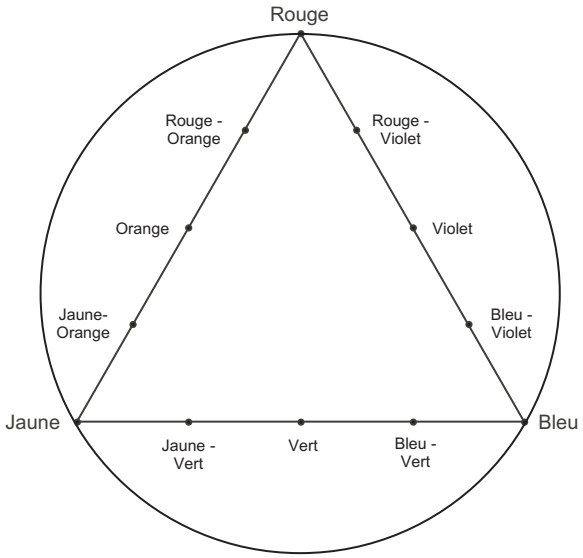


FIG. *48 – Maximisation de la diversité chromatique. Tous les points représentant les couleurs primaires, secondaires et intermédiaires se situent sur un triangle. Le plus petit cercle contenant tous ces points, et aussi tous ceux qui pourraient être rajoutés dans le triangle, est représenté. La valeur maximale de l'entropie quadratique est obtenue lorsque seules les trois couleurs primaires sont utilisées, en proportions égales.*

$S-1$ dimensions pour laquelle la distance entre deux sommets est toujours égale à 1. Donc plus c est grande, plus la représentation graphique associée à un $\mathbf{\Delta}'$ ressemblera à celle de $c\,(\mathbf{1}\mathbf{1}^t - \mathbf{I})$ puisque plus c augmente plus $\mathbf{\Delta}$ devient négligeable dans la définition de $\mathbf{\Delta}'$. Il semble exister une plus petite valeur de c à partir de laquelle la matrice $\mathbf{\Delta}'$ devient SEH-circum-euclidienne.

Lorsque seulement trois catégories sont considérées, la valeur exacte de la constante rajoutée à $\mathbf{D}$ dans la deuxième méthode présente une solution simple. Considérons que la catégorie A se situe au sommet d'un angle obtus et rajoutons une constante c aux termes non diagonaux de $\mathbf{D}$

$$\mathbf{D} + c\,(\mathbf{1}_3\mathbf{1}_3^t - \mathbf{I}_3) = \begin{pmatrix} 0 & d_{AB}+c & d_{AC}+c \\ d_{AB}+c & 0 & d_{BC}+c \\ d_{AC}+c & d_{BC}+c & 0 \end{pmatrix}$$

La constante c la plus petite est celle qui placera la catégorie A au sommet d'un angle droit. Cette constante est telle que

$$\left(2\sqrt{d_{AB}+c}\right)^2 + \left(2\sqrt{d_{AC}+c}\right)^2 = \left(2\sqrt{d_{BC}+c}\right)^2$$

c'est-à-dire

$$c = d_{BC} - d_{AB} - d_{AC}$$

Quand $c \to \infty$ le triangle tend vers un triangle équilatéral et $\mathbf{p}_{\max} \to (1/3, 1/3, 1/3)$. Prenons le triangle obtus de la figure 45a. Les matrices $\mathbf{\Delta}$ et $\mathbf{D}$ correspondant à cette figure sont

$$\mathbf{\Delta} = \begin{pmatrix} 0 & 0.9 & 1.3 \\ 0.9 & 0 & 2 \\ 1.3 & 2 & 0 \end{pmatrix} \text{ et } \mathbf{D} = \begin{pmatrix} 0 & 0.405 & 0.845 \\ 0.405 & 0 & 2 \\ 0.845 & 2 & 0 \end{pmatrix}$$

La constante c la plus petite est égale à 0.75. Elle est plus de deux fois plus élevée que $d_{AB} = 0.405$. La déformation est importante. Une représentation euclidienne comportant beaucoup de points dans un espace à peu de dimensions nécessitera une constante très élevée. L'information donnée par la matrice de départ en sera donc très modifiée.

En conclusion, les transformations d'une matrice quelconque en une matrice SEH-circum-euclidienne peuvent dans certains cas déformer fortement la matrice de départ. Existe-t-il alors des matrices connues qui seraient, de par leur définition, déjà SEH-circum-euclidiennes ?

5.2.3 Importance des matrices de dissimilarités ultramétriques

Rappelons la définition d'une matrice de dissimilarités ultramétrique. Une matrice $\mathbf{D} = [d_{kl}]$ de dissimilarités est ultramétrique si et seulement si

$$d_{kl} \leq \max(d_{ki}, d_{il}) \text{ pour tout } k, l \text{ et } i.$$

Une matrice ultramétrique est associée à un arbre raciné que l'on qualifie aussi d'ultramétrique. Sur cet arbre, la plus petite somme des longueurs de branches reliant deux feuilles k et l est égale à $2d_{kl}$. La valeur d_{kl} est précisément la plus petite somme de longueurs de branches joignant k et l à leur nœud interne commun le plus proche (cf. Fig. 49).

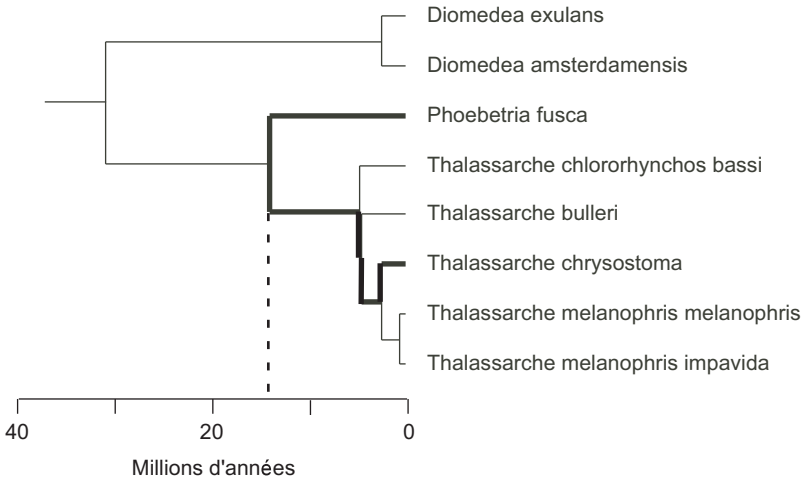


FIG. *49 – Exemple d'arbre ultramétrique : la phylogénie. Sur cet arbre, la distance entre les espèces* Phoebetria fusca *et* Thalassarche chrysostoma*, par exemple, est égale à 14 millions d'années. Données extraites de Bried* et al. *(2003), jeu de données 'procella' dans le package 'ade4' de R (Ihaka et Gentleman 1996, Chessel* et al. *2004).*

Toute matrice uniforme $\lambda(\mathbf{1}\mathbf{1}^t - \mathbf{I})$, où λ appartient à $\mathbb{R}^{+*}$, est ultramétrique. La matrice de dissimilarités utilisée pour la variance d'une variable qualitative (indice de Gini-Simpson), est $(\mathbf{1}\mathbf{1}^t - \mathbf{I})$, c'est-à-dire une matrice uniforme où $\lambda = 1$. Elle est donc ultramétrique. L'arbre ultramétrique correspondant à une matrice de distance uniforme a la forme d'un râteau (Fig. 50).

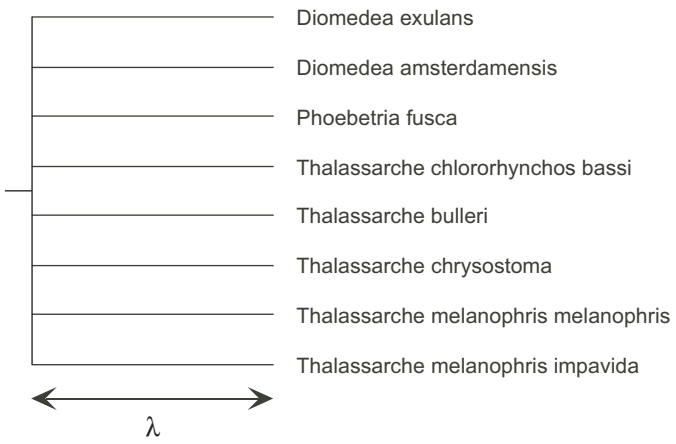


FIG. *50 – Exemple d'arbre ultramétrique associé à une matrice de distances uniformes notée* $\mathbf{D} = \lambda(\mathbf{1}\mathbf{1}^t - \mathbf{I})$*, où* $\lambda > 0$ *: arbre en forme de râteau. Sur cet arbre, la distance entre deux espèces est toujours égale à* λ*. Les étiquettes, noms des espèces, sont interchangeables, puisque leurs permutations ne changent pas la structure de l'arbre.*

Toute matrice ultramétrique appartient à la sous-classe forte de l'ensemble des matrices SEH-circum-euclidiennes (proposition 4 dans Pavoine *et al.* 2005b, cf. annexe 3).

Dans le cas particulier de l'indice de Gini-Simpson, la matrice $\mathbf{D}_{\text{G-S}}^{-1}$ contient la valeur $-(S-2)/(S-1)$ sur toute la diagonale, et la valeur $1/(S-1)$ en dehors de la diagonale :

$$\mathbf{D}_{\text{G-S}}^{-1} = (\mathbf{1}\mathbf{1}^t - \mathbf{I})^{-1} = \begin{pmatrix} 0 & 1 & \cdots & 1 \\ 1 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 1 & \cdots & 1 & 0 \end{pmatrix}^{-1} = \begin{pmatrix} -\frac{S-2}{S-1} & \frac{1}{S-1} & \cdots & \frac{1}{S-1} \\ \frac{1}{S-1} & -\frac{S-2}{S-1} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \frac{1}{S-1} \\ \frac{1}{S-1} & \cdots & \frac{1}{S-1} & -\frac{S-2}{S-1} \end{pmatrix}$$

La valeur maximale sur $\mathcal{P}$ de $H_{\mathbf{D}_{\text{G-S}}}(\mathbf{p})$ est $\left(\mathbf{1}_S^t \mathbf{D}_{\text{G-S}}^{-1} \mathbf{1}_S\right)^{-1}$, c'est-à-dire l'inverse de la somme totale des termes de $\mathbf{D}_{\text{G-S}}^{-1}$. Ici

$$\max_{\mathbf{p}}(H_{\mathbf{D}_{\text{G-S}}}(\mathbf{p})) = \left(-\frac{S-2}{S-1}S + S(S-1)\frac{1}{S-1}\right)^{-1} = \frac{S-1}{S}.$$

Et la distribution de fréquences au maximum est

$$\begin{aligned} \mathbf{p}_{\max} &= \mathbf{D}_{\text{G-S}}^{-1}\mathbf{1}_S \left(\mathbf{1}_S^t \mathbf{D}_{\text{G-S}}^{-1}\mathbf{1}_S\right)^{-1} \\ &= \left(\left[-\frac{S-2}{S} + (S-1)\frac{1}{S-1}\right]\left[\frac{S-1}{S}\right]^{-1}, ..., \left[-\frac{S-2}{S} + (S-1)\frac{1}{S-1}\right]\left[\frac{S-1}{S}\right]^{-1}\right) \\ &= \left(\frac{1}{S}, ..., \frac{1}{S}\right) \end{aligned}$$

Nous retrouvons ces résultats connus (partie 2.2.1).

Dans la partie précédente, il a été dit que $\mathbf{p}_{\max}$ est la (*ou une*) distribution de fréquences conduisant au maximum. En effet, pour des dissimilarités quelconques, parfois, plusieurs distributions de fréquences peuvent conduire à la valeur maximale de l'entropie quadratique.

Considérons la matrice de distances suivante entre quatre catégories théoriques A, B, C et D.

$$\boldsymbol{\Delta}^{\text{cat}} = \begin{array}{c} \\ \text{A} \\ \text{B} \\ \text{C} \\ \text{D} \end{array} \begin{array}{c} \begin{array}{cccc} \text{A} & \text{B} & \text{C} & \text{D} \end{array} \\ \begin{pmatrix} 0 & \sqrt{2} & 2 & \sqrt{2} \\ \sqrt{2} & 0 & \sqrt{2} & 2 \\ 2 & \sqrt{2} & 0 & \sqrt{2} \\ \sqrt{2} & 2 & \sqrt{2} & 0 \end{pmatrix} \end{array}$$

Dans leur représentation euclidienne, ces quatre catégories forment les quatre sommets A, B, C, D d'un carré de côté de longueur $\sqrt{2}$. Le plus petit cercle entourant les points A, B, C, D passe par ces points, possède son centre au centre du carré et a un rayon de 1 unité (Fig. 51).

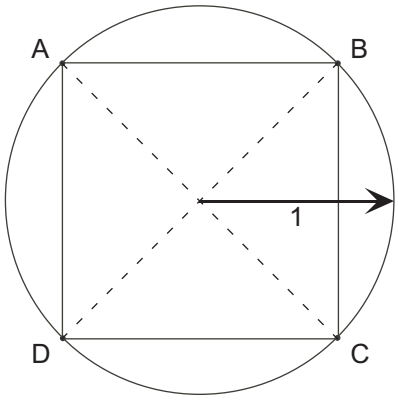


FIG. *51 – Représentation euclidienne des quatre catégories A, B, C et D.*

La valeur maximale de $H_{\boldsymbol{\Delta}}$ est le carré du rayon du plus petit cercle contenant les 4 catégories A, B, C et D. Elle vaut donc 1. Soit $\mathbf{p} = (p_{\text{A}}, p_{\text{B}}, p_{\text{C}}, p_{\text{D}})$. La distribution de fréquences

$$\mathbf{p}_{\max} = \mathbf{D}^{-1}\mathbf{1}_S \,/\, \left(\mathbf{1}_S^t \mathbf{D}^{-1}\mathbf{1}_S\right) = (1/4, 1/4, 1/4, 1/4)^t$$

conduit à ce maximum. Mais elle n'est pas unique. Par exemple, les vecteurs de fréquences $\mathbf{p'}_{\max} = (1/2, 0, 1/2, 0)^t$ et $\mathbf{p''}_{\max} = (0, 1/2, 0, 1/2)^t$ donnant des poids de 1/2 à deux sommets opposés, conduisent aussi au maximum : $H_{\boldsymbol{\Delta}}(\mathbf{p}_{\max}) = H_{\boldsymbol{\Delta}}(\mathbf{p'}_{\max}) = H_{\boldsymbol{\Delta}}(\mathbf{p''}_{\max}) = 1$. Le centre O du plus petit cercle entourant les points A, B, C, D est le barycentre de ces points, qu'ils soient pondérés par $\mathbf{p}_{\max}$, par $\mathbf{p'}_{\max}$ ou par $\mathbf{p''}_{\max}$. N'importe quel vecteur de fréquences défini par $(a/2, b/2, a/2, b/2)^t$, tels que $a, b \geq 0$ et $a + b = 1$, conduit au maximum. Il en existe donc une infinité. Cette situation porte bien sûr sur un cas très particulier avec quatre points en dimension 2, et la forme très particulière d'un carré. Mais elle prouve que plusieurs distributions

peuvent conduire au maximum. Les études précédentes (Shimatani 2001, Izsak et Szeidl 2002, Champely et Chessel 2002) de la valeur maximale de l'entropie quadratique étaient basées sur des processus itératifs. La distribution de fréquences qui était ainsi obtenue pour chaque cas n'était peut-être pas unique. Et lorsque Shimatani (2001), Izsak et Szeidl (2002) et Champely et Chessel (2002) ont observé des valeurs nulles dans la distribution trouvée par le processus itératif au maximum peut-être qu'une autre distribution conduisant à la même valeur maximale de diversité mais ne présentant aucune valeur nulle existait aussi, et inversement. La distribution obtenue par une méthode itérative n'est donc pas obligatoirement unique et est fortement dépendante de la distribution choisie au départ de l'itération.

Dans le cas des catégories A, B, C et D formant un carré, le fait qu'aucun des points ne chevauche un autre signifie que chaque catégorie est unique. Si ces catégories représentaient des espèces, elles auraient donc chacune des caractères propres. Et si une espèce était éliminée, tous ses caractères uniques seraient perdus de l'ensemble. D'après les informations dont nous disposons ici, la perte d'une des espèces serait équivalente à la perte d'une autre car la symétrie de la représentation des dissimilarités entre les catégories fait qu'une permutation quelconque des étiquettes "A", "B", "C" et "D" entre les sommets du carré ne changerait pas la valeur de l'entropie quadratique. Ces espèces devraient donc avoir le même poids dans la mesure de la diversité, ce qui correspond au vecteur $\mathbf{p}_{\max}$. Si plusieurs distributions de fréquences très différentes peuvent conduire à la valeur maximale de l'entropie quadratique, finalement, l'entropie quadratique est-elle un indice de diversité ?

De plus, les matrices de dissimilarités calculées entre espèces biologiques par exemple, sont beaucoup plus complexes. Il n'est cependant pas impossible et certainement pas rare d'obtenir plusieurs distributions de fréquences au maximum. Dans ce cas, comment interpréter les résultats fournis par l'entropie quadratique ?

Nous venons de mettre en évidence deux problèmes liés à la maximisation de l'entropie quadratique :

1. la possibilité de maximiser l'indice en réduisant la richesse ;
2. la présence de plusieurs distributions de fréquences radicalement différentes au maximum.

Nous avons vu que le premier problème est évité en utilisant des dissimilarités ultramétriques. Nous allons voir que le deuxième l'est aussi :

Pour toute matrice ultramétrique $\mathbf{D}$, il existe une unique distribution de fréquences $\mathbf{p}_{\max}$ vérifiant

$$H_{\mathbf{D}}(\mathbf{p}_{\max}) = \max_{\mathbf{p}} \left[H_{\mathbf{D}}(\mathbf{p}) \right].$$

En effet, toute matrice ultramétrique $\mathbf{D}$ de dimensions $S \times S$ est de rang $S - 1$. La distribution $\mathbf{p}_{\max}$ est unique car $\mathbf{D}$ est SEH-circum-euclidienne et chaque catégorie définit une dimension.

Puisque cette distribution est unique et que toutes les espèces sont incluses dans cette distribution, alors la valeur maximale de l'entropie quadratique est une fonction qui augmente par l'ajout d'une espèce dans un ensemble :

Soient E un ensemble de S espèces et j une autre espèce ($j \notin E$), soient $\mathbf{p}_E$ une distribution de fréquences de E et $\mathbf{p}_{E\cup j}$ une distribution de fréquences de $E \cup j$, alors

$$\max_{\mathbf{p}_E} (\mathbf{p}_E^t \mathbf{D} \mathbf{p}_E) < \max_{\mathbf{p}_{E\cup j}} \left(\mathbf{p}_{E\cup j}^t \mathbf{D} \mathbf{p}_{E\cup j}\right).$$

Démonstration : Soit **D** la matrice contenant l'ensemble des dissimilarités entre les espèces de $E \cup j$. Tous les vecteurs $\mathbf{p}_E$ s'écrivent sous la forme $\mathbf{p} = (p_1, ..., p_S, 0)^t$, 0 étant la fréquence de l'espèce j. Le vecteur $\mathbf{p}_{\max E}$ qui maximise $\mathbf{p}_E^t \mathbf{D} \mathbf{p}_E$ sur $\mathcal{P}$ s'écrit donc $(p_{\max 1}, ..., p_{\max S}, 0)^t$ où $p_{\max k} > 0$ pour tout k, $1 \leq k \leq S$. Or l'unique distribution de fréquence qui maximise $\mathbf{p}_{E\cup j}^t \mathbf{D} \mathbf{p}_{E\cup j}$ ne contient pas de valeur nulle donc

$$\mathbf{p}_{\max E}^t \mathbf{D} \mathbf{p}_{\max E} < \max_{\mathbf{p}_{E\cup j}} \left(\mathbf{p}_{E\cup j}^t \mathbf{D} \mathbf{p}_{E\cup j}\right),$$

c'est-à-dire

$$\max_{\mathbf{p}_E} \left(\mathbf{p}_E^t \mathbf{D} \mathbf{p}_E\right) < \max_{\mathbf{p}_{E\cup j}} \left(\mathbf{p}_{E\cup j}^t \mathbf{D} \mathbf{p}_{E\cup j}\right).$$

Les distances uniformes entre catégories, quelle que soit la nature de ces catégories, et les distances phylogénétiques entre espèces ont été citées au début de cette partie comme exemple de dissimilarités ultramétriques. Elles ne sont pas les seuls cas connus de dissimilarités ultramétriques. Les distances taxonomiques sont ultramétriques, ce qui confirme et généralise l'observation de Shimatani (2001) : avec des dissimilarités taxonomiques aucune fréquence nulle ne peut apparaître dans l'unique vecteur qui maximise l'entropie quadratique. D'une façon générale toute distance définie comme la plus petite somme des longueurs de branches qui séparent deux catégories dans un arbre ultramétrique où toutes les feuilles (catégories) sont alignées est ultramétrique.

Lorsque les dissimilarités entre catégories sont ultramétriques, $\mathbf{p}_{\max}$ possède des propriétés intéressantes : les catégories les plus communes sur l'arbre ultramétrique doivent avoir les plus faibles fréquences, et les catégories les plus isolées les plus fortes fréquences pour maximiser l'entropie quadratique. Nous allons maintenant étudier l'intérêt pratique de ces propriétés pour la mesure de la biodiversité. L'attribution de poids plus importants à des espèces originales est une des facettes de la biologie de la conservation.

5.3 I ’ ´ -

´

Pour cette partie, seules la diversité taxonomique et la diversité phylogénétique ont été considérées.

5.3.1 Diversités taxonomique et phylogénétique, comparaison avec l'indice de Barker

Faith (1992a) propose que l'unité de base de la conservation biologique soit le caractère (phénotypique, physiologique, etc.). Il mesure alors la diversité phylogénétique comme une estimation de la diversité totale en caractères d'un ensemble d'espèces. Il base son indice, appelé "diversité phylogénétique" (PD) (cf. partie 2.3.3), sur des arbres phylogénétiques en supposant que les longueurs de branches sont proportionnelles à un nombre donné de caractères. En fait Faith (1992b) parle de diversité en caractères mais "richesse en caractères" serait plus juste car le terme "diversité en catégories" est généralement réservé aux indices qui prennent en compte des fréquences ou des abondances. Or dans l'indice PD, chaque espèce n'est considérée qu'une seule fois et chaque branche n'est également considérée qu'une seule fois dans le but de compter le nombre absolu (sans données de fréquences) de caractères. Barker (2002) propose alors une méthode pour pouvoir prendre en compte les abondances des espèces dans la mesure de l'indice PD. Cette méthode peut être décrite de la façon suivante :

- A une branche donnée est attribuée une valeur égale au produit de sa longueur et de l'abondance de l'espèce la plus abondante, parmi toutes les espèces qui descendent de cette branche.
- L'indice de Barker est égal à la somme des valeurs obtenues pour toutes les branches.

L'indice de Barker mesure donc la diversité phylogénétique a partir des mêmes données que l'entropie quadratique. Nous allons comparer ces deux indices. Pour des raisons pratiques, nous utiliserons pour cela la diversité taxonomique plutôt que phylogénétique. Les résultats obtenus s'appliquent à la fois à l'ensemble des taxonomies et à l'ensemble des phylogénies ultramétriques.

Les développements phylogénétiques sont très récents. Pour beaucoup d'organismes les phylogénies proposées sont plutôt des cladistiques (description des liens évolutifs entre espèces, mais absence de longueurs de branches) et sont, de plus, souvent très controversées. Des données comportant à la fois des distances phylogénétiques et des estimations d'abondance des espèces sont actuellement rares pour la plupart des organismes. L'arbre idéal représentant la relation entre les espèces serait un cladogramme parfaitement résolu dans lequel les longueurs des branches ont été déterminées par des méthodes moléculaires. Bien que ce soit possible pour certains organismes, pour la grande majorité des taxa une telle information n'est tout simplement pas disponible (Warwick et Clarke 2001). Les classifications linnéennes conventionnelles sont arbitraires (Williams et Humphries 1994). Gayon (1996) note que, dans les premiers volumes d'"Histoire Naturelle", l'objectif principal de Buffon est de réfuter le système de classification de Linné, qu'il présente comme résumé, arbitraire et stérile. Les classifications linnéennes sont pourtant toujours mieux que rien (Humphries *et al.* 1995) et beaucoup de biologistes sont d'accord pour dire qu'elles sont assez réalistes en termes phylogénétiques (Warwick et Clarke 2001).

Pour illustrer cette comparaison de l'indice de Barker et de l'entropie quadratique, seule la taxonomie sera donc utilisée. Etudions la diversité taxonomique de Trichoptères et Coléoptères de la Loire (fleuve) (données de Ivol *et al.* 1997). Dans les arbres taxonomiques, une longueur de branche de une unité sera considérée entre deux niveaux taxonomiques consécutifs (par exemple entre espèce et genre). Dans Pavoine et Dolédec (2005, cf. annexe 2) ces données

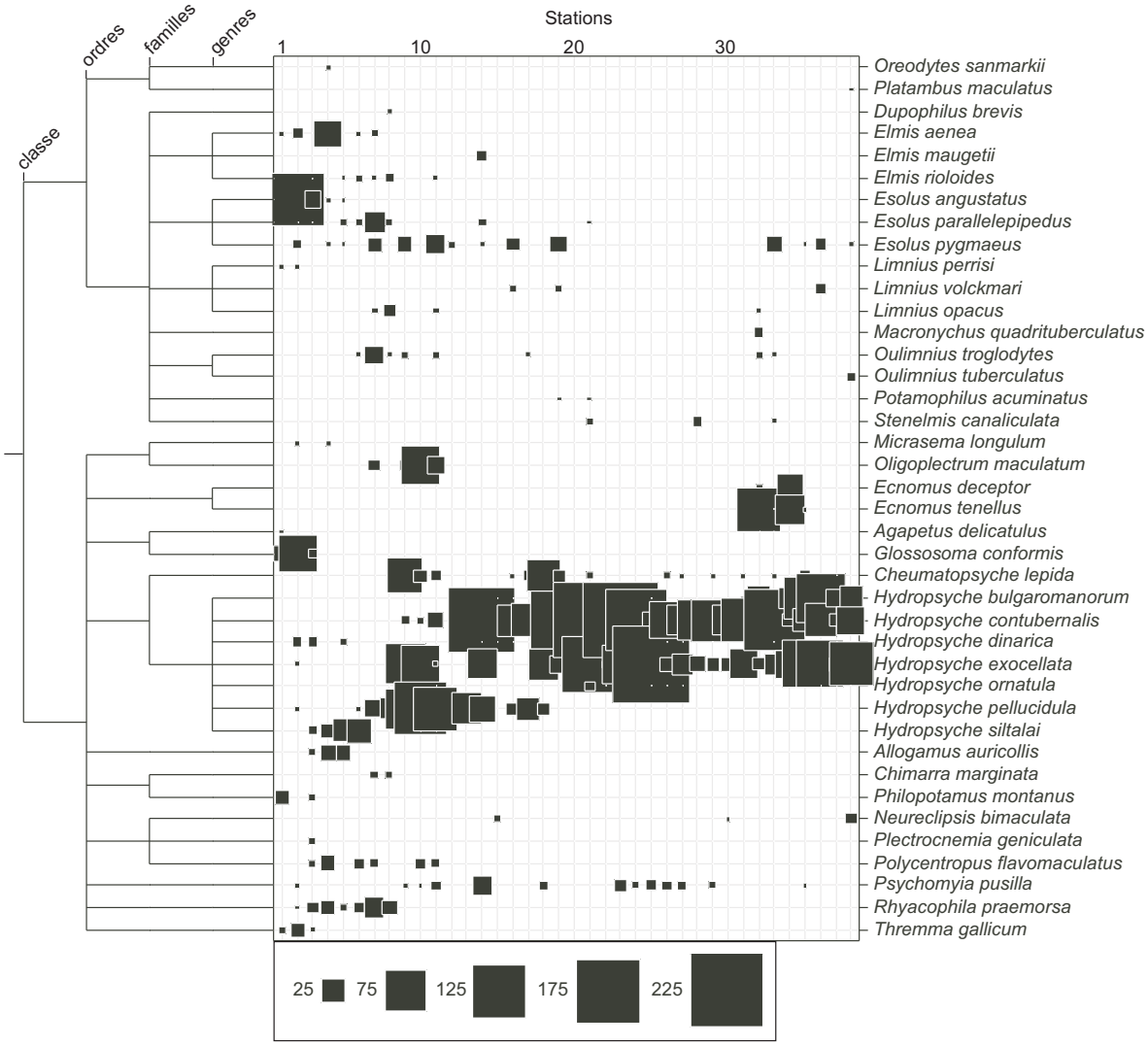


FIG. *52 – Echantillonnage de Trichoptères et Coléoptères dans 38 stations réparties le long de la Loire (fleuve). A gauche est schématisée la taxonomie des 40 espèces observées. Le tableau de droite donne les effectifs observés dans chaque station pour chaque espèce (38 stations réparties d'amont en aval, numérotées à partir de la source). Cette figure a été réalisée avec les fonctions 'as.taxo', 'taxo2phylog', 'table.phylog' du package ade4 de R (Ihaka et Gentleman 1996, Chessel* et al. *2004).*

ont servi à l'étude de la décomposition de la diversité selon l'entropie quadratique avec quatre critères de dissimilarité entre espèces : uniformité (indice de Gini-Simpson), tailles du corps, habitudes alimentaires, et taxonomie. Ici, nous allons nous intéresser à la diversité taxonomique au sein des stations, pour comparer deux indices de diversité : celui de Barker et l'entropie quadratique. Sur l'ensemble des stations, 40 espèces ont été observées. Les données sont résumées dans la figure 52. Les stations sont numérotées à partir de la source.

Le fonctionnement de l'indice de Barker est explicité dans la figure 53 sur la première station. Cette première station contient trois espèces de Coléoptères et quatre espèces de Tri-

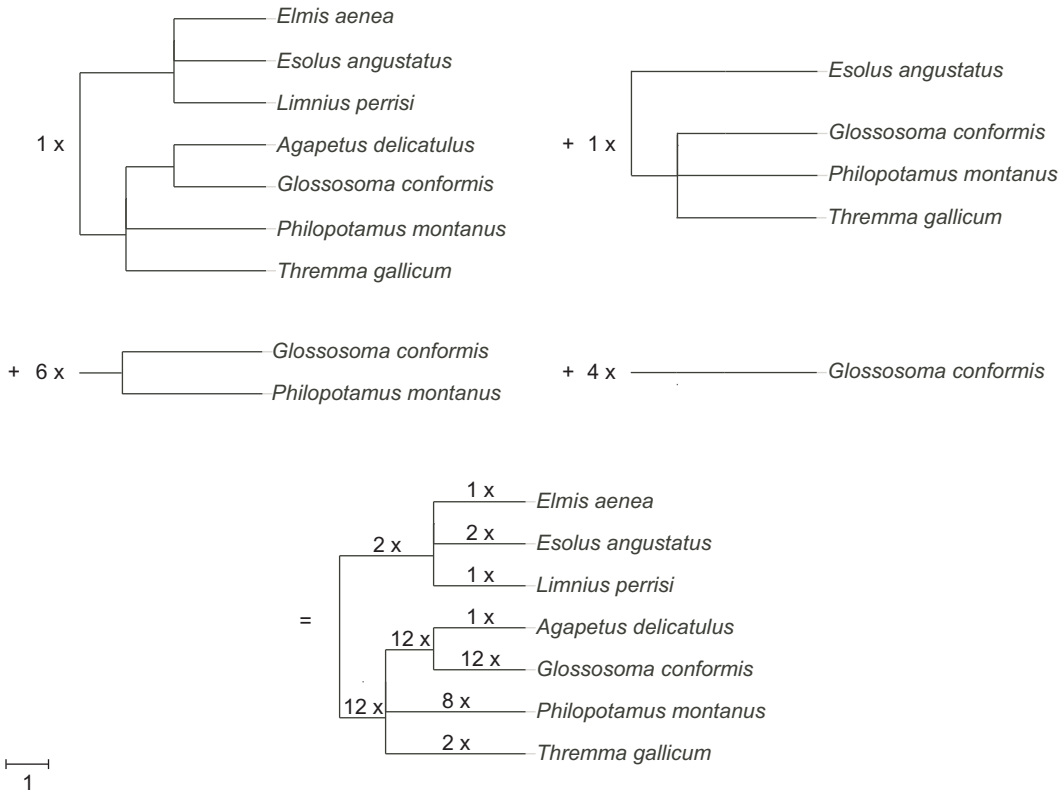


FIG. *53 – Calcul de l'indice de Barker, exemple de la station 1. La valeur de l'indice de Barker pour la station #1 est égale à la somme pondérée des valeurs de PD des quatre premiers arbres. Elle est également égale à la somme pondérée des longueurs de branches du cinquième arbre. En considérant que la longueur de la branche séparant deux niveaux taxonomiques juxtaposés (par exemple espèce et genre) est égale à 1, on trouve que la valeur de l'indice de Barker pour la station #1 est de 92.*

choptères :

Nom	Abondance
Coleoptères	
Elmis aenea	1
Esolus angustatus	2
Limnius perrisi	1
Tricoptères	
Agapetus delicatulus	1
Glossosoma conformis	12
Philopotamus montanus	8
Thremma gallicum	2

Les valeurs de l'indice de Barker pour les 38 stations sont données dans la figure 54a. Aucune tendance ne peut être dégagée des changements de cette valeur le long de la Loire. L'indice

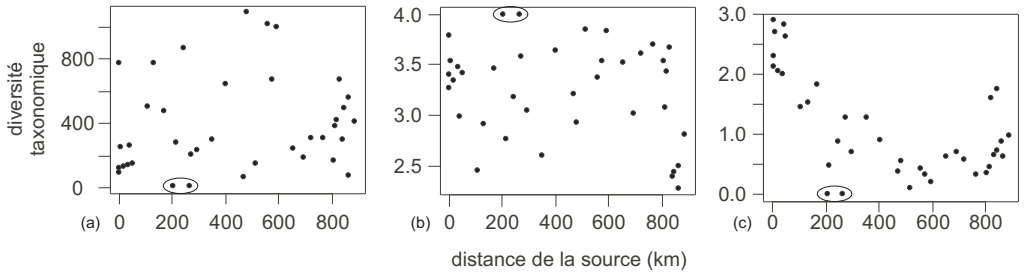


FIG. *54 – Diversité taxonomique au sein des stations le long de la Loire : (a) indice de Barker, (b) indice de Barker divisé par le nombre d'individus échantillonnés dans chaque station (c) entropie quadratique. Les deux stations encerclées sont situées immédiatement en aval d'un barrage.*

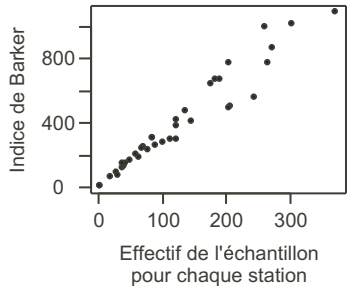


FIG. *55 – Valeurs observées de l'indice de Barker en fonction du nombre d'individus échantillonnés dans chaque station.*

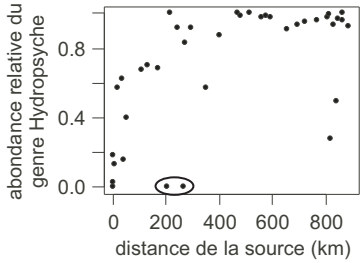


FIG. *56 – Augmentation de la dominance du genre Hydropsyche de l'amont à l'aval. Les deux stations encerclées se situent immédiatement en aval d'un barrage.*

de Barker est en fait fortement lié au nombre d'individus échantillonnés (Fig. 55). Pour éliminer cette dépendance, l'indice peut être divisé par l'effectif de l'échantillon. Mais là encore, aucune tendance n'est visible (Fig. 54b). Les résultats sont même assez surprenants puisque la diversité maximale serait obtenue pour les deux stations (#12 et #15) situées immédiatement en aval d'un barrage. Or il s'agit des deux stations les plus pauvres : dans la station #12 seuls deux individus de l'espèce *Esolus pygmaeus* ont été observés ; et dans la station #15, deux individus de l'espèce *Neureclipsis bimaculata* seulement. Au contraire, La mesure de la diversité taxonomique par l'entropie quadratique révèle une diminution de la diversité taxonomique de l'amont

(station #1) jusqu'à environ 600km (station #25) de la source (Fig. 54c). Cette diminution est due en grande partie à une forte augmentation de la dominance du genre Hydropsyche (Fig. 52 et 56). Les valeurs de l'entropie quadratique reflètent assez bien la diversité taxonomique contenue dans les données échantillonnées. La valeur la plus grande de l'entropie quadratique (2.90, ce qui signifie une différence moyenne entre individus de 2.90 niveaux hiérarchiques) est obtenue pour la station #3 à trois kilomètres seulement de la source. Dans cette station, une espèce de Coléoptères et neuf espèces de Trichoptères ont été observées :

Nom	Abondance
Coléoptères	
Esolus angustatus	12
Trichoptères	
Allogamus auricollis	2
Glossosoma conformis	4
Hydropsyche dinarica	4
Hydropsyche siltalai	3
Philopotamus montanus	2
Plectrocnemia geniculata	2
Polycentropus flavomaculatus	2
Rhyacophila praemorsa	6
Thremma gallicum	1

En mettant de côté les deux stations immédiatement en aval d'un barrage, la station ayant la plus faible diversité taxonomique selon l'entropie quadratique et les données échantillonnées est la station #22 qui contient seulement 2 espèces du genre *Hydropsyche* :

Nom	Abondance
Hydropsyche contubernalis	36
Hydropsyche exocellata	2

FIG. *57 – L'indice de Barker donne la même valeur de diversité à deux stations théoriques S1 et S2, la première constituée de 10 individus d'une espèce a, 10 individus d'une espèce b, et 10 d'une espèce c, et la seconde simplement de 30 individus de l'espèce a. Les espèces a, b, et c sont liées par un arbre constitué d'un seul nœud et dont la longueur des branches est L.*

Au moins d'un point de vue théorique, mais aussi d'un point de vue pratique pour les organismes pour lesquels nous connaissons une phylogénie bien justifiée, Faith, en parlant de richesse en caractères, a efficacement résolu le problème de refléter les différences d'impacts

des espèces dans la mesure de biodiversité. L'indice PD qu'il propose a pour but d'estimer cette richesse en caractères. Le souhait de Barker était de pouvoir tenir compte des abondances des espèces. Il s'agit alors de passer de la richesse en caractères à la diversité en caractères, exactement comme précédemment il y avait eu passage de la richesse en espèces à la diversité en espèces par les indices de Gini-Simpson et Shannon par exemple. Ces indices ne peuvent être utilisés au niveau des caractères puisqu'il est impossible de connaître ou d'identifier tous les caractères d'un ensemble d'espèces. Utiliser une mesure de diversité sur un arbre phénotypique, phylogénétique ou taxonomique va dans le sens d'essayer de mesurer la diversité en caractères. Barker a échoué dans cette démarche (Fig. 57). Par contre l'entropie quadratique reflète assez bien à la fois la répartition des abondances entre espèces et les différences phylogénétiques ou, ici, taxonomiques.

5.3.2 Mesurer l'originalité d'une espèce par l'entropie quadratique

Un des problèmes rencontrés en biologie de la conservation est de donner une importance, une valeur relative, à une espèce pour définir des priorités de conservation.

Le fait de prendre en compte les fréquences des espèces pour mesurer la diversité, plutôt que de simplement compter le nombre d'espèces, correspondait déjà à l'attribution d'une valeur à une espèce. Ensuite, l'intervention de facteurs biologiques et écologiques pour classer les espèces apparaît dans la prise en compte des niches, ensemble des ressources que des espèces utilisent. Dans l'hyperespace défini par les ressources disponibles pour une communauté, les espèces peuvent être classées en fonction de la proportion de l'hyperespace qu'elles utilisent. L'importance d'une espèce évoque alors une fonction écologique. Si nous connaissions suffisamment les relations entre niches, une telle ordination serait possible. Comme ce n'est souvent pas le cas, des alternatives sont adoptées telles que par exemple la prise en compte de la productivité. Hurlbert (1971) propose que l'importance d'une espèce soit la somme des changements en productivité qui auraient lieu si cette espèce était éliminée de la communauté. La productivité représente mieux, que la biomasse et le nombre d'espèces, les changements qui apparaissent dans l'abondance relative de différents niveaux trophiques dans les échantillons. Cependant, la productivité étant elle-même difficile à évaluer, d'autres mesures d'importance sont alors utilisées telles que la densité, la biomasse (Wilhm 1968) et la fréquence (Whittaker 1972). C'est ainsi que l'importance ou la valeur d'une espèce dans un écosystème a été, depuis quelques décennies, fréquemment associée à sa fonction dans un écosystème. Notons cependant que ces mesures sont souvent appliquées à la fraction d'une communauté correspondant à un taxon donné (taxocenes (Whittaker 1972)) et non à l'ensemble de la communauté.

Dans un écosystème, les espèces ne sont donc pas perçues comme ayant la même valeur. Des espèces "clés de voûte", aussi qualifiées de "pivots", sont distinguées : la disparition de l'une d'elles entraîne des bouleversements considérables dans une grande partie de la communauté. Pour identifier ces espèces, il faut estimer l'effet de la perte de chacune et donc comprendre le fonctionnement d'un écosystème. Pour cela, une technique consiste à reconstituer la séquence dans laquelle les espèces se sont ajoutées les unes aux autres quand la communauté s'est édifiée. Roman *et al.* (2001) étudient la diversité des plantes dans la partie péruvienne de l'Amazonie. Moins de 1% de cette diversité semble supporter presque tous les frugivores (primates, oiseaux

et rongeurs) pendant trois mois entiers de l'année. Si ces plantes pivots étaient enlevées, les frugivores disparaîtraient, ce qui provoquerait un effet de cascade. D'autres plantes qui dépendent des frugivores pour la dispersion des graines pourraient également être perdues, ce qui tour à tour pourrait causer des effets faisant échos dans tout le réseau trophique. Ainsi l'élimination d'une espèce clé (ou pivot) pourrait engendrer l'effondrement d'un écosystème entier.

Pour Erwin (1991), les activités de conservation doivent prendre pour cible des zones géographiques contenant une forte concentration de lignées qui évoluent actuellement (à l'échelle des temps géologiques), et les espèces endémiques trouvées dans des zones géographiques restreintes seraient des reliques vouées à une extinction naturelle. Les sauvegarder seraient comme

> "saving living fossils, something of human interest, but perhaps not beneficial to the protection of evolutionary processes and environmental systems that will generate future biodiversity."

Pour lui, le but de la stratégie de conservation doit être de sauvegarder le maximum de biodiversité future et de préserver les espèces contemporaines qui ont un intérêt pour l'homme.

Cette notion d'intérêt pour l'homme a abouti à définir une autre classification des espèces, en référence à l'espèce humaine. Elle est déterminée par le concept de "valeur d'option" d'une espèce, qui a été créé pour justifier économiquement la conservation comme une forme d'assurance, afin de conserver des organismes qui pourront plus tard être utiles à l'homme. La valeur d'option est donc une expression surtout utilisée par les économistes mais de plus en plus présente dans la littérature écologique. Trois critères sont considérés : l'utilité, le caractère plaisant (esthétique), et la valeur morale.

L'utilité d'une espèce peut être définie en écologie comme la fonction de cette espèce dans les écosystèmes dans lesquels elle vit. Ici, le terme "utilité" fait plutôt référence à une utilisation directe par l'homme. La fonction écologique permettrait de rajouter son utilité indirecte pour l'homme à travers sa participation à la pérennité d'écosystèmes et donc des composants de ces écosystèmes. Une des principales justifications des politiques de conservation est l'utilité pharmaceutique d'espèces connues et l'utilité pharmaceutique potentielle d'espèces inconnues, notamment des régions tropicales.

Le caractère plaisant d'une espèce a induit l'identification d'"espèces-parapluies". Elles représentent des espèces vedettes comme *Ailuropoda melanoleuca* le panda géant, *Gorilla gorilla* le gorille et *Aquila heliaca* l'aigle impérial, ou des espèces relativement bien connues, comme l'ensemble des mammifères et beaucoup d'angiospermes, qui si elles sont préservées permettront la sauvegarde des espèces qui les entourent. Beaucoup d'espèces d'insectes, par contre, sont inconnues. Pour ces insectes, les plantes ont le rôle de parapluies : si nous devions perdre la moitié des espèces de plantes endémiques, nous pourrions bien perdre aussi une grande proportion, peut-être similaire, d'espèces d'insectes (Myers *et al.* 2000).

La notion de valeur morale d'une espèce est discutable et certains affirment que toutes les espèces ont de ce point de vue la même valeur.

La principale critique qui doit être faite à ces valeurs d'option est que notre ignorance du futur rend impossible leur évaluation. Pour Erwin, les espèces endémiques et phylogénétiquement isolées sont des reliques et les espèces évoluant fortement actuellement seront à l'origine

de la biodiversité future. Mais le problème est beaucoup plus compliqué. L'histoire de la vie, qui se déroule sur une longue échelle de temps, a connu des phénomènes de radiations massives à partir de groupes minoritaires, suite à des périodes courtes de grandes extinctions (Cooper et Fortey 1998). La diversification s'est faite de façon discontinue - "conclusion qu'il n'était pas possible d'avancer à partir de principes évolutionnistes premiers" (Barbault 1997) - montrant combien il est difficile de prédire quels organismes vont gagner sur le long-terme (Mace *et al.* 2003). Finalement, faut-il ou ne faut-il pas attribuer une valeur à une espèce dans une optique de conservation ?

Faith résout efficacement le problème en proposant de donner des valeurs égales à tous les caractères des espèces, ce qui oblige à donner des valeurs différentes aux espèces. Selon Faith (1992a, 1994b) la diversité en caractères dans un sous-ensemble d'espèces fournit une valeur d'option non seulement parce qu'elle assure qu'un ou plus des membres du sous-ensemble peuvent s'adapter à des conditions changeantes, mais aussi parce que la société peut être amenée à bénéficier des caractères de ces espèces en réponse aux besoins futurs. Williams et Humphries (1996) complexifient le problème en répliquant que seuls les caractères qui s'expriment dans le phénotype d'un organisme ou dans celui de sa progéniture pourront avoir de la valeur en terme de capacité d'adaptation. Ils distinguent trois grandes classes de caractères : génétiques, phénotypiques, fonctionnels. La diversité génétique est en pratique souvent reliée aux cultures sélectives de plantes et aux prospections pharmaceutiques. Donc ce sont en fait les produits phénotypiques des gènes, incluant les molécules, qui ont été directement valués. La diversité des caractères phénotypiques, et en particulier morphologiques, est celle dont la valeur est la plus directement perçue. La diversité des caractères fonctionnels est celle qui assure le maintien de l'intégrité des écosystèmes.

Que l'on souhaite considérer la totalité ou une partie des caractères, il reste une contrainte majeure : les caractères ne peuvent ni être comptés ni être exhaustivement connus. Une solution proposée par Faith (1992a) est d'utiliser les phylogénies. Prédire la distribution des caractères entre organismes requiert alors non-seulement une estimation fortement justifiée de l'arbre phylogénétique, mais aussi la sélection d'un modèle évolutif décrivant la façon dont les caractères changent le long des branches de l'arbre (Williams et Humphries 1996). L'avantage de la mesure PD pour prédire la richesse en caractères pourrait être diminué par l'absence de bonnes estimations des longueurs de branches. L'approche phylogénétique doit supposer à la fois un échantillonnage non-biaisé des caractères, dans une branche et entre les branches, et des changements dans le temps se faisant au même rythme pour les caractères connus et pour les caractères inconnus (Williams et Humphries 1994). Peu de données vérifient actuellement toutes ces propriétés.

Lorsque des informations sur les caractères peuvent être obtenues, des questions se posent sur la façon de traiter les homoplasies (occurrences de caractères similaires chez deux taxa par évolution convergente plutôt que par ascendance commune) par rapport aux vraies homologies (ressemblances entre deux taxa dues à une ascendance commune) (Faith 1996). Les caractères obtenus par convergence ne peuvent intervenir dans la prédiction du nombre total de caractères qui séparent deux espèces. En revanche, ils sont importants lorsqu'ils correspondent à des fonctions particulières des espèces dans leurs écosystèmes.

Si les changements de caractères arrivaient en même temps que les spéciations, alors, sur

les arbres ultramétriques en dehors de quelques branches cachées dues à des espèces éteintes, les nombres de nœuds seraient une meilleur estimation des nombres de caractères que les longueurs de branches (Williams *et al.* 1994).

Faith (1992b) s'intéresse à l'unicité d'une espèce, i.e. à tout ce que cette espèce possède d'unique, et donc à tout ce qu'elle peut apporter de nouveau à un ensemble de référence. Pour un sous-ensemble donné de taxa (*e.g.* déjà impliquées dans un programme de protection) il devrait être possible d'identifier le taxon qui une fois ajouté au sous-ensemble, apporterait le plus grand nombre de nouveaux caractères.

L'unicité d'une espèce n'est qu'une partie de son originalité qui peut être définie ainsi

> L'originalité d'une espèce est la rareté moyenne de ses caractères (Pavoine *et al.* 2005a, annexe 4), la rareté étant une fonction décroissante de la fréquence (Patil et Taillie 1982).

Si tous les caractères d'une espèce sont partagés par les autres espèces de l'ensemble considéré, cette espèce est peu originale. Si beaucoup de ses caractères ne sont pas partagés ou sont partagés par peu d'espèces alors son originalité est grande. Comme il est impossible d'avoir accès à l'ensemble des caractères d'une espèce, la mesure de l'originalité, comme celle de la richesse en caractères, nécessite l'intervention de la phylogénie.

Nous avons vu que la distribution de fréquences qui maximise l'entropie quadratique appliquée à des dissimilarités ultramétriques possède des propriétés intéressantes. Les arbres considérés par Faith ne sont généralement pas ultramétriques. Cependant il existe un type d'arbres phylogénétiques qui lui est ultramétrique : celui qui décrit les histoires évolutives des espèces et sur lequel les unités des longueurs de branches sont des millions d'années. Nous avons démontré qu'avec des dissimilarités phylogénétiques exprimées en termes d'histoire évolutive, la distribution de fréquences qui maximise l'entropie quadratique possède des propriétés très intéressantes puisqu'elle reflète les originalités relatives des espèces les unes par rapport aux autres (Pavoine *et al.* 2005a, cf. annexe 4). Nous avions vu dans la partie 2.3 que des indices de diversité ont été développés en attribuant des valeurs aux espèces : l'indice de Vane-Wright *et al.* (1991), celui de May (1990), les indices non-pondéré et pondéré de Nixon et Wheeler (1992), et la mesure de spécificité taxonomique de Ricotta (2004). La possibilité pour ces indices de mesurer de la diversité a été remise en cause parce qu'ils n'attribuent pas la plus grande diversité relative à un ensemble qui contiendrait les espèces les plus divergentes (Solow *et al.* 1993, Humphries et Williams 1994). Cette critique est très justifiée. Ces indices ne mesurent pas de la diversité mais de l'originalité (Pavoine *et al.* 2005a, cf. annexe 4). Je propose alors (Pavoine *et al.* 2005a, cf. annexe 4) de les renommer indices d'originalité d'un ensemble d'espèces. Ils prennent ainsi tout leur sens. Par exemple, dans l'arbre théorique de la figure 58, la région 1 contient des espèces très divergentes entre elles mais proches de beaucoup d'autres. Elle est donc diverse mais peu originale. A l'inverse, la région 2 comprend trois espèces isolées des autres sur l'arbre mais proches les unes des autres. Elle est donc peu diverse mais très originale. Les indices de Vane-Wright *et al.* (1991), May (1990) et Nixon et Wheeler (1992) mesurent donc bien l'originalité des espèces. Cependant, ils ne permettent de travailler que sur les nœuds des arbres sans considération des longueurs de branches. La distribution de fréquences qui maximise l'entropie quadratique appliquée aux distances entre espèces sur l'arbre ultramétrique permet de mesurer

l'originalité en tenant compte des longueurs de branches sur l'arbre.

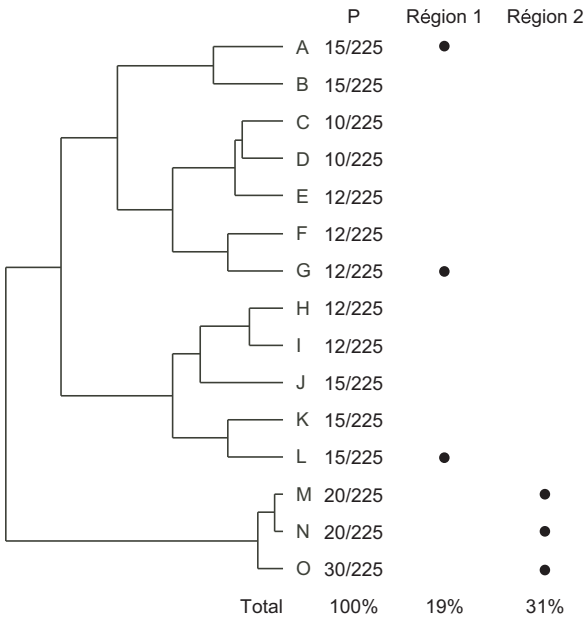


FIG. *58 – Diversité versus originalité. A gauche, une phylogénie théorique entre 15 espèces est représentée. Dans le tableau de droite, la première colonne donne les valeurs P, poids des espèces, dans l'indice de Vane-Wright. Les deux autres colonnes indiquent la composition de deux régions théoriques. La dernière ligne du tableau est la somme des valeurs P, respectivement pour toutes les espèces, pour les espèces de la région 1 et pour celles de la région 2.*

Pour pouvoir être comparés, tous ces indices (l'indice de Vane-Wright *et al.* (1991), celui de May (1990), les indices non-pondéré et pondéré de Nixon et Wheeler (1992)) sont convertis en pourcentages : soit λ_k la mesure de l'originalité selon un des indices, la valeur associée en pourcentage est $\lambda_k / \sum_k \lambda_k$. Illustrons cette comparaison par l'arbre phylogénétique de la famille des Viverridae (Carnivora) (Bininda-Emonds *et al.* 1999). Il s'agit de mammifères mesurant généralement entre 40 et 70cm et appelés couramment genettes ou civettes. Les mesures d'originalité pour ces espèces sont données dans la figure 59. Une étude plus détaillée de ce type de mesures sera trouvée dans Pavoine *et al.* (2005a, cf. annexe 4). En résumé, les indices de Nixon et Wheeler ont la particularité d'attribuer un grand poids aux clades isolés. Ainsi, les classements des espèces déduits de ces indices ordonnent beaucoup plus, les uns par rapport aux autres, les cinq grands clades définis par les cinq branches qui partent de la racine, qu'ils n'ordonnent les espèces au sein des clades. L'allure de la distribution de fréquences obtenue en maximisant l'entropie quadratique est proche de celle de May mais s'en éloigne par la prise en compte des longueurs de branches. Par exemple tout le groupe comprenant six espèces de *Paradocurinae* (*Paradoxurus jerdoni, P. zeylonensis, P. hermaphroditus, Arctictis binturong, Paguma larvata* et *Macrogalidia musschenbroekii*) et l'unique espèce de *Nandininae* (*Nandinia binotata*) possède des poids plus grands selon l'entropie quadratique que selon les autres indices, du fait des grandes longueurs de branches qui séparent les espèces de ce groupe les unes des autres.

Une idée très répandue est que la contribution d'une espèce donnée à la diversité d'un ensemble est définie par la diminution de la diversité due à l'élimination de cette espèce de l'ensemble (Krajewski 1991, Reist-Marti *et al.* 2003) (sans considération de réaction en chaîne pouvant entraîner l'extinction d'autres espèces). Il est bon de noter que les résultats obtenus par ce procédé ont en fait des significations différentes selon l'indice considéré. Ces indices peuvent éventuellement estimer la contribution relative totale d'une espèce à la diversité d'un ensemble dans le seul cas où la somme des contributions de toutes les espèces est égale à la diversité totale de l'ensemble. Prenons l'indice PD de Faith (1992a). Barker propose d'utiliser cet indice et le processus décrit ci-dessus (étude de l'effet, sur la valeur de l'indice, de l'élimination d'une espèce) pour mesurer la contribution de chaque espèce à la diversité. Il calcule donc la contribution d'une espèce à la diversité par la diminution de la valeur prise par PD après élimination de cette espèce. Il multiplie ensuite la valeur obtenue par la probabilité d'extinction de l'espèce pour définir des priorités de conservation. Prenons 12 espèces de félins du Nouveau-Monde (données extraites de Diniz-Filho et Tôrres (2002) à partir de l'arbre phylogénétique de l'ensemble des Carnivores (Bininda-Emonds *et al.* 1999). Considérons trois types de mesures. La première est la contribution d'une espèce selon le calcul de Barker. Pour la seconde mesure, je propose de calculer la diversité perdue pour chacune des combinaisons possibles d'espèces. La contribution d'une espèce est alors la somme des valeurs obtenues pour toutes les combinaisons d'espèces dans lesquelles elle intervient. Cette valeur est ensuite divisée par la somme des valeurs obtenues pour toutes les combinaisons possibles. La troisième mesure est l'indice basé sur l'entropie quadratique ("QE-based index") (Fig. 60). Le résultat est que si l'on classe les espèces selon leurs contributions à la diversité, la première mesure, celle de Barker est très différente des deux autres qui, elles, aboutissent sur ces données exactement au même classement. La correspondance entre la deuxième mesure et celle qui est basée sur l'entropie quadratique n'est pas toujours aussi nette, surtout lorsque plusieurs espèces ont des degrés d'originalité très proches. Mais ce qu'il faut retenir c'est que la contribution d'une espèce à la biodiversité ne peut pas être obtenue par l'observation des effets de sa seule perte. Ces résultats sont très théoriques puisqu'ils ne font appel qu'à une partie de la biodiversité, la diversité phylogénétique. Mais ils montrent justement que l'on ne peut pas juger l'importance des espèces seulement par l'effet d'une perte unique. Et, dans un domaine qui dépasse l'étude des indices de biodiversité, il est évident que le terme d'"espèce clé de voûte" doit sous-entendre aussi "ensemble, ou combinaison, d'espèces clé de voûte".

La même critique s'applique à Hurlbert (1971) qui propose que l'importance d'une espèce soit la somme des changements en productivité qui aurait lieu si cette espèce était éliminée de la communauté.

L'idée de regarder différentes combinaisons d'espèces (ou populations) perdues, plutôt que chaque espèce séparément, est utilisée avec l'indice de Weitzman par Laval *et al.* (2000) et Thaon d'Arnoldi *et al.* (1998). Mais malgré cela, les mesures de conservation restent surtout basées sur l'unicité des espèces et non sur leur originalité.

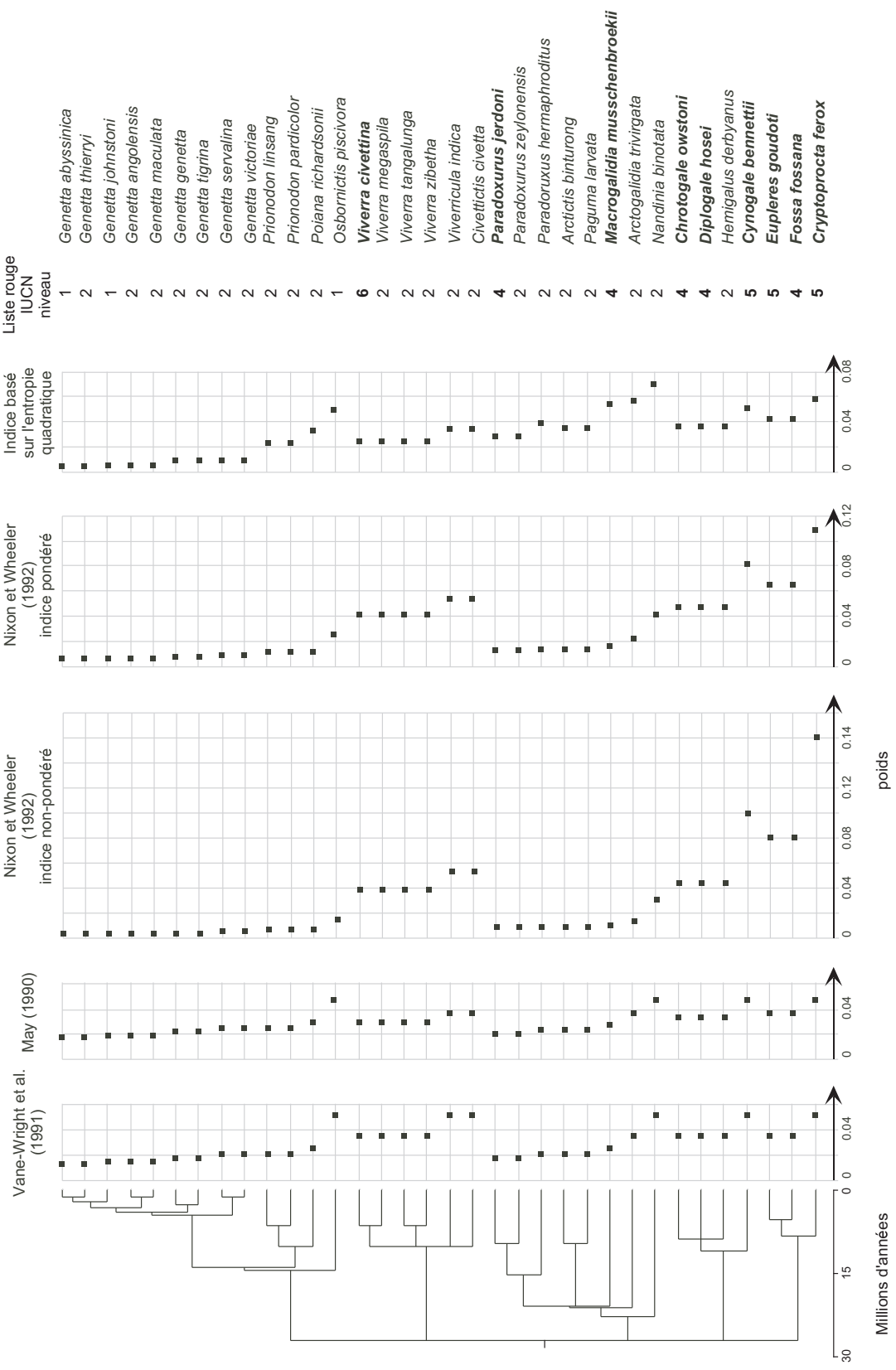
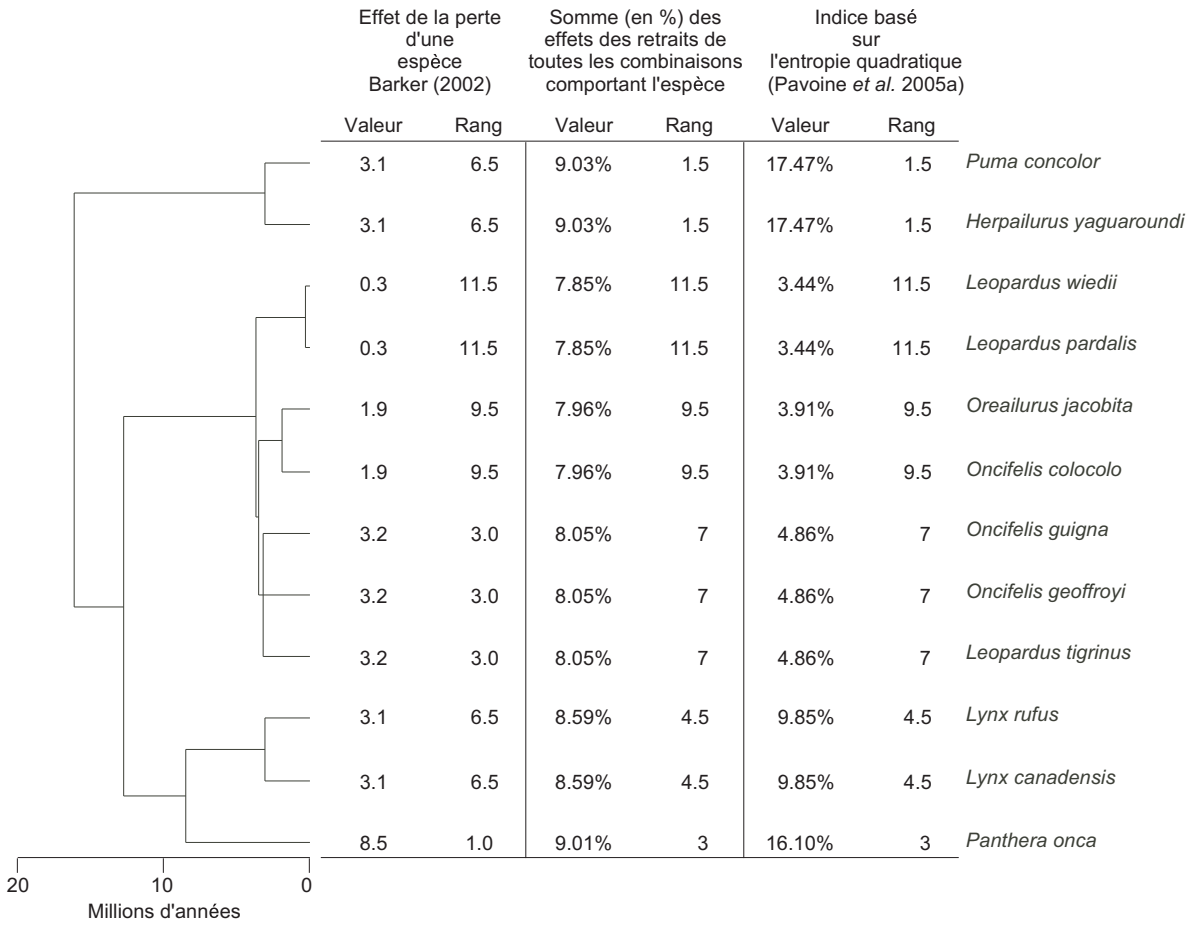


Fig 57 : Mesures d'originalité à partir de l'arbre phylogénétique des Viverridae. Six indices sont utilisés. Les distributions de poids sont données par des diagrammes de Cleveland. La dernière colonne à droite donne le niveau de pérennité des espèces selon la liste rouge de l'IUCN : de 0 non menacée à 7 éteinte à l'état sauvage. Les espèces en caractères gras ont des niveaux supérieurs à 4, elles sont toutes menacées d'extinction.

FIG. *59* –



Effet de la perte d'une espèce Barker (2002)		Somme (en %) des effets des retraits de toutes les combinaisons comportant l'espèce		Indice basé sur l'entropie quadratique (Pavoine *et al.* 2005a)		
Valeur	Rang	Valeur	Rang	Valeur	Rang	
3.1	6.5	9.03%	1.5	17.47%	1.5	*Puma concolor*
3.1	6.5	9.03%	1.5	17.47%	1.5	*Herpailurus yaguaroundi*
0.3	11.5	7.85%	11.5	3.44%	11.5	*Leopardus wiedii*
0.3	11.5	7.85%	11.5	3.44%	11.5	*Leopardus pardalis*
1.9	9.5	7.96%	9.5	3.91%	9.5	*Oreailurus jacobita*
1.9	9.5	7.96%	9.5	3.91%	9.5	*Oncifelis colocolo*
3.2	3.0	8.05%	7	4.86%	7	*Oncifelis guigna*
3.2	3.0	8.05%	7	4.86%	7	*Oncifelis geoffroyi*
3.2	3.0	8.05%	7	4.86%	7	*Leopardus tigrinus*
3.1	6.5	8.59%	4.5	9.85%	4.5	*Lynx rufus*
3.1	6.5	8.59%	4.5	9.85%	4.5	*Lynx canadensis*
8.5	1.0	9.01%	3	16.10%	3	*Panthera onca*

FIG. *60 – Originalité versus unicité d'une espèce. A gauche est donnée la phylogénie des 12 félins considérés. Le tableau fournit ensuite les valeurs attribuées aux espèces, selon l'indice de Barker, par la somme des effets de la perte d'une espèce donnée plus toutes combinaisons possibles des autres espèces, et selon la distribution de fréquences qui maximise l'entropie quadratique appliquée aux dissimilarités phylogénétiques. A partir de ces valeurs, les espèces sont classées dans les colonnes intitulées "Rang". Le premier rang est attribué à la plus grande valeur d'unicité ou d'originalité selon l'indice.*

5.3.3 Préserver diversité et originalité

Originalité et diversité sont deux mesures distinctes et doivent être considérées comme telles.

En biologie de la conservation, souvent ce ne sont pas les abondances des espèces qui sont considérées mais les probabilités d'extinction de ces espèces, déduites entre autres de leurs abondances. On prend alors en compte les probabilités d'extinction pour calculer la diversité attendue ou la perte de diversité attendue (Witting et Loeschke (1995), avec l'indice de Weitzman (Reist-Marti *et al.* 2003)).

Nee et May (1997) s'intéressent aux arbres ultramétriques dont les longueurs de branches sont calculées en temps de divergence. La somme totale des longueurs des branches sur ces arbres est une mesure de la quantité d'histoire évolutive représentée par un ensemble d'espèces.

Nee et May (1997) recherchent quelle quantité de cette histoire évolutive serait sauvée si seulement k espèces sur n étaient préservées. Ils déterminent que la quantité maximale d'histoire évolutive est sauvée en prenant une espèce dans chacun des k plus grands clades. Nee et May comparent cette quantité optimale à la quantité moyenne sauvée en tirant au hasard ces k espèces parmi les n possibles. Ils concluent qu'une grande quantité d'histoire évolutive pourra être préservée par peu d'espèces même si nous laissons les extinctions se produire de façon aléatoire. Nee et May considèrent deux modèles théoriques de croissance des clades : (1) augmentation exponentielle du nombre d'espèces au fil du temps ; (2) clades de tailles constantes (l'apparition d'une nouvelle espèce par spéciation est compensée par l'extinction d'une autre espèce). Leurs simulations suggèrent que, selon le deuxième modèle, dans le cas d'un scénario catastrophe où 5% seulement d'un grand nombre d'espèces survivraient, en moyenne 81% de l'histoire évolutive seraient conservés par ces 5%. Ce résultat se justifie par le fait que pour qu'une branche proche de la racine disparaisse il faut que toutes les espèces sans exception qui en descendent s'éteignent. Mais alors, si seulement 5% des espèces soutiennent en moyenne 81% de l'histoire évolutive, est-il raisonnable d'entamer des stratégies orientées de conservation ?

Reprenons l'arbre évolutif des Viverridae et appliquons les méthodes de Nee et May (Fig. 61). Pour préserver 80% de l'histoire évolutive, il faudrait sauver 53% des espèces selon l'algorithme d'optimalité et 74% en moyenne si les espèces sont tirées aléatoirement. Ces résultats sont très différents de ceux du modèle de Nee et May.

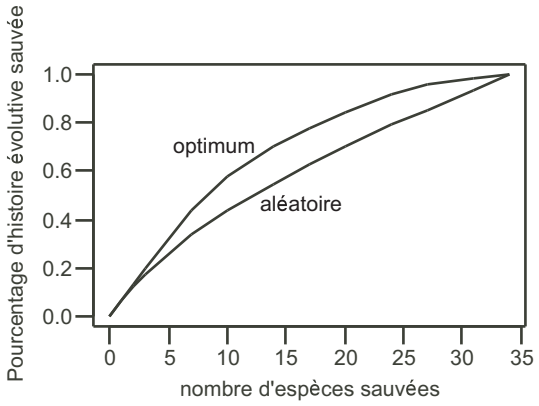


FIG. *61 – Histoire évolutive préservée en fonction du nombre d'espèces sauvées. L'histoire évolutive sauvée est exprimée en pourcentage de l'histoire évolutive totale des 34 espèces de Viverridae. La courbe du haut correspond au schéma d'optimisation, il s'agit donc de la quantité maximale d'histoire évolutive qui peut être sauvée. La courbe du bas est une moyenne obtenue en prenant 1000 échantillons aléatoires pour 0, 1, 2, 3, 7, 10, 14, 17, 20, 24, 27, 31 et 34 espèces.*

L'union mondiale pour la conservation recense dans la liste rouge de l'IUCN les états connus de toutes les espèces vivantes actuellement observées. Ces états sont divisés en huit niveaux : non menacée (rang 0), au statut inconnu (rang 1), moins concernée (rang 2), près d'être menacée (rang 3), vulnérable (rang 4), en danger (rang 5), en grave danger (rang 6), éteinte à l'état sauvage (rang 7). Les espèces classées au rang 1 sont celles pour lesquelles nous avons peu de connaissances. Elles peuvent appartenir réellement à n'importe quel niveau. Cependant

il est fort probable que beaucoup de ces espèces soient à un fort risque d'extinction. Un des critères utilisés pour définir ces niveaux est le nombre d'individus en âge de se reproduire : vulnérable $]2500, 10000]$, en danger d'extinction $]250 - 2500]$, en grave danger d'extinction ≤ 250. L'utilisation générale de telles listes présente quelques limites (Possingham *et al.* 2002). Elles ne peuvent être utilisées seules pour sélectionner des réserves parce qu'un grand nombre d'organismes notamment parmi les invertébrés et les plantes non-vasculaires sont absents de ces listes. Elles ne peuvent non plus être utilisées pour étudier l'évolution de l'état d'une espèce car cette évolution dans la liste dépend essentiellement de l'évolution des connaissances sur cette espèce. Cependant la liste rouge de l'IUCN est utile pour évaluer la validité d'un modèle qui contient des composants de degré d'extinction des espèces (Diniz-Filho 2004). Les niveaux des 34 espèces de Viverridae ont été obtenus pour la liste IUCN à partir de données variant de 1996 à 2004 (Fig. 59). Trois espèces sur trente-quatre ont des statuts inconnus. Le tirage aléatoire de Nee et May suppose que les extinctions se produisent aléatoirement le long des phylogénies. Pour certaines familles, par exemple les Félidés, les valeurs IUCN ne sont effectivement pas phylogénétiquement autocorrélées, c'est-à-dire que deux espèces phylogénétiquement proches n'ont pas particulièrement des valeurs égales ou proches (Diniz-Filho 2004). Cependant ce résultat n'est pas généralisable et, au contraire, souvent les extinctions ne sont pas phylogénétiquement indépendantes. Chez les mammifères et les oiseaux, la probabilité qu'une espèce soit menacée diminue avec le nombre d'espèces dans ses genres, familles et ordres (Purvis *et al.* 2000), alors que ce serait le contraire chez les plantes (Schwartz et Simberloff 2001). Si une espèce a un risque d'extinction, les espèces proches qui lui sont apparentées ont plus de chance qu'en moyenne d'être aussi à risque (Mace *et al.* 2003). Les analyses phylogénétiques récentes montrent clairement que les extinctions actuelles touchent les espèces endémiques de grande taille de corps, qui vivent longtemps, se reproduisent lentement et vivent dans des habitats spécialisés (Mace *et al.* 2003).

