N d'ordre 207-2008 Année 2008

Thèse

Présentée
devant l'Université Claude Bernard - Lyon 1

pour l'obtention
du Diplôme de Doctorat
(arrêté du 7 août 2006)

et soutenue publiquement le
14 novembre 2008

par

Thibaut Jombart

Analyses multivariées de marqueurs génétiques : développements méthodologiques, applications et extensions.

Directrice de thèse : Anne-Béatrice Dufour
Co-directrice : Dominique Pontier

Jury :	Anne-Béatrice Dufour,	Directrice
	Dominique Pontier,	Co-directrice
	Nigel Gilles Yoccoz,	Rapporteur
	François Bonhomme,	Rapporteur
	Benjamin Bolker,	Examinateur
	Sylvain Dolédec,	Examinateur

Remerciements

A l'instant où cette thèse s'achève, il est difficile de recenser l'ensemble des personnes qui devraient y être remerciées, que ce soit pour leur contribution scientifique ou humaine. J'entame donc ce délicat recensement, en adressant par avance mes excuses pour les quelques oublis éventuels.

Puisqu'il s'agit d'une thèse de science, mes premiers remerciements vont aux personnes qui ont contribué scientifiquement à mon travail. A ce titre, je suis plus que redevable à Daniel Chessel, qui m'a encadré en Master 2 Recherche ainsi qu'en première année de thèse, et m'a énormément apporté. Plus que de m'avoir transmis un savoir, il m'a communiqué une passion — celle de l'analyse de données — et m'a donné les armes pour explorer mon sujet. S'il m'arrive un jour d'encadrer un étudiant, je m'estimerai heureux de lui apporter le dixième de ce que Daniel a pu m'apporter. Je tiens également à remercier mes encadrantes, Anne-Béatrice Dufour et Dominique Pontier, pour m'avoir accordé leur attention et leur soutien pendant ces trois ans sans pour autant restreindre une liberté qui m'a été précieuse. J'adresse également un grand "merci" à Stéphane Dray, Sébastien Devillard, Denis Laloë, Katayoun Moazami-Goudarzi et Sandrine Pavoine, avec qui j'ai eu le plaisir de collaborer. En particulier, merci à Stéphane pour m'avoir fait participer au SEDAR, et à Sébastien et Denis pour avoir servi de beta-testeurs à adegenet. Merci également à Hilmar Lapp et aux organisateurs et aux participants du 'R Hackathon on Comparative Methods', avec lesquels ce fut un réel plaisir de travailler. Merci à Emmanuel Paradis et à Jérome Sueur pour le séjour de formation 'APE' à Graz : j'en garde un excellent souvenir. Merci à Stéphanie Manel et Marie-José Fortin pour le temps passé ensemble à Grenoble. Merci aux personnes qui se sont intéressées à mon travail et m'ont souvent prodigué de précieux conseils, en particulier Christian Biémont, Steve Dobson et Gilles Yoccoz. Pour la même raison, je tiens à remercier les rapporteurs et les examinateurs de ma thèse. Merci enfin à celles qui ont facilité mon quotidien administratif : Misou (avant sa retraite), Nathalie Arbasetti, Isabelle Ravis et Agnès Python.

Il me reste à remercier toutes les personnes qui ont fait de cette thèse trois ans de plaisir, de bonheur, et parfois aussi de mal de cheveux. Merci à Simon et Sophie (in grind we trust !), à Elise et Eric (130-carreau), à Liliana et Andrès (on se retrouve en Colombie), à Vincent (Lombard) pour s'être laissé entraîné vers le free-fight, à Ludo et Émilie (et Jeanne), à Claire pour sa gentillesse proverbiale (vive la piscine), à Manu et Élo (et Lise), à Patricia pour sa bonne-humeur, à Vicente pour nos discussions math tardives, à Marilia (viva footcheubol), à Marie-France et au reste des Baobab ; merci à Vincent (Daubin) et Christine pour avoir été nos compagnons gastronomes ; merci à Vincent (Navrat) et Sandy (vive les mariés !), merci à Stéphane et Fredo (et leurs enfants), à Lionel et Mado (idem), et à Bruno pour avoir supporté mes intrusions intempestives au service informatique (ce n'est pas moi !) ; merci à Louise et Deborah pour l'espagnol ; merci à Nathalie pour sa patience. Merci aux tourangeaux : Soli (hacker), Joce (amen) et Sim (je n'ai rien contre les polaks) ; Philippe (Philiiiippe), François, Mickael et Min-Jae (mort aux druides). Merci à ceux qui ont su m'inspirer dans mon travail de

tous les jours : Mark Greenway, Julien Truchan, Dave Hunt, Devin Townsend, Adam Darski, Christof Dekronos, Franz Schuller, Vince Peake. Merci à mes frères Arnaud et Vincent, et à ma mère pour m'avoir appris à travailler pour moi-même et permis d'arriver jusqu'ici. Et merci à Perrine, pour tout.

Table des matières

Table des figures

Avant-propos

Quelques précisions méritent d'être apportées quant au contexte dans lequel cette thèse s'est déroulée, et plus précisément quant à son encadrement, puisque cet aspect revêt peut-être une certaine originalité. Bien qu'Anne-Béatrice Dufour et Dominique Pontier soient à juste titre respectivement directrice et co-directrice de ce travail, il me faut souligner que c'est à l'origine Daniel Chessel qui a dirigé cette thèse pendant la première année, jusqu'à son départ à la retraite. Il est difficile de quantifier son influence sur les travaux qui sont ici présentés, mais elle a été déterminante, et il m'a souvent été impossible d'en faire état en citant des travaux précédents auxquels il a participé. Il est à l'origine de la collaboration avec deux chercheurs de l'INRA de Jouy-en-Josas dont les fruits sont présentés au chapitre 2 de cette thèse. Il est également l'auteur des fondements mathématiques de l'analyse en composantes principales spatiales, présentée au chapitre 3, et en a suggéré l'extension à l'investigation de structures phylogénétiques (chapitre 6). Plus largement, Daniel Chessel a posé en bonne partie les bases de la démarche scientifique que j'ai adoptée, et qui a façonné de nombreux aspects de ce travail de thèse.

Forewords

About the supervision of this PhD thesis

Some light deserve to be cast onto the context in which this PhD thesis took place, and more precisely onto the supervision of the thesis. Despite Anne-Béatrice Dufour and Dominique Pontier supervised most of the work presented in this manuscript, I have to underline that Daniel Chessel originally supervised the first year of my PhD thesis, until his retirement. It is somewhat difficult to quantify the influence he had on my work, but his role has been central. Daniel Chessel originated the collaboration with two researchers from the INRA of Jouy-en-Josas, which eventually led to the results presented in chapter 2. He was also the author of the mathematical foundations of the spatial principal component analysis, presented in chapter 3, and he also suggested the extension of this approach to the investigation of phylogenetic structures (chapter 6). More generally, Daniel Chessel set up the basis of the scientific approach I adopted, which designed many aspects of this PhD thesis.

About this manuscript

This manuscript is written in French, but is based on seven articles written in English. Each chapter, apart from the conclusion, consists in an introduction, an article, and a discussion. If the essential of the work is thus accessible to non-French readers, it may be relevant to provide some insights about what makes the remaining of the manuscript.

The first chapter aims at providing an overview of the current use of multivariate analysis to extract biological information from genetic markers. After insisting on the fact that genetics has long been a field of statistical interest (as illustrated by Sir R.A. Fisher's carrier), we explain why multivariate analysis is a relevant tool to analyse molecular markers. Our approach of multivariate methods is based on the *duality diagram*, a framework which unifies most multivariate analyses as particular cases of a general algorithm. After presenting this framework, we focus on some biological questions that seemed of particular interest. First, when using multiallelic genetic markers to infer the biodiversity between a set of species, populations, or genotypes, it happens that different markers provide different, potentially inconsistent information. Tools are therefore needed to study the coherence of the information provided by a set of markers. Second, when trying to uncover the structure of biological populations from the analysis of genetic markers, it is frequent that spatial information about genotypes or populations

is known but not used, despite most genetic structures are also expected to be spatial structures. Hence, tools are needed to take both the genetic and the spatial information into account when analysing georeferenced genotypes or populations. After providing these mathematical and biological backgrounds, a review paper details current applications of multivariate analysis to genetic markers data.

The second chapter is devoted to the question of the coherence of the information provided by a set of multiallelic markers. Our approach consists in seeking similarities between the typologies provided by different markers, using K-table methods. This class of multivariate analyses includes different methods, three of which are presented in introduction. One of them, the multiple co-inertia analysis, is proposed for the analysis of genetic markers in the presented paper, introducing K-table methods in genetics. In the discussion, the three K-table methods presented previously are compared through the analysis of a microsatellite dataset.

In the third chapter, *spatial principal component analysis* (sPCA) is proposed as a new tool to investigate spatial patterns in the genetic variability. The introduction provides some insights about the origins of this methodology, which relies on a multivariate extension of Moran's index of spatial autocorrelation. After presenting the paper introducing the sPCA, the discussion first provides an overview of the implementation of the method, and then illustrates the sPCA by analysing a dataset of georeferenced genotypes of Chamois (*Rupicapra rupicapra*) in the Bauges mountains. We end the chapter by discussing some issues of the sPCA. The first issue relates to the asymmetry of the distribution of Moran's index, which can result in difficulties for identifying some kinds of spatial structures. The second issue concerns some artefactual patterns arising in non-structured datasets.

The fourth chapter focuses on practical aspects of the multivariate analysis of genetic markers within the R software. After exposing the main concerns and needs that existed at the beginning of this PhD thesis, we present a paper introducing the *adegenet* package, a R package implementing classes and methods to facilitate the analysis of molecular markers using multivariate tools. An example illustrating some functionalities of the package is also provided.

The remaining of the manuscript diverges gradually from the analysis of genetic markers. The fifth chapter tackles the question of scales of spatial patterns in biological data, a problematic which received a great interest in ecology, although it is also relevant in population genetics. We introduce the *multi-scale pattern analysis* (MSPA) as a new method to identify the main scales of spatial variation in a multivariate dataset, that can be composed of quantitative as well as of qualitative variables. In the discussion, we show that this method can provide insights about the scales at which a set of genotypes is spatially structured by analysing a Scandinavian Brown Bear dataset, a famous dataset which has already illustrated several methods in landscape genetics.

The sixth chapter consists in the extension of the sPCA to the analysis of phylogenetic patterns in a set of biological traits. As an introduction, we highlight similarities existing between some spatial and some phylogenetic methods, which all basically rely on a measure of autocorrelation. The *phylogenetic principal component analysis* (pPCA) is then presented as an extension of the sPCA based on the measurement of phylogenetic autocorrelation using Moran's index. In the discussion, the pPCA is illustrated using a dataset of life history traits in Lacertid lizards, and some issues relating to the measure of phylogenetic proximities are then

discussed.

In the seventh and last chapter, we show that the principle of sPCA can be generalized to investigate autocorrelated structures in quantitative as well as qualitative data. Autocorrelation is defined with respect to any measure of proximity between the objects under study ; it can thus be applied in various contexts, to measure the non-independence of observations in time, space, among the tips of phylogeny, or on the vertices of an interaction graph.

Notations mathématiques

- $\mathbb{R}$: l'ensemble des nombres réels.

$\mathbf{x} \in \mathbb{R}^n$: un vecteur de n nombres réels ; sauf spécification contraire, tout vecteur est considéré comme vecteur-colonne ; son terme général est noté x_i (on notera $\mathbf{x} = [x_i]$, $i = 1, \ldots, n$)

- $\mathbf{x}^T$: le transposé du vecteur $\mathbf{x}$, considéré comme vecteur-ligne si $\mathbf{x}$ est un vecteur colonne, et vice-versa.
- $\mathbf{X} \in \mathbb{R}^{n \times p}$: une matrice de nombres réels comportant n lignes et p colonnes ; son terme général est noté x_{ij} (on notera $\mathbf{X} = [x_{ij}]$, $i = 1, \ldots, n$, $j = 1, \ldots, p$).
- $\mathbf{X}^j$: le vecteur correspondant à la $j$ème colonne de la matrice $\mathbf{X}$.
- $\mathbf{X}_{[i]}$: le vecteur correspondant à la $i$ème ligne de la matrice $\mathbf{X}$.
- $\mathbf{X}^T$: la transposée de la matrice $\mathbf{X}$, définie par $\mathbf{X}^T = [x_{ji}]$.
- $\mathbb{R}^{n*}$: espace dual de $\mathbb{R}^n$, c'est-à-dire l'ensemble des applications linéaires allant de $\mathbb{R}^n$ dans $\mathbb{R}$ (ou formes linéaires sur $\mathbb{R}^n$).
- $\langle \mathbf{x}, \mathbf{y} \rangle$: le produit scalaire canonique entre les vecteurs $\mathbf{x}$ et $\mathbf{y}$ ($\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \mathbf{y}$).
- $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{M}}$: le produit scalaire entre les vecteurs $\mathbf{x}$ et $\mathbf{y}$ au sens de la métrique $\mathbf{M}$ ($\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \mathbf{M} \mathbf{y}$).
- $\mathrm{tr}(\mathbf{A})$: la trace de la matrice carrée $\mathbf{A} \in \mathbb{R}^{n \times n}$, *i.e.* la somme de ses éléments diagonaux ($\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}$).
- $\langle \mathbf{X}, \mathbf{Y} \rangle$: le produit de Frobenius entre les matrices $\mathbf{X} \in \mathbb{R}^{n \times p}$ et $\mathbf{Y} \in \mathbb{R}^{n \times p}$ ($\langle \mathbf{X}, \mathbf{Y} \rangle = \mathrm{tr}(\mathbf{X}^T \mathbf{Y})$).
- $\langle \mathbf{X}, \mathbf{Y} \rangle_{\mathbf{M}}$: le produit de Frobenius entre les matrices $\mathbf{X}$ et $\mathbf{Y}$ au sens de la métrique $\mathbf{M}$ ($\langle \mathbf{X}, \mathbf{Y} \rangle_{\mathbf{M}} = \mathrm{tr}(\mathbf{X}^T \mathbf{M} \mathbf{Y})$).
- $\|\mathbf{X}\|$: la norme de Frobenius de la matrice $\mathbf{X}$ ($\|\mathbf{X}\| = \langle \mathbf{X}, \mathbf{X} \rangle^{1/2}$).
- $\|\mathbf{X}\|_{\mathbf{M}}$: la norme de Frobenius de la matrice $\mathbf{X}$ au sens de la métrique $\mathbf{M}$ ($\|\mathbf{X}\|_{\mathbf{M}} = \langle \mathbf{X}, \mathbf{X} \rangle_{\mathbf{M}}^{1/2}$).
- $\mathbf{I}_n$: la matrice identité de dimension n, de terme général $\mathbf{I}_n = [\delta_{ij}]$ où δ_{ij} est le symbole de Kronecker ($\delta_{ij} = 1$ si $i = j$, et $\delta_{ij} = 0$ si $i \neq j$).
- $\mathbf{1}_n$: le vecteur de $\mathbb{R}^n$ dont toutes les composantes valent 1.
- ${\mathbf{1}_n}^{\perp}$: le sous-espace vectoriel orthogonal au vecteur $\mathbf{1}_n$.
- $\mathbf{X} \bullet \mathbf{Y}$: le produit d'Hadamard entre les matrices $\mathbf{X} \in \mathbb{R}^{n \times p}$ et $\mathbf{Y} \in \mathbb{R}^{n \times p}$ (soit $\mathbf{Z} = \mathbf{X} \bullet \mathbf{Y}$, on a $z_{ij} = x_{ij} y_{ij}$).

Mathematical notations

- $\mathbb{R}$: the set of real numbers.

$\mathbf{x} \in \mathbb{R}^n$: a vector of n real numbers ; unless otherwise specified, vectors are taken as column vectors ; the general term of $\mathbf{x}$ is denoted x_i (so, $\mathbf{x} = [x_i]$, $i = 1, \ldots, n$)

- $\mathbf{x}^T$: the transposed of $\mathbf{x}$, considered as a row vector if $\mathbf{x}$ is a column vector, and conversely.
- $\mathbf{X} \in \mathbb{R}^{n \times p}$: a matrix of real numbers with n rows and p columns ; its general term is denoted x_{ij} (so, $\mathbf{X} = [x_{ij}]$, $i = 1, \ldots, n$, $j = 1, \ldots, p$).
- $\mathbf{X}^j$: the vector corresponding to the jth column of matrix $\mathbf{X}$.
- $\mathbf{X}_{[i]}$: the vector corresponding to the ith row of matrix $\mathbf{X}$.
- $\mathbf{X}^T$: the transposed matrix of $\mathbf{X}$, defined as $\mathbf{X}^T = [x_{ji}]$.
- $\mathbb{R}^{n*}$: the dual space of $\mathbb{R}^n$, *i.e.* the set of linear maps from $\mathbb{R}^n$ to $\mathbb{R}$ (or linear forms on $\mathbb{R}^n$).
- $\langle \mathbf{x}, \mathbf{y} \rangle$: the canonical dot product between vectors $\mathbf{x}$ and $\mathbf{y}$ ($\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \mathbf{y}$).
- $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{M}}$: the dot product between vectors $\mathbf{x}$ and $\mathbf{y}$ computed with the metric $\mathbf{M}$ ($\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \mathbf{M} \mathbf{y}$).
- $\mathrm{tr}(\mathbf{A})$: the trace of the square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, *i.e.* the sum of the diagonal terms ($\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}$).
- $\langle \mathbf{X}, \mathbf{Y} \rangle$: the Frobenius product between matrices $\mathbf{X} \in \mathbb{R}^{n \times p}$ and $\mathbf{Y} \in \mathbb{R}^{n \times p}$ ($\langle \mathbf{X}, \mathbf{Y} \rangle = \mathrm{tr}(\mathbf{X}^T \mathbf{Y})$).
- $\langle \mathbf{X}, \mathbf{Y} \rangle_{\mathbf{M}}$: the Frobenius product between matrices $\mathbf{X}$ and $\mathbf{Y}$ computed with the metric $\mathbf{M}$ ($\langle \mathbf{X}, \mathbf{Y} \rangle_{\mathbf{M}} = \mathrm{tr}(\mathbf{X}^T \mathbf{M} \mathbf{Y})$).
- $\|\mathbf{X}\|$: the Frobenius norm of matrix $\mathbf{X}$ ($\|\mathbf{X}\| = \langle \mathbf{X}, \mathbf{X} \rangle^{1/2}$).
- $\|\mathbf{X}\|_{\mathbf{M}}$: the Frobenius norm of matrix $\mathbf{X}$ computed with the metric $\mathbf{M}$ ($\|\mathbf{X}\| = \langle \mathbf{X}, \mathbf{X} \rangle_{\mathbf{M}}^{1/2}$).
- $\mathbf{I}_n$: the identity matrix of dimension n, with general term $\mathbf{I}_n = [\delta_{ij}]$ where δ_{ij} is the Kronecker symbol ($\delta_{ij} = 1$ if $i = j$, and $\delta_{ij} = 0$ if $i \neq j$).
- $\mathbf{1}_n$: the vector in $\mathbb{R}^n$ whose components are all 1.
- ${\mathbf{1}_n}^{\perp}$: the vectorial sub-space orthogonal to the vector $\mathbf{1}_n$.
- $\mathbf{X} \bullet \mathbf{Y}$: the Hadamard product between matrices $\mathbf{X} \in \mathbb{R}^{n \times p}$ and $\mathbf{Y} \in \mathbb{R}^{n \times p}$ (denoting $\mathbf{Z} = \mathbf{X} \bullet \mathbf{Y}$, we have $z_{ij} = x_{ij} y_{ij}$).