Pour l'arbre des Viverridae, il est clair que les valeurs IUCN ne sont pas indépendantes des relations phylogénétiques entre les espèces puisque sur les neuf espèces menacées d'extinction, trois forment un clade (*Eupleres goudoti*, *Fossa fossana* et *Oryptoprocta ferox*) et trois autres appartiennent à un clade contenant quatre espèces au total (*Chrotogale owstoni*, *Diplogale hosei* et *Cynogale bennetti*). Si ces neufs espèces venaient à s'éteindre dans un futur proche, les populations des 25 espèces de Viverridae qui continueraient à évoluer représenteraient environ 74% de l'histoire évolutive totale des 34 espèces actuelles de Viverridae. En moyenne, la perte de 9 espèces provoqueraient la perte de 19% de l'histoire évolutive, donc un maintien de 81% (Fig. 61). Seulement 2% des 1000 tirages aléatoires de 25 espèces effectués contenaient moins d'histoire évolutive que les 25 espèces dont le rang IUCN est inférieur strictement à 4 (Fig 62). Selon le schéma d'optimisation, la perte minimale serait de 7%, elle correspondrait à la perte de toutes les espèces du genre *genetta* sauf une, et à la perte soit de *Eupleres goudoti* soit de *Fossa fossana*. Pourquoi ? parce que seraient perdues des espèces très proches, tout en conservant 14 années d'évolution grâce à une seule espèce du genre Genetta, et 3.1 millions d'années d'évolution représentée soit par *Eupleres goudoti*, soit par *Fossa fossana*. Ces 3.1 millions d'années seraient évidemment menacés puisque *Fossa fossana* est vulnérable et *Eupleres goudoti* en danger d'extinction. L'espèce restante deviendrait alors très originale puisqu'elle n'aurait plus d'espèce sœur parmi les Viverridae. De même, une seule espèce de genette représenterait à elle seule 14 millions d'années d'histoire évolutive si toutes les autres étaient perdues. Même si la

quantité d'histoire évolutive était maximisée par cette situation, l'espérance de cette histoire évolutive (cf. Witting et Loeschke 1995) ne serait sans doute pas optimisée.

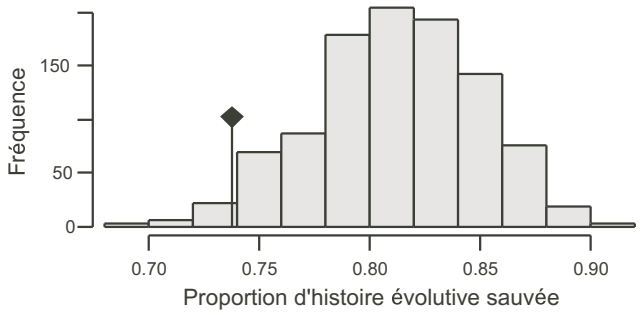


FIG. *62 – Histogramme de la proportion d'histoire évolutive sauvée par 25 espèces tirées au hasard parmi des 34 espèces de Viverridae. 1000 tirages ont été effectués. Le losange noir indique la proportion d'histoire évolutive qui resterait si les neufs espèces actuellement menacées d'extinction venaient à s'éteindre.*

Les neuf espèces menacées d'extinction ont des liens phylogénétiques, en particulier pour les deux groupes d'espèces : *Eupleres goudoti*, *Fossa fossana*, *Cryptoprocta ferox* d'une part, et *Chrotogale owstoni*, *Diplogale hosei*, *Cynogale bennetti* d'autre part. Elles ont aussi en commun leurs habitats et la région dans laquelle elles vivent. Elles sont pour la plupart endémiques, c'est-à-dire qu'elles ne vivent que dans un seul lieu restreint. Les espèces *Eupleres goudoti*, *Fossa fossana* et *Cryptoprocta ferox* sont endémiques sur l'île de Madagascar. Leur perte provoquerait l'élimination de tout un sous-arbre évolutif. *Cynogale bennettii* vit actuellement dans la péninsule malaise, et sur les îles de Bornéo et Sumatra. *Macrogalidia musschenbroekii* évolue sur l'île de Célèbes (Sulawesi). Classée comme vulnérable, elle représente à elle seule 21 millions d'années d'évolution. *Diplogale hosei* a été observée sur l'île de Bornéo, de même que *Cynogale bennetti* dont l'aire de répartition comprend les îles de Bornéo et Sumatra ainsi que la péninsule malaise. *Chrotogale owstoni* vit dans le nord du Vietnam, le nord du Laos et le sud de la Chine. *Paradaxurus jerdoni* et *Viverra civettina* vivent dans les Ghats, escarpement montagneux situés dans la partie ouest de l'Inde. Toutes ces espèces se situent donc dans six (Madagascar, Western Ghats and Sri Lanka, South-Central China, Indo-Burma, Sundaland, Wallacea) des vingt cinq "points chauds de biodiversité", zones présentant à la fois une exceptionnelle concentration d'espèces de plantes endémiques (au moins 0.5% de la richesse spécifique mondiale des plantes) et une exceptionnelle perte de l'habitat (Myers *et al.* 2000). Les régions de Madagascar (comprenant également les îles Maurice, Réunion, Seychelles et Comores) et Sundaland (moitié ouest de l'archipel indo-malais) ont respectivement les première et troisième priorités de conservation. Ces priorités ont été définies selon cinq facteurs : nombre d'espèces endémiques et rapports espèces endémiques / surface pour les plantes d'une part, et les vertébrés (sauf poissons, par manque de données) d'autre part, et perte d'habitat (Myers *et al.* 2000). A Madagascar, il ne reste plus que 9.9% de la forêt primaire. De plus la région de Madagascar se distingue également si l'on considère des taxa plus élevés dans la classification : cette région contient 11 familles endémiques et 310 genres endémiques de plantes, 5 familles

endémiques et 14 genres endémiques de primates, 5 familles endémiques et 35 genres endémiques d'oiseaux. Sur les 12000 espèces de plantes qu'elle contient 9704 sont endémiques et sur les 987 espèces de vertébrés (hors poissons) 771 sont endémiques (Myers *et al.* 2000).

Selon l'IUCN, les causes majeures de leur survie menacée sont la dégradation et la fragmentation de leur habitat par l'homme, en particulier à cause de l'exploitation forestière à Madagascar, de l'introduction d'espèces exogènes prédatrices ou compétitrices pour *Eupleres goudoti* et *Viverra civettina*, de la chasse. *Fossa fossana* est également victime du changement de comportement d'une espèce native devenue compétitive, et *Cryptoprocta ferox* est persécutée par l'homme à cause d'une réputation exagérée de férocité et de caractère destructeur (Nowak 1991).

Les régions contenant à la fois beaucoup d'espèces originales et beaucoup d'espèces menacées d'extinction devraient avoir la priorité dans les stratégies de conservation ; surtout si ce sont les mêmes espèces qui sont à la fois originales et menacées. Pour la famille Viverridae, Madagascar constitue une telle région. L'originalité d'une espèce doit donc être un critère pris en compte dans les méthodes de conservation, en plus des critères tels que son endémisme et sa rareté en terme d'abondance, sa fonction au sein d'un écosystème et ses interactions avec les autres espèces.

Il faut maximiser la diversité tout en minimisant la fragilité des représentants de caractères originaux, en ce sens que si une seule espèce vulnérable est la représentante unique de dizaines de millions d'années d'évolution, les caractères originaux qu'elle porte ont de fortes chances d'être perdus. Cela signifie que, dans les procédures d'optimisation, il est nécessaire de protéger les espèces originales, et aussi d'éviter d'en créer artificiellement d'autres en provoquant l'extinction de toutes leurs espèces sœurs.

5.4 P ´ ´ ´

Lande (1996) donne les qualités requises suivantes pour une mesure de diversité spécifique : être applicable à n'importe quelle communauté indépendamment de la distribution d'abondance des espèces, stricte concavité, faible biais, petite variance d'échantillonnage dans des échantillons de tailles modérées. Ces qualités sont orientées sur la qualité d'un indice en tant qu'estimateur. Lande (1996) montre que, entre la richesse, l'indice de Shannon et l'indice de Gini-Simpson, seul l'indice de Gini-Simpson vérifie toutes ces propriétés. Une de ses généralisations, l'entropie quadratique vérifie ces propriétés, sauf peut-être la dernière. La variance de l'entropie quadratique est asymptotiquement nulle, mais son amplitude pour des échantillons de tailles modérées ne semble pas avoir été, pour l'instant, étudiée.

Les indices de Shannon et Gini-Simpson ont aussi été caractérisés par leurs propriétés en tant que fonctions dont le domaine de départ est l'ensemble $\mathcal{P}$ des fréquences, ce qui donne les représentations suivantes de leur manière de mesurer la diversité.

Soit la fonction H définie sur $\mathcal{P} = \left\{\mathbf{p} = (p_1, ..., p_S)^t, p_k \geq 0, \sum_{k=1}^{S} p_k = 1\right\}$ et telle que
– H est continue en toutes les valeurs p_k,
– Si tous les p_k sont égaux ($p_k = 1/n$) alors H est une fonction monotone croissante de

n. Avec des événements équiprobables, il y a plus de choix ou d'incertitudes avec plus d'événements possibles.
- Si un choix est divisé en deux choix successifs, la valeur de départ de H est la somme pondérée de ses valeurs individuelles.

Shannon (1948) affirme que la seule fonction H satisfaisant ces trois propriétés est de la forme $H(\mathbf{p}) = -K \sum_k p_k \ln(p_k)$, où K est une constante représentant un choix d'unité de mesure (Washington 1984).

Pielou (1975) ajuste légèrement ces trois propriétés (partie 3.2.2) sous la forme des trois postulats suivants. Soit H définie sur $\mathcal{P}$ et telle que

1. H prend sa valeur maximale pour distribution uniforme.
2. Entre deux distributions uniformes, la première de longueur S et la deuxième de longueur $S + 1$, H attribue une valeur plus grande à la deuxième distribution.
3. H peut être décomposée selon deux classifications croisées des catégories (cf. partie 3.2.2).

La seule fonction définie sur $\mathcal{P}$ vérifiant ces propriétés, est de la forme $H(\mathbf{p}) = -C \sum_k p_k \ln(p_k)$, $C > 0$ (Pielou 1975).

Rao (1982c) donne les trois postulats qui caractérisent l'indice de Gini-Simpson. Soit $H(\mathbf{p})$ une mesure de diversité, où $\mathbf{p} \in \mathcal{P}$,

1. $H(\mathbf{p})$ est symétrique dans p_1, ..., p_S et atteint son extremum lorsque $\mathbf{p} = \mathbf{e} = \left(\frac{1}{S}, ..., \frac{1}{S}\right)$.
2. $H(\mathbf{p})$ fonction de p_1, ..., p_{S-1} en substituant p_S par $1 - p_1 - ... - p_{S-1}$, admet des dérivées partielles jusqu'à l'ordre 2, et la matrice des dérivées secondes

$$H''(\mathbf{p}) = \left[\frac{\partial^2 H}{\partial p_k \partial p_l}\right]$$

   est continue et non nulle lorsque $\mathbf{p} = \mathbf{e}$.
3. $H\left(\frac{1}{2}(\mathbf{p} + \mathbf{e})\right) - \frac{1}{2}(H(\mathbf{p}) + H(\mathbf{e})) = c(H(\mathbf{e}) - H(\mathbf{p}))$, où c est une constante.

Le postulat (1) est selon Rao un postulat "naturel", on le retrouve ci-dessus dans la caractérisation de l'indice de Shannon. Le postulat (2) assure que la mesure de diversité soit sensible à l'éloignement de $\mathbf{p}$ par rapport à $\mathbf{e}$ où elle atteint son maximum. Le postulat (3) introduit deux façons de calculer la distance entre $\mathbf{p}$ et $\mathbf{e}$. Il demande l'égalité de ces deux distances à une constante près. Sous ces trois postulats, $H(\mathbf{p})$ est de la forme

$$H(\mathbf{p}) = a\left(1 - \sum_{k=1}^{S} p_k^2\right) + b,$$

où $a > 0$ et b sont deux constantes (Rao 1982c). En choisissant $a = 1$ et $b = 0$, on obtient l'indice de Gini-Simpson.

Si au lieu d'exiger que l'entropie quadratique soit une mesure parfaite de diversité selon l'axiomatisation de Rao, i.e., qu'elle soit complètement concave, les deux propriétés suivantes seulement sont demandées

1. $H_{\mathbf{D}}(\mathbf{p})$ doit valoir 0, minimum, quand toutes les entités appartiennent à la même catégorie ;
2. $H_{\mathbf{D}}(\mathbf{p})$ doit être une fonction concave sur l'ensemble $\mathcal{P}$;

alors les deux seules contraintes sur le choix de la matrice **D** sont

$$d_{11} = ... = d_{SS}$$

et la matrice $(S-1 \times S-1)$

$$\{d_{kS} + d_{lS} - d_{kl} - d_{SS}\} \ (k, l = 1, ..., S-1)$$

est définie négative. La matrice **D**, dans ce cas, n'est plus une matrice de dissimilarités et les valeurs d_{kk} peuvent être non nulles. Sous ces contraintes, la mesure $H_{\mathbf{D}}(\mathbf{p})$ est appelée fonction d'entropie quadratique généralisée. D'après Rao (1982c), cette fonction généralisée peut être caractérisée par les mêmes postulats que l'indice de Gini-Simpson, en remplaçant **e** par un point fixé $\mathbf{q} \in \mathcal{P}$ dans le postulat 2.

Les trois propriétés intéressantes que possède l'entropie quadratique utilisée avec des dissimilarités ultramétriques sont les suivantes :

1. La valeur maximale
$$\max_{\mathbf{p}} (\mathbf{p}^t \mathbf{D} \mathbf{p})$$
est obtenue lorsque les espèces sont présentes avec des fréquences reflétant leurs originalités relatives (les espèces les plus originales ont les plus grandes fréquences).
2. Soient E un ensemble de S espèces et j une autre espèce ($j \notin E$), soient $\mathbf{p}_E$ une distribution de fréquence de E et $\mathbf{p}_{E \cup j}$ une distribution de fréquence de $E \cup j$, alors
$$\max_{\mathbf{p}_E} (\mathbf{p}_E^t \mathbf{D} \mathbf{p}_E) < \max_{\mathbf{p}_{E \cup j}} \left(\mathbf{p}_{E \cup j}^t \mathbf{D} \mathbf{p}_{E \cup j}\right)$$
3. $\mathbf{p}^t \mathbf{D} \mathbf{p}$ est complètement concave, donc "décomposable".

On peut rajouter également que l'entropie quadratique est une mesure d'inertie associée à des méthodes statistiques descriptives d'ordination. Sa décomposition peut permettre de tester l'existence de différences entre plusieurs sites dans leur composition spécifique au regard des liens phylogénétiques entre espèces.

5.5 P

Appliquée à des dissimilarités ultramétriques, l'entropie quadratique est donc un indice de diversité qui possède de bonnes propriétés, aussi bien du point de vue de la mesure et de la précision de l'estimation d'un aspect de la diversité d'une collection que du point de vue de son appartenance à un schéma fondamental : l'axiomatisation de Rao. Il possède un autre avantage : il est applicable à n'importe quelle nature de données et à n'importe quelle discipline scientifique. Il peut donc servir à l'étude de nombreux aspects de la biodiversité qui a différents objectifs, la biologie de la conservation n'en étant qu'un parmi beaucoup d'autres. Cet indice

est peut-être le seul actuellement qui tienne compte, avec autant de qualités, à la fois de mesures de dissimilarité et de mesures d'abondance dans l'évaluation de la diversité.

Cependant, nous avons pointé du doigt une limite de cet indice en tant que mesure de biodiversité : sa maximisation peut diminuer la richesse. Nous avons donc remis en question l'interprétabilité et la pertinence de cet indice pour la mesure de la biodiversité. Fort heureusement, nous avons vu que l'entropie quadratique est très adaptée aux mesures de diversités taxonomique et phylogénétique, deux critères qui s'adaptent parfaitement à la contrainte mathématique d'obtention d'une matrice de dissimilarités ultramétrique. Néanmoins, si l'on souhaite regarder d'autres critères de diversité, à tout type de données ne peut être associé une matrice ultramétrique, et nous avons vu que la transformation d'une matrice de dissimilarités quelconque en une matrice ultramétrique peut modifier fortement les données d'origine. L'entropie quadratique ne peut donc s'appliquer, dans le cadre de la mesure de la biodiversité, à tout type de données.

En conclusion, l'entropie quadratique a incontestablement une place centrale dans les mesures de variabilité. Son étude a permis d'ouvrir des discussions dans la littérature et au cours de cette thèse sur l'importance des abondances et des différences entre espèces, sur leurs intérêts respectifs pour aider à comprendre l'évolution actuelle de la biodiversité des communautés et sur la nécessité ou non de les inclure ensemble dans une étude étant donnée son objectif. La place de l'entropie quadratique dans la mesure de la biodiversité doit être discutée et offre la possibilité de débats sur ce que l'on veut réellement désigner, évaluer en parlant de biodiversité.

L'intérêt d'avoir affaire à un indice qui dépend d'une matrice de dissimilarités est qu'il nous permet d'affirmer que les catégories choisies, et surtout les espèces, ne sont pas interchangeables. La grande majorité des indices actuellement sont basés sur la propriété communément admise, que la diversité est maximale pour une distribution uniforme des fréquences des espèces. Cette propriété ne peut être acceptée que faute de mieux, par manque de connaissances dans les cas où nous ne pouvons caractériser précisément les espèces autrement que par leurs abondances. L'entropie quadratique appliquée à des dissimilarités ultramétriques rompt avec cette traditionnelle propriété et constitue ainsi une grande étape franchie vers ce à quoi nous devons tendre.

Enfin, cet indice m'a permis d'introduire une nouvelle mesure de l'originalité d'une espèce, montrant ainsi l'importance de cette mesure en biologie de la conservation. L'originalité relative d'une espèce n'est certainement pas le seul critère. Son intérêt est relatif au point de vue que l'on adopte pour regarder la biodiversité, disons grossièrement par exemple écologique, économique, philosophique ou sociale, et à la complexité des objets biologiques que l'on manipule. Mais l'originalité d'une espèce en tant que degré d'isolement phylogénétique est certainement l'un des critères à ne pas éluder, les espèces les plus isolées étant les plus susceptibles de posséder des caractères phénotypiques ou moléculaires rares ou même uniques.

Chapitre 6

Conclusion

Avant de conclure, reprenons succinctement les points les plus essentiels de ce travail.

Rao s'inspirant des travaux de Nei développe l'entropie quadratique, fonction mesurant la diversité d'une collection à partir d'un vecteur de fréquences et d'une matrice de dissimilarités entre catégories. Puis, autour de cette fonction, il développe l'axiomatisation des mesures de diversité. Centrée sur l'entropie quadratique et englobant l'indice de Gini-Simpson, cette axiomatisation permet de définir quels indices peuvent être décomposés selon des facteurs hiérarchiques ou croisés. Son élève Nayak étudie alors les propriétés de l'entropie quadratique : un estimateur non biaisé et de variance asymptotiquement nulle peut être obtenu. Des procédures de tests sont ensuite développées : à partir de l'entropie quadratique, peut-on affirmer que les collections étudiées diffèrent ? L'entropie quadratique joue un rôle central dans cette axiomatisation. Elle repose sur des bases statistiques depuis longtemps établies : variance sur variable quantitative, ANOVA ; variance sur variable qualitative, indice de Gini-Simpson, CATANOVA. Si les dissimilarités entre catégories ont des propriétés euclidiennes, elle est décomposable selon n'importe quel nombre de facteurs hiérarchiques et/ou croisés. Elle unifie les concepts de dissimilarité et de diversité. Et enfin elle possède une définition claire : l'entropie quadratique est la dissimilarité attendue entre deux entités tirées au hasard dans une collection.

Nous avons placé l'entropie quadratique dans le cadre des méthodes d'ordination. L'entropie quadratique est alors une mesure d'inertie d'un nuage de points dans un espace euclidien. Nous avons défini, à travers une double analyse en coordonnées principales, des espaces euclidiens dans lesquels la décomposition d'inertie correspond à la décomposition de l'entropie quadratique. Nous relions ainsi les concepts de diversité, inertie, dissimilarité et ordination. La double analyse en coordonnées principales généralise l'analyse en composantes principales inter-classes, l'analyse non-symétrique des correspondances, l'analyse discriminante, l'analyse canonique sur coordonnées principales et l'analyse canonique des correspondances ou analyse factorielle des correspondances sur variables instrumentales. Nous montrons ainsi que la décomposition de l'entropie quadratique est en filigrane dans ces cinq analyses.

Toutes ces propriétés font que l'entropie quadratique s'impose naturellement comme un moyen efficace de tenir compte à la fois des fréquences des catégories et des dissimilarités entre les catégories pour mesurer et décrire la biodiversité.

L'entropie quadratique possède une définition unique, et pourtant elle se comporte différemment à son maximum en fonction des propriétés mathématiques de la matrice de dissimilarités entre catégories. Le choix de cette matrice, qui se fait selon un intérêt biologique, a donc des conséquences sur ce qui est finalement mesuré. Pourquoi est-ce si important ? L'entropie quadratique n'est pas une mesure corrélée à la richesse, et c'est ce que nous recherchions. Sa valeur dépend de trois éléments : de la richesse d'une part mais aussi des fréquences des catégories et des dissimilarités entre les catégories. Mais ce n'est pas suffisant. La critique faite par Shimatani et Izsák et Szeidl à l'entropie quadratique d'être maximisée en réduisant considérablement la richesse ne concerne que son utilisation en tant que "mesure de diversité". Il a été dit qu'une mesure de diversité doit être maximale pour une richesse maximisée et pour une distribution uniforme des fréquences entre les composants de cette richesse. Si aucune caractérisation des catégories n'est faite et si les catégories sont effectivement interchangeables alors cette affirma-

tion est raisonnable. Dans les autres cas, la distribution uniforme ne devrait pas être exigée.

En effet, diversité signifie variété. Pour maximiser l'entropie quadratique, il faut augmenter la dissimilarité entre deux entités tirées au hasard. Reprenons les interprétations données aux indices de Gini-Simpson et de Shannon. L'indice de Gini-Simpson est la probabilité de tirer deux entités appartenant à deux catégories différentes. L'indice de Shannon est interprété comme la perte d'information due au retrait d'une entité. L'entropie quadratique généralise le concept de l'indice de Gini-Simpson mais pas celui de Shannon. Ce concept n'est pas repris dans la forme la plus générale de l'entropie quadratique puisque, l'entropie quadratique peut être maximum avec un nombre réduit de catégories alors que plus il existe de catégories dans une collection et plus ces catégories sont différentes, plus la perte d'une entité représente une perte d'information sur la collection.

Lorsque l'entropie quadratique est appliquée à des matrices de dissimilarités ultramétriques, en particulier phylogénétique, la critique de Shimatani, Izsák et Szeidl tombe. L'entropie quadratique alors est maximisée en même temps que la richesse. La distribution de fréquences n'est pas uniforme, ce qui est tout à fait logique puisqu'elle reflète les originalités relatives des catégories. La catégorie la plus distincte, la plus isolée sur un arbre ultramétrique doit avoir la plus grande fréquence pour maximiser l'entropie quadratique. Cette distribution de fréquences est unique, et cohérente avec les indices d'originalité de Vane-Wright *et al.* (1991) et May (1990). Avec les dissimilarités ultramétriques, l'entropie quadratique est donc un indice de diversité aux propriétés très intéressantes : son biais dépendant d'un paramètre connu (la taille de l'échantillon), un estimateur de l'entropie quadratique, non-biaisé et de variance asymptotiquement nulle, existe ; l'entropie quadratique est maximisée en même temps que la richesse pour une distribution de fréquences qui reflète les originalités relatives des catégories ; sa valeur maximale augmente par l'ajout d'une catégorie dans l'ensemble des catégories possibles ; et enfin l'entropie quadratique est décomposable selon un nombre quelconque de facteurs hiérarchiques (*e.g.* stations $\subset$ régions) et croisés (*e.g.* sites $\times$ dates).

L'axiomatisation de Rao est un schéma fondamental qui inclut tous les indices de diversité définis sur l'ensemble $\mathcal{P}$ des distributions de fréquences, l'ensemble des distributions de probabilité discrètes et finies, et même aussi des indices définis sur des distributions de probabilités continues. Dans ce schéma général, celui de la décomposition de l'entropie quadratique inclut les décompositions des indices de variances sur variables quantitatives et qualitatives. Et dans le schéma de la décomposition de l'entropie quadratique, l'entropie quadratique appliquée à des dissimilarités ultramétriques généralise l'indice de Gini-Simpson pour la mesure de la diversité à différentes échelles.

L'axiomatisation de Rao est un schéma fondamental non seulement parce qu'elle inclut beaucoup d'indices mais aussi parce qu'elle est applicable à n'importe quelle nature de données pourvu que la structure de ces données comprenne une ou plusieurs distributions de fréquences ou de probabilités. Pour la biodiversité, la nature de ces données comprend le niveau dans l'échelle bio-écologique (allant des gènes jusqu'aux régions, ou aires biogéographiques), et le critère (morphologie, physiologie, etc) considéré.

Une seule étude ne peut révéler toute la diversité d'une région. Les indices traditionnels de biodiversité (richesse, indice de Gini-Simpson, indice de Shannon, etc.) sont basés sur deux

hypothèses implicites (Hendrickson et Ehrlich 1971) : (1) tous les individus d'une même espèce sont identiques, (2) toutes les paires d'espèces distinctes sont également différentes. Or dans bien des groupes, les différences entre individus de la même espèce sont essentielles pour le maintien de cette espèce dans la nature, ce qui contredit l'hypothèse 1. Et, du point de vue d'un systématicien, l'hypothèse 2 est clairement fausse (Cousins 1991).

Nous avons vu que l'entropie quadratique possède des qualités indéniables et utiles dans son rôle d'estimateur de la diversité. Elle permet d'éviter la deuxième hypothèse. Cependant, des limites fortes ont été rajoutées à son utilisation. La propriété ultramétrique des dissimilarités contraint considérablement son analyse, notamment en génétique où les modèles basés sur l'horloge moléculaire, qui correspond à des dissimilarités ultramétriques, sont souvent réfutés et réorientés vers des modèles avec des taux de mutations différents selon les espèces, qui eux ne correspondent pas à des dissimilarités ultramétriques. Mais l'essentiel est que l'axiomatisation de Rao ouvre la voie pour d'autres recherches et montre combien il est important de considérer que les catégories, les séquences, les populations, les espèces étudiées ne sont pas interchangeables et sont plus ou moins ressemblantes, dissemblables, connectées.

Un indice unifiant les concepts de dissimilarité et de diversité, comme celui de Rao ou un autre, est essentiel puisqu'il affiche clairement que les catégories définies ne sont pas les plus petites unités uniformes que l'on puisse identifier et donc qu'elles ne sont pas interchangeables. En génétique, un tel indice se place au niveau des différences entre individus, et même entre les deux chromosomes de leur génome diploïde, en restant généralement à l'échelle des populations. En écologie, il permet une évaluation à grande échelle de la diversité morphologique, fonctionnelle ou phylogénétique d'espèces. En écologie du paysage, il permet l'évaluation d'un aspect choisi de la diversité des régions.

Un des objectifs de la biologie de la conservation est de définir une unité à partir de laquelle on puisse développer des stratégies pour la préservation de la biodiversité. Et en général, l'unité choisie est l'espèce. Dans la pratique, c'est le nombre d'espèces qui est utilisé pour mesurer la biodiversité surtout parce qu'il s'agit d'une unité plus facilement mesurable et visible que le nombre total d'allèles, de caractères ou d'interactions entre individus. Mais nous avons vu que le terme d'espèce est une notion dont les délimitations restent floues. Daugherty *et al.* (1990) découvrent que toutes les populations de *Sphenodon* tuatara n'appartiennent pas à la même espèce. La tuatara était considérée comme l'unique espèce représentant toute une longue histoire évolutive. La découverte d'une seconde espèce diminua son originalité. Ainsi Crozier (1992) suggère qu'en général, la considération des populations comme unités de base donnera un résultat plus raisonnable que l'unité espèce à travers laquelle les populations sont toutes traitées de façon identique. Tout est dit dans ces derniers mots. Si la diversité est mesurée par le nombre d'espèces, les différences entre populations sont ignorées. Mais si elle est mesurée par le nombre de populations, les différences entre individus d'une même population sont ignorées. La question de l'unité pour la conservation pose le problème de ce qu'il est raisonnable d'étudier à grande échelle avec les moyens actuels. Finalement pourquoi utiliser les espèces plutôt que directement les gènes ou les caractères ? Une autre façon de poser la question serait donc plutôt : jusqu'à quel niveau de précision mesure-t-on la diversité ? au niveau des gènes ? des populations ? des espèces ? La diversité génétique est-elle le reflet de toute la variabilité actuelle et permet-elle des prédictions sur la variabilité future des êtres vivants ? Quelles critères

avons-nous pour mesurer la biodiversité ? les gènes ? la morphologie ? la physiologie ? le comportement ? Tout est emboîté, connecté. Dans quelle mesure pouvons-nous prendre en compte toutes ces informations dans l'évaluation de la diversité ?

La biologie de la conservation affirme clairement son but : sauver le plus de diversité au moindre coût (Simianer 2002, Simianer *et al.* 2003). Le problème premier est de savoir quelle est cette diversité qu'il nous faut préserver. Sur cette question, les points de vue divergent.

Préserver des pools de gènes plutôt que des espèces est plus abstrait et implique des questions scientifiques non encore résolues. Pour certains, préserver des écosystèmes est plus difficile que de préserver des espèces (May 1995). Pour d'autres, la conservation des habitats est plus rentable en terme de résultats par rapport à la somme d'argent investie qu'une approche centrée sur des espèces (Mace *et al.* 2003).

Les connaissances du monde vivant que nous avons sont inégales. De nombreuses méthodologies ont été développées pour que les sélections des zones à préserver soient faites objectivement. La prise en compte des liens phylogénétiques entre espèces a été demandée. Mais si les phylogénies sont disponibles uniquement pour des groupes attractifs, nous n'avons pas avancé (Mace *et al.* 2003).

La préservation de la biodiversité demande la prise en compte des fonctions des espèces et de leurs interactions. Conserver un ensemble d'espèces très divers sans conserver aussi les espèces dont elles dépendent fournira une faible espérance de diversité (Solow et Polasky 1994). L'approche écosystémique est basée sur la reconnaissance que le fonctionnement et la résilience d'un écosystème dépendent d'une relation dynamique au sein des espèces, entre les espèces, et entre les espèces et leur environnement abiotique, et que la conservation de ces interactions et procédés est d'une plus grande signification que la simple protection des espèces (Roman *et al.* 2001).

La biologie de la conservation balance ainsi entre la conservation du profil de la biodiversité ou la préservation des processus qui l'ont générée (Mace *et al.* 2003). Pour Norton (Norton 2001, Faith 2003), il ne faut pas valuer les objets (espèces) mais les processus, les fonctions qui maintiennent la santé des écosystèmes, telles que les provisions d'air pur et d'eau potable.

> "The use of measures other than species counts also forces us to ask what it is that we are trying to preserve." (Agapow *et al.* 2004)

Il est difficile d'être convainquant si on ne sait pas de quoi on parle. Je suis obligée de clore cette thèse par la question que j'ai posée à son début : qu'est-ce que la biodiversité ? Il est souvent dit qu'il n'existe pas actuellement de définition universelle du mot biodiversité.

> "Biodiversity is 'all things to all men'." (Williams *et al.* 1991)

La biodiversité est un terme dont les utilisateurs supposent que tous partagent la même définition intuitive (Williams et Humphries 1994).

> "La « biodiversité » n'est pas un concept, encore moins un paradigme ; c'est une coquille vide où chacun met ce qu'il veut, un « mot de passe »." (Blondel 2000)

Williams *et al.* (1993) commentent la définition de McNeely *et al.* (1990, page 17),

> "Umbrella term for the degree of nature's variety, including both the number and frequency of ecosystems, species, or genes in a given assemblage."

Ils précisent que c'est une définition globale qui ne permet pas d'arriver à une mesure pratique et rigoureuse de la biodiversité. Les méthodes statistiques développées depuis plusieurs années pour mesurer la biodiversité sont tellement différentes entre elles et pourtant toutes rangées sous les termes de "mesure de biodiversité" ou plus souvent "mesure de diversité". En fait chaque mesure décrit un aspect de la variabilité du monde vivant, et en plus chacune à sa façon. Le nom de chaque fonction développée dans le cadre de la mesure de la biodiversité devrait refléter cet aspect et cette façon de mesurer. La biodiversité a été identifiée en suivant l'immense complexité de la totalité de la vie non réductible à une formule mathématique. C'est pourquoi, il ne peut y avoir une unique et objective mesure de biodiversité, mais seulement des mesures appropriées à des objectifs restreints (Norton 1994, Williams *et al.* 1994). Il faut décrire la diversité du vivant de façon différente pour chaque but. Le choix d'un modèle de mesure pour la biodiversité dépend de la valeur qui est importante pour celui qui prend la décision de mesurer une portion choisie de la biodiversité. Le terme biodiversité est donc trop complexe pour qu'il ne soit pas dissocié en plusieurs aspects, et pas seulement en se référant à des échelles biologiques mais aussi en faisant appel à des caractéristiques de ce qui peut être considéré comme de la variété. Par exemple, l'étude de Vane-Wright *et al.* (1991) a stimulé le développement d'un grand nombre de mesures alternatives reliées à la diversité taxonomique (Altschul et Lipman 1990, May 1990, Williams *et al.* 1991, Faith 1992a). Cependant, parce que chacune de ces mesures englobe une notion différente de la diversité, ce qu'est la "diversité taxonomique" reste flou (Faith 1992b). Il faudrait que le nom et la définition de chaque indice explicitent clairement cette facette pour que puissent être connus ce que l'indice mesure, ses avantages et ses limites.

Si un tel soin était pris, chacun de nous saurait beaucoup mieux ce que chaque indice mesure. Par exemple, l'indice Δ^+ de Clarke et Warwick (1998) n'est pas basé sur un vecteur de fréquences, et pourtant il peut être diminué par l'ajout d'une espèce. Si un ensemble contient deux espèces taxonomiques très différentes, la valeur de Δ^+ sur cet ensemble est grande. L'ajout dans cet ensemble d'une espèce intermédiaire dans la taxonomie diminue Δ^+. Comme la variance, Δ^+ mesure un "écart moyen". C'est exactement "l'éloignement taxonomique moyen entre deux espèces d'un assemblage". C'est une dénomination un peu longue mais beaucoup plus juste et explicite que "diversité taxonomique". La variance a un nom bien particulier qui signifie "la somme pondérée des carrés des écarts à la moyenne". C'est aussi la moyenne des carrés des différences entre deux mesures, ce qui se rapproche de la définition de Δ^+. Ce n'est pas un indice de diversité. L'indice de Rao est souvent appelé "diversité quadratique" car le terme "entropie" est plutôt réservé à la thermodynamique pour la mesure de la dégradation de l'énergie d'un système, et à la théorie de la communication pour mesurer l'incertitude de la nature d'un message donné à partir de celui qui le précède. J'ai choisi de ne pas employer ce qualificatif de diversité quadratique, car dans sa forme générale, sans restriction, l'entropie quadratique n'est pas un indice de diversité.

La biodiversité est souvent définie comme la variété de toutes les formes de vie, de l'échelle des gènes jusqu'à celles des espèces et écosystèmes.

> "The nud of the problem of defining biodiversity is that it is hard to exclude anything from a concept that is taken so easily to mean 'everything'." (Faith 2003)

Le problème n'est pas seulement sémantique. Le terme biodiversité est en fait clair. Il englobe tout ce qui est « bio » et divers (Blondel 2000), c'est-à-dire une multitude d'objets et de phénomènes très différents. Par contre les connaissances concrètes que nous avons de la biodiversité, de toute la variété du monde vivant, sont très maigres. Mais le problème est uniquement que dans le cas particulier d'une étude, il est primordial de décrire précisément l'infime portion de l'ensemble de la biodiversité que l'on souhaite réellement étudier. Les points chauds définis par Myers *et al.* (2000), par exemple, sont très informatifs puisque Myers précise exactement sur quels critères ces points chauds sont fondés. Selon que les critères sont considérés comme bons, suffisants ou au contraire erronés ou insuffisants, les emplacements des points chauds peuvent être soutenus ou réfutés car ils sont clairement définis. Pour la définition de ces points chauds, Myers *et al.* (2000) considèrent les espèces comme la forme la plus importante et la plus facilement reconnaissable de biodiversité, mais s'empressent d'ajouter qu'ils ne suggèrent pas que les populations et même les processus écologiques ne sont pas des manifestations importantes de la biodiversité ; cependant elles n'apparaissent pas dans leurs estimations. Dans l'article, ils donnent néanmoins comme informations complémentaires, les nombres de familles et de genres endémiques à chacun des points chauds définis. La mesure directe de l'histoire évolutive représentée par les organismes d'une région pourrait être une solution au problème du concept d'espèces (Mace *et al.* 2003). Les conflits de définition d'espèces devraient surtout avoir lieu entre groupes d'organismes partageant une longue histoire évolutive commune et représentant moins d'histoire évolutive unique. L'erreur faite dans l'évaluation de la biodiversité devrait donc être moins grande en considérant l'histoire évolutive qu'en comptant simplement le nombre d'espèces. Une question intéressante serait alors de savoir si les points chauds de la diversité en espèces sont aussi des points chauds de l'histoire évolutive (Mace *et al.* 2003).

Tout ce qu'on doit exiger d'une mesure de diversité, c'est la connaissance exacte de ce qu'elle mesure, de son domaine d'application, de ses qualités et surtout de ses limites, le problème étant de choisir la bonne mesure selon le but poursuivi.

En conclusion, la biodiversité englobe tout ce qui est « bio » et divers mais chaque discipline scientifique n'en recouvre qu'une infime partie. Lorsqu'on évalue la biodiversité d'une région, il faudrait prendre en compte l'ensemble du vivant (archébactéries, eubactéries, eucaryotes), leurs interactions et leurs interactions avec leur environnement abiotique. Il faudrait aussi ne pas se restreindre à mesurer la diversité à un temps *t*, mais plutôt suivre les fluctuations et les grands changements de cette diversité à court et à long termes. C'est actuellement impossible.

> "Si un nombre suffisant d'espèces subissent l'extinction, les écosystèmes vont-ils s'effondrer, et l'extinction de la plupart des autres espèces va-t-elle s'ensuivre ? "peut-être" est la seule réponse que l'on puisse donner actuellement. [...] Il n'y a qu'une seule planète et qu'une seule expérience à observer. [...] La solution demandera que coopèrent des disciplines séparées depuis longtemps par leurs traditions pratiques et universitaires. La biologie, l'anthropologie, l'économie, l'agriculture, la politique et le droit devront trouver un langage commun." (Wilson 1993).

Tout est dit. Jusqu'à maintenant, chaque discipline a développé ses méthodes, son vocabulaire

et sa (ou ses) vision(s) de ce qu'est la biodiversité. Nous avons vu que des disciplines souvent séparées, telles que la biologie moléculaire et l'écologie, travaillent avec des notations et des dénominations différentes sur des structures de données et des outils semblables.

> "Une ancienne fable indienne raconte que six hommes, tous aveugles, rencontrèrent un jour un éléphant. Le premier toucha les flancs de l'animal : « Il s'agit d'un mur », déclara-t-il. Le deuxième saisit la trompe et conclut que l'éléphant était un serpent géant. Le troisième sentit la pointe de la défense : « Cette créature est aussi dangereuse qu'un sabre », dit-il. Le quatrième homme, après avoir enlacé une des pattes de l'animal, pensa avoir affaire à un arbre. Le cinquième, sentant l'énorme oreille du pachyderme, déclara qu'il s'agissait d'un éventail, peut-être d'un tapis volant. Le sixième, après avoir attrapé la queue de l'animal, fut convaincu qu'un éléphant n'était rien d'autre qu'une veille corde. Ils commencèrent alors à se quereller sur la nature de cet animal étrange. Réveillé par leurs cris, un rajah arriva et leur dit : « Comment pouvez-vous chacun être certain d'avoir raison ? L'éléphant est un grand animal et vous n'avez chacun touché qu'une partie de son corps. Si vous mettez les parties ensemble, vous verrez peut-être la vérité. » (Lambin 2004, page 217)"

Nous sommes tous de tels aveugles, nous ne touchons qu'une infime partie d'un problème vaste et éminemment complexe : la connaissance du monde vivant sous toutes ses formes et à toutes ses échelles.

Perspectives :

Les variations spatio-temporelles ont juste été abordées au cours de cette thèse. L'étude de l'impact de l'entropie quadratique dans ce cadre est une première perspective de prolongement de ce travail. L'entropie quadratique présente en effet de bonnes propriétés, mais aussi des limites qui nous ont amenés, dans le cadre de la mesure de la biodiversité, à réduire son utilisation aux dissimilarités ultramétriques. Cela représente pour certains jeux de données une forte contrainte. C'est parce qu'il existe un cadre, donc des limites, à chaque indice, qu'il nous faut les dépasser en cherchant d'autres possibilités, donc d'autres indices. L'entropie quadratique est au cœur d'un schéma mathématique solide. Elle ouvre donc un chemin vers d'autres recherches pour mesurer et décrire la biodiversité.

Les procédures de tests paramétriques présentées dans cette thèse ont souvent été développées par des statisticiens. Elles sont basées sur l'hypothèse multinomiale, c'est-à-dire que pour que les procédures de tests développées soient valides il faut que les variables étudiées, ici des effectifs, suivent la loi multinomiale. Souvent, elles supposent aussi que les individus échantillonnés soient indépendants. Or les données biologiques vérifient rarement ces hypothèses. Les organismes peuvent par exemple être agrégés dans leur répartition spatiale. Ou bien dans des études pluri-spécifiques, différentes espèces peuvent avoir différentes probabilités d'être vue ou capturées, elles peuvent aussi inter-agir. Une méthode statistique d'étude de la variation est susceptible d'avoir une application pratique dans n'importe quel domaine de la biologie puisque la variation est intrinsèque à tout jeu de données échantillonnées à partir d'une population non

uniforme. Cependant, chaque jeu de données en biologie possède des propriétés particulières qui dépendent du plan d'échantillonnage et du support biologique et nécessite ainsi une adaptation des méthodes statistiques ou encore l'élaboration de nouvelles méthodes.

Ainsi, nous avons vu que des biologistes peuvent avec certains types de données avoir besoin d'un nouvel indice et le créer, on aboutit à des indices ad hoc. Une fois la formule obtenue, il leur faut étudier les propriétés de l'indice, cherchant à le replacer dans un schéma statistique. Pour un indice de diversité, trois de ces propriétés sont, selon Lande (1996), son espérance, sa variance et son éventuelle concavité. La question de la concavité peut se révéler complexe et demander l'intervention d'un mathématicien. Le problème qu'un biologiste rencontre alors est qu'il est forcément difficile d'intéresser un mathématicien à une question souvent éloignée de sa propre sphère de recherche, question peut-être naïve, et sûrement secondaire face au mouvement actuel de l'avancée des connaissances mathématiques. Je me suis heurtée à ce genre de difficultés après que Carlo Ricotta, Université de Rome, Italie, m'ait contactée au cours de cette thèse pour quelques questions au sujet de propriétés mathématiques d'indices, propriété métrique et euclidienne pour des fonctions de dissimilarité et propriété de concavité pour des indices de diversité. Certaines réponses sont encore à trouver. Pour illustrer mon propos, j'aimerais parler d'un des nouveaux indices de diversité qu'il propose (Ricotta 2005). Ce nouvel indice est très intéressant parce qu'il généralise dans une même formule, pour la mesure de la diversité basée sur une matrice de dissimilarité, à la fois l'indice de Shannon par

$$Q_1 = -\sum_{j=1}^{S} p_j \ln\left(1 - \sum_{i=1}^{S} d_{ij}\right)$$

et celui de Gini-Simpson par l'entropie quadratique

$$Q_2 = \sum_{i=1}^{S}\sum_{j=1}^{S} p_i p_j d_{ij}.$$

L'indice, nommé Q_α dépend d'un paramètre α. Lorsque $\alpha \to 1$, $Q_\alpha \to Q_1$, et Q_2 ($\alpha = 2$) est l'entropie quadratique. L'article de C. Ricotta qui présente ce nouvel indice général est en révision. La question posée est : l'indice est-il concave ? La réponse à cette question dépend bien sûr, dans la forme générale de Q_α, de la valeur de α. Nous avons la réponse pour $\alpha = 2$, puisqu'il s'agit de l'entropie quadratique. Pour les autres valeurs de α, les démonstrations étaient à rechercher. J'ai essayé sans succès de collaborer avec des mathématiciens. C'est finalement auprès du mathématicien László Szeidl, université de Pécs, Hongrie, que Carlo Ricotta obtenu gain de cause, les résultats seront bientôt publiés.

Ce type de collaboration est fondamental pour évaluer les domaines d'application d'un indice, ses limites et aboutir ainsi à des schémas fondamentaux solides mathématiquement et dont on saura interpréter biologiquement les résultats. Il y a donc nécessité d'un dialogue entre biologistes, statisticiens, et mathématiciens.

Avec János Izsák, Berzsenyi College, Hongrie, nous examinons actuellement l'impact de deux nouveaux indices de variation sur l'analyse des données biologiques. L'étude de ces

indices est en cours. De plus, avec Eric Marcon et Jean-Christophe Roggy, INRA, Kourou, Guyane, et Christopher Baraloto, Université du Michigan, Etats-Unis, nous développons un projet d'étude des liens entre diversité spécifique, taxonomique et diversité fonctionnelle des arbres dans certains sites de la forêt Guyanaise. Les principaux collaborateurs associés à ce projet sont Stéphane Guitet, ONF, Guyane, Robert Lensi, CEFE, Montpellier, Jean-François Molino, AMAP, Montpellier, João Ferraz, INPA, Manaus, Michelle Mack et Ted Schuur, Université de Floride, et Cam Webb, Université de Yale. Cette étude s'insère dans l'ensemble des recherches entreprises pour mieux comprendre le fonctionnement de la forêt et arriver à des connaissances nécessaires pour mettre en oeuvre des stratégies de conservation. En particulier ses résultat s'ajouteront à ceux déjà rassemblés afin de parvenir à régénérer la forêt sur les zones dénudées par les sites d'orpaillages.

Le lecteur aura compris que mon but est de travailler dans cette vaste discipline qu'est la biologie, et d'y travailler avec des moyens fournis par les statistiques, l'algèbre, la géométrie ou encore l'informatique. Je dois donc conserver les deux langages, celui de la biologie et celui des statistiques. Il m'a été en effet jusqu'à maintenant nécessaire de pouvoir travailler dans les deux registres, et ainsi aussi de pouvoir comprendre et lire à la fois les littératures écologique, génétique, et statistique, et je souhaite continuer. La connaissance de plusieurs disciplines apporte de nouveaux éclairages à l'étude des données biologiques. Ma démarche va donc dans le sens de la pluridisciplinarité.

Ce travail de thèse va me permettre de collaborer à deux projets d'étude financés par l'Institut Français de la Biodiversité. Ces deux projets s'articulent autour des quatre axes de réflexions suivants :

1. "développer des outils et des méthodes permettant la mesure d'indices variés de la biodiversité à partir de données de nature différentes : génétique, morphologie, taxonomie, mais aussi « perceptive » tenant compte aussi bien de la richesse des interactions interspécifiques que des représentations dont la diversité biologique fait l'objet auprès des communautés humaines concernées ;
2. apprécier et étudier la façon dont les populations locales et les systèmes écologiques s'adaptent à des perturbations extérieures, entraînant des transformations significatives dans les modes de gestion des territoires ;
3. évaluer la façon dont les politiques de conservation et de gestion liées à l'application de nouvelles réglementation juridiques tiennent compte des différentes interactions dont dépend le maintien de la diversité biologique ;
4. développer des outils et des méthodes de concertation et de négociation capables de favoriser les stratégies d'appropriation par les acteurs locaux des politiques de conservation, en proposant des recommandations pour l'élaboration d'une gestion intégrée."

Ces deux projets réunissent des jeunes chercheurs de spécialités variées : anthropologie, bioacoustique, biostatistique, droit, écologie, économie, éthnobotanique, génétique, géographie, horticulture, philosophie, politique, sociologie.

Le premier projet est intitulé "Modifications des sociétés et des écosystèmes du Rufiji liées aux politiques d'aménagement du Sud Tanzanien". Ma collaboration interviendra au niveau

de l'analyse statistique de données obtenues par une nouvelle méthode d'échantillonnage non invasive et non destructrice pour l'étude de la biodiversité : la bioacoustique. Cette technique a permis très récemment de découvrir de nouvelles espèces de primates dans les forêts côtières du Kenya et de la Tanzanie (Perkin *et al.* 2002). Le résumé global du projet est le suivant :

> La basse vallée du Rufiji est un espace en pleine mutation. La construction d'un pont et d'une nouvelle route, les modifications du cadre réglementaire, la création d'une nouvelle aire marine protégée, la construction prévue d'un barrage, sont autant d'événements susceptibles de modifier à la fois les écosystèmes et les sociétés. Dans le cadre plus large d'une réflexion sur la notion de potentiel adaptatif, notre équipe pluridisciplinaire et franco-tanzanienne souhaite développer des méthodes et des outils d'analyse des interactions Société-Environnement et 1/porter des regards croisés sur la dynamique des représentations, pratiques et savoirs locaux et leur évolution récente ; 2/ développer une nouvelle méthode acoustique et statistique pour estimer la biodiversité animale en milieu de forêt et ainsi estimer les perturbations récentes des écosystèmes forestiers ; 3/ analyser les stratégies des acteurs locaux face aux politiques de gestion de la biodiversité.

Le deuxième projet intitulé "Biodiversité, perceptions et usages : des parcs urbains de la ville d'Angers au Parc naturel régional de Camargue" a pour but d'observer les conséquences des modes de gestion des parcs urbains et naturels sur plusieurs aspects de leur biodiversité en regardant notamment, la diversité esthétique ou perçue, la diversité morphologique, la diversité génétique. Deux modèles seront utilisés : les parcs urbains de la ville d'Angers et le parc naturel régional de Camargue. Ma collaboration interviendra au niveau de la mesure des différents aspects de la biodiversité. Ci-dessous est donné un extrait du résumé globale du projet :

> "Les modes de gestion d'un territoire dépendent de ses usages, de ses représentations sociales, du contexte socio-économique et historique et des interactions avec les autres territoires. [...] Notre proposition vise à mobiliser diverses disciplines (écologie, géographie, sociologie, modélisation multi-agents) afin d'étudier quels sont les effets des modes de gestion sur la biodiversité mesurée et la biodiversité perçue par les usagers et les non usagers des territoires. [...] Nous développerons des indices de biodiversité ainsi que des indicateurs permettant de caractériser la variété des usages associés aux systèmes étudiés. [...] La démarche proposée est de nature à éclairer la décision publique sur le choix des méthodes et procédures pour répondre aux enjeux de protection et de développement durable des territoires. Elle doit également fournir aux acteurs impliqués localement dans leur gestion une voie pour rechercher un compromis dans la négociation sur les règles de gestion durable des écosystèmes, tenant compte des incertitudes scientifiques et des risques inhérents à leurs dynamiques. [...]"

Ces projets sont, pour leurs acteurs, un réel défi à relever puisqu'ils vont mettre en œuvre une collaboration entre des chercheurs d'une dizaine de disciplines scientifiques, qui comme le souligne Wilson (1993), ont été longtemps séparés par les traditions universitaires. Nous devrons donc, avec nos habitudes et nos langages, trouver les moyens d'aller, chacun par son apport, et ensemble par nos interactions, tous dans une même direction.

R´ ´

Agapow P.-M., Bininda-Emonds O. R. P., Crandall K. A., Gittleman J. L., Mace G. M., Marshall J. C. et Purvis A. (2004). The impact of species concept on biodiversity studies. *The Quaterly Review of Biology*, 79 : 161–179.

Agresti A. et Agresti B. (1978). Statistical analysis of qualitative variation. In Schuessler K., editor, *Statistical Methodology*, pages 204–237.

Allan J. D. (1975). Components of diversity. *Oecologia*, 18 : 359–367.

Altschul S. et Lipman D. (1990). Equal animals. *Nature*, 348 : 493–494.

Anderson M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, 26 : 32–46.

Anderson M. J. et Willis T. (2003). Canonical analysis of principal coordinates : a useful method of constrained ordination for ecology. *Ecology*, 84 : 511–525.

Balakrishnan V. et Sanghvi L. D. (1968). Distance between populations on the basis of attribute data. *Biometrics*, 24 : 859–865.

Barbault R. (1997). *Biodiversité : Introduction à la biologie de la conservation*. Hachette, Paris.

Barker G. M. (2002). Phylogenetic diversity : a quantitative framework for measurement of priority and achievement in biodiversity conservation. *Biological Journal of the Linnean Society*, 76 : 165–194.

Bartlett M. (1937). Some examples of statistical methods of research on agriculture and applied biology. *Journal of the Royal Statistical Society*, Supplement 4 : 137–170.

Ben-Moshe A., Dayan T. et Simberloff D. (2001). Convergence in mophological patterns and community organization between old and new world rodent guilds. *The American Naturalist*, 158 : 484–495.

Bhattacharyya A. (1946). On a measure of divergence between two multinomial populations. *Sankhya*, 7 : 407–406.

Bininda-Emonds O. R. P., Gittleman J. L. et Purvis A. (1999). Building large trees by combining phylogenetic information : a complete phylogeny of the extant Carnivora (Mammalia). *Biological Reviews of the Cambridge Philosophical Society*, 74 : 143–175.

Blondel J. (2000). *Biogéographie : Approche écologique et évolutive*, volume 27 de *Collection écologique*. Masson.

Blondel J., Vuilleumier F., Marcus L. et Terouanne E. (1984). Is there ecomorphological convergence among mediterranean bird communities of Chile, California, and France. In Hecht M., Wallace B. et MacIntyre R., editors, *Evolutionary Biology*, volume 18, pages 141–213. Plenum Press, New York.

Bosch E., Calafell F., Santos F., Perez-Lezaun A., Comas D., Benchemsi N., Tyler-Smith C. et Bertranpetit J. (1999). Variation in short tandem repeats is deeply structured by

genetic background on the human Y chromosome. *American Journal of Human Genetics*, 65 : 1623–1638.

Bray J. R. et Curtis J. (1957). An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs*, 27 : 325–349.

Bremer B. K., Chase M., Reveal J., Soltis D., Soltis P., Stevens P., Anderberg A., Fay M., Goldblatt P., Judd W., Kallersjo M., Karehed J., Kron K., Lundberg J., Nickrent D., Olmstead R., Oxelman B., Pires J., Rodman J., Rudall P., Savolainen V., Sytsma K., van der Bank M., Wurdack K., Xiang J. et Zmarzty S. (2003). An update of the angiosperm phylogeny group classification for the orders and families of flowering plants : APG II. *Botanical Journal of the Linnean Society, London*, 141 : 399–436.

Bried J., Pontier D. et Jouventin P. (2003). Mate fidelity in monogamous birds : a re-examination of the Procellariiformes. *Animal Behaviour*, 65 : 235–246.

Buzas M. et Gibson T. (1969). Species diversity : benthonic Foraminifera in western North Atlantic. *Science*, 163 : 72–75.

Cailliez F. (1983). The analytic solution of the additive constant problem. *Psychometrika*, 48 : 305–310.

Carnes B. et Slade N. (1982). Some comments on niche analysis in canonical space. *Ecology*, 63 : 888–893.

Caron H. (2000). *Organisation et dynamique de la diversité génétique de cinq espèces arborées de la forêt guyanaise*. Thèse de doctorat, Université Montpellier II.

Carrascal L. M., Moreno E. et Telleria J. L. (1990). Ecomorphological relationships in a group of insectivorous birds of temperate forests in winter. *Holarctic Ecology*, 13 : 105–111.

Chakraborty R. et Rao C. (1991). Measurement of genetic variation for evolutionary studies. In Rao C. et Chakraborty R., editors, *Handbook of statistics : Statistical Methods in Biological and Medical Sciences*, volume 8, pages 271–316. Elsevier Science Publishers B.V.

Champely S. et Chessel D. (2002). Measuring biological diversity using Euclidean metrics. *Environmental and Ecological Statistics*, 9 : 167–177.

Chao A. (1984). Non-parametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics*, 11 : 265–270.

Chao A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, 43 : 783–791.

Chao A., Chazdon R. L., Colwell R. K. et Shen T.-J. (2005). A new statistical approach for assessing similarity of species composition with incidence and abundance data. *Ecology Letters*, 8 : 148–159.

Chessel D., Dufour A.-B. et Thioulouse. J. (2004). The ade4 package-i- one-table methods. *R News*, 4 : 5–10.

Chessel D. et Hanafi M. (1996). Analyses de la co-inertie de k nuages de points. *Revue de Statistique Appliquée*, 44 : 35–60.

Chessel D., Lebreton J. et Prodon R. (1982). Mesures symétriques d'amplitude d'habitat et de diversité intra-échantillon dans un tableau espèces-relevés : cas d'un gradient simple. *Compte rendu hebdomadaire des séances de l'Académie des sciences. Paris, D*, III : 83–88.

Chessel D., Lebreton J. et Yoccoz N. (1987). Propriétés de l'analyse canonique des correspondances. une utilisation en hydrobiologie. *Revue de Statistique Appliquée*, 35 : 55–72.

Cho A. (2002). A fresh take on disorder, or disorderly science ? *Science*, 297 : 1268–1269.

Chown S. L., Pistorius P. et Scholtz C. H. (1998). Morphological correlates of flightlessness in southern African Scarabaeinae (Coleoptera : Scarabaeidae) : testing a condition of the water-conservation hypothesis. *Canadian Journal of Zoology - Journal Canadien de Zoologie*, 76 : 1123–1133.

Clarke K. R. et Warwick R. M. (1998). A taxonomic distinctness index and its statistical properties. *Journal of Applied Ecology*, 35 : 523–531.

Clarke K. R. et Warwick R. M. (1999). The taxonomic distinctness measure of biodiversity : weighting of step lengths between hierarchical levels. *Marine Ecology - Progress Series*, 184 : 21–29.

Clarke K. R. et Warwick R. M. (2001). A further biodiversity index applicable to species lists : variation in taxonomic distinctness. *Marine Ecology - Progress Series*, 216 : 265–278.

Colwell R. K. (2000). Rensch's rule crosses the line : convergent allometry of sexual size dimorphism in hummingbirds and flower mites. *The American Naturalist*, 156 : 495–510.

Colwell R. K. et Coddington J. A. (1994). Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society, B*, 345 : 101–118.

Colwell R. K. et Futuyma D. (1971). On the measurement of niche breadth and niche overlap. *Ecology*, 52 : 567–576.

Comas D., Calafell F., Bendukidze N., Fananas L. et Bertranpetit J. (2000). Georgian and Kurd mtDNA sequence analysis shows a lack of correlation between languages and female genetic lineages. *American Journal of Physical Anthropology*, 112 : 5–16.

Comas D., Calafell F., Mateu E., Pérez-Lezaun A. et Bertranpetit J. (1996). Geographical variation in human mitochondrial DNA control region sequence : the population history of Turkey and its relationship to the European populations. *Molecular Biology and Evolution*, 13 : 1067–1077.

Comas D., Calafell F., Mateu E., Pérez-Lezaun A., Bosch E., Martinez-Arias R., Clarimon J., Facchini F., Fiori G., Luiselli D., Pettener D. et Bertranpetit J. (1998). Trading genes along the silk road : mtDNA sequences and the origin of Central Asian populations. *American Journal of Human Genetics*, 63 : 1824–1838.

Cooper A. et Fortey R. (1998). Evolutionary explosions and the phylogenetic fuse. *Trends in Ecology and Evolution*, 13 : 151–156.

Cousins S. (1991). Species diversity measurement : choosing the right index. *Trends in Ecology and Evolution*, 6 : 190–192.

Couteron P. et Pélissier R. (2004). Additive apportionment of species diversity : towards more sophisticated models and analyses. *Oikos*, 107 : 215–221.

Couteron P., Pélissier R., Mapaga D., Molino J.-F. et Teillier L. (2003). Drawing ecological insights from a management-oriented forest inventory in French Guiana. *Forest Ecology and Management*, 172 : 89–108.

Crandall K., Bininda-Emonds O. R., Mace G. M. et Wayne R. K. (2000). Considering evolutionary processes in conservation biology. *Trends in Ecology and Evolution*, 15 : 290.

Crease T., Lynch M. et Spitze K. (1990). Hierarchical analysis of population genetic variation in mitochondrial and nuclear genes of *Daphnia pulex*. *Molecular Biology and Evolution*, 7 : 444–458.

Critchley F. et Fichet B. (1997). On (super-)spherical distance matrices and two results from Schoenberg. *Linear Algebra and its Applications*, 251 : 145–165.

Cronquist A. (1981). *An Integrated System of Classification of Flowering Plants*. Columbia University Press, New York.

Crozier R. (1992). Genetic diversity and the agony of choice. *Biological Conservation*, 61 : 11–15.

Crozier R. et Kusmierski R. (1994). Genetic distances and the setting of conservation priorities. In Loeschcke V., Tomiuk J. et Jain S., editors, *Conservation Genetics*, pages 227–237. Birkhäser Verlag, Basel.

Curtis T. P. et Sloan W. T. (2004). Prokaryotic diversity and its limits : microbial community structure in nature and implications for microbial ecology. *Current Opinion in Microbiology*, 7 : 221–226.

Daugherty C., Cree A., Hay J. et Thompson M. (1990). Neglected taxonomy and continuing extinctions of tuatara (Sphenodon). *Nature*, 347 : 177–179.

DeSalle R. et Amato G. (2004). The expansion of conservation genetics. *Nature*, 5 : 702–712.

DeVries P. J., Murray D. et Lande R. (1997). Species diversity in vertical, horizontal, and temporal dimensions of a fruit-feeding butterfly community in an Ecuadorian rainforest. *Biological Journal of the Linnean Society*, 62 : 343–364.

DeVries P. J., Walla T. R. et Greeney H. F. (1999). Species diversity in spatial and temporal dimensions of fruit-feeding butterflies from two Ecuadorian rainforests. *Biological Journal of the Linnean Society*, 68 : 333–353.

Dice L. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26 : 297–302.

Diniz-Filho J. A. F. (2004). Phylogenetic autocorrelation analysis of extinction risks and the loss of evolutionary history in Felidae (Carnivora : Mammalia). *Evolutionary Ecology*, 18 : 273–282.

Diniz-Filho J. A. F. et Tôrres N. M. (2002). Phylogenetic comparative methods and the geographic range size - body size relationship in new world terrestrial Carnivora. *Evolutionary Ecology*, 16 : 351–367.

Dolédec S. et Chessel D. (1987). Rhythmes saisonniers et composantes stationnelles en milieu aquatique. I- Description d'un plan d'observation complet par projection de variables. *Acta Œcologica - Œcologia Generalis*, 8 : 403–426.

Drouet d'Aubigny G. (1989). *L'analyse multidimensionelle des données de dissimilarité*. Thèse de doctorat, Université Grenoble I.

Eckburg P. B., Bik E. M., Bernstein C. N., Purdom E., Dethlefsen L., Sargent M., Gill S. R., Nelson K. E. et Relman D. A. (2005). Diversity of the human intestinal microbial flora. *Science*, 308 : 1635–1638.

Edwards A. (1971). Distance between populations on the basis of gene frequencies. *Biometrics*, 27 : 873–881.

Efron B. (1982). *The jackknife, the bootstrap and other resampling plans*. Society for Industrial and Applied Mathematics, Philadelphia.

Eiswerth M. et Haney J. (1992). Allocating conservation expenditures across habitats : Accounting for inter-species genetic distinctiveness. *Ecological Economics*, 5 : 235–250.

Erwin T. (1991). An evolutionary basis for conservation strategies. *Science*, 253 : 750–752.

Escudero A., Iriondo J. M. et Elena T. M. (2003). Spatial analysis of genetic diversity as a tool for plant conservation. *Biological diversity*, 113 : 351–365.

Excoffier L. (1994). The statistical analysis of molecular data for inferring population genetic structure : the AMOVA framework. In Perrin N., Keller L. et Goudet J., editors, *Evolution in structured populations*, volume 83, pages 159–160. Bull. Soc. Vaud. Sc. Nat.

Excoffier L. (2001). Analysis of population subdivision. In Balding D., Bishop M. et Cannings C., editors, *Handbook of Statistical Genetics*, pages 271–307. John Wiley and Sons, Ltd, New York.

Excoffier L., Smouse P. et Quattro J. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes : application to human mitochondrial DNA restriction data. *Genetics*, 131 : 479–491.

Faith D. P. (1992a). Conservation evaluation and phylogenetic diversity. *Biological Conservation*, 61 : 1–10.

Faith D. P. (1992b). Systematics and conservation : on predicting the feature diversity of subsets of taxa. *Cladistics*, 8 : 361–373.

Faith D. P. (1993). Biodiversity and systematics : the use and misuse of divergence information in assessing taxonomic diversity. *Pacific Conservation Biology*, 1 : 53–57.

Faith D. P. (1994a). Genetic diversity and taxonomic priorities for conservation. *Biological Conservation*, 68 : 69–74.

Faith D. P. (1994b). Phylogenetic diversity : a general framework for the prediction of fea-

ture diversity. In FOREY P., HUMPHRIES C. et VANE-WRIGHT R., editors, *Systematics and Conservation Evaluation*, pages 251–268. Oxford University Press.

FAITH D. P. (1995). Phylogenetic pattern and the quantification of organismal biodiversity. In HAWKSWORTH D., editor, *Biodiversity Measurement and Estimation*, volume 345, pages 45–58. Chapman and Hall, The Royal Society, London.

FAITH D. P. (1996). Conservation priorities and phylogenetic pattern. *Conservation Biology*, 10 : 1286–1289.

FAITH D. P. (2003). *Biodiversity*. The Stanford Encyclopedia of Philosophy (Summer 2003 Edition), URL = <http ://plato.stanford.edu/archives/sum2003/entries/biodiversity/>.

FAITH D. P., MINCHIN P. et BELBIN L. (1987). Compositional dissimilarity as a robust measure of ecological distance. *Vegetatio*, 69 : 57–68.

FINDLEY J. S. (1973). Phenetic packing as a measure of faunal diversity. *The American Naturalist*, 107 : 580–584.

FINDLEY J. S. (1976). The structure of bat communities. *The American Naturalist*, 110 : 129–139.

FINKELDEY R. (1994). A simple derivation of the partitioning of genetic differentiation within subdivided populations. *Theoretical and Applied Genetics*, 89 : 198–200.

FISHER R. (1925). *Statistical methods for research workers*. Oliver and Boyd, Edinburgh.

FISHER R. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7 : 179–188.

FISHER R., STEVEN CORBET A. et WILLIAMS C. (1943). The relationship between the number of species and the number of individuals in a random sample of an animal population. *Journal of animal Ecology*, 12 : 42–58.

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

FOURNIER E. et LOREAU M. (2001). Respective roles of recent hedges and forest patch remnants in the maintenance of ground-beetle (Coleoptera : Carabidae) diversity in an agricultural landscape. *Landscape Ecology*, 16 : 17–32.

GASTON K. (1996). What is biodiversity ? In GASTON K., editor, *Biodiversity : a biology of numbers and difference*, pages 1–9. Blackwell Science, Owford.

GAYON J. (1996). The individuality of the species : a Darwinian theory ? from Buffon to Ghiselin, and back to Darwin. *Biology and philosophy*, 11 : 215–244.

GERING J. C., CRIST T. O. et VEECH J. A. (2003). Additive partitioning of species diversity across multiple spatial scales : implications for regional conservation of biodiversity. *Conservation Biology*, 17 : 488–499.

GIMARET-CARPENTIER C., CHESSEL D. et PASCAL J.-P. (1998). Non-symmetric correspondence analysis : an alternative for species occurences data. *Plant Ecology*, 138 : 97–112.

Gini C. (1912). Variabilita e mutabilita, studi economicoaguridici della facotta di ginrisprudenza dell. Technical report, Universite di Cagliari III.

Gleason H. (1920). Some applications of the quadrat method. *Bulletin of the Torrey Botanical Club*, 47 : 21–33.

Gore A. P. (1994). Comment to Solow, A.R. and Polaski, S. 1994 Measuring biological diversity Environmental and Ecological Statistics 1 : 95-107. *Environmental and Ecological Statistics*, 1 : 106–107.

Gower J. C. (1982). Euclidean distance geometry. *Mathematical Scientist*, 7 : 1–14.

Gower J. C. (1984). Distance matrices and their Euclidean approximation. In Diday E., Jambu M., Lebart L., Pagès J. et Tomassone R., editors, *Data Analysis and Informatics III*, pages 3–21. Elsevier, Amsterdam, North-Holland.

Gower J. C. et Legendre P. (1986). Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification*, 3 : 5–48.

Grant P. R. et Grant B. R. (2002). Unpredictable evolution in 30-year study of Darwin's Finches. *Science*, 296 : 707–711.

Grassle J. et Smith W. (1976). A similarity measure sensitive to the contribution of rare species and its use in investigation of variation in marine benthic communities. *Oecologia*, 25 : 13–22.

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

Harper J. L. et Hawksworth D. L. (1995). Preface. In Hawksworth D. L., editor, *Biodiversity mesurement and estimation*, pages 5–12. Chapman and Hall, London.

Hattemer H. H. (1982). Genetic distance between populations. Part 3 : Wahlund's principle as related to genetic distance and an application. *Theoretical and Applied Genetics*, 62 : 219–223.

Havrda M. et Charvat F. (1967). Quantification method of classification processes : concept of structural α-entropy. *Kybernatica*, 3 : 30–35.

Hendrickson J. A. J. et Ehrlich P. R. (1971). An expanded concept of "species diversity". *Notulae Naturae*, 439 : 1–6.

Herrel A., Meyers J. J. et Vanhooydonck B. (2001). Correlations between habitat use and body shape in a Phrynosomatid lizard (Urosaurus ornatus) : a population-level analysis. *Biological Journal of the Linnean Society*, 74 : 305–314.

Hespenheide H. A. (1973). Ecological inferences from morphological data. *Annual Review of Ecology and Systematics*, 4 : 213–229.

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

Hill T. C. J., Walsh K. A., Harris J. A. et Moffett B. F. (2003). Using ecological diversity measures with bacterial communities. *Fems Microbiology Ecology*, 43 : 1–11.

Holsinger K. E. et Mason-Gamer R. J. (1996). Hierarchical analysis of nucleotide diversity in geographically structured populations. *Genetics*, 142 : 629–639.

Hosken D. J. et Balloux F. (2002). Thirty years of evolution in Darwin's Finches. *Trends in Ecology and Evolution*, 17 : 447–448.

Hudson R., Boos D. et Kaplan N. (1992). A statistical test for detecting geographical subdivision. *Molecular Biology and Evolution*, 9 : 138–151.

Hugues J. B., Daily G. C. et Ehrlich P. R. (1997). Population diversity : its extent and extinction. *Science*, 278 : 689–692.

Humphries C. et Williams P. (1994). Cladograms and trees in biodiversity. In Scotland R., Siebert D. et Williams D., editors, *Models in phylogenetic construction*, pages 335–352. Clarendon, Oxford.

Humphries C., Williams P. et Vane-Wright R. (1995). Measuring biodiversity value for conservation. *Annual Review of Ecology and Systematics*, 26 : 93–111.

Hurlbert S. (1971). The non-concept of species diversity : a critique and alternative parameters. *Ecology*, 52 : 577-586 : 52 : 577–586.

Hwang U. W., Friedrich M., Tautz D., Park C. J. et Kim W. (2001). Mitochondrial protein phylogeny joins myriapods with chelicerates. *Nature*, 413 : 154–157.

Ihaka R. et Gentleman R. (1996). R : a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5 : 299–314.

Irschick D. J. et Losos J. B. (1998). A comparative analysis of the ecological significance of maximal locomotor performance in Caribbean Anolis lizards. *Evolution*, 52 : 219–226.

Irschick D. J. et Losos J. B. (1999). Do lizards avoid habitats in which performance is submaximal ? the relationship between sprinting capabilities and structural habitat use in Caribbean anoles. *The American Naturalist*, 154 : 293–305.

IUCN/UNEP/WWF (1980). *World conservation strategy, living resource conservation for sustainable development*. IUCN, UNEP and WWF, Gland, Switzerland.

Ivol J. M., Guinand B., Richoux P. et Tachet H. (1997). Longitudinal changes in Trichoptera and Coleoptera assemblages and environmental conditions in the Loire River (France). *Archiv für Hydrobiologie*, 138 : 525–557.

Izsak C. et Price A. (2001). Measuring β-diversity using a taxonomic index, and its relation to spatial scale. *Marine Ecology - Progress Series*, 215 : 69–77.

Izsak J. et Papp L. (1995). Application of the quadratic entropy indices for diversity studies of drosophilid assemblages. *Environmental and Ecological Statistics*, 2 : 213–224.

Izsak J. et Papp L. (2000). A link between ecological diversity indices and measures of biodiversity. *Ecological Modelling*, 130 : 151–156.

Izsak J. et Szeidl L. (2002). Quadratic diversity : its maximization can reduce the richness of species. *Environmental and Ecological Statistics*, 9 : 423–430.

JACCARD P. (1901). Etude comparative de la distribution florale dans une portion des Alpes et du Jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37 : 547–579.

JACKSON D. A., SOMERS K. M. et HARVEY H. H. (1989). Similarity coefficients : measures of co-occurrence and association or simply measures of occurrence ? *The American Naturalist*, 133 : 436–453.