Chapitre 1

L'analyse multivariée appliquée aux marqueurs génétiques

Sommaire

1.1 Introduction générale

Génétique et statistique ont été étroitement liées depuis le début du XXe siècle, et ont entretenu à l'origine des relations houleuses, si ce n'est conflictuelles. Joshi (1997) fait état de l'opposition qui a marqué les premiers rapports entre **biométrie** — alors sous l'égide de Karl Pearson — et **génétique mendélienne** — menée par William Bateson — sur la question centrale des mécanismes de l'évolution. Chaque protagoniste s'intéresse alors à la variabilité des traits observés dans les populations biologiques, et en particulier à la mesure de cette variabilité, d'où le terme de « biométrie », qu'on pourrait finalement attribuer aux deux écoles. Mais si les deux partis s'accordent alors sur le caractère passionnant du sujet d'étude, ils proposent deux interprétations opposées des mécanismes menant à cette variabilité : les biométriciens considèrent qu'un ensemble de facteurs discrets et héritables (les gènes) ne sauraient être le support de l'évolution de traits biologiques pour la plupart continus, alors que les mendéliens, partant du paradigme du gène, contestent la vision darwinienne d'une évolution continue. On peut peut-être voir dans cette opposition d'école une opposition sur le fond, comme le suggère Benzecri (1976b) dans son histoire de l'analyse des données [1] :

> *l'expérimentation mendélienne s'oppose au principe de la biométrie selon lequel la composition de la population (ou de l'échantillon qui la représente) est elle-même un phénomène naturel ; et de plus, digne d'être étudié* [car] *les échantillons non-naturels brouillent tout et ne découvrent rien.*

D'après Joshi (1997), la controverse a alors une influence considérable sur l'avancée de la biologie :

> *as long as the controversy was not resolved in some kind of grand synthesis, any significant advancement of our understanding of evolution, the most important of all biological phenomena, would not have been possible.*

Fort heureusement, cette synthèse fut réalisée par Fisher (1918), qui a réconcilié les deux points de vue en montrant qu'un ensemble de facteurs discrets et héritables combinés pouvaient engendrer des distributions continues de traits, réconciliant ainsi biométriciens et mendéliens (Piegorsch, 1990; Joshi, 1997). Notons que Fisher avait dès 1911 compris l'importance de cette réconciliation : Piegorsch (1990) rapporte que dans un discours donné à la société eugénique de l'université de Cambridge (où il n'était alors qu'étudiant) intitulé « Mendelism and Biometry », Fisher soutenait déjà la compatibilité des deux écoles de pensée. Outre cette réconciliation cruciale, la contribution scientifique majeure de Fisher (1918) pose les bases de la génétique quantitative, et introduit notamment le terme de **variance** [2], appelé à un avenir radieux en statistique :

> *It is therefore desirable in analysing the causes of variability to deal with the square of the standard deviation as the measure of variability. We shall term this quantity*

[1] On pourrait d'ores-et-déjà contester l'opposition entre biologie expérimentale et analyse de données : le fruit d'une expérience au sens premier, c'est-à-dire de la manipulation d'objets naturels, est encore un objet naturel. Le seul écueil à craindre est que les caractères que l'on observe sur cet objet soient des conséquences triviales du protocole expérimental. A la suite de Legay (1997, p.59), on serait d'ailleurs tenté de considérer que le recueil de données dans le cadre d'un plan expérimental, tel qu'il existe en écologie, et finalement aussi en génétique, est encore une expérience : *j'ai proposé d'appeler expérience "toute procédure organisée d'acquisition de l'information qui comporte, dans la perspective d'un objectif exprimé, une confrontation avec la réalité".*

> *the Variance of the normal population to which it refers, and we may now ascribe to the constituent causes fractions or percentages of the total variance which they together produce.*

Le fait que cette contribution scientifique majeure ait tout d'abord reçu un accueil défavorable peut surprendre, mais par le rejet initial de son manuscrit, Fisher avait déjà d'une certaine façon réuni — contre lui — deux éminents représentants des écoles ennemies : un biométricien (Pearson) et un généticien (Punnett) (Piegorsch, 1990).

L'opposition initiale entre biométrie et génétique était donc révolue, comme l'atteste l'immense apport de Fisher dans les deux domaines. En effet, si Fisher est connu comme un des pères de la statistique, il a également profondément marqué la génétique : comme Edwards (1990) et Thompson (1990) l'ont remarqué, Fisher, bien qu'ayant une formation de mathématicien, n'a jamais occupé que des postes de professeur en génétique. Thompson (1990) le place d'ailleurs au côté de Haldane et de Wright comme fondateur de la génétique des populations, et Piegorsch (1990) lui attribue environ 150 articles, livres et revues dans le domaine de la génétique. Sans rentrer plus avant dans la présentation d'une oeuvre qui nous dépasse (on renverra pour ce faire aux références citées par Piegorsch (1990)), nous nous contenterons de retenir qu'il s'agit là de la preuve par excellence que la génétique est un domaine éminemment biométrique.

C'est ici qu'il convient peut-être de préciser ce que l'on entend par « biométrie », en tant que domaine de recherche et en tant qu'activité. D'un point de vue étymologique, la biométrie peut être décrite comme la (science de la) mesure du vivant. La pratique de la biométrie est par contre plus difficile à cerner. On trouvera dans Chessel (1992) une discussion lumineuse de la nature de la recherche biométrique, que l'on réduira ici de manière incompressible à un **dialogue interdisciplinaire** entre biologie et statistique, dans lequel idéalement une relation symbiotique s'installe : la statistique trouve dans la biologie de nouvelles problématiques, et la biologie trouve dans la statistique des réponses aux questions posées via ou par les données. La carrière de Fisher ne laisse pas de doute sur l'existence effective d'un tel dialogue (Piegorsch, 1990) :

> [Fisher's] *advancements in the field of genetics and heredity are rivaled only by those in the field of biometry and statistics. The two efforts were often intricately related. Indeed, major advances in biometry are often catalyzed by subject-matter problems, and the study of genetics and heredity has provided great motivation for such advances*

Il est donc essentiel pour définir le travail de cette thèse en biométrie d'avoir conscience du dialogue interdisciplinaire dans lequel elle s'inscrit, et d'identifier les objets autour desquels la discussion s'organise, car *dans le dialogue qui est l'essence de l'analyse des données, le premier élément est formé par les données numériques* (Chessel, 1992, p.8).

Ces objets, Benzecri (1976b) pressent leur émergence et les besoins qu'ils créeront du point de vue méthodologique :

> *à la génétique mendélienne elle-même, l'analyse de données peut s'associer. D'une*

[2] En réalité, Fisher (1918) introduit également l'analyse de la variance (ANOVA).

> *part les caractères génétiques sont recensés sur des échantillons de population de toute provenance géographique ; d'où matière à des analyses révélant les parentés entre ces populations. D'autre part* [...] *le généticien doit collectionner des informations multidimensionnelles pour distinguer les locus et établir l'inventaire des allèles.*

En effet, s'il est vrai que Fisher a considérablement contribué à poser les bases de l'analyse statistique des données génétiques, ces données ont considérablement changé depuis ; en particulier, l'avènement relativement récent des marqueurs génétiques (Schlötterer, 2004) a permis d'aborder de nouvelles questions biologiques, mais a en retour interrogé la statistique. Ces objets sont par essence multivariés : un jeu de données type comprend aisément des centaines de génotypes et des centaines d'allèles. C'est donc logiquement que Benzecri (1976b) y voit un champ d'application de l'**analyse de données**, qui est décrite par G. Morlat dans la préface de l'ouvrage de Cailliez & Pages (1976) comme un champ de la biométrie *s'appuyant sur un outil mathématique purement algébrique, et visant à décrire, réduire, classer, des observations multidimensionnelles* [en considérant que] *le statisticien doit mettre en oeuvre ses techniques d'analyse sans faire aucune hypothèse sur les phénomènes observés.* La méthodologie mise en oeuvre, que l'on nommera indifféremment **analyse multivariée** ou **ordination en espace réduit**, visera donc à explorer les données de marqueurs moléculaires afin d'identifier des structures, c'est-à-dire des ressemblances ou des dissemblances entre génotypes, populations, allèles ou locus, selon l'objectif affiché. Cette démarche s'inscrit bel et bien dans la tradition biométrique initiée par Fisher : elle découle en fait directement de son école. En effet, le premier article faisant usage d'une analyse multivariée pour extraire de l'information biologique de marqueurs génétiques est dû à Cavalli-Sforza (1966), qui utilise une analyse en composantes principales (ACP, Pearson, 1901; Hotelling, 1933a,b) pour résumer les relations génétiques entre trente-cinq populations humaines typées pour un ensemble de marqueurs sérologiques. Mais il n'est pas surprenant que L.L. Cavalli-Sforza — à l'origine expert en génétique bactérienne — ait marqué les débuts de l'analyse multivariée de marqueurs génétiques, compte tenu de son séjour de 1948 à 1950 dans un laboratoire de biométrie de Cambridge ... où il fut encadré par Sir R.A. Fisher en personne (Edwards, 1990).

Cette thèse s'inscrit exactement dans la voie ouverte par Cavalli-Sforza, suivie par d'autres depuis, et dont nous essayerons de montrer la pertinence et la fécondité, tant pour la biologie que pour la biométrie. Nous tenterons d'évaluer la portée de l'analyse de données (au sens de Cailliez & Pages, 1976) pour extraire de l'information biologique des données de marqueurs génétiques, en faisant d'une part le point sur l'état actuel du dialogue instauré entre génétique et statistique, et en proposant d'autre part de nouveaux éléments méthodologiques pour l'enrichir. Certains de ces éléments ont également permis d'ouvrir de nouvelles voies, hors de l'analyse de données génétiques, qui constituent des extensions méthodologiques que l'on considérera comme faisant partie intégrante de cette thèse.

Chessel (1992, p.42) remarque qu'*on n'institutionnalise des échanges qu'en reconnaissant des frontières, ce qui permet d'en fréquenter le voisinage mais oblige à savoir où l'on est.* Avant de présenter les outils que nous avons développés pour l'analyse multivariée des marqueurs génétiques, il convient donc de clarifier les fondements mathématiques servant de support

au développement méthodologique, ainsi que les problématiques biologiques qui motivent ce développement. Lorsque ces « frontières » auront été identifées, nous présenterons un état des lieux du dialogue interdisciplinaire auquel nous participons, au travers d'une revue bibliographique de l'application de l'ordination en espace réduit aux données de marqueurs moléculaires.

1.2 Fondements mathématiques

1.2.1 L'analyse multivariée

L'analyse multivariée est un champ de la statistique visant à extraire une information humainement compréhensible à partir d'un grand ensemble de variables et d'observations. Par « humainement », on entend un individu averti sur le plan méthodologique comme sur le plan empirique ce qui, en biométrie, demande à l'analyste de maîtriser à la fois les propriétés mathématiques et le contexte biologique. Par « compréhensible », on signifie que l'information primairement multivariée, donc multidimensionnelle, a été réduite et simplifiée de façon à être perceptible par des « organismes » dont la perception est tridimensionnel. Enfin, il y a autant de façons de définir ce qu'on entend par « information » qu'il existe de points de vue différents d'un problème donné, ce qui justifie la multitude des approches existantes et constitue un point central de l'analyse des données.

Bien que les deux notions soient souvent confondues — y compris dans ce manuscrit — nous pouvons distinguer l'analyse multivariée au sens large, qui inclut toutes les méthodes traitant des jeux de données multivariés, et l'**ordination en espace réduit**, qui constitue un ensemble d'analyses multivariées visant à résumer l'information contenue dans les données en quelques variables de synthèse. Ces variables de synthèse sont en général obtenues en formant des combinaisons linéaires les données (variables ou observations) et choisies de façon à ce qu'elles représentent une part importante d'information (en acceptant encore une fois ce que ce terme recèle d'arbitraire) et à ce qu'elles soient non redondantes entre elles.

Sans pour autant retracer l'histoire de l'analyse multivariée (voir Benzecri, 1976a,b), on peut reconnaître plusieurs courants de pensée, plusieurs visions de ce domaine de la statistique, et nous situer par rapport à celles-ci. Au sein de ces différents points de vue, une distinction paraît essentielle entre ce que l'on pourrait nommer une **approche inférentielle** et une **approche descriptive**. Dans les deux cas, on considère un jeu de données constitué de plusieurs individus statistiques et plusieurs variables.

La première approche considère que les individus forment un échantillon d'une population (ce qui est presque toujours vrai), et cherche à estimer les relations entre variables au sein de la population. L'ouvrage de Anderson (1958) et, dans une moindre mesure, celui de Seal (1966) illustrent clairement cette démarche. Cette approche conduit à formuler des hypothèses sur la distribution des variables : par exemple dans le cas de l'ACP, on supposera que celles-ci suivent une loi normale multivariée, et l'on cherchera à estimer les covariances (ou corrélations) entre variables au sein de la population (Anderson, 1958). Il s'agit donc bien de statistique

inférentielle : la partie est utilisée pour décrire le tout. Une conséquence immédiate est que le nombre d'individus doit être grand devant le nombre de variables, afin d'avoir des estimations robustes des covariances entre variables.

La seconde approche est descriptive, en ce sens qu'elle s'intéresse d'abord à la structure de l'échantillon, et dans une moindre mesure à celle du tout. Cette démarche repose sur des considérations géométriques, qui sont centrales dans l'ouvrage de Takeuchi *et al.* (1984) : les données sont vues commes des nuages de points situés dans un espace multivarié, et l'ordination en espace réduit consiste à traduire en quelques dimensions les caractéristiques géométriques essentielles du nuage de points. C'est le point de vue adopté par Pearson (1901) dans son article fondateur sur l'ACP :

> *In many physical, statistical, and biological investigations it is desirable to represent a system of points in plane, three, or higher dimensioned space by the "best-fitting" straight line or plane.*

Dès lors, les hypothèses distributionnelles s'effacent et il n'est plus nécessaire que le nombre d'individus soit grand devant le nombre de variables, puisque les conclusions ne portent que sur le jeu de données étudié. Bien qu'il présente les deux points de vue, on trouve dans Joliffe (2004, pp.49-50) des éléments pour préférer l'approche descriptive à l'approche inférentielle :

> *The major assumption that* $\mathbf{x}$ *has a multivariate normal distribution is often not satisfied and the practical value of the results is therefore limited. It can be argued that PCA should only ever be done for data that are, at least approximately, multivariate normal, for it is only then that 'proper' inferences can be made regarding the underlying population* [...] *this is a narrow view of what PCA can do, as it is a much more widely applicable tool whose main use is descriptive rather than inferential. The majority of applications of PCA successfully treat the technique as a purely descriptive tool.*

Ce faisant, l'auteur concède à l'approche inférentielle :

> *Although purely inferential side of PCA is a very small part of the overall picture, the ideas of inference can sometimes be useful*

On pourra donc retenir que la distinction entre approche inférentielle et approche descriptive a ceci d'intéressant qu'elle attire notre attention sur la portée des résultats obtenus, et sur les limites des conclusions qui peuvent être tirées à partir de l'analyse d'un jeu de données.

Le point de vue choisi ici appartient clairement à l'approche descriptive et reflète une pratique de la biométrie avant tout francophone (Cailliez & Pages, 1976; Pontier *et al.*, 1990; Saporta, 1990; Legendre & Legendre, 1998; Lebart *et al.*, 2004). Cette orientation est à la fois un héritage d'école et un choix motivé par une application plus large et sans doute plus riche des méthodes d'ordination à l'analyse de données biologiques. Par ailleurs, l'analyse multivariée prise en tant qu'outil descriptif met en exergue des fondements géométriques qui permettent d'unifier les différentes méthodes sous un même cadre théorique. La présentation n'en est pas unique : Lebart *et al.* (2004, pp.15-31) le décrivent en tant qu'**analyse générale**, mais une présentation plus

complète, et sans doute plus élégante sur le plan théorique est celle du **schéma de dualité** (Cailliez & Pages, 1976, pp.194-220), dont nous abordons à présent la description.

1.2.2 Le schéma de dualité

Le schéma de dualité est un formalisme mathématique donnant un cadre général à l'analyse multivariée (Cailliez & Pages, 1976), introduit en écologie par Escoufier (1987) et présenté à nouveau récemment par Holmes (2006) et Dray & Dufour (2007). Des travaux précédents ont déjà couvert l'essentiel du sujet : Yoccoz (1988) en a détaillé les fondements mathématiques et en a souligné l'intérêt pour l'analyse de données écologiques, et Dray (2003) a replacé les méthodes d'ordination les plus couramment utilisées en écologie dans le cadre du schéma de dualité. La présentation que nous proposons ici, volontairement différente des travaux précédents, est donc nécessairement incomplète. Nous avons choisi d'illustrer d'abord la signification générale de l'objet mathématique, en insistant sur le sens des opérations décrites du point de vue empirique. Par la suite, nous décrivons quelques résultats connus qui paraissent fondamentaux et sont utilisés, explicitement ou implicitement, par les méthodes mises en oeuvre ou développées dans le reste du manuscrit.

Le résultat principal du schéma de dualité est que la majorité des ordinations en espace réduit peut se résumer à l'analyse d'un triplet de matrices (le **triplet statistique**), qu'on note $\mathfrak{T} = (\mathbf{X}, \mathbf{Q}, \mathbf{D})$. La matrice $\mathbf{X}$ est le tableau de données portant n **individus** (au sens statistique) en lignes et p **variables** (ou descripteurs) en colonnes. On associe aux colonnes de $\mathbf{X}$ une matrice de pondération symétrique $\mathbf{Q}$ de dimension p et définie positive ($\mathbf{a}^T\mathbf{Q}\mathbf{a} > 0 \ \forall \mathbf{a} \in \mathbb{R}^p, \mathbf{a} \neq 0$). De façon similaire, $\mathbf{D}$ est une matrice symétrique de dimension n et définie positive ($\mathbf{b}^T\mathbf{D}\mathbf{b} > 0$ $\forall \mathbf{b} \in \mathbb{R}^n, \mathbf{b} \neq 0$), portant des pondérations des lignes de $\mathbf{X}$. On verra plus loin que les matrices $\mathbf{Q}$ et $\mathbf{D}$ servent à définir des distances (ou des similarités) entre les variables et entre les individus contenus dans $\mathbf{X}$. Le schéma de dualité est un ensemble d'applications linéaires, détaillées plus bas, qui résume cette situation (Fig. 1.1).