JAMES F. C. (1982). The ecological morphology of birds : a review. *Annales Botanici Fennici*, 19 : 265–275.

JAMES F. C. (1983). Environmental component of morphological differentiation in birds. *Science*, 221 : 184–186.

JENSEN J. (1906). Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica*, 30 : 175–193.

JIN L. et NEI M. (1990). Limitations of the evolutionary parsimony method of phylogenetic analysis. *Molecular Biology and Evolution*, 7 : 82–102.

JUKES T. et CANTOR C. (1969). Evolution of protein molecules. In MUNRO H., editor, *Mammaliam protein metabolism*, pages 21–132. Academic press, New York.

KARP A., SEBERG O. et BUIATTI M. (1996). Molecular techniques in the assessment of botanical diversity. *Annals of Botany Company*, 78 : 143–149.

KAWANO K. (2002). Character displacement in giant rhinoceros beetles. *The American Naturalist*, 159 : 255–271.

KIMURA M. (1980). A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16 : 111–120.

KOCH L. F. (1957). Index of biotal dispersion. *Ecology*, 38 : 145–148.

KRAJEWSKI C. (1991). Phylogeny and diversity. *Science*, 254 : 918–819.

KREMER A., PETIT R. et PONS O. (1998). Measures of polymorphism within and among populations. In KARP A., ISAAC P. et INGRAM D., editors, *Molecular tools for screening biodiversity, plants and animals*. Chapman and Hall, London.

KRUSKAL J. (1964a). Multidimensional scaling by optimizing goodness of fit to nonmetric hypothesis. *Psychometrika*, 29 : 1–27.

KRUSKAL J. (1964b). Nonmetric multidimensional scaling : a numerical method. *Psychometrika*, 29 : 115–129.

KUHNLEIN U., DAWE Y. et GAVORA J. (1989). DNA fingerprinting : a tool for determining genetic distances between strains of poultry. *Theoretical and Applied Genetics*, 77 : 669–672.

LAMBIN E. (2004). *La Terre sur un fil*. Le Pommier, Paris.

LANCE G. et WILLIAMS W. (1966). A generalized sorting strategy for computer classifications. *Nature*, 212 : 218.

LANCE G. et WILLIAMS W. (1967). A general theory of classificatory sorting strategies. I Hierar-

chical systems. *Computer Journal*, 9 : 373–380.

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

Lauro N. et D'Ambra L. (1984). L'analyse non symétrique des correspondances. In Diday E., Jambu M., Lebart L., Pages J. et Tomassone R., editors, *Data Analysis and Informatics, III*, pages 433–446. Elsevier, North-Holland.

Laval G., Iannuccelli N., Legault C., Milan D., Groenen M. A. M., Giuffra E., Andersson L., Nissen P. H., Jorgensen C. B., Beeckmann P., Geldermann H., Foulley J.-L., Chevalet C. et Ollivier L. (2000). Genetic diversity of eleven European pig breeds. *Genetics Selection Evolution*, 32 : 187–203.

Laval G., SanCristobal M. et Chevalet C. (2002). Measuring genetic distances between breeds : use of some distances in various short term evolution models. *Genetics Selection Evolution*, 34 : 481–507.

Leal M., Knox A. K. et Losos J. B. (2002). Lack of convergence in aquatic anolis lizards. *Evolution*, 56 : 785–791.

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

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

Lebreton J., Chessel D., Prodon R. et Yoccoz N. (1988b). L'analyse des relations espèces-milieu par l'analyse canonique des correspondances. II. Variables de milieu qualitatives. *Acta Œcologica, Œcologia Generalis*, 9 : 137–151.

Legendre P. et Anderson M. (1999). Distance-based redundancy analysis : testing multispecies responses in multifactorial ecological experiments. *Ecological Monographs*, 69 : 1–24.

Legendre P. et Legendre L. (1998). *Numerical ecology*. Elsevier Science BV, Amsterdam, 2nd English edition.

Leisler B. et Winkler H. (1985). Ecomorphology. *Current Ornithology*, 2 : 155–186.

Lennon J. J., Koleff P., Greenwood J. J. D. et Gaston K. J. (2001). The geographical structure of british bird distributions : diversity, spatial turnover and scale. *Journal of Animal Ecology*, 70 : 966–979.

Levins R. (1968). *Evolution in changing environments*. Princeton University Press, Princeton, New Jersey.

Lewontin R. C. (1972). The apportionment of human diversity. *Evolutionary Biology*, 6 : 381–398.

Light R. et Margolin B. (1971). An analysis of variance for categorical data. *Journal of the American Statistical Association*, 66 : 534–544.

Lingoes J. (1971). Some boundary conditions for a monotone analysis of symmetric matrices.

Psychometrika, 36 : 195–203.

Liu Z. J. et Rao C. R. (1995). Asymptotic distribution of statistics based on quadratic entropy and bootstrapping. *Journal of Statistical Planning and Inference*, 43 : 1–18.

Lloyd M. et Ghelardhi R. (1964). A table for calculating the "equitability" component of species diversity. *Journal of Animal Ecology*, 33 : 217–225.

Loreau M. (2000). Are communities saturated ? On the relationship between α, β and γ diversity. *Ecology letters*, 3 : 73–76.

Losos J. B. (1990a). Concordant evolution of locomotor behaviour, display rate and morphology in Anolis lizards. *Animal Behaviour*, 39 : 879–890.

Losos J. B. (1990b). Ecomorphology, performance capacity, and scaling of west Indian Anolis lizards : an evolutionary analysis. *Ecological Monographs*, 60 : 369–388.

Losos J. B. (1990c). The evolution of form and function : morphology and locomotion performance in west Indian Anolis lizards. *The American Naturalist*, 44 : 1189–1203.

Losos J. B. (1992). The evolution of convergent structure in Caribbean Anolis communities. *Systematic Biology*, 41 : 403–420.

Losos J. B. et Miles D. B. (2002). Testing the hypothesis that a clade has adaptatively radiated : iguanid lizard clades as a case study. *The American Naturalist*, 160 : 147–157.

Lynch M. et Crease T. (1990). The analysis of population survey data on DNA sequence variation. *Molecular Biology and Evolution*, 7 : 377–394.

MacArthur R., Recher H. et Cody M. (1966). On the relation between habitat selection and species diversity. *American Naturalist*, 100 : 319–332.

Mace G. M., Gittleman J. L. et Purvis A. (2003). Preserving the tree of life. *Science*, 300 : 1707–1709.

Magurran A. (1988). *Ecological diversity and its measurement*. Croom Helm Limited, London.

Magurran A. (2004). *Measuring biological diversity*. Blackwell Publishing.

Manly B. F. (1994). *Multivariate Statistical Methods*. Chapman and Hall, London.

Margalef R. (1958). Information theory in ecology. *General Systems*, 3 : 36–71.

Margalef R. et Gutierrez E. (1983). How to introduce connectance in the frame of an expression for diversity. *The American Naturalist*, 121 : 601–607.

May R. (1990). Taxonomy as destiny. *Nature*, 347 : 129–130.

May R. M. (1995). Conceptual aspects of the quantification of the extent of biological diversity. In Hawksworth D., editor, *Biodiversity measurement and estimation*, volume 345, pages 13–20. Chapman and Hall, The Royal Society, London.

May R. M. (2002). The future of biological diversity in a crowded world. *Current Science*, 82 : 1325–1331.

Mayr E. (1942). *Systematics and the origin of species*. Columbia University Press, New York.

McArdle B. et Anderson M. (2001). Fitting multivariate models to community data : comment on distance-based redundancy analysis. *Ecology*, 82 : 290–297.

McNeely J., Miller K., Reid W., Mittermeier R. et Werner T. (1990). Conserving the world's biodiversity. Technical report, IUCN, WRI, CI, WWF and World Bank.

Menhinick E. P. (1964). A comparison of some species-individuals diversity indices applied to samples of field insects. *Ecology*, 45 : 859–861.

Merilä J., Sheldon B. C. et Kruuk L. E. B. (2001). Explaining stasis : microevolutionary studies in natural populations. *Genetica*, 112-113 : 199–222.

Michalakis Y. et Excoffier L. (1996). A generic estimation of population subdivision using distances between alleles with special reference for microsatellite loci. *Genetics*, 142 : 1061–1064.

Milano D., Cussac V. E., Macchi P. M., Ruzzante D. E., Alonso F. M., Vigliano P. H. et Denegri M. A. (2002). Predator associated morphology in *Galaxias platei* in Pantagonian lakes. *Journal of Fish Biology*, 61 : 138–156.

Mougel C., Thioulouse J., Perrière G. et Nesme X. (2002). A mathematical method for determining genome divergence and species delineation using AFLP. *International Journal of Systematic and Evolutionary Microbiology.*, 52 : 573–586.

Myers N., Mittermeier R. A., Mittermeier C. G., da Fonseca G. A. B. et Kent J. (2000). Biodiversity hotspots for conservation priorities. *Nature*, 403 : 853–858.

Nayak T. K. (1983). *Applications of entropy functions in measurement and analysis of diversity*. Thesis, University of Pittsburgh.

Nayak T. K. (1985). On diversity measures based on entropy functions. *Communications in Statistics - Theory and Methods*, 14 : 203–215.

Nayak T. K. (1986a). An analysis of diversity using Rao's quadratic entropy. *Sankhya : The indian Journal of Statistics*, 48B : 315–330.

Nayak T. K. (1986b). Sampling distributions in analysis of diversity. *Sankhya : The indian Journal of Statistics*, 48B : 1–9.

Nee S. et May R. M. (1997). Extinction and the loss of evolutionary history. *Science*, 278 : 692–694.

Nei M. (1972). Genetic distance between populations. *The American Naturalist*, 106 : 283–292.

Nei M. (1973). Analysis of gene diversity in subdivised populations. *Proceedings of the National Academy of Sciences of the United States of America*, 70 : 3321–3323.

Nei M. (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics*, 89 : 583–590.

Nei M. (1987). *Molecular evolutionary genetics*. Columbia University Press, New York, NY, USA.

Nei M. et Jin L. (1989). Variances of the average numbers of nucleotide substitutions within

and between populations. *Molecular Biology and Evolution*, 6 : 290–300.

Nei M. et Li W.-H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America*, 76 : 5269–5273.

Nei M. et Miller J. C. (1990). A simple method for estimating average number of nucleotide substitutions within and between populations from restriction data. *Genetics*, 125 : 873–879.

Nei M. et Tajima F. (1981). DNA polymorphism detectable by restriction endonucleases. *Genetics*, 97 : 145–163.

Neigel J. (2002). Is Fst obsolete ? *Conservation Genetics*, 3 : 167–173.

Nixon K. et Wheeler Q. (1992). Measures of phylogenetic diversity. In Novacek M. et Wheeler Q., editors, *Extinction and Phylogeny*, pages 216–234. Columbia University Press, New York.

Norton B. (1994). On what we should save : the role of culture in determining conservation targets. In Forey P., Humphries C. et Vane-Wright R., editors, *Systematics and conservation evaluation*, pages 23–39. Oxford University Press, Oxford.

Norton B. (2001). Conservation biology and environmental values : can there be a universal earth ethic ? In Potvin C., Kraenzel M. et Seutin G., editors, *Protecting biological diversity : roles and responsibilities*. McGill-Queen's university Press, Montréal.

Nowak R. M. (1991). *Walker's mammals of the world*. The Johns Hopkins University Press, Baltimore and London.

Ochiai A. (1957). Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions. *Bulletin of the Japanese Society of Scientific Fisheries*, 22 : 526–530.

O'Neill R., Gardner R., Milne, B.T. T. M. et Jackson B. (1991). Heterogeneity and spatial hierarchies. In Kosala J. et Pickett S., editors, *Ecological heterogeneity*, volume 86, pages 85–96. Springer-Verlag, New York.

Orlóci L. (1967). An agglomerative method for classification of plant communities. *Journal of Ecology*, 55 : 193–206.

Patil G. et Taillie C. (1982). Diversity as a concept and its measurement. *Journal of the American Statistical Association*, 77 : 548–561.

Pavoine S., Blondel J. et Chessel D. (en révision). A new technique for ordering multiple independent three-dimensional data sets in convergence studies. *Ecology*.

Pavoine S. et Dolédec S. (2005). The apportionment of quadratic entropy : a useful alternative for partitioning diversity in ecological data. *Environmental and Ecological Statistics*, 12 : 125–138.

Pavoine S., Dufour A. B. et Chessel D. (2004). From dissimilarities among species to dissimilarities among communities : a double principal coordinate analysis. *Journal of Theoretical Biology*, 228 : 523–537.

Pavoine S., Ollier S. et Dufour A. B. (2005a). Is the originality of a species measurable ? *Ecology Letters*, 8 : 579–586.

Pavoine S., Ollier S. et Pontier D. (2005b). Measuring diversity from dissimilarities with Rao's quadratic entropy : are any dissimilarity indices suitable ? *Theoretical Population Biology*, 67 : 231–239.

Peet R. (1974). The measurement of species diversity. *Annual Review of Ecology and Systematics*, 5 : 285–307.

Perkin A., Bearder S., Butynski T., Agwanda B. et Bytebier B. (2002). The taita mountain dwarf galago galagoides sp. : a new primate for kenya. *Journal of East African Natural History*, 91 : 1–13.

Petchey O. L. et Gaston K. (2002). Functional diversity (FD), species richness and community composition. *Ecology Letters*, 5 : 402–411.

Pielou E. (1969). *An introduction to mathematical ecology*. John Wiley and Sons, New York.

Pielou E. (1975). *Ecological diversity*. Wiley and Sons, New York.

Pillar V. D. P. et Orlóci L. (1996). On randomization testing in vegetation science : multifactor comparisons of relevé groups. *Journal of Vegetation Science*, 7 : 585–592.

Pélissier R., Couteron P., Dray S. et Sabatier D. (2003). Consistency between ordination techniques and diversity measurements : two strategies for species occurence data. *Ecology*, 84 : 242–251.

Possingham H. P., Andelman S., Burgman M. A., Medellin R. A., Master L. L. et Keith D. A. (2002). Limits to the use of threatened species lists. *Trends in Ecology and Evolution*, 17 : 503–507.

Prevosti A., Ocana J. et Alonso G. (1975). Distances between populations of *Drosophila subobscura* based on chromosome arrangement frequencies. *Theoretical and Applied Genetics*, 45 : 231–241.

Price A., Keeling M. et O'Callaghan C. (1999). Ocean-scale patterns of 'biodiversity' of atlantic asteroids determined from taxonomic distinctness and other measures. *Biological Journal of the Linnean Society*, 66 : 187–203.

Purvis A., Agapow P.-M., Gittleman J. L. et Mace G. M. (2000). Nonrandom extinction and the loss of evolutionary history. *Science*, 288 : 328–330.

Rao C. (1952). *Advanced statistical methods in biometric research*. Wiley, New York.

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

Rao C. R. (1982a). Diversity and dissimilarity coefficients : a unified approach. *Theoretical Population Biology*, 21 : 24–43.

Rao C. R. (1982b). Diversity : its measurement, decomposition, apportionment and analysis. *Sankhya : The Indian Journal of Statistics*, A44 : 1–22.

Rao C. R. (1982c). Gini-simpson index of diversity : a characterization, generalization and applications. *Utilitas Mathematics*, 21 : 273–282.

Rao C. R. (1986). Rao's axiomatization of diversity measures. In Kotz S. et Johnson N., editors, *Encyclopedia of Statistical Sciences*, volume 7, pages 614–617. Wiley and Sons, New York.

Rao C. R. et Boudreau R. (1984). Diversity and cluster analyses of blood group data on some human populations. In Chakravarti A., editor, *Symposium on human population genetics : the Pittsburgh symposium. 1982*, pages 277–296. Van Nostrand Reinhold, University of Pittsburgh.

Rao C. R. et Nayak T. K. (1985). Cross entropy, dissimilarity measures, and characterizations of quadratic entropy. *IEEE Transactions on Information Theory*, IT-31 : 589–593.

Rüber L. et Adams D. C. (2001). Evolutionary convergence of body shape and trophic morphology in Cichlids from Lake Tanganyika. *Journal of Evolutionary Biology*, 14 : 325–332.

Reist-Marti S. B., Simianer H., Gibson J., Hanotte O. et Rege J. (2003). Weitzman's approach and conservation of breed diversity : an application to African Cattle Breeds. *Conservation Biology*, 17 : 1299–1311.

Reynolds J., Weir B. et Cockerham C. (1983). Estimation of the coancestry coefficient : basis for a short-term genetic distance. *Genetics*, 105 : 767–779.

Richmond J. Q. et Reeder T. W. (2002). Evidence for parallele ecological speciation in scincid lizards of the Eumeces Skiltonianus species group (Squamata : Scincidae). *Evolution*, 56 : 1498–1513.

Ricotta C. (2002). Bridging the gap between ecological diversity indices and measures of biodiversity with shannon's entropy : comment to Izsák and Papp. *Ecological Modelling*, 152 : 1–3.

Ricotta C. (2003). Additive partition of parametric information and its associated beta-diversity measure. *Acta Biotheoretica*, 51 : 91–100.

Ricotta C. (2004). A parametric diversity measure combining the relative abundances and taxonomic distinctiveness of species. *Diversity and Distributions*, 10 : 143–146.

Ricotta C. (2005). A non-probabilistic information-theoretical framework for multiscale landscape analysis. In *90th ESA Annual Meeting - IX International Congress of Ecology*, Montréal, Canada.

Ricotta C. et Avena G. C. (2003). An information-theoretical measure of taxonomic diversity. *Acta Biotheoretica*, 51 : 35–41.

Rogers D. et Tanimoto T. (1960). A computer program for classifying plants. *Science*, 132 : 1115–1118.

Rogers J. S. (1972). Measures of genetic similarity and genetic distance. In Wheeler M. R., editor, *Studies in Genetics VII*, volume 7213, pages 145–153. The University of Texas Publication, Austin.

ROMAN G., EMERSON L. et FAIRWEATHER K. (2001). *Forest fragmentation and biodiversity conservation : case studies of Costa Rica and Vancouver Island*. Envr 400 thesis, UBC Faculty of Science.

ROOT R. B. (2001). Guilds. In LEVIN S., editor, *Enclopedia of Biodiversity*, volume 3, pages 295–302. Academic Press.

ROTHSTEIN S. I. (1973). Relative variation of avian morphological characters : relation to the niche. *The American Naturalist*, 107 : 796–799.

RUSSELL J., FULLER J., MACAULAY M., HATZ B., JAHOOR A., POWELL W. et WAUGH R. (1997). Direct comparaison of levels of genetic variation among barley accessions detected by RFLPs, AFLPs, SSRs and RAPDs. *Theoretical and Applied Genetics*, 95 : 714–722.

RUSSELL P. et RAO T. (1940). On habitat and association of species of anopheline larvae in south-eastern Madras. *Journal of the Malaria Institute, India*, 3 : 153–178.

SAITOU N. et NEI M. (1987). The neighbor-joining method : a new method for reconstructiong phylogenetic trees. *Molecular Biology and Evolution*, 4 : 406–425.

SARKAR S. et MARGULES C. (2002). Operationalizing biodiversity for conservation planning. *Journal of Biosciences*, 27 : 299–308.

SCHLUTER D. (2000a). Ecological character displacement in adaptative radiation. *The American Naturalist*, 156 : S4–S16.

SCHLUTER D. (2000b). Introduction to the symposium : species interactions and adaptative radiation. *The American Naturalist*, S156 : S1–S3.

SCHLUTER D. et RICKLEFS R. (1993). Convergence and regional component of species diversity. In RICKLEFS R. et SCHLUTER D., editors, *Species diversity in ecological communities : historical and geographical perspectives*, pages 230–242. The University of Chicago Press, Chicago.

SCHNEIDER S., ROESSLI D. et EXCOFFIER L. (2000). *Arlequin Ver 2.000 : A software for population genetics data analysis*. Genetics and Biometry Laboratory, Dept. of Anthropology, University of Geneva, Switzerland.

SCHWARTZ M. et SIMBERLOFF D. (2001). Taxon size predicts rates of rarity in vascular plants. *Ecology Letters*, 4 : 464–469.

SHANNON C. E. (1948). A mathematical theory of communication. *Bell System technical journal*, 27 : 379–423, 623–656.

SHIMATANI K. (2001). On the measurement of species diversity incorporating species differences. *Oikos*, 93 : 135–147.

SIMIANER H. (2002). Noah's dilemna : which breeds to take aboard the ark ? In ORGANISING COMMITTEE, editor, *7th World Congress on Genetics Applied to Livestock Production*, volume 33, pages Communication 26–02, Montpellier, France. INRA, CIRAD, France.

SIMIANER H., MARTI S., GIBSON J., HANOTTE O. et REGE J. (2003). An approach to the optimal allocation of conservation funds to minimize loss of genetic diversity between livestock

breeds. *Ecological Economics*, 45 : 377–392.

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

Simpson G. G. (1961). *Principles of animal taxonomy*. Columbia University Press, New York.

Slatkin M. (1995). A measure of population subdivision based on microsatellite allele frequencies. *Genetics*, 139 : 457–462.

Smith E. P., Pontasch K. W. et Cairns, John J. (1990). Community similarity and the analysis of multispecies environmental data : a unified statistical approach. *Water Research*, 24 : 507–514.

Smith W. (1995). Discussion about the paper entitled "Application of the quadratic entropy indices for diversity studies of drosophilid assemblages" written by Izsák, J. and Papp, L. and published in Environmental and Ecological Statistics, 2 : 213-224. *Environmental and Ecological Statistics*, 2 : 221–222.

Sneath P. H. A. et Sokal R. R. (1973). *Numerical taxonomy*. W. H. Freeman and company, San Francisco.

Sokal R. et Michener C. (1957). The effects of different numerical techniques on the phenetic classification of bees of the Hoplitis complex (Megachilidae). *Proceedings of the Linnean Society of London*, 178 : 59–74.

Sokal R. et Sneath P. (1963). *Principles of numerical taxonomy*. Witt. Freeman and Co., San Francisco.

Solow A. et Polasky S. (1994). Measuring biological diversity. *Environmental and Ecological Statistics*, 1 : 95–107.

Solow A., Polasky S. et Broadus J. (1993). On the measurement of biological diversity. *Journal of Environmental Economics and Management*, 24 : 60–68.

Southwood R. et Henderson P. (2000). *Ecological methods*. Blackwell Science, Oxford.

Sørensen T. (1948). A method for establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Biologiske Skrifter*, 5 : 1–34.

Stander M. (1970). *Diversity and similarity of benthic fauna of Oregon*. M. Sc. thesis, Oregon State University.

Staudhammer C. et LeMay V. (2001). Introduction and evaluation of possible indices of stand structural diversity. *Canadian Journal of Forest Research - Revue Canadienne De Recherche Forest*, 31 : 1105–1115.

Stewart, C.N. J. et Excoffier L. (1996). Assessing population genetic structure and variability with RAPD data : application to *Vaccinium macrocarpon* (American Cranberry). *Journal of Evolutionary Biology*, 9 : 153–171.

Swingland I. (2001). Biodiversity, definition of. In Levin S., editor, *Enclopedia of Biodiversity*, volume 1, pages 381–. Academic Press.

Tajima F. et Nei M. (1984). Estimation of evolutionary distance between nucleotide sequences. *Molecular Biology and Evolution*, 1 : 269–285.

Tamura K. (1992). Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. *Molecular Biology and Evolution*, 9 : 678–687.

Tamura K. et Nei M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in Humans and Chimpanzees. *Molecular Biology and Evolution*, 10 : 512–526.

Templeton A. R., Routman E. et Phillips C. A. (1995). Separating population structure from population history : a cladistic analysis of the geographical distribution of mitochondrial DNA haplotypes in the Tiger Salamander, *Ambystoma tigrinum*. *Genetics*, 140 : 767–782.

ter Braak C. (1986). Canonical correspondence analysis : a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, 67 : 1167–1179.

ter Braak C. (1987). The analysis of vegetation-environment relationships by canonical correspondence analysis. *Vegetatio*, 69 : 69–77.

ter Braak C. (1992). Permutation versus bootstrap significance tests in multiple regression and ANOVA. In Jöckel K.-H., Rothe G. et Sendler W., editors, *Bootstrapping and related techniques*, pages 79–86. Springer-Verlag.

Thaon d'Arnoldi C., Foulley J.-L. et Ollivier L. (1998). An overview of the Weitzman approach to diversity. *Genetics Selection Evolution*, 30 : 149–161.

Turner T. F., Trexler J. C., Harris J. L. et Haynes J. L. (2000). Nested cladistic analysis indicates population fragmentation shapes genetic diversity in a freshwater mussel. *Genetics*, 154 : 777–785.

Van Valen L. (1976). Ecological species, multispecies, and oaks. *Taxon*, 25 : 233–239.

Vane-Wright R., Humphries C. et Williams P. (1991). What to protect ? Systematics and the agony of choice. *Biological Conservation*, 55 : 235–254.

Veech J. A., Summerville K. S., Crist T. O. et Gering J. C. (2002). The additive partitioning of species diversity : recent revival of an old idea. *Oikos*, 99 : 3–9.

Ver Hoef J., Cressie N. et Glenn-Lewin D. (1993). Spatial models for spatial statistics : some unification. *Journal of Vegetation Science*, 4 : 441–452.

Vona G., Ghiani M., Calo C., Vacca L., Memmi M. et Varesi L. (2001). Mitochondrial DNA sequence analysis in Sicily. *American Journal of Human Biology*, 13 : 576–589.

Wagner H. H., Wildi O. et Ewald K. C. (2000). Additive partitioning of plant species diversity in an agricultural mosaic landscape. *Landscape Ecology*, 15 : 219–227.

Warwick R. et Clarke K. (1995). New 'biodiversity' measures reveal a decrease in taxonomic distinctness with increasing stress. *Marine Ecology Progress Series*, 129 : 301–305.

Warwick R. M., Ashman C. M., Brown A. R., Clarke K. R., Dowell B., Hart B., Lewis R. E., Shillabeer N., Somerfield P. et Tapp J. F. (2002). Inter-annual changes in the biodiversity

and community structure of the macrobenthos in Tees Bay and the Tees estuary, UK, associated with local and regional environmental events. *Marine Ecology - Progress Series*, 234 : 1–13.

Warwick R. M. et Clarke K. R. (1998). Taxonomic distinctness and environmental assessment. *Journal of Applied Ecology*, 35 : 532–543.

Warwick R. M. et Clarke K. R. (2001). Practical measures of marine biodiversity based on relatedness of species. *Oceanography and Marine Biology Annual Review*, 39 : 207–231.

Washington H. (1984). Diversity, biotic and similarity indices. A review with special relevance to aquatic ecosystems. *Water Resources*, 15 : 653-694 : 15 : 653–694.

Webb C. O. (2000). Exploring the phylogenetic structure of ecological communities : an exemple for rain forest trees. *The American Naturalist*, 156 : 145–155.

Weir B. (1996). *Genetic data analysis II : Methods for discrete population genetic data*. Sunderland, Massachusetts, Sinauer Associates, Inc. Publishers.

Weir B. et Cockerham C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution*, 38 : 1358–1370.

Weitzman M. L. (1992). On diversity. *The Quarterly Journal of Economics*, 107 : 363–406.

Whittaker R. (1972). Evolution and measurement of species diversity. *TAXON*, 21 : 213–251.

Whittaker R. H. (1960). Vegetation of the Siskiyou Mountains, Oregon and California. *Ecological Monographs*, 30 : 279–338.

Wilhm J. L. (1968). Use of biomass units in Shannon's formula. *Ecology*, 49 : 153–156.

Williams D., Humphries C. et Vane-Wright R. (1991). Measuring biodiversity : taxonomic relatedness for conservation priorities. *Australian Systematic Botany*, 4 : 665–669.

Williams P. et Gaston K. (1994). Measuring more of biodiversity : can higher-taxon richness predict wholesale species richness ? *Biological Conservation*, 67 : 211–217.

Williams P., Gaston K. et Humphries C. (1994). Do conservationists and molecular biologists value differences between organisms in the same way ? *Biodiversity Letters*, 2 : 67–78.

Williams P. et Humphries C. (1994). Biodiversity, taxonomic relatedness, and endemism in conservation. In Forey P., Humphries C. et Vane-Wright R., editors, *Systematics and Conservation Evaluation*, volume 31, pages 269–287. Clarendon, Oxford.

Williams P. et Humphries C. (1996). Comparing character diversity among biotas. In Gaston K., editor, *Biodiversity. A biology of numbers and differences*, pages 54–76. Blackwell Science, Oxford.

Williams P., Vane-Wright R. et Humphries C. (1993). Measuring biodiversity for choosing conservation areas. In LaSalle J. et Gauld I., editors, *Hymenoptera and Biodiversity*, pages 309–328. CAB International, Wallingford, UK.

Willson M. F. (1969). Avian niche size and morphological variation. *The American Naturalist*, 103 : 531–542.

Wilson E. O. (1993). *La diversité de la vie*. Odile Jacob, Paris.

Wilson E. O. (2003). *L'avenir de la vie*. Science ouverte, Seuil.

Wilson E. O. et Postel-Vinay O. (2000). L'enjeu écologique n˚1. *La Recherche*, 333 : 14–16.

Witting L. et Loeschke V. (1995). The optimization of biodiversity conservation. *Biological Conservation*, 71 : 205–207.

Wright S. (1951). The genetical structure of populations. *Annals of Eugenics*, 15 : 323–354.

Yoccoz N. G., Nichols J. D. et Boulinier T. (2001). Monitoring of biological diversity in space and time. *TRENDS in Ecology and Evolution*, 16 : 446–453.

Yule G. (1912). On the methods of measuring association between two attributes. *Journal of the Royal Society*, 75 : 579–642.

Zahl S. (1977). Jackknifing an index of diversity. *Ecology*, 58 : 907–913.

Zhang Q. et Allard R. (1986). Sampling variance of the genetic diversity index. *Journal of Heredity*, 77 : 54–55.

Zhu J., Gale M., Quarrie S., Jackson M. et Bryan G. (1998). AFLP markers for the study of rice biodiversity. *Theoretical and Applied Genetics*, 96 : 602–611.

Les Annexes

Annexe 1
Pavoine *et al.* 2004 - Journal of Theoretical Biology

Pavoine, S., A. B. Dufour, and D. Chessel. 2004. From dissimilarities among species to dissimilarities among communities : a double principal coordinate analysis. Journal of Theoretical Biology 228 :523-537.

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Theoretical Biology 228 (2004) 523–537
Journal of Theoretical Biology
www.elsevier.com/locate/jtbi

From dissimilarities among species to dissimilarities among communities: a double principal coordinate analysis

Sandrine Pavoine*,*, Anne-Béatrice Dufour*, Daniel Chessel*

* *Laboratoire de Biométrie et Biologie Evolutive, UMR CNRS 5558, Université Claude Bernard LYON I, 43, boulevard du 11 novembre 1918, 69622 Villeurbanne Cedex, France*

Received 18 June 2003; received in revised form 28 November 2003; accepted 9 February 2004

Abstract

This paper presents a new ordination method to compare several communities containing species that differ according to their taxonomic, morphological or biological features. The objective is first to find dissimilarities among communities from the knowledge about differences among their species, and second to describe these dissimilarities with regard to the feature diversity within communities. In 1986, Rao initiated a general framework for analysing the extent of the diversity. He defined a diversity coefficient called quadratic entropy and a dissimilarity coefficient and proposed a decomposition of this diversity coefficient in a way similar to ANOVA. Furthermore, Gower and Legendre (1986) built a weighted principal coordinate analysis. Using the previous context, we propose a new method called the double principal coordinate analysis (DPCoA) to analyse the relation between two kinds of data. The first contains differences among species (dissimilarity matrix); the second the species distribution among communities (abundance or presence/absence matrix). A multidimensional space assembling the species points and the community points is built. The species points define the original differences between species and the community points define the deduced differences between communities. Furthermore, this multidimensional space is linked with the diversity decomposition into between-community and within-community diversities. One looks for axes that provide a graphical ordination of the communities and project the species onto them. An illustration is proposed comparing bird communities which live in different areas under mediterranean bioclimates. Compared to some existing methods, the double principal coordinate analysis can provide a typology of communities taking account of an abundance matrix and can include dissimilarities among species. Finally, we show that such an approach generalizes some of these methods and allows us to develop new analyses.

Keywords: Dissimilarity; Diversity; Quadratic entropy; PCoA

1. Introduction

Studying communities for species composition is a central topic in ecology. According to Gittins (1985), when several communities are compared, we can study (1) the relationships between variables characterizing species and variables characterizing communities, (2) the structure, i.e. the way the data set is organized or built. Communities are defined as collections of species found in the same habitat. The study of relationship was well discussed by Dolédec et al. (1996); Legendre et al. (1997) and Legendre and Legendre (1998, p. 565–574). The study of the structure is the aim of this paper and a new method based on Rao's axiomatization is proposed.

The data are composed of two matrices (Fig. 1): the first contains distances or dissimilarities between species; the second contains abundance (or presence/absence) of species (row) in communities (column). Differences among species are assessed, for example, according to their taxonomy (Izsak and Papp, 1995; Warwick and Clarke, 1995), their morphology (Blondel et al., 1984; Cody and Mooney, 1978; Losos, 1992), or their biological traits (Lamouroux et al., 2002). The dissimilarities are evaluated using indices either directly from practical or experimental observations or indirectly from an observed data matrix.

Many studies raise the question of distinguishing differences between communities from differences

*Corresponding Author. Tel.: 33-4-72-43-27-57; Fax: 33-4-72-43-13-88.
E-mail addresses: pavoine@biomserv.univ-lyon1.fr (S. Pavoine), dufour@biomserv.univ-lyon1.fr (A.-B. Dufour), chessel@biomserv.univ-lyon1.fr (D. Chessel).

0022-5193/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2004.02.014

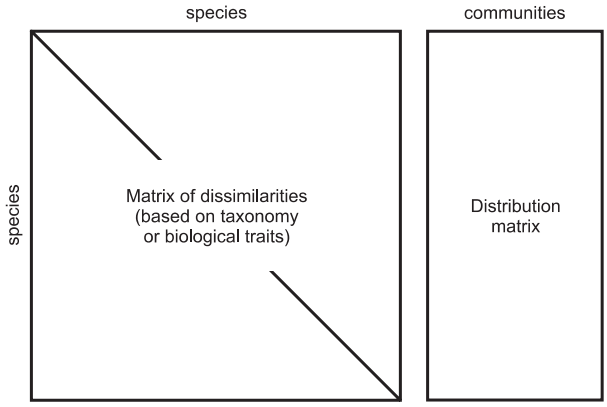


Fig. 1. Original data.

between species (Lande, 1996; Whittaker, 1972). The species diversity indices are usually calculated as the relationship of the number of species (richness) to the number of individuals per species (abundance) for a given community. The most common species indices are Gini-Simpson diversity (Gini, 1912; Simpson, 1949) and Shannon information (Shannon, 1948). However, in using species diversity, they did not calculate the differences between species. We define the term "feature diversity" to denote indices including differences among species in one or several biological traits. While properties and the biological meaning of Shannon information have been argued by many authors (see for instance Hurlbert, 1971; Lande, 1996; Rao, 1982), many developments of Gini–Simpson index have been made (e.g. Hendrickson and Ehrlich, 1971; Rao, 1982; Warwick and Clarke, 1995). These authors make innovations by introducing differences between species in the within-community diversity measures.

A new approach is also possible. Gimaret-Carpentier et al. (1998) and Pélissier et al. (2003) recently showed that ordination techniques and several usual species diversity measurements could be related. They demonstrated that inertia measurements for correspondence analysis and non-symmetric correspondence analysis equate common species diversity indices such as species richness, Gini–Simpson diversity and Shannon information. Our subject matter is in line with these ordination methods.

A new method, called the double principal coordinate analysis (DPCoA), is proposed. Its main objective is to obtain a community typology from the heterogeneity of species identities, but also from differences between species and from the relative abundances of each species. First of all, we present the origin of the (DPCoA) using Rao's axiomatization (1986). We compare this with other ordination methods, revealing its originality. Finally, we give an ecological example: comparison of bird communities in three regions under mediterranean bioclimates and a control region under temperate bioclimate (Blondel et al., 1984).

2. Rao's axiomatization

For years, the question of diversity measurements has been producing an incredible variety of solutions coming from ecology, population biology, genetics, molecular biology, and now molecular ecology. This is why a general framework must be defined. Rao (1986) was the first to begin such research. He characterized the measure of diversity of a distribution, defined a diversity coefficient called quadratic entropy and a dissimilarity coefficient. Finally, he proposed a decomposition of the diversity coefficient in a way similar to ANOVA (Fisher, 1925).

2.1. The measure of diversity of a distribution

The definition of the measure of diversity of a distribution comes from Rao's axiomatization (1986), based on a convex set $\mathbb{P}$ of probability distributions, i.e.

$$\forall P\in\mathbb{P}, \forall Q\in\mathbb{P}, \forall \alpha\in[0,1], \quad (\alpha P+(1-\alpha)Q)\in\mathbb{P}.$$

Let $\mathbb{P}$ be the following convex set of frequency distributions:

$$\mathbb{P}=\left\{P=(P_1,\ldots,P_n), P_k\geqslant 0, \sum_{k=1}^{n} P_k=1\right\}.$$

To be characterized as a measure of diversity of a distribution, H, a real-valued function defined on $\mathbb{P}$, has to verify at least two axioms:

First H must be nonnegative,

$$H(P)\geqslant 0 \quad \forall P\in\mathbb{P}.$$

And secondly H must be concave,

$$\forall P\in\mathbb{P},\ \forall Q\in\mathbb{P},\ \forall \mu_1\in\mathbb{R}^+,\ \forall \mu_2\in\mathbb{R}^+,\ \mu_1+\mu_2=1, \quad H(\mu_1 P+\mu_2 Q)\geqslant \mu_1 H(P)+\mu_2 H(Q).$$

In concrete terms, this last property of H means that diversity increases by mixing.

2.2. The diversity and dissimilarity coefficients

Consider r communities and n different species. The frequency distribution of the species in the community j is denoted by the probability vector $\mathbf{p}_j=(p_{1/j},\ldots,p_{n/j})\left(\mathbf{p}_j\in\mathbb{P}\right)$. Rao (1982) defined a diversity coefficient (DIVC), also called quadratic entropy, by

$$H_{\mathbf{\Delta}_n}(\mathbf{p}_j)=\sum_{k=1}^{n}\sum_{l=1}^{n} p_{k/j}p_{l/j}\delta_{kl}^{SP}. \tag{1}$$

$\mathbf{\Delta}_n=\left[\delta_{kl}^{SP}\right]_{1\leqslant k\leqslant n,1\leqslant l\leqslant n}$ is the $n\times n$ symmetric matrix containing dissimilarities between species, where $\delta_{kk}^{SP}=0$ for all k and $\delta_{kl}^{\mathrm{SP}}>0$ for all k and $l\neq k$. $\delta_{kl}^{\mathrm{SP}}$ is a conditionally negative definite function so that $H_{\mathbf{\Delta}}$ is concave (Rao, 1986).

Rao's DIVC may be considered as an expansion of the Gini-Simpson index (Gini, 1912; Simpson, 1949). In fact, if $\delta_{kl}^{SP} = 1$ for all $l \neq k$, then

$$H_{\Delta_n}(\mathbf{p}_j) = \sum_{k=1}^{n} \sum_{l=1, l\neq k}^{n} p_{k/j} p_{l/j} = 1 - \sum_{k=1}^{n} p_{k/j}^2,$$

is the Gini–Simpson index. In the literature, a great number of diversity indices which have been calculated from a distance or dissimilarity matrix can be translated in DIVC form. For example, we find that the Hendrickson and Ehrlich index (1971) is a non-biased version of Rao's DIVC. In fact, it corresponds to Rao's DIVC multiplied by a constant depending on the total number of species. The Warwick and Clarke index (1995) called taxonomic diversity turns out to be a special use of the Hendrickson and Ehrlich index with an arbitrary taxonomic dissimilarity.

Recently many authors have suggested the use of Rao's DIVC in order to assess the diversity within communities, taking into account differences between species. Izsak and colleagues (Izsak and Papp, 1995, 2000; Izsak and Szeidl, 2002) proposed to use Rao's index by integrating taxonomic dissimilarities between species. These dissimilarities are arbitrarily obtained from a taxonomic tree and are equal to those proposed by Warwick and Clarke (1995). In microbial ecology, Watve and Gangal (1996) suggested that indices should consider taxonomic dissimilarities. From this prospect, they introduced an index derived from Rao's DIVC that includes genetic dissimilarities between pairs of isolates. In ecology, Shimatani (2001) suggested the use of Rao's index by integrating between-species taxonomic dissimilarities and dissimilarities based on the study of amino acids. He outlined the link between Rao's index and Warwick and Clarke's index. By applying this index to tree populations with the goal of observing the effects of thinning operation for promoting survival of specific desirable species, he concluded that it can be expected that biodiversity indices incorporating species differences have more applications in ecology.

Rao (1982) went a step further by suggesting a unifying approach of diversity and dissimilarity measures. He introduced a dissimilarity coefficient (DISC) between two communities i and j with the respective species frequency vectors $\mathbf{p}_i$ and $\mathbf{p}_j$:

$$D_{H_{\Delta_n}}\left(\mathbf{p}_i, \mathbf{p}_j\right) = 2H_{\Delta_n}\left(\frac{\mathbf{p}_i + \mathbf{p}_j}{2}\right) - H_{\Delta_n}(\mathbf{p}_i) - H_{\Delta_n}\left(\mathbf{p}_j\right). \tag{2}$$

2.3. Decomposition of quadratic entropy

Rao (1982, 1984, 1986) then proposed a decomposition of quadratic entropy in a way similar to ANOVA. Let $\boldsymbol{\mu} = (\mu_1, ..., \mu_r) (\boldsymbol{\mu} \in \mathbb{P})$ be the community weight vector. Let $\mathbf{p}_\bullet = (p_{1\bullet}, ..., p_{n\bullet})$ $(\mathbf{p}_\bullet = \sum_{i=1}^{r} \mu_i \mathbf{p}_i,\ (\mathbf{p}_\bullet \in \mathbb{P}))$ be the species frequency vector in mixed communities, i.e. in the whole data set. This decomposition is

$$H_{\Delta_n}(\mathbf{p}_\bullet) = \sum_{i=1}^{r} \mu_i H_{\Delta_n}(\mathbf{p}_i) + \sum_{i=1}^{r} \sum_{j=1}^{r} \mu_i \mu_j D_{H_{\Delta_n}}\left(\mathbf{p}_i, \mathbf{p}_j\right).$$

The total quadratic entropy $H_{\Delta_n}(\mathbf{p}_\bullet)$ is divided into the within-community quadratic entropy (the first term) and the between-community quadratic entropy (second term).

The interest of these two partitions is to assess the part of the total quadratic entropy due to differences among the communities compared to the differences among the species within communities. These measures can be quantified, for example, to identify communities of high conservation value. Such communities have a great internal diversity and are very different from the remaining communities.

3. Description of the DPCoA

3.1. Entries

Recall that the data to be analysed arise in two matrices.

Consider first $\mathbf{A} = \left[a_{kj}\right]_{1\leqslant k\leqslant n, 1\leqslant j\leqslant r}$, in which a_{kj} is the abundance of the species k in the community j and $\mathbf{P} = \left[p_{kj}\right]_{1\leqslant k\leqslant n, 1\leqslant j\leqslant r}$, in which p_{kj} is the frequency of the species k in the community j. These matrices are linked by the following relations:

$$\left.\begin{array}{c} a_{\bullet j} = \sum_{k=1}^{n} a_{kj} \\ a_{k\bullet} = \sum_{j=1}^{r} a_{kj} \\ a_{\bullet\bullet} = \sum_{j=1}^{r} a_{\bullet j} = \sum_{k=1}^{n} a_{k\bullet} \end{array}\right\} \Rightarrow \left\{\begin{array}{l} p_{\bullet j} = a_{\bullet j}/a_{\bullet\bullet}, \\ p_{k\bullet} = a_{k\bullet}/a_{\bullet\bullet}, \\ p_{kj} = a_{kj}/a_{\bullet\bullet}. \end{array}\right.$$

Let $\mathbf{D}_n = diag(p_{1\bullet}, ..., p_{n\bullet})$ and $\mathbf{D}_r = diag(p_{\bullet 1}, ..., p_{\bullet r})$ be the diagonal matrices containing the marginal weighting associated with $\mathbf{P}$. Species and communities each have a natural weighting matrix, $\mathbf{D}_n$ and $\mathbf{D}_r$, respectively.

Consider secondly $\boldsymbol{\Delta}_n = \left[\delta_{kl}^{SP}\right]_{1\leqslant k\leqslant n, 1\leqslant l\leqslant n}$ the $n \times n$ matrix containing dissimilarities between species. We choose a Euclidean matrix because this type of matrix is associated with a typology and because $\left(\delta_{kl}^{SP}\right)^2$ is a conditionally negative definite function. We show the importance of this last property in section 6. $\boldsymbol{\Delta}_n$ is said to be Euclidean if and only if n points M_k $(k = 1, 2, ..., n)$ can be embedded in a Euclidean space such that the Euclidean distance between M_k and M_l is δ_{kl}^{SP}, i.e. $\delta_{kl}^{SP} = |M_k - M_l|$ (Gower and Legendre, 1986). This, of course, implies that δ_{kl}^{SP} must be nonegative. We consider that calling $\boldsymbol{\Delta}_n = \left[\delta_{kl}^{SP}\right]$ Euclidean is synonymous with stating that δ_{kl}^{SP} has Euclidean properties. These two expressions are used in the following process.

3.2. Building a common space

The n points M_k can be obtained by a principal coordinate analysis (PCoA). In PCoA, a projector $\mathbf{Q}$ centers the scatter of points. This projector is usually equal to $\mathbf{Q} = \mathbf{I}_n - (1/n)\mathbf{1}_n\mathbf{1}_n^t$, and gives a uniform weighting to the species. From now on, the weighting inserted in the centering projector is arbitrary (Gower, 1982, 1984; Gower and Legendre, 1986), and leads to a weighted PCoA. The theoretical reasons for this freedom were discussed by d'Aubigny (1989), but do not seem to have provided any concrete result. We decided to use this weighting to DPCoA.

Let us denote $\mathbf{Q} = \mathbf{I}_n - \mathbf{1}_n\mathbf{1}_n^t\mathbf{D}_n$ the new projector, and $\mathbf{\Omega}_n = \left[\left(\delta_{kl}^{SP}\right)^2/2\right]_{1\leqslant k\leqslant n, 1\leqslant l\leqslant n}$, where $\mathbf{I}_n$ is the $n\times n$ identity matrix, and $\mathbf{1}_n$ is the unit n-vector. The following matrix $-\mathbf{D}_n^{1/2}\mathbf{Q}\mathbf{\Omega}_n\mathbf{Q}^t\mathbf{D}_n^{1/2}$ is built. Let $\mathbf{\Lambda}$ and $\mathbf{U}$ be the eigenvalues and eigenvectors of this matrix. We can write

$$
\begin{aligned}
&-\mathbf{D}_n^{1/2}\mathbf{Q}\mathbf{\Omega}_n\mathbf{Q}^t\mathbf{D}_n^{1/2} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^t\\
&\Rightarrow -\mathbf{Q}\mathbf{\Omega}_n\mathbf{Q}^t = \mathbf{D}_n^{-1/2}\mathbf{U}\mathbf{\Lambda}\mathbf{U}^t\mathbf{D}_n^{-1/2}\\
&= \mathbf{D}_n^{-1/2}\mathbf{U}\mathbf{\Lambda}^{1/2}\left(\mathbf{D}_n^{-1/2}\mathbf{U}\mathbf{\Lambda}^{1/2}\right)^t = \mathbf{X}\mathbf{X}^t.
\end{aligned}
$$

The rows of the obtained matrix $\mathbf{X}$ give the coordinates of the species. The Euclidean distance between rows k and l of $\mathbf{X}$ provides exactly δ_{kl}^{SP}. Therefore, the communities may be represented by points whose coordinates are given by $\mathbf{Y} = \mathbf{D}_r^{-1}\mathbf{P}^t\mathbf{X}$. According to $\mathbf{Y}$, communities are placed on the barycenter of their species points. Furthermore, the coordinates of species are then $\mathbf{D}_n$-centered and coordinates of the communities are $\mathbf{D}_r$-centered (Appendix A). We call this space shared by species and communities the "space of the double principal coordinate analysis".

3.3. Defining a typology

The principal axes of the species points enable us to obtain a typology of the species with a reduced number of dimensions. To obtain a typology of the communities with a reduced number of dimensions, we look for the orthogonal principal axes of the scatter of community points. Let $\mathbf{I}_f$ be the $f\times f$ identity matrix. The generalized singular value decomposition (GSVD) (Greenacre, 1984) of the triplet $(\mathbf{Y}, \mathbf{I}_f, \mathbf{D}_r)$, i.e. the PCA of $\mathbf{Y}$ weighted by $\mathbf{D}_r$, gives the principal axes of community points. These axes are contained in a $f\times g$ matrix $\mathbf{V}$ defined by

$$\mathbf{Y}^t\mathbf{D}_r\mathbf{Y} = \mathbf{X}^t\mathbf{P}\mathbf{D}_r^{-1}\mathbf{P}^t\mathbf{X} = \mathbf{V}\mathbf{\Psi}\mathbf{V}^t,$$

where $\mathbf{\Psi}$ contains the eigenvalues of $\mathbf{Y}^t\mathbf{D}_r\mathbf{Y}$. Each of these axes sequentially explains as much of the variance of the community points as possible and the amount of the variance explained by an axis is given by its associated eigenvalue. We choose two (or more) of these principal axes and project on them the species and community points as well as the principal axes of the scatter of species points, so that their coordinates are given by the rows of $\mathbf{XV}$, $\mathbf{YV}$, and $\mathbf{I}_f\mathbf{V} = \mathbf{V}$, respectively.

This methodology leads to the representations of the dissimilarities between communities on which species are positioned.

4. DPCoA and apportionment of quadratic entropy

4.1. Distances among the community points

Changing $\mathbf{\Delta}_n = \left[\delta_{kl}^{SP}\right]_{1\leqslant k\leqslant n, 1\leqslant l\leqslant n}$ for $\mathbf{\Omega}_n = \left[\left(\delta_{kl}^{SP}\right)^2/2\right]_{1\leqslant k\leqslant n, 1\leqslant l\leqslant n}$ leads to the following diversity index according to Rao's DIVC:

$$H_{\mathbf{\Omega}_n}(\mathbf{p}_j) = \frac{1}{2}\sum_{k=1}^{n}\sum_{l=1}^{n} p_{k/j}p_{l/j}\left(\delta_{kl}^{SP}\right)^2 = \mathbf{p}_j^t\mathbf{\Omega}_n\mathbf{p}_j. \tag{3}$$

Let $\mathbf{\Delta}_r = \left[\delta_{ij}^{CO}\right]_{1\leqslant i\leqslant r, 1\leqslant j\leqslant r}$ be the matrix containing dissimilarities between the communities and $\mathbf{\Omega}_r = \left[\left(\delta_{ij}^{CO}\right)^2/2\right]_{1\leqslant i\leqslant r, 1\leqslant j\leqslant r}$. We define δ_{ij}^{CO} by

$$\delta_{ij}^{CO} = \sqrt{2\left(2H_{\mathbf{\Omega}_n}\left(\frac{\mathbf{p}_i+\mathbf{p}_j}{2}\right) - H_{\mathbf{\Omega}_n}(\mathbf{p}_i) - H_{\mathbf{\Omega}_n}\left(\mathbf{p}_j\right)\right)}. \tag{4}$$

This expression can be rewritten with matrices:

$$\delta_{ij}^{CO} = \sqrt{\left(\mathbf{p}_i - \mathbf{p}_j\right)^t(-\mathbf{\Omega}_n)\left(\mathbf{p}_i - \mathbf{p}_j\right)}. \tag{5}$$

If $\mathbf{\Delta}_n = \left[\delta_{kl}^{SP}\right]_{1\leqslant k\leqslant n, 1\leqslant l\leqslant n}$ is a Euclidean matrix, then $H_{\mathbf{\Omega}_n}$ is concave (Rao and Nayak, 1985) and $\mathbf{\Delta}_r = \left[\delta_{ij}^{CO}\right]_{1\leqslant i\leqslant r, 1\leqslant j\leqslant r}$ is Euclidean (Champely and Chessel, 2002). The concavity assures that $H_{\mathbf{\Omega}_n}$ can be partitioned. Moreover δ_{ij}^{CO} is the Euclidean distance between the point of community i and the point of community j.

4.2. Decomposition of inertia

Let $\mathbf{\mu} = (p_{\bullet 1}, \ldots, p_{\bullet r})(\mathbf{\mu}\in\mathbb{P})$ be the community weight vector. The quadratic entropy decomposition becomes

$$H_{\mathbf{\Omega}_n}(\mathbf{p}_\bullet) = \sum_{i=1}^{r} p_{\bullet i}H_{\mathbf{\Omega}_n}(\mathbf{p}_i) + H_{\mathbf{\Omega}_r}(\mathbf{\mu}). \tag{6}$$

where $H_{\mathbf{\Omega}_r}(\mathbf{\mu}) = \sum_{i=1}^{r}\sum_{j=1}^{r} p_{\bullet i}p_{\bullet j}\left(\delta_{ij}^{CO}\right)^2/2$. The overall quadratic entropy, $H_{\mathbf{\Omega}_n}(\mathbf{p}_\bullet)$, is divided into within-community entropy, $\sum_{i=1}^{r} p_{\bullet i}H_{\mathbf{\Omega}_n}(\mathbf{p}_i)$, and between-community entropy, $H_{\mathbf{\Omega}_r}(\mathbf{\mu})$. Since $\mathbf{\Delta}_r$ is Euclidean,

$H_{\mathbf{\Omega}_r}$ is concave. Therefore $H_{\mathbf{\Omega}_r}$ is actually a measure of diversity.

The diversity $H_{\mathbf{\Omega}_n}(\mathbf{p}_j)$ within a community j is the inertia (variance) of the species points weighted by $\mathbf{p}_j$ in the space of the DPCoA. The overall diversity $H_{\mathbf{\Omega}_n}(\mathbf{p}_\bullet)$ is the inertia of all the species points weighted by $\mathbf{p}_\bullet$ in the space of the DPCoA. And finally the between-community diversity $H_{\mathbf{\Omega}_r}(\boldsymbol{\mu})$ is the inertia of all the community points weighted by $\boldsymbol{\mu}$ in the space of the DPCoA (Appendix B).

By selecting two principal axes of the points of the communities in the DPCoA space, we obtain a two-dimensional typology of the communities and the species are positioned on this typology. These two axes explain a great part of the between-community diversity. We then give an index of the diversity within a community on the typology by an ellipse. Its center is at the community point and its amplitude depends on the positions of the species and the abundance of the species in the community.

5. Relationships between DPCoA and other ordination methods

The multivariate ordinations which are concerned with data structure can maximize the variance among communities, and allow the projection of the species on the typology of the communities. Four methods are studied in this paper: the canonical variate analysis or discriminant analysis (Fisher, 1936; Rao, 1952), the canonical analysis of principal coordinates used with a discriminant analysis (Anderson and Willis, 2003), the canonical correspondence analysis (CCA) (ter Braak, 1986, 1987) and the between-class principal component analysis (BPCA) (Dolédec and Chessel, 1987). We present succinctly these methods, their relationships and the relationships between these methods and our DPCoA.

5.1. Special use of the four methods connected with DPCoA

For a start, the canonical variate analysis or discriminant analysis (CVA) and the BPCA are used in their original meanings. The other two methods are presented in a transformed way. The CCA is usually performed from a relevé-species matrix and a relevé-environmental variables matrix (ter Braak, 1986, 1987). The goal of CCA is to maximize the variance between species by choosing ordination axes that are linear combinations of environmental variables. In this paper, CCA is performed from a species-community matrix and a species-biological variables matrix. The goal is then to maximize the variance between communities by choosing axes that are linear combinations of biological variables. The canonical analysis of principal coordinates (CAP) combined with a discriminant analysis is usually performed from a matrix of distances or dissimilarities among sites and a partition of sites into groups. Its goal is to maximize the variance between groups. Here we perform a specific application from a matrix of dissimilarities among species and a partition of species into communities.

5.2. Relationship between the four methods

The four above methods (CVA, CAP, CCA and BPCA) can be interconnected in three different ways. First, they can be decomposed into two families: the canonical (CVA-DA, CCA, CAP) and the between-class (BPCA) methods. The canonical methods studied here eliminate the redundancy between the variables characterizing species whereas the BPCA does not eliminate this redundancy. We note that BPCA is equal to redundancy analysis which is also called principal component analysis in respect to instrumental variables (Israels, 1984; Johansson, 1981; Rao, 1964; van den Wollenberg, 1977) when the explanatory variables are categorical. Secondly, all these methods differ in the original data form they consider. In CVA, CAP and BPCA, the species are divided into communities. This means that the occurrence rather than the abundance of the species is considered and that one species cannot belong to more than one community. The matrix that tells which community each species belongs to is called the indicator matrix. We distinguish this matrix from the abundance (or presence/absence) matrix. In fact, with the abundance matrix, occurrences as well as abundances of the species may be regarded and any species may belong to more than one community. Abundance matrix is used in CCA. Finally, in CVA, CCA and BPCA, the species are characterized by several variables put in a matrix for which a principal component analysis would be an appropriate separated analysis. Distances between species are implicitly computed from these matrices. Conversely, in CAP, species are distinguished with the help of any distance or dissimilarity matrix.

5.3. Affinities with DPCoA

We now study the relationships between these four methods and DPCoA. Consider that the data on the features of the species are organized in a matrix with species (row) and Gaussian variables (column). We perform an indirect distance matrix from these variables. In fact we compute the principal component analysis of the variable matrix, get the matrix containing the standardized coordinates of the species, and then compute the Euclidean distances between the rows of this matrix. The resulting distance metric eliminates the redundancy between the Gaussian variables. It is a Mahalanobis metric. We then perform a DPCoA on the resulting Mahalanobis distance matrix and the indicator

matrix. This process is exactly equal to CVA (Appendix C1).

Consider now that we have a between-species dissimilarity matrix. Suppose that we perform the PCoA of this matrix and keep some resulting orthonormal axes. If we compute the Euclidean distances between the standardized coordinates of the species on the retained axes (instead of raw species coordinates), we obtain a decorrelated dissimilarity matrix according to Anderson and Willis (2003). Performing DPCoA on the resulting distance matrix and the matrix of indicator variables is exactly similar to the computing of CAP on the raw distance matrix and the indicator matrix (Appendix C2).

Assume that data are organized in a matrix with species (row) and biological variables (column). If we simply compute the Euclidean distances between the rows of this matrix and perform DPCoA on the resulting distance matrix and the indicator matrix, the process is exactly equal to BPCA (Appendix C3).

Finally, the matrix giving species composition within communities contains abundance or presence/absence variables, i.e. a correspondence analysis would be appropriate for an ordination of this matrix. Consider that the data about the features of the species are organized in a matrix with species (row) and biological variables (column). If we compute the Mahalanobis distances between the rows of this matrix, weighting each species by its abundance in the whole data set, and perform DPCoA on the resulting weighted Mahalanobis distance matrix and the matrix of abundance (or presence/absence) variables, the process is exactly equal to CCA. The community coordinates are then linear combinations of the biological variables (Appendix C4).

All the cited ordination analyses are thus special applications of DPCoA. The advantage of DPCoA is to allow the choice of appropriate dissimilarities or distances between the species, not only Euclidean, weighted or unweighted Mahalanobis distances. The appropriate dissimilarity or distance matrix may be decorrelated or not. The main interest of DPCoA with regard to these ordination methods appears when the data lead directly to between-species distances without considering feature variables. Taxonomic or phylogenetic distances are good examples. Furthermore, our method also allows the choice of an appropriate distribution matrix: indicator matrix or abundance matrix (Table 1).

Table 1
The relationships between four ordination methods and the double principal coordinate analysis

	Dissimilarity between species	Distribution matrix of species into communities
Canonical analyses		
CAP	Euclidean properties	Indicator
CVA-DA	Mahalanobis	Indicator
CCA	Weighted Mahalanobis	Abundance
Between-class analyses		
BPCA	Euclidean	Indicator

5.4. Two other methods

Some other methods found in the literature can be linked to DPCoA. For instance, we present the non-symmetric correspondence analysis (NSCA) (Lauro and D'Ambra, 1984) and the distance-based redundancy analysis (dbRDA) (Legendre and Anderson, 1999).

In NSCA, the data are made up of a species × communities matrix. There are no distances between species. Correspondence analysis studies contingency tables or tables containing abundances (or presence/absence). We can analyse the simultaneous discrimination of the communities and the species. The result is a compromise between these two discriminations. NSCA carries out only one of the discriminations. And our interest focuses on the discrimination of the communities. So, NSCA is equal to DPCoA with an artificial matrix containing equal distances between species (Appendix C5).

In dbRDA, the data are made up of sites and groups. Ter Braak (1986) defines a site as a basic sampling unit, separated in space or time from other sites. A group is a set of sites characterized by geography or experimentation. Two matrices are built: a dissimilarity matrix for sites and an indicator matrix giving which site belongs to which group. The objective is different from ours. In fact, this method is focused on the statistical test of the difference between the groups. It performs a decomposition of diversity, whose components are used to build F-like statistics. This decomposition is the partition of the inertia in the space of the redundancy analysis applied on the principal coordinates of the dissimilarity matrix and on the indicator matrix. The decomposition of the diversity given by DPCoA is in this case that provided by the distance-based redundancy analysis (Legendre and Anderson, 1999) when only one factor is of concern (Appendix C6).

6. Illustration: bird communities and mediterranean area

Computations and graphical displays were done with R (Ihaka and Gentleman, 1996). The "dpcoa" function and the data called "ecomor" are available in the ade4 package (http://cran.r-project.org).

In the framework of the evolutionary convergence paradigm which was very popular in the seventies,

Blondel et al. (1984) evaluated the soundness of the concept of ecomorphological convergence by studying bird communities living in different parts of the world but in similar types of environments, namely three regions under mediterranean bioclimates: central Chile, California (United States), and Provence (France). These regions are compared to a control region under temperate bioclimate: Burgundy (France). In each region, these authors determined four habitats corresponding to a vegetation gradient. They tried to find a precise correspondence between equivalent habitats among the four regions in terms of structure, height and physiognomy of vegetation.

6.1. Previous studies

Blondel et al. (1984) took four kinds of information: the species composition for each community (i.e. in each habitat of each region), the foraging sites, the diet habits, and the morphometric characteristics of each species. They concluded that a morphometric convergence was not controversial on the scales of species and guilds. But, using discriminant analysis, on the scale of communities, they found that the phylogenetic relationship between the species of Burgundy and Provence seemed stronger than a hypothetical mediterranean convergence.

Schluter and Ricklefs (1993) reanalysed these data in another way. They studied the number of species found in diet categories in the three mediterranean regions by using a test similar to an analysis of variance in order to find possible effects of categories and of regions. They found a strong level of convergence of the number of species per diet category for the three mediterranean regions. By adding Burgundy to their analysis, we found that the level of convergence for the four regions was still strong (unpublished data) indicating that the nature of the convergence is not climatic.

6.2. Description of the data

We will not reopen here the question of ecomorphological convergence. To illustrate our method, we choose to analyse data structure dealing with the taxonomy, morphology and foraging substrate of the species. We denote **A** the matrix giving the species present in each community and **P** the corresponding frequency matrix. Only French regions share species. But habitats within a region share many species. The data we use here are slightly different from those of Blondel et al. (1984). Blondel informed us that the repartition of the species, shared by the two European regions, among the habitats is different according to the region. We took this fact into account.

To compute taxonomic dissimilarities between species, we assign the value 1 between two species of the same genus, 2 between two species of different genera belonging to the same family, 3 for two species of different families belonging to the same order, 4 for two species belonging to different orders. Pairwise dissimilarities between species build the following matrix $\left[\left(\delta_{kl}^{SP}\right)_{TAX}^{2}/2\right]$ so that $\mathbf{\Delta}_n^{TAX} = \left[\left(\delta_{kl}^{SP}\right)_{TAX}\right]$ is Euclidean and the diversity index given by formula (3) is exactly the index of taxonomic diversity proposed both by Izsak and Papp (1995) and Warwick and Clarke (1995).

To compute substrate dissimilarities between species, we resolve foraging sites into the six following modalities: aerial, foliage feeders, twig feeders, bush feeders, trunk feeders, ground feeders. For each species, a percentage is assigned to each modality, according to its affinities. From these percentages, we calculate the Edwards distance (Edwards, 1971) between species. The resulting distance matrix is denoted $\mathbf{\Delta}_n^{SUB}$.

To compute morphometric dissimilarities between species, we studied the following traits: wing length, tail length, total culmen length, bill height, bill width, tarsus length, length of middle toe and claw, and weight. Pairwise dissimilarities between species are estimated by working out Mahalanobis distances on the mean logarithmic morphometric measures of the species. The resulting distance matrix is denoted $\mathbf{\Delta}_n^{MOR}$.

6.3. Decomposition of diversity

For each biological and ecological trait, we compute a within-community diversity coefficient with formula (3) and the apportionment of the diversity coefficient with formula (6). On the whole, species richness, taxonomic diversity, foraging-substrate diversity and morphometric diversity increase with the complexity of the vegetation. More precisely, there are two exceptions. First, the species richness within Burgundy is roughly constant along the gradient of habitats. Second, the morphometric diversity within California decreases with the complexity of the vegetation. We also note that the within-community morphometric diversity converges in the complex habitats for the four regions (Fig. 2).

Decomposition of the diversity coefficient (Table 2) shows that the differences between communities correspond to 5%, 8% and 9% of the total diversity according to the taxonomy, morphology and foraging substrate, respectively. To test the significance of these differences, we compute the following permutation test. For each iteration, we perform a random permutation of each species occurrence across communities and calculate a statistic corresponding to the division of between-community diversity by within-community diversity. After 999 iterations, we find that only 3% of the random values are superior to the observed value with taxonomy, and less than 1% with morphology and foraging substrate. We conclude that differences

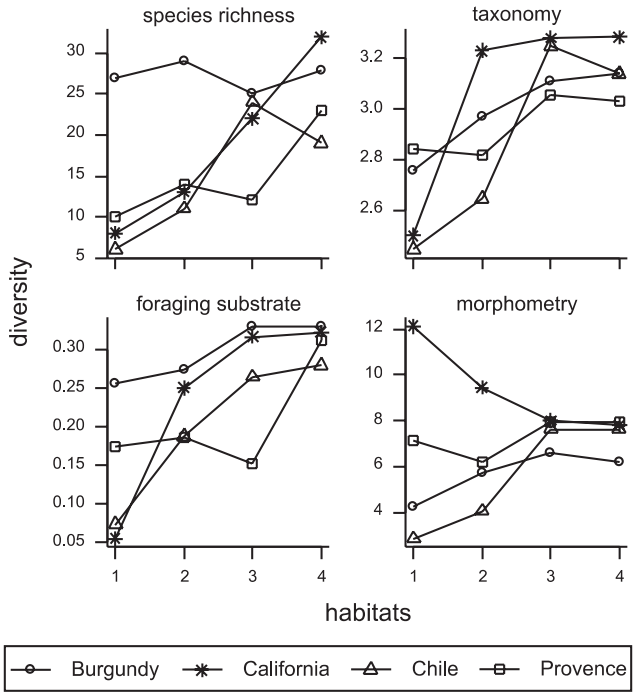


Fig. 2. Diversity patterns along the habitat gradient within regions. Four criteria are concerned: species richness, taxonomic diversity, foraging-substrate diversity and morphometric diversity.

Table 2
Apportionment of quadratic entropy

Diversity source	Taxonomy		Morphology		Foraging substrate	
Among communities	0.176	(5%)	0.587	(8%)	0.027	(9%)
Within communities	3.039	(95%)	6.893	(92%)	0.267	(91%)
Total	3.215		7.480		0.294	

between communities are significant for all dissimilarity criteria.

We also compute a matrix containing the Euclidean Jaccard dissimilarities (Gower and Legendre, 1986; Jaccard, 1901) between the rows of the presence/absence matrix **A** and perform the procruste (Jackson, 1995) and the RV tests (Heo and Gabriel, 1998) on this matrix and $\mathbf{\Delta}_n^{TAX}$, $\mathbf{\Delta}_n^{MOR}$, and $\mathbf{\Delta}_n^{SUB}$, respectively. The aim is to assess the coherence between the species structure given by **A** and the species structure given by $\mathbf{\Delta}_n^{TAX}$, $\mathbf{\Delta}_n^{MOR}$, and $\mathbf{\Delta}_n^{SUB}$, respectively. For both tests and the three dissimilarity criteria, the coherence is highly significant.

6.4. Community typology

The above results reveal differences between the communities and the DPCoA allows us to describe them. We detail the results for the taxonomy. The data are summarized in Fig. 3. The objective is to put the species occurrence matrix **A** (left part of Fig. 3) in order according to the species taxonomy (right part of Fig. 3).

We compute a DPCoA on **P** and $\mathbf{\Delta}_n^{TAX}$ (Fig. 4). The four plots in Fig. 4 are superimposable. They focus either on species points (Fig. 4(a) and (b)), on community points (Fig. 4(d)), or simultaneously on species and community points (Fig. 4(c)). The first two axes of community points explain 36% and 17% of the between-community taxonomic diversity, respectively. Fig. 4(a) and (b) inform us about the role of the species on the typology of the communities. Only three families contribute to it: Sylvidae, Emberizidae, and Picidae. Fig. 4(c) contains both species and community points and shows the distribution ellipses centered on community points. These ellipses indicate the area in which species from a given community are likely to be located. Traits also connect a community point to its species. Fig. 4(d) shows that the first axis highly separates the European communities from the American communities. And the second axis distributes communities according to the gradient of vegetation complexity. It is worth noticing that the gradient of vegetation in Burgundy extends the gradient of vegetation in Provence. So superimposing Fig. 4(b) and (d) indicates that species from Sylvidae are mainly found in the studied European regions, species from Emberizidae in the studied American regions and species from Picidae in the communities with complex vegetation regardless of the region.

We have also performed the DPCoA on P and the dissimilarity matrices: morphometry ($\mathbf{\Delta}_n^{MOR}$), foraging substrate ($\mathbf{\Delta}_n^{SUB}$) and a new matrix of equidistances between species ($\mathbf{\Delta}_n^{EQU}$). For each trait studied, we found a different pattern of differences between communities. The structure given by Fig. 5(a) is trivial. The regions are well separated, except for Burgundy and Provence which have many species in common. The foraging substrate (Fig. 5(b)) shows the stratification of the vegetation. As underlined in Fig. 4(d), this plot suggests that the gradient of vegetation in Burgundy follows the gradient of vegetation in Provence. The morphometry (Fig. 5(c)) separates the four regions and particularly the continents. Finally the taxonomy (Fig. 5(d)) simultaneously separates the continent and shows the vegetation structure.

This illustration shows the importance of integrating the differences between species into diversity indices. The DPCoA expresses the modification of the diversity pattern according to the nature of the differences between the concerned species. This illustration also shows that DPCoA can compare communities with no species in common, as for instance those of California and Chile.

7. Conclusion

In this paper, we propose a new method, the double principal coordinate analysis (DPCoA), based on a

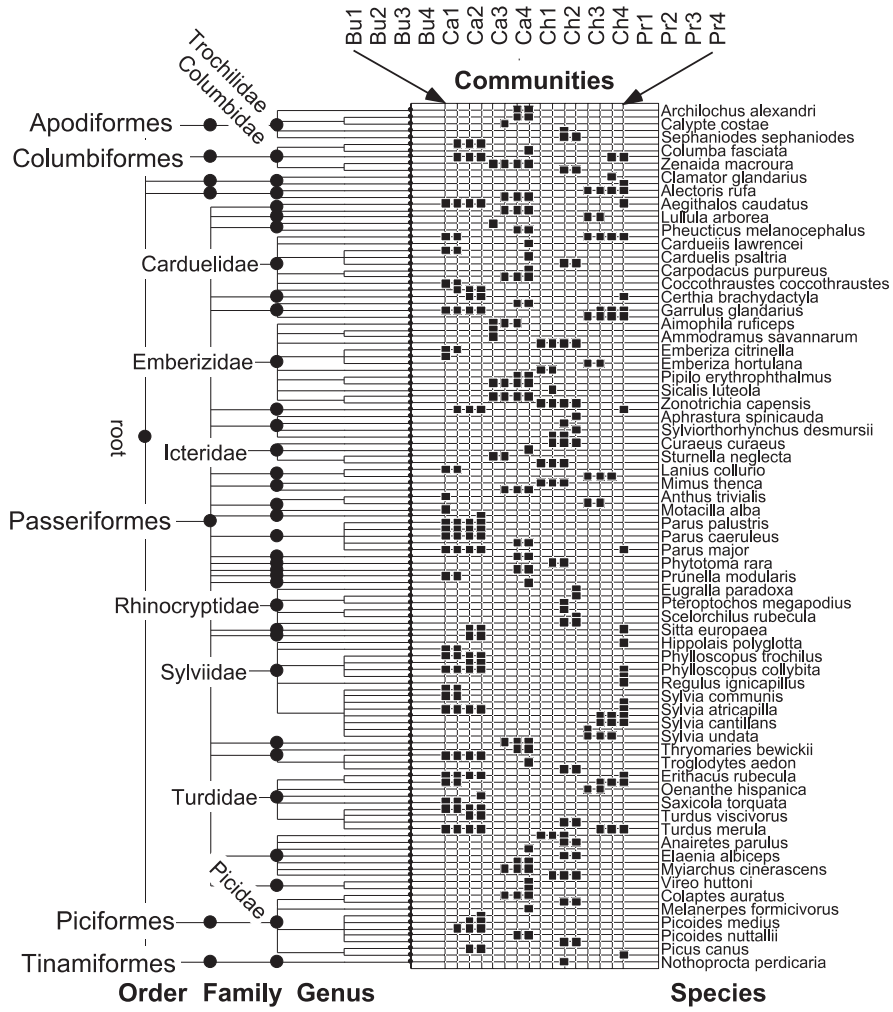


Fig. 3. Summary of the data set. On the right a representation of the species occurrence matrix with species as rows and communities as columns. Each community is labeled by the first two letters of its region name (Bu, Ca, Ch, Pr) and its rank along the gradient of increasing complex vegetation structure (1,2,3,4). For the sake of legibility, one species name out of two is indicated. A black square indicates the presence of a species. On the left: the associated species taxonomic tree.

weighted PCoA, which enables us to compare communities from two kinds of data: a matrix giving the abundance or presence/absence of the species within the communities and a matrix containing distances or dissimilarities between the species. The dissimilarity matrix is obtained either directly from experimental observations or indirectly from an observed data matrix.