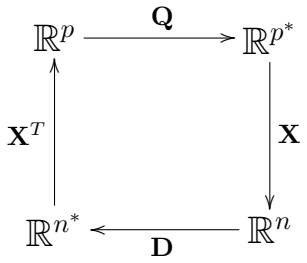


Fig. 1.1: Schéma de dualité du triplet $\mathfrak{T} = (\mathbf{X}, \mathbf{Q}, \mathbf{D})$

Le terme de **dualité** provient du fait qu'on puisse considérer ces données de deux façons, en s'intéressant aux relations entre les colonnes de $\mathbf{X}$ (les variables) ou entre ses lignes (les individus). Il en résulte deux analyses pour un seul et même tableau, dont les procédures sont parfaitement symétriques, et qui correspondent à deux points de vue complémentaires. Néanmoins, l'existence de cette dualité pour chaque triplet statistique n'implique pas que les

deux analyses aient un sens, et il arrivera souvent en pratique qu'une seule soit exploitée.

a. Définition des relations entre variables

Une première approche considère $\mathbf{X}$ comme un ensemble de variables, c'est-à-dire p vecteurs de $\mathbb{R}^n$:

$$\mathbf{X} = [\mathbf{X}^1|\cdots|\mathbf{X}^j|\cdots|\mathbf{X}^p] \text{ avec } \mathbf{X}^j \in \mathbb{R}^n,\ j = 1,\ldots,p$$

Le schéma de dualité définit une application f_1 :

$$\begin{array}{rccc} f_1 : & \mathbb{R}^n & \xrightarrow{\mathbf{D}} & \mathbb{R}^{n*} \\ & \mathbf{X}^j & \longmapsto & \langle \mathbf{X}^j, \cdot \rangle_{\mathbf{D}} = \cdot\, \mathbf{D}\mathbf{X}^j \end{array} \tag{1.1}$$

où $\mathbb{R}^{n*}$ est le dual de $\mathbb{R}^n$, c'est-à-dire l'ensemble des applications linéaires de $\mathbb{R}^n$ dans $\mathbb{R}$. Par la fonction f_1, on associe à une variable $\mathbf{X}^j$ un produit scalaire ; $\mathbf{X}^j$ est donc dans un espace euclidien.

La second opération définie est une application f_2 :

$$\begin{array}{rccc} f_2 : & \mathbb{R}^{n*} & \xrightarrow{\mathbf{X}^T} & \mathbb{R}^p \\ & \langle \mathbf{X}^j, \cdot \rangle_{\mathbf{D}} & \longmapsto & [\langle \mathbf{X}^j, \mathbf{X}^1 \rangle_{\mathbf{D}} \cdots \langle \mathbf{X}^j, \mathbf{X}^j \rangle_{\mathbf{D}} \cdots \langle \mathbf{X}^j, \mathbf{X}^p \rangle_{\mathbf{D}}]^T = \mathbf{X}^T\mathbf{D}\mathbf{X}^j \end{array} \tag{1.2}$$

Par l'application $f_2 \circ f_1$ (*i.e.*, $f_2(f_1(\mathbf{X}^j))$), on définit les relations entre les variables dans leur espace euclidien à l'aide du produit scalaire $\mathbf{D}$; ce produit est une mesure de **similarité** entre les variables qui prend en général ses valeurs dans $\mathbb{R}$. Par exemple, si $\mathbf{D} = \frac{1}{n}\mathbf{I}_n$ et $\mathbf{X}^j \perp \mathbf{1}_n$ $\forall j = 1,\ldots,p$ (*i.e.*, les variables sont centrées), alors $f_2 \circ f_1$ associe à une variable le vecteur de ses covariances avec l'ensemble des variables (y compris elle-même, c'est-à-dire sa variance). Si en plus de la condition précédente les variables sont normées ($\|\mathbf{X}^j\|^2_{\mathbf{D}} = 1$), alors $f_2 \circ f_1$ associe à une variable ses corrélations avec l'ensemble des variables.

b. Définition des relations entre individus

Le second point de vue considère $\mathbf{X}$ comme un ensemble d'individus, c'est-à-dire n vecteurs de $\mathbb{R}^p$ (on notera $\mathbf{X}_{[i]}$ la $i$ème ligne de $\mathbf{X}$) :

$$\mathbf{X}^T = [\mathbf{X}_{[1]}|\cdots|\mathbf{X}_{[i]}|\cdots|\mathbf{X}_{[n]}] \text{ avec } \mathbf{X}_{[i]} \in \mathbb{R}^p,\ i = 1,\ldots,n$$

Le schéma de dualité définit une application f_3 :

$$\begin{array}{rccc} f_3 : & \mathbb{R}^p & \xrightarrow{\mathbf{Q}} & \mathbb{R}^{p*} \\ & \mathbf{X}_{[i]} & \longmapsto & \langle \mathbf{X}_{[i]}, \cdot \rangle_{\mathbf{Q}} = \cdot\, \mathbf{Q}\mathbf{X}_{[i]} \end{array} \tag{1.3}$$

où $\mathbb{R}^{p*}$ est le dual de $\mathbb{R}^p$. Par la fonction f_3, on associe à un individu statistique $\mathbf{X}_{[i]}$ un produit scalaire ; $\mathbf{X}_{[i]}$ est donc également dans un espace euclidien.

L'opération suivante est définie par l'application f_4 :

$$\begin{array}{rccl} f_4 : & \mathbb{R}^{p*} & \xrightarrow{\mathbf{X}} & \mathbb{R}^n \\ & \langle \mathbf{X}_{[i]}, \cdot \rangle_{\mathbf{Q}} & \longmapsto & [\langle \mathbf{X}_{[i]}, \mathbf{X}_{[1]} \rangle_{\mathbf{Q}} \cdots \langle \mathbf{X}_{[i]}, \mathbf{X}_{[i]} \rangle_{\mathbf{Q}} \cdots \langle \mathbf{X}_{[i]}, \mathbf{X}_{[n]} \rangle_{\mathbf{Q}}]^T = \mathbf{X}\mathbf{Q}\mathbf{X}_{[i]} \end{array} \tag{1.4}$$

Comme pour les variables, l'application $f_4 \circ f_3$ définit les relations entre les individus dans leur espace euclidien, par le produit scalaire $\mathbf{Q}$; ce produit est une mesure de **similarité** entre individus qui prend également ses valeurs dans $\mathbb{R}$. Cependant, on préfère souvent considérer non pas les similarités entre individus, mais leurs **distances**. Ceci n'est qu'un changement de point de vue : $\mathbf{Q}$ étant définie positive, le produit scalaire sous-tend une mesure de distance d entre paires d'individus (respectivement $\mathbf{X}_{[i]}$ et $\mathbf{X}_{[k]}$) :

$$\begin{array}{rccl} d : & \mathbb{R}^p \times \mathbb{R}^p & \xrightarrow{\mathbf{Q}} & \mathbb{R}_+ \\ & (\mathbf{X}_{[i]}, \mathbf{X}_{[k]}) & \longmapsto & \|\mathbf{X}_{[i]} - \mathbf{X}_{[k]}\|_{\mathbf{Q}} \end{array} \tag{1.5}$$

En particulier, si $\mathbf{Q}$ est la matrice identité, alors d mesure la distance euclidienne canonique entre individus.

1.2.3 Dualité des analyses

La dualité entre nuage des individus dans l'espace engendré par les variables et nuage de variables dans l'espace engendré par les individus se retrouve au niveau de l'analyse : le schéma de dualité permet d'analyser les deux nuages séparemment, mais propose aussi des formules de transition d'un espace vers un autre. Une description récente des propriétés mathématiques du schéma de dualité peut être trouvée dans Holmes (2006) et Dray & Dufour (2007). La présentation qui suit est volontairement parcellaire et n'aborde que quelques-unes de ses propriétés qui paraissent fondamentales.

L'analyse des deux nuages de points repose sur le même principe : on recherche dans l'espace multivarié des directions orthogonales dans lesquelles la dispersion des points est maximale. Ces directions correspondent à des axes sur lesquels le nuage de points est projeté. La dispersion des coordonnées obtenues (ou **scores**) est mesurée par leur inertie, c'est-à-dire le carré de leur norme pour une métrique donnée.

a. Analyse des relations entre variables

Comme il a été vu plus haut, les variables peuvent être considérées comme p vecteurs de $\mathbb{R}^n$ muni d'une métrique $\mathbf{D}$ (Fig. 1.2). L'analyse de ce nuage de points consiste à trouver une base de vecteurs propres $\mathbf{D}$-orthonormés (en rouge sur la figure 1.2) maximisant l'inertie des projections du nuage de variables (en bleu sur la figure 1.2). Ceci revient à rechercher une base $\mathbf{V} = [\mathbf{v}_1 | \cdots | \mathbf{v}_k | \cdots \mathbf{v}_r]$ de vecteurs $\mathbb{R}^n$ vérifiant :

$$\|\mathbf{X}^T\mathbf{D}\mathbf{v}_k\|^2_{\mathbf{Q}} = \lambda_k \text{ avec } \lambda_k > \lambda_{k+1} \forall k = 1, \ldots, r-1 \text{ et } \mathbf{V}^T\mathbf{D}\mathbf{V} = \mathbf{I}_r \tag{1.6}$$

La solution est donnée par la diagonalisation de l'opérateur $\mathbf{D}$-symétrique $\mathbf{X}\mathbf{Q}\mathbf{X}^T\mathbf{D}$, fournissant des vecteurs propres $\mathbf{D}$-orthonormés maximisant l'inertie des scores des variables

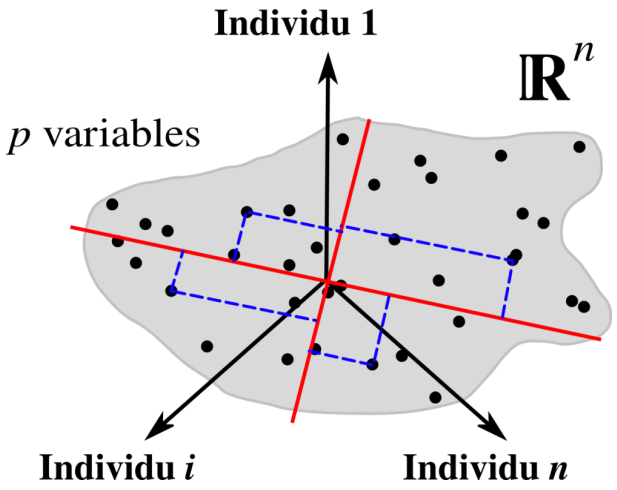


FIG. 1.2: Analyse du nuage des variables. Les axes rouges correspondent aux composantes principales normées à 1 ; les coordonnées des variables sur ces axes sont obtenues par projection orthogonale (en bleu).