We link this new method to four ordination methods. We show that the canonical correspondence analysis (CCA) and the canonical variate analysis (CVA) are particular applications of DPCoA when a Mahalanobis metric is used. This metric is interesting when correlated Gaussian variables are under study. The same is true for the between-class principal component analysis (BPCA), which is under the same pattern as CVA but is based on Euclidean distance. DPCoA allows us to choose the appropriate metric to compute distances. For example, when categorical variables are of concern, dissimilarity indices based on frequencies may be more appropriate (Manly, 1994, formula 5.8 p. 68). When a direct distance or dissimilarity matrix is under study, Anderson and Willis (2003) propose using PCoA as a first step before a CVA. We state that this same process may be used before a CCA with a weighted PCoA. These methods eliminate the redundancy in the knowledge obtained about species. This operation is not always interesting. Anderson and Willis (2003) state that the canonical analysis of principal coordinates (CAP), which uses any distance or dissimilarity matrix, takes into account the correlation structure in the response data cloud; but it provides us no information concerning the overall pattern of dispersion points in the multivariate cloud, or potential differences in multivariate variability, or dispersion among groups. Conversely, the principal component analysis or the principal coordinate analysis for the study of one matrix and BPCA for the study of two matrices is interested in the description of the overall raw data. The DPCoA thus enables us to

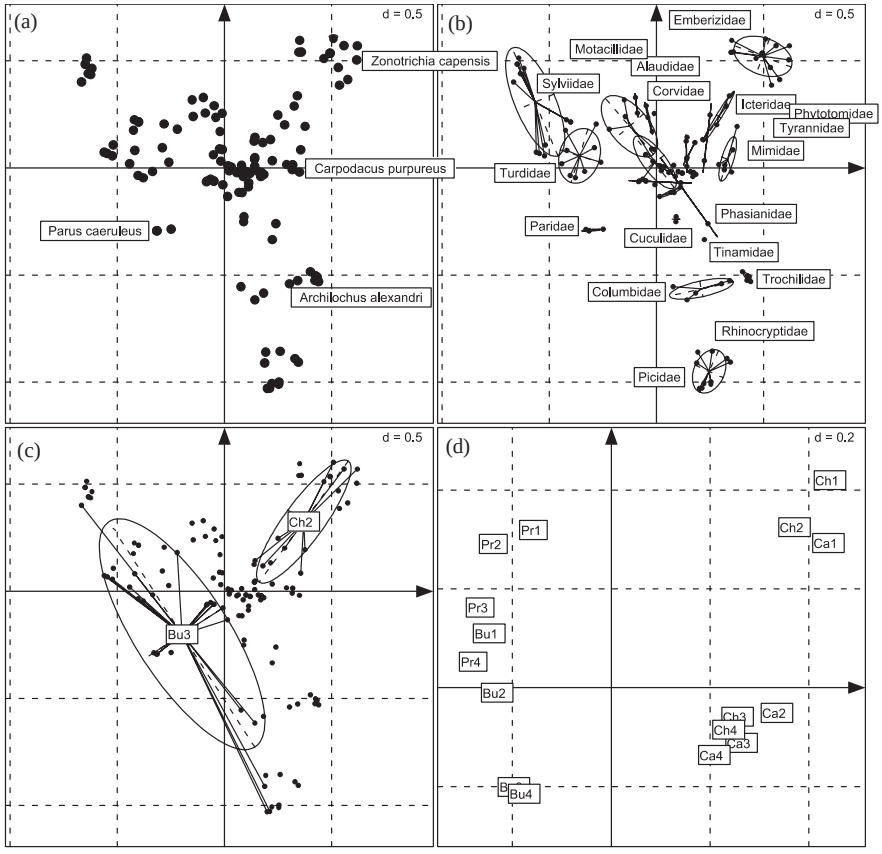


Fig. 4. DPCoA on **P** and Δ_n^{TAX}: (a) species points; (b) species points grouped by families; (c) species points grouped for two communities; (d) community points. In each figure, a grid indicates the scale; the length of a square side is indicated by the d value. Each community is labeled by the first two letters of its region's name (Bu, Ca, Ch, Pr) and its rank along the gradient of increasing complex vegetation structure (1,2,3,4). All the scatters are superimposable (by adjusting the scale if necessary). Distribution ellipses are centered on (b) family points or (c) community points. They give an index of the species distribution for each community or family. Lines connect (b) family or (c) community to their species.

develop the BPCA principle for (1) a matrix giving the abundance (or presence/absence) of the species within communities instead of an indicator matrix and (2) a matrix containing distances or dissimilarities between the species.

The second interest of the developed DPCoA is to generate all its possible extensions. In fact, the organization, showed in Table 1, allows a natural expression of new methods. In the scheme of canonical analyses, taking account of an abundance matrix and a Euclidean matrix for the distance between species leads to a new method: the CCA of weighted principal coordinate. This is equal to the computation of CCA after a weighted PCoA.

Furthermore, in the scheme of between-class analyses, three new methods can be exposed. A between-class analysis of principal coordinates can be built using Euclidean properties for distances between species and an indicator distribution matrix. A between-class correspondence analysis can be built with weighted Euclidean distance between species and an abundance distribution matrix. Likewise, a between-class correspondence analysis of weighted principal coordinates can be built with Euclidean properties for the distance between species and an abundance distribution matrix. This last method can be viewed as an extension of the between-class analysis of principal coordinates for abundance matrices instead of indicator matrices.

In this paper, we perform a descriptive study of the diversity focusing on the graphic display. DPCoA takes into account the dissimilarities between species and describes the diversity of a community (or a site) and the differences between two communities. With the same goal as Legendre, Anderson and McArdle (Anderson, 2001; Legendre and Anderson, 1999; McArdle and Anderson, 2001), we are now able to test the effects of a factor on the differences between communities by focusing on the analytical aspect instead of the graphic aspect. In this case, we have the possibility of taking into account the differences between the species in measurements of differences between communities.

The DPCoA, with related methods, can improve the analyses of diversity by providing further information concerning the overall pattern of dispersion between communities. This means that they describe the differences and the relationships among communities by

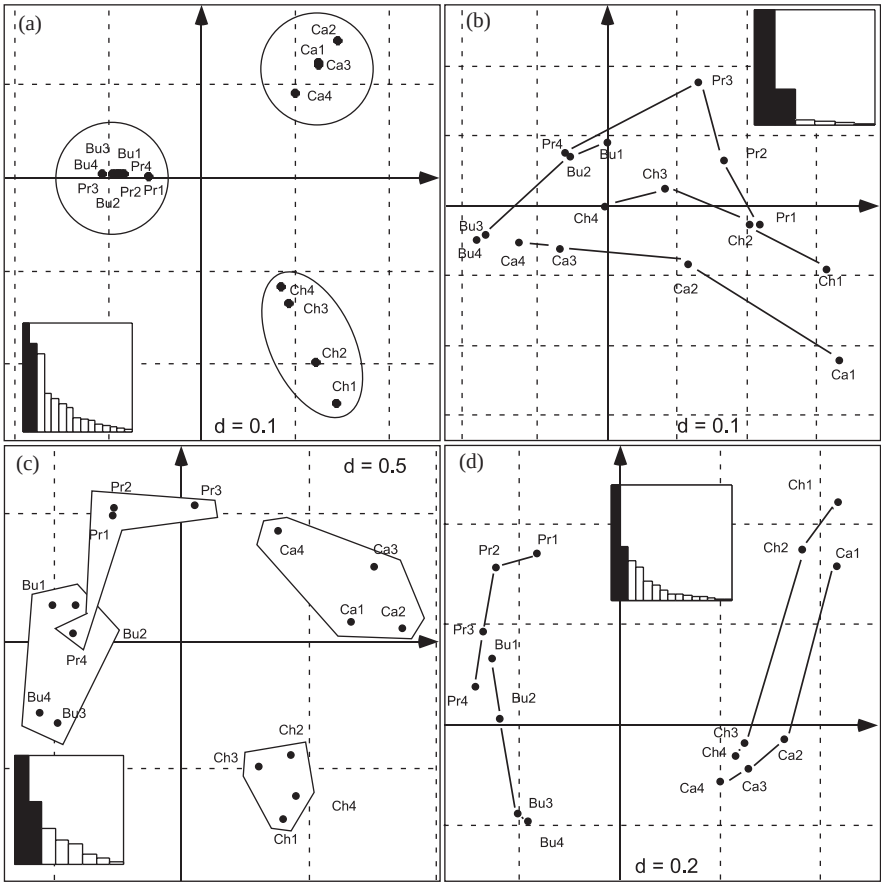


Fig. 5. Community points on the plane given by the first two principal axes of the DPCoA on **P** and (a) Δ_n^{EQU}; (b) Δ_n^{SUB}; (c) Δ_n^{MOR}, (d) Δ_n^{TAX}. In each figure, a grid indicates the scale; the length of a square side is indicated by the *d* value. Each community is labeled by the first two letters of its region's name (Bu, Ca, Ch, Pr) and its rank along the gradient of increasing complex vegetation structure (1,2,3,4). In Fig. 6(b) and (d), lines connect habitats from the same regions. In Fig. 6(a) and (c), ellipses or polygons surround the communities of a single region. In each figure, a box gives barplot of eigenvalues (in black, the retained axes).

showing the degree of overlap between communities in terms of features of species (taxonomy, morphology or otherwise). They also supply information about the role of each species in the differences among communities. Endemic species with special features are highlighted. Another important point is that the method is flexible enough to allow its use in any scope of study: for examples, biology, ecology, economics, and sociology. In population genetics, a similar concern arises (e.g. Nei, 1987; Weir, 1996; Weir and Cockerham, 1984; Wright, 1951, 1965). Several populations of a species may be compared according to genetic traits of their individuals (DNA sequence, fingerprinting pattern, or microsatellites). The DPCoA can provide interesting concrete results for the future.

Acknowledgements

The authors would like to thank C. Ter Braak and two anonymous referees for their relevant comments, J. Blondel for helpful discussions about data sets and R. Grantham for his councils on editorial quality. They all contributed to improving the presentation of this paper.

Appendix A. Centering points in DPCoA

*A.1. Proof that coordinates of the species (***X***) are* $\mathbf{D}_n$*-centered*

$$\left|\mathbf{X}^t\mathbf{D}_n\mathbf{1}_n\right|^2 = \mathbf{1}_n^t\mathbf{D}_n\mathbf{X}\mathbf{X}^t\mathbf{D}_n\mathbf{1}_n = -\mathbf{1}_n^t\mathbf{D}_n\mathbf{Q}\boldsymbol{\Omega}_n\mathbf{Q}^t\mathbf{D}_n\mathbf{1}_n$$
$$= trace\left(-\mathbf{1}_n\mathbf{1}_n^t\mathbf{D}_n\mathbf{Q}\boldsymbol{\Omega}\mathbf{Q}^t\mathbf{D}_n\right) = 0.$$

A.2. Proof that coordinates of the communities ($\mathbf{Y} = \mathbf{D}_r^{-1}\mathbf{P}^t\mathbf{X}$) *are* $\mathbf{D}_r$*-centered*

$$\mathbf{Y}^t\mathbf{D}_r\mathbf{1}_r = \mathbf{X}^t\mathbf{P}\mathbf{1}_r = \mathbf{X}^t\mathbf{D}_n\mathbf{1}_n.$$

Appendix B. Decomposition of the inertia in DPCoA

The link between this decomposition and the DPCoA can be made in the space we call DPCoA space, where

scatters of species points and community points are centered. In this space, we denote M_k the point of the species k, G_j the point of the community j. The scores of M_k are given by the kth row of $\mathbf{X}$ and denoted by $\mathbf{x}_k$. The scores of G_j are given by the jth row of $\mathbf{Y}$ and denoted by $\mathbf{y}_j$. As previously mentioned, G_j is the barycenter of the species points weighted by $\mathbf{p}_j$: $\mathbf{y}_j = \mathbf{X}^t\mathbf{p}_j$. We denote $G_\bullet$ the null coordinate point. $G_\bullet$ is the barycenter of all the species points weighted by $\mathbf{p}_\bullet$, and the barycenter of all the community points weighted by $\boldsymbol{\mu}$.

As previously mentioned, the Euclidean distance between M_k and M_l ($|M_k - M_l| = \sqrt{(\mathbf{x}_k - \mathbf{x}_l)^t(\mathbf{x}_k - \mathbf{x}_l)}$) equals δ_{kl}^{SP}. As $\mathbf{p}_j^t\mathbf{D}_n\mathbf{1}_n = 1$ for all j, formula (5) can be rewritten as

$$\delta_{ij}^{CO} = \sqrt{\left(\mathbf{p}_i - \mathbf{p}_j\right)^t(-\mathbf{Q}\boldsymbol{\Omega}_n\mathbf{Q})\left(\mathbf{p}_i - \mathbf{p}_j\right)}.$$

This can be rewritten as

$$\begin{aligned}\delta_{ij}^{CO} &= \sqrt{\left(\mathbf{p}_i - \mathbf{p}_j\right)^t(\mathbf{X}\mathbf{X}^t)\left(\mathbf{p}_i - \mathbf{p}_j\right)}\\ &= \sqrt{\left(\mathbf{X}^t\mathbf{p}_i - \mathbf{X}^t\mathbf{p}_j\right)^t\left(\mathbf{X}^t\mathbf{p}_i - \mathbf{X}^t\mathbf{p}_j\right)}\\ &= \sqrt{\left(\mathbf{y}_i - \mathbf{y}_j\right)^t\left(\mathbf{y}_i - \mathbf{y}_j\right)}\\ &= \left|G_i - G_j\right|.\end{aligned}$$

The components of diversity can thus be rewritten as follows:

$$\begin{aligned}H_{\boldsymbol{\Omega}_n}(\mathbf{p}_\bullet) &= \frac{1}{2}\sum_{k=1}^{n}\sum_{l=1}^{n}p_{k\bullet}p_{l\bullet}\left(\delta_{kl}^{SP}\right)^2 = \frac{1}{2}\sum_{k=1}^{n}\sum_{l=1}^{n}p_{k\bullet}p_{l\bullet}|M_k - M_l|^2\\ &= \sum_{k=1}^{n}p_{k\bullet}|M_k - G_\bullet|^2,\end{aligned}$$

$$\begin{aligned}H_{\boldsymbol{\Omega}_n}(\mathbf{p}_j) &= \frac{1}{2}\sum_{k=1}^{n}\sum_{l=1}^{n}p_{kj}p_{lj}\left(\delta_{kl}^{SP}\right)^2 = \frac{1}{2}\sum_{k=1}^{n}\sum_{l=1}^{n}p_{kj}p_{lj}|M_k - M_l|^2\\ &= \sum_{k=1}^{n}p_{kj}\left|M_k - G_j\right|^2,\end{aligned}$$

$$\begin{aligned}H_{\boldsymbol{\Omega}_r}(\boldsymbol{\mu}) &= \frac{1}{2}\sum_{i=1}^{r}\sum_{j=1}^{r}p_{\bullet i}p_{\bullet j}\left(\delta_{ij}^{CO}\right)^2 = \frac{1}{2}\sum_{i=1}^{r}\sum_{j=1}^{r}p_{\bullet i}p_{\bullet j}\left|G_i - G_j\right|^2\\ &= \sum_{i=1}^{r}p_{\bullet i}|G_i - G_\bullet|^2.\end{aligned}$$

Appendix C. Clarification of the relationship between DPCoA and other methods

Let $\mathbf{L}$ be an indicator matrix and $\mathbf{A}$ be an abundance (or presence/absence) matrix. Let $\mathbf{Z}$ be a matrix of centered variables. Let $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{V}_*$ be three matrices containing eigenvectors and $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ be two matrices containing eigenvalues. Note that when $\mathbf{L}$ is of concern, we have the following relations: $\mathbf{D}_n = (1/n)\mathbf{I}_n$, $\mathbf{D}_n\mathbf{L} = \mathbf{P}$ and $\mathbf{D}_r = \mathbf{L}^t\mathbf{D}_n\mathbf{L}$.

C.1. CVA

Process: Let $\mathbf{T} = \mathbf{Z}^t\mathbf{D}_n\mathbf{Z}$ be the total covariance matrix and $\mathbf{B} = \mathbf{Z}^t\mathbf{D}_n\mathbf{L}(\mathbf{D}_r)^{-1}\mathbf{D}_r(\mathbf{D}_r)^{-1}\mathbf{L}^t\mathbf{D}_n\mathbf{Z}$ the between-community covariance matrix. The CVA on $\mathbf{Z}$ and $\mathbf{L}$ is the eigenanalysis of $\mathbf{T}^{-1}\mathbf{B}$. They are computational advantages in working with the symmetric matrix $\mathbf{T}^{-1/2}\mathbf{B}\mathbf{T}^{-1/2}$, rather than $\mathbf{T}^{-1}\mathbf{B}$. The eigenvalues of $\mathbf{T}^{-1/2}\mathbf{B}\mathbf{T}^{-1/2}$ are identical to those of $\mathbf{T}^{-1}\mathbf{B}$, while its eigenvectors, $\mathbf{V}$, are connected with those of $\mathbf{T}^{-1}\mathbf{B}$, $\mathbf{V}_0$, by the relation $\mathbf{V}_0 = \mathbf{T}^{-1/2}\mathbf{V}$:

$$\begin{aligned}\mathbf{T}^{-1/2}\mathbf{B}\mathbf{T}^{-1/2} &= \left(\mathbf{Z}^t\mathbf{D}_n\mathbf{Z}\right)^{-1/2}\mathbf{Z}^t\mathbf{D}_n\mathbf{L}(\mathbf{D}_r)^{-1}\mathbf{D}_r(\mathbf{D}_r)^{-1}\\ &\quad\mathbf{L}^t\mathbf{D}_n\mathbf{Z}\left(\mathbf{Z}^t\mathbf{D}_n\mathbf{Z}\right)^{-1/2} = \mathbf{V}\boldsymbol{\Psi}\mathbf{V}^t.\end{aligned}$$

Affinity with DPCoA:

(1) Consider the GSVD of $(\mathbf{Z},\mathbf{I}_m,\mathbf{D}_n)$:

$$\mathbf{Z}^t\mathbf{D}_n\mathbf{Z} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^t.$$

Consider $\mathbf{X} = \mathbf{Z}\mathbf{U}\boldsymbol{\Lambda}^{-1/2}$. The Euclidean distances between the rows of $\mathbf{X}$ are the Mahalanobis distances between the rows of $\mathbf{Z}$. Let $\boldsymbol{\Delta}$ contain these distances so that the species coordinates provided by the PCoA of $\boldsymbol{\Delta}$ weighted by $\mathbf{D}_n$ are given by $\mathbf{X}$.

(2) The matrix $\mathbf{T}^{-1/2}\mathbf{B}\mathbf{T}^{-1/2}$ can be rewritten as

$$\begin{aligned}\mathbf{T}^{-1/2}\mathbf{B}\mathbf{T}^{-1/2} &= \mathbf{U}\boldsymbol{\Lambda}^{-1/2}\mathbf{U}^t\mathbf{Z}^t\mathbf{P}(\mathbf{D}_r)^{-1}\mathbf{D}_r(\mathbf{D}_r)^{-1}\mathbf{P}^t\mathbf{Z}\mathbf{U}\boldsymbol{\Lambda}^{-1/2}\mathbf{U}^t\\ &= \mathbf{U}\mathbf{Y}^t\mathbf{D}_r\mathbf{Y}\mathbf{U}^t.\end{aligned}$$

Consider $\mathbf{V}_* = \mathbf{U}^t\mathbf{V}$, then

$$\mathbf{Y}^t\mathbf{D}_r\mathbf{Y} = \mathbf{V}_*\boldsymbol{\Psi}\mathbf{V}_*^t,$$

which is the GSVD of $(\mathbf{Y},\mathbf{I}_k,\mathbf{D}_r)$ and thus the DPCoA of $\mathbf{P}$ and $\boldsymbol{\Delta}$. The coordinates of the communities are in $\mathbf{Y}\mathbf{V}_* = (\mathbf{D}_r)^{-1}\mathbf{P}^t\mathbf{Z}\left(\mathbf{Z}^t\mathbf{D}_n\mathbf{Z}\right)^{-1/2}\mathbf{V}$, and those of the species are in $\mathbf{X}\mathbf{V}_* = \mathbf{Z}\left(\mathbf{Z}^t\mathbf{D}_n\mathbf{Z}\right)^{-1/2}\mathbf{V}$. The DPCoA on $\mathbf{P}$ and $\boldsymbol{\Delta}$ is thus equal to the CVA on $\mathbf{Z}$ and $\mathbf{L}$.

C.2. CAP

CAP computes CVA after having performed PCoA on a dissimilarity matrix. In order to find its link with DPCoA, in the previous part, replace $\mathbf{Z}$ with a matrix containing m principal coordinates of a dissimilarity matrix.

C.3. BPCA

Process: Let $\hat{\mathbf{Z}}$ be defined as

$$\hat{\mathbf{Z}} = \mathbf{L}\left(\mathbf{L}^t\mathbf{D}_n\mathbf{L}\right)^{-1}\mathbf{L}^t\mathbf{D}_n\mathbf{Z}.$$

The BPCA on $\mathbf{Z}$ and $\mathbf{L}$ is the GSVD of $(\hat{\mathbf{Z}},\mathbf{I}_m,\mathbf{D}_n)$, i.e.

$$\begin{aligned}\hat{\mathbf{Z}}^{\mathrm{t}}\mathbf{D}_n\hat{\mathbf{Z}} &= \mathbf{Z}^{\mathrm{t}}\mathbf{D}_n\mathbf{L}\left(\mathbf{L}^{\mathrm{t}}\mathbf{D}_n\mathbf{L}\right)^{-1}\mathbf{L}^{\mathrm{t}}\mathbf{D}_n\mathbf{L}\left(\mathbf{L}^{\mathrm{t}}\mathbf{D}_n\mathbf{L}\right)^{-1}\mathbf{L}^{\mathrm{t}}\mathbf{D}_n\mathbf{Z}\\ &= \mathbf{Z}^{\mathrm{t}}\mathbf{P}(\mathbf{D}_r)^{-1}\mathbf{D}_r(\mathbf{D}_r)^{-1}\mathbf{P}^{\mathrm{t}}\mathbf{Z}\\ &= \mathbf{V}\boldsymbol{\Psi}\mathbf{V}^{\mathrm{t}}.\end{aligned}$$

Affinity with DPCoA:

(1) $\boldsymbol{\Delta}$ is defined by computing the Euclidean distances between the rows of $\mathbf{Z}$.

(2) The PCoA on $\boldsymbol{\Delta}$ weighted by $\mathbf{D}_n$ is equal to the GSVD of $(\mathbf{Z},\mathbf{I}_m,\mathbf{D}_n)$:

$$\mathbf{Z}^{\mathrm{t}}\mathbf{D}_n\mathbf{Z} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{t}}.$$

The species coordinates are thus given by the rows of $\mathbf{X} = \mathbf{Z}\mathbf{U}$.

(3) The covariance matrix $\hat{\mathbf{Z}}^{\mathrm{t}}\mathbf{D}_n\hat{\mathbf{Z}}$ can be rewritten as

$$\begin{aligned}\hat{\mathbf{Z}}^{\mathrm{t}}\mathbf{D}_n\hat{\mathbf{Z}} &= \mathbf{U}\mathbf{X}^{\mathrm{t}}\mathbf{P}(\mathbf{D}_r)^{-1}\mathbf{D}_r(\mathbf{D}_r)^{-1}\mathbf{P}^{\mathrm{t}}\mathbf{X}\mathbf{U}^{\mathrm{t}}\\ &= \mathbf{U}\mathbf{Y}^{\mathrm{t}}\mathbf{D}_r\mathbf{Y}\mathbf{U}^{\mathrm{t}}.\end{aligned}$$

This implies that

$$\mathbf{Y}^{\mathrm{t}}\mathbf{D}_r\mathbf{Y} = \mathbf{U}^{\mathrm{t}}\mathbf{V}\boldsymbol{\Psi}\mathbf{V}^{\mathrm{t}}\mathbf{U},$$

which is the GSVD of $(\mathbf{Y},\mathbf{I}_k,\mathbf{D}_r)$. So the coordinates of the communities are in $\mathbf{Y}\mathbf{U}^{\mathrm{t}}\mathbf{V} = (\mathbf{D}_r)^{-1}\mathbf{P}^{\mathrm{t}}\mathbf{X}\mathbf{U}^{\mathrm{t}}\mathbf{V} = (\mathbf{D}_r)^{-1}\mathbf{P}^{\mathrm{t}}\mathbf{Z}\mathbf{V}$, and those of the species are in $\mathbf{X}\mathbf{U}^{\mathrm{t}}\mathbf{V} = \mathbf{Z}\mathbf{V}$. The DPCoA on $\mathbf{P}$ and $\boldsymbol{\Delta}$ is thus equal to the BPCA on $\mathbf{Z}$ and $\mathbf{L}$.

C.4. CCA

Process: We define

$$\bar{\mathbf{P}} = \mathbf{D}_n^{1/2}\left(\mathbf{D}_n^{-1}\mathbf{P}\mathbf{D}_r^{-1}\right)\mathbf{D}_r^{1/2}$$

and

$$\hat{\mathbf{P}} = \mathbf{D}_n^{1/2}\mathbf{Z}\left(\mathbf{Z}^{\mathrm{t}}\mathbf{D}_n\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{t}}\mathbf{D}_n^{1/2}\bar{\mathbf{P}}.$$

CCA on $\mathbf{P}$ and $\mathbf{Z}$ is the GSVD of $(\hat{\mathbf{P}}^{\mathrm{t}},\mathbf{I}_n,\mathbf{I}_r)$:

$$\hat{\mathbf{P}}\hat{\mathbf{P}}^{\mathrm{t}} = \mathbf{V}\boldsymbol{\Psi}\mathbf{V}^{\mathrm{t}}.$$

Affinity with DPCoA:

(1) Consider the GSVD of $(\mathbf{Z},\mathbf{I}_m,\mathbf{D}_n)$:

$$\mathbf{Z}^{\mathrm{t}}\mathbf{D}_n\mathbf{Z} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{t}}.$$

Consider $\mathbf{X} = \mathbf{Z}\mathbf{U}\boldsymbol{\Lambda}^{-1/2}$. The Euclidean distances between the rows of $\mathbf{X}$ are Mahalanobis distances between the rows of $\mathbf{Z}$ weighted by $\mathbf{D}_n$. $\boldsymbol{\Delta}$ contains these distances so that the species coordinates provided by the PCoA of $\boldsymbol{\Delta}$ weighted by $\mathbf{D}_n$ are in the rows of $\mathbf{X}$.

(3) Matrix $\hat{\mathbf{P}}$ can be rewritten as

$$\hat{\mathbf{P}} = \mathbf{D}_n^{1/2}\mathbf{Z}\left(\mathbf{Z}^{\mathrm{t}}\mathbf{D}_n\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{t}}\mathbf{D}_n^{1/2}\mathbf{D}_n^{1/2}\left(\mathbf{D}_n^{-1}\mathbf{P}\mathbf{D}_r^{-1}\right)\mathbf{D}_r^{1/2},$$

$$\Leftrightarrow \hat{\mathbf{P}} = \mathbf{D}_n^{1/2}\mathbf{Z}\mathbf{U}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{\Lambda}^{-1/2}\mathbf{U}^{\mathrm{t}}\mathbf{Z}^{\mathrm{t}}\mathbf{P}\mathbf{D}_r^{-1/2},$$

$$\Leftrightarrow \hat{\mathbf{P}} = \mathbf{D}_n^{1/2}\mathbf{X}\mathbf{Y}^{\mathrm{t}}\mathbf{D}_r^{1/2}.$$

The GSVD of $\hat{\mathbf{P}}\hat{\mathbf{P}}^{t}$ is

$$\hat{\mathbf{P}}\hat{\mathbf{P}}^{\mathrm{t}} = \mathbf{D}_n^{1/2}\mathbf{X}\mathbf{Y}^{\mathrm{t}}\mathbf{D}_r^{1/2}\mathbf{D}_r^{1/2}\mathbf{Y}\mathbf{X}^{\mathrm{t}}\mathbf{D}_n^{1/2} = \mathbf{V}\boldsymbol{\Psi}\mathbf{V}^{\mathrm{t}},$$

$$\Leftrightarrow \mathbf{Y}^{\mathrm{t}}\mathbf{D}_r\mathbf{Y} = \mathbf{X}^{\mathrm{t}}\mathbf{D}_n^{1/2}\mathbf{V}\boldsymbol{\Psi}\mathbf{V}^{\mathrm{t}}\mathbf{D}_n^{1/2}\mathbf{X}.$$

Consider $\mathbf{V}_* = \mathbf{X}'\mathbf{D}_n^{1/2}\mathbf{V}$, then

$$\mathbf{Y}^{\mathrm{t}}\mathbf{D}_r\mathbf{Y} = \mathbf{V}_*\boldsymbol{\Psi}\mathbf{V}_*^{\mathrm{t}},$$

which is the GSVD of $(\mathbf{Y},\mathbf{I}_k,\mathbf{D}_r)$ and thus the DPCoA of $\mathbf{P}$ and $\boldsymbol{\Delta}$. The coordinates of the communities are in

$$\begin{aligned}\mathbf{Y}\mathbf{V}_* &= \mathbf{D}_r^{-1}\mathbf{P}\mathbf{X}\mathbf{X}^{\mathrm{t}}\mathbf{D}_n^{1/2}\mathbf{V}\\ &= \mathbf{D}_r^{-1}\mathbf{P}\mathbf{Z}\left(\mathbf{Z}^{\mathrm{t}}\mathbf{D}_n\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{t}}\mathbf{D}_n^{1/2}\mathbf{V},\end{aligned}$$

and those of the species are in

$$\begin{aligned}\mathbf{X}\mathbf{V}_* &= \mathbf{X}\mathbf{X}^{\mathrm{t}}\mathbf{D}_n^{1/2}\mathbf{V}\\ &= \mathbf{Z}\left(\mathbf{Z}^{\mathrm{t}}\mathbf{D}_n\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{t}}\mathbf{D}_n^{1/2}\mathbf{V}.\end{aligned}$$

The DPCoA on $\mathbf{P}$ and $\boldsymbol{\Delta}$ is thus equal to the CCA on $\mathbf{Z}$ and $\mathbf{P}$.

C.5. NSCA

Process: We define

$$\bar{\mathbf{P}} = \mathbf{P}\mathbf{D}_r^{-1} - \mathbf{D}_n\mathbf{1}_n\mathbf{1}_r^{\mathrm{t}} = \left(\mathbf{I}_n - \mathbf{D}_n\mathbf{1}_n\mathbf{1}_n^{\mathrm{t}}\right)\mathbf{P}\mathbf{D}_r^{-1}.$$

The NSCA of $\mathbf{P}$ is the GSVD of $(\bar{\mathbf{P}}^{\mathrm{t}},\mathbf{I}_n,\mathbf{D}_r)$:

$$\bar{\mathbf{P}}\mathbf{D}_r\bar{\mathbf{P}}^{\mathrm{t}} = \mathbf{V}\boldsymbol{\Psi}\mathbf{V}^{\mathrm{t}}.$$

Affinity with DPCoA:

(1) We take a matrix of equidistances among the species, say, $\boldsymbol{\Delta} = \left(\mathbf{1}_n\mathbf{1}_n^{\mathrm{t}} - \mathbf{I}_n\right)*\sqrt{2}$. The multiplicative factor $\sqrt{2}$ leads to a total inertia equal to the between-communities Gini-Simpson diversity.

(2) The PCoA on $\boldsymbol{\Delta}$ weighted by $\mathbf{D}_n$ is equal to the centered PCA of $\mathbf{I}_n$ weighted by $\mathbf{D}_n$ i.e.:

$$\left(\mathbf{I}_n - \mathbf{1}_n\mathbf{1}_n^{\mathrm{t}}\mathbf{D}_n\right)^{\mathrm{t}}\mathbf{D}_n\left(\mathbf{I}_n - \mathbf{1}_n\mathbf{1}_n^{\mathrm{t}}\mathbf{D}_n\right) = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{t}}.$$

The species coordinates are given by $\mathbf{X} = \left(\mathbf{I}_n - \mathbf{1}_n\mathbf{1}_n^{\mathrm{t}}\mathbf{D}_n\right)\mathbf{U}$.

(3) $\bar{\mathbf{P}}\mathbf{D}_r\bar{\mathbf{P}}^{\mathrm{t}}$ can be rewritten as

$$\begin{aligned}\bar{\mathbf{P}}\mathbf{D}_r\bar{\mathbf{P}}^{\mathrm{t}} &= \left(\mathbf{I}_n - \mathbf{D}_n\mathbf{1}_n\mathbf{1}_n^{\mathrm{t}}\right)\mathbf{P}\mathbf{D}_r^{-1}\mathbf{D}_r\mathbf{D}_r^{-1}\mathbf{P}^{\mathrm{t}}\left(\mathbf{I}_n - \mathbf{1}_n\mathbf{1}_n^{\mathrm{t}}\mathbf{D}_n\right)\\ &= \mathbf{U}\mathbf{X}^{\mathrm{t}}\mathbf{P}\mathbf{D}_r^{-1}\mathbf{D}_r\mathbf{D}_r^{-1}\mathbf{P}^{\mathrm{t}}\mathbf{X}\mathbf{U}^{\mathrm{t}}\\ &= \mathbf{U}\mathbf{Y}^{\mathrm{t}}\mathbf{D}_r\mathbf{Y}\mathbf{U}^{\mathrm{t}}.\end{aligned}$$

Consider $\mathbf{V}_* = \mathbf{U}^{\mathrm{t}}\mathbf{V}$

$$\mathbf{Y}\mathbf{D}_r\mathbf{Y} = \mathbf{V}_*\boldsymbol{\Psi}\mathbf{V}_*^{\mathrm{t}},$$

which is the GSVD of $(\mathbf{Y},\mathbf{I}_k,\mathbf{D}_r)$ and thus the DPCoA of $\mathbf{P}$ and $\boldsymbol{\Delta}$. The coordinates of the communities are given by

$$\mathbf{Y}\mathbf{V}_* = (\mathbf{D}_r)^{-1}\mathbf{P}^{\mathrm{t}}\mathbf{X}\mathbf{U}^{\mathrm{t}}\mathbf{V} = \bar{\mathbf{P}}^{\mathrm{t}}\mathbf{V},$$

and those of the species are in

$$\mathbf{X}\mathbf{V}_* = \mathbf{X}\mathbf{U}^{\mathrm{t}}\mathbf{V} = \left(\mathbf{I}_n - \mathbf{1}_n\mathbf{1}_n^{\mathrm{t}}\mathbf{D}_n\right)\mathbf{V}.$$

Then the DPCoA on $\mathbf{P}$ and $\mathbf{\Delta}$ is equal to the NSCA on $\mathbf{P}$.

C.6. dbRDA

Let $\mathbf{\Delta}$ be a matrix of distances among replicates. Let $\mathbf{L}$ be an indicator matrix corresponding to the design of the experiment (only one factor is concerned).

Let $\mathbf{X}$ be the principal coordinates of $\mathbf{\Delta}$. Compute the redundancy analysis (i.e. BPCA) on $\mathbf{X}$ and $\mathbf{L}$. Let $\hat{\mathbf{X}}$ be defined as

$$\hat{\mathbf{X}} = \mathbf{L}\left(\mathbf{L}^{t}\mathbf{D}_{n}\mathbf{L}\right)^{-1}\mathbf{L}^{t}\mathbf{D}_{n}\mathbf{X}.$$

The BPCA on $\mathbf{X}$ and $\mathbf{L}$ is the GSVD of $(\hat{\mathbf{X}},\mathbf{I}_{m},\mathbf{D}_{n})$, i.e.

$$\begin{aligned}\hat{\mathbf{X}}^{t}\mathbf{D}_{n}\hat{\mathbf{X}} &= \mathbf{X}^{t}\mathbf{D}_{n}\mathbf{L}\left(\mathbf{L}^{t}\mathbf{D}_{n}\mathbf{L}\right)^{-1}\mathbf{L}^{t}\mathbf{D}_{n}\mathbf{L}\left(\mathbf{L}^{t}\mathbf{D}_{n}\mathbf{L}\right)^{-1}\mathbf{L}^{t}\mathbf{D}_{n}\mathbf{X}\\ &= \mathbf{X}^{t}\mathbf{P}(\mathbf{D}_{r})^{-1}\mathbf{D}_{r}(\mathbf{D}_{r})^{-1}\mathbf{P}^{t}\mathbf{X}\\ &= \mathbf{Y}^{t}\mathbf{D}_{r}\mathbf{Y}\\ &= \mathbf{V}\mathbf{\Psi}\mathbf{V}^{t}.\end{aligned}$$

Thus, as shown in Appendix C.3, this process is equal to the GSVD of $(\mathbf{Y},\mathbf{I}_{k},\mathbf{D}_{r})$ and so to the DPCoA on $\mathbf{\Delta}$ and $\mathbf{L}$. Matrices $\mathbf{X}$ and $\mathbf{Y}$ provides coordinates in DPCoA space.

The decomposition of the diversity performed in dbRDA uses three components. The first is the inertia of $\mathbf{X}$ (sum of all unconstrained eigenvalues). The second is the inertia of $\mathbf{Y}$ (sum of the canonical eigenvalues). And the last is equal to the inertia of $\mathbf{X}$ minus the inertia of $\mathbf{Y}$. This process is exactly the decomposition of the inertia in the space of the DPCoA applied to $\mathbf{\Delta}$ and $\mathbf{L}$ (cf. Appendix B).

References

Anderson, M.J., 2001. A new method for non-parametric multivariate analysis of variance. Austral Ecol. 26, 32–46.

Anderson, M.J., Willis, T., 2003. Canonical analysis of principal coordinates: a useful method of constrained ordination for ecology. Ecology 84, 511–525.

Blondel, J., Vuilleumier, F., Marcus, L.F., Terouanne, E., 1984. Is there ecomorphological convergence among mediterranean bird communities of Chile, California, and France. In: Hecht, M.K., Wallace, B., MacIntyre, R.J. (Eds.), Evolutionary Biology. Plenum Press, New York, pp. 141–213.

Champely, S., Chessel, D., 2002. Measuring biological diversity using Euclidean metrics. Environ. Ecol. Stat. 9, 167–177.

Cody, M.L., Mooney, H.A., 1978. Convergence versus nonconvergence in mediterranean-climate ecosystems. Annu. Rev. Ecol. Syst. 9, 265–321.

Dolédec, S., Chessel, D., 1987. Rhythmes saisonniers et composantes stationnelles en milieu aquatique I– Description d'un plan d'observation complet par projection de variables. Acta Oecol.—Oecol. Gener. 8, 403–426.

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.

Drouet d'Aubigny, G. 1989. L'analyse multidimensionelle des données de dissimilarité. Thèse de Doctorat, Université Grenoble 1.

Edwards, A.W.F., 1971. Distance between populations on the basis of gene frequencies. Biometrics 27, 873–881.

Fisher, R.A., 1925. Statistical Methods for Research Workers. Oliver & Boyd, Edinburgh.

Fisher, R.A., 1936. The use of multiple measurements in taxonomic problems. Ann. Eugen. 7, 179–188.

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

Gini, C. 1912. Variabilitá e mutabilitá. Studi economicoaguridici delle facoltta di giurizprudenza dell, Universite di Cagliari III, Parte II.

Gittins, R., 1985. Canonical Analysis: A Review with Applications in Ecology. Springer, Berlin.

Gower, J.C., 1982. Euclidean distance geometry. Math. Scientist 7, 1–14.

Gower, J.C., 1984. Distance matrices and their Euclidean approximation. In: Diday, E., Jambu, M., Lebart, L., Pagès, J., Tomassone, R. (Eds.), Data Analysis and Informatics III. Elsevier, North-Holland, Amsterdam, pp. 3–21.

Gower, J.C., Legendre, P., 1986. Metric and Euclidean properties of dissimilarity coefficients. J. Classif. 3, 5–48.

Greenacre, M.J., 1984. Theory and Applications of Correspondence Analysis. Academic Press, London.

Hendrickson, J.A.J., Ehrlich, P.R., 1971. An expanded concept of "species diversity". Notulae Naturae 439, 1–6.

Heo, M., Gabriel, K.R., 1998. A permutation test of association between configurations by means of the RV coefficient. Commun. Stat. — Simul. Comput 27, 843–856.

Hurlbert, S.H., 1971. The non-concept of species diversity: a critique and alternative parameters. Ecology 52, 577–586.

Ihaka, R., Gentleman, R., 1996. R: a language for data analysis and graphics. J. Comput Graph. Stat. 5, 299–314.

Israels, A.Z., 1984. Redundancy analysis for qualitative variables. Psychometrika 49, 346–661.

Izsak, J., Papp, L., 1995. Application of the quadratic entropy indices for diversity studies of drosophilid assemblages. Environ. Ecol. Stat. 2, 213–224.

Izsak, J., Papp, L., 2000. A link between ecological diversity indices and measures of biodiversity. Ecol. Model. 130, 151–156 doi:10.1016/S0304-3800(00)00203-9.

Izsak, J., Szeidl, L., 2002. Quadratic diversity: its maximization can reduce the richness of species. Environ. Ecol. Stat. 9, 423–430.

Jaccard, P., 1901. Distribution de la flore alpine dans le Bassin des Dranses et dans quelques régions voisines. Bull. Soc. Vaudoise Sci. Natur. 37, 241–272.

Jackson, D.A., 1995. Protest: a PROcustean randomization TEST of community environment concordance. Ecosciences 2, 297–303.

Johansson, J.K., 1981. An extension of Wollenberg's redundancy analysis. Psychometrika 46, 93–103.

Lamouroux, N., LeRoy Poff, N., Angermeier, P.L., 2002. Intercontinental convergence of stream fish community traits along geomorphic and hydraulic gradients. Ecology 83, 1792–1807.

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

Lauro, N., D'Ambra, L., 1984. Non-symmetrical correspondence analysis. In: Tomassone, R. (Ed.), Data Analysis and Informatics, III. Elsevier, North-Holland, Amsterdam, pp. 433–446.

Legendre, P., Anderson, M.J., 1999. Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. Ecol. Monogr. 69, 1–24.

Legendre, P., Legendre, L., 1998. Numerical ecology. Elsevier Science BV, Amsterdam.

Legendre, P., Galzin, R., Harmelin-Vivien, M.L., 1997. Relating behavior to habitat: solutions to the fourth-corner problem. Ecology 78, 547–562.

Losos, J.B., 1992. The evolution of convergent structure in Caribbean *Anolis* communities. Syst. Biol. 41, 403–420.

Manly, B.F., 1994. Multivariate Statistical Methods. A primer 2nd Edition.. Chapman & Hall, London.

McArdle, B.H., Anderson, M.J., 2001. Fitting multivariate models to community data: comment on distance-based redundancy analysis. Ecology 82, 290–297.

Nei, M., 1987. Molecular evolutionary genetics. Columbia University Press, New York, NY, USA.

Pélissier, R., Couteron, P., Dray, S., Sabatier, D., 2003. Consistency between ordination techniques and diversity measurements: two strategies for species occurence data. Ecology 84, 242–251.

Rao, C.R., 1952. Advanced Statistical Methods in Biometric Research. Wiley, New York.

Rao, C.R., 1964. The use and interpretation of principal component analysis in applied research. Sankhya A 26, 329–359.

Rao, C.R., 1982. Diversity and dissimilarity coefficients: a unified approach. Theor. Popul. Biol. 21, 24–43.

Rao, C.R., 1984. Convexity properties of entropy functions and analysis of diversity. Inequalities Statist. Probab. 5, 68–77.

Rao, C.R., 1986. Rao's axiomatization of diversity measures. In: Kotz, S., Johnson, N.L. (Eds.), Encyclopedia of Statistical Sciences. Wiley, New York, pp. 614–617.

Rao, C.R., Nayak, T.K., 1985. Cross entropy, dissimilarity measures, and characterizations of quadratic entropy. IEEE Trans. Inf. Theory IT-31, 589–593.

Schluter, D., Ricklefs, R.E., 1993. Convergence and regional component of species diversity. In: Ricklefs, R.E., Schluter, D. (Eds.), Species Diversity in Ecological Communities: Historical and Geographical Perspectives. The University of Chicago Press, Chicago, pp. 230–242.

Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. 27, 379–423 623–656.

Shimatani, K., 2001. On the measurement of species diversity incorporating species differences. Oikos 93, 135–147.

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

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.

van den Wollenberg, A.L., 1977. Redundancy analysis, an alternative for canonical analysis. Psychometrika 42, 207–219.

Warwick, R.M., Clarke, K.R., 1995. New 'biodiversity' measures reveal a decrease in taxonomic distinctness with increasing stress. Mar. Ecol. Prog. Ser. 129, 301–305.

Watve, M.G., Gangal, R.M., 1996. Problems in measuring bacterial diversity and a possible solution. Appl. Environ. Microbiol. 62, 4299–4301.

Weir, B.S., 1996. Genetic Data Analysis II: Methods for Discrete Population Genetic Data. Sinauer Associates, Inc Publishers., Sunderland, MA.

Weir, B.S., Cockerham, C.C., 1984. Estimating F-statistics for the analysis of population structure. Evolution 38, 1358–1370.

Whittaker, R.H., 1972. Evolution and measurement of species diversity. Taxon 21, 213–251.

Wright, S., 1951. The genetical structure of populations. Ann. Eugen. 15, 323–354.

Wright, S., 1965. The interpretation of population structure by F-statistics with special regard to systems of mating. Evolution 19, 395–420.

Annexe 2
Pavoine et Dolédec 2005 - Environmental and Ecological Statistics

Pavoine, S., and S. Dolédec. 2005. The apportionment of quadratic entropy : a useful alternative for partitioning diversity in ecological data. Environmental and Ecological Statistics 12 :125-138.

Environmental and Ecological Statistics **12,** 125–138, 2005

The apportionment of quadratic entropy: a useful alternative for partitioning diversity in ecological data

SANDRINE PAVOINE[1]* and SYLVAIN DOLÉDEC[2]

[1]*Laboratoire de biométrie et biologie évolutive, UMR CNRS 5558, Université Lyon I, 69622, Villeurbanne Cedex, France;*
E-mail: pavoine@biomserv.univ-lyon1.fr
[2]*Laboratoire d'écologie des hydrosystèmes fluviaux UMR CNRS 5023, Université Lyon I, 69622, Villeurbanne Cedex, France*

Received September 2003; Revised September 2004

Many methods that study the diversity within hierarchically structured populations have been developed in genetics. Among them, the analysis of molecular variance (AMOVA) (Excoffier *et al.*, 1992) has the advantage of including evolutionary distances between individuals. AMOVA is a special case of a far more general statistical scheme produced by Rao (1982a; 1986) and called the apportionment of quadratic entropy (APQE). It links diversity and dissimilarity and allows the decomposition of diversity according to a given hierarchy. We apply this framework to ecological data showing that APQE may be very useful for studying diversity at various spatial scales. Moreover, the quadratic entropy has a critical advantage over usual diversity indices because it takes into account differences between species. Finally, the differences that can be incorporated in APQE may be either taxonomic or functional (biological traits), which may be of critical interest for ecologists.

Keywords: dissimilarity, functional diversity, macroinvertebrates, Rao's axiomatization, taxonomy

1. Introduction

Biodiversity means variability of life in all its form, levels and combination including genetic diversity, species diversity and ecosystem diversity (see e.g. Heywood and Watson, 1995). The study of biodiversity thus covers a wide range of disciplines. Within each discipline, several indices or statistical methods have been developed to measure biodiversity. This enormous quantity of specific biodiversity measurements

*Corresponding author.

1352-8505 © 2005 Springer Science + Business Media, Inc.

requires a more general framework since diversity is a property common to any biological element whatever its scale and type. Such a framework was initiated by Rao (1986) who proposed a set of axioms listing the properties required for a given measure to be considered as a measure of diversity.

Several methods dealing with genetic diversity in subdivided populations have been developed (Excoffier, 2001). Every method decomposes genetic diversity into an average genetic diversity within demes and a genetic diversity among demes. If demes are themselves hierarchically structured, genetic diversity within demes can be decomposed as well. Such approaches are currently used for decomposing gene diversity. (Nei, 1973; Weir and Cockerham, 1984; Finkeldey, 1994), nucleotide diversity (Nei and Li, 1979; Nei and Tajima, 1981; Nei and Jin, 1989; Crease *et al*., 1990; Lynch and Crease, 1990; Nei and Miller, 1990; Holsinger and Mason-Gamer, 1996), microsatellite diversity (Slatkin, 1995), and any kind of genetic diversity (Excoffier *et al*., 1992).

Similar to genetic data, ecological data are hierarchically structured according to spatial and temporal scale (Frissell *et al*., 1986; Kolasa, 1989) and taxonomy in case of community ecology. Whittaker (1960, 1972) defined important concepts stating that the total γ-diversity of a region includes two components: α-diversity which represents within-community diversity and β-diversity which characterizes the degree of change in species diversity along environmental gradients. In the traditional multiplicative approach (Whittaker, 1960, 1972), β-diversity is the ratio between total diversity (γ) and α-diversity. Recently many authors stated that the decomposition of the total diversity into within- and between-community diversity should be additive (e.g., Veech *et al*., 2002; Ricotta, 2003), as formerly suggested by Allan (1975) and Pielou (1975) for example. Such an additive partition of diversity can be expanded to various levels of organization and has potential application to multiple scales (Lande, 1996). Moreover it provides commensurable measures of within- and between-community diversity (Wagner *et al*., 2000; Veech *et al*., 2002). Despite the bewildering range of diversity indices, ecologists lack methods that can simultaneously analyze all components of diversity by taking into account both the abundance of the species and the dissimilarities among species.

In this paper, we highlight that Rao's apportionment of quadratic entropy (APQE, Rao, 1982a) is a fundamental basis allowing the partition of diversity suited to any kind of data. Although this method has been used extensively in genetics under the name of AMOVA, it is rather new in Ecology. We illustrate the potential of APQE using an ecological data set including two scales and various types of biological information such as species composition and trait composition.

2. The apportionment of quadratic entropy

We will restrict our approach to hierarchically structured data (e.g., nested sampling design, taxonomic hierarchy). With this type of data, Rao (1986) provided two rules needed to characterize a measure of diversity (H). The first one is that H must be obviously nonnegative. The second one concerns concavity, which means that the diversity in two mixed sets (communities, regions or taxonomic levels, for example)

must be higher than the average diversity within each set in order to avoid negative values for the components of diversity (see also Lande, 1996).

In this context, Rao (1982b, 1984) developed the apportionment of diversity (APDIV) as an appropriate method for partitioning diversity in hierarchically structured data.

Let us consider any N entities distributed among subsets nested into r groups. Group i contains s_i subsets. Each entity belongs to one out of n categories. Let and μ_{ij} be *a priori* probabilities associated with group i and the subset j of group i, respectively. These probabilities usually are the relative size (entity number) of each group or subset. Let $\mathbf{p}_{ij}$ be a vector that contains the frequencies of the categories in the subset j of group i. The frequencies of the categories in group i as a whole are given by the vector $\mathbf{p}_{i\bullet} = \sum_{j=1}^{s_i} \mu_{ij}\mathbf{p}_{ij}$; and their frequencies in all the groups mixed together are given by the vector $\mathbf{p}_{\bullet\bullet} = \sum_{i=1}^{r} \lambda_i \mathbf{p}_{i\bullet}$. Further consider H as a measure of diversity being nonnegative and concave. Then the APDIV is defined as

$$\begin{aligned} H(\mathbf{p}_{\bullet\bullet}) = & \sum_{i=1}^{r} \lambda_i \sum_{j=1}^{s_i} \mu_{ij} H(\mathbf{p}_{ij}) \\ & + \sum_{i=1}^{r} \lambda_i H(\mathbf{p}_{i\bullet}) - \sum_{i=1}^{r} \lambda_i \sum_{j=1}^{s_i} \mu_{ij} H(\mathbf{p}_{ij}) \\ & + H(\mathbf{p}_{\bullet\bullet}) - \sum_{i=1}^{r} \lambda_i H(\mathbf{p}_{i\bullet}). \end{aligned} \tag{1}$$

The index $H(\mathbf{p}_{\bullet\bullet})$ measures the total diversity within all the groups mixed together. In the right-hand side of equation (1), the first row represents the diversity within subsets and within groups, the second corresponds to the diversity among subsets but within groups, and the last one stands for the diversity among subsets and among groups.

Rao and colleagues (Rao and Nayak, 1985; Liu and Rao, 1995) applied this general approach to a particular diversity index called ''quadratic entropy''. This index was introduced by Rao (1982b) to link diversity measurements with dissimilarity coefficients. Let $\mathbf{D}$ be a $n \times n$ matrix containing the dissimilarities d_{kl} between any categories k and l ($1 \leq k \leq n$ and $1 \leq l \leq n$). Matrix $\mathbf{D}$ is symmetric with null values on the diagonal. The quadratic entropy is defined as

$$H_{\mathbf{D}}(\mathbf{p}) = \mathbf{p}^t\mathbf{D}\mathbf{p} = \sum_{k=1}^{n}\sum_{l=1}^{n} p_k p_l d_{kl}$$

where $\mathbf{p} = (p_1 \cdots p_k \cdots p_n)$ is a frequency vector, either p_{ij}, $\mathrm{p}_{i\bullet}$ or $\mathrm{p}_{\bullet\bullet}$ with the above notations.

As stated by Rao (1986), a diversity measure can be decomposed along a nested sampling provided that it is nonnegative and concave. H_{D} is always nonnegative because it only sums up frequencies and distances which are nonnegative. In order to assure its concavity, we only consider dissimilarity matrices $\mathbf{D}$ such that the matrix noted $\mathbf{D}^{\frac{1}{2}}$, which contains the square root of the values in $\mathbf{D}$, is Euclidean (Rao, 1984; Rao and Nayak, 1985; Champely and Chessel, 2002) that is to say n points M_k ($k = 1,2,...,n$) can be embedded in a Euclidean space so that the Euclidean distance between M_k and M_l is $\sqrt{d_{kl}}$. (Gower and Legendre, 1986). Note that a nonEuclidean

dissimilarity matrix can be transformed into a Euclidean dissimilarity matrix (Lingoes, 1971; Cailliez, 1983). The APDIV applied to the quadratic entropy is called apportionment of quadratic entropy (APQE). APQE generalizes other types of decomposition of diversity indices: the categorical analysis of variance (CATANOVA, Light and Margolin, 1971), which decomposes Gini–Simpson index, and the analysis of variance (ANOVA, Fisher, 1925), which partitions the variance of a quantitative variable. APQE is equal to CATANOVA when the dissimilarities among the categories are all equal to 1 ($\mathbf{D} = \mathbf{1}\mathbf{1}^t - \mathbf{I}$, where $\mathbf{1}$ is a $n \times 1$ vector of units and $\mathbf{I}$ is the $n \times n$ identity matrix) (Nayak, 1986a). It is equal to the ANOVA when the dissimilarity between the entities k and l is equal to $(y_k - y_l)^2$, where y_k and y_l are the values taken by a quantitative variable for the entities k and l, respectively (Rao, 1984).

3. Use of APQE: from genetics to ecology

Fifteen years before Rao's axiomatization the quadratic entropy was introduced probably for the first time by two ecologists (Hendrickson and Ehrlich, 1971) to take into account differences between species in a diversity index. However, for the last 25 years, this index has been given success mostly in genetics. Two teams of geneticists have contributed to the use of quadratic entropy. First, Nei and collaborators (Nei and Li, 1979; Nei and Tajima, 1981) designed indices similar to quadratic entropy in order to measure nucleotide diversity. In that case, the entities are organisms and the categories represent particular DNA sequences. Second, Excoffier *et al.* (1992) developed a decomposition similar to the apportionment of the quadratic entropy widely used in genetics (see for example Bosch *et al.*, 1999; Olsen *et al.*, 2003; Lecis and Norris, 2004; Qiu *et al.*, 2004; Vences *et al.*, 2004). This decomposition, called analysis of molecular variance (AMOVA), treats data where individuals (entities) generally belonging to the same species are sampled from several populations (subsets). These populations may be grouped into clusters, which may be themselves grouped into larger clusters thus generating a hierarchical structure. Furthermore, each individual is characterized by one or two genetic traits (categories) and genetic dissimilarities among traits are computed.

This data scheme may be easily transposable to ecological data since the analysis of species diversity should take into account the dissimilarities between species, the abundance of species within communities and the hierarchical structure of communities.

Several ecologists have recently rediscovered Rao's work and have suggested applying quadratic entropy to their particular ecological data (Izsak and Papp, 1995; Watve and Gangal, 1996; Izsak and Papp, 2000; Shimatani, 2001; Champely and Chessel, 2002; Izsak and Szeidl, 2002). In that case, the diversity within a fauna or a community is under concern and differences between species are estimated in terms of taxonomy or trait. Izsak and Szeidl (2002) even showed that quadratic entropy may decrease with the number of species. Indeed in a community where species are different in terms of genetics or traits, if the additional species are very close to the others the mean of the between-species dissimilarities decreases. To our knowledge, only Champely and Chessel (2002) have applied the partition of Rao's quadratic entropy to ecological data, as

formerly suggested by Woollcott Smith in his discussion following Izsak and Papp (1995) paper.

4. An ecological application in hydrobiology

To illustrate the potential of APQE for analyzing and decomposing diversity in ecological data, we selected a data set published in aquatic ecology by Ivol *et al.* (1997). These authors aimed at analyzing changes in macroinvertebrate assemblages along the course of a large river. Fluvial hydrosystems are in essence hierarchically organized from microhabitats to entire watersheds (Frissell *et al.*, 1986) and thus provide an adequate model for decomposing diversity. Furthermore, approaches at the community level presuppose the use of species lists to compare assemblages among various environmental situations. Such approaches thus imply to consider how species aggregate. Taxonomic aggregation has been mostly used (Corkum and Ciborowski, 1988), however functional aggregation according to biological traits such as size or reproduction (Statzner *et al.*, 1997) or feeding types (Vannote *et al.*, 1980) may enable larger scale comparisons and provide much more general information about ecosystem functioning.

Computations and graphical displays were done using the R statistical software (Ihaka and Gentleman, 1996). Programs and functions for computing AMOVA and quadratic entropy are available in the ade4 package of the R environment; and the APQE function is available by request to the first author of the paper.

4.1 *Data set*

A total of 38 sites were surveyed along 800 km of the Loire River yielding 40 species of Trichoptera and Coleoptera sampled from riffle habitats (see Ivol *et al.*, 1997 for further details on sampling). The river was divided into three regions according to geology: granitic highlands (Region#1), limestone lowlands (Region#2) and granitic lowlands (Region#3).

Two species traits were selected from existing databases (Usseglio-Polatera *et al.*, 2000; Statzner *et al.*, 2001; Gayraud *et al.*, 2003). These trait databases summarise the available European knowledge accumulated over the 20th century for all easily identifiable freshwater invertebrates of France. We selected maximal size, which usually indicates the ratio of production/biomass and of production/respiration in lotic invertebrate (Statzner, 1987) and has many implications for many other functions in the ecosystem. We also considered the functional feeding groups, which have been largely documented since the works of Cummins (1974) and represent cornerstone in the river continuum concept (Vannote *et al.*, 1980). In these databases, for each of the two traits, the affinity of each species to each trait category is quantified using a fuzzy coding approach, (Chevenet *et al.*, 1994). The maximal size achieved by the species is divided into five length categories ranging from ≤ 5 to >40 mm. Feeding habits comprise seven categories: engulfers, shredders, scrapers, deposit-feeders, active filter-feeders, passive filter-feeders and piercers. A score is

assigned to each species for describing its affinity for a given trait category from "0" which indicates no affinity to "3" which indicates high affinity. These affinities are further transformed into percentage per trait per species. The percentage of affinity of the species k for the category m is noted q_{km}.

4.2 *Dissimilarity measurements*

We used four criteria to compute the dissimilarities among species: equidistance, body size, feeding habits and taxonomy. For the equidistance, the dissimilarity between two species was arbitrarily set to 1 meaning an equivalence between species. To compute dissimilarities from the fuzzy variables (body size and feeding habit), we selected the Manly's distance. (Manly, 1994, formula 5.8 p. 68) defined as

$$d_{kl} = 1 - \sum_{m=1}^{M} q_{km} q_{lm} \Bigg/ \sqrt{\left\{ \sum_{m=1}^{M} q_{km}^2 \sum_{m=1}^{M} q_{lm}^2 \right\}}$$

where d_{kl} is the dissimilarity between species k and l, M is the number of categories (five for the maximal-size trait and seven for the feeding-habit trait), and 9 km and q_{lm} are the percentages of affinities of species k and l for the category m of either the body size or the feeding habit depending upon which criteria is concerned. Quadratic entropy applied to dissimilarities taking into account species traits yields a measure of functional diversity (Petchey and Gaston, 2002). For computing taxonomic dissimilarities we used the index proposed independently by Izsak and Papp (1995) and Warwick and Clarke (1995): the dissimilarity equals 1 between two species of the same genus, 2 between two species of different genera belonging to the same family, 3 between two species of different families belonging to the same order and 4 between two species belonging to a different order.

4.3 *Decomposition of the quadratic entropy*

Since the value of total diversity depends on the type of biological variable we compared the decomposition of the total diversity in terms of percentage. Within-site diversity incorporated from 50 to about 65% of the total diversity (Table 1). Values were very stable across the indices. Only feeding groups demonstrated a lower diversity at this scale linked to a higher diversity among sites. The diversity among sites within regions contained from 24 to 36% of the total diversity. Finally, the diversity among regions ranged from 9 to 14% of the total diversity. For comparing values of diversity across spatial scales, we had to take into account the number of independent items (degree of freedom) at each scale. To test the within-scale differences several methods have been proposed. Nayak's process (1986a, b) tests if the organisms are independently divided into the sites according to a multinomial distribution. This distribution is assumed to be constant first across regions and second across sites within each region. Our data highly differs from both assumptions. Two interpretations are possible: either the sites and regions are indeed different, or the

Table 1. Decomposition of the total diversity according to (a) the frequencies of the species (Gini–Simpson index), the taxonomy, and (b) two biological traits.

Diversity Source	*d.f.*	*Gini–Simpson*	*Ratio*	*P*	*Taxonomy*	*Ratio*	*P*
(a)							
Among regions	2	0.087 (11%)	8.322	< 0.001	0.210 (9%)	5.797	0.022
Among sites/Within regions	35	0.183 (24%)	48.07	0.894	0.635 (28%)	59.84	0.277
Within sites	4628	0.504 (65%)			1.403 (63%)		
Total	4665	0.774			2.248		
	d.f.	Size	Ratio	*P*	Diet	Ratio	*P*
(b)							
Among regions	2	0.013 (9%)	6.558	0.022	0.020 (14%)	6.721	0.012
Among sites/Within regions	35	0.036 (26%)	52.53	0.545	0.051 (36%)	96.08	0.017
Within sites	4628	0.090 (65%)			0.070 (50%)		
Total	4665	0.139			0.141		

The degrees of freedom (d.f.) are indicated. Statistics and results of the permutation tests are given.

distribution of the organisms across the sites is not multinomial. Macroinvertebrates show patchy distributions according to levels of environmental disturbance (Levins and Paine, 1974). Those aggregations imply that, sampling an individual from a species increases the chance of observing other individuals from the same species. The multinomial assumption is thus here invalidated. Excoffier *et al.* (1992) suggested permutation tests to avoid distribution assumptions. They performed test on the differences between regions by permuting sites across regions. We choose this permutation scheme because it suits our data by taking into account the aggregation of individuals. For each permutation, we compute the ratio of the diversity between regions to the diversity between sites within regions. Excoffier *et al.* (1992) tested the differences between sites within regions by permuting the individuals across the sites within each region. This type of permutations does not take into account aggregation and thus could overestimate the real between-site differences. We choose to permute each species' abundance across the sites within each region. For example, for species k in region i, the abundance values n_{ijk} are permuted over sites $1 \leq j \leq s_i$. Once the permutations are done for all species, the ratio of the diversity between sites within regions to the diversity within sites is computed. For selected permutation scheme, the number of simulated values (out of 1000 samples) higher than the observed one is given in Table 1. For the between-region diversity from 12 to 22 simulated values exceeded the observed reference ratio. This suggests significant differences between regions in terms of frequency distributions, taxonomic and size compositions and diets. By contrast, the only significant differences between sites within regions are due to feeding habits.

According to Lande (1996), gamma diversity equals the weighted average alpha diversity plus beta diversity. Computing such values for each of our three regions (Table 2) showed that the three regions were approximately equally balanced according to the abundance of species (Gini–Simpson total diversity). Taxonomic, size, and diet diversities discriminated Region#1 situated upstream better than Gini–Simpson diversity due to higher regional species richness about twice that of the two other regions. Though differing in species richness, Region#2 and Region#3 had similar low diversity in size and diet, whereas similar to other indices Region#1 had the largest value. This result, which contradicts usual knowledge (e.g., Statzner, 1987), should be associated with the further downstream impact of human activities (Region#2 and Region#3) involving a reduction of environmental heterogeneity.

5. Discussion

Few methods for quantifying gamma diversity (i.e., diversity at the landscape level allowing comparison among regions) exist in ecology (Sweeney and Cook, 2001). The selected ecological example was intended to show how the APQE may help to partition diversity. In this case the APQE allowed the computation of two types of global diversity. Diversity at the scale of the entire stream or total diversity could be valuably compared to a similar value computed for some other stream. Regional gamma diversity allowed us the comparison of diversity among regions. As a result, the APQE is a useful tool for estimating biodiversity at a variety of spatial scales, a major issue in both basic and applied ecology (Vinson and Hawkins, 1998).

Table 2. Values of the diversity among sites within regions, the diversity within sites and the total diversity in each selected region.

	d.f.	*Gini–Simpson*	*Taxonomy*	*Size*	*Diet*	*Richness*
Region#1						
Among sites	15	0.25 (28%)	1.28 (33%)	0.08 (31%)	0.11 (44%)	24.00
Within sites	1579	0.63 (72%)	2.64 (67%)	0.18 (69%)	0.14 (56%)	7.00
Total	1594	0.88	3.92	0.26	0.25	31
Region#2						
Among sites	16	0.18 (32%)	0.30 (31%)	0.02 (33%)	0.02 (40%)	13.69
Within sites	2307	0.37 (68%)	0.66 (69%)	0.04 (67%)	0.03 (60%)	4.31
Total	2323	0.55	0.96	0.06	0.05	18
Region#3						
Among sites	4	0.05 (7%)	0.29 (22%)	0.01 (17%)	0.02 (29%)	6.60
Within sites	742	0.66 (93%)	1.05 (78%)	0.05 (83%)	0.05 (71%)	5.40
Total	746	0.71	1.34	0.06	0.07	12

Three diversity indices are used: the Gini–Simpson index, the quadratic entropy applied to three dissimilarity criteria (taxonomy, body size and feeding habits), and the richness. Richness decomposition was given by Lande (1996). The degrees of freedom (d.f.) are indicated.

In our study, diversity was significantly different between regions whereas diversity between sites within regions was significantly dissimilar only for diet composition. Differences in diet composition and resulting diversity are predicted along rivers by the river continuum concept (Vannote *et al.*, 1980). This may explain our results since our sites within regions are distributed from up- to downstream over large distances (>40 km). The differences observed between regions are probably due to the environmental characteristics of each region such as altitude for example. This latter result is supported by Parsons *et al.* (2003) who have demonstrated a greater similarity in macroinvertebrate assemblages at the site (riffle) scale than at the catchment scale.

Veech *et al.* (2002) states that ecologists should use diversity partitioning as a conceptual framework and an analytical method to address questions pertaining to the relationship between local and regional species diversity. We think that Rao's axiomatization appears as a fundamental basis for analyzing patterns of diversity. Quadratic entropy has an advantage over species richness and Gini–Simpson index because it takes into consideration the dissimilarities among species.

Other recent studies are linked to APQE. Their goal was not to describe diversity but to test differences between groups of sites estimated through Bray–Curtis index (Bray and Curtis, 1957). Legendre and Anderson (1999), McArdle and Anderson (2001) and Anderson (2001) have tackled the question of factorial designs instead of nested designs. In fact, their analyses correspond to another part of Rao's axiomatization namely the analysis of diversity (ANODIV). In that case, up to now the constraint of orthogonal sampling is needed. Ways for performing permutation tests are still debated. Anderson (2001) did not restrict the partition of diversity to Euclidean matrix of dissimilarity, but stated that any matrix can be used. Since the primary focus of the author was to test differences between groups of sites, then the question of negative diversity was of minor importance. In this paper, we have restricted our analysis to Euclidean matrix (e.g., Rao and Nayak, 1985; Schneider *et al.*, 2000) to remove the non-interpretable result such as negative diversity between communities.

Cousins (1991) underlined the contradiction that traditional ecological indices treat species on an equal basis whereas species identification put to the fore differences between species. Shimatani (2001) also considered that species differences should be included in biodiversity indices to provide better ecological applications. Finally, Watve and Gangal (1996) noticed that "an information-based index would treat a community of four different biotypes of coliforms identical to another community consisting of one species of coliforms, one of actinomycetes, one of myxobacteria, and one of archaebacteria, whereas we feel that the latter should be treated as more diverse."

The use of APQE allows the introduction of phylogenetic or taxonomic distances between species, which helps to overcome the above drawbacks. In our example, we used a taxonomic distance computed according to the taxonomic tree. An alternative could be to decompose diversity according to the taxonomic level (e.g., genus, family). Furthermore, ecologists do need methods that simultaneously evaluate diversity at different scales (Ricklefs, 1987) and APQE may help to reach this objective.

As stated in the introduction, several methods for decomposing diversity were developed mainly in genetics and transferred to ecology. For example, Allan (1975)

compared two methods developed in ecology by Pielou (1967) and Levins (1968) to one method designed in genetics by Lewontin (1972). Furthermore, Lande (1996) introduced in ecology the additive decomposition of the Gini–Simpson index proposed in genetics by Nei (1973). In this paper, we suggest that the apportionment of the quadratic entropy of Rao (1982a), also at work in AMOVA (Excoffier *et al*., 1992) a method designed for genetics, can be efficient for ecological data. All these decompositions of diversity represent particular cases of Rao's apportionment of diversity.

Acknowledgments

Many thanks to the "BIOFUN" group who kindly provided some additional information on the trait data. The data set on the Loire River was kindly provided by J.M. Ivol.

References

Allan, J.D. (1975) Components of diversity. *Oecologia*, **18**, 359–67.

Anderson, M.J. (2001) A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, **26**, 32–46.

Bosch, E., Calafell, F., Santos, F.R., Perez-Lezaun, A., Comas, D., Benchemsi, N., Tyler-Smith, C., and Bertranpetit, J. (1999) Variation in short tandem repeats is deeply structured by genetic background on the human Y chromosome. *American Journal of Human Genetics*, **65**, 1623–38.

Bray, J.R. and Curtis, J.T. (1957) An ordination of the upland forest communities of Southern Wisconsin. *Ecological Monographs*, **27**, 325–49.

Cailliez, F. (1983) The analytical solution of the additive constant problem. *Psychometrika*, **48**, 305–10.

Champely, S. and Chessel, D. (2002) Measuring biological diversity using Euclidean metrics. *Environmental and Ecological Statistics*, **9**, 167–77.

Chevenet, F., Dolédec, S., and Chessel, D. (1994) A fuzzy coding approach for the analysis of long-term ecological data. *Freshwater Biology*, **31**, 295–309.

Corkum, L.D. and Ciborowski, J.J.H. (1988) Use of alternative classifications in studying broad-scale distributional patterns of lotic invertebrates. *Journal of the North American Benthological Society*, **7**, 167–79.

Cousins, S.H. (1991) Species diversity measurement: Choosing the right index. *Trends in Ecology and Evolution*, **6**, 190–92.

Crease, T.J., Lynch, M., and Spitze, K. (1990) Hierarchical analysis of population genetic variation in mitochondrial and nuclear genes of *Daphnia pulex*. *Molecular Biology and Evolution*, **7**, 444–58.

Cummins, K.W. (1974) Structure and function of stream ecosystems. *BioScience*, **24**, 631–41.

Excoffier, L. (2001) Analysis of population subdivision in *Handbook of Statistical Genetics*, D.J. Balding, M. Bishop and C. Cannings (eds.), John Wiley & Sons, Ltd, New York, pp. 271–307.

Excoffier, L., Smouse, P.E., and Quattro, J.M. (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. *Genetics*, **131**, 479–91.

Finkeldey, R. (1994) A simple derivation of the partitioning of genetic differentiation within subdivided populations. *Theoretical and Applied Genetics*, **89**, 198–200.

Fisher, R.A. (1925) *Statistical Methods for Research Workers*, Oliver & Boyd, Edinburgh.

Frissell, C.A., Liss, W.J., Warren, C.E., and Hurley, M.D. (1986) A hierarchical framework for stream habitat classification: Viewing streams in a watershed context. *Environmental Management*, **10**, 199–214.

Gayraud, S., Statzner, B., Bady, P., Haybach, A., Schöll, F., Usseglio-Polatera, P., and Bachi, M. (2003) Invertebrate traits for the biomonitoring of European large rivers: An initial assessment of alternative metrics. *Freshwater Biology*, **48**, 2045–64.

Gower, J.C. and Legendre, P. (1986) Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification*, **3**, 5–48.

Hendrickson, J.A.J. and Ehrlich, P.R. (1971) An expanded concept of ''species diversity''. *Notulae Naturae*, **439**, 1–6.

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

Holsinger, K.E. and Mason-Gamer, R.J. (1996) Hierarchical analysis of nucleotide diversity in geographically structured populations. *Genetics*, **142**, 629–39.

Ihaka, R. and Gentleman, R. (1996) R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, **5**, 299–314.

Ivol, J.M., Guinand, B., Richoux, P., and Tachet, H. (1997) Longitudinal changes in Trichoptera and Coleoptera assemblages and environmental conditions in the Loire River (France). *Archiv für Hydrobiologie*, **138**, 525–57.

Izsak, J. and Papp, L. (1995) Application of the quadratic entropy indices for diversity studies of drosophilid assemblages. *Environmental and Ecological Statistics*, **2**, 213–24.

Izsak, J. and Papp, L. (2000) A link between ecological diversity indices and measures of biodiversity. *Ecological Modelling*, **130**, 151–56.

Izsak, J. and Szeidl, L. (2002) Quadratic diversity: Its maximization can reduce the richness of species. *Environmental and Ecological Statistics*, **9**, 423–30.

Kolasa, J. (1989) Ecological systems in hierarchical perspective: breaks in community structure and other consequences. *Ecology*, **70**, 36–47.

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

Lecis, R. and Norris, K. (2004) Population genetic diversity of the endemic Sardinian newt *Euproctus platycephalus*: Implications for conservation. *Biological Conservation*, **119**, 263–70.

Legendre, P. and Anderson, M.J. (1999) Distance-based redundancy analysis: Testing multispecies responses in multifactorial ecological experiments. *Ecological Monographs*, **69**, 1–24.

Levins, R. (1968) *Evolution in changing environments. Monographs in population biology*, Princeton University Press, Princeton.

Levins, S.A. and Paine, R.T. (1974) Disturbance, patch formation, and community structure. *Proceedings of the National Academy of Sciences of the USA*, **71**, 2744–47.

Lewontin, R.C. (1972) The apportionment of human diversity. *Evolutionary Biology*, **6**, 381–98.

Light, R.J. and Margolin, B.H. (1971) An analysis of variance for categorical data. *Journal of the American Statistical Association*, **66**, 534–44.

Lingoes, J.C. (1971) Some boundary conditions for a monotone analysis of symmetric matrices. *Psychometrika*, **36**, 195–203.

Liu, Z.J. and Rao, C.R. (1995) Asymptotic distribution of statistics based on quadratic entropy and bootstrapping. *Journal of Statistical Planning and Inference*, **43**, 1–18.

Lynch, M. and Crease, T.J. (1990) The analysis of population survey data on DNA sequence variation. *Molecular Biology and Evolution*, **7**, 377–94.

Manly, B.F. (1994) *Multivariate Statistical Methods. A primer*, 2 Chapman & Hall, London.

McArdle, B.H. and Anderson, M.J. (2001) Fitting multivariate models to community data: comment on distance-based redundancy analysis. *Ecology*, **82**, 290–97.

Nayak, T.K. (1986a) An analysis of diversity using Rao's quadratic entropy. *Sankhya: The Indian Journal of Statistics*, **48B**, 315–30.

Nayak, T.K. (1986b) Sampling distributions in analysis of diversity. *Sankhya: The Indian Journal of Statistics*, **48B**, 1–9.

Nei, M. (1973) Analysis of gene diversity in subdivised populations. *Proceedings of the National Academy of Sciences of the United States of America*, **70**, 3321–23.

Nei, M. and Jin, L. (1989) Variances of the average numbers of nucleotide substitutions within and between populations. *Molecular Biology and Evolution*, **6**, 290–300.

Nei, M. and Li, W.-H. (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America*, **76**, 5269–73.

Nei, M. and Miller, J.C. (1990) A simple method for estimating average number of nucleotide substitutions within and between populations from restriction data. *Genetics*, **125**, 873–9.

Nei, M. and Tajima, F. (1981) DNA polymorphism detectable by restriction endonucleases. *Genetics*, **97**, 145–63.

Olsen, J.B., Miller, S.J., Spearman, W.J., and Wenburg, J.K. (2003) Patterns of intra- and inter-population genetic diversity in Alaska coho salmon: Implications for conservation. *Conservation Genetics*, **4**, 557–69.

Parsons, M., Thoms, M.C., and Norris, R.H. (2003) Scales of macroinvertebrate distribution in relation to the hierarchical organization of river systems. *Journal of the North American Benthological Society*, **22**, 105–22.

Petchey, O.L. and Gaston, K. (2002) Functional diversity (FD), species richness and community composition. *Ecology Letters*, **5**, 402–11.

Pielou, E.C. (1967) The use of information theory in the study of the diversity of biological populations. In *Proc. Fifth Berkeley Symposium on Math. Stat. and Prob.* pp. 163–77.

Pielou, E.C. (1975) *Ecological diversity*, Wiley & Sons, New York.

Qiu, Y.-X., Hong, D.-Y., Fu, C.-X., and Cameron, K.M. (2004) Genetic variation in the endangered and endemic species *Changium smyrnioides* (Apiaceae). *Biochemical Systematics and Ecology*, **32**, 583–96.

Rao, C.R. (1982a) Diversity and dissimilarity coefficients: A unified approach. *Theoretical Population Biology*, **21**, 24–43.

Rao, C.R. (1982b) Diversity: Its measurement, decomposition, apportionment and analysis. *Sankhya: The Indian Journal of Statistics*, **A44**, 1–22.

Rao, C.R. (1984) Convexity properties of entropy functions and analysis of diversity. *Inequalities in statistics and probability*, **5**, 68–77.

Rao, C.R. (1986) Rao's axiomatization of diversity measures in *Encyclopedia of Statistical Sciences*, S. Kotz and N.L. Johnson (eds.), Wiley & Sons, New York, pp. 614–17.

Rao, C.R. and Nayak, T.K. (1985) Cross entropy, dissimilarity measures, and characterizations of Quadratic Entropy. *IEEE Transactions on Information Theory*, **IT-31**, 589–93.

Ricklefs, R.E. (1987) Community diversity: Relative roles of local and regional processes. *Science*, **235**, 167–71.

Ricotta, C. (2003) Additive partition of parametric information and its associated beta-diversity measure. *Acta Biotheoretica*, **51**, 91–100.

Schneider, S., Roessli, D., and Excoffier, L. (2000) *Arlequin Ver 2.000: A Software for Population Genetics Data Analysis*, Genetics and Biometry Laboratory, Department of Anthropology, University of Geneva, Switzerland.

Shimatani, K. (2001) On the measurement of species diversity incorporating species differences. *Oikos*, **93**, 135–47.

Slatkin, M. (1995) A measure of population subdivision based on microsatellite allele frequencies. *Genetics*, **139**, 457–62.

Statzner, B. (1987) Characteristics of lotic ecosystem and consequences for future research directions. *Ecological studies*, **61**, 365–90.

Statzner, B., Bis, B., Dolédec, S., and Usseglio-Polatera, P. (2001) Perspectives for biomonitoring at large spatial scales: a unified measure for the functional composition of invertebrate communities in European running waters. *Basic and Applied Ecology*, **2**, 73–85.

Statzner, B., Hoppenhaus, K., Arens, M.-F., and Richoux, P. (1997) Reproductive traits, habitat use and templet theory: A synthesis of world-wide data on aquatic insects. *Freshwater Biology*, **38**, 109–35.

Sweeney, B.A. and Cook, E. (2001) A landscape-level assessment of understory diversity in upland forests of North-Central Wisconsin, USA. *Landscape Ecology*, **16**, 55–69.

Usseglio-Polatera, P., Bournaud, M., Richoux, P., and Tachet, H. (2000) Biological and ecological traits of benthic freshwater macroinvertabrates: relationships and definition of groups with similar traits. *Freshwater Biology*, **43**, 175–205.

Vannote, R.L., Minshall, G.W., Cummins, K.W., Sedell, J.A., and Cushing, C.E. (1980) The river continuum concept. *Canadian Journal of Fisheries and Aquatic Sciences*, **37**, 130–37.

Veech, J.A., Summerville, K.S., Crist, T.O., and Gering, J.C. (2002) The additive partitioning of species diversity: recent revival of an old idea. *Oikos*, **99**, 3–9.

Vences, M., Chiari, Y., Raharivololoniaina, L., and Meyer, A. (2004) High mitochondrial diversity within and among populations of Malagasy poison frogs. *Molecular Phylogenetics and Evolution*, **30**, 295–307.

Vinson, M.R. and Hawkins, C.P. (1998) Biodiversity of stream insects: variation at local, basin, and regional scales. *Annual Review of Entomology*, **43**, 271–93.

Wagner, H.H., Wildi, O., and Ewald, K.C. (2000) Additive partitioning of plant species diversity in an agricultural mosaic landscape. *Landscape Ecology*, **15**, 219–27.

Warwick, R.M. and Clarke, K.R. (1995) New 'biodiversity' measures reveal a decrease in taxonomic distinctness with increasing stress. *Marine Ecology Progress Series*, **129**, 301–5.

Watve, M.G. and Gangal, R.M. (1996) Problems in measuring bacterial diversity and a possible solution. *Applied and Environmental Microbiology*, **62**, 4299–301.

Weir, B.S. and Cockerham, C.C. (1984) Estimating F-statistics for the analysis of population structure. *Evolution*, **38**, 1358–70.

Whittaker, R.H. (1960) Vegetation of the Siskiyou Mountains, Oregon and California. *Ecological Monographs*, **30**, 279–338.

Whittaker, R.H. (1972) Evolution and measurement of species diversity. *TAXON*, **21**, 213–51.

Biographical sketches

S. Pavoine is Ph.D. student at the ''statistical ecology'' team in the UMR CNRS 5558. Her research interest concerns the statistical methods for measuring diversity including test procedures and validation with the aim of providing a theoretical unifying point of view available to various almost separate disciplines such as genetics and community ecology.

S. Dolédec teaches biology, ecology and data analysis at the University of Lyon1. He belongs to a CNRS ecological research unit in the Laboratoire d'Ecologie des Hydrosystèmes Fluviaux of the same university. His research incorporates data analysis and current ecological theories to serve ecologically oriented management.

Annexe 3
Pavoine *et al.* 2005 - Theoretical Population Biology

Pavoine, S., S. Ollier, and D. Pontier. 2005. Measuring diversity from dissimilarities with Rao’s quadratic entropy : are any dissimilarity indices suitable ? Theoretical Population Biology 67 :231-239.

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Theoretical Population Biology 67 (2005) 231–239
Theoretical Population Biology
www.elsevier.com/locate/ytpbi

Measuring diversity from dissimilarities with Rao's quadratic entropy: Are any dissimilarities suitable?

S. Pavoine*, S. Ollier, D. Pontier

Laboratoire de Biométrie et Biologie Evolutive, UMR CNRS 5558, Université Claude Bernard LYON I, 43, boulevard du 11 novembre 1918, Villeurbanne cedex 69622, France

Received 7 June 2004
Available online 5 April 2005

Abstract

Rao has developed quadratic entropy to measure diversity in a set of entities divided up among a fixed set of categories. This index depends on a chosen matrix of dissimilarities among categories and a frequency distribution of these categories. With certain choices of dissimilarities, this index could be maximized over all frequency distributions by eliminating several categories. This unexpected result is radically opposite to those obtained with usual diversity indices. We demonstrate that the elimination of categories to maximize the quadratic entropy depends on mathematical properties of the chosen dissimilarities. In particular, when quadratic entropy is applied to ultrametric dissimilarities, all categories are retained in order to reach its maximal value. Three examples, varying from simple one-dimensional to ultrametric dissimilarity matrices, are provided. We conclude that, as far as diversity measurement is concerned, quadratic entropy is most relevant when applied to ultrametric dissimilarities.

Keywords: Dissimilarity; Diversity; Quadratic entropy; Smallest enclosing hypersphere; Ultrametric

1. Introduction

Diversity is traditionally measured in a set of entities divided up among categories. From mathematical point of view a diversity measure is a function mapping the points of the probability simplex $u = \{\mathbf{p} = (p_1, \ldots, p_n), p_i \geqslant 0, \sum_{i=1}^{n} p_i = 1\}$ into the set of positive scalars. Note that taking the frequencies of the categories in a sample with N entities, the vector $(N_1/N, N_2/N, \ldots, N_n/N)$ belongs to u, where N_i is the number of entities in a category i. For example, in ecology it is measured in a community of organisms divided up among species (Pielou, 1975); in genetics it is usually assessed in a population of individuals divided up among nucleomorphs (Nei and Tajima, 1981). The most widespread indices are the Gini–Simpson index, reciprocate Simpson index and Shannon index (Magurran, 1988). An important property of these indices is that they reach their maximum when all the existing categories have the same frequency.

Many authors have argued against these indices by stating that differences among categories should also be taken into account (e.g., Nei and Li, 1979; Rao, 1982a; Cousins, 1991; Watve and Gangal, 1996). Quoting Watve and Gangal (1996), the usual indices cited above would treat a community of four different biotypes of coliform bacteria identical to another community consisting of one species of coliform bacteria, one of actinomycetes, one of myxobacteria, and one of archaebacteria, whereas we feel that the latter should be treated as more diverse. Among these researchers, Rao (1982a) defined a general index for the measurement of diversity. The domain of this index, called quadratic entropy, is also the simplex u. However, a matrix of dissimilarities among categories (e.g., in ecology, phenetic differences among species) plays a role in the index value formation.

*Corresponding author. Fax: +33 472 431 388.
E-mail addresses: pavoine@biomserv.univ-lyon1.fr (S. Pavoine), ollier@biomserv.univ-lyon1.fr (S. Ollier), dpontier@biomserv.univ-lyon1.fr (D. Pontier).

0040-5809/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.tpb.2005.01.004

Many types of dissimilarities are often available according to the precise objective of the research. For example, in ecology among-species dissimilarities can be computed with genetic, morphologic, taxonomic or phylogenetic criteria. For instance, genetic dissimilarities among species can be calculated from different markers such as amino acids, DNA sequences and AFLP fingerprints. Furthermore, once a criterion and a marker have been chosen, there are still many developed metrics for providing a dissimilarity matrix (Edwards, 1971; Nei, 1972). Dissimilarity matrices and frequency distributions do not have an identical role in quadratic entropy. Indeed the dissimilarity matrix, which characterizes the categories, remains constant over sets. Conversely, the frequency distribution, which characterizes the sets, varies over sets.

Quadratic entropy is not a bounded index because its range of values depends on the scale of the dissimilarities. Consequently, it is difficult to compare results obtained from different choices of dissimilarities. To overcome this limitation one may change to a relative diversity index. To do this the observed measure of the quadratic entropy should be divided by its maximal theoretical value in the set of frequency distributions (Champely and Chessel, 2002). As for this maximal value, a surprising result appeared: this maximizing frequency distribution often contains several, or even many, zero values (Champely and Chessel, 2002; Izsak and Szeidl, 2002). This observation means that the maximal diversity can be reached with only a few out of all the categories concerned. As far as maximization in the set of frequency distributions is concerned, the quadratic entropy is thus very different from the usual indices cited above.

When applied to biological conservation, quadratic entropy might suggest that several species must often be eliminated in order to maximize biodiversity. This is unacceptable and thus invalidates such a use of the quadratic entropy, at least for the data sets for which this surprising result appears. Indeed, in ecology, Shimatani (2001) made an important observation by applying quadratic entropy to a range of dissimilarity matrices: the behavior of quadratic entropy can be very different depending upon the kind of dissimilarity matrix. For example, he first calculated specific genetic dissimilarities among species by Kimura's formula (Nei and Tajima, 1981; Kimura, 1983): each species is characterized by an amino-acid sequence, the dissimilarity between two species is the number of substitutions which are expected to have occurred between their respective sequences. Shimatani (2001) obtained null frequencies to reach the maximal value of the quadratic entropy applied to these dissimilarities. Conversely, when measuring the quadratic entropy from taxonomic dissimilarities among species, he obtained no null frequencies to reach this maximum. The use of the quadratic entropy to measure diversity is thus not invalidated with all kinds of dissimilarity. This observation raises the question: what makes the difference between these genetic dissimilarities and the taxonomic dissimilarities?.

As far as we know, no method has been developed to detect a priori whether a dissimilarity matrix, of interest from a biological point of view, will provide relevant measure of diversity when associated with quadratic entropy. It is thus useful to define a class of dissimilarity matrices which never suggest eliminating species to maximize biological diversity. In this paper, we firstly explain the mathematical reasons for the differences found in the behavior of quadratic entropy according to the dissimilarity matrix and then define a suitable class of dissimilarity matrices.

2. Rao's quadratic entropy and its maximization: a geometrical interpretation

Denote $\mathbf{p} = (p_1 \cdots p_n)$ the vector containing the frequencies of n categories in a set, where $p_i \geqslant 0$ and $\sum_{i=1}^{n} p_i = 1$, d_{ij} the dissimilarity between categories i and j and $\mathbf{D}$ the matrix $[d_{ij}]$ with d_{ij} as entries where $0 \leqslant i \leqslant n$ and $0 \leqslant j \leqslant n$. Rao (1982b) defined the quadratic entropy as

$$H_{\mathbf{D}}(\mathbf{p}) = \sum_{i=1}^{n} \sum_{j=1}^{n} p_i p_j d_{ij}.$$

In ecology, $\mathbf{p}$ may be the vector containing occurrence probabilities in a multispecies population and d_{ij} a measure of the dissimilarity between species i and j.

The index d_{ij} must be a conditionally negative definite function (Rao and Nayak, 1985) that is to say the matrix $\left[\sqrt{d_{ij}}\right]$ must be Euclidean (Critchley and Fichet, 1997), which means that n points M_i $(i = 1, \ldots, n)$ can be embedded in a Euclidean space so that the Euclidean distance between M_i and M_j is $\sqrt{d_{ij}}$, i.e. $\sqrt{d_{ij}} = ||M_i - M_j||$ (Gower, 1966). Quadratic entropy measures the expected difference between two entities randomly drawn from the set. When $d_{ij} = 1$ for all $i \neq j$ and $d_{ii} = 0$ for all i, quadratic entropy is identical with the Gini–Simpson index (Rao, 1982a). When $d_{ij} = \frac{1}{2}(y_i - y_j)^2$, where y_i and y_j are the values of a quantitative variable Y for the elements or specimens in categories i and j, respectively, quadratic entropy is equal to the variance of Y (Rao, 1986).

Champely and Chessel (2002) provided an optimization technique to assess the maximum value of $H_{\mathbf{D}}(\mathbf{p})$ in the set $\{\mathbf{p} = (p_1 \cdots p_n), p_i \geqslant 0, \sum_{i=1}^{n} p_i = 1\}$. They verified this technique using the Kuhn-Tuckey conditions. We propose below a new solution for finding this maximum. It is based on geometrical properties instead of an iterative process. Following the notations of Champely

and Chessel (2002), we write $\delta_{ij} = \sqrt{2d_{ij}}$. The $n \times n$ matrix $\Delta = [\delta_{ij}]$ is now the reference. Rao's quadratic entropy becomes

$$H_{\Delta}(\mathbf{p}) = \sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j \frac{(\delta_{ij})^2}{2}$$

which does not change the value of the quadratic entropy: $H_{\Delta}(\mathbf{p}) = H_{\mathbf{D}}(\mathbf{p})$. As stated above, Δ must be Euclidean. The coordinates of the n points can be obtained by principal coordinate analysis (see e.g., Gower and Legendre, 1986).

We developed three propositions as follows and a fourth proposition in the next part (proofs are given in our Appendix).

Proposition 1. *The maximal value of* $H_{\Delta}(\mathbf{p})$ *in the set* $\{\mathbf{p} = (p_1 \cdots p_n), p_i \geqslant 0, \sum_{i=1}^{n} p_i = 1\}$ *is the squared radius of the smallest hypersphere containing the n points* M_i.

The smallest enclosing hypersphere of the n points M_i is defined as the ball of minimal radius which contains these n points. It has k points on the boundary (the support set T, $0 \leqslant k \leqslant n$) and $n-k$ points strictly enclosed inside (Fisher et al., 2003). Let $\Delta_T = [\delta_{ij}]$ be the $k \times k$ matrix containing the Euclidean distances among the points in the support set T. Let $\mathbf{D}_T = [\frac{1}{2}\delta_{ij}^2]$ where $1 \leqslant i \leqslant k$ and $1 \leqslant j \leqslant k$, and $\mathbf{1}_k$ be the $k \times 1$ vector of units.

Proposition 2. *The squared radius of the smallest hypersphere containing the n points* M_i *is equal to* $(\mathbf{1}_k^t \mathbf{D}_T^{-1} \mathbf{1}_k)^{-1}$.

The proof of Proposition 2 can be found in Gower (1982, 1984) considering that the k points in the support set T lie on the surface of a hypersphere. Let $\mathbf{p}_{\max} = (p_{\max(1)} \cdots p_{\max(n)})$ be the vector which maximizes $H_{\Delta}(\mathbf{p})$ in the set $\{\mathbf{p} = (p_1 \cdots p_n), p_i \geqslant 0, \sum_{i=1}^{n} p_i = 1\}$.

Proposition 3. *The frequencies in* $\mathbf{p}_{\max}$ *of the k points in the support set T are given by the vector* $\mathbf{p}_T = (\mathbf{D}_T^{-1}\mathbf{1}_k)/(\mathbf{1}_k^t \mathbf{D}_T^{-1}\mathbf{1}_k)$ *and the frequencies in* $\mathbf{p}_{\max}$ *of the* $n-k$ *points located strictly inside the smallest enclosing hypersphere equal zero.*

3. Particular class of dissimilarity matrices and the special case of ultrametric dissimilarities

Let $\Delta_S = [\delta_{ij}]$ be a $n \times n$ Euclidean matrix of dissimilarities. By definition, it is associated with a set S of n points M_i $(i = 1, \ldots, n)$ embedded in a Euclidean space so that the Euclidean distance between M_i and M_j is δ_{ij}. Consider $\mathbf{D}_S = [\frac{1}{2}\delta_{ij}^2]$, where $1 \leqslant i \leqslant n$ and $1 \leqslant j \leqslant n$. The matrix Δ_S is circum-Euclidean if and only if it is Euclidean and the points in S lie on the boundary (surface) of a hypersphere, which is called a circumsphere (Critchley and Fichet, 1997). The center O of that circumsphere belongs to conv(S) if and only if it belongs to the set of all the barycenters of the points in S weighted by nonnegative coefficients. From the above propositions we deduce that $H(\mathbf{p}_{\max}) = (\mathbf{1}_n^t \mathbf{D}_S^{-1} \mathbf{1}_n)^{-1}$ and $\mathbf{p}_{\max} = (\mathbf{D}_S^{-1}\mathbf{1}_n)/(\mathbf{1}_n^t \mathbf{D}_S^{-1}\mathbf{1}_n)$ if and only if Δ_S is circum-Euclidean and its circumcenter belongs to conv(S).

These two conditions define a particular class of dissimilarity matrices. A matrix of that class will be said to be "SEH-circum-Euclidean" because it is circum-Euclidean and its associated circumsphere is equal to its smallest enclosing hypersphere. We distinguish two subclasses. The first is said to be "weak" because the vector $\mathbf{p}_{\max}$ may have several null values. The second is said to be "strong" because the vector $\mathbf{p}_{\max}$ has only positive values.

A $n \times n$ matrix $\Delta = [\delta_{ij}]$ is ultrametric if and only if $\delta_{ij} \geqslant 0$, for all i and j, $\delta_{ij} \leqslant \max(\delta_{ik}, \delta_{kj})$, for all i, j and k, and $\delta_{ii} < \min_{j \neq i}(\delta_{ij})$, for all j $(\delta_{ii} = 0)$. The ultrametric property is possessed by all dissimilarities which can be directly associated with rooted trees in which all the end nodes are equidistant from the root of the tree, e.g., taxonomic trees and also particular phylogenetic trees (Van de Peer, 2003).

Proposition 4. *Any ultrametric matrix belongs to the strong subclass of the set of SEH-circum-Euclidean matrices.*

Any uniform dissimilarity matrix $\lambda(\mathbf{1}_n \mathbf{1}_n^t - \mathbf{I}_n)$, where λ belongs to $\mathbb{R}^{+*}$ and $\mathbf{I}_n$ is the $n \times n$ identity matrix, is ultrametric. The hierarchical trees corresponding to such uniform matrices are the most simple: they have only one node and uniform branch lengths (equal to λ). Consequently, we demonstrate again with Proposition 4 that all categories are retained to maximize the Gini–Simpson index, which is one of the well-known properties of this index. The consequences of this proposition are more important because it shows that this property of the Gini–Simpson index is actually satisfied by any index resulting from the application of an ultrametric dissimilarity matrix to quadratic entropy.

4. Case studies

We illustrate the influence of the dissimilarity matrix on the maximization of quadratic entropy with three different examples. All computations and graphical displays involved in the preparation of this paper were carried out using program language R (Ihaka and Gentleman, 1996), with both pre-programmed and personal routines. The data and routines are available in the ade4 package at http://lib.stat.cmu.edu/R/CRAN/.

In the first example we define one-dimensional dissimilarities by applying the Euclidean distance to a single quantitative variable: the weight of 70 species of terrestrial Carnivora (Diniz-Filho and Tôrres, 2002)

('carni70' data in the ade4 package). The dissimilarity between two species is simply calculated by their difference in weight (kg). An index of weight diversity is defined by applying quadratic entropy to these dissimilarities. This index is actually equal to the variance in weight. To reach the maximal value of this index in the set of all theoretical frequency distribution of these 70 species, we must eliminate all the species except the lightest (*Mustela nivalis*, the least weasel) and the heaviest (*Ursus arctos*, the grizzly bear) whose relative frequencies must equal 0.5.

The second example considers two-dimensional dissimilarities by applying the Euclidean distance to two quantitative variables: the length of the tarsus and the length of the median toe of 42 bird species living in Burgundy (France) (ecomor data (Blondel et al., 1984) in the ade4 package). These two variables provides a good indication of the relative hardness of foraging substrates. Hard substrate is associated with a short tarsus and a relatively long median toe, whereas soft substrate is associated with a long tarsus and relatively short median toe (Blondel et al., 1984). Both variables are divided by the cube root of weight, which is a way of eliminating size effect. Note t_i, c_i and w_i the tarsus length, median toe length and weight of species i, respectively. The dissimilarity between species i and j is

$$\delta_{ij} = \sqrt{\left(\frac{t_i}{w_i^{1/3}} - \frac{t_j}{w_j^{1/3}}\right)^2 + \left(\frac{c_i}{w_i^{1/3}} - \frac{c_j}{w_j^{1/3}}\right)^2}.$$

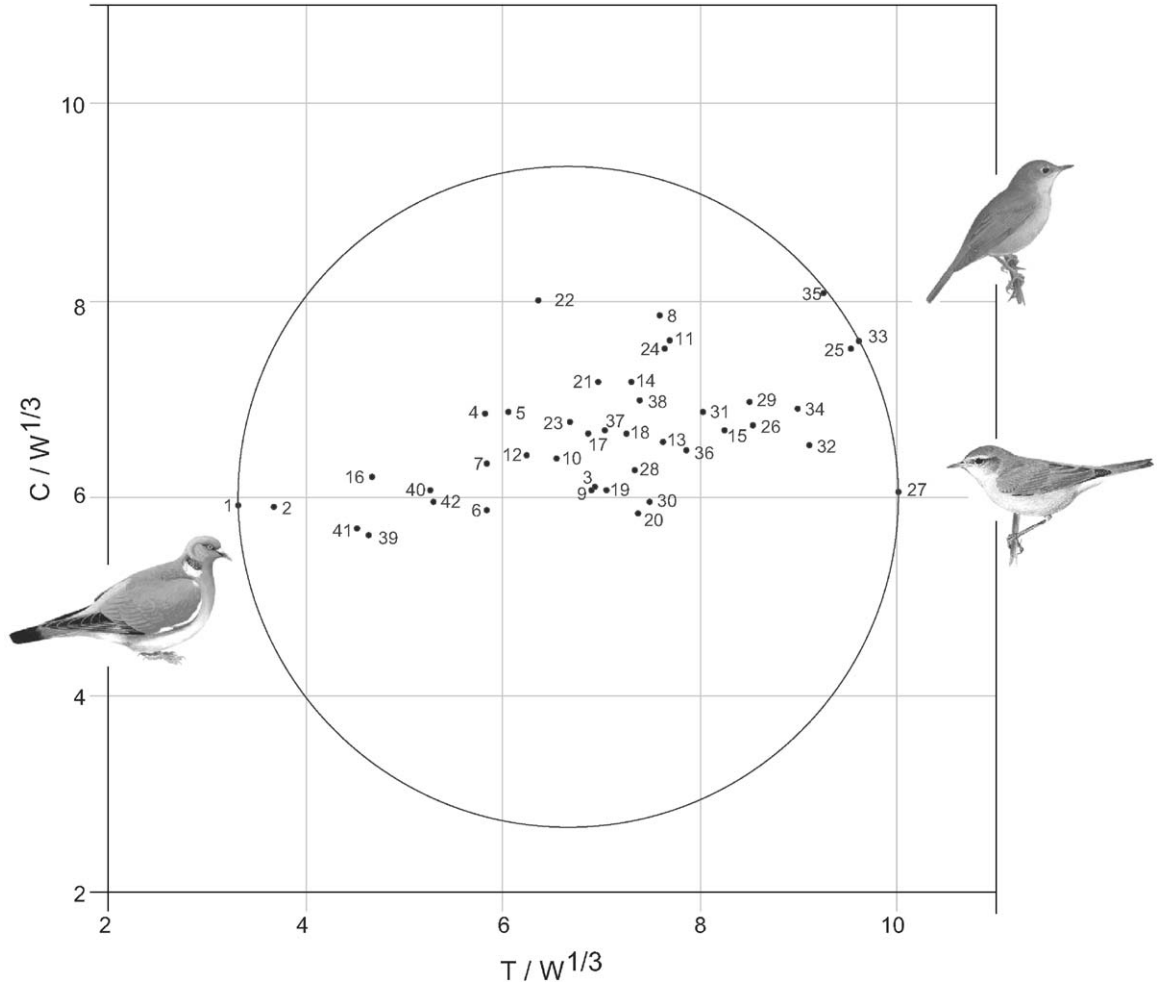


Fig. 1. Positions of the Burgundy bird species in the morphometric space defined by tarsus length (T) and median-toe length (C), both divided by the cube root of weight (W). The drawn circle is the smallest circle which encloses all bird species. Numbers indicate the following species : (1) *C. palumbus*, (2) *Streptopelia turtur,* (3) *Aegithalos caudatus*, (4) *Acanithis cannabina*, (5) *Carduelis carduelis*, (6) *Coccothraustes coccothraustes*, (7) *Pyrrhula pyrrhula*, (8) *Certhia brachydactyla*, (9) *Garrulus glandarius*, (10) *Emberiza citrinella*, (11) *Emberiza schoeniclus*, (12) *Fringilla coelebs*, (13) *Lanius collurio*, (14) *Anthus trivialis*, (15) *Motacilla alba*, (16) *Oriolus oriolus*, (17) *Parus palustris*, (18) *Parus montanus*, (19) *Parus caeruleus*, (20) *Parus major*, (21) *Prunella modularis*, (22) *Sitta europaea*, (23) *Sturnus vulgaris*, (24) *Locustella naevia*, (25) *Phylloscopus trochilus*, (26) *Phylloscopus sibilatrix*, (27) *P. collybita*, (28) *Sylvia borin*, (29) *Sylvia communis*, (30) *Sylvia atricapilla*, (31) *Troglodytes troglodytes*, (32) *Erithacus rubecula*, (33) *E. megarhynchos*, (34) *Phoenicurus phoenicurus*, (35) *Saxicola torquata*, (36) *Turdus philomelos*, (37) *Turdus viscivorus*, (38) *Turdus merula*, (39) *Picoides minor*, (40) *Picoides medius*, (41) *Picoides major*, (42) *Picus canus*. Bird pictures have been taken from Bird Guides http://www.birdguides.com.

The diversity index obtained by applying quadratic entropy to these dissimilarities is a simple morphological diversity index. The interest of considering only two variables is that the dissimilarity matrix can be represented by a scatter of points which can be drawn in a plane. The smallest circle enclosing all the points can thus be displayed (Fig. 1). We search for the theoretical frequencies of the bird species which maximizes this index of morphological diversity. These theoretical frequencies retain only three species: *Columba palumbus* with a frequency of 0.50, *Phylloscopus collybita* with a frequency of 0.49 and *Erithacus megarhynchos* with a frequency of 0.01. These three species lie on the boundary of the smallest circle enclosing all the species under consideration (see Fig. 1). Moreover, over all the species of concern, *C. palumbus* has the shortest tarsus proportionally to its weight, whereas *P. collybita* and *E. megarhynchos* have the longest. All intermediate sizes have been eliminated to reach the maximal value of the morphological diversity. *C. palumbus* has its median toe longer than its related tarsus whereas *P. collybita* and *E. megarhynchos* have their median toe shorter than their tarsus. The species for which such differences in length are small must be eliminated to reach the maximal value of the morphological diversity calculated with quadratic entropy. To complete this second example, two-dimensional simulations are provided in Fig. 2 and Table 1. They show the behavior of the quadratic entropy with a chosen sample of two-dimensional dissimilarities.

As a third example, we search for the theoretical frequency distribution between 18 cattle breeds which maximizes quadratic entropy applied to ultrametric genetic dissimilarities among these cattle breeds (Fig. 3) ('microsatt' data (Moazami-Goudarzi et al., 1997) in the ade4 package). The ultrametric dissimilarities among cattle breeds are computed from the minimal spanning tree obtained with Nei's genetic distance (Nei, 1972, 1978) as the path length between them considered in a pairwise fashion. Those dissimilarities can be associated with a scatter of points in a 17-dimensional Euclidean space. All the points are located on the boundary of the smallest enclosing hypersphere. Consequently, the theoretical distribution

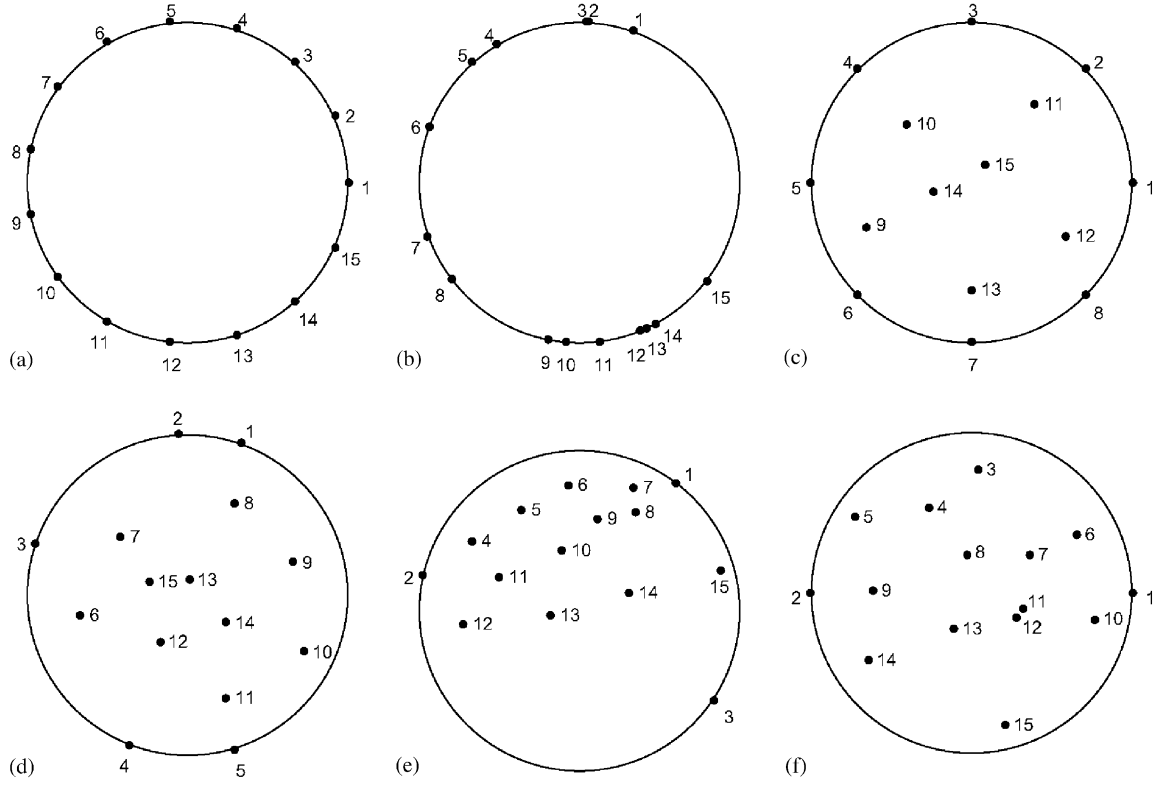


Fig. 2. Two-dimensional simulations. Each simulation involves 15 points. The circle of each simulation is the smallest circle including the 15 points. (a) Figure was built by uniformly distributing the points on a unit circle. In (b), we randomly placed these points on a unit circle until we found that this circle was the smallest including all the points. In (c), eight points were uniformly placed on the unit circle and seven points were randomly positioned into that circle. In (d), five points were randomly placed on the unit circle and ten points were randomly positioned into that circle; this operation was repeated until we found that the circle corresponds to the smallest circle including all the 15 points. A similar process was used for (e) except that only three points were located on the unit circle. In (f), two points were placed on each side of the diameter of the unit circle; the other 13 points were randomly placed into that circle. This last figure illustrates that only two points can support the smallest circle enclosing a large set of points.

Table 1
$\mathbf{p}_{max}$ vectors corresponding to simulations in Fig. 2 (labels of the simulations in columns and labels of points in rows)

	a	b	c	d	e	f
1	0.0667	0.1037	0.1250	0.2443	0.1600	0.5
2	0.0667	0.0962	0.1250	0.1916	0.4393	0.5
3	0.0667	0.0958	0.1250	0.0874	0.4007	0
4	0.0667	0.0747	0.1250	0.1937	0	0
5	0.0667	0.0671	0.1250	0.2830	0	0
6	0.0667	0.0502	0.1250	0	0	0
7	0.0667	0.0356	0.1250	0	0	0
8	0.0667	0.0348	0.1250	0	0	0
9	0.0667	0.0462	0	0	0	0
10	0.0667	0.0497	0	0	0	0
11	0.0667	0.0564	0	0	0	0
12	0.0667	0.0661	0	0	0	0
13	0.0667	0.0677	0	0	0	0
14	0.0667	0.0699	0	0	0	0
15	0.0667	0.0859	0	0	0	0

that maximizes the resulting genetic diversity retains all the species. It also matches the tree. For example, the Zébu Choa, Zébu M'Bororo, Borgou and Zébu Peul, which are isolated from the other breeds in the tree, are given high weights, and the Borgou and Zébu Peul, which are more different from each other than are the Zébu Choa and Zébu M'Bororo, have the highest weights.

5. Conclusion

When dissimilarities are defined to characterize the differences among species (e.g., from morphological or genetic criteria), it is not assured that they belong to the class of dissimilarity defined in this paper. Consequently, if it has been proven that the dissimilarities of interest for a particular data set are not ultrametric, or

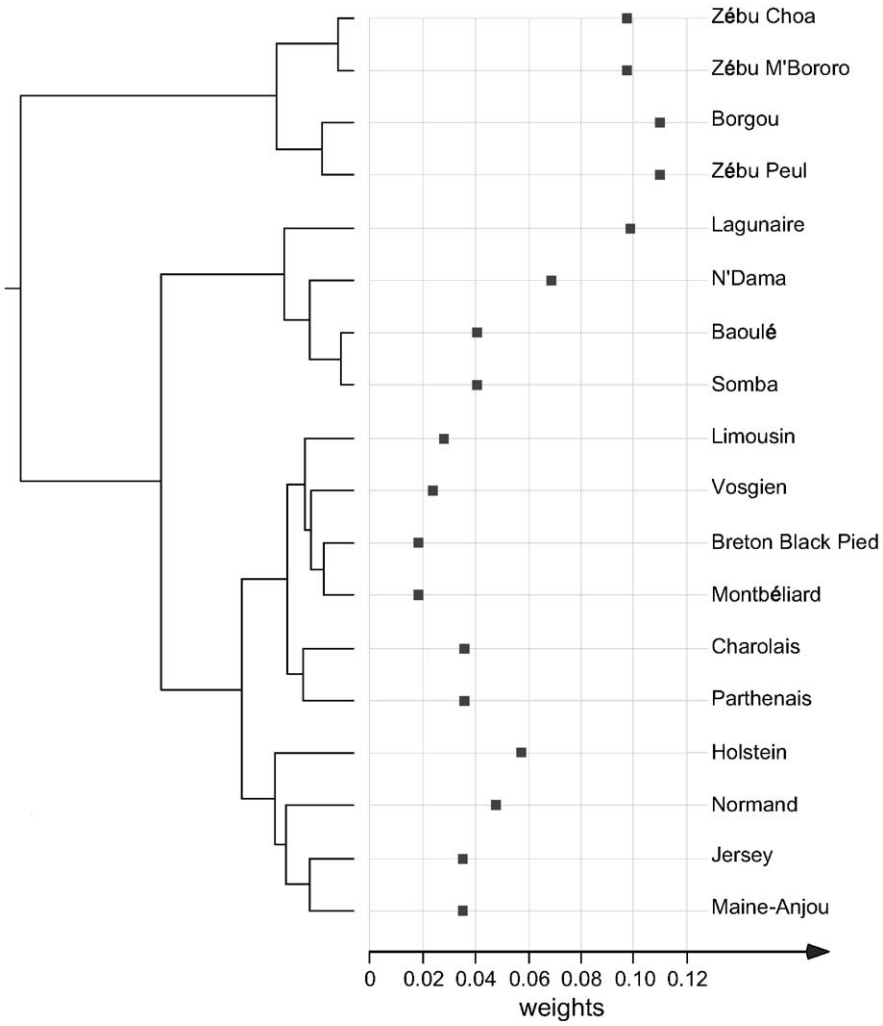


Fig. 3. Frequency distribution among cattle breeds to maximize the genetic diversity among them according to the quadratic entropy. On the left the tree is drawn describing the genetic relationships among the cattle breeds. On the right a Cleveland's dot plot (Cleveland, 1994) gives the frequency distribution of the cattle breeds which maximizes the genetic diversity. The scale of the Cleveland's dot plot is horizontally indicated; each weight is given by the position of a closed square. For example, the frequency of Baoulé is 0.04.

more generally that they do not belong to the strong subclass of the set of SEH-circum-Euclidean matrices, then the interpretation of results obtained from quadratic entropy must be made carefully.

One particular dissimilarity, highlighted by the first example, is the Euclidean metric applied to only one variable. In this case the quadratic entropy is equal to the variance. A set in which 50% of the entities have the lowest possible value and the remaining 50% have the highest possible value would have the greatest variance. If this variable is the body size of a species, the community containing only the smallest and the largest species in equal proportion will have the highest size diversity according to the quadratic entropy. However, it is reasonable to believe that a community in which species have a large range of body size should have the greatest measure of size diversity. From a biological conservation point of view, such a community is less likely to be threatened with extinction because it has a high adaptability while having limited competition (Etienne and Olff, 2004).

A similar result is observed when the Euclidean distance is applied to two variables: to maximize the morphological diversity of birds in Burgundy, all species are eliminated except three. In the morphologic space defined by the lengths of bird tarsus and median toe, these three species lie on the boundary of the smallest circle including all the 42 bird species living in Burgundy. The simulations associated with this example show that the points strictly enclosed in the circle are eliminated no matter their exact positions (Figs. 2c–f). Because the simulations are in two dimensions, they permit the graphical comparison among two SEH-circum-Euclidean matrices of dissimilarities (Figs. 2a and b) and show with graphical display that many matrices of dissimilarities do not belong to the SEH-circum-Euclidean matrices (Figs. 2c and f). Although these simulated matrices are not based on biological criteria, similar situations can appear with biological data.

In our third example, the genetic distances among cattle breeds have been slightly changed to become ultrametric. We have demonstrated that ultrametric distances have an interesting property: when using them, all categories are needed to maximize the diversity as measured by quadratic entropy. Sometimes dissimilarities are ultrametric without any transformation as when, in community, organisms are divided up among species which are differentiated by their taxonomic or phylogenetic connections (Izsak and Papp, 1995; Warwick and Clarke, 1995). Hence taxonomic and phylogenetic dissimilarities obtained from rooted trees in which all the end nodes are equidistant from the root of the tree are ultrametric (Van de Peer, 2003, p. 103). Thus it is expected that quadratic entropy associated with ultrametric matrices will have an important impact on the measure of diversity, especially in ecology.

To sum up, the behavior of quadratic entropy in general depends on the choice of dissimilarity index. Differences in measures of diversity have been illustrated by the distinction between the Gini–Simpson index and the variance, both of which are particular cases of quadratic entropy: the former considers that the set with maximal diversity has the maximal number of evenly represented categories; the latter is maximal only when the two most extreme categories are evenly represented. Many intermediate grades exist. Ultrametric dissimilarities offer a relevant compromise between these two extremes.

Acknowledgments

We thank S. Champely for helpful discussions. This work was partially supported by the BRG and the French Ministry in charge of ecology (MEDD)—contract 14-A/2003.

Appendix. Proofs

Proof of Proposition 1.

We search for the maximal value of the quadratic entropy in the set $\{\mathbf{p} = (p_1 \cdots p_n), p_i \geqslant 0, \sum_{i=1}^{n} p_i = 1\}$. Denote $\mathbf{x}_i$ and $\mathbf{x}_j$ the $K \times 1$ vectors containing the coordinates of the points M_i and M_j, respectively. The quadratic entropy can be written as

$$H_{\Delta}(\mathbf{p}) = \sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j \frac{\delta_{ij}^2}{2} = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j (\mathbf{x}_i - \mathbf{x}_j)^t(\mathbf{x}_i - \mathbf{x}_j). \tag{A.1}$$

Let O be the center of the smallest hypersphere which encloses the n points M_i. Insert the coordinates of O into the above expression

$$\begin{aligned} H_{\Delta}(\mathbf{p}) &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j (\mathbf{x}_i - \mathbf{x}_j)^t(\mathbf{x}_i - \mathbf{x}_j) \\ &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j ((\mathbf{x}_i - \mathbf{x}_O) \\ &\quad - (\mathbf{x}_j - \mathbf{x}_O))^t((\mathbf{x}_i - \mathbf{x}_O) - (\mathbf{x}_j - \mathbf{x}_O)) \\ &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j (\mathbf{x}_i - \mathbf{x}_O)^t(\mathbf{x}_i - \mathbf{x}_O) \\ &\quad + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j (\mathbf{x}_j - \mathbf{x}_O)^t(\mathbf{x}_j - \mathbf{x}_O) \\ &\quad - \sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j (\mathbf{x}_i - \mathbf{x}_O)^t(\mathbf{x}_j - \mathbf{x}_O) \end{aligned} \tag{A.2}$$

which is equal to

$$H_{\Delta}(\mathbf{p}) = \sum_{i=1}^{n} p_i(\mathbf{x}_i - \mathbf{x}_O)^t(\mathbf{x}_i - \mathbf{x}_O) - \left(\sum_{i=1}^{n} p_i\mathbf{x}_i - \mathbf{x}_O\right)^t \left(\sum_{j=1}^{n} p_j\mathbf{x}_j - \mathbf{x}_O\right). \quad (A.3)$$

Let $B(\mathbf{p})$ be the barycenter whose vector of coordinates is $\sum_{i=1}^{n} p_i\mathbf{x}_i$. From (1), (2) and (3)

$$H_{\Delta}(\mathbf{p}) = \sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j \frac{\delta_{ij}^2}{2} = \sum_{i=1}^{n} p_i||M_i - O||^2 - ||B(\mathbf{p}) - O||^2.$$

Let R_O be the radius of the smallest hypersphere which encloses the n points M_i. Insert R_O into the above expression

$$H_{\Delta}(\mathbf{p}) = \sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j \frac{\delta_{ij}^2}{2} = R_O^2 - \sum_{i=1}^{n} p_i(R_O^2 - ||M_i - O||^2) - ||B(\mathbf{p}) - O||^2. \quad (A.4)$$

Because the n points M_i are inside or on the boundary of their smallest enclosing hypersphere, the term $\sum_{n=1}^{n} p_i(R_O^2 - ||M_i - O||^2)$ is nonnegative. The smallest enclosing hypersphere of the n points M_i is the circumscribed sphere of the support set T (Gower, 1982, 1984). Moreover, the circumcenter of this support set T is contained in the convex hull of T, which means that O is a barycenter of the points in T weighted by nonnegative coefficients.

Hence, we deduce from (4) that the maximal value of the quadratic entropy in the set $\{\mathbf{p} = (p_1 \cdots p_n), p_i \geqslant 0, \sum_{i=1}^{n} p_i = 1\}$ is reached at $\mathbf{p}_{\max} = (p_{\max(1)} \cdots p_{\max(n)})$ such that $B(\mathbf{p}) = \sum_{i=1}^{n} p_{\max(i)} x_i$ is equal to O, and consequently $H_{\Delta}(\mathbf{p}_{\max}) = R_O^2$.

Proof of Proposition 3.

By proving Proposition 1, we showed that the center of the smallest enclosing hypersphere is a barycenter of the points in T weighted by nonnegative coefficients. Consequently, the frequencies in $\mathbf{p}_{\max}$ of the k points in the support set T are nonnegative and the frequencies in $\mathbf{p}_{\max}$ of the $n-k$ points located strictly inside the smallest enclosing hypersphere are null. The proof of $\mathbf{p}_T = (\mathbf{D}_T^{-1}\mathbf{1}_k)/(\mathbf{1}_k^t\mathbf{D}_T^{-1}\mathbf{1}_k)$ is given by Gower (Martinez et al., 1994; Nabben and Varga, 1994) considering that the k points in the support set T lie on the surface of a hypersphere.

Proof of Proposition 4.

Let us recall that a $n \times n$ matrix $\Delta = [\delta_{ij}]$ is ultrametric if and only if $\delta_{ij} \geqslant 0$, for all i and j, $\delta_{ij} \leqslant \max(\delta_{ik}, \delta_{kj})$, for all i, j and k, and $\delta_{ii} < \min_{j \neq i}(\delta_{ij})$, for all j ($\delta_{ii} = 0$). A $n \times n$ matrix $\mathbf{A} = [a_{ij}]$ is strictly ultrametric if and only if $a_{ij} \geqslant 0$, for all i and j, $a_{ij} \geqslant \min(a_{ik}, a_{kj})$, for all i, j and k, and $a_{ii} > \max_{j \neq i}(a_{ij})$, for all j. A strictly ultrametric matrix is nonsingular, and its inverse $\mathbf{A}^{-1} = [\alpha_{ij}]$ is a strictly diagonally dominant Stieltjes matrix, i.e. $\alpha_{ij} \leqslant 0$ for all $i, j, j \neq i$, and $\alpha_{ii} > \sum_{j \neq i}|\alpha_{ij}|$, for all i (Critchley and Fichet, 1997).

A $n \times n$ ultrametric matrix $\Delta = [\delta_{ij}]$ is circum-Euclidean (Gower, 1982, 1984). Consequently, a set S of n points M_i ($i = 1, \ldots, n$) can be embedded in a Euclidean space so that the Euclidean distance between M_i and M_j is δ_{ij}. By definition, these points are on the boundary of their circumsphere. Let $\mathbf{X}$ be the matrix whose ith row gives the coordinates of M_i. The coordinates of the circumcenter of the points in S are given by $\mathbf{s}^t\mathbf{X}$, where $\mathbf{s} = \mathbf{D}^{-1}\mathbf{1}_n/\mathbf{1}_n^t\mathbf{D}^{-1}\mathbf{1}_n$ and $\mathbf{D} = [\frac{1}{2}\delta_{ij}^2]$ with $1 \leqslant i \leqslant n$ and $1 \leqslant j \leqslant n$ (Gower, 1982, 1984). This circumcenter belongs to conv(S) if and only if $\mathbf{s}$ has only positive values. Hence, Δ is SEH-circum-Euclidean if and only if $\mathbf{s}$ contains only positive values.

As Δ is ultrametric, $\mathbf{D}$ is also ultrametric (the proof is obtained by the fact that the relations defining the class of ultrametric dissimilarities (see above) are not changed by applying an increasing function to δ_{ij}). Denote $\sigma(\mathbf{D}) = \max_{i,j}(\frac{1}{2}\delta_{ij}^2)$ and $\tilde{\mathbf{D}} = \sigma(\mathbf{D})\mathbf{1}_n\mathbf{1}_n^t - \mathbf{D}$. If $\mathbf{D}$ is ultrametric then $\tilde{\mathbf{D}}$ is strictly ultrametric.

According to the Sherman-Morrison formula

$$\mathbf{D}^{-1} = -\tilde{\mathbf{D}}^{-1} - \sigma(\mathbf{D})\frac{\tilde{\mathbf{D}}^{-1}\mathbf{1}_n\mathbf{1}_n^t\tilde{\mathbf{D}}^{-1}}{1 - \sigma(\mathbf{D})\mathbf{1}_n^t\tilde{\mathbf{D}}^{-1}\mathbf{1}_n}$$

Consequently,

$$\mathbf{D}^{-1}\mathbf{1}_n = \frac{-\tilde{\mathbf{D}}^{-1}\mathbf{1}_n(1 - \sigma(\mathbf{D})\mathbf{1}_n^t\tilde{\mathbf{D}}^{-1}\mathbf{1}_n) - \sigma(\mathbf{D})\tilde{\mathbf{D}}^{-1}\mathbf{1}_n\mathbf{1}_n^t\tilde{\mathbf{D}}^{-1}\mathbf{1}_n}{1 - \sigma(\mathbf{D})\mathbf{1}_n^t\tilde{\mathbf{D}}^{-1}\mathbf{1}_n} = \frac{-\tilde{\mathbf{D}}^{-1}\mathbf{1}_n}{1 - \sigma(\mathbf{D})\mathbf{1}_n^t\tilde{\mathbf{D}}^{-1}\mathbf{1}_n} \quad (A.5)$$

$$\mathbf{1}_n^t\mathbf{D}^{-1}\mathbf{1}_n = \frac{-\mathbf{1}_n^t\tilde{\mathbf{D}}^{-1}\mathbf{1}_n}{1 - \sigma(\mathbf{D})\mathbf{1}_n^t\tilde{\mathbf{D}}^{-1}\mathbf{1}_n}. \quad (A.6)$$

Because the matrix Δ is 'circum-Euclidean', $\mathbf{1}_n^t\mathbf{D}^{-1}\mathbf{1}_n$ is positive. Moreover, as the matrix $\tilde{\mathbf{D}}$ is strictly ultrametric, by definition $\mathbf{1}_n^t\mathbf{D}^{-1}\mathbf{1}_n$ is positive. Consequently, according to (6), $1 - \sigma(\mathbf{D})\mathbf{1}_n^t\tilde{\mathbf{D}}^{-1}\mathbf{1}_n$ is negative. The matrix $\tilde{\mathbf{D}}$ being strictly ultrametric, $\tilde{\mathbf{D}}^{-1}\mathbf{1}_n$ has only positive values. Therefore according to (5), $\mathbf{D}^{-1}\mathbf{1}_n$ has only positive values. As the vector $\mathbf{s} = \mathbf{D}^{-1}\mathbf{1}_n/\mathbf{1}_n^t\mathbf{D}^{-1}\mathbf{1}_n$ contains only positive values, any ultrametric matrix is

SEH-circum-Euclidean and belongs to the strong subclass defined in the main text.

References

Blondel, J., Vuilleumier, F., Marcus, L.F., Terouanne, E., 1984. Is there ecomorphological convergence among mediterranean bird communities of Chile, California, and France. In: Hecht, M.K., Wallace, B., MacIntyre, R.J. (Eds.), Evolution Biology. Plenum Press, New York, pp. 141–213.

Champely, S., Chessel, D., 2002. Measuring biological diversity using Euclidean metrics. Environ. Ecol. Stat. 9, 167–177.

Cleveland, W.S., 1994. The Elements of Graphing Data. AT&T Bell Laboratories, Murray Hill, New Jersey.

Cousins, S.H., 1991. Species diversity measurement: choosing the right index. Trend. Ecol. Evol. 6, 190–192.

Critchley, F., Fichet, B., 1997. On (super-)spherical distance matrices and two results from Schoenberg. Linear Algebra Appl. 251, 145–165.

Diniz-Filho, J.A.F., Tôrres, N.M., 2002. Phylogenetic comparative methods and the geographic range size—body size relationship in New World terrestrial Carnivora. Evol. Ecol. 16, 351–367.

Edwards, A.W.F., 1971. Distance between populations on the basis of gene frequencies. Biometrics 27, 873–881.

Etienne, R.S., Olff, H., 2004. How dispersal limitation shapes species—body size distributions in local communities. Am. Naturalist 163, 69–83.

Fisher, K., Gärtner, B., Kutz, M., 2003. Fast smallest-Enclosing-Ball computation in High dimensions. In: Di Battista, G., Zwick, U. (Eds.), Proceedings of the 11th Annual European Symposium on Algorithms (ESA 2003). Springer, Budapest, pp. 630–641.

Gower, J.C., 1982. Euclidean distance geometry. Math. Scientist 7, 1–14.

Gower, J.C., 1984. Distance matrices and their Euclidean approximation. In: Diday, E., Jambu, M., Lebart, L., Pagès, J., Tomassone, R. (Eds.), Data Analysis and Informatics III. Elsevier, Amsterdam, North-Holland, pp. 3–21.

Gower, J.C., 1966. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53, 325–338.

Gower, J.C., Legendre, P., 1986. Metric and Euclidean properties of dissimilarity coefficients. J. Classif. 3, 5–48.

Ihaka, R., Gentleman, R., 1996. A language for data analysis and graphics. J. Comp. Graph. Stat. 5, 299–314.

Izsak, J., Papp, L., 1995. Application of the quadratic entropy indices for diversity studies of drosophilid assemblages. Environ. Ecol. Stat. 2, 213–224.

Izsak, J., Szeidl, L., 2002. Quadratic diversity: its maximization can reduce the richness of species. Environ. Ecol. Stat. 9, 423–430.

Kimura, M., 1983. The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge.

Magurran, A.E., 1988. Ecological Diversity and its Measurement. Croom Helm Limited, London.

Martinez, S., Michon, G., San Martin, J., 1994. Inverse of strictly ultrametric matrices are of Stieltjes type. SIAM J. Matrix Anal. A 15, 98–106.

Moazami-Goudarzi, K., Laloë, D., Furet, J.P., Grosclaude, F., 1997. Analysis of genetic relationships between 10 cattle breeds with 17 microsatellites. Anim. Genetics 28, 338–345.

Nabben, R., Varga, R.S., 1994. A linear algebra proof that the inverse of a strictly ultrametric matrix is a strictly diagonally dominant Stieltjes matrix. SIAM J. Matrix Anal. A 15, 107–113.

Nei, M., 1972. Genetic distance between populations. Am. Naturalist 106, 283–292.

Nei, M., 1978. Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics 89, 583–590.

Nei, M., Li, W.-H., 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA 76, 5269–5273.

Nei, M., Tajima, F., 1981. DNA polymorphism detectable by restriction endonucleases. Genetics 97, 145–163.

Pielou, E.C., 1975. Ecological Diversity. Wiley, New York.

Rao, C.R., 1982a. Diversity and dissimilarity coefficients: a unified approach. Theor. Popul. Biol. 21, 24–43.

Rao, C.R., 1982b. Diversity: its measurement, decomposition, apportionment and analysis. Sankhya: Ind. J. Stat. A 44, 1–22.

Rao, C.R., 1986. Rao's axiomatization of diversity measures. In: Kotz, S., Johnson, N.L. (Eds.), Encyclopedia of Statistical Sciences. Wiley, New York, pp. 614–617.

Rao, C.R., Nayak, T.K., 1985. Cross entropy, dissimilarity measures, and characterizations of Quadratic Entropy. IEEE Trans. Inf. Theory IT-31, 589–593.

Shimatani, K., 2001. On the measurement of species diversity incorporating species differences. Oikos 93, 135–147.

Van de Peer, Y., 2003. Phylogeny inference based on distance methods. In: Salemi, M., Vandamme, A.-M. (Eds.), The Phylogenetic Handbook, a Practical Approach to DNA and Protein Phylogeny. Cambridge University Press, Cambridge, pp. 101–136.

Warwick, R.M., Clarke, K.R., 1995. New 'biodiversity' measures reveal a decrease in taxonomic distinctness with increasing stress. Mar. Ecol. Prog. Ser. 129, 301–305.

Watve, M.G., Gangal, R.M., 1996. Problems in measuring bacterial diversity and a possible solution. Appl. Environ. Microbiol. 62, 4299–4301.

Annexe 4
Pavoine *et al.* 2005 - Ecology Letters

Pavoine, S., S. Ollier, and A. B. Dufour. 2005. Is the originality of a species measurable ? Ecology letters 8 :579-586.

Ecology Letters, (2005) **8**: 579–586 doi: 10.1111/j.1461-0248.2005.00752.x

IDEAS AND PERSPECTIVES

Is the originality of a species measurable?

Sandrine Pavoine,* Sébastien Ollier and Anne-Béatrice Dufour
Laboratoire de Biométrie et Biologie Evolutive, UMR CNRS 5558, Université Claude Bernard LYON I, France
**Correspondence: E-mail: pavoine@biomserv.univ-lyon1.fr*

Abstract

In this paper, we introduce the concept of 'originality of a species within a set' in order to indicate the average rarity of all the features belonging to this species. Using a phylogenetic tree of 70 species of New World terrestrial Carnivora, we suggest measuring the originality by a probability distribution. This maximizes the expected number of features shared by two species randomly drawn from the set. By using this new index, we take account of branch lengths whereas current indices of originality focus on tree topology. As a supplement to Nee and May's optimizing algorithm, we find that originality must be one of the criteria used in conservation planning.

Keywords

Biodiversity, Carnivora, conservation biology, divergence times, feature richness, phylogenetics, quadratic entropy, uniqueness.

Ecology Letters (2005) 8: 579–586

INTRODUCTION

Faced with the acceleration of species extinctions over recent decades, many indices have been introduced which allow us to evaluate the loss of biological diversity from a very local scale to a global scale. In the absence of other information, biodiversity is often measured by species richness i.e. the number of species (Purvis & Hector 2000). However, this index suffers from not considering differences among species (Nei & Li 1979; Cousins 1991; Watve & Gangal 1996). For example, a set containing a species of Felidae, one of Canidae and one of Ursidae should be treated as more diverse than a set containing three different species of Felidae. To quote Cousins (1991) 'it is ironic that species are treated as equal in conventional indices when the very basis of the identification of species is that they are different from each other'. A particular species can be described by its characters; these are its constitutive biochemical products, morphological structures and behavioural traits (Williams & Humphries 1996). In this context, we defer to Faith (1992) by defining diversity as feature richness, where a feature means a particular state of a character (see discussion by Humphries *et al.* 1995).

In order to lay down policies of biological conservation, other indices have been introduced as well which measure the value of the conservation of a species. Faith (1992, 1995) introduced the concept of the complementarity of a given species when compared with a set of species of reference. This complementarity is measured by the strict uniqueness of a species, that is to say the number of features possessed by this species yet not those shared with the others. This strict uniqueness of a species is currently calculated using two methods: the length of the branch or branches of a phylogenetic tree gained from the addition of this species within the set (Faith 1992, 1995) and the probability that this species has a unique feature (Crozier 1992; Crozier & Kusmierski 1994).

The features exhibited by only one species are the least frequent while those shared by all the species are the most frequent. This leads us to introduce the concept of 'rarity of a feature' within a set. Patil & Taillie (1982) defined the concept of rarity of a species by particular decreasing function of its abundance or its frequency in an area. In the same way, the rarity of a feature could be calculated by a decreasing function of its frequency. We define the originality of a species by the average rarity of its features. The whole contribution of a species to feature richness depends on its originality. For example, Fig. 1 shows a hypothetical tree on which features are indicated. Suppose that these features are states of homologies, meaning that a particular feature is only possessed by organisms of common ancestry. All the species have three unique features i.e. they have equal strict uniqueness, but the species A and B have more rare features than the other species because they are isolated on the tree, hence they have greater originality. Consequently, if both species A and B were to become extinct, the feature richness of these 21 hypothetical species would decrease dramatically.

Each species is a unique combination of DNA units contributing to its physiology, its morphology and its

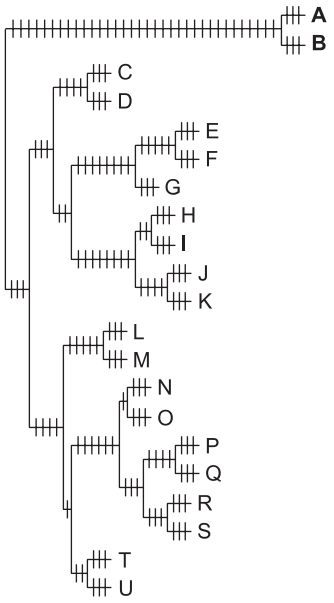


Figure 1 Illustration of the difference between 'strict uniqueness' and 'originality' using a hypothetical tree. Each solid bar along a branch indicates the origin of a new feature (a character change from an ancestral status). Each species has three unique features, that is to say features owned solely by this very species. But contrary to other species, species A and B have 34 rare but not unique features because they are isolated on the tree.

behaviour. Each state of a physiological, morphological and behavioural character is a particular feature. Consequently, all species features cannot usually be counted because they are numerous and also many of them are probably unknown. If it was possible to know all features of all species, we could measure the originality by the rarity functions of Patil & Taillie (1982) applied to feature level instead of species level. As this is not the case, we have to find alternatives, such as using phylogenetic or taxonomic trees to find substitutes.

We start by considering four existing indices based on taxonomic or cladistic trees: Vane-Wright *et al.* (1991) node-counting index, May (1990) branch-counting index and Nixon & Wheeler (1992) unweighted and weighted indices. These indices evaluate the originality of a species without taking into account the lengths of branches in trees. On phylogenetic trees, the length of the branch or branches shared by two species indicates the expected number of features that these species have inherited from their common ancestor; the length of the branch for a single species is a substitute for the number of features that only this species owns. Here we develop a new index, called the quadratic entropy (QE)-based index, which considers these lengths. Nee & May (1997) introduced a process making it possible to find, among several options, the subsets of species which optimize the sum of the branch lengths preserved on a phylogenetic tree when only the species included in these subsets are saved. We show how originality can aid in supplementing this process of optimization. These mathematical models are concretely illustrated using a set of 70 species of New World terrestrial Carnivora.

MATERIALS AND METHODS

Data

A total of 70 species of terrestrial Carnivora living in the New World are considered: 12 Felidae, 14 Canidae, three Ursidae, 13 Procyonidae and 28 Mustelidae. We considered their phylogenetic tree given by Diniz-Filho & Tôrres (2002) from Bininda-Emonds *et al.* (1999) who derived a complete phylogeny for all 271 extant species of the Carnivora. The branch lengths on this tree are expressed as divergence times based on fossil and molecular data.

Existing indices

Four indices have been developed to estimate the originality of a species in a set. Vane-Wright *et al.* (1991) index is proportional to the inverse of the number of nodes between this species and the root of the tree. The most distinct (close-to-root) species have the highest weights. May (1990) proposed counting the number of branches at each node instead of the number of nodes itself. This modification takes into consideration the nodes which are not entirely defined i.e. of which more than two branches diverge. Thus more numerous the species connected to a given node are, the less original they are. Nixon & Wheeler (1992) suggested ranking species according to the amount of phylogenetic diversity in the clades to which they belong and hence defined two indices. For their unweighted index, each node of the phylogenetic tree is characterized using a binary variable: a given node has the value 1 if there are more species in all clades descended from this node than from its sister-nodes and 0 otherwise. With this method, the originality of a particular species is inversely proportional to the sum of the values attributed to all nodes from this species to the root of the tree. The species that belong to the most species-poor clades have the highest originality. According to their weighted index, the originality of a species is inversely proportional to the sum of the species numbers for each subclade to which that species belongs. We will now look at how we determined the need for the new index.

New index

The four indices described above are useful when information on the phylogenetic relationships between species is incomplete. In order to include existing branch lengths, we must take into account the dissimilarity d_{ij} between two

species *i* and *j*. This dissimilarity is calculated from the sum of the length of the branches, which connect each of these species with their first common ancestor on the phylogenetic tree, which is to say by the time period, which separates these species from their common ancestry.

Rao (1982b) introduced QE by suggesting that the diversity of a set of entities divided into categories could be expressed as:

$$Q(\mathbf{p}) = \sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j d_{ij},$$

where p_i is the frequency of the category i of n categories in the set, $\mathbf{p} = (p_1,\ldots, p_i,\ldots, p_n)$ is the frequency distribution, d_{ij} is the dissimilarity between the categories i and j, and n is the number of categories. The index d_{ij} must be a conditionally negative definite function (Rao & Nayak 1985), meaning the matrix $\left[\sqrt{d_{ij}}\right]$ must be euclidean (Critchley & Fichet 1997). Matrix $\left[\sqrt{d_{ij}}\right]$ is euclidean if and only if n points M_i $(i = 1,\ldots,n)$ can be embedded in a euclidean space so that the euclidean distance between M_i and M_j is $\sqrt{d_{ij}}$ (Gower 1966). Each point here represents a category.

Concretely, the QE is equal to the expected dissimilarity between two entities randomly selected using replacement. This was first introduced to ecology by rewriting Simpson (1949) index in order to measure the diversity without imposing the following: (i) the similarity of individuals within a species; (ii) the equality of differences amongst all pairs of species (Hendrickson & Ehrlich 1971). We apply this index to our sample of Carnivora where the categories are the 70 species. As stated above, the dissimilarity between two species is the time period, which separates them from their first common ancestor. The square root of these dissimilarities is ultrametric, i.e. $\sqrt{d_{ij}} \leq \max\left(\sqrt{d_{ik}}, \sqrt{d_{kj}}\right)$, for all i, j and k. One can verify the ultrametric property through dissimilarities associated with rooted trees in which all the end nodes are equidistant from the root of the tree (Van de Peer 2003; Pavoine *et al.* 2005). These dissimilarities appear in taxonomic trees, and some phylogenetic trees such as the tree of 70 species of Carnivora described above. We describe ultrametric dissimilarities as euclidean but, in fact, they are circum-euclidean (Critchley & Fichet 1997), which means that the n points M_i lie on the boundary of a hypersphere or rather a ball in a multidimensional space. The species weights that maximize the QE applied to such dissimilarities are equal to the weights which the n points M_i must have so their barycentre is the centre of their enclosing hypersphere (Pavoine *et al.* 2005). This property is not assured when the phylogenetic tree is not ultrametric or when the leaves of the tree are not equidistant from the root.

In this context, we define a new index of originality, calling it the 'QE-based index'. We will explain this in detail. Let λ_i be the originality of species i and $\lambda = (\lambda_1,\ldots, \lambda_i,\ldots, \lambda_n)$ the distribution of the originalities of all species. We suggest measuring the originalities of the species using frequency distribution which maximizes $Q : Q(\boldsymbol{\lambda}) = \max(Q(\mathbf{p}))$. Let $\mathbf{D}$ be the matrix $[d_{ij}]$, $\boldsymbol{\lambda} = \mathbf{D}^{-1}\mathbf{1}/\mathbf{1}'\mathbf{D}^{-1}\mathbf{1}$ (Pavoine *et al.* 2005). In concrete terms, these species frequencies maximize the expected dissimilarity between two species randomly drawn from the set. Close species are partially redundant (Shimatani 2001). If such species are abundant to the detriment of more distinct species, then the average phylogenetic diversity measured by Q is low. The maximal value of Q is obtained when the most distinct species have the highest weights.

Spearman's rank correlations are computed using this new index as well as the four existing indices described above. All computations and graphical displays were carried out using 'R' (Ihaka & Gentleman 1996), with both pre-programmed and personal routines. The data and routines are available in the 'ade4' package (version >1.3–3) at http://lib.stat.cmu.edu/R/CRAN/ (Chessel *et al.* 2004). They are also available in Appendixes S1–S3: Appendix S1 gives instructions and help, Appendix S2 provides the data set and Appendix S3 contains the functions.

Optimizing feature richness

Faith (1992) estimated feature richness using the sum of the length of all the branches in a phylogenetic tree. He called this index 'phylogenetic diversity'. Nee & May (1997) named it 'amount of evolutionary history' and considered trees whereby all tips were equidistant from the root and where branch lengths were expressed as divergence times. They introduced an algorithm to optimize the amount of independent evolutionary history preserved if only k species out of a total of n were to be saved: the k−1 closest-to-root nodes in a tree are selected, which defines k clades; one species from each clade is picked. This optimizing process also maximizes the average dissimilarity among the species saved and thus the amount of QE retained. Consequently, QE and evolutionary history are optimized using the same subsets of species. By maximizing the amount of QE, we can provide two kinds of information: the subsets of k species which optimize the amount of diversity saved and the respective originalities of these k species in the subsets. We apply Nee & May optimizing process to our sample of Carnivora for various numbers of species saved, then results are compared with random choices of species.

If at any step of the process a species is picked from a clade which has more than one species in it, Nee & May (1997) suggested that the one species be picked at random. This process considers that all these species are equivalent

because they result in the same amount of evolutionary history being preserved. Nevertheless, these species differ by their total originalities, which is to say in the whole set of 70 species. For each percentage of species saved, we calculate the minimal and maximal amounts of originality preserved (sum of species originalities measured by the QE-based index) above all the combinations of species, which optimize the amount of evolutionary history conserved.