(EQN. 1.6). Ces vecteurs sont les **composantes principales de norme unitaire** associée au triplet $\mathfrak{T} = (\mathbf{X}, \mathbf{Q}, \mathbf{D})$. En pratique, comme les programmes de diagonalisation ne fournissent que des vecteurs propres orthonormés pour la métrique canonique, on calculera d'abord les vecteurs propres $\mathbf{b}_k$ $(k = 1, \ldots, r)$ de $\mathbf{D}^{1/2}\mathbf{X}\mathbf{Q}\mathbf{X}^T\mathbf{D}^{1/2}$. Par définition, ces vecteurs vérifient :

$$\|\mathbf{X}^T\mathbf{D}^{1/2}\mathbf{b}_k\|^2_{\mathbf{Q}} = \lambda_k \text{ avec } \|\mathbf{b}_k\|^2 = 1 \text{ et } \mathbf{b}_k^T\mathbf{b}_l = 0 \ \forall k \neq l \tag{1.7}$$

En prenant $\mathbf{v}_k = \mathbf{D}^{-1/2}\mathbf{b}_k$ $(\Leftrightarrow \mathbf{b}_k = \mathbf{D}^{1/2}\mathbf{v}_k)$, on obtient bien les vecteurs recherchés (EQN. 1.6). La signification de cette opération dépend évidemment de la nature du triplet $\mathfrak{T}$. Par exemple, dans le cas de l'ACP centrée où $\mathbf{Q} = \mathbf{I}_p$ et $\mathbf{D} = \frac{1}{n}\mathbf{I}_n$, les vecteurs $\mathbf{v}_k$ sont des variables de synthèse normées maximisant la somme des carrés des covariances avec toutes les autres variables $(\frac{1}{n^2}\|\mathbf{X}^T\mathbf{v}_k\|^2 = \sum_{j=1}^p \text{cov}^2(\mathbf{X}^j, \mathbf{v}_k))$. En ACP normée, les variables de synthèse $\mathbf{v}_k$ maximisent la somme des carrés de corrélations avec toutes les autres variables de $\mathbf{X}$ $(\frac{1}{n^2}\|\mathbf{X}^T\mathbf{v}_k\|^2 = \sum_{j=1}^p \text{cor}^2(\mathbf{X}^j, \mathbf{v}_k))$.

b. Analyse des relations entre individus

Le nuage des individus peut être analysé de la même façon que le nuage des variables. On étudie alors les relations typologiques entre n vecteurs de $\mathbb{R}^p$ muni d'une métrique $\mathbf{D}$ (FIG. 1.3). Comme pour le nuage des variables, on recherche une base de vecteurs propres $\mathbf{Q}$-orthonormés (en bleu sur la figure 1.3) maximisant l'inertie des projections du nuage d'individus (en rouge sur la figure 1.3). Ceci revient à rechercher une base $\mathbf{U} = [\mathbf{u}_1|\cdots|\mathbf{u}_k|\cdots\mathbf{u}_r]$ de vecteurs $\mathbb{R}^p$ vérifiant :

$$\|\mathbf{X}\mathbf{Q}\mathbf{u}_k\|^2_{\mathbf{D}} = \lambda_k \text{ avec } \lambda_k > \lambda_{k+1} \forall k = 1, \ldots, r-1 \text{ et } \mathbf{U}^T\mathbf{Q}\mathbf{U} = \mathbf{I}_r \tag{1.8}$$

La solution est donnée par la diagonalisation de l'opérateur $\mathbf{Q}$-symétrique $\mathbf{X}^T\mathbf{D}\mathbf{X}\mathbf{Q}$, fournissant des vecteurs propres $\mathbf{Q}$-orthonormés maximisant (EQN. 1.6). Ces vecteurs sont les **axes principaux** associés à $\mathfrak{T}$. La procédure utilisée pour les obtenir est en tout point similaire à la procédure mise en oeuvre pour les variables. Les projections des individus sur les axes

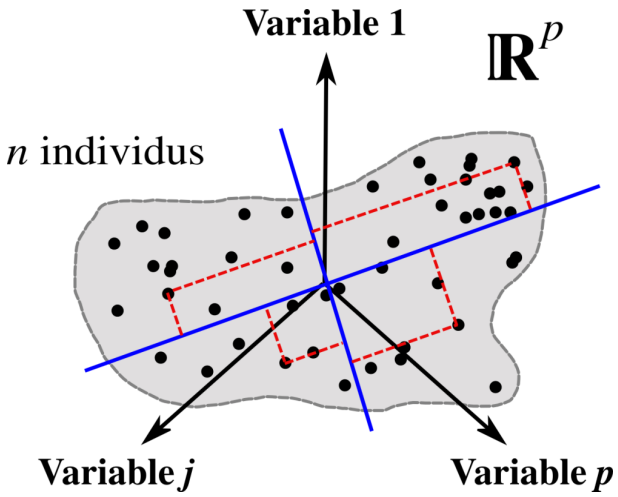


FIG. 1.3: Analyse du nuage des individus. Les axes bleus correspondent aux axes principaux ; les coordonnées des individus sur ces axes sont obtenues par projection orthogonale (en rouge).

principaux (**XQU**) sont les **composantes principales de norme maximale**. La signification de ces composantes principales dépend également de $\mathfrak{T}$. En ACP centrée comme normée, $\mathbf{Xu}_k$ est un score d'individus obtenu par combinaison linéaire des variables de variance maximale ($\frac{1}{n}\|\mathbf{Xu}_k\|^2 = \text{var}(\mathbf{Xu}_k)$).

c. D'une analyse à l'autre

L'ambiguïté terminologique entre les composantes principales de norme unitaire (**V**) et de norme maximale (**XQU**) n'est pas fortuite, et reflète le fait que l'on puisse obtenir l'une à partir de l'autre. En effet, les deux sont liées par une simple relation de normalisation :

$$\mathbf{V} = \mathbf{XQU}\boldsymbol{\Lambda}^{-1/2} \tag{1.9}$$

Cette relation procède de :

$$\begin{aligned} \mathbf{X}^T\mathbf{DXQU} &= \mathbf{U}\boldsymbol{\Lambda} && (1.10)\\ \mathbf{XQX}^T\mathbf{DXQU} &= \mathbf{XQU}\boldsymbol{\Lambda} && (1.11)\\ \mathbf{XQX}^T\mathbf{DXQU}\boldsymbol{\Lambda}^{-1/2} &= \mathbf{XQU}\boldsymbol{\Lambda}^{1/2} && (1.12) \end{aligned}$$

en posant $\mathbf{XQU}\boldsymbol{\Lambda}^{-1/2} = \mathbf{V}$ ($\Leftrightarrow \mathbf{XQU} = \mathbf{V}\boldsymbol{\Lambda}^{1/2}$) et on obtient :

$$\mathbf{XQX}^T\mathbf{DV} = \mathbf{V}\boldsymbol{\Lambda} \tag{1.13}$$

La même relation existe pour les axes principaux :

$$\mathbf{U} = \mathbf{X}^T\mathbf{DV}\boldsymbol{\Lambda}^{-1/2} \tag{1.14}$$

Ces relations de transition sont particulièrement intéressantes car elles permettent d'obtenir les deux analyses d'un tableau en ne procédant qu'à une seule diagonalisation. Cette propriété

est utile lorsque les deux analyses sont d'égale importance comme en analyse de co-inertie (Dolédec & Chessel, 1994), et précieuse pour l'analyse de grands jeux de données, puisqu'on pourra diagonaliser la plus petite des matrices $\mathbf{X}^T\mathbf{DXQ}$ et $\mathbf{XQX}^T\mathbf{D}$.

Par ailleurs, ces relations de transition entre les espaces permettent d'identifier les maximums des formes ($\langle \mathbf{XQu}|\mathbf{v}\rangle_\mathbf{D}$) et ($\langle \mathbf{X}^T\mathbf{Dv}|\mathbf{u}\rangle_\mathbf{Q}$) :

$$\begin{aligned} \|\mathbf{XQu}_k\|^2_\mathbf{D} &= \lambda_k \\ \langle \mathbf{XQu}_k|\mathbf{XQu}_k\rangle_\mathbf{D} &= \lambda_k \\ \langle \mathbf{XQu}_k|\sqrt{\lambda}\mathbf{v}_k\rangle_\mathbf{D} &= \lambda_k \\ \langle \mathbf{XQu}_k|\mathbf{v}_k\rangle_\mathbf{D} &= \sqrt{\lambda_k} \end{aligned} \tag{1.15}$$

et inversement :

$$\begin{aligned} \|\mathbf{X}^T\mathbf{Dv}_k\|^2_\mathbf{Q} &= \lambda_k \\ \langle \mathbf{X}^T\mathbf{Dv}_k|\mathbf{X}^T\mathbf{Dv}\rangle_\mathbf{Q} &= \lambda_k \\ \langle \mathbf{X}^T\mathbf{Dv}_k|\sqrt{\lambda}\mathbf{u}_k\rangle_\mathbf{Q} &= \lambda_k \\ \langle \mathbf{X}^T\mathbf{Dv}_k|\mathbf{u}\rangle_\mathbf{Q} &= \sqrt{\lambda_k} \end{aligned} \tag{1.16}$$

qui sont utilisées notamment par plusieurs méthodes K-tableaux dont certaines seront abordées au chapitre 2, telles que l'analyse de co-inertie multiple (Chessel & Hanafi, 1996) ou l'analyse canonique généralisée (Carroll, 1968).