RESULTS

The relative originalities of the 70 species of Carnivora are given in Fig. 2. They are expressed as percentages $\lambda_i^* = (\lambda_i(\sum_{i=1}^{n} \lambda_i))$ to facilitate comparisons among indices. The rank correlations among the five indices are relatively high (Table 1). The greatest correlations are between Vane-Wright *et al.* index and Nixon & Wheeler unweighted index (0.94) and between May's index and the QE-based index (0.93). Although the QE-based index is in rough concordance with May's index, it is more sensitive to the individual variation of species originality as opposed to previous metrics, which valued all species equally. According to Vane-Wright *et al.* index, May's index and the QE-based index, from the set of 70 species, the *Tremarctos ornatus* appears to have one of the greatest originalities. According to the QE-based index, it even has the greatest originality. The indices from Nixon & Wheeler (1992), especially their weighted index, apply different ranks to species than those obtained with indices from Vane-Wright *et al.* and May and the QE-based index. In fact, they give higher weights to species of the Felidae, which constitutes a large isolated clade on this tree. Using the weighted index of Nixon &

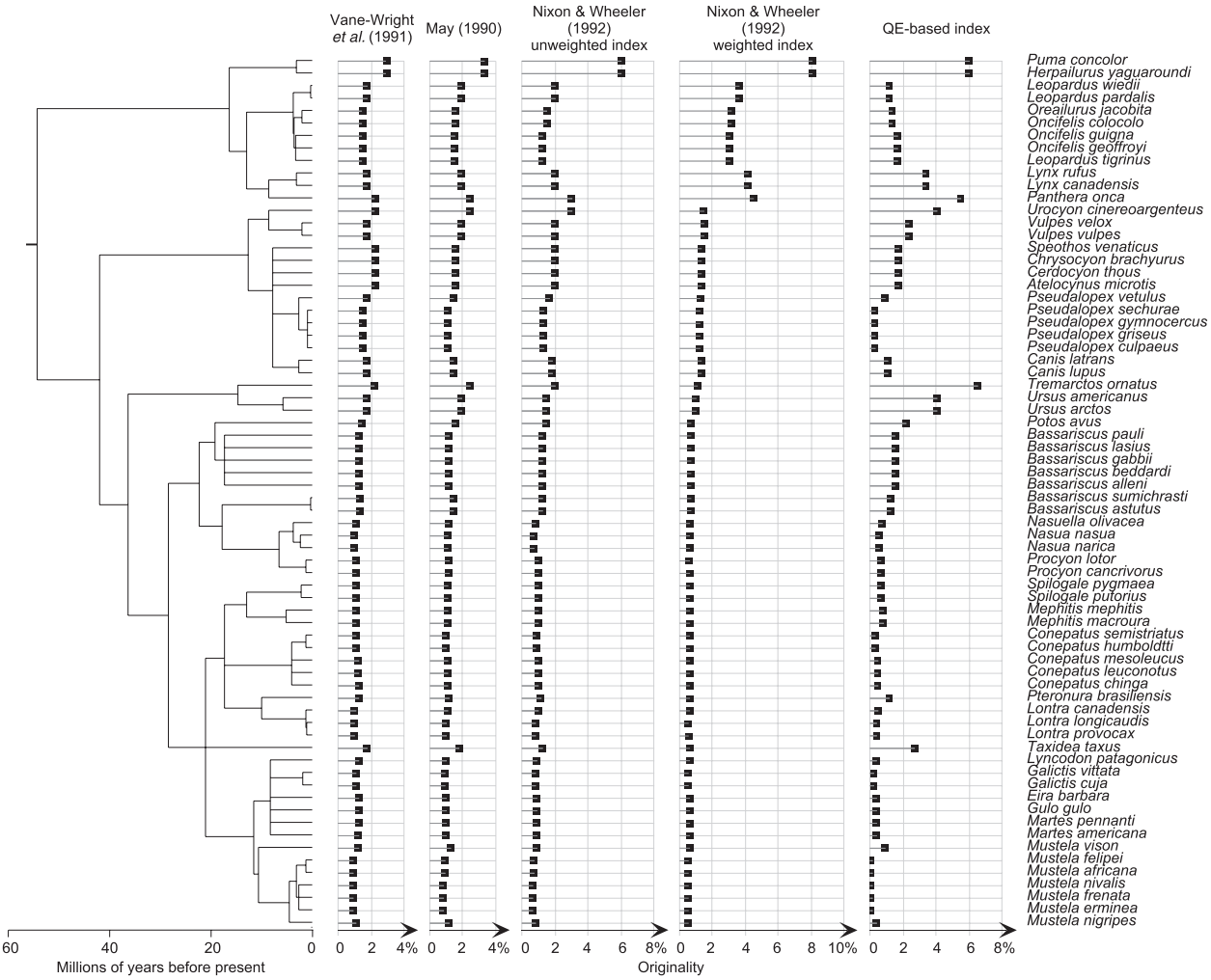


Figure 2 Measures of the originality for 70 species of Carnivora: the phylogeny of these species, Vane-Wright *et al.* (1991) node-counting index, May (1990) branch-counting index, Nixon & Wheeler (1992) unweighted index, Nixon & Wheeler (1992) weighted index, the QE-based index. Originality indices are given by Cleveland's dot plots (Cleveland 1994).

Table 1 Matrix of Spearman rank correlation coefficients among the 70 Carnivora species rankings given by five indices of originality

	VW	M	NWU	NWW
M	0.8503			
NWU	0.9365	0.8844		
NWW	0.8573	0.8733	0.9188	
QE-based	0.7779	0.9328	0.8154	0.7775

VW, Vane-Wright *et al.* (1991) index; M, May (1990); QE-based, QE-based index; NWU, Nixon & Wheeler (1992) unweighted index; NWW, Nixon & Wheeler (1992) weighted index.

Wheeler the differences in originality between species from a single clade are hardly distinguishable and appear 'among' clades much more than 'within' clades.

Interestingly, when 20% of the species are saved at random, 43% of the evolutionary history is preserved on average (Fig. 3a). Yet when these same 20% are saved according to the algorithm of Nee & May, 55% of the evolutionary history is conserved. The difference between the average quantity of evolutionary history preserved from random samples and the quantity of evolutionary history preserved from a sample obtained by the process of optimization is highest when between 30 and 60% of the species are saved. With all the combinations of species optimizing the amount of independent evolutionary history preserved, the difference between the minimal and maximal amounts of originality (sum of species originalities) preserved increases when the number of species saved decreases (Fig. 3b).

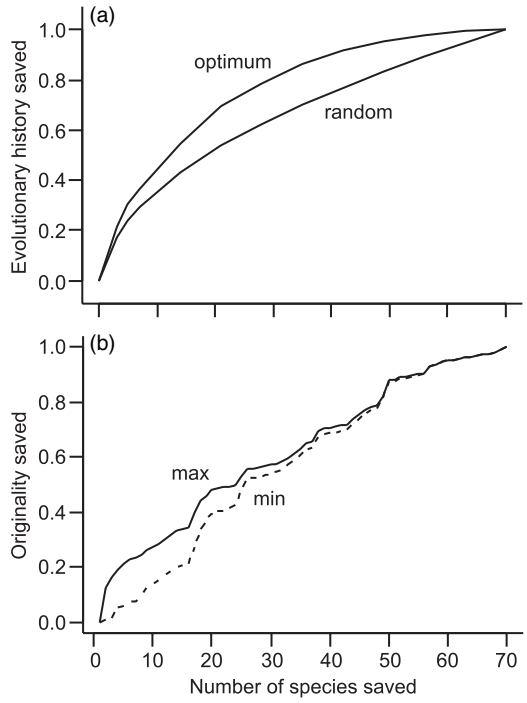


Figure 3 Saving species for biological conservation: (a) evolutionary history saved as a function of the number of species saved. Evolutionary history saved is expressed as a percentage of the total evolutionary history of the 70 species. The top curve is for the optimizing scheme. The bottom curve is for random samples; (b) amount of originality preserved as a function of the number of species saved. Out of all combinations of species optimizing the amount of evolutionary history saved, the solid line is for the maximal amount of originality saved and the dashed line is for the minimal amount of originality saved.

DISCUSSION

Tremarctos ornatus has one of the greatest originalities, first because the three species of Ursidae considered (*T. ornatus*, *Ursus americanus*, *Ursus arctos*) are relatively isolated from the other species on the phylogenetic tree, and second because this species was the first to diverge from the two other species of Ursidae (Bininda-Emonds *et al.* 1999). Depending on indices, it does not always result in the greatest contribution because the Felidae and Canidae species, relatively isolated from other species, are also expected to contribute rare features to the set of 70 Carnivora species. In particular, *Puma concolor* and *Herpailurus yaguaroundi* have the greatest contribution measured by the indices of Vane-Wright *et al.*, May and Nixon & Wheeler because they are the two most original species of the most isolated clade, i.e. Felidae.

The indices of Nixon & Wheeler give higher weights to Felidae species which diverged from the other species approximately 53.8 million years ago (Bininda-Emonds *et al.* 1999). On the contrary, the behaviour of the QE-based index is linked to the behaviour of Vane-Wright *et al.* and May's indices which in fact simultaneously discriminate species within and among clades rather than primarily among large clades. The QE-based index describes the structure of the tree in a way close to May's index; but it has the advantage of considering branch lengths on a phylogenetic tree. Consider for instance the subtree containing *Leopardus wiedii*, *Leopardus pardalis*, *Leopardus tigrinus*, *Oncifelis colocolo*, *Oncifelis guigna*, *Oncifelis geoffroyi* and *Oreailurus jacobita*. The ranks of these species calculated by the QE-based index are in opposition to those given by the indices of Vane-Wright *et al.*, May and Nixon & Wheeler. These species are divided by two nodes forming three groups: group 1 with *L. wiedii* and *L. pardalis*, group 2 containing *Oreailurus jacobita* and *Oncifelis colocolo* and group 3 embedding *Oncifelis guigna*, *Oncifelis geoffroyi* and *L. tigrinus*. Because there are five nodes between the species of group 1 and the root of the tree whereas six nodes separate the species of groups 2 and 3 and the root of the tree, the indices of Vane-Wright *et al.*,

May and Nixon & Wheeler give the first rank to group 1. Because group 3 contains three species while group 2 holds only two species, the indices of May and Nixon & Wheeler give the second rank to group 2 and the third rank to group 3. Actually, only 0.2 million years separate the two nodes which divide the seven species. Consequently, by taking branch lengths into account, species of group 3 are treated as more isolated than those of groups 1 and 2. Thus, the QE-based index gives the first rank to the species of group 3 which diverged 3.2 million years ago, the second rank to the species of group 2 which diverged 1.9 million years ago and the third rank to the species of group 1 which diverged 0.3 million years ago.

This leads us to an important conclusion. When branch lengths are available, the QE-based index provides a relevant measure of species originality. Its sensitivity to branch length makes it dependent on the process providing these lengths as tree topology and branch lengths are both estimated quantities. It is worth noting that because of reversion or convergence, the chronological distance between species will overestimate the unique feature richness separating two species. Conversely, because of saturation of the phylogenetic signal, molecular distance may underestimate chronological distance. Divergence times are not available or even easily knowable for many taxa. Moreover, especially in the case of highly divergent species, branch lengths are more robustly estimated than topology (Crozier 1997). It is interesting to note, then, that the QE-based index is less vulnerable to phylogenetic instability, which makes it greatly superior to the structure based indices of Vane-Wright *et al.* (1991) and May (1990). The QE-based index is also dependent on a model using tree structure and branch length to infer feature-information. Although ultrametric trees provide rough substitutes for the number of features shared by two species, they suppose that all species have similar evolutionary rates. Therefore, we are now studying the behaviour of the QE-based index on non-ultrametric phylogenies to extend its use to the whole set of possible phylogenies and to take into consideration the potentially various evolutionary rates. This research will aim to measure the whole contribution of a species to the feature richness of a set by estimating the number of features it has developed or inherited as well as by evaluating the degree of rarity of these features within the set.

Vane-Wright *et al.* (1991) were the first to define an index of originality, where the weights of the species introduced are one step in a process defining a measure of diversity sensitive to both taxonomic classification and number of species. The sum of weights λ_i is equal to this measure of diversity $\left(\sum_{i=1}^{n} \lambda_i\right)$ and the contribution of each species is expressed as a percentage $\lambda_i / \sum_{i=1}^{n} \lambda_i$. The indices of Vane-Wright *et al.*, May and Nixon & Wheeler were often criticized for the following reason: by measuring the diversity of a set by summing the values of the species within the set, these measures fail to give greater weight to those sets with the most divergent species (Solow *et al.* 1993; Humphries & Williams 1994). However, these indices should not be rejected because if they do not measure diversity, they still measure originality. Very specific attention must be given to this originality which Vane-Wright *et al.* (1991) called contribution to diversity. If the sum of these absolute contributions does not measure diversity it is because these contributions are correlated. It is therefore better to talk about relative contributions expressed as percentages.

When a conservation strategy is developed, several criteria are retained, especially value, vulnerability, endemism and complementarity. As well the value of a species can be economic, pharmaceutical or aesthetic (for a discussion on these aspects see Faith 1995 and references therein). In this paper, we concentrate on the features of the species to evaluate this value. Some of these features can indeed have socio-economic repercussions like the fur of *Mustela vison*, which was for years considered one of the most luxurious on the market. Williams & Humphries (1996) even suggested giving various weights to features by privileging those which are expressed in the phenotype of an organism, or in the phenotypes of its progeny. They then defined three different classes of features: genetic, phenotypic and functional. We cannot take this differentiation into account here because we only have approximate and quantitative information in terms of number of features possessed by each species.

The vulnerability of a species is in general evaluated by the size of its populations and by the expected evolution of the habitats in which it lives. Several attempts to integrate probabilities of extinction into strategies of conservation have been made (Witting & Loeschke 1995). However, our inability to predict the future makes estimation of these probabilities difficult (Pimm *et al.* 1995). In their simulation of random sampling of species, Nee & May (1997) implicitly assumed random extinctions of species. Moreover, numerous recent studies have shown that these extinctions are not independent of taxonomic or phylogenetic relationships (Purvis *et al.* 2000) and that species interactions can result in correlated extinction events (Witting & Loeschke 1995). Species extinctions are thus definitely not random. Taking into account the probability of extinction is certainly a complex problem, which largely exceeds the framework of this paper.

Endemism is generally employed to indicate the species which are present in only one area. By analogy we can use it to indicate the features owned by only one species. Likewise complementarity is generally used to indicate the number of new species, which an area contributes to a set of reference areas. So, by the same analogy, the complementarity of a

species is the number of new features it brings to a set of species (Faith 1992).

Each index studied or quoted in this paper measures one of these four criteria. The strict uniqueness of a species is a measure of complementarity, whereas the originality of this species, which contains, amongst other things, this information of strict uniqueness, characterizes the total value of this species. We stated above that the index of diversity suggested by Vane-Wright *et al.* (1991) has undesirable properties. The diversity indices of Faith (1992) and Nee & May (1997) or those suggested by Barker (2002) and Rao (1982a) are more relevant. On the contrary, the index measuring the contribution of a species to diversity (node-counting index) proposed by Vane-Wright *et al.* (1991), as well as the four indices that we studied, which all used as a starting point the work by Vane-Wright *et al.* (1991) (the index of May, the two indices of Nixon & Wheeler and the QE-based index), are good indices of originality. The QE-based index, which we propose here, reflects the isolation of a species on all levels of the tree. An index of species value based only on strict uniqueness, like that proposed by Barker (2002), is incomplete because it takes into consideration only 'endemism of characters' which is essential yet does not contain all information on the contribution of a species to diversity.

The algorithm of optimization enables us to highlight which species must be saved in order to maximize the amount of evolutionary history preserved. This situation may correspond to the dilemma which consists of taking k species on board Noah's Ark when the $n - k$ other species will die. In reality the situation is as follows where k species are under a plan of conservation and the other ones are left to struggle for existence. At worst, all the $n - k$ other species will die, which corresponds to the Noah's Ark situation. At best, all will survive. Our ignorance of the future does not allow reliable long-term predictions as far as species survival is concerned. Nevertheless, the features of the most original species have a greater probability of being lost than the features of common species (Witting & Loeschke 1995). We discover then, that among all the combinations of species, which optimize the amount of evolutionary history preserved, the conservation of the combination containing the highest originality of the species must be prioritized. From a conservation standpoint, the originality of a species should be one of the criteria used to decide which species must be protected first. Indeed, an original species threatened with short-term extinction must be more carefully protected as opposed to a very common species; this must be performed even if many very common species are threatened with extinction in the future in ways that we presently can not anticipate. Likewise, rare features must be saved rather than common features. The best means we have to save rare features is the actual safeguard of the most original species.

ACKNOWLEDGEMENTS

The authors would like to thank Dominique Pontier, Robert M. May and two anonymous referees for their helpful comments. They all contributed to improve the presentation of this paper.

SUPPLEMENTARY MATERIAL

The following material is available from http://www.blackwellpublishing.com/products/journals/suppmat/ELE/ELE752/ELE752sm.htm

Appendix S1 Instructions for using functions and data of appendixes S2 and S3.
Appendix S2 Data set.
Appendix S3 Functions.

REFERENCES

Barker, G.M. (2002). Phylogenetic diversity: a quantitative framework for measurement of priority and achievement in biodiversity conservation. *Biol. J. Linn. Soc.*, 76, 165–194.

Bininda-Emonds, O.R.P., Gittleman, J.L. & Purvis, A. (1999). Building large trees by combining phylogenetic information: a complete phylogeny of the extant Carnivora (Mammalia). *Biol. Rev. Camb. Philos. Soc.*, 74, 143–175.

Chessel, D., Dufour, A.-B. & Thioulouse, J. (2004). The ade4 package-I – one-table methods. *R News*, 4, 5–10.

Cleveland, W.S. (1994). *The Elements of Graphing Data*. AT & T Bell Laboratories, Murray Hill, NJ.

Cousins, S.H. (1991). Species diversity measurement: choosing the right index. *Trend. Ecol. Evol.*, 6, 190–192.

Critchley, F. & Fichet, B. (1997). On (super-)spherical distance matrices and two results from Schoenberg. *Linear Algebra Appl.*, 251, 145–165.

Crozier, R.H. (1992). Genetic diversity and the agony of choice. *Biol. Conserv.*, 61, 11–15.

Crozier, R.H. (1997). Preserving the information content of species: genetic diversity, phylogeny, and conservation worth. *Annu. Rev. Ecol. Syst.*, 28, 243–268.

Crozier, R.H. & Kusmierski, R.M. (1994). Genetic distances and the setting of conservation priorities. In: *Conservation Genetics* (eds Loeschcke, V., Tomiuk, J. & Jain, S.K). Birkhäser Verlag, Basel, Switzerland, pp. 227–237.

Diniz-Filho, J.A.F. & Tôrres, N.M. (2002). Phylogenetic comparative methods and the geographic range size – body size relationship in New World terrestrial Carnivora. *Evol. Ecol.*, 16, 351–367.

Faith, D.P. (1992). Conservation evaluation and phylogenetic diversity. *Biol. Conserv.*, 61, 1–10.

Faith, D.P. (1995). Phylogenetic pattern and the quantification of organismal biodiversity. In: *Biodiversity Measurement and Estimation* (ed. Hawksworth, D.L.). Chapman & Hall, The Royal Society, London, UK, pp. 45–58.

Gower, J.C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53, 325–338.

Hendrickson, J.A.J. & Ehrlich, P.R. (1971). An expanded concept of "species diversity". *Notul. Nat.*, 439, 1–6.

Humphries, C.J. & Williams, P.H. (1994). Cladograms and trees in biodiversity. In: *Models in Phylogenetic Construction* (eds Scotland, R.W., Siebert, D.M. & Williams, D.). Clarendon, Oxford, UK, pp. 335–352.

Humphries, C.J., Williams, P.H. & Vane-Wright, R.I. (1995). Measuring biodiversity value for conservation. *Annu. Rev. Ecol. Syst.*, 26, 93–111.

Ihaka, R. & Gentleman, R. (1996). R: a language for data analysis and graphics. *J. Comp. Graph. Stat.*, 5, 299–314.

May, R.M. (1990). Taxonomy as destiny. *Nature*, 347, 129–130.

Nee, S. & May, R.M. (1997). Extinction and the loss of evolutionary history. *Science*, 278, 692–694.

Nei, M. & Li, W.-H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. USA*, 76, 5269–5273.

Nixon, K.C. & Wheeler, Q.D. (1992). Measures of phylogenetic diversity. In: *Extinction and Phylogeny* (eds Novacek, M.J. & Wheeler, Q.D.). Columbia University Press, New York, NY, pp. 216–234.

Patil, G.P. & Taillie, C. (1982). Diversity as a concept and its measurement. *J. Am. Stat. Assoc.*, 77, 548–561.

Pavoine, S., Ollier, S. & Pontier, D. (2005). Measuring diversity from dissimilarities with Rao's quadratic entropy: are any dissimilarity indices suitable?. *Theor. Popul. Biol.* (in press).

Pimm, S.L., Russell, G.J., Gittleman, J.L. & Brooks, T.M. (1995). The future of biodiversity. *Science*, 269, 347–350.

Purvis, A. & Hector, A. (2000). Getting the measure of biodiversity. *Nature*, 405, 212–219.

Purvis, A., Agapow, P.-M., Gittleman, J.L. & Mace, G.M. (2000). Nonrandom extinction and the loss of evolutionary history. *Science*, 288, 328–330.

Rao, C.R. (1982a). Diversity and dissimilarity coefficients: a unified approach. *Theor. Popul. Biol.*, 21, 24–43.

Rao, C.R. (1982b). Diversity: its measurement, decomposition, apportionment and analysis. *Sankhya: Ind. J. Stat.*, A44, 1–22.

Rao, C.R. & Nayak, T.K. (1985). Cross entropy, dissimilarity measures, and characterizations of quadratic entropy. *IEEE Trans. Inf. Theory*, IT-31, 589–593.

Shimatani, K. (2001). On the measurement of species diversity incorporating species differences. *Oikos*, 93, 135–147.

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

Solow, A., Polasky, S. & Broadus, J. (1993). On the measurement of biological diversity. *J. Environ. Econ. Manage.*, 24, 60–68.

Van de Peer, Y. (2003). Phylogeny inference based on distance methods. In: *The Phylogenetic Handbook, a Practical Approach to DNA and Protein Phylogeny* (eds Salemi, M. & Vandamme, A.-M.). Cambridge University Press, Cambridge, UK, pp. 101–136.

Vane-Wright, R.I., Humphries, C.J. & Williams, P.H. (1991). What to protect? Systematics and the agony of choice. *Biol. Conserv.*, 55, 235–254.

Watve, M.G. & Gangal, R.M. (1996). Problems in measuring bacterial diversity and a possible solution. *Appl. Environ. Microbiol.*, 62, 4299–4301.

Williams, P.H. & Humphries, C.J. (1996). Comparing character diversity among biotas. In: *Biodiversity. A Biology of Numbers and Differences* (ed. Gaston, K.J.). Blackwell Science, Oxford, UK, pp. 54–76.

Witting, L. & Loeschke, V. (1995). The optimization of biodiversity conservation. *Biol. Conserv.*, 71, 205–207.

Editor, Ross Crozier
Manuscript received 7 January 2005
First decision made 31 January 2005
Manuscript accepted 25 February 2005

Annexe 5
Pavoine, S. 2004. La biodiversité, ça se mesure ? - Prix de l'IFB

Premier prix du concours des jeunes chercheurs 2004 par l'institut français de la biodiversité.
Pavoine, S.. 2004. La biodiversité, ça se mesure ?

La biodiversité, ça se mesure ?

par

Sandrine Pavoine

INTRODUCTION

La biodiversité qu'est-ce ? pourquoi l'étudier ? La diversité est l'état ou le caractère de ce qui est varié (cf. hétérogénéité, pluralité, variété ; antonymes : ressemblance, uniformité, unité) (Girodet 1976). La biodiversité ou diversité biologique est donc la variabilité du monde vivant à toutes les échelles : de l'écosystème à l'ADN. Chaque espèce résulte d'une combinaison unique de plusieurs milliards de bases d'ADN déterminant sa physiologie, sa morphologie, son comportement, son adaptation à un environnement et sa résistance à une modification de cet environnement. Pour définir des priorités de conservation face à l'extinction massive des espèces, il faut pouvoir chiffrer et comparer la biodiversité dans plusieurs régions du monde.

Pour ce faire, beaucoup d'indices ont été proposés. Dans ce contexte, nous présentons un développement d'un indice de diversité. A partir des indices traditionnels, plusieurs auteurs issus de champs différents de la biologie ont participé à la construction d'un outil de mesure de la diversité permettant de tenir compte de plusieurs échelles telles que les habitats, les

espèces et leurs différences phylogénétiques ou taxonomiques. Nous montrons notre participation à cette recherche et nos perspectives.

1. DEFINITION(S) DE LA BIODIVERSITE ET MESURES TRADITIONNELLES

1.1. Les trois indices les plus répandus

En pratique la diversité se mesure dans une collection à partir d'un regroupement de ses entités en catégories. Ces collections peuvent être des zones d'études, les entités des individus et les catégories des espèces. A un autre niveau, les collections peuvent être des populations d'une même espèce, les entités des individus et les catégories des profils d'ADN.

L'indice de diversité le plus utilisé, appelé richesse, est le nombre de catégories (S) moins une afin qu'une collection ne possédant qu'une seule catégorie ait une diversité nulle : $H_r = S - 1$. Deux autres indices corrigent la richesse par les fréquences relatives des catégories. Le premier, l'indice de Shannon-Wiener (Shannon 1948) est issu d'une théorie qui suppose que la diversité peut être mesurée de la même façon que l'information contenue dans un code ou message : $H_{S-W}(\mathbf{p}) = -\sum_{k=1}^{S} p_k \ln(p_k)$, où $\mathbf{p}$ est la distribution de fréquences des catégories et p_k la fréquence de la catégorie k. Le deuxième, l'indice de Gini-Simpson (Gini 1912, Simpson 1949), est égal à la probabilité de tirer dans une collection deux entités appartenant à deux catégories différentes : $H_{G-S}(\mathbf{p}) = 1 - \sum_{k=1}^{S} (p_k)^2$.

1.2. Discussion sur ces indices

Ces trois indices peuvent être réécrits comme des moyennes d'une fonction de rareté $R(p_k)$ (Patil & Taillie 1982), la rareté d'une catégorie diminuant lorsque sa fréquence augmente :

$$S-1=\sum_{k=1}^{S} p_k\left(\frac{1}{p_k}-1\right)\Rightarrow R_1(p_k)=\frac{1}{p_k}-1$$

$$-\sum_{k=1}^{S} p_k \ln(p_k)=\sum_{k=1}^{S} p_k \ln\left(\frac{1}{p_k}\right)\Rightarrow R_2(p_k)=\ln\left(\frac{1}{p_k}\right)$$

$$1-\sum_{k=1}^{S}(p_k)^2=\sum_{k=1}^{S} p_k(1-p_k)\Rightarrow R_3(p_k)=1-p_k$$

Ces trois indices diffèrent par leur sensibilité vis à vis des espèces rares (cf. figure 1). Le plus sensible est la richesse et le moins sensible l'indice de Simpson.

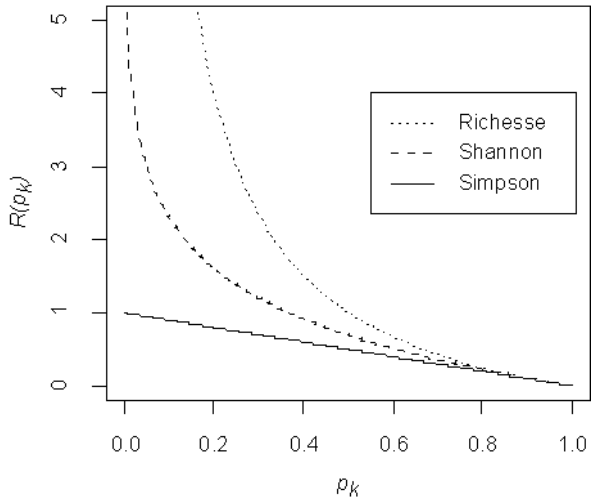


FIGURE 1 : Fonctions de rareté dans la richesse, l'indice de Shannon-Wiener et l'indice de Gini-Simpson.

Deux critiques peuvent être faites à ces indices. La première est que plus une collection possède de catégories, moins les indices de Shannon et de Simpson sont sensibles aux différences de fréquences entre catégories dans cette collection. La mesure de diversité qu'ils fournissent est alors très proche de la richesse. La deuxième critique est que ces indices attribueraient la même diversité à une région dans laquelle seraient présents une autruche, un crocodile et un lion, qu'à une région dans laquelle se trouveraient un campagnol, un mulot et un rat.

2. La biodiversite a plusieurs echelles : l'approche unifiee de Rao

2.1. Mesurer la diversité à partir de dissimilarités

Le premier indice qui peut permettre de tenir compte des différences entre catégories pour mesurer la diversité d'une collection est la variance. La variance d'une variable Y est habituellement vue comme la moyenne des carrés des écarts entre les valeurs y_k pour chaque catégorie et la valeur moyenne $y_\bullet$: $Var(Y) = \sum_{k=1}^{S} p_k \left(y_k - y_\bullet\right)^2$. Cette formule peut être réécrite comme le carré de la différence attendue entre deux entités tirées au hasard dans cette collection : $Var(Y) = \frac{1}{2} \sum_{k=1}^{S} \sum_{l=1}^{S} p_k p_l \left(y_k - y_l\right)^2$.

Si au lieu de mesurer la différence entre catégories à partir d'une variable qualitative, nous considérons n'importe quel type de dissimilarité $\delta_{kl}^{\mathrm{CA}}$ entre deux catégories k et l, cette mesure devient $H_{\Delta^{\mathrm{CA}}}(\mathbf{p}) = \frac{1}{2} \sum_{k=1}^{S} \sum_{l=1}^{S} p_k p_l \left(\delta_{kl}^{\mathrm{CA}}\right)^2$, où Δ^{CA} désigne l'ensemble des

dissimilarités entre catégories. En choisissant $\delta_{kl}^{\mathrm{CA}} = \sqrt{2}$ pour tout $k \neq l$, nous retrouvons l'indice de Gini-Simpson (Rao 1982a).

Cet indice a été indépendamment défini en écologie (Hendrickson & Ehrlich 1971, Warwick & Clarke 1995) et en génétique (Nei & Li 1979). Rao (1982b) le nomme "entropie quadratique" et le replace dans un cadre beaucoup plus général. Il affirme qu'il peut être utilisé dans des champs aussi divers que l'anthropologie, la génétique, l'économie, la sociologie et la biologie.

2.2. Décomposer la diversité

Rao (1982a) définit à partir de cet indice une mesure de la dissimilarité entre deux collections dont les compositions en catégories sont $\mathbf{p}_i$ et $\mathbf{p}_j$: $\delta_{ij}^{\mathbf{CO}} = \sqrt{2\left(2H_{\boldsymbol{\Delta}^{\mathrm{CA}}}\left(\frac{1}{2}\mathbf{p}_i + \frac{1}{2}\mathbf{p}_j\right) - H_{\boldsymbol{\Delta}^{\mathrm{CA}}}\left(\mathbf{p}_i\right) - H_{\boldsymbol{\Delta}^{\mathrm{CA}}}\left(\mathbf{p}_j\right)\right)}$. Il définit alors la diversité entre collections par $H_{\boldsymbol{\Delta}^{\mathrm{CO}}}\left(\boldsymbol{\mu}\right) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mu_i\mu_j\left(\delta_{ij}^{\mathrm{CO}}\right)^2$ (Rao & Nayak 1985), où μ_i et μ_j sont les tailles relatives des collections, $\boldsymbol{\mu}$ l'ensemble de ces tailles relatives et Δ^{CO} désigne l'ensemble des dissimilarités entre collections.

Reprenons le cas particulier où l'entropie quadratique est égale à la variance d'une variable Y. Notons $y_{\bullet i}$ et $y_{\bullet j}$ les valeurs moyennes de Y dans les collections i et j. Cet indice $\delta_{ij}^{\mathbf{CO}}$ est égal à $\left|y_{\bullet i} - y_{\bullet j}\right|$.

La diversité totale (dans l'ensemble des collections mélangées) est divisée en la diversité moyenne au sein des collections plus la diversité entre les collections. Cette décomposition peut s'étendre à d'autres facteurs en plus de la répartition en collections. Elle rassemble dans un même schéma l'analyse de variance classique (Fisher 1925), l'analyse de variance sur variable qualitative (Light & Margolin 1971), la décomposition de la diversité génique (Nei 1973) et l'analyse de variance moléculaire (Excoffier et al. 1992).

3. DEVELOPPEMENTS

3.1. Diversité et typologie

Rao (1986) a placé une contrainte sur le choix des dissimilarités pour que l'entropie quadratique puisse être décomposée. Cette contrainte est que les dissimilarités doivent correspondre à un nuage de points (Rao & Nayak 1985, Rao 1986), chaque point représentant une catégorie.

Champely & Chessel (2002) ont montré que si les dissimilarités entre les catégories ont cette propriété, les dissimilarités entre collections déduites par la théorie de Rao l'ont aussi. A partir des travaux de Gower (1982), nous avons développé une double analyse en coordonnées principales (DPCoA) qui représente à la fois les dissimilarités entre les collections et celles entre les catégories (Pavoine et al., in press). Cette méthode graphique permet une analyse de la structure de la diversité (cf. exemple figure 2).

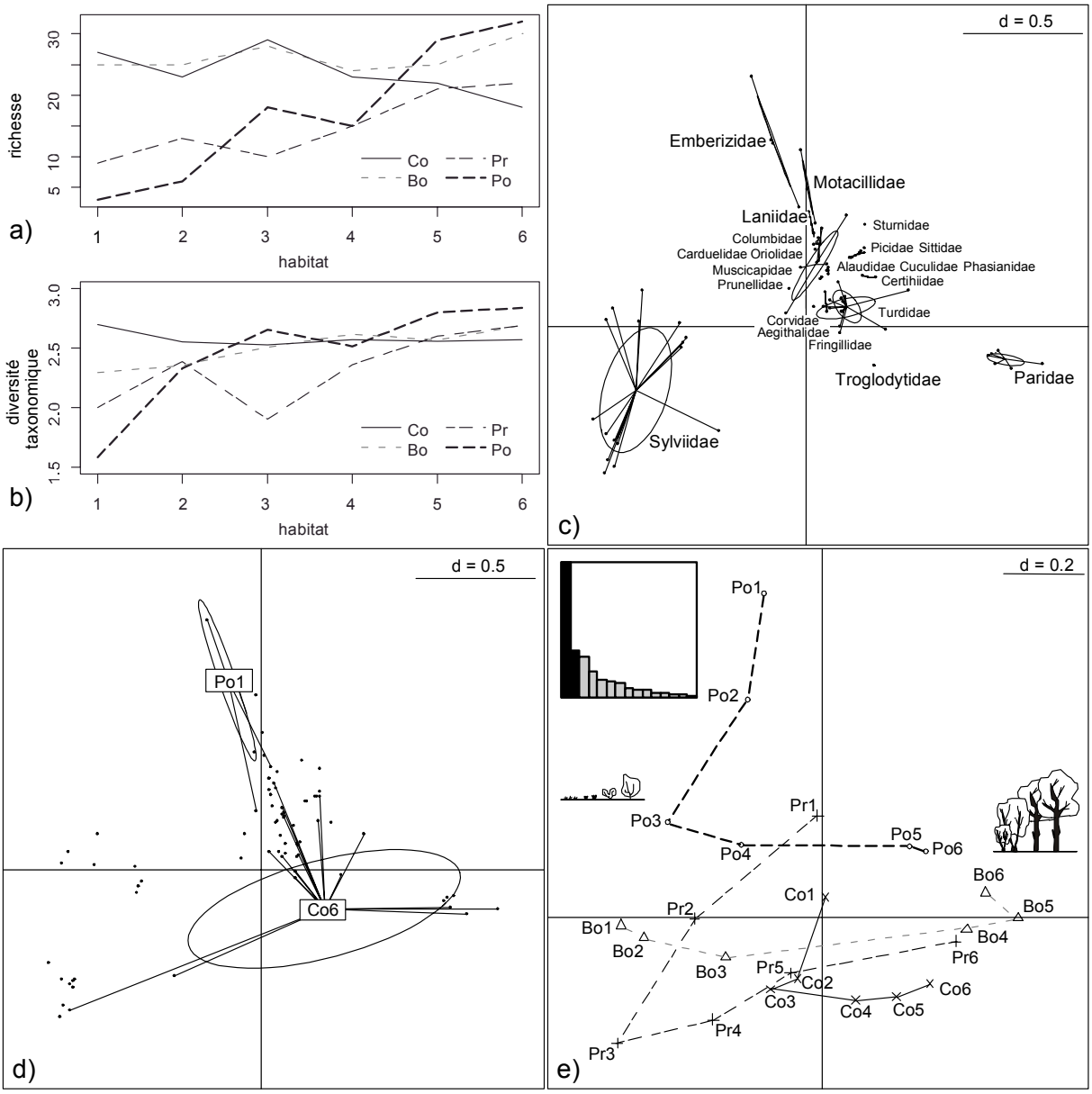


FIGURE 2 : Structure de la diversité taxonomique dans des communautés d'oiseaux (données de Blondel & Farré (1988)). Quatre régions sont comparées : Bourgogne (Bo), Corse (Co), Pologne (Po) et Provence (Pr). Chacune est divisée en six habitats choisis dans un gradient d'ouverture de la végétation (de 1 ouvert à 6 fermé). La diversité taxonomique est mesurée par l'entropie quadratique à partir des dissimilarités entre espèces définies par Warwick & Clarke (1995). Les figures représentent : les modifications de la richesse (a) et la diversité taxonomique (b) dans chaque région ; la structure taxonomique sur la typologie des espèces (c) ; un regroupement des espèces par leur présence dans deux communautés (d), la typologie des communautés (e). Les figures (c) à (e) sont superposables (à un ajustement d'échelle près). Elles résultent de la double analyse en coordonnées principales. La valeur 'd' indique l'échelle. L'encadré de la figure (e) donne la quantité de variation entre communautés représentée par chaque axe.

Exemples d'analyse : Les habitats 2 de ces quatre régions illustrent la propriété suivante : de même que des communautés peuvent avoir la même richesse avec des espèces différentes, elles peuvent avoir la même diversité taxonomique (b) avec des compositions taxonomiques différentes (c+e). Différentes dans les habitats ouverts, les compositions avifaunistiques de ces régions convergent dans les habitats fermés 5 et 6.

3.2. Diversité et richesse

Ainsi grâce à l'entropie quadratique, nous avons un indice de diversité qui généralise la variance et révèle la structure de la biodiversité. Cependant, la valeur de l'entropie quadratique n'est pas bornée entre 0 et 1. Elle dépend du choix des dissimilarités. Pour comparer plusieurs jeux de données, Champely & Chessel (2002) suggèrent de diviser la valeur observée de l'entropie quadratique par la valeur théorique maximale obtenue en faisant varier les fréquences des catégories. Un résultat surprenant est alors apparu : l'entropie quadratique serait souvent maximisée en éliminant des catégories. L'indice de Rao est-il invalidé ?

Shimatani (2001) observe que ce phénomène dépend du choix des dissimilarités. Nous en avons démontré les raisons mathématiques (Pavoine & Ollier, soumis). Cette démonstration nous a amené à rechercher, dans la représentation graphique des catégories, la plus petite boule (ou cercle en dimension 2) contenant toutes les catégories et à constater que seules les catégories situées sur la surface (périmètre) de cette boule (cercle) sont retenues pour atteindre la valeur maximale de l'entropie quadratique (cf. exemple figure 3). Les deux cas extrêmes de cette démonstration sont la variance, où seules deux catégories sont retenues, et l'indice de Simpson, où toutes les catégories sont retenues avec des fréquences égales. Pour la dissimilarité taxonomique et certaines dissimilarités phylogénétiques entre

espèces qualifiées d'ultramétriques, toutes les espèces sont retenues avec des fréquences inégales (cf. exemple figure 4).

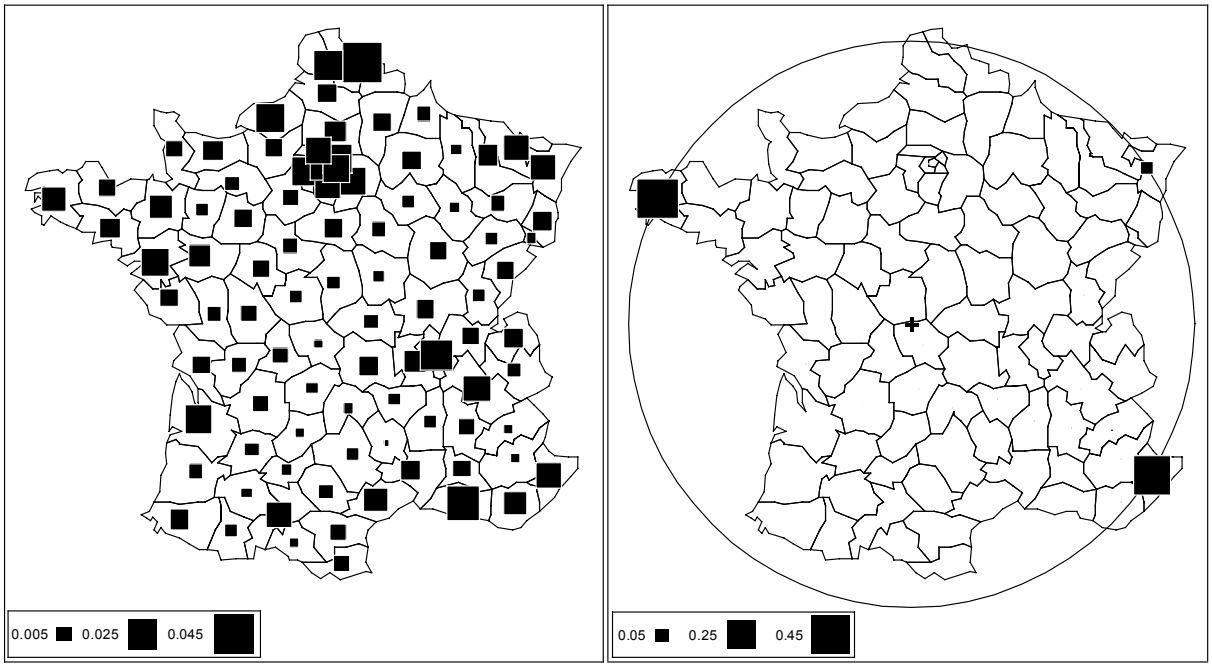


FIGURE 3: Distribution des habitants en France métropolitaine : lors du recensement de 1999 (à gauche) ; pour maximiser la diversité spatiale selon l'entropie quadratique (à droite). Les carrés indiquent l'abondance relative des habitants dans un département. La France (collection) est considérée comme un ensemble d'habitants (entités) regroupés en départements (catégories). Les dissimilarités introduites dans l'entropie quadratique sont les distances à vol d'oiseau entre les départements représentés par leur centre géographique. Le cercle (à droite), dont le centre est indiqué par une croix, englobe strictement les centres de 91 départements et passe par ceux des 3 départements restants : les Alpes-Maritimes, le Finistère et le Bas-Rhin. Pour maximiser la diversité spatiale selon l'entropie quadratique, les habitants doivent être répartis uniquement entre ces trois départements avec les proportions indiquées sur la figure de droite. La distribution des habitants lors du recensement de 1999 ne représentait que 40% de la diversité spatiale maximale selon l'entropie quadratique.

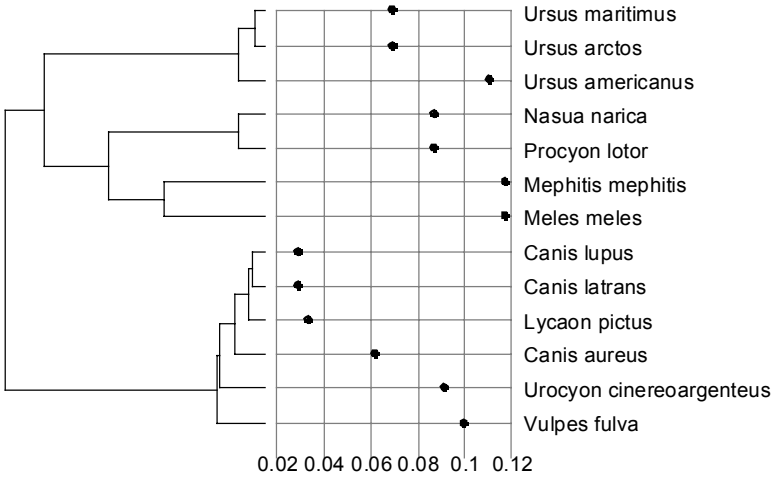


FIGURE 4 : Distribution de fréquences entre 13 espèces de carnivores maximisant leur diversité phylogénétique selon l'entropie quadratique (données de Diniz-Filho et al. 1998) : à gauche la phylogénie, à droite les fréquences. L'échelle des fréquences est horizontale. La dissimilarité entre deux espèces est égale aux longueurs des branches qui les séparent. Cette distribution de fréquences montre l'originalité de chaque espèce.

CONCLUSION ET PERSPECTIVES

Regroupant et généralisant l'indice de Gini-Simpson et la variance, l'entropie quadratique rassemble dans une même théorie les concepts de diversité et dissimilarité. Sa décomposition ainsi que son lien avec une représentation graphique, généralisant des techniques développées en statistiques et dans plusieurs champs des sciences naturelles, permettent une analyse de la structure de la biodiversité et même de la diversité au sens large.

Les recherches pour la mesure de la biodiversité ressemblent à la construction de la tour de Babel. Certains travaillent sur les gènes, d'autres au niveau du paysage. Certains travaillent sur les plantes, d'autres sur les animaux, d'autres encore sur les micro-organismes. Chacun a son langage, ses outils de travail. Pour mesurer et gérer la biodiversité en son ensemble, il nous faut savoir si les outils des uns sont applicables aux autres et si deux outils développés dans deux champs différents ne sont pas en réalité les mêmes écrits dans deux langues différentes. La théorie de Rao permet d'avoir un tel langage commun.

La suite de ce travail nous amènera à comparer l'entropie quadratique à d'autres indices tels que la diversité phylogénétique de Faith (1992) et la dispersion cladistique de Humphries & Williams (1994). Puis nous appliquerons la théorie de Rao en rajoutant la notion de complémentarité : sachant qu'une région contenant telles espèces est déjà

protégée, quelle autre région choisir pour augmenter la biodiversité de l'ensemble ?

REFERENCES

BLONDEL, J. & FARRE, H. (1984). *Oecologia.* **75**, 83-93.

CHAMPELY, S. & CHESSEL, D. (2002). *Environ. Ecol. Stat.* **9**, 167-177.

DINIZ-FILHO, J.A.F., DE SANT'ANA, C.E.R. & BINI, L.M. (1998). *Evolution*, **52**, 1247-1262.

EXCOFFIER, L., SMOUSE, P.E. & QUATTRO, J.M. (1992). *Genetics* **131**, 479-491.

FAITH, D.P. (1992). *Biol. Conserv.* **61**, 1-10.

FISHER, R.A. (1925). Statistical Methods for Research Workers. Edinburgh: Oliver & Boyd.

GINI, C. (1912). Variabilitá e mutabilitá. Studi economicoaguridici delle facoltta di giurizprudenza dell, Universite di Cagliari III, Parte II.

GIRODET, J. (1976). Dictionnaire de la langue française. Paris : Bordas.

GOWER, J.C. (1982). *Math. Scientist* **7**, 1-14.

HENDRICKSON, J.A.J. & EHRLICH, P.R. (1971). *Notulae Naturae* **439**, 1-6.

HUMPHRIES, C.J. & WILLIAMS, P.H. (1994). In: *Models in phylogenetic construction*, Ed. R. W. Scotland, D. M. Siebert and D. Williams, pp. 335-352. Oxford: Clarendon.

LIGHT, R.J. & MARGOLIN, B.H. (1971). *J. Amer. Statist. Assn.* **66**, 534-544.

NEI, M. (1973). *Proc. Natl. Acad. Sci. USA* **70**, 3321-3323.

NEI, M. & LI, W.-H. (1979). *Proc. Natl. Acad. Sci. USA* **76**, 5269-5273.

PATIL, G.P. & TAILLIE, C. (1982). *J. Amer. Statist. Assn.* **77**, 548-561.

PAVOINE, S., DUFOUR, A.B. & CHESSEL, D. (in press). From dissimilarities among species to dissimilarities among communities: a double principal coordinate analysis. *J. Theor. Biol.*

PAVOINE, S. & OLLIER, S. (soumis). Measuring diversity from dissimilarities with Rao's quadratic entropy: are any dissimilarity indices suitable? *Biometrika*.

RAO, C.R. (1982a). *Theor. Popul. Biol.* **21**, 24-43.

RAO, C.R. (1982b). *Sankhya: Ind. J. Stat.* **A44**, 1-22.

RAO, C.R. (1986). In: *Encyclopedia of Statistical Sciences*, Ed. S. Kotz and N. L. Johnson, pp. 614-617. New York: Wiley & Sons.

RAO, C.R. & NAYAK, T.K. (1985). *IEEE Transactions on Information Theory* **IT-31**, 589-593.

SHANNON, C.E. (1948). *Bell System Tech.* **27**, 379-423, 623-656.

SHIMATANI, K. (2001). *Oikos* **93**, 135-147.

SIMPSON, E.H. (1949). *Nature* **163**, 688.

WARWICK, R.M. & CLARKE, K.R. (1995). *Marine Ecology Progress Series* **129**, 301-305.

Annexe 6
Pavoine, S., J. Blondel et D. Chessel 2006 - Ecology - en révision

Pavoine, S., J. Blondel, and D. Chessel. 2006. A new technique for ordering three-dimensional data sets in convergence studies. Ecology. En révision

A NEW TECHNIQUE FOR ORDERING THREE-DIMENSIONAL DATA SETS IN CONVERGENCE STUDIES

by

SANDRINE PAVOINE[1,4], JACQUES BLONDEL[2] AND DANIEL CHESSEL[3]

[1] *Centre de Recherches sur la Biologie des Populations d'Oiseaux, Muséum National d'Histoire Naturelle, 55, Rue Buffon, F-75005 Paris, France.*

[2] *CEFE/CNRS, 1919 Route de Mende, 34293 Montpellier cedex 05, France.*

[3] *Laboratoire de Biométrie et biologie évolutive (UMR 5558), CNRS, Univ. Lyon 1 43 bd du 11 novembre 1918, F-69622 Villeurbanne cedex.*

[4] Corresponding author: Sandrine Pavoine, UMR 5173 MNHN-CNRS-P6 '*Conservation des espèces, restauration et suivi des populations',* Muséum National d'Histoire Naturelle, CRBPO, 55, Rue Buffon, 75005 Paris, France
Telephone: (33-1)-4079-3073, Fax: (33-1)-4079-3835, E-mail: pavoine@biomserv.univ-lyon1.fr.

Running head: ordering 3D data sets in convergence studies

Manuscript type: article

Abstract. Using a method we call Foucart's correspondence analysis, the aim of this paper is to coordinate the correspondence analyses of several independent species × site matrices which share the same species and sites which are ecologically (or evolutionary) similar. Such coordination makes it possible to compare the characteristics of each individual matrix with the structure inferred by the average matrix. This technique proved to be particularly powerful in the analyses of evolutionary convergence of biotas which include several distinct data sets. An example is provided by a study of the convergence of ecological trajectories of bird communities among mediterranean (Provence, Corsica and Algeria) and non-mediterranean (Burgundy and Poland) ecological successions in Europe. The problem which arises from this study corresponds to a cube of data formed by two factors fixed by the experimenter (stage of vegetation along ecological successions and different geographically distinct successions) and a factor which is not controlled (species in these vegetation stages and successions). The entries of this cube are the densities of the species in each vegetation stage and each succession. Two analyses are of particular interest: (i) the analysis of the differences in the species-vegetation stage structure between successions; (ii) the analysis of the changes in the species-succession structure across vegetation stages. Each of the two analyses is performed by Foucart's correspondence analysis. The combined two analyses are much more insightful than a single correspondence analysis applied to the global matrix which is obtained by simply juxtaposing the individual species × stage matrices through successions. This is because the single, global correspondence analysis combines three effects of factors (vegetation stage, succession and vegetation stage × succession interaction) whereas the two applications of Foucart's correspondence analysis clearly discriminate issues: first, the pattern of species distribution along stages fairly well remains across successions; second, the structure of the differences among the avifauna of distinct successions remains unchanged

from stage to stage, but the amount of differences decreases with the complexity of the vegetation.

Key words: birds; community convergence; correspondence analysis; crossed-factor analysis; multivariate analysis; ordination; species abundances

INTRODUCTION

A three dimensional data set, often called data cube, consists of stacked two-dimensional data matrices as a function of a third coordinate. The rows and columns of each two-dimensional data matrix are given by two coordinates so that the data cube is defined by three coordinates called "modes". The ordination of multi-arrays is a crutial question in ecology (Potvin and Travis 1993). The analysis of data cubes covers a widespread field of research in which the three-mode principal component analysis plays a large part (Kroonenberg 1983, Coppi and Bolasco 1989). This method assumes that the three modes represent three comparable factors symmetrically treated. However, ecologists, especially those working on convergence studies, face a very particular asymmetrical situation. Their data include three very different modes: the species, and two crossed factors defining the zones and/or dates of experiment (Swaine and Greig-Smith 1980). The entries of the cube are the abundances or presences/absences of species. The two modes corresponding to the two crossed factors are fixed by the observer. The third mode, on the other hand, is not controlled: it is the list of the species determined by the investigated ecosystems. We address the problems raised by this type of asymmetric data cube using the study by Blondel and Farré (1988) on the ecological convergence of bird communities in forested areas of the western Palaearctic.

Within the framework of the history of bird fauna since the Pleistocene, Blondel et al. (Blondel 1986b, Blondel and Farré 1988, Blondel 1995, Blondel and Mourer-Chauviré 1998) tested the hypothesis of an increasing similarity in the composition of bird communities as vegetation becomes taller and more complex along ecological successions located in various parts of Europe. They studied four successions, two in the mediterranean region and two in the medioeuropean region. Then they choose, in each succession, six stages of vegetation as similar as possible between successions. Finally they took a census of bird communities from

point count methods. The problem arising from the resulting data sets is based on three crossed factors forming a data cube: species, vegetation stage, and succession.

Assuming that the different successions reasonably well match one another in terms of habitat structure, two questions emerged from a regional meta-analysis of several bird successions such as that addressed by Blondel and Farré: (i) to which extent the distribution of species between stages of vegetation depends on the geographical position of the successions? (ii) do the differences in bird faunas between successions depend on stages of vegetation?

To answer the first question, we must compare the effect of vegetation stages on bird distribution in an individual succession with their effects in the other successions. In each succession, the matrix of data includes a species × vegetation stage type. The relationships between bird communities and architecture of vegetation are defined in each succession by a structure which must be revealed from the matrix. This matrix may be justifiably processed by a correspondence analysis (CA) (Benzécri and Coll. 1973, Greenacre 1984) which purpose is to construct an ordination and thus a typology of species according to their stage-specific response to habitat structure and also symmetrically a typology of vegetation stages according to their species compositions. As there are four such matrices, there is an average structure, called trade-off structure, and local region-specific departures from this global mean structure.

Answering the second question requires to compare the differences between the successions in their bird composition for an individual stage of vegetation with these differences for the other vegetation stages. In each stage, the matrix of data is of species × succession type. The relationships between fauna and successions are defined in each stage of vegetation by a structure contained in the data matrix and are therefore treated by CA. This analysis provides a typology of species according to their geographical distribution and also symmetrically a typology of successions according to their species contents. As there are six

matrices of this type, there is an average global structure, the so-called “trade-off structure”, and depatures from this mean according to vegetation factors.

The purpose of this paper is to find, for each of the two questions, the average structure and to describe the departures from this average structure according to each modality of the factor: each succession for the first question and each vegetation stage for the second question. To reach this purpose, the main difficulty is the coordination of several correspondence analyses. Similar situations in social sciences (for example, age-group × region × year with population density as entry) have been investigated by Foucart (1978). The main difficulty in coordinating the correspondence analyses of K matrices is due to the fact that each matrix is characterized by its own weightings of rows and columns. The rationale of Foucart (1978) was to work first on an average matrix (“trade-off” matrix) and then to look at the differences between the resulting global structure of the combined matrices and that given by each of the K constitutive matrices. We call this method "Foucart's CA" which is not very known, most probably because the French journal where it was published ceased its activity several years ago. The same concept of average or "trade-off" matrix constitutes a basis for the partial triadic analysis also called "STATIS on tables" which calculates an average analysis for several principal component analyses applied to the same individuals and the same variables (Escoufier 1973, Lavit et al. 1994).

We will not come back here on the paradigm of convergence which is discussed at length by Blondel and Farré (1988). Instead, we will show how rejuvenating an old concept using old data can contribute to decipher new lines of reasoning and put new wine in old bottles. In order to widen the discussion, we add to the data from Blondel and Farré a succession which had been studied in Algeria by Benyacoub (1993) following exactly the same rationale and the same methods.

MATERIALS AND METHODS

Data from five ecological successions have been selected to carry out this study. Three of them are located in the mediterranean region: Provence (southern France, Blondel 1979), the island of Corsica (France, Blondel 1979) and Algeria (Benyacoub 1993). The two other successions are located in the medioeuropean region, one in Burgundy (central France, Ferry and Frochot 1970) and the other in Poland (Glowacinski 1975). Each succession has been conventionally divided into six serial stages in such a way that all five successions match one another in respect to the number and structure of habitats. Their selection has been made using classical criteria of habitat structure, especially the complexity and height of the vegetation (ranging from low bushy vegetation, height < 1 m, stage 1, to forests with trees at least 20 meters high, stage 6). There remain, however, inescapable differences between gradients. On average, the vegetation is higher and lusher, the structure more complex and the plant species with deciduous morphotypes more numerous in the medioeuropean successions as compared to the mediterranean successions. In addition, bushy habitats are on average more clear cut than forested habitats in the mediterranean gradients than in their medioeuropean counterparts (Blondel and Farré 1988). To be consistent with Blondel and Farré's data, Benyacoub (1993) chose a similar ordination of habitats: low maquis, medium maquis, high maquis, maquis with dense stratum of trees, suber oak forest with undergrowth, zeen oak forest.

The composition (species richness) and abundance (number of breeding pairs/km²) of bird communities have been determined from bird censuses in every succession and stage. A total of 89 distinct species were listed; 39 species in Poland, 39 in Provence, 39 in Corsica, 42 in Algeria and 45 in Burgundy. A similar number of species in the successions is a general feature in the forest habitats of the western Palearctic region (Blondel 1986a, Blondel and Mourer-Chauviré 1998) so that the Algerian succession belongs to the same biogeographical

100 realm as the European ones. The number of species in common in pairs of successions is
101 shown in Table 1.

102 *Global correspondence analysis*

103 The methodologies used are summarized in Fig. 1. The first option proposed by
104 Blondel and Farré (1988) for analyzing these data consisted of applying CA to a matrix which
105 included all data from 30 sites (5 successions of 6 stages) as columns, 89 species as rows and
106 densities as entries. We call this method “global CA”. This analysis results in a simultaneous
107 ordination of species and sites (Benzecri 1973, Hill 1974) and provides factorial maps in
108 which species and sites are represented by points (Greenacre and Hastie 1987).

109 *Foucart's correspondence analysis*

110 Consider n matrices with the same p rows and the same q columns. The rows
111 represent the species. Either the columns stand for the vegetation stages and each matrix
112 corresponds to a succession, or the columns are the successions and each matrix corresponds
113 to a stage. Foucart's CA studies changes in the structure generated by a factor through the
114 levels of a second factor. The first choice, therefore, results in studying changes in the
115 species-stage structure between successions; and the second choice leads to analyzing changes
116 in the species-succession structure between vegetation stages. Both configurations are
117 performed. The entries of these matrices are the densities and are noted $\mathbf{A}_{(k)} = \left[a_{ij(k)} \right]$ with
118 $1 \le i \le p$, $1 \le j \le q$ and $1 \le k \le n$. Each matrix is changed to be expressed as a percentage:
119 $\mathbf{P}_{(k)} = \left[a_{ij(k)} / a_{\bullet\bullet(k)} \right]$, where $a_{\bullet\bullet(k)}$ is the sum of the terms in $\mathbf{A}_{(k)}$. Let $p_{ij(k)}$ be equal to
120 $a_{ij(k)} / a_{\bullet\bullet(k)}$.

121 Foucart's CA coordinates the separate analysis of the n matrices using a "trade-off
122 matrix" which is the uniformly-weighted mean of the n matrices $\mathbf{P}_{(k)}$:

123 $$\bar{\mathbf{P}} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{P}_{(k)} = \left[\frac{1}{n}\sum_{k=1}^{n} p_{ij(k)}\right]_{\substack{1\le i\le p \\ 1\le j\le q}}.$$

124 The application of CA to $\bar{\mathbf{P}}$ provides a trade-off structure of the distribution of the species
125 between the levels of the column-factor. To obtain the modifications of this structure between
126 matrices, the method consists of projecting the rows and columns of the n individual matrices
127 as additional elements. The weights of the trade-off CA are used as reference. The addition of
128 supplementary elements in CA is described in LeRoux et Rouanet (2004).

129 The data set is given in Appendix A. All computations and graphical displays were
130 carried out using R (Ihaka and Gentleman 1996), with both pre-programmed and personal
131 routines available in the ade4 package at http://lib.stat.cmu.edu/R/CRAN/ (Chessel et al.
132 2004) (see Appendix B).

133 ## RESULTS

134 In Fig. 2 the typology of sites is displayed on the first four axes of the global CA with
135 15.4%, 13.7%, 11.8% and 9.1% of the variation pertaining to F1, F2, F3, and F4, respectively.
136 The ordination is similar to that obtained by Blondel and Farré (1988) with the Algerian
137 succession being close to the Corsican succession. F1 and F2 separate the two medioeuropean
138 successions from the three mediterranean successions. The successions of the non-
139 mediterranean region vs the mediterranean region start from points opposed on the map and
140 then converge in the forested habitats. F1 and F3 express a strong originality of the first two
141 vegetation stages in Provence. On F3, the ordination of stages 1 and 2 of Provence even runs
142 counter to gradients of vegetation. F4 separates the successions within each region, opposing
143 Poland to Burgundy and discriminating Algeria, Provence and Corsica.

144 Fig. 3 illustrates the first two axes of Foucart's CA which emphasizes the differences
145 in the species-vegetation stage structure between the successions (F1: 63.5%, F2: 20.1% of
146 the variation). In spite of the experimental difficulties in finding habitat gradients of exactly

147 similar vegetation structure in different geographical situations, the successions reproduce the
148 same structure fairly well. The species point clouds have similar shapes. In the open habitats,
149 the Algerian and Provençal successions depart from the trade-off, i.e. the average pattern.
150 Bushy habitats are closer from each other in the Algerian succession than on average in the
151 other successions and conversely they are more different in the succession of Provence.
152 Fig. 4 displays the factorial map F1 × F2 of Foucart's CA studying the changes in the
153 species-succession structure along the vegetation stages (F1: 37%, F2: 22.4% of the
154 variation). F1 discriminates the medioeuropean region from the mediterranean region while
155 F2 expresses differences between successions within each of the two regions. On the whole,
156 the typology of the differences between successions remains unchanged in the six stages but
157 the magnitude of differences gradually decreases as the complexity of the vegetation
158 increases. In the forested habitat, the four successions become very similar. A similar
159 convergence is observed from the axes F3 and F4 (Fig. 5).

160 ## DISCUSSION

161 The global correspondence analysis provides both inter-succession information
162 (regional variation of avifauna) and inter-vegetation stage information (stratification of
163 avifauna on gradients of vegetation height), as well as information on simultaneous variation
164 or interaction of the vegetation stage and succession factors. It is impossible to extract each
165 kind of information when interpreting the factorial maps of global CA because all are mixed
166 in a way that cannot be determined. On the bivariate space F3 × F4 (Fig. 2b), the mixture of
167 the two crossed factors, succession and vegetation stage, creates a spurious ordination from
168 Poland to Corsica via Algeria, Provence and Burgundy in this particular order. At this point,
169 the results obtained with five successions do not depart from those obtained by Blondel and
170 Farré (1988) with four successions only: "the best visualization of the biogeographical and
171 ecological convergence of bird communities in the forested stages of the successions is given

by the display of the data on the bivariate space F1 × F4". However, this choice of axes F1 and F4 is not objective. It has been done because these two axes provides an ordination of the bird communities which can be explained with ecological cases. In addition, it provides some odd results. For example, the mediterranean successions on this bivariate space are ordered from Algeria to Corsica via Provence, whereas the Foucart's analysis (Fig. 4) clearly shows that, despite the existence differences between Provence and Corsica (Blondel et al. 1988), these differences are certainly less significant than indicated on this graph, resulting in an opposition between Provence and Corsica on the one hand and Algeria on the other hand.

The display of data on Fig. 2a suggests that the open habitats (stage 1 and 2) are more similar between Burgundy and Poland than between Provence and Corsica. This could be explained by a better discrimination of bushy *vs* pre-forested and forested habitats in the medioeuropean region than in the mediterranean region (Blondel and Farré 1988). The particular position of stages 1 and 2 on the Provence habitat gradient can in fact be explained by a more marked distinction between garrigue (scrubland in southern France) and forest (Blondel 1986a). This conclusion mixes up two effects: the effects of the geographical localization of successions and the effect of stages of vegetation on the species composition. Because both kinds of effects are mixed, it is quite impossible to explain the results from the global CA in terms of either difference between successions or difference between stages. In contrast Foucart's analysis separates the effects of successions from the effects of vegetation stages, and reciprocally, reveals (Figs. 4 and 5) that, considering the species compositions for separate stages, there are at least as many differences in open habitats between Burgundy and Poland than between Corsica and Provence.

In fact, open habitats (vegetation stages 1 and 2) in Provence and Corsica share more species in common (9 species) than Burgundy and Poland (6 species). Poland has the most distinct succession, especially in stages 1 and 2, presumably as a consequence of a

197 geographical effect. Indeed only three scarce species have been observed in habitat 1 of
198 Poland. This habitat is very different in term of species composition from its counterpart of
199 Burgundy which includes 25 species, several of which being abundant (up to 99 breeding
200 pairs/km^2). In addition, this habitat does not share any species with the gradients of the
201 mediterranean region. The Polish vegetation gradient keeps in the last forested stage some
202 habitat patches of early stages. Comparing to the other gradient, there is in Poland a shifting
203 of the species from earlier to older stages (Blondel and Farré 1988).

204 On another hand, the particular positions of stages 1 and 2 in the Provence gradient do
205 not result from genuine differences between successions but rather from spurious
206 discrepancies between stages along the succession of Provence as shown on Fig. 3. In open-
207 habitats, the succession of Provence has thus the most distinct species-stage structure,
208 whereas the succession of Poland has the most special species composition. Consequently, the
209 conclusion, made by Blondel and Farré from the results of the global CA, of a stronger
210 similarity in terms of species composition within open habitats between Burgundy and Poland
211 than between Corsica and Provence, is thus true only if we consider similarity in the
212 distribution of species between stages within successions. It is wrong if similarity in species
213 composition between successions is of concern. As a result, Fig. 2a, which combines factors
214 (succession, vegetation stage and succession × vegetation stage interaction), cannot be
215 properly interpreted without strong risks of misinterpretation.

216 The first computed Foucart's CA explains changes in the species-stage structure
217 between the successions. Differences in this structure between successions have effectively
218 been noted. In particular, the strong similarity among the first four stages in Algeria is due to
219 the high abundance of *Sylvia melanocephala* in these four stages. Other inconsistencies of the
220 species-stage structure between successions mainly result from the difficulties in finding sites
221 that fairly well match each other in very different series of vegetation. In spite of these

222 difficulties, the average species × vegetation stage structure remains steady through
223 successions. This Foucart's CA by definition does not take into consideration differences in
224 avifauna between the five successions. It extracts the analysis of the structure of the
225 distribution of species along vegetation gradients independently of existing similarities and
226 dissimilarities in the identity of these species between distinct successions. As a result, we are
227 now able to conclude that the species are distributed according to the same patterns along the
228 gradients, whatever the identity of the species.
229 The second computed Foucart's CA describes the evolution of differences between
230 successions along the stages. A spectacular result provided by this analysis is that the
231 structure of differences between successions is maintained from stage to stage while the
232 amount of these differences decreases along habitat gradients. The ecological trajectories of
233 the bird communities along the successions result in a homogenization of bird communities in
234 the old forests, independently of the area considered. In particular the three forested stages,
235 with a low, albeit non null interregional disparity, are almost equivalent. On the other hand,
236 site-specific typology is reinforced in the first stages of the gradients so that one can say that
237 communities converge in complex mature habitats, or that they diverge in open habitats of
238 early successional stages, which better fits the discussion of Blondel and Farré (1988). A
239 detailed discussion on the biogeographic and historical grounds explaining these patterns can
240 be found in Blondel et al. (1988) and Blondel and Farré (1988) who explained why and how
241 the discrimination between communities is much more pronounced in the early stages of the
242 successions than in the final forested stages, including those of the Algerian succession.
243 Indeed, this succession shares the same history as those of other Mediterranean regions, and
244 in addition, shows some components of insularity (Blondel et al. 1988, Benyacoub 1993)
245 because the Maghreb (Morocco, Algeria and Tunisia) is isolated by sea and desert which act
246 as barriers to colonization (Blondel 1986a).

Foucart's CA proves that the global CA basically captures between-succession information, but in the same time this information clearly depends primarily on the stage of vegetation. This explains why Blondel and Farré (1988) used a F1-F4 display of the data but the resulting map included only a part of the vegetation stage × succession interaction. Therefore the use of such a global analysis is not recommended. In order to separate the effects of the two crossed factors (succession and vegetation stage), Foucart's CA must be used.

Foucart's CA makes it possible to consider a three-dimensional set of data as a combination of several two-dimensional matrices. When one mode of the set is species and the other two succession and vegetation stage, two models are possible: (1) one two-dimensional species × stage matrix for each succession, (2) one two-dimensional species × succession matrix for each vegetation stage. Processing Foucart's CA on various combinations of matrices by definition depends on the corresponding chosen model, whereas both models are mixed in the global correspondence analysis. Consequently the structures revealed by Foucart's method are undoubtedly much clearer.

In conclusion, the Foucart's method is mathematically simple and proves to be particularly effective in observing the separated effects of two crossed factors on the species composition of a flora or fauna. It will prove to be very useful and attractive for spatiotemporal analyses aiming at deciphering the dynamics of spatial structures over time as well as changes in temporal fluctuation of species composition in space.

LITTERATURE CITED

Benyacoub, S. 1993. Ecologie de l'avifaune forestière nicheuse de la région d'El Kala (Nord-Est algérien). Doctorat thesis. University of Burgundy.

Benzecri, J. P. 1973. L'analyse des données. II L'analyse des correspondances. Bordas, Paris.

Benzécri, J. P., and Coll. 1973. L'analyse des données. II L'analyse des correspondances. Bordas, Paris.

Blondel, J. 1979. Biogéographie et écologie. Masson, Paris.

Blondel, J. 1986a. Biogéographie évolutive. Masson, Paris.

Blondel, J. 1986b. Biogéographie évolutive à différentes échelles : l'histoire des avifaunes méditerranéennes. Pages 155-188 *in* H. Ouellet, editor. Acta XIX Congressus Internationalis Ornithologici. National Museum of Natural Science, University of Ottawa Press, Ottawa.

Blondel, J. 1995. Biogéographie, Approche Ecologique et Evolutive. Masson, Paris.

Blondel, J., D. Chessel, and B. Frochot. 1988. Bird species impoverishment, niche expension, and density inflation in mediterranean island habitats. Ecology **69**:1899-1917.

Blondel, J., and H. Farré. 1988. The convergent trajectories of bird communities along ecological successions in European forests. Œcologia (Berlin) **75**:83-93.

Blondel, J., and C. Mourer-Chauviré. 1998. Evolution and history of the western Palaearctic avifauna. Trends in Ecology and Evolution **13**:488-492.

Chessel, D., A.-B. Dufour, and J. Thioulouse. 2004. The ade4 package -I- One-table methods. R News **4**:5-10.

Coppi, R., and S. E. Bolasco. 1989. Multiway Data Analysis. Elsevier Science Publishers B.V., North-Holland.

Escoufier, Y. 1973. Le traitement des variables vectorielles. Biometrics **29**:750-760.

Ferry, C., and B. Frochot. 1970. L'avifaune nidificatrice d'une forêt de Chênes pédonculés en Bourgogne: étude de deux successions écologiques. La Terre et la Vie (Revue d'Ecologie) **24**:153-250.

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

297 Glowacinski, Z. 1975. Succession of bird communities in the Niepolomice Forest (Southern
298 Poland). Ekologia Polska **23**:231-263.
299 Greenacre, M. J. 1984. Theory and applications of correspondence analysis. Academic Press,
300 London.
301 Greenacre, M. J., and T. Hastie. 1987. The geometric interpretation of correspondence
302 analysis. Journal of the American Statistical Association **82**:437-447.
303 Hill, M. O. 1974. Correspondence analysis : A neglected multivariate method. Applied
304 Statistics **23**:340-354.
305 Ihaka, R., and R. Gentleman. 1996. R: a language for data analysis and graphics. Journal of
306 Computational and Graphical Statistics **5**:299-314.
307 Kroonenberg, P. M. 1983. Three-mode principal component analysis. DSWO Press, Leiden.
308 Lavit, C., Y. Escoufier, R. Sabatier, and P. Traissac. 1994. The ACT (Statis method).
309 Computational Statistics and Data Analysis **18**:97-119.
310 LeRoux, B., and H. Rouanet. 2004. Geometric data analysis. Kluwer Academic Publishers,
311 Dordrecht, The Netherlands.
312 Potvin, C., and J. Travis. 1993. Concluding remarks: a drop in the ocean. Ecology **74**:1674-
313 1676.
314 Swaine, M. D., and P. Greig-Smith. 1980. An application of principal components analysis to
315 vegetation changes in permanent plot. Journal of Ecology **68**:33-41.
316

317 **APPENDIX A**

318 The data set is available in ESA's Electronic Data Archive.

319

320 **APPENDIX B**

Instructions for resuming the calculations of this paper with R are available in ESA's Electronic Data Archive.

TABLE

Table 1. Number of species shared by paired successions at the scale of the whole habitat gradient

	Poland	Burgundy	Corsica	Provence
Burgundy	23			
Corsica	24	20		
Provence	18	33	15	
Algeria	25	21	23	14

FIGURE LEGENDS:

FIG. 1. Schema summing up the three executed analyses of the species × stage × succession data cube.

FIG. 2. Display of sites on the bivariate spaces (a) F1 × F2 (b) F3 × F4 of global correspondence analysis. The eigenvalue barplot is given. The first stage of each succession is indicated. Solid lines link the stages of a succession from bushy to pre-forested habitats. "Pro1" and "Pro2" stand for the first and the second vegetation stages of Provence. Their particular position on F3 is highlighted. In each figure, a grid indicates the scale; the length of a square side is indicated by the “d” value.

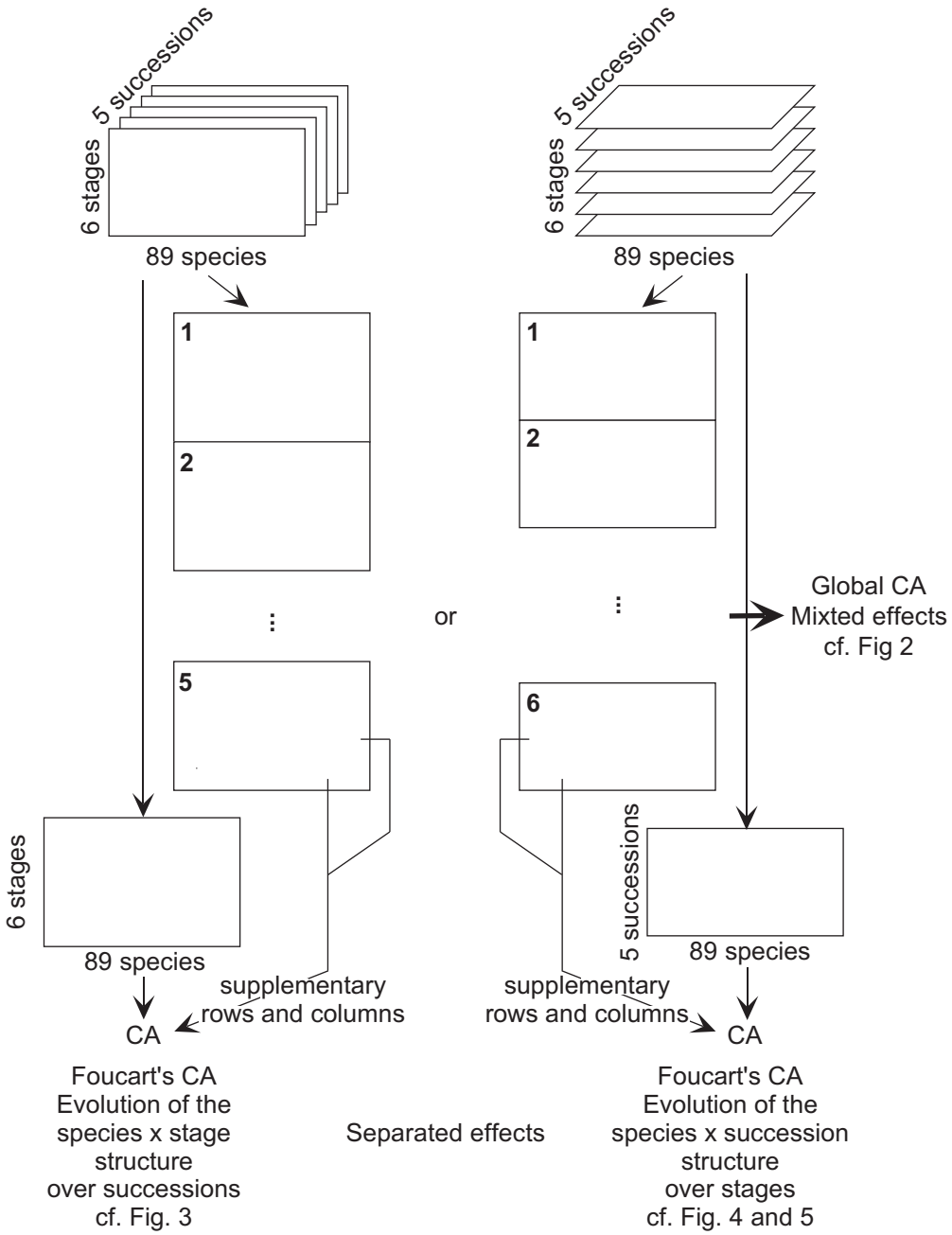
FIG. 3. Display of vegetation stages and species per succession on the bivariate space F1 × F2 of Foucart's correspondence analysis applied to five matrices (corresponding to the

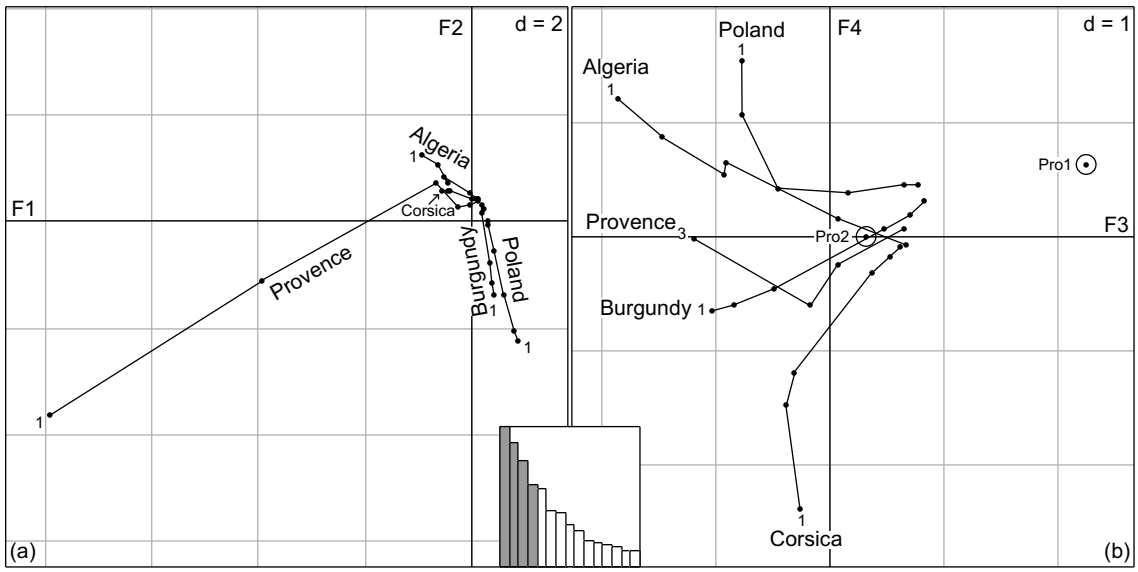
five successions) with stages as columns and species as rows. Solid lines link the stages of a succession from bushy to pre-forested habitats. Species far from the origin of the displays are indicated. The point located at the origin stands for species which are not present in the succession of interest. The last display is for the compromise. Arrows highlight the shape of species points, which is strongly conserved across successions. The eigenvalue barplot is given in the middle of this figure. In each figure, a grid indicates the scale; the length of a square side is indicated by the “d” value.

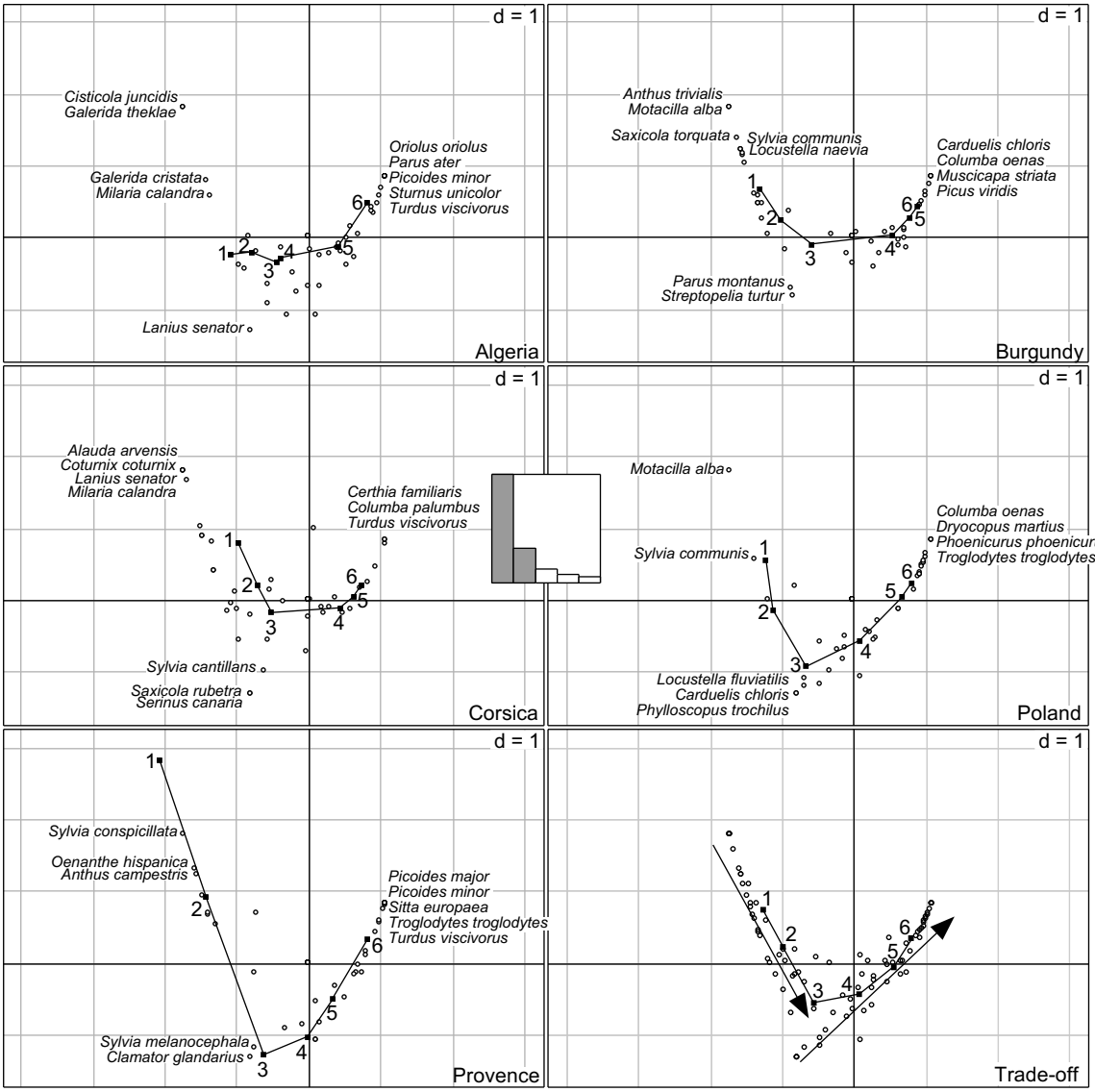
FIG. 4. Display of successions and species per vegetation stage on the bivariate space F1 × F2 of Foucart's correspondence analysis applied to six matrices (corresponding to the six vegetation stages) with successions as columns and species as rows. Labels for successions are 'Alg' (Algeria), 'Bur' (Burgundy), 'Cor' (Corsica), 'Pol' (Poland) and 'Pro' (Provence). Species far from the origin of the displays are indicated. The point located at the origin stands for species which are not present in the stage of interest. The eigenvalue barplot is given in the middle of this figure. In each figure, a grid indicates the scale; the length of a square side is indicated by the “d” value.

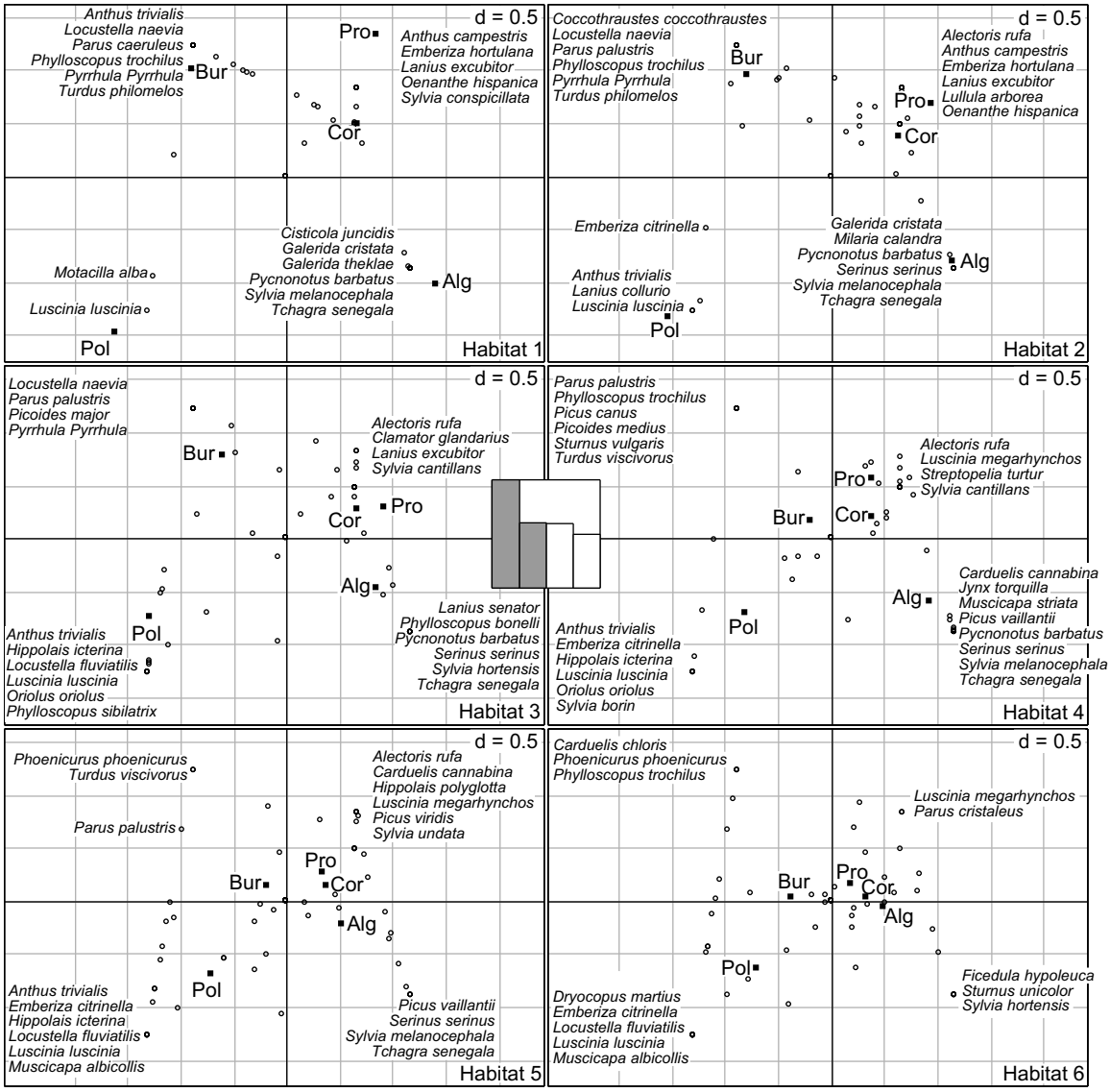
FIG. 5. Convex hull of the five successions positioned per stage on the bivariate spaces (a) F1 × F2 and (b) F3 × F4 of Foucart's correspondence analysis applied to the six matrices (corresponding to the six vegetation stages) with successions as columns and species as rows. Numbers indicate the stage (from 1 bushy to 6 forested). This figure shows how much Foucart's analysis brings an appropriate answer to the ecological question of convergence. It highlights the amplitude of the avifauna-succession structure as a function of the habitat. We can now see how much the six individual matrices express the same structure in different ways. These differences are mainly due to a variability of the degree of differences among

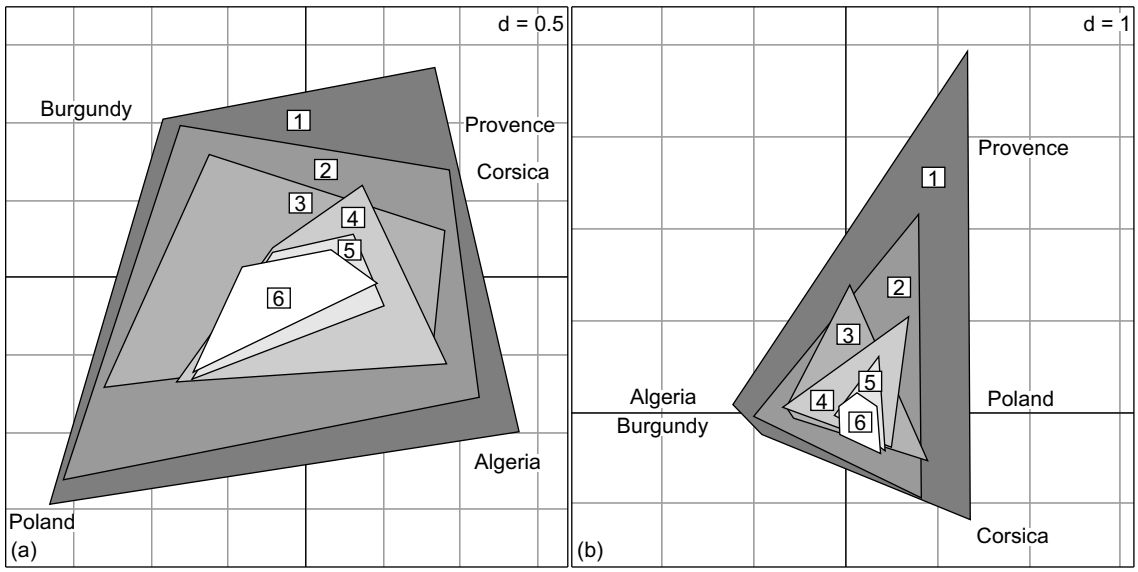
successions as a function of the habitat height. In each figure, a grid indicates the scale; the length of a square side is indicated by the “d” value.


5 successions
6 stages
89 species
1
2
5
or
6
Global CA
Mixted effects
cf. Fig 2
6 stages
89 species
supplementary
rows and columns
CA
Foucart's CA
Evolution of the
species x stage
structure
over successions
cf. Fig. 3
Separated effects
5 successions
89 species
supplementary
rows and columns
CA
Foucart's CA
Evolution of the
species x succession
structure
over stages
cf. Fig. 4 and 5


F2
d = 2
F1
Algeria
1
Corsica
Provence
Burgundy
Poland
1
1
1
(a)
Poland
1
F4
d = 1
Algeria
1
Pro1
Provence 3
Pro2
F3
Burgundy 1
1
Corsica
(b)


d = 1
Cisticola juncidis
Galerida theklae
Oriolus oriolus
Parus ater
Picoides minor
Sturnus unicolor
Turdus viscivorus
Galerida cristata
Milaria calandra
Lanius senator
Algeria
Anthus trivialis
Motacilla alba
Saxicola torquata
Sylvia communis
Locustella naevia
Carduelis chloris
Columba oenas
Muscicapa striata
Picus viridis
Parus montanus
Streptopelia turtur
Burgundy
Alauda arvensis
Coturnix coturnix
Lanius senator
Milaria calandra
Certhia familiaris
Columba palumbus
Turdus viscivorus
Sylvia cantillans
Saxicola rubetra
Serinus canaria
Corsica
Motacilla alba
Sylvia communis
Columba oenas
Dryocopus martius
Phoenicurus phoenicurus
Troglodytes troglodytes
Locustella fluviatilis
Carduelis chloris
Phylloscopus trochilus
Poland
Sylvia conspicillata
Oenanthe hispanica
Anthus campestris
Picoides major
Picoides minor
Sitta europaea
Troglodytes troglodytes
Turdus viscivorus
Sylvia melanocephala
Clamator glandarius
Provence
Trade-off


d = 0.5
Anthus trivialis
Locustella naevia
Parus caeruleus
Phylloscopus trochilus
Pyrrhula Pyrrhula
Turdus philomelos
Bur
Pro
Anthus campestris
Emberiza hortulana
Lanius excubitor
Oenanthe hispanica
Sylvia conspicillata
Cor
Cisticola juncidis
Galerida cristata
Galerida theklae
Pycnonotus barbatus
Sylvia melanocephala
Tchagra senegala
Alg
Motacilla alba
Luscinia luscinia
Pol
Habitat 1
Coccothraustes coccothraustes
Locustella naevia
Parus palustris
Phylloscopus trochilus
Pyrrhula Pyrrhula
Turdus philomelos
Alectoris rufa
Anthus campestris
Emberiza hortulana
Lanius excubitor
Lullula arborea
Oenanthe hispanica
Emberiza citrinella
Galerida cristata
Milaria calandra
Pycnonotus barbatus
Serinus serinus
Sylvia melanocephala
Tchagra senegala
Anthus trivialis
Lanius collurio
Luscinia luscinia
Habitat 2
Locustella naevia
Parus palustris
Picoides major
Pyrrhula Pyrrhula
Alectoris rufa
Clamator glandarius
Lanius excubitor
Sylvia cantillans
Lanius senator
Phylloscopus bonelli
Pycnonotus barbatus
Serinus serinus
Sylvia hortensis
Tchagra senegala
Anthus trivialis
Hippolais icterina
Locustella fluviatilis
Luscinia luscinia
Oriolus oriolus
Phylloscopus sibilatrix
Habitat 3
Parus palustris
Phylloscopus trochilus
Picus canus
Picoides medius
Sturnus vulgaris
Turdus viscivorus
Alectoris rufa
Luscinia megarhynchos
Streptopelia turtur
Sylvia cantillans
Carduelis cannabina
Jynx torquilla
Muscicapa striata
Picus vaillantii
Pycnonotus barbatus
Serinus serinus
Sylvia melanocephala
Tchagra senegala
Anthus trivialis
Emberiza citrinella
Hippolais icterina
Luscinia luscinia
Oriolus oriolus
Sylvia borin
Habitat 4
Phoenicurus phoenicurus
Turdus viscivorus
Parus palustris
Alectoris rufa
Carduelis cannabina
Hippolais polyglotta
Luscinia megarhynchos
Picus viridis
Sylvia undata
Picus vaillantii
Serinus serinus
Sylvia melanocephala
Tchagra senegala
Anthus trivialis
Emberiza citrinella
Hippolais icterina
Locustella fluviatilis
Luscinia luscinia
Muscicapa albicollis
Habitat 5
Carduelis chloris
Phoenicurus phoenicurus
Phylloscopus trochilus
Luscinia megarhynchos
Parus cristaleus
Ficedula hypoleuca
Sturnus unicolor
Sylvia hortensis
Dryocopus martius
Emberiza citrinella
Locustella fluviatilis
Luscinia luscinia
Muscicapa albicollis
Habitat 6


d = 0.5
Burgundy
Provence
Corsica
1
2
3
4
5
6
Algeria
Poland
(a)
d = 1
Provence
1
2
3
5
4
6
Algeria
Burgundy
Poland
Corsica
(b)

APPENDIX B - Instructions for resuming the calculations of the paper.

In your working directory, copy the file 'AppendixA.txt' (Appendix A). Then, write the following instructions on your R console:

```
library(ade4)
ecoconv <- read.table("AppendixA.txt", h = TRUE, row.names = 1)
names(ecoconv)
[1] "Alg1" "Alg2" "Alg3" "Alg4" "Alg5" "Alg6" "Bur1" "Bur2" "Bur3" "Bur4"
[11] "Bur5" "Bur6" "Cor1" "Cor2" "Cor3" "Cor4" "Cor5" "Cor6" "Pol1" "Pol2"
[21] "Pol3" "Pol4" "Pol5" "Pol6" "Pro1" "Pro2" "Pro3" "Pro4" "Pro5" "Pro6"
```

The object `'ecoconv'` is the global species × site matrix. The notations are: 'Alg' Algeria, 'Bur' Burgundy, 'Cor' Corsica, 'Pol' Poland, 'Pro' Provence, and from 1-low bush to 6-preforest for the vegetation stages.

From the global matrix, obtain two lists of individual matrices:

The following object `'list1'` contains the list of the five species × stage matrices.

```
list1 <- list(Alg = ecoconv[, 1:6], Bur = ecoconv[, 7:12], Cor =
    ecoconv[, 13:18], Pol = ecoconv[, 19:24], Pro = ecoconv[, 25:30])
names(list1$Alg) <- c("S1", "S2", "S3", "S4", "S5", "S6")
names(list1$Bur) <- c("S1", "S2", "S3", "S4", "S5", "S6")
names(list1$Cor) <- c("S1", "S2", "S3", "S4", "S5", "S6")
names(list1$Pol) <- c("S1", "S2", "S3", "S4", "S5", "S6")
names(list1$Pro) <- c("S1", "S2", "S3", "S4", "S5", "S6")
```

The following object `'list2'` contains the list of the six species × succession matrices.

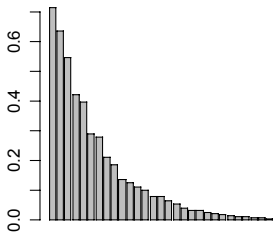
```
list2 <- list(S1 = ecoconv[, c(1, 7, 13, 19, 25)],
    S2 = ecoconv[, c(2, 8, 14, 20, 26)],
    S3 = ecoconv[, c(3, 9, 15, 21, 27)],
    S4 = ecoconv[, c(4, 10, 16, 22, 28)],
    S5 = ecoconv[, c(5, 11, 17, 23, 29)],
    S6 = ecoconv[, c(6, 12, 18, 24, 30)])
names(list2$S1) <- c("Alg", "Bur", "Cor", "Pol", "Pro")
names(list2$S2) <- c("Alg", "Bur", "Cor", "Pol", "Pro")
names(list2$S3) <- c("Alg", "Bur", "Cor", "Pol", "Pro")
names(list2$S4) <- c("Alg", "Bur", "Cor", "Pol", "Pro")
names(list2$S5) <- c("Alg", "Bur", "Cor", "Pol", "Pro")
names(list2$S6) <- c("Alg", "Bur", "Cor", "Pol", "Pro")
```

Obtaining Fig. 2

```
ca1 <- dudi.coa(ecoconv, scan = F, nf = 4)
```

Eigenvalue barplot:

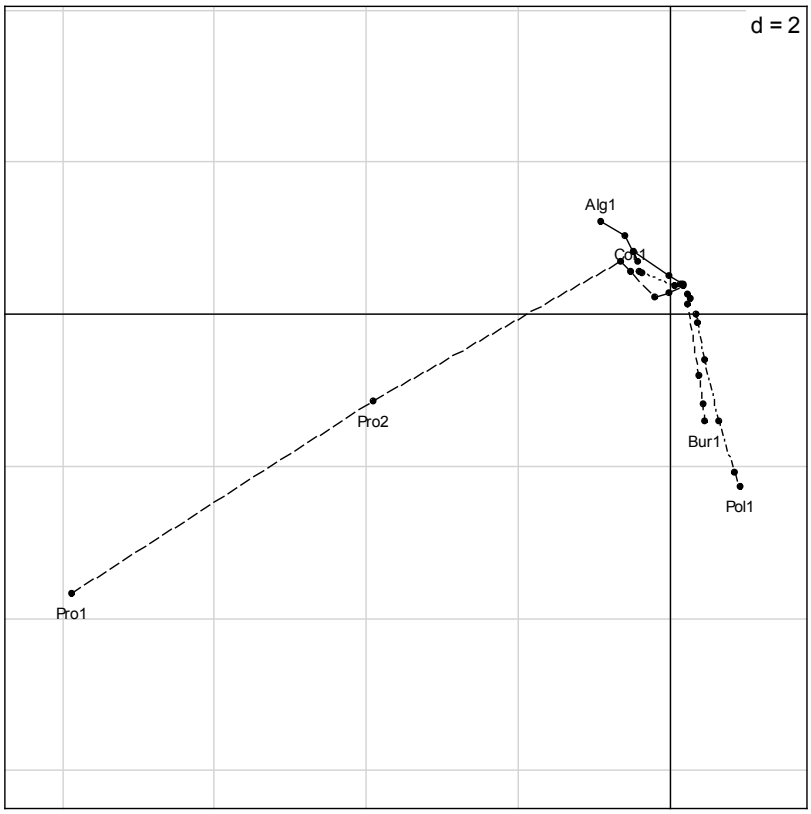
```
barplot(ca1$eig)
```




```
par.sauv<-par()$mar # mar gives the lines of margin for the plot
par(mar=c(0.1,0.1,0.1,0.1))
```

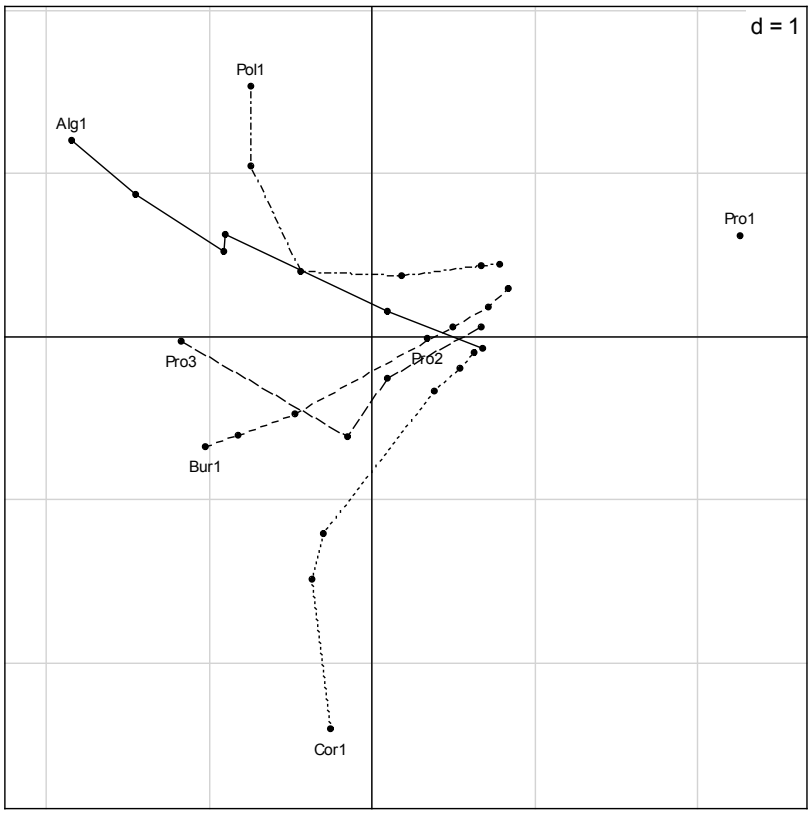
Factorial map F1 × F2:

```
s.label(ca1$co, clab=0)
lines(ca1$co[1:6, ])
text(ca1$co[1, ], "Alg1", pos = 3, cex=0.7)
lines(ca1$co[7:12, ], lty = 2)
text(ca1$co[7, ], "Bur1", pos = 1, cex=0.7)
lines(ca1$co[13:18, ], lty = 3)
text(ca1$co[13, ], "Cor1", pos = 3, cex=0.7)
lines(ca1$co[19:24, ], lty = 4)
text(ca1$co[19, ], "Pol1", pos = 1, cex=0.7)
lines(ca1$co[25:30, ], lty = 5)
text(ca1$co[c(25, 26), ], c("Pro1", "Pro2"), pos = 1, cex=0.7)
```



Factorial map F3 × F4:

```
s.label(ca1$co, clab = 0, xax = 3, yax = 4)
lines(ca1$co[1:6, 3:4])
text(ca1$co[1, 3:4], "Alg1", pos = 3, cex=0.7)
lines(ca1$co[7:12, 3:4], lty = 2)
text(ca1$co[7, 3:4], "Bur1", pos = 1, cex=0.7)
lines(ca1$co[13:18, 3:4], lty = 3)
text(ca1$co[13, 3:4], "Cor1", pos = 1, cex=0.7)
lines(ca1$co[19:24, 3:4], lty = 4)
text(ca1$co[19, 3:4], "Pol1", pos = 3, cex=0.7)
lines(ca1$co[27:30, 3:4], lty = 5)
text(ca1$co[c(25, 26, 27), 3:4], c("Pro1", "Pro2", "Pro3"), pos = c(3, 1,
    1), cex=0.7)

par(mar = par.sauv)
```

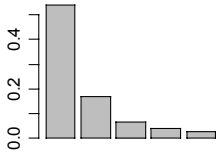


Obtaining Fig. 3:

```
fouc1 <- foucart(list1, scan = F, nf = 2)
```

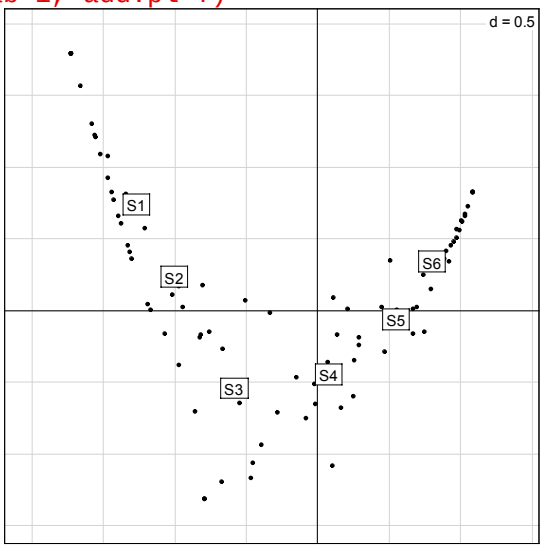
Eigenvalue barplot:

```
barplot(fouc1$eig)
```



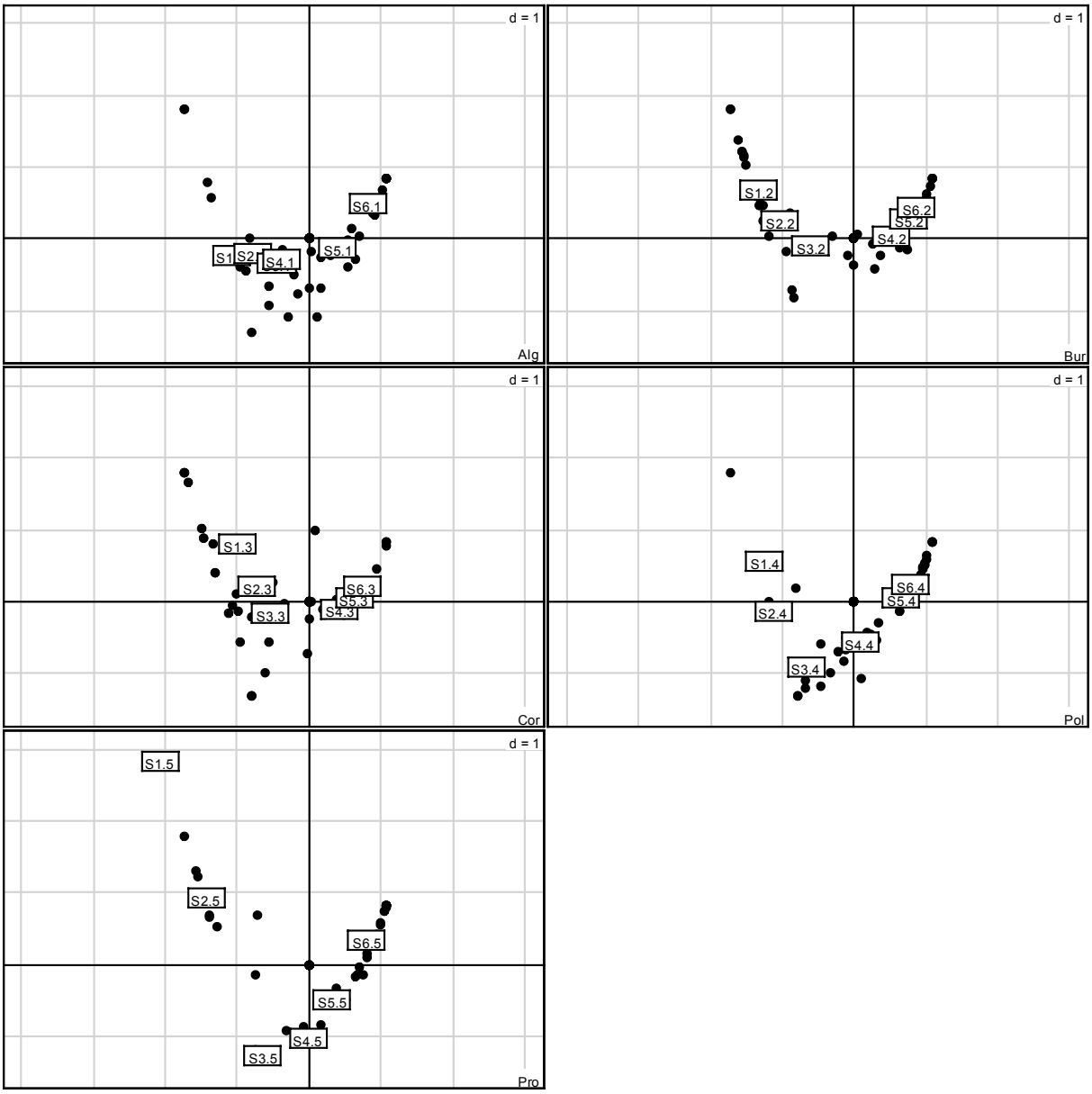
Trade-off:

```
s.label(fouc1$li, clab=0)
s.label(fouc1$co, clab=1, add.pl=T)
```



Projection of the individual matrices:

```
kplot(fouc1, clab.c = 1, clab.r = 0, csub = 1)
```

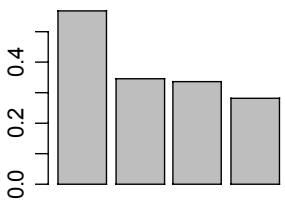


Obtaining Fig. 4:

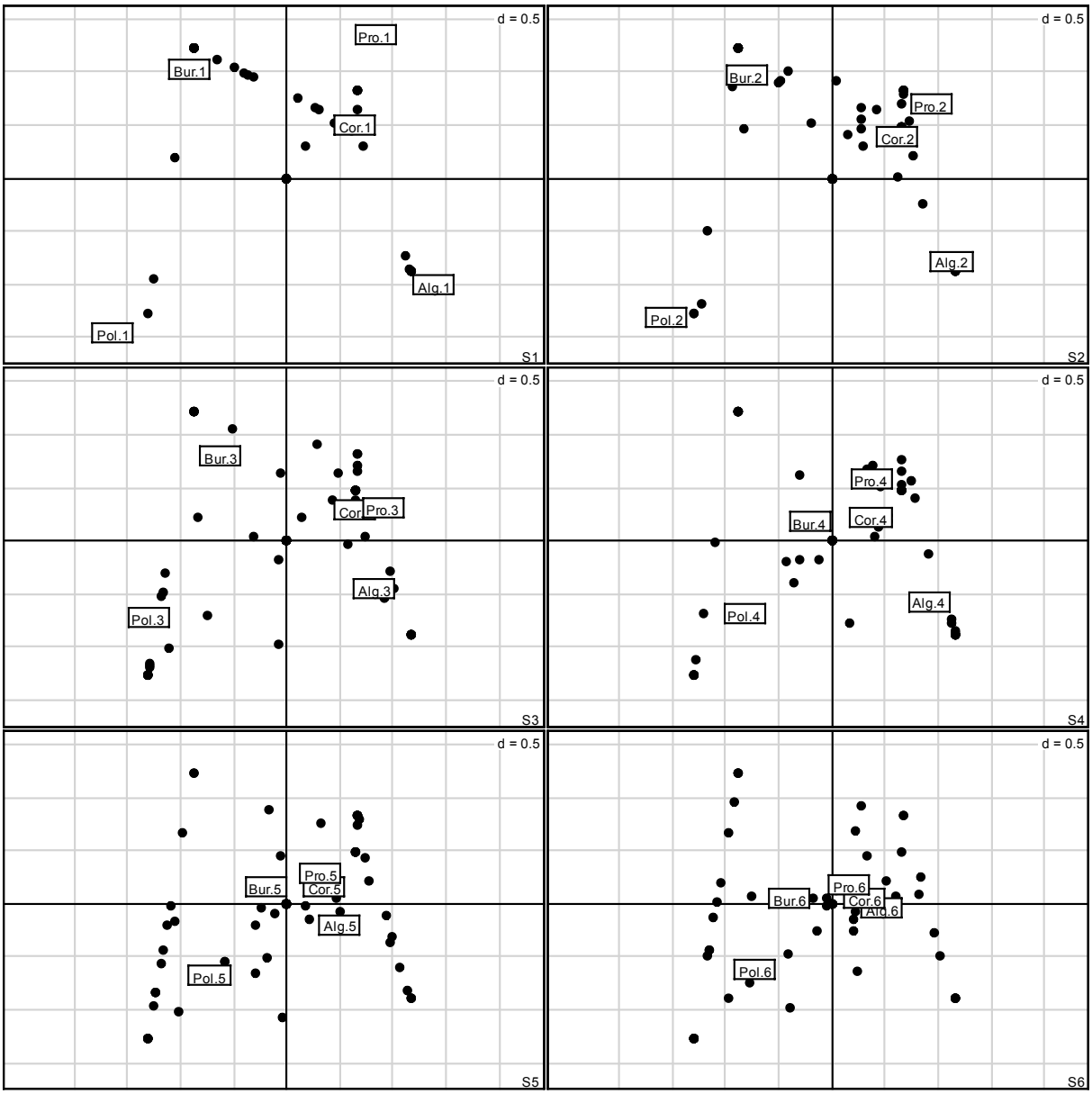
```
fouc2 <- foucart(list2, scan = F, nf = 4)
```

Eigenvalue barplot:

```
barplot(fouc2$eig)
```



```
kplot(fouc2, clab.c = 1, clab.r = 0, csub = 1)
```

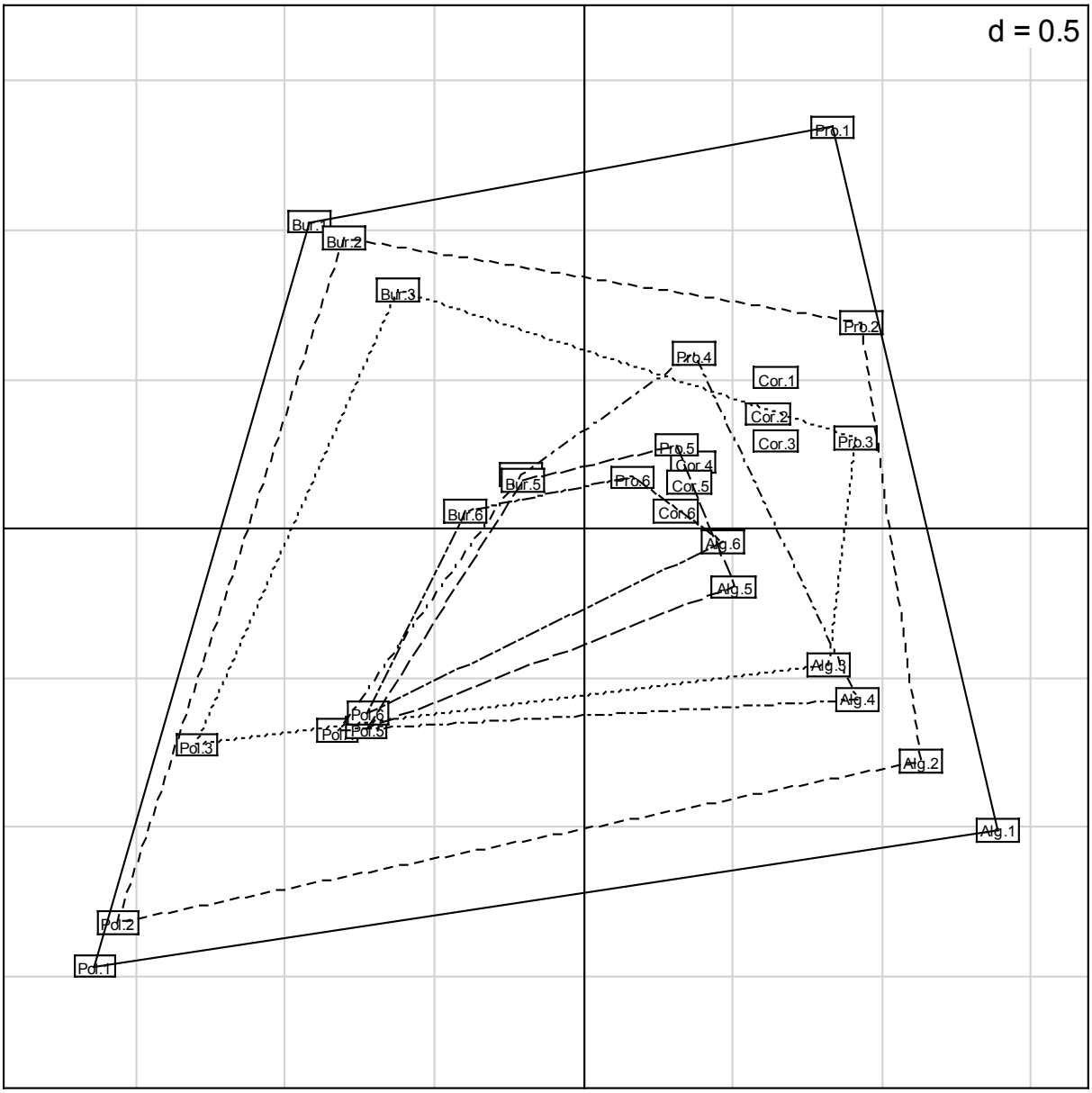


Obtaining Fig. 5:

```
rownames(fouc2$Tco)
 [1] "Alg.1" "Bur.1" "Cor.1" "Pol.1" "Pro.1" "Alg.2" "Bur.2" "Cor.2"
 [9] "Pol.2" "Pro.2" "Alg.3" "Bur.3" "Cor.3" "Pol.3" "Pro.3" "Alg.4"
[17] "Bur.4" "Cor.4" "Pol.4" "Pro.4" "Alg.5" "Bur.5" "Cor.5" "Pol.5"
[25] "Pro.5" "Alg.6" "Bur.6" "Cor.6" "Pol.6" "Pro.6"
```

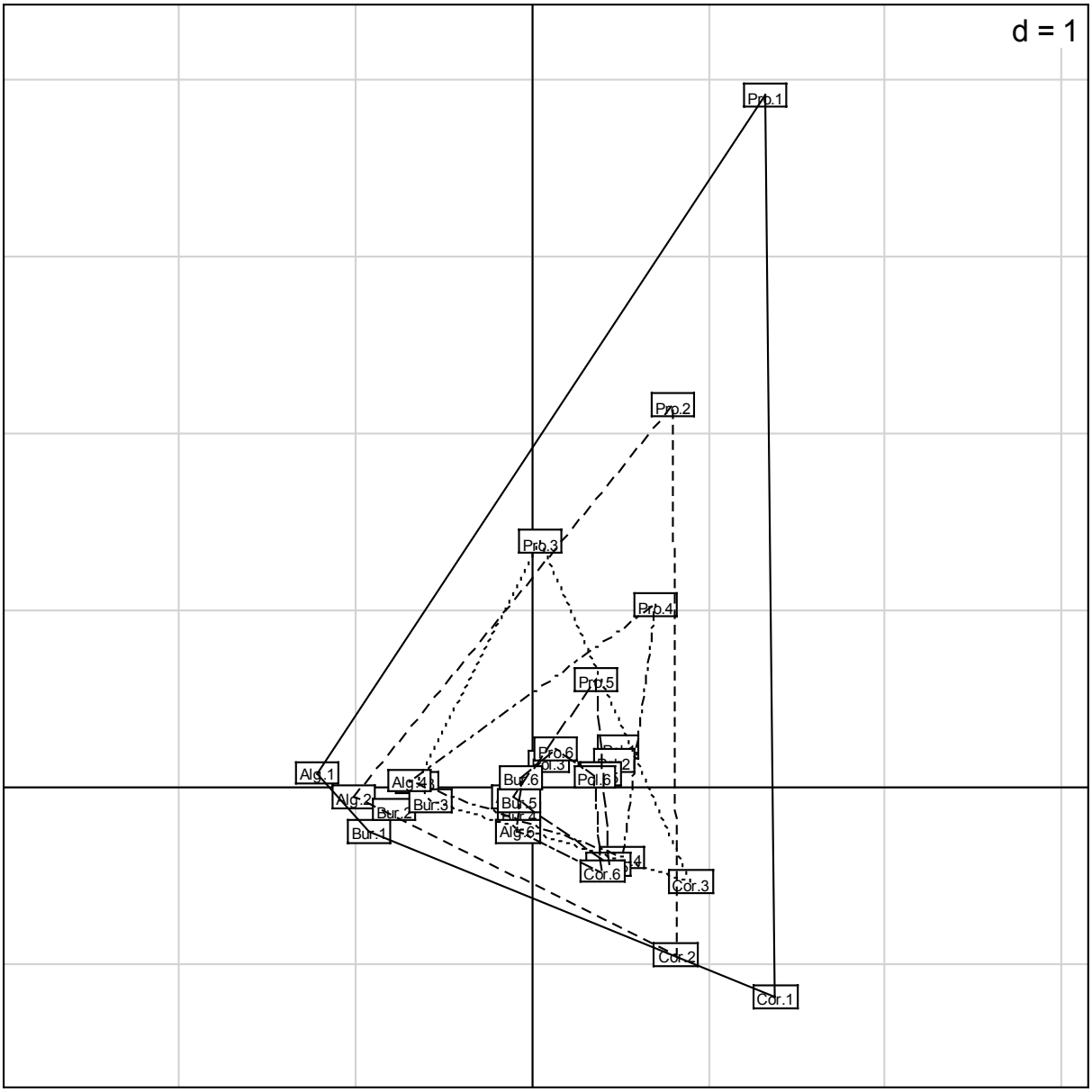
Factorial map F1 × F2:

```
par.sauv<-par()$mar
par(mar=c(0.1,0.1,0.1,0.1))

s.label(fouc2$Tco, clab = 0.5)

for(i in 0:5) {
    index <- (1:5) + 5 * i
    transi <- chull(fouc2$Tco[index, ])
    transi <- index[c(transi, transi[1])]
    lines(fouc2$Tco[transi, ], lty = i + 1)
}
```



Factorial map F3 × F4:

```
s.label(fouc2$Tco[, 3:4], clab = 0.5)

for(i in 0:5) {
    index <- (1:5) + 5 * i
    transi <- chull(fouc2$Tco[index, 3:4])
    transi <- index[c(transi, transi[1])]
    lines(fouc2$Tco[transi, 3:4], lty = i + 1)
}
```




```
par(mar = par.sauv)
```

Annexe 7
"Dissimilarity Coefficient" pour "The Encyclopedia of Measurement and Statistics"

Article sous presse à Sage Publications suite à l'invitation de Neil J. Salkind, éditeur de "The Encyclopedia of Measurement and Statistics", dont la publication est prévue pour 2007.

Dissimilarity Coefficient

Definition

A dissimilarity coefficient is a function which measures the difference between two objects. It is defined from a set $E \times E$ (e.g. $\mathbb{R} \times \mathbb{R}$, $\mathbb{R}^2 \times \mathbb{R}^2$, $\mathbb{R}^n \times \mathbb{R}^n$) to the non negative real numbers $\mathbb{R}^+$. Let g be a dissimilarity coefficient. Let x and y be two elements from E, g verifies the following properties:

$$g(x,x) = 0 \text{ (C1)},$$

$$g(x,y) = g(y,x) \text{ (C2: symmetry)},$$

$$g(x,y) \geq 0 \text{ (C3: positivity)}.$$

The function g is said to be a pseudo-metric if and only if g verifies C1, C2, C3 and the following property, let z be another element from E,

$$g(x,y) + g(y,z) \geq g(x,z) \text{ (C4: triangle inequality)}.$$

Furthermore, the function g is said to be a metric if and only if g verifies C1, C2, C3, C4 and the following additional property

$$g(x,y) = 0 \Rightarrow x = y \text{ (C5)}.$$

The value taken by g for two elements x and y is called "dissimilarity" if g is simply a dissimilarity coefficient, "semi-distance" if g is in addition a pseudo-metric, and "distance" if g is a metric.

The application of the function g to a finite set of S elements $\{x_1,...,x_k,...,x_S\}$ leads to a matrix of dissimilarities (or semi-distances, or distances) between pairs of the elements. This matrix is said to be Euclidean if and only if one can find S points M_k $(k = 1,...,S)$ which can be embedded in a Euclidean space so that the Euclidean distance between M_k and M_l is $g(x_k,x_l) = \|M_k M_l\|$; $g(x_k,x_l) = \sqrt{(\mathbf{c}_k - \mathbf{c}_l)^t (\mathbf{c}_k - \mathbf{c}_l)}$, where $\mathbf{c}_k$ and $\mathbf{c}_l$ are the vectors

of coordinates for M_k and M_l, respectively, in the Euclidean space. These vectors of coordinates can be obtained by a Principal Coordinate Analysis. Consequently, the interest of this Euclidean property is the direct association between the dissimilarities and the obtention of a typology, a graphical representation of the dissimilarities among elements. Other types of graphical displays can be obtained with any dissimilarity coefficient by hierarchical cluster analysis and nonmetric multidimensional scaling.

Examples

Example 1: Let E be the Euclidean space $\mathbb{R}^n$, vector space of all n-tuples of real numbers $(x_1,...,x_i,...,x_n)$. An element of this space is noted $\mathbf{x}_k$. In that case, each element may be characterized by n quantitative variables X_1, ..., X_i, ..., X_n. Let $\mathbf{x}_k = (x_{1k},...,x_{ik},...,x_{nk})^t$ and $\mathbf{x}_l = (x_{1l},...,x_{il},...,x_{nl})^t$ be two vectors containing the values taken by the objects k and l, respectively, for each of the variables considered; $\mathbf{x}_k, \mathbf{x}_l \in \mathbb{R}^n$. The following dissimilarity coefficients can be used to measure the difference between the objects k and l:

- the Euclidean metric

$$g_1(\mathbf{x}_k, \mathbf{x}_l) = \sqrt{(\mathbf{x}_k - \mathbf{x}_l)^t (\mathbf{x}_k - \mathbf{x}_l)};$$

- the Joreskog distance

$$g_2(\mathbf{x}_k, \mathbf{x}_l) = \sqrt{(\mathbf{x}_k - \mathbf{x}_l)^t \mathbf{V}^{-1} (\mathbf{x}_k - \mathbf{x}_l)},$$

where $\mathbf{V} = diag(V(Y_1),...,V(Y_i),...,V(Y_n))$ is the diagonal matrix containing the variances of the n variables;

- the Mahalanobis distance

$$g_3(\mathbf{x}_k, \mathbf{x}_l) = \sqrt{(\mathbf{x}_k - \mathbf{x}_l)^t \mathbf{W}^{-1} (\mathbf{x}_k - \mathbf{x}_l)},$$

where $\mathbf{W}$ is the variance-covariance matrix for the n variables.

All these dissimilarity coefficients are metrics and provide Euclidean dissimilarity matrices.

Example 2: Let E be the set of frequency vectors:

$$E = \left\{ \mathbf{p} = (p_1, ..., p_k, ..., p_S) \mid p_k \geq 0, \sum_{k=1}^{S} p_k = 1 \right\}.$$

In that case, let $\mathbf{p}$ and $\mathbf{q}$ be two vectors from E. Several functions can be used to measure the dissimilarity between the two frequency vectors:

- the Euclidean metric $g_1(\mathbf{p}, \mathbf{q})$,

- the taxicab metric, also called Manhattan distance

$$g_4(\mathbf{p}, \mathbf{q}) = \sum_{k=1}^{S} |p_k - q_k|.$$

Modifications of these functions have been proposed in genetics and ecology so that their values lie between 0 and 1:

- the Rogers distance

$$g_5(\mathbf{p}, \mathbf{q}) = \sqrt{\frac{1}{2}(\mathbf{p} - \mathbf{q})^t (\mathbf{p} - \mathbf{q})};$$

- the minimal distance from Nei

$$g_6(\mathbf{p}, \mathbf{q}) = \frac{1}{2}(\mathbf{p} - \mathbf{q})^t (\mathbf{p} - \mathbf{q});$$

- and the absolute genetic distance from Gregorius

$$g_7(\mathbf{p}, \mathbf{q}) = \frac{1}{2} \sum_{k=1}^{S} |p_k - q_k|.$$

The dissimilarity coefficients g_4, g_5 and g_7 are metrics, but g_6 is not because it does not verify the triangle inequality (property C4).

Other dissimilarity coefficients have been developed exclusively for frequency vectors. In 1946, Bhattacharyya introduced the notion of angular distance, considering two multinomial sets characterized by two frequency vectors $\mathbf{p}$ and $\mathbf{q}$. The two vectors $\left(\sqrt{p_1},...,\sqrt{p_S}\right)$ and $\left(\sqrt{q_1},...,\sqrt{q_S}\right)$ can be considered as the directions of two lines starting from the origin of a multidimensional space and separated by an angle θ whose cosine is

$$\cos\theta = \sum_{k=1}^{S} \sqrt{p_k q_k} \ .$$

The coefficient of dissimilarity proposed by Bhattacharyya is the squared value of this angle:

$$g_8(\mathbf{p},\mathbf{q}) = \theta^2 = \left[\cos^{-1}\left(\sum_{k=1}^{S} \sqrt{p_k q_k}\right)\right]^2 .$$

Another series of dissimilarity coefficients stems from the probability of drawing two similar objects from two populations with respective frequency vectors $\mathbf{p}$ and $\mathbf{q}$:

$$\sum_{k=1}^{S} p_k q_k \ .$$

Among the dissimilarity coefficients developed from this probability are Nei dissimilarity index in genetics

$$g_9(\mathbf{p},\mathbf{q}) = -\ln\left(\frac{\sum_{k=1}^{S} p_k q_k}{\sqrt{\sum_{k=1}^{S} p_k^2}\sqrt{\sum_{k=1}^{S} q_k^2}}\right),$$

and Manly overlap index in ecology

$$g_{10}(\mathbf{p},\mathbf{q}) = 1 - \frac{\sum_{k=1}^{S} p_k q_k}{\sqrt{\sum_{k=1}^{S} p_k^2}\sqrt{\sum_{k=1}^{S} q_k^2}} .$$

Example 3: Let E be the set of binary vectors:

$$E = \left\{ \mathbf{u} = \left(u_1, ..., u_k, ..., u_S \right) | \, u_k \in \left\{ 0,1 \right\} \right\}.$$

In that case many dissimilarity coefficients, whose values lie between 0 and 1, have been developed in ecology where a vector $\mathbf{u}$ gives the presence/absence of species in a community. They have been defined from similarity coefficients. Let $\mathbf{u}$ and $\mathbf{v}$ be two vectors from E. For each position in the vectors, that is to say for $k = 1, ..., S$, the coefficients look at the similarity of the values taken by $\mathbf{u}$ and $\mathbf{v}$: u_k and v_k. Let a be the number of positions, for $k = 1, ..., S$, where $u_k = 1$ and $v_k = 1$, b the number of positions where $u_k = 1$ and $v_k = 0$, c the number of positions where $u_k = 0$ and $v_k = 1$ and d the number of positions where $u_k = 0$ and $v_k = 0$. The most used similarity coefficient is Jaccard similarity coefficient

$$s_{11}\left(\mathbf{u}, \mathbf{v} \right) = \frac{a}{a + b + c}.$$

A modification of s_{11}, the Sokal and Michener similarity coefficient

$$s_{12}\left(\mathbf{u}, \mathbf{v} \right) = \frac{a + d}{a + b + c + d}$$

takes into account the number of species absent from the two communities compared but present in other comparable communities.

Two modifications of coefficients s_{11} and s_{12}, the Sokal and Sneath similarity coefficient

$$s_{13}\left(\mathbf{u}, \mathbf{v} \right) = \frac{a}{a + 2\left(b + c \right)},$$

and the Rogers and Tanimoto similarity coefficient

$$s_{14}\left(\mathbf{u}, \mathbf{v} \right) = \frac{a + d}{a + 2\left(b + c \right) + d},$$

give to the difference (measures b and c) twice as much weight as to the similitude (measures a and d) between the two communities compared. The most common dissimilarity coefficient associated to each similarity coefficient is equal to $g = 1 - s$. It is

metric for $g_{11} = 1 - s_{11}$, $g_{12} = 1 - s_{12}$, $g_{13} = 1 - s_{13}$ and $g_{14} = 1 - s_{14}$. For all these coefficients, a matrix of dissimilarities calculated by $g^* = \sqrt{1 - s}$ is Euclidean.

Dissimilarity and diversity coefficients: Rao's unified approach

The concepts of dissimilarity and diversity have been linked together by C. R. Rao in a unified theoretical framework. The diversity is the character of objects that exhibit variety. A population in which the objects are numerous and different possesses variety. This variety depends on the relative abundance of the objects and the dissimilarity among these objects. This assertion is at the root of Rao's unified approach.

Consider a population i of S elements $\{x_1, ..., x_k, ..., x_S\}$ from any set E. Suppose that these elements are distributed in the population i according to the frequency vector $\mathbf{p}_i = (p_{1i}, ..., p_{ki}, ..., p_{Si})^t$. One can calculate $\Delta = [\delta_{kl}]$, $1 \le k \le S$, $1 \le l \le S$, a matrix of dissimilarities among the elements, by g, a chosen dissimilarity coefficient: $\delta_{kl} = g(x_k, x_l)$. The diversity in the population i depends on the frequency vector $\mathbf{p}_i$ and the dissimilarity matrix Δ. Rao defined a diversity coefficient, also called quadratic entropy, as

$$H(\mathbf{p}_i) = \sum_{k=1}^{S} \sum_{l=1}^{S} p_{ki} p_{li} \frac{[g(x_k, x_l)]^2}{2}.$$

Consider the matrix $\mathbf{D} = [\delta_{kl}^2 / 2]$, i.e. $\mathbf{D} = [g(x_k, x_l)^2 / 2]$, changing Δ for $\mathbf{D}$, in the notations, leads to

$$H(\mathbf{p}_i) = \mathbf{p}_i^t \mathbf{D} \mathbf{p}_i.$$

This coefficient has the special feature of being associated with the Jensen difference, a dissimilarity coefficient which calculates dissimilarities among two populations:

$$J\left(\mathbf{p}_i,\mathbf{p}_j\right)=2H\left(\frac{\mathbf{p}_i+\mathbf{p}_j}{2}\right)-H\left(\mathbf{p}_i\right)-H\left(\mathbf{p}_j\right)$$
$$=-\frac{1}{2}\left(\mathbf{p}_i-\mathbf{p}_j\right)^t\mathbf{D}\left(\mathbf{p}_i-\mathbf{p}_j\right).$$

The sign of J depends on g via $\mathbf{D}$. It is positive if g is a metric leading to Euclidean matrices. In addition, where g is a metric leading to Euclidean matrices, the dissimilarity coefficient is

$$f\left(\mathbf{p}_i,\mathbf{p}_j\right)=\sqrt{2J\left(\mathbf{p}_i,\mathbf{p}_j\right)}\,.$$

Interestingly, f, as g, is in that case a metric leading to Euclidean matrices. Consequently the coefficients g and f are measures of inertia (dispersion of points) in a Euclidean space. This result is the heart of the new ordination method called double Principal Coordinate Analysis. It allows a graphical representation of the dissimilarities among populations (coefficient f), together with a projection of the constituting elements.

Thus the coefficients g and f are connected in a framework of diversity decomposition. The total diversity over several populations is equal to the sum of the average diversity within populations and the diversity among populations. Each component of the decomposition is measured by the quadratic entropy but its formula depends either on g when it represents diversity among elements (total and intra diversity) or on f when it represents diversity among populations (inter diversity).

Consider r populations. A weight μ_i is attributed to population i so that $\sum_{i=1}^{r}\mu_i=1$. The diversity and dissimilarity coefficients are connected in the following diversity decomposition:

$$H\left(\sum_{i=1}^{r}\mu_i\mathbf{p}_i\right)=\sum_{i=1}^{r}\mu_i H\left(\mathbf{p}_i\right)+\sum_{i=1}^{r}\sum_{j=1}^{r}\mu_i\mu_j\frac{f\left(\mathbf{p}_i,\mathbf{p}_j\right)^2}{2}.$$

The component

$$H\left(\sum_{i=1}^{r}\mu_i\mathbf{p}_i\right)=\sum_{k=1}^{S}\sum_{l=1}^{S}\left(\sum_{i=1}^{r}\mu_i p_{ki}\right)\left(\sum_{i=1}^{r}\mu_i p_{li}\right)\frac{\left[g\left(x_k,x_l\right)\right]^2}{2}$$

stems from the total diversity irrespective of populations. It is measured by the quadratic entropy from g and the global frequencies of the S objects. The mean

$$\sum_{i=1}^{r}\mu_i H\left(\mathbf{p}_i\right)=\sum_{i=1}^{r}\mu_i\left[\sum_{k=1}^{S}\sum_{l=1}^{S}p_{ki}p_{li}\frac{g\left(x_k,x_l\right)^2}{2}\right]$$

is the average diversity within populations, also measured by the quadratic entropy from the dissimilarity coefficient g and from the frequencies of the objects within populations. Finally, the last term

$$\sum_{i=1}^{r}\sum_{j=1}^{r}\mu_i\mu_j\frac{f\left(\mathbf{p}_i,\mathbf{p}_j\right)^2}{2}$$

denotes the diversity among populations and is measured by the quadratic entropy from the dissimilarity coefficient f and the relative weights attributed to populations.

This general framework has two interesting specific cases.

Where E is the set of values taken by a qualitative variable X, and

$$\frac{g\left(x_k,x_l\right)^2}{2}=\begin{cases}1 & \text{if } x_k\neq x_l\\ 0 & \text{if } x_k=x_l\end{cases},$$

then

$$H\left(\mathbf{p}_i\right)=1-\sum_{k=1}^{S}p_{ki}^2$$

which is known as the Gini-Simpson diversity index, and,

$$f\left(\mathbf{p}_i,\mathbf{p}_j\right)=\sqrt{\left(\mathbf{p}_i-\mathbf{p}_j\right)^t\left(\mathbf{p}_i-\mathbf{p}_j\right)}.$$

is the Euclidean distance between $\mathbf{p}_i$ and $\mathbf{p}_j$.

Where E is the set of values taken by a quantitative variable X, and

$$g\left(x_k, x_l\right) = \left|x_k - x_l\right|, \tag{1}$$

then

$$H\left(\mathbf{p}_i\right) = \sum_{k=1}^{S} p_{ki}\left(x_k - \sum_{k=1}^{S} p_{ki} x_k\right)^2$$

which is the variance of the quantitative variable X. Let $\mathbf{x}$ be the vector $\left(x_1, ..., x_k, ..., x_S\right)^t$,

$$f\left(\mathbf{p}_i, \mathbf{p}_j\right) = \sqrt{\left(\mathbf{p}_i^t\mathbf{x} - \mathbf{p}_j^t\mathbf{x}\right)^t\left(\mathbf{p}_i^t\mathbf{x} - \mathbf{p}_j^t\mathbf{x}\right)},$$

which can be simply written as the absolute difference between two means

$$f\left(\mathbf{p}_i, \mathbf{p}_j\right) = \left|\sum_{k=1}^{S} p_{ki} x_k - \sum_{k=1}^{S} p_{kj} x_k\right|. \tag{2}$$

This second writting highlights the consistency between coefficient g (eq. (1)) measuring distances between elements and coefficient f (eq. (2)) measuring distances between populations.

In conclusion, the dissimilarity coefficients are functions which may correspond to inertia in Euclidean spaces provided that they verify additional properties. They are used in many disciplines and fit in perfectly with any diversity studies.

Sandrine Pavoine

Further Readings and References

Gower, J. C. 1966. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika **53**:325-338.

Legendre, P., and L. Legendre. 1998. Numerical ecology, 2nd English edition. Elsevier Science BV, Amsterdam.

Nei, M. 1987. Molecular evolutionary genetics. Columbia University Press, New York, NY, USA.

Pavoine, S., A. B. Dufour, and D. Chessel. 2004. From dissimilarities among species to dissimilarities among communities: a double principal coordinate analysis. Journal of Theoretical Biology **228**:523-537.

Rao, C. R. 1982. Diversity and dissimilarity coefficients: a unified approach. Theoretical Population Biology **21**:24-43.

From the Internet

The ade4 package for R at http://pbil.univ-lyon1.fr/R/rplus/ade4dsR.html which enables you to enter data and compute dissimilarity coefficients, diversity coefficients, the Principal Coordinates Analysis and the double Principal Coordinates Analysis. Details in:

Chessel, D., A.-B. Dufour, and J. Thioulouse. 2004. The ade4 package-I- One-table methods. R News **4**:5-10.

Annexe 8
Fonctions développées pour ade4

Fonctions développées dans le package ade4 de R
(Ihaka and Gentleman, 1996).
Programmation en langage R et C.

Les pages suivantes sont extraites du manuel de la nouvelle version d'ade4 qui sera bientôt disponible. Dans ces pages, les fonctions développées personnellement apparaissent en noir et les autres fonctions d'ade4 en gris. Ces fiches d'aide sont suivies par deux documents dans lesquels ces fonctions sont utilisées explicitement. Le premier contient les instructions qui permettent de refaire tous les calculs de l'article "Pavoine, S., S. Ollier, and A. B. Dufour. 2005. Is the originality of a species measurable ? Ecology letters 8 :579-586." Le second illustre l'existence de liens entre la mesure de la diversité par l'entropie quadratique, la mesure de la dissimilarité par le coefficient de Rao (différence de Jensen appliquée à l'entropie quadratique), l'inertie dans la double analyse en coordonnées principales, et la mesure de la variation dans l'AMOVA.

Ces fonctions sont les suivantes :

1. amova - analyse de variance moléculaire - manuel d'ade4 - page 5 ;
2. disc - coefficient de dissimilarité de Rao : différence de Jensen appliquée à l'entropie quadratique. Notons d cette différence. La fonction calcule la dissimilarité δ associée à ce coefficient : $\delta = \sqrt{(2d)}$ - manuel d'ade4 - page 49 ;
3. divc - coefficient de diversité de Rao. Il s'agit de l'entropie quadratique - manuel d'ade4 - page 60 ;
4. divcmax - valeur maximale de l'entropie quadratique pour une matrice de dissimilarités donnée - manuel d'ade4 - page 61 ;
5. dpcoa - double analyse en coordonnées principales - manuel d'ade4 - page 66 ;
6. plot.DPCoA - représentation graphique associée à la double analyse en coordonnées principales - manuel d'ade4 - page 66 ;
7. print.DPCoA - description dans la fenêtre de R des différents objets construits par la fonction dpcoa - manuel d'ade4 - page 66 ;
8. EH - quantité d'histoire évolutive - somme des longueurs de branches dans une phylogénie ou dans une partie de la phylogénie - manuel d'ade4 - page 86 ;

9. humDNA - données extraites de Excoffier *et al.* (1992) - manuel d'ade4 - page 102 ;
10. optimEH - processus d'optimisation de Nee et May (1997) - manuel d'ade4 - page 167 ;
11. originality - somme des valeurs d'originalité pour un ensemble d'espèces, calculée à partir d'un arbre phylogénétique - manuel d'ade4 - page 170 ;
12. orisaved - quantité maximale ou minimale d'originalité sauvée par un ensemble d'espèces choisies selon le critère d'optimalité de Nee et May (1997) - manuel d'ade4 - page 171 ;
13. randEH - processus aléatoire de Nee et May (1997), - manuel d'ade4 - page 201 ;
14. randtest.amova - tests associés à l'analyse moléculaire de variance - manuel d'ade4 - page 202 ;
15. plot.randtest.amova - histogrammes des valeurs théoriques, et valeur observée pour les tests de permutation associés à l'analyse moléculaire de variance - manuel d'ade4 - page 202.

Les fonctions 'amova', 'disc', 'divc', 'divcmax', 'dcpoa', 'plot.dpcoa', 'print.dpcoa', 'randtest.amova', et le jeu de données 'humDNAm' sont disponibles dans la version actuelle d'ade4 (vers. 1.3-3). Les autres seront disponible dans la nouvelle version d'ade4 déposée fin septembre 2005 et pourront aussi être trouvées à l'adresse suivante : http ://pbil.univ-lyon1.fr/R/donnees. Ces autres fonctions ont demandé la mise à jour des fonctions 'dotchart.phylog' et 'symbols.phylog' qui associent entre autres un diagramme de Cleveland à une phylogénie, et des données 'carni70', également disponibles à cette adresse.

Examples

```
data(aminoacyl)
aminoacyl$genes
aminoacyl$usage.codon
dudi.coa(aminoacyl$usage.codon, scannf = FALSE)
```

`amova` *Analysis of molecular variance*

Description

The analysis of molecular variance tests the differences among population and/or groups of populations in a way similar to ANOVA. It includes evolutionary distances among alleles.

Usage

```
amova(samples, distances, structures)
print.amova(x, full = FALSE, ...)
```

Arguments

`samples`	a data frame with haplotypes (or genotypes) as rows, populations as columns and abundance as entries
`distances`	an object of class `dist` computed from Euclidean distance. If `distances` is null, equidistances are used.
`structures`	a data frame containing, in the jth row and the kth column, the name of the group of level k to which the jth population belongs
`x`	an object of class `amova`
`full`	a logical value indicating whether the original data ('distances', 'samples', 'structures') should be printed
`...`	further arguments passed to or from other methods

Value

Returns a list of class `amova`

`call`	call
`results`	a data frame with the degrees of freedom, the sums of squares, and the mean squares. Rows represent levels of variability.
`componentsofcovariance`	a data frame containing the components of covariance and their contribution to the total covariance
`statphi`	a data frame containing the phi-statistics

Author(s)

Sandrine Pavoine ⟨pavoine@biomserv.univ-lyon1.fr⟩

References

Excoffier, L., Smouse, P.E. and Quattro, J.M. (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, **131**, 479–491.

See Also

randtest.amova

Examples

```
data(humDNAm)
amovahum <- amova(humDNAm$samples, sqrt(humDNAm$distances), humDNAm$structures)
amovahum
```

`apis108` *Allelic frequencies in ten honeybees populations at eight microsatellites loci*

Description

This data set gives the occurences for the allelic form on 8 loci in 10 populations of honeybees.

Usage

```
data(apis108)
```

Format

A data frame containing 180 rows (allelic forms on 8 loci) and 10 columns (populations of honeybees : El.Hermel, Al.Hoceima, Nimba, Celinda, Pretoria, Chalkidiki, Forli, Valenciennes, Umea and Seville).

Source

http://www.montpellier.inra.fr/URLB/apis/libanfreq.pdf

Franck P., Garnery L., Solignac M. and Cornuet J.M. (2000) Molecular confirmation of a fourth lineage in honeybees from the Near-East. *Apidologie*, **31**, 167–180.

Description

Calculates the root square of Rao's dissimilarity coefficient between samples.

Usage

```
disc(samples, dis = NULL, structures = NULL)
```

Arguments

`samples`	a data frame with elements as rows, samples as columns, and abundance, presence-absence or frequencies as entries
`dis`	an object of class `dist` containing distances or dissimilarities among elements. If `dis` is NULL, equidistances are used.
`structures`	a data frame containing, in the jth row and the kth column, the name of the group of level k to which the jth population belongs.