1.2.4 Du rôle du schéma de dualité

Nous avons vu que le schéma de dualité permet de généraliser les ordinations en espace réduit en les voyant comme des cas particuliers de l'analyse d'un triplet statistique. Outre une formulation mathématique élégante, on retiendra d'une part les propriétés d'orthonormalité des axes principaux et composantes principales normées ainsi que la maximisation de l'inertie des scores projetés sur ces axes, et d'autre part l'utilité de l'objet en tant que représentation d'une analyse. En effet, devant la complexité de certaines ordinations en espace réduit, le fait de pouvoir se ramener à un schéma général permet de s'orienter. Par ailleurs, par sa nature générique, le schéma de dualité forme un carcan sur lequel on pourra se baser pour développer de nouvelles méthodes ou adapter des méthodes existantes à des problématiques particulières.

Cepdendant, il est manifeste que la généralité des propriétés qui ont été énoncées n'a que peu de sens du point de vue empirique, et c'est l'interactions d'une méthode avec un type de données qui possède un sens. Ce sont ces significations particulières qui feront d'une méthode un outil pertinent, ou au contraire inutile, pour une problématique donnée. Il sera donc crucial de les identifier si l'on veut dresser un bilan de l'utilité de l'analyse multivariée dans l'extraction de structures biologiques des données de marqueurs génétiques. Notons que ces interactions ne relèvent plus complètement des mathématiques (les fondements mathématiques ne sont plus en cause) mais ne relèvent pas plus de la biologie : il s'agit donc bien de biométrie.

Maintenant que nous avons identifié les bases théoriques de l'analyse de données telles que nous l'entendons, c'est-à-dire au sens de Cailliez & Pages (1976), nous pouvons nous intéresser à l'autre composante du dialogue interdisciplinaire qui est le coeur de l'activité biométrique : les questions biologiques.

1.3 Motivations biologiques

Le champ des questions biologiques qui peuvent être abordées via l'analyse de marqueurs moléculaires est immense, et il n'est pas question ici d'en faire le tour, ou même d'en dresser une carte générale. Il semble plus raisonnable et plus pertinent de se borner à expliciter les quelques thématiques qui ont motivé les propositions méthodologiques auxquelles ce travail de thèse a abouti. Le choix de ces thématiques provient du fait qu'elles répondent à une problématique biologique féconde et nécessitent l'apport de méthodologies nouvelles. Deux sujets déjà vastes en eux-mêmes seront abordés : la **mesure de la biodiversité** et l'**identification de la structure des populations naturelles**.

1.3.1 Mesurer la biodiversité

La mesure de la biodiversité est en soi un sujet trop riche pour être ici convenablement résumé : on renverra à la thèse de Sandrine Pavoine pour un traitement approfondi (Pavoine, 2005). L'analyse de marqueurs génétiques est sans doute aujourd'hui une des approches courantes pour mesurer la biodiversité au sein d'un ensemble d'organismes, le plus souvent partitionnés en espèces ou en « races » dans le cas de populations domestiques ayant des phénotypes bien différenciés. L'ordination en espace réduit peut alors être utilisée pour rechercher une image de cette biodiversité. Cette recherche, bien qu'elle puisse émaner d'une simple volonté de connaître les relations génétiques entre les objets étudiés (*e.g.,* Mitton, 1978; MacHugh *et al.*, 1998; Laval *et al.*, 2000; Pariset *et al.*, 2003; Xuebin *et al.*, 2005)[3], est souvent motivée par des objectifs de conservation (Toro, 2006). Il s'agit alors de décider si l'originalité d'une espèce ou d'une race domestique est suffisante pour justifier sa conservation (*e.g.*, Moazami-Goudarzi *et al.*, 2001), c'est-à-dire si son potentiel évolutif est suffisamment différent des autres pour acquérir le statut d'**unité évolutive significative** (*Evolutionary Significant Units* (ESU), Moritz, 1994). Néanmoins, on contestera la vision de Moritz (1994) qui compte définir ces unités évolutives significatives sur un critère algorithmique ; d'une façon générale, l'objet biologique est sans doute trop complexe et trop variable pour qu'il puisse rentrer dans un moule universel. En l'occurence, la question de la conservation d'espèces semble en effet trop complexe pour se contenter d'examiner le seul critère génétique (Paetkau, 1999; Fraser & Bernatchez, 2001), mais on peut accorder à celui-ci qu'il constitue au moins une partie de l'information à prendre en compte. McKay & Latta (2002) notent à ce sujet :

> *We emphasize that none of the foregoing is intended to argue against the use of molecular markers or translocations, both of which can be extremely beneficial in ecological, evolutionary or conservation studies.* [...] *However, we caution against an oversimplified interpretation of the results, in which it is assumed that a low marker*

differentiation inevitably precludes adaptative differentiation.

La question n'étant pas tranchée, on retiendra l'idée que des unités évolutives significatives sont des groupes d'organismes différenciés au plan génétique et fonctionnel. Du point de vue biométrique, l'analyse des données de marqueurs moléculaires dans l'inférence de la biodiversité est une problématique riche (Pavoine, 2005). Un aspect intéressant et relativement peu traité est celui de la façon dont on considère l'information provenant de différents marqueurs génétiques, en particulier lorsque ceux-ci sont multialléliques (*e.g.*, allozymes, microsatellites). Dans de tels cas, chaque marqueur peut être suffisamment informatif pour fournir une image de la biodiversité, et il peut être intéressant de comparer les images fournies par différents marqueurs (Moazami-Goudarzi & Laloë, 2002). Néanmoins, la pratique courante consiste à réunir tous les marqueurs en une seule analyse, ce qui implique qu'on considère, souvent à tort, qu'ils portent tous la même information. La question de la cohérence de l'information provenant de différents marqueurs moléculaires est avant tout méthodologique : elle demande des méthodes pour extraire l'information biologique de chaque marqueur, pour rechercher une information commune, et situer chaque marqueur par rapport à cette information consensuelle. Nous verrons au chapitre 2 que l'analyse multivariée offre un certain nombre de réponses prometteuses à cette problématique. Notons que la question de la cohérence typologique de l'information génétique, abordée dans le cadre d'étude d'unités évolutives significatives (Moazami-Goudarzi & Laloë, 2002), pourrait également être posée au niveau populationnel : les réponses méthodologiques proposées seront valables dans un cas comme dans l'autre.

1.3.2 Identifier la structure des populations naturelles

Nombre d'études utilisant des marqueurs génétiques pour répondre à des problématiques biologiques sont conduites au sein d'une même unité évolutive significative (souvent une espèce) : on s'intéresse alors à la structuration génétique d'individus ou de groupes d'individus d'une même espèce, dont les aires de répartition sont souvent connexes ou communes. En biologie de la conservation, on voudra alors identifier des **unités de gestion** (*management units* (MU), Moritz, 1994), c'est-à-dire des groupes d'individus d'une même espèce indépendants sur le plan démographique. L'identification de ces unités de gestion se fait en général sur des critères génétiques, à partir de l'analyse de données de marqueurs moléculaires (Palsboll *et al.*, 2006; Schwartz *et al.*, 2006; Toro, 2006). Notons qu'ici encore, l'information génétique n'est pas la seule qui soit nécessaire à la prise de décision au niveau de la gestion des populations, mais intervient au sein d'un faisceau de considérations politiques, écologiques, économiques et démographiques (Paetkau, 1999; Taylor & Dizon, 2002). Paetkau (1999) et Palsboll *et al.* (2006) soulignent le fait que l'outil moléculaire n'est qu'un moyen d'inférer des caractéristiques démographiques, qui seules définissent les unités de gestion. De ce point

[3]On notera que la recherche de la nature des relations génétiques n'est pas toujours naïve lorsqu'il s'agit de génétique humaine : par exemple, Mitton (1978) cherche à démontrer le bien-fondé du concept de races humaines sur des critères génétiques par une interprétation frauduleuse d'une analyse en composantes principales. Ce point est abordé dans l'article de synthèse bibliographique présenté plus bas dans ce chapitre (voir 'Interpreting genetic structures'). Ce mode de pensée semble avoir disparu de l'esprit scientifique actuel (Edwards, 2003; Bamshad *et al.*, 2004), et il serait aujourd'hui surprenant de voir paraître un article intitulé « *Are human races "substantially" different genetically ?* » (Powell & Taylor, 1978).

de vue, les marqueurs génétiques sont des outils imparfaits, puisqu'ils ne reflètent jamais la situation démographique telle qu'elle est, mais telle qu'elle a été. Ainsi, tout événement récent affectant la démographie d'un ensemble d'unités de conservation sera invisible du point de vue moléculaire. Par exemple, des populations très différentes génétiquement peuvent être réunies (*e.g.*, par une réintroduction) et fonctionner comme une seule unité démographique ; inversement, une population panmictique peut être scindée en plusieurs sous-populations (*e.g.*, par une fragmentation de l'habitat), qui deviennent indépendantes sur le plan démographique mais restent génétiquement proches. Dans les deux cas, l'analyse de données génétiques sera prise en défaut si trop peu de générations séparent l'observateur de l'événement : on identifiera trop d'unités de gestion dans le premier cas, et trop peu dans le second. Néanmoins, l'approche génétique demeure efficace dans la majorité des cas (Palsboll *et al.*, 2006; Schwartz *et al.*, 2006), et constitue parfois la seule approche possible du problème (Schwartz *et al.*, 2006). Par ailleurs, une fois les unités de gestion identifiées, il peut être intéressant de savoir quels groupes d'individus sont les plus différents sur le plan génétique, ou recèlent le plus de variabilité génétique entre individus, pour éventuellement orienter le choix des unités conservées.