Value

Returns a list of objects of class `dist`

Author(s)

Sandrine Pavoine ⟨pavoine@biomserv.univ-lyon1.fr⟩

References

Rao, C.R. (1982) Diversity and dissimilarity coefficients: a unified approach. *Theoretical Population Biology*, **21**, 24–43.

Examples

```
data(humDNAm)
humDNA.dist <- disc(humDNAm$samples, sqrt(humDNAm$distances), humDNAm$structures)
humDNA.dist
is.euclid(humDNA.dist$samples)
is.euclid(humDNA.dist$regions)

## Not run:
data(ecomor)
ecomor.phylog <- taxo2phylog(ecomor$taxo)
ecomor.dist <- disc(ecomor$habitat, ecomor.phylog$Wdist)
ecomor.dist
is.euclid(ecomor.dist)
## End(Not run)
```

Author(s)

Daniel Chessel ⟨chessel@biomserv.univ-lyon1.fr⟩
Stéphane Dray ⟨dray@biomserv.univ-lyon1.fr⟩

Examples

```
data(ecomor)
par(mfrow = c(2,2))
scatter(dudi.pco(dist.quant(ecomor$morpho,3), scan = FALSE))
scatter(dudi.pco(dist.quant(ecomor$morpho,2), scan = FALSE))
scatter(dudi.pco(dist(scalewt(ecomor$morpho)), scan = FALSE))
scatter(dudi.pco(dist.quant(ecomor$morpho,1), scan = FALSE))
par(mfrow = c(1,1))
```

---

divc *Rao's diversity coefficient also called quadratic entropy*

Description

Calculates Rao's diversity coefficient within samples.

Usage

```
divc(df, dis, scale)
```

Arguments

`df`	a data frame with elements as rows, samples as columns, and abundance, presence-absence or frequencies as entries
`dis`	an object of class `dist` containing distances or dissimilarities among elements. If `dis` is NULL, Gini-Simpson index is performed.
`scale`	a logical value indicating whether or not the diversity coefficient should be scaled by its maximal value over all frequency distributions.

Value

Returns a data frame with samples as rows and the diversity coefficient within samples as columns

Author(s)

Sandrine Pavoine ⟨pavoine@biomserv.univ-lyon1.fr⟩

References

Rao, C.R. (1982) Diversity and dissimilarity coefficients: a unified approach. *Theoretical Population Biology*, **21**, 24–43.

Gini, C. (1912) Variabilitá e mutabilitá. *Universite di Cagliari III*, Parte II.

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

Champely, S. and Chessel, D. (2002) Measuring biological diversity using Euclidean metrics. *Environmental and Ecological Statistics*, **9**, 167–177.

Examples

```
data(ecomor)
ecomor.phylog <- taxo2phylog(ecomor$taxo)
divc(ecomor$habitat, ecomor.phylog$Wdist)

data(humDNAm)
divc(humDNAm$samples, sqrt(humDNAm$distances))
```

`divcmax` *Maximal value of Rao's diversity coefficient also called quadratic entropy*

Description

For a given dissimilarity matrix, this function calculates the maximal value of Rao's diversity coefficient over all frequency distribution. It uses an optimization technique based on Rosen's projection gradient algorithm and is verified using the Kuhn-Tucker conditions.

Usage

```
divcmax(dis, epsilon, comment)
```

Arguments

`dis`	an object of class `dist` containing distances or dissimilarities among elements.
`epsilon`	a tolerance threshold : a frequency is non null if it is higher than epsilon.
`comment`	a logical value indicating whether or not comments on the optimization technique should be printed.

Value

Returns a list

`value`	the maximal value of Rao's diversity coefficient.

`vectors` a data frame containing four frequency distributions : sim is a simple distribution which is equal to $\frac{D1}{1^t D1}$, pro is equal to $\frac{z}{1^t z1}$, where z is the nonnegative eigenvector of the matrix containing the squared dissimilarities among the elements, met is equal to z^2, num is a frequency vector maximizing Rao's diversity coefficient.

Author(s)

Stéphane Champely ⟨Stephane.Champely@univ-lyon1.fr⟩
Sandrine Pavoine ⟨pavoine@biomserv.univ-lyon1.fr⟩

References

Rao, C.R. (1982) Diversity and dissimilarity coefficients: a unified approach. *Theoretical Population Biology*, **21**, 24–43.

Gini, C. (1912) Variabilitá e mutabilitá. *Universite di Cagliari III*, Parte II.

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

Champely, S. and Chessel, D. (2002) Measuring biological diversity using Euclidean metrics. *Environmental and Ecological Statistics*, **9**, 167–177.

Pavoine, S., Ollier, S. and Pontier, D. (2005) Measuring diversity from dissimilarities with Rao's quadratic entropy: are any dissimilarities suitable? *Theoretical Population Biology*, **67**, 231–239.

Examples

```
par.safe <- par()$mar
data(elec88)
par(mar = c(0.1, 0.1, 0.1, 0.1))
# Departments of France.
area.plot(elec88$area)

# Dissimilarity matrix.
d0 <- dist(elec88$xy)

# Frequency distribution maximizing spatial diversity in France
# according to Rao's quadratic entropy.
France.m <- divcmax(d0)
w0 <- France.m$vectors$num
v0 <- France.m$value
(1:94) [w0 > 0]

# Smallest circle including all the 94 departments.
# The squared radius of that circle is the maximal value of the
# spatial diversity.
w1 = elec88$xy[c(6, 28, 66), ]
w.c = apply(w1 * w0[c(6, 28, 66)], 2, sum)
symbols(w.c[1], w.c[2], circles = sqrt(v0), inc = FALSE, add = TRUE)
s.value(elec88$xy, w0, add.plot = TRUE)
par(mar = par.safe)
```

`doubs$poi` contains the abundance of the following fish species: *Cottus gobio* (CHA), *Salmo trutta fario* (TRU), *Phoxinus phoxinus* (VAI), *Nemacheilus barbatulus* (LOC), *Thymallus thymallus* (OMB), *Telestes soufia agassizi* (BLA), *Chondrostoma nasus* (HOT), *Chondostroma toxostoma* (TOX), *Leuciscus leuciscus* (VAN), *Leuciscus cephalus cephalus* (CHE), *Barbus barbus* (BAR), *Spirlinus bipunctatus* (SPI), *Gobio gobio* (GOU), *Esox lucius* (BRO), *Perca fluviatilis* (PER), *Rhodeus amarus* (BOU), *Lepomis gibbosus* (PSO), *Scardinius erythrophtalmus* (ROT), *Cyprinus carpio* (CAR), *Tinca tinca* (TAN), *Abramis brama* (BCO), *Ictalurus melas* (PCH), *Acerina cernua* (GRE), *Rutilus rutilus* (GAR), *Blicca bjoerkna* (BBO), *Alburnus alburnus* (ABL), *Anguilla anguilla* (ANG).

Source

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

References

See a French description of fish species at `http://pbil.univ-lyon1.fr/R/articles/arti049.pdf`.

Chesse, D., Lebreton, J.D. and Yoccoz, N.G. (1987) Propriétés de l'analyse canonique des correspondances. Une illustration en hydrobiologie. *Revue de Statistique Appliquée*, **35**, 4, 55–72.

Examples

```
data(doubs)
pca1 <- dudi.pca(doubs$mil, scan = FALSE)
pca2 <- dudi.pca(doubs$poi, scale = FALSE, scan = FALSE)
coiner1 <- coinertia(pca1, pca2, scan = FALSE)
par(mfrow = c(3,3))
s.corcircle(coiner1$aX)
s.value(doubs$xy, coiner1$lX[,1])
s.value(doubs$xy, coiner1$lX[,2])
s.arrow(coiner1$c1)
s.match(coiner1$mX, coiner1$mY)
s.corcircle(coiner1$aY)
s.arrow(coiner1$l1)
s.value(doubs$xy, coiner1$lY[,1])
s.value(doubs$xy, coiner1$lY[,2])
par(mfrow = c(1,1))
```

---

dpcoa *Double principal coordinate analysis*

Description

Performs a double principal coordinate analysis

Usage

```
dpcoa (df, dis = NULL, scannf = TRUE, nf = 2, full = FALSE, tol = 1e-07)
plot.dpcoa (x, xax = 1, yax = 2, option = 1:4, csize = 2, ...)
print.dpcoa (x, ...)
```

Arguments

`df`	a data frame with elements as rows, samples as columns and abundance or presence-absence as entries
`dis`	an object of class `dist` containing the distances between the elements.
`scannf`	a logical value indicating whether the eigenvalues bar plot should be displayed
`nf`	if scannf is FALSE, an integer indicating the number of kept axes
`full`	a logical value indicating whether all non null eigenvalues should be kept
`tol`	a tolerance threshold for null eigenvalues (a value less than tol times the first one is considered as null)
`x`	an object of class `dpcoa`
`xax`	the column number for the x-axis
`yax`	the column number for the y-axis
`option`	the function `plot.dpcoa` produces four graphs, `option` allows us to choose only some of them
`csize`	a size coefficient for symbols
`...`	`...` further arguments passed to or from other methods

Value

Returns a list of class `dpcoa` containing:

`call`	call
`nf`	a numeric value indicating the number of kept axes
`w1`	a numeric vector containing the weights of the elements
`w2`	a numeric vector containing the weights of the samples
`eig`	a numeric vector with all the eigenvalues
`RaoDiv`	a numeric vector containing diversities within samples
`RaoDis`	an object of class `dist` containing the dissimilarities between samples
`RaoDecodiv`	a data frame with the decomposition of the diversity
`l1`	a data frame with the coordinates of the elements
`l2`	a data frame with the coordinates of the samples
`c1`	a data frame with the scores of the principal axes of the elements

Author(s)

Daniel Chessel ⟨chessel@biomserv.univ-lyon1.fr⟩
Sandrine Pavoine ⟨pavoine@biomserv.univ-lyon1.fr⟩
Anne B Dufour ⟨dufour@biomserv.univ-lyon1.fr⟩

References

Pavoine, S., Dufour, A.B. and Chessel, D. (2004) From dissimilarities among species to dissimilarities among communities: a double principal coordinate analysis. *Journal of Theoretical Biology*, **228**, 523–537.

Examples

```
data(humDNAm)
dpcoahum <- dpcoa(humDNAm$samples, sqrt(humDNAm$distances), scan = FALSE, nf = 2)
dpcoahum
plot(dpcoahum, csize = 1.5)
## Not run:
data(ecomor)
ecomor.phylog <- taxo2phylog(ecomor$taxo)
dpcoaeco <- dpcoa(ecomor$habitat, ecomor.phylog$Wdist, scan = FALSE, nf = 2)
dpcoaeco
plot(dpcoaeco, csize = 1.5)
## End(Not run)
```

---

dudi.acm *Multiple Correspondence Analysis*

Description

`dudi.acm` performs the multiple correspondence analysis of a factor table.
`acm.burt` an utility giving the crossed Burt table of two factors table.
`acm.disjonctif` an utility giving the complete disjunctive table of a factor table.
`boxplot.acm` a graphic utility to interpret axes.

Usage

```
dudi.acm (df, row.w = rep(1, nrow(df)), scannf = TRUE, nf = 2)
acm.burt (df1, df2, counts = rep(1, nrow(df1)))
acm.disjonctif (df)
boxplot.acm (x, xax = 1, ...)
```

Arguments

`df, df1, df2`	data frames containing only factors
`row.w, counts`	vector of row weights, by default, uniform weighting

```
     clabel.c = 0.75, clabel.r = 0.5, csi = 0.75, cleg = 0)
plot.phylog(ecophy, clabel.n = 0.75, clabel.l = 0.75,
     labels.l = ecomor$labels[,"latin"])
dtaxo <- ecophy$Wdist
mantel.randtest(dmorpho, dtaxo)
mantel.randtest(dhabitat, dtaxo)
## End(Not run)
```

EH *Amount of Evolutionary History*

Description

computes the sum of branch lengths on an ultrametric phylogenetic tree.

Usage

```
EH(phyl, select = NULL)
```

Arguments

`phyl`	an object of class phylog
`select`	a vector containing the numbers of the leaves (species) which must be considered in the computation of the amount of Evolutionary History. This parameter allows the calculation of the amount of Evolutionary History for a subset of species.

Value

returns a real value.

Author(s)

Sandrine Pavoine ⟨pavoine@biomserv.univ-lyon1.fr⟩

References

Nee, S. and May, R.M. (1997) Extinction and the loss of evolutionary history. *Science*, **278**, 692–694.

Examples

```
data(carni70)
carni70.phy <- newick2phylog(carni70$tre)
EH(carni70.phy)
EH(carni70.phy, select = 1:15) # Felidae
```

housetasks *Contingency Table*

Description

The housetasks data frame gives 13 housetasks and their repartition in the couple.

Usage

```
data(housetasks)
```

Format

This data frame contains four columns : wife, alternating, husband and jointly. Each column is a numeric vector.

Source

Kroonenberg, P. M. and Lombardo, R. (1999) Nonsymmetric correspondence analysis: a tool for analysing contingency tables with a dependence structure. *Multivariate Behavioral Research*, **34**, 367–396

Examples

```
data(housetasks)
nsc1 <- dudi.nsc(housetasks, scan = FALSE)
s.label(nsc1$c1, clab = 1.25)
s.arrow(nsc1$li, add.pl = TRUE, clab = 0.75)
```

---

humDNAm *human mitochondrial DNA restriction data*

Description

This data set gives the frequencies of haplotypes of mitochondrial DNA restriction data in ten populations all over the world.
It gives also distances among the haplotypes.

Usage

```
data(humDNAm)
```

Format

`humDNAm` is a list of 3 components.

distances is an object of class `dist` with 56 haplotypes. These distances are computed by counting the number of differences in restriction sites between two haplotypes.

samples is a data frame with 56 haplotypes, 10 abundance variables (populations). These variables give the haplotype abundance in a given population.

structures is a data frame with 10 populations, 1 variable (classification). This variable gives the name of the continent in which a given population is located.

Source

Excoffier, L., Smouse, P.E. and Quattro, J.M. (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, **131**, 479–491.

Examples

```
data(humDNAm)
dpcoahum <- dpcoa(humDNAm$samples,
    sqrt(humDNAm$distances), scan = FALSE, nf = 2)
plot(dpcoahum, csize = 1.5)
```

---

ichtyo *Point sampling of fish community*

Description

This data set gives informations between a faunistic array, the total number of sampling points made at each sampling occasion and the year of the sampling occasion.

Usage

```
data(ichtyo)
```

Format

`ichtyo` is a list of 3 components.

tab is a faunistic array with 9 columns and 32 rows.

eff is a vector of the 32 sampling effort.

dat is a factor where the levels are the 10 years of the sampling occasion.

Details

The value *n(i,j)* at the *ith* row and the *jth* column in `tab` corresponds to the number of sampling points of the *ith* sampling occasion (in `eff`) that contains the *jth* species.

Source

Example 357 in:
Hand, D.J., Daly, F., Lunn, A.D., McConway, K.J. and Ostrowski, E. (1994) *A handbook of small data sets*, Chapman & Hall, London. 458 p.

Lunn, A. D. and McNeil, D.R. (1991) *Computer-Interactive Data Analysis*, Wiley, New York

Examples

```
data(olympic)
pca1 <- dudi.pca(olympic$tab, scan = FALSE)
par(mfrow = c(2,2))
barplot(pca1$eig)
s.corcircle(pca1$co)
plot(olympic$score, pca1$l1[,1])
abline(lm(pca1$l1[,1]~olympic$score))
s.label(pca1$l1, clab = 0.5)
s.arrow(2 * pca1$co, add.p = TRUE)
par(mfrow = c(1,1))
```

---

optimEH *Nee and May's optimizing process*

Description

performs Nee and May's optimizing scheme. When branch lengths in an ultrametric phylogenetic tree are expressed as divergence times, the total sum of branch lengths in that tree expresses the amount of evolutionary history. Nee and May's algorithm optimizes the amount of evolutionary history preserved if only k species out of n were to be saved. The k-1 closest-to-root nodes are selected, which defines k clades; one species from each clade is picked. At this last step, we decide to select the most original species of each from the k clades.

Usage

```
optimEH(phyl, nbofsp, tol = 1e-8, give.list = TRUE)
```

Arguments

`phyl`	an object of class phylog
`nbofsp`	an integer indicating the number of species saved (k).
`tol`	a tolerance threshold for null values (a value less than `tol` in absolute terms is considered as NULL).
`give.list`	logical value indicating whether a list of optimizing species should be provided. If `give.list = TRUE`, `optimPD` provides the list of the k species which optimize the amount of evolutionary history preserved and are the most original species in their clades. If `give.list = FALSE`, `optimPD`

returns directly the real value giving the amount of evolutionary history preserved.

Value

Returns a list containing:

`value`	a real value providing the amount of evolutionary history preserved.
`selected.sp`	a data frame containing the list of the k species which optimize the amount of evolutionary history preserved and are the most original species in their clades.

Author(s)

Sandrine Pavoine ⟨pavoine@biomserv.univ-lyon1.fr⟩

References

Nee, S. and May, R.M. (1997) Extinction and the loss of evolutionary history. *Science* **278**, 692–694.

Pavoine, S., Ollier, S. and Dufour, A.-B. (2005) Is the originality of a species measurable? *Ecology Letters*, **8**, 579–586.

See Also

`randEH`

Examples

```
data(carni70)
carni70.phy <- newick2phylog(carni70$tre)
optimEH(carni70.phy, nbofsp = 7, give.list = TRUE)
```

---

oribatid *Oribatid mite*

Description

This data set contains informations about environmental control and spatial structure in ecological communities of Oribatid mites.

Usage

```
data(oribatid)
```

`originality` *Originality of a species*

Description

computes originality values for species from an ultrametric phylogenetic tree.

Usage

```
originality(phyl, method = 5)
```

Arguments

`phyl`	an object of class phylog
`method`	a vector containing integers between 1 and 5.

Value

Returns a data frame with species in rows, and the selected indices of originality in columns. Indices are expressed as percentages.

Author(s)

Sandrine Pavoine ⟨pavoine@biomserv.univ-lyon1.fr⟩

References

Pavoine, S., Ollier, S. and Dufour, A.-B. (2005) Is the originality of a species measurable? *Ecology Letters*, **8**, 579–586.

Vane-Wright, R.I., Humphries, C.J. and Williams, P.H. (1991). What to protect? Systematics and the agony of choice. *Biological Conservation*, **55**, 235–254.

May, R.M. (1990). Taxonomy as destiny. *Nature*, **347**, 129–130.

Nixon, K.C. & Wheeler, Q.D. (1992). Measures of phylogenetic diversity. In: *Extinction and Phylogeny* (eds. Novacek, M.J. and Wheeler, Q.D.), 216–234, Columbia University Press, New York.

Examples

```
data(carni70)
carni70.phy <- newick2phylog(carni70$tre)
ori.tab <- originality(carni70.phy, 1:5)
names(ori.tab)
dotchart.phylog(carni70.phy, ori.tab, scaling=FALSE, yjoining=0,
    cleaves=0, ceti=0.5, csub=0.7, cdot=0.5)
```

`orisaved` *Maximal or minimal amount of originality saved under optimal conditions*

Description

computes the maximal or minimal amount of originality saved over all combinations of species optimizing the amount of evolutionary history preserved. The originality of a species is measured with the QE-based index.

Usage

```
orisaved(phyl, rate = 0.1, method = 1)
```

Arguments

`phyl`	an object of class phylog
`rate`	a real value (between 0 and 1) indicating how many species will be saved for each calculation. For example, if the total number of species is 70 and 'rate = 0.1' then the calculations will be done at a rate of 10 % i.e. for 0 (= 0 %), 7 (= 10 %), 14 (= 20 %), 21 (= 30 %), ..., 63 (= 90 %) and 70(= 100 %) species saved. If 'rate = 0.5' then the calculations will be done for only 0 (= 0 %), 35 (= 50 %) and 70(= 100 %) species saved.
`method`	an integer either 1 or 2 (see details).

Details

1 = maximum amount of originality saved 2 = minimum amount of originality saved

Value

Returns a numeric vector.

Author(s)

Sandrine Pavoine ⟨pavoine@biomserv.univ-lyon1.fr⟩

References

Pavoine, S., Ollier, S. and Dufour, A.-B. (2005) Is the originality of a species measurable? *Ecology Letters*, **8**, 579–586.

Examples

```
data(carni70)
carni70.phy <- newick2phylog(carni70$tre)
tmax <- orisaved(carni70.phy, rate = 1 / 70, method = 1)
tmin <- orisaved(carni70.phy, rate = 1 / 70, method = 2)
plot(c(0, 1:70), tmax, xlab = "nb of species saved", ylab = "Originality saved", type = "l")
lines(c(0, 1:70), tmin, lty = 2)
```

`orthobasis` *Orthonormal basis for orthonormal transform*

Description

These functions returns object of class `'orthobasis'` that contains data frame with n rows and *n-1* columns. Each data frame defines an orthonormal basis for the uniform weights.

`orthobasic.neig` returns the eigen vectors of the matrix N-M where M is the symmetric n by n matrix of the between-sites neighbouring graph and N is the diagonal matrix of neighbour numbers.
`orthobasis.line` returns the analytical solution for the linear neighbouring graph.
`orthobasic.circ` returns the analytical solution for the circular neighbouring graph.
`orthobsic.mat` returns the eigen vectors of the general link matrix M.
`orthobasis.listw` returns the eigen vectors of the general link matrix M associated to a `listw` object.
`orthobasis.haar` returns wavelet haar basis.

Usage

```
orthobasis.neig(neig)
orthobasis.line(n)
orthobasis.circ(n)
orthobasis.mat(mat, cnw=TRUE)
orthobasis.listw(listw)
orthobasis.haar(n)
print.orthobasis(x,...)
```

Arguments

`neig`	is an object of class `neig`
`n`	is an integer that defines length of vectors
`mat`	is a n by n phylogenetic or spatial link matrix
`listw`	is a `'listw'` object
`cnw`	if TRUE, the matrix of the neighbouring graph is modified to give Constant Neighbouring Weights
`x`	is an object of class `orthobasis`
`...`	: further arguments passed to or from other methods

Description

When branch lengths in an ultrametric phylogenetic tree are expressed as divergence times, the total sum of branch lengths in that tree expresses the amount of evolutionary history. The function `randPD` calculates the amount of evolutionary history preserved when k random species out of n original species are saved.

Usage

```
randEH(phyl, nbofsp, nbrep = 10)
```

Arguments

`phyl`	an object of class phylog
`nbofsp`	an integer indicating the number of species saved (k).
`nbrep`	an integer indicating the number of random sampling.

Value

Returns a numeric vector

Author(s)

Sandrine Pavoine ⟨pavoine@biomserv.univ-lyon1.fr⟩

References

Nee, S. and May, R.M. (1997) Extinction and the loss of evolutionary history. *Science* **278**, 692–694.

Pavoine, S., Ollier, S. and Dufour, A.-B. (2005) Is the originality of a species measurable? *Ecology Letters*, **8**, 579–586.

See Also

optimEH

Examples

```
data(carni70)
carni70.phy <- newick2phylog(carni70$tre)
mean(randEH(carni70.phy, nbofsp = 7, nbrep = 1000))

## Not run:
# the folowing instructions can last about 2 minutes.
data(carni70)
```

```
carni70.phy <- newick2phylog(carni70$tre)
percent <- c(0,0.04,0.07,seq(0.1,1,by=0.1))
pres <- round(percent*70)
topt <- sapply(pres, function(i) optimEH(carni70.phy, nbofsp = i, give = F))
topt <- topt / EH(carni70.phy)
tsam <- sapply(pres, function(i) mean(randEH(carni70.phy, nbofsp = i, nbrep = 1000)))
tsam <- tsam / EH(carni70.phy)
plot(pres, topt, xlab = "nb of species saved", ylab = "Evolutionary history saved", type = "l")
lines(pres, tsam)
## End(Not run)
```

`randtest-internal` *Internal Permutation Tests functions (in C).*

Description

Internal Permutation Tests functions (in C)

Usage

Details

These are not to be called by the user.

`randtest.amova` *Permutation tests on an analysis of molecular variance (in C).*

Description

Tests the components of covariance with permutation processes described by Excoffier et al. (1992).

Usage

```
randtest.amova(xtest, nrepet = 99, ...)
```

Arguments

`xtest`	an object of class `amova`
`nrepet`	the number of permutations
`...`	further arguments passed to or from other methods

Value

returns an object of class `krandtest` or `randtest`

Author(s)

Sandrine Pavoine ⟨pavoine@biomserv.univ-lyon1.fr⟩

References

Excoffier, L., Smouse, P.E. and Quattro, J.M. (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, **131**, 479–491.

Examples

```
data(humDNAm)
amovahum <- amova(humDNAm$samples, sqrt(humDNAm$distances), humDNAm$structures)
amovahum
randtesthum <- randtest.amova(amovahum, 49)
plot(randtesthum)
```

`randtest.between` *Monte-Carlo Test on the between-groups inertia percentage (in C).*

Description

Performs a Monte-Carlo test on the between-groups inertia percentage.

Usage

```
randtest.between(xtest, nrepet = 999, ...)
```

Arguments

`xtest`	an object of class `between`
`nrepet`	the number of permutations
`...`	further arguments passed to or from other methods

Value

a list of the class `randtest`

Author(s)

Jean Thioulouse ⟨ade4-jt@biomserv.univ-lyon1.fr⟩

References

Romesburg, H. C. (1985) Exploring, confirming and randomization tests. *Computers and Geosciences*, **11**, 19–37.

Dernière mise à jour de l'annexe électronique pour l'article
Pavoine, S., S. Ollier, and A. B. Dufour. 2005. Is the originality of a species measurable ?
Ecology letters 8 :579-586.

Ce document utilise les fonctions 'EH', 'optimEH', 'originality' et 'randEH'.

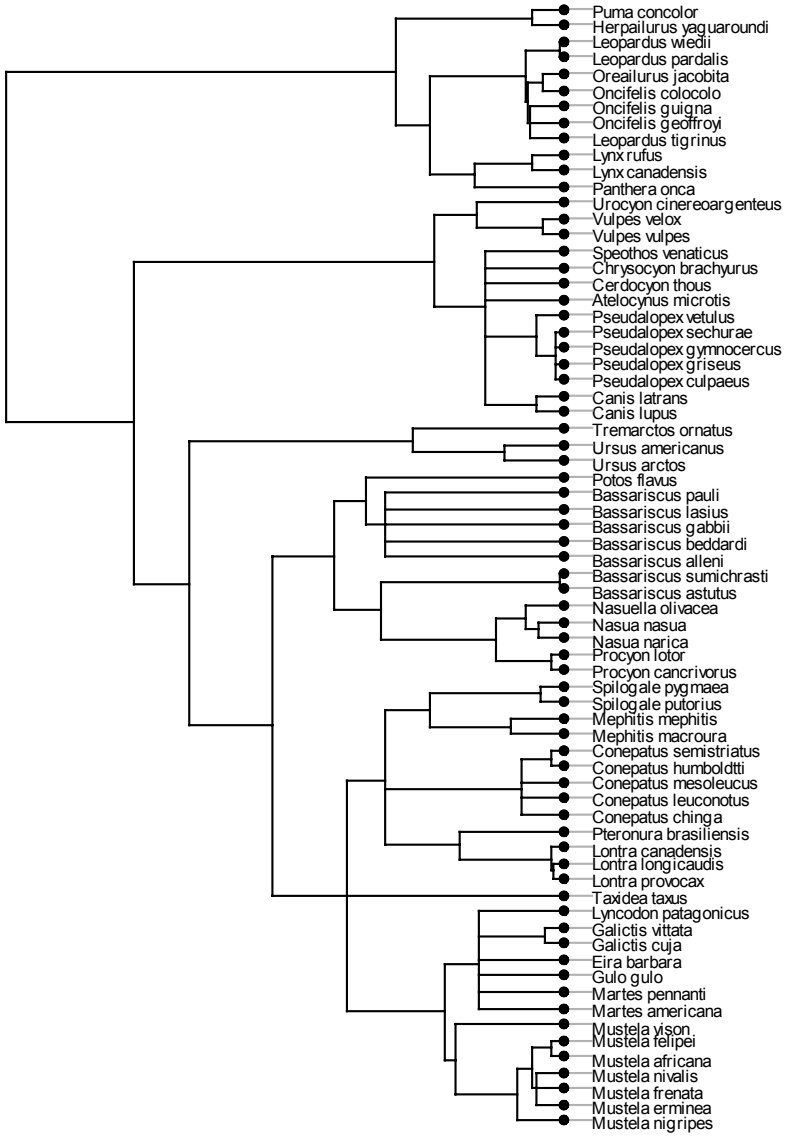
Figures from Pavoine, S., S. Ollier, and A. B. Dufour. 2005. Is the originality of a species measurable? Ecology letters **8**:579-586.

In your working directory, write the following instructions on your R console:

```
library(ade4)
data(carni70)
carni70.phy <- newick2phylog(carni70$tre)
```

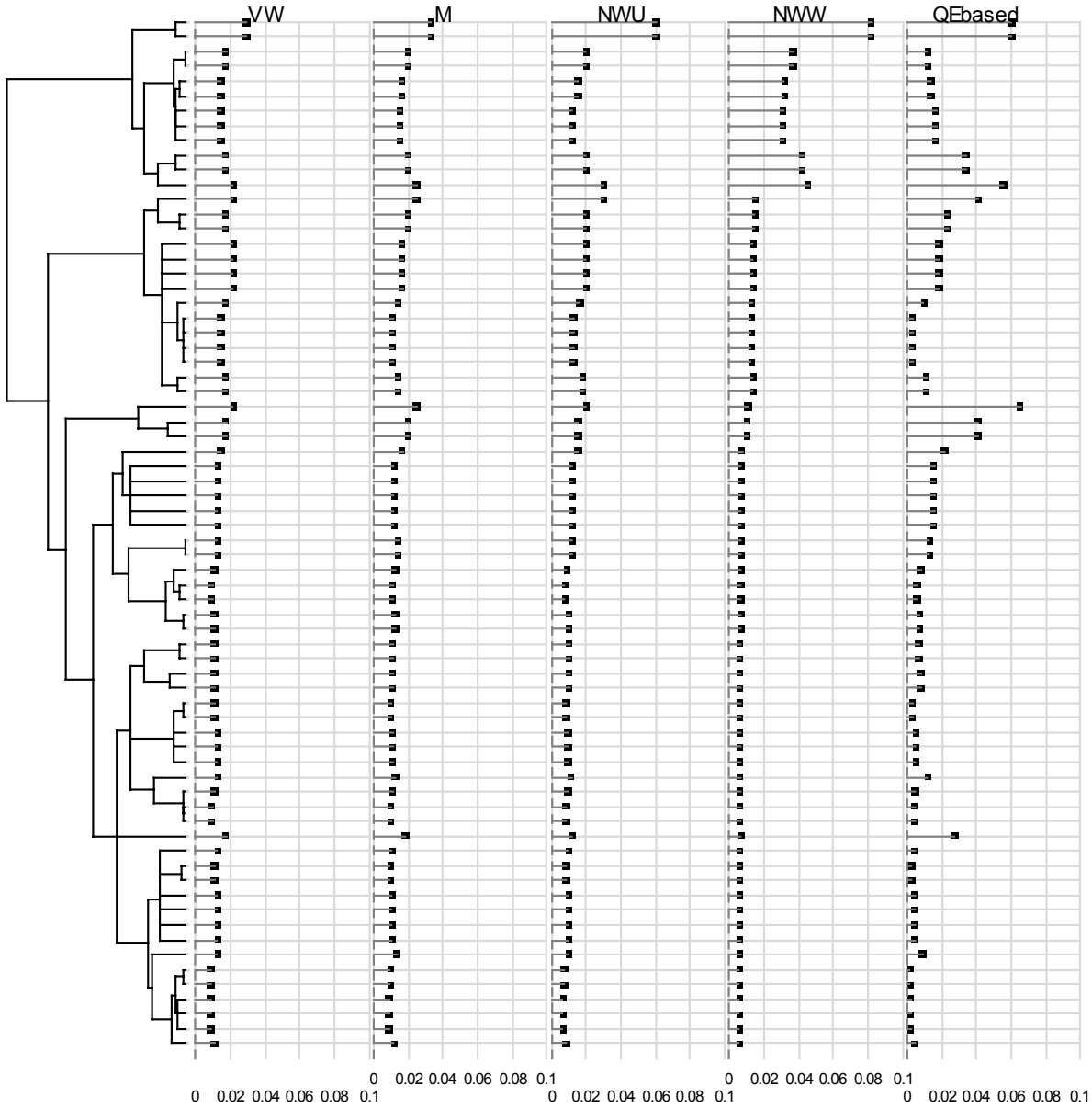
Display of the phylogeny:

```
plot.phylog(carni70.phy, cleaves = 0.5, clabel.leaves = 0.5)
```



Obtaining the figure 2 of the manuscript:

```
ori.tab <- originality(carni70.phy, 1:5)
names(ori.tab)
[1] "VW"      "M"       "NWU"     "NWW"     "QEbased"
dotchart.phylog(carni70.phy, originality(carni70.phy, 1:5),
    scaling=FALSE, yjoining=0, cleaves=0, ceti=0.5, csub=0.7, cdot=0.5)
```



Obtaining the figure 3a of the manuscript:

It is possible to obtain the k species which optimize the amount of evolutionary history and have the highest originality of their clade.
For example, if k = 7:

```
optimEH(carni70.phy, nbofsp = 7, give.list = TRUE)
$value
[1] 256.1

$selected.sp
                                        names
1       Puma_concolor OR Herpailurus_yaguaroundi
2                     Urocyon_cinereoargenteus
3                           Tremarctos_ornatus
4                                 Potos_flavus
5 Bassariscus_sumichrasti OR Bassariscus_astutus
6                                 Taxidea_taxus
7                                 Mustela_vison
```

The following instructions give the amount of evolutionary history preserved if 0%, 4%, 7%, 10%, 20%, 30%, 40%, ... and 100% of the species are saved according to the optimizing process.

```
percent <- c(0,0.04,0.07,seq(0.1,1,by=0.1))
percent
[1] 0.00 0.04 0.07 0.10 0.20 0.30 0.40 0.50 0.60 0.70 0.80 0.90 1.00
pres <- round(percent*70)
pres
[1]  0  3  5  7 14 21 28 35 42 49 56 63 70
topt <- sapply(pres, function(i)
    optimEH(carni70.phy, nbofsp = i, give = F))
topt
[1]   0.0 149.1 213.2 256.1 381.9 484.9 549.9 604.9 642.1 666.5 685.2 696.5
[13] 701.6
```

The vector topt can be expressed as percentages:

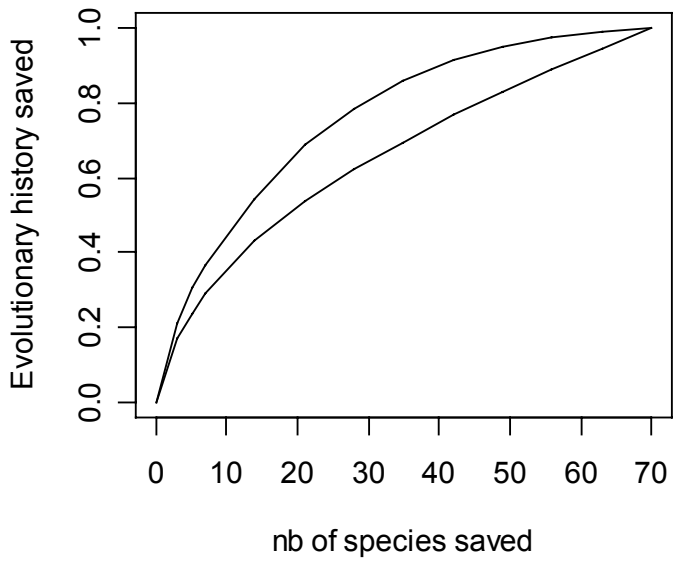
```
topt <- topt / EH(carni70.phy)
topt
[1] 0.0000 0.2125 0.3039 0.3650 0.5443 0.6911 0.7838 0.8622 0.9152 0.9500
[11] 0.9766 0.9927 1.0000
```

With the function 'randEH', it is possible to calculate the average amount of evolutionary history preserved by saving k random species.
For example, if k = 7, over 1000 random samples:

```
mean(randEH(carni70.phy, nbofsp = 7, nbrep = 1000))
[1] 203.2
```

The following instruction gives the amount of evolutionary history preserved if 0%, 4%, 7%, 10%, 20%, 30%, 40%, ... and 100% of the species are saved at random. Please, be aware that this instruction can last about 2 minutes.

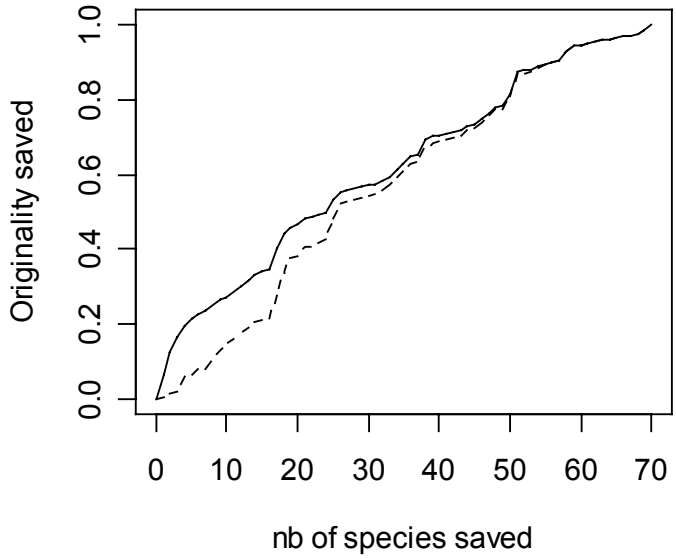
```
tsam <- sapply(pres, function(i)
    mean(randEH(carni70.phy, nbofsp = i, nbrep = 1000)))
tsam
[1]   0.0 119.1 165.7 204.8 303.4 375.7 436.8 487.6 538.2 583.4 625.3 664.5
[13] 701.6
tsam <- tsam / EH(carni70.phy)
tsam
[1] 0.0000 0.1698 0.2362 0.2919 0.4324 0.5355 0.6226 0.6950 0.7671 0.8315
[11] 0.8913 0.9472 1.0000

plot(pres, topt, xlab = "nb of species saved",
    ylab = "Evolutionary history saved", type = "l")

lines(pres, tsam)
```



In this figure, the top curve is for the optimizing scheme; and the bottom curve is for random samples.

Obtaining the figure 3b of the manuscript:

```
tmax <- orisaved(carni70.phy, rate = 1 / 70, method = 1)
tmin <- orisaved(carni70.phy, rate = 1 / 70, method = 2)

plot(c(0, 1:70), tmax, xlab = "nb of species saved", ylab =
    "Originality saved", type = "l")
lines(c(0, 1:70), tmin, lty = 2)
```



Document illustrant les liens entre décomposition de la diversité et ordination, avec des données génétiques.

Ce document utilise les fonctions ‘amova’, ‘randtest.amova’, ‘plot.randtest.amova’, ‘disc’, ‘divc’, ‘dpcoa’, ‘print.dpcoa’, ‘plot.dpcoa’ et le jeu de données ‘humDNAm’.

Décomposition de la diversité et représentations graphiques des différences
Exemple de données génétiques

Ce texte illustre l'existence de liens entre la mesure de la diversité par l'entropie quadratique, la mesure de la dissimilarité par le coefficient de Rao (différence de Jensen appliquée à l'entropie quadratique), l'inertie dans la double analyse en coordonnées principales, et la mesure de la variation dans l'AMOVA. Il permet de se familiariser avec les fonctions 'amova', 'randtest.amova', 'plot.randtest.amova', 'disc', 'divc', 'dpcoa', 'print.dpcoa', 'plot.dpcoa' à l'aide du jeu de données 'humDNAm'.

1. Les données

Ces données proviennent de l'article :
Excoffier, L., P. E. Smouse, and J. M. Quattro. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics **131**:479-491.

Des haplotypes d'ADNmt humain ont été échantillonnés dans dix populations représentant cinq groupes régionaux de deux populations chacun : "Asia" (populations "Tharu" et "Oriental"), "West Africa" (populations "Wolof" et "Peul"), "America" (populations "Pima" et "Maya"), "Europe" (populations "Finnish" et "Sicilian"), et "Middle-East" (populations "Israeli Jews" et "Israeli Arabs"). Les dissimilarités entre haplotypes sont calculées en nombres de sites de restriction différents entre séquences. Excoffier *et al.* proposent une représentation des liens entre haplotypes sous la forme d'un réseau de longueur minimale (Fig. 1).

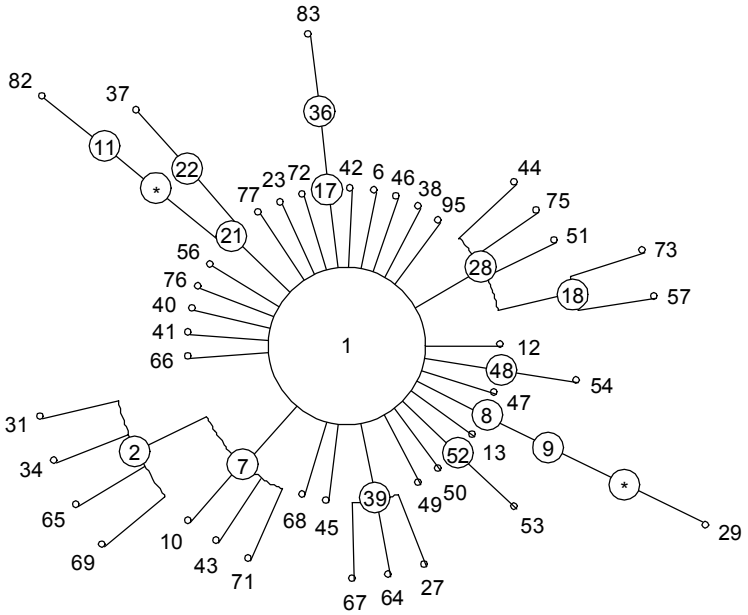


Fig 1: Réseau de longueur minimale (minimum spanning network) des 56 haplotypes trouvés dans les 10 populations. Chaque lien représente une mutation. Les astérisques indiquent deux haplotypes non trouvés dans les populations échantillonnées. L'haplotype 1 central a été trouvé dans toutes les populations. Les feuilles (haplotypes situés en extrémité du réseau) sont indiquées par des petits ronds, les noeuds intermédiaires par des ronds moyens et l'haplotype central par un grand rond.

Trois questions peuvent être posées : (1) Existe-t-il des différences entre populations ou entre groupes ? (2) Quels sont les principaux haplotypes impliqués dans ces différences ? (3) Quelle est la typologie des différences entre populations et entre goupes, en particulier quelles sont les populations et groupes qui se distinguent le plus des autres ? L'AMOVA permet de répondre à la première question. La dPCoA va permettre de répondre aux deux autres questions. Les calculs qui suivent ont été réalisés dans R avec les fonctions 'amova', 'randtest.amova', 'plot.randtest.amova', 'dpcoa' et 'plot.dpcoa' du package 'ade4.

```
data(humDNAm)
names(humDNAm)
[1] "distances"  "samples"    "structures"
sapply(humDNAm, class)
   distances      samples   structures
      "dist" "data.frame" "data.frame"
```

'distances' est un objet de classe 'dist', et 'samples' et 'stuctures' sont des tableaux.

```
summary(humDNAm$distances)
Class: dist
Distance matrix by lower triangle : d21, d22, ..., d2n, d32, ...
Size: 56
Labels: 1 2 6 7 8 9 10 11 12 13 17 18 21 22 23 27 28 29 31 34 36 37 38 39 40 41
42 43 44 45 46 47 48 49 50 51 52 53 54 56 57 64 65 66 67 68 69 71 72 73 75 76 77
82 83 95
call: as.dist(m = distance)
method:
Euclidean matrix (Gower 1966): FALSE
```

'distances' contient les dissimilarités entre 56 haplotypes.
Les noms de ces haplotypes sont donnés par

```
attributes(humDNAm$distances)$Labels
[1] "1"  "2"  "6"  "7"  "8"  "9"  "10" "11" "12" "13" "17" "18" "21"
[14] "22" "23" "27" "28" "29" "31" "34" "36" "37" "38" "39" "40" "41"
[27] "42" "43" "44" "45" "46" "47" "48" "49" "50" "51" "52" "53" "54"
[40] "56" "57" "64" "65" "66" "67" "68" "69" "71" "72" "73" "75" "76"
[53] "77" "82" "83" "95"
```

Les dissimilarités ont été calculées par Excoffier *et al*. à partir de profils de restriction. La dissimilarité entre deux haplotypes est le nombre de différences entre leurs profils.

```
humDNAm$samples[1:5, ]
  oriental tharu wolof peul pima maya finnish sicilian israelij israelia
1       32    48    23   11   59   30      87       50       15       22
2        0     0    39   19    0    0       0        3        0        1
6        1     0     0    0    2    0       2        9       14        1
7        0     0    29   12    0    0       0        0        0        1
8        2     2     0    2    0    0       0        0        0        0
```

Le tableau 'samples' fournit les abondances de chaque haplotype dans chaque population.

```
humDNAm$structures
      regions
1        asia
2        asia
3      africa
4      africa
5     america
6     america
```

```
7       europe
8       europe
9  middleeast
10 middleeast
```
Le tableau 'structure' décrit la répartition des populations en groupes.

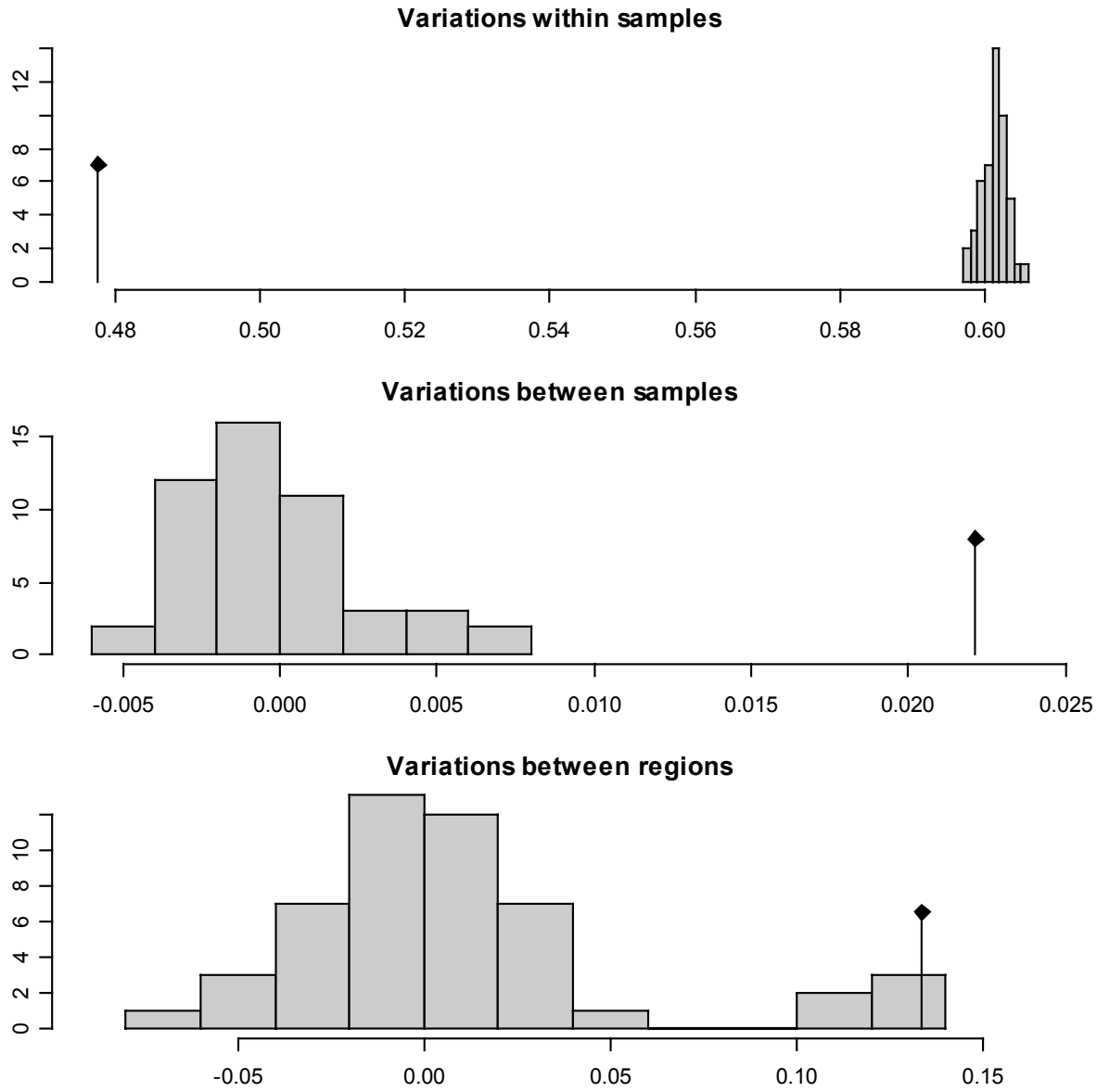
2. AMOVA

Excoffier *et al*. analysent ces données par l'AMOVA, analyse de variance moléculaire :

Pour éviter l'apparition de composants négatifs, la fonction 'amova' ne permet l'accès qu'à des matrices de dissimilarités euclidiennes.

```
is.euclid(humDNAm$distances)
[1] FALSE
is.euclid(sqrt(humDNAm$distances))
[1] TRUE
```
Cette transformation (racine) a également été faite par Excoffier *et al*.

```
amovahum <- amova(humDNAm$samples, sqrt(humDNAm$distances),
    humDNAm$structures)
amovahum
$call
amova(samples = humDNAm$samples, distances = sqrt(humDNAm$distances),
    structures = humDNAm$structures)

$results
                                Df  Sum Sq Mean Sq
Between regions                  4  78.238 19.5595
Between samples Within regions   5   9.285  1.8569
Within samples                 662 316.197  0.4776
Total                          671 403.720  0.6017

$componentsofcovariance
                                                 Sigma       %
Variations  Between regions                    0.13381  21.119
Variations  Between samples Within regions     0.02213   3.493
Variations  Within samples                     0.47764  75.387
Total variations                               0.63358 100.000

$statphi
                        Phi
Phi-samples-total   0.24613
Phi-samples-regions 0.04429
Phi-regions-total   0.21119

randtesthum <- randtest.amova(amovahum, 49)
randtesthum
class: krandtest
test number:   3
permutation number:   49
  test                        obs  P(X<=obs) P(X>=obs)
1 Variations within samples  0.48 0.02      1
2 Variations between samples 0.02 1         0.02
3 Variations between regions 0.13 1         0.02

plot(randtesthum)
```



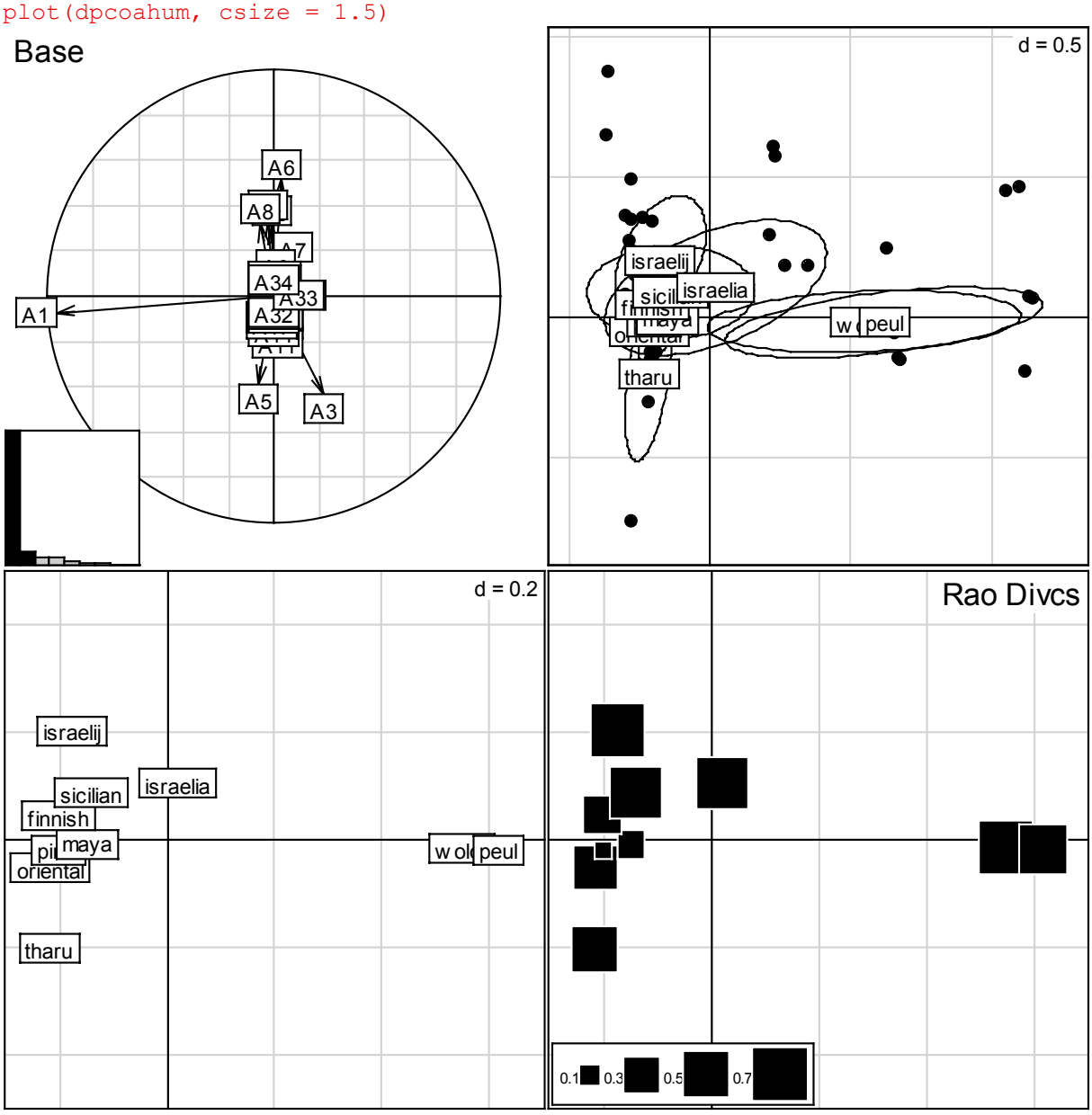
Il existe des différences significatives entre populations au sein des groupes et également entre groupes. Peut-on décrire ces différences ?

3. Double analyse en coordonnées principales

```
dpcoahum <- dpcoa(humDNAm$samples, sqrt(humDNAm$distances), scan = FALSE,
    nf = 2)
dpcoahum
double principal coordinate analysis
class: dpcoa
$call: dpcoa(df = humDNAm$samples, dis = sqrt(humDNAm$distances), scannf =
FALSE,
    nf = 2)

$nf: 2 axis-components saved
eigen values: 0.1018 0.01035 0.006281 0.005602 0.003179 ...
  vector  length mode    content
1 $w1     56     numeric weights of species
```

```
2 $w2     10     numeric weights of communities
3 $eig    9      numeric eigen values
4 $RaoDiv 10     numeric diversity coefficients within communities

  dist    Size content
1 $RaoDis 10   dissimilarities among communities

  data.frame  nrow ncol content
1 $RaoDecodiv 3    1    decomposition of diversity
2 $l1         56   2    coordinates of the species
3 $l2         10   2    coordinates of the species
4 $c1         34   2    scores of the principal axes of the species
```

```
plot(dpcoahum, csize = 1.5)
```



Fonctionnement de la double PCoA :

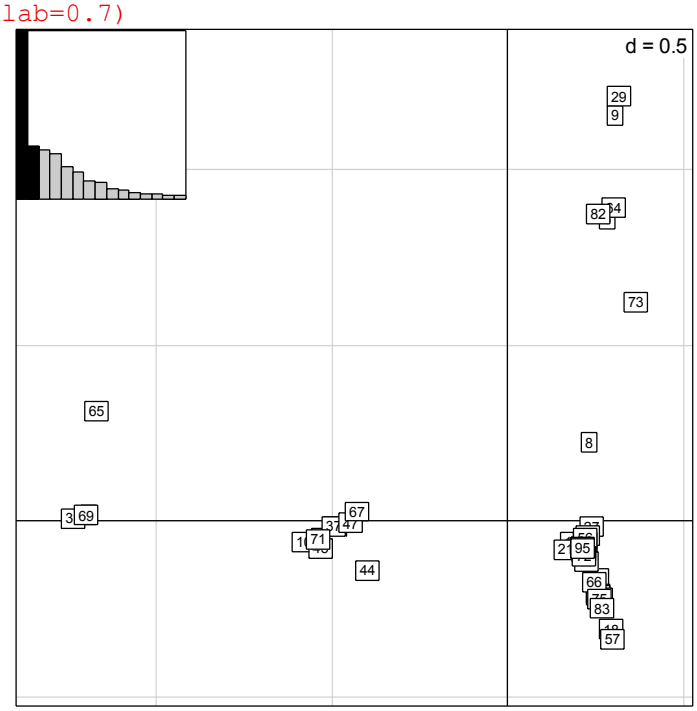
1ère étape : Représenter les dissimilarités entre haplotypes, utilisation d'une analyse en coordonnées principales pondérée :

```
args(dudi.pco)
```

```
function (d, row.w = "uniform", scannf = TRUE, nf = 2, full = FALSE,
    tol = 1e-07)
pcowHAP <- dudi.pco(sqrt(humDNAm$distances),
    row.w = apply(humDNAm$samples, 1, sum)/sum(humDNAm$samples),
    scan=FALSE, full = TRUE)
```

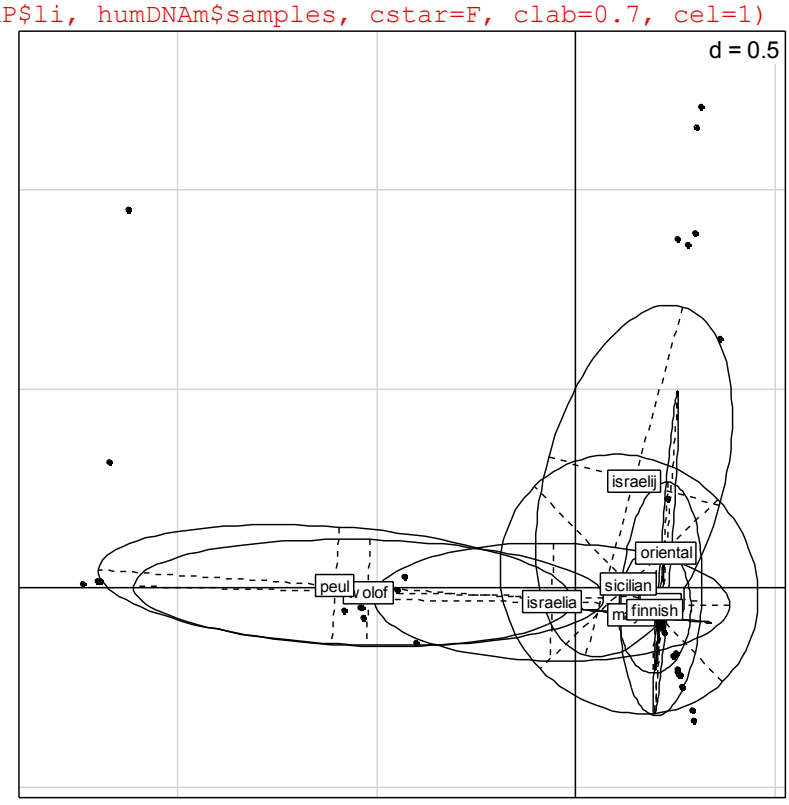
L'argument `'full = TRUE'` indique que nous gardons tous les axes.

```
scatter(pcowHAP, clab=0.7)
```



2ème étape : Chaque population est placée au barycentre de ses haplotypes.

```
s.distri(pcowHAP$li, humDNAm$samples, cstar=F, clab=0.7, cel=1)
```



Nous sommes dans l'espace des haplotypes. Pour représenter au mieux les différences entre populations, il faut chercher les axes principaux des populations.

3ème étape, recherche des axes principaux des points populations :

Poids des populations :

```
popw <- apply(humDNAm$samples, 2, sum) / sum(humDNAm$samples)
```

Coordonnées des populations dans l'espace des haplotypes :

```
Ypop <- diag(popw^{-1}) %*% t(humDNAm$samples / sum(humDNAm$samples)) %*%
    as.matrix(pcowHAP$li)

rownames(Ypop) <- names(humDNAm$samples)

rownames(Ypop)
 [1] "oriental" "tharu"    "wolof"    "peul"     "pima"     "maya"
 [7] "finnish"  "sicilian" "israelij" "israelia"

args(dudi.pca)
function (df, row.w = rep(1, nrow(df))/nrow(df), col.w = rep(1,
    ncol(df)), center = TRUE, scale = TRUE, scannf = TRUE, nf = 2)

pcawPOP <- dudi.pca(as.data.frame(Ypop), row.w = popw, scale = FALSE)
Select the number of axes: 2
```

Les coordonnées des deux premiers axes principaux des populations dans l'espace des haplotypes sont :

```
pcawPOP$c1[1:2, ]
           CS1         CS2
A1 -0.96513188 -0.07215368
A2 -0.01061850  0.31176536
```

4ème étape : projection des populations et des haplotypes sur ces axes.

```
pcawPOP$li
               Axis1       Axis2
oriental -0.21682997 -0.05214105
tharu    -0.21887770 -0.20389638
wolof     0.54767115 -0.01483056
peul      0.61771052 -0.01845884
pima     -0.20184976 -0.01962178
maya     -0.15006730 -0.00924097
finnish  -0.20348564  0.04554415
sicilian -0.14037075  0.08858604
israelij -0.17583049  0.20332671
israelia  0.01946232  0.10558291

dpcoahum$l2
               Axis1       Axis2
oriental -0.21682997 -0.05214105
tharu    -0.21887770 -0.20389638
wolof     0.54767115 -0.01483056
peul      0.61771052 -0.01845884
pima     -0.20184976 -0.01962178
maya     -0.15006730 -0.00924097
finnish  -0.20348564  0.04554415
sicilian -0.14037075  0.08858604
israelij -0.17583049  0.20332671
```

```
israelia  0.01946232  0.10558291
```

Les coordonnées précédentes (pcawPOP$li) provenaient de la projection des populations sur ces axes.

Les haplotypes sont également projetés dans l'espace des populations :

```
# Coordonnées des six premiers haplotypes
(as.matrix(pcowHAP$li)%*%as.matrix(pcawPOP$c1))[1:6, ]
```

```
           CS1         CS2
  1 -0.2001648 -0.03254234
  2  1.1359227  0.07176296
  6 -0.2832474  0.34741764
  7  0.6675842 -0.14676321
  8 -0.2186866 -0.29758432
  9 -0.3017693  0.08237566
```

```
dpcoahum$l1[1:6, ]
         CS1         CS2
1 -0.2001648 -0.03254234
2  1.1359227  0.07176296
6 -0.2832474  0.34741764
7  0.6675842 -0.14676321
8 -0.2186866 -0.29758432
9 -0.3017693  0.08237566
```

Les axes principaux des haplotypes peuvent également être projetés dans l'espace des populations afin d'évaluer la qualité de la représentation des différences entre haplotypes dans le nouvel espace.

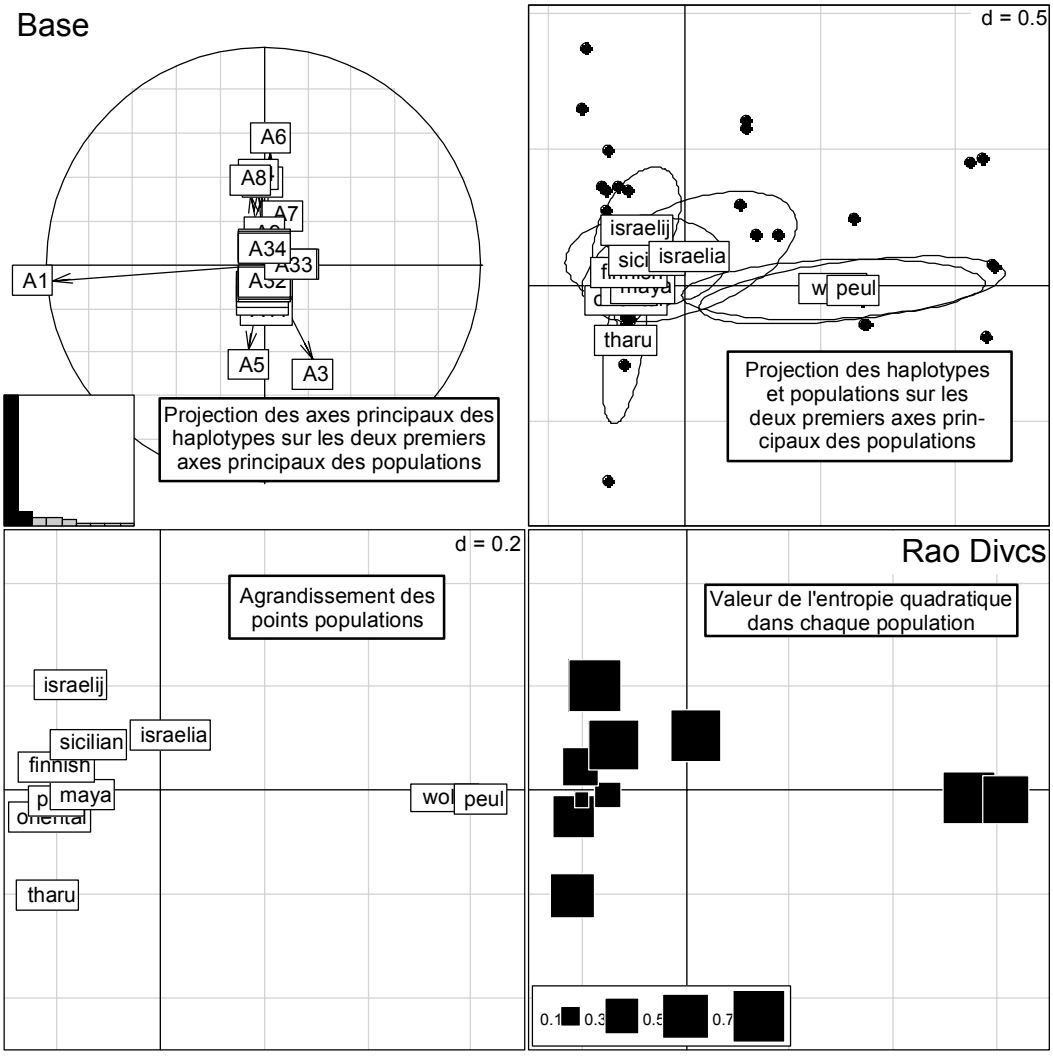
```
# Projection des deux premiers axes :
(as.matrix(pcowHAP$c1)%*%as.matrix(pcawPOP$c1))[1:2, ]
```

```
             CS1         CS2
  A1 -0.96513188 -0.07215368
  A2 -0.01061850  0.31176536
```

```
dpcoahum$c1[1:2, ]
           CS1         CS2
A1 -0.96513188 -0.07215368
A2 -0.01061850  0.31176536
```

Il s'agit des lignes de la matrice ' dpcoahum$c1' . 'dpcoahum$c1' est une "matrice de passage", ses colonnes donnent les coordonnées des nouveaux axes (axes principaux des populations) dans l'espace défini par les anciens (axes principaux des haplotypes). Inversement, ces lignes donnent les coordonnées des anciens axes dans l'espace défini par les nouveaux axes.

Ainsi les résultats fournis par la double analyse en coordonnées principales sont



4. Lien avec l'entropie quadratique

Distances entre populations dans l'espace euclidien déterminé par la dPCoA.

```
dist(Ypop)
          oriental     tharu     wolof      peul      pima      maya   finnish
tharu    0.2520200
wolof    0.7821356 0.7965116
peul     0.8506600 0.8632233 0.1251814
pima     0.1994968 0.2755218 0.7582823 0.8338317
maya     0.2559697 0.3023221 0.7177691 0.7857762 0.1406173
finnish  0.2455583 0.3043545 0.7615230 0.8367302 0.1416780 0.1899101
sicilian 0.2966927 0.3553515 0.7194281 0.7871977 0.2731589 0.2783955 0.2593187
israelij 0.3578283 0.4564520 0.7963554 0.8604647 0.3906841 0.4075339 0.3601095
israelia 0.4102655 0.4363589 0.5822519 0.6437361 0.3588124 0.3027937 0.2984287
          sicilian  israelij
tharu
wolof
peul
pima
maya
finnish
sicilian
israelij 0.3690240
```

```
israelia 0.3541134 0.3947386
```

Dissimilarités entre populations selon la formule de Rao :

```
disc(humDNAm$samples, sqrt(humDNAm$distances))
          oriental     tharu     wolof      peul      pima      maya   finnish
tharu    0.2520200
wolof    0.7821356 0.7965116
peul     0.8506600 0.8632233 0.1251814
pima     0.1994968 0.2755218 0.7582823 0.8338317
maya     0.2559697 0.3023221 0.7177691 0.7857762 0.1406173
finnish  0.2455583 0.3043545 0.7615230 0.8367302 0.1416780 0.1899101
sicilian 0.2966927 0.3553515 0.7194281 0.7871977 0.2731589 0.2783955 0.2593187
israelij 0.3578283 0.4564520 0.7963554 0.8604647 0.3906841 0.4075339 0.3601095
israelia 0.4102655 0.4363589 0.5822519 0.6437361 0.3588124 0.3027937 0.2984287
          sicilian  israelij
tharu
wolof
peul
pima
maya
finnish
sicilian
israelij 0.3690240
israelia 0.3541134 0.3947386
```

Diversité inter-populations :

```
divc(as.data.frame(popw), disc(humDNAm$samples, sqrt(humDNAm$distances))
     diversity
popw 0.1302423

sum(dpcoahum$eig)
[1] 0.1302423
```

→ L'inertie totale de la dPCoA est la diversité inter-populations.

```
amovahum$results
                                  Df     Sum Sq    Mean Sq
Between regions                    4  78.238115 19.5595288
Between samples Within regions     5   9.284744  1.8569488
Within samples                   662 316.197379  0.4776395
Total                            671 403.720238  0.6016695
sum(amovahum$results$"Sum Sq"[1:2])/672
[1] 0.1302423
```

Diversité dans chaque population :

```
divc(humDNAm$samples, sqrt(humDNAm$distances))
          diversity
oriental 0.43100189
tharu    0.48255042
wolof    0.65884298
peul     0.55952920
pima     0.06198035
maya     0.17092768
finnish  0.33280992
sicilian 0.61913580
israelij 0.67061144
israelia 0.64036818
```

Moyenne des diversités intra-populations :

```
sum(divc(humDNAm$samples, sqrt(humDNAm$distances))*popw)
[1] 0.4705318

(amovahum$results$"Sum Sq"[3])/672
[1] 0.4705318
```

Diversité totale :

```
divc(as.data.frame(apply(humDNAm$samples, 1, sum)), sqrt(humDNAm$distances))
                                diversity
apply(humDNAm$samples, 1, sum) 0.6007742

(amovahum$results$"Sum Sq"[4])/672
[1] 0.6007742
```

Ainsi la décomposition de la variation utilisée dans l'AMOVA est celle de l'entropie quadratique :

```
amovahum$results[2]/672
                                        Sum.Sq
Between regions                      0.11642577
Between samples Within regions       0.01381658
Within samples                       0.47053181
Total                                0.60077416
```

Nous avons donc relié la mesure de la diversité par l'entropie quadratique, la mesure de la dissimilarité par le coefficient de Rao (différence de Jensen appliquée à l'entropie quadratique), l'inertie dans la double analyse en coordonnées principales, et la mesure de la variation dans l'analyse moléculaire de variance.

RESUME

Face à l'accumulation des indices développés pour mesurer la biodiversité, la détermination de schémas fondamentaux est devenue nécessaire. Cette thèse démontre que : 1) l'axiomatisation de Rao constitue un schéma statistique pour l'analyse de la variation, en particulier variance et diversité; 2) au cœur de ce schéma, un indice, l'entropie quadratique, basé sur une matrice de dissimilarités est défini sur l'ensemble des distributions de fréquences; 3) la décomposition de cet indice généralise des méthodes utilisées pour l'analyse de la variation en statistique (ANOVA), génétique (AMOVA) et écologie, et est égale à la décomposition de l'inertie d'un nuage de points dans un espace euclidien déterminé; 4) l'entropie quadratique appliquée à des dissimilarités ultramétriques présente trois propriétés qui sont fondamentales pour un indice de biodiversité. Cette thèse analyse l'unité de ce schéma qui réunit les concepts de diversité, inertie, dissimilarité, ordination et originalité.

TITLE
Statistical methods for the measurement of biodiversity

ABSTRACT

Given the accumulation of indices developed for measuring biodiversity, the determination of fundamental patterns has become necessary. This thesis demonstrates that: 1) Rao's axiomatization constitutes a statistical framework for the analysis of variation, especially variance and diversity; 2) At the heart of this framework, an index, the quadratic entropy, which is based on a matrix of dissimilarities, is defined on the set of frequency distributions; 3) the decomposition of this index generalizes methods used to analyze variation in statistics (ANOVA), genetics (AMOVA) and ecology, and it is equal to the decomposition of the inertia of a cloud of points in a specified Euclidean space; 4) the quadratic entropy applied to ultrametric matrices has three properties which are fundamental for an index of biodiversity. This thesis analyzes the unity of this framework, which assembles the concepts of diversity, inertia, dissimilarity, ordination and originality.

DISCIPLINE
Ecologie statistique

MOTS-CLES

Analyse multivariée, ANOVA, CATANOVA, entropie quadratique, espèce rare, indice de dissimilarité, indice de diversité, ordination, test d'hypothèses, statistique

INTITULE ET ADRESSE DU LABORATOIRE :

Laboratoire de Biométrie et Biologie Evolutive (UMR 5558); CNRS; Univ. Lyon 1, 43 bd 11 nov, 69622, Villeubanne Cedex, France.