Il reviendra donc au biométricien de rechercher des structures génétiques au sein d'un ensemble d'individus, c'est-à-dire d'identifier des groupes de génotypes relativement semblables entre eux et différents de ceux des autres groupes. Cet énoncé met en évidence une certaine subjectivité, et comme pour l'identification d'unités évolutives significatives, il faudra renoncer à un critère universel (Palsboll *et al.*, 2006) :

> *At what amount of population genetic divergence should populations be assigned to different MUs ? A 'one size fits all' answer is not possible given that it depends upon the specific conservation context, as well as on the biological characteristics and populations history of the target species.* [...] *More importantly, as Waples and Gaggiotti* [2006, Molecular Ecology 15 : 1419-1439] *point out, there is currently no general framework for determining at which dispersal rate populations become demographically correlated.*

L'exploration de la structure génétique des populations naturelles se fait aujourd'hui le plus souvent par le biais d'une information génétique spatialisée (Manel *et al.*, 2003) : une étude typique implique la recherche de structures dans un ensemble de génotypes géoréférencés (*e.g.*, Taberlet *et al.*, 1995; Waits *et al.*, 2000; Manel *et al.*, 2004, 2007). Il sera donc important que la méthode utilisée pour analyser les données prenne en compte l'**information spatiale** au même titre que l'**information génétique**. On note par ailleurs que dans le cas général, la distribution géographique des génotypes ne permet pas d'inférer des groupes *a priori* (Paetkau, 1999; Manel *et al.*, 2003) : la définition des groupes existants ne saurait être une donnée du problème. Enfin, on notera les difficultés posées par l'utilisation de modèles dont les hypothèses de départ sont parfois peu réalistes ou non vérifiées (Palsboll *et al.*, 2006) :

> *Current population genetic inferences rely upon highly idealized and simplistic population models that do not apply to most natural populations.* [...] *It is important to know when the conclusions from these models are robust with regard to deviations*

> *from the underlying assumptions. The sensitivity of biologically feasible deviations from the underlying population genetic model should be assessed* [...] *before making firm recommendations.*

On retiendra donc que l'identification d'unités de gestion requiert en général une approche exploratoire, qui ne nécessite que peu de connaissances préalables à propos du système étudié, et qui fasse peu d'hypothèses quant au modèle de génétique des populations sous-jacent (voire, qui n'en utilise pas). L'analyse de données, au sens de Cailliez & Pages (1976), semble donc un outil de choix pour répondre à cette problématique. Le chapitre 3, et d'une façon moins particulière le chapitre 5, présentent des éléments méthodologiques permettant d'aborder cette problématique sous un angle nouveau.

A présent que les bases méthodologiques et les motivations biologiques ont été établies, c'est-à-dire maintenant que nous savons avec quel équipement et dans quelle direction nous partons, nous pouvons nous intéresser au paysage, c'est-à-dire à l'état du monde situé entre statistique et génétique tel que nous avons pu le percevoir. La revue bibliographique qui suit, soumise à *Heredity*, dresse un état des lieux des applications de l'analyse multivariée à l'extraction d'information biologique des données de marqueurs génétiques.

1.4 Article 1 : Genetic markers in the playground of multivariate analysis

Revue bibliographique commandée par **Heredity**, en révision.

Genetic markers in the playground of multivariate analysis

T Jombart[1], D Pontier[1] and A-B Dufour[1]

[1] Université de Lyon, F-69000, Lyon ; Université Lyon 1 ; CNRS, UMR5558, Laboratoire de Biométrie et Biologie Evolutive, F-69622, Villeurbanne, France.

Keywords: Multivariate analysis, Ordination, Genetic markers, Spatial, Principal Component Analysis, Statistics, Methods

Corresponding author:
Thibaut Jombart
UMR CNRS 5558 - LBBE
"Biométrie et Biologie Évolutive"
UCB Lyon 1 - Bât. Grégor Mendel
43 bd du 11 novembre 1918
69622 VILLEURBANNE cedex
FRANCE

Phone: +33 (0)4 72 43 29 35

Fax: +33 (0)4 72 43 13 88

Email: jombart@biomserv.univ-lyon1.fr

Running Title: Multivariate analysis of genetic markers

Word count (excluding references and legends): 6899

Abstract

Multivariate analyses such as principal component analysis were among the first statistical methods employed to extract information from genetic markers. From their early applications to current innovations, these approaches have proven to be efficient for the analysis of the genetic variability in various contexts such as human genetics, conservation, and adaptation studies. However, because multivariate analysis is a wide and diversified area of statistics, choosing a method appropriate to both the data and to the question being asked can be difficult. Moreover, some particularities of genetic markers need to be taken into account when using multivariate methods. As a consequence, multivariate analyses are often used as black boxes, which results in frequent mistakes in the literature. In this review, we provide a critical analysis of the application of multivariate methods to genetic markers, using a general framework that unifies all these methods for the sake of clarity. First, we focus on some common mistakes in these applications and ways to avoid these pitfalls. We then detail the most critical particularities of allele frequencies that demand adaptations of multivariate methods, and we propose solutions to the subsequent problems. Finally, we tackle several questions of interest in which multivariate analysis has a great role to play, such as the study of the typological coherence of different genetic markers, or the investigation of spatial genetic patterns.

Introduction

Statistical methods have long become an essential component of the toolbox of population geneticists (Fisher, 1952). Developments in statistical theories and the continual increases in cheap computing power provide numerous tools for genetic marker analysis, allowing geneticists to address new and challenging questions. *Multivariate analyses* (also called *ordinations in reduced space*) such as principal component analysis (Pearson, 1901) have been shown to be efficient in extracting information from genetic markers (Cavalli-Sforza, 1966; Johnson *et al.*, 1969; Smouse *et al.*, 1982) because of their ability to summarize multivariate genetic information into a few synthetic variables. From these early applications to current innovative developments (Patterson *et al.*, 2006; Pavoine and Bailly, 2007; Jombart *et al.*, 2008), these methods have proven to be useful in various fields, such as human genetics (Menozzi *et al.*, 1978; Bertranpetit and Cavalli-Sforza, 1991; Cavalli-Sforza *et al.*, 1993), conservation (Moazami-Goudarzi *et al.*, 1997; Escudero *et al.*, 2003; Laloë *et al.*, 2007), phylogeography (Hanotte *et al.*, 2002; Matsuoka *et al.*, 2002; Ciofi *et al.*, 2006), landscape genetics (Angers *et al.*, 1999; Mcrae *et al.*, 2005), and the identification of adaptations (Johnson *et al.*, 1969; Mulley *et al.*, 1979; Barker *et al.*, 1986).

Multivariate analysis has several advantages over other classical approaches used in population genetics, like the Bayesian clustering implemented in the software STRUCTURE (Pritchard *et al.*, 2000; Falush *et al.*, 2003). First, multivariate methods are exploratory, *i.e.*, they do not require strong assumptions about an underlying genetic model, such as the Hardy-Weinberg equilibrium or the absence of linkage disequilibrium. While clustering approaches suppose that genotypes are structured in discrete populations, ordinations in reduced space simply aim at summarizing the genetic variability, and can therefore reveal any kind of genetic structuring including clines (for example, Jombart *et al.*, 2008). As stressed by Patterson *et al.* (2006), multivariate methods are not computer-intensive, and can be applied to huge datasets (such as '*hundreds of thousands of markers and thousands of samples*' in Patterson *et al.* (2006)), for which Bayesian clustering would be impractical. Moreover, multivariate analysis can

address complex questions such as identification of adaptation, by linking genetic variability to environmental data (Barker *et al.*, 1986; Angers *et al.*, 1999), while the impossibility of formulating an explicit model of adaptation would make Bayesian clustering methods inapplicable, in most cases. Lastly, multivariate methods have been developped and used extensively for more than a century in various fields, such as psychometry and ecology (Pearson, 1901). Currently, multivariate analysis represents a whole, rich, and diversified area of statistics offering a wide choice of methods, each with its own properties (Takeuchi *et al.*, 1984; Jambu, 1991; Legendre and Legendre, 1998).

The unfortunate consequence of this diversity of methods is that multivariate analyses are often used as black boxes when applied to genetic markers, leading to frequent mistakes that sometimes question the results of an entire study. In fact, it can be difficult to know which method can be efficiently applied to extract information from genetic markers, which precautions should be taken, and how the results should be interpreted. Moreover, there is no doubt that multivariate analysis has been under-utilized and has much more to offer to the study of the genetic variability. The purpose of this paper is to critically review the use of ordination in reduced space to infer biological structures from genetic markers.

First, we attempt to clarify the rationale for these methods and provide an overview of their current application to genetic markers. Frequent mistakes regarding the utilization of these methods are then detailed, and guidelines are provided to avoid these pitfalls. The following section focuses on some particularities of genetic markers that should be taken into account to improve their multivariate analysis. The rest of this review covers the use of multivariate analyses to tackle specific questions of interest, such as the coherence of the information of different genetic markers, linkage of genetic markers to other types of data, and the study of spatial genetic patterns. We conclude by examining some promising perspectives offered by these approaches to answer challenging questions in various fields, such as conservation, spatial genetics, and molecular ecology.

Multivariate analysis of genetic markers

Rationale of multivariate analysis

Throughout this paper, the terms 'ordination in reduced space' and 'multivariate analysis' are used indifferently. However, the first term is certainly more accurate than the second because ordinations in reduced space represent a particular class of multivariate methods, another being, for instance, hierarchical clustering. The purpose of these methods is to summarize a strongly multivariate dataset into a small set of uncorrelated *synthetic variables*. In other words, ordinations in reduced space aim to provide a simplified, yet meaningful, picture of complex information that is impossible to perceive. This task implies a necessary loss of information, and the crucial point in all these methods is to define a criterion that is optimised by the synthetic variables seeked. For instance, in principal component analysis (PCA, Pearson, 1901; Takeuchi *et al.*, 1984, pp.185-224), synthetic variables best preserve the variance among observations, while the chi-squared distances are preserved in the correspondence analysis (Greenacre, 1984). Below, we introduce general concepts required to describe multivariate analyses with accuracy.

As formalized by the *duality diagram* framework (Escoufier, 1987; Dray and Dufour, 2007), most multivariate analyses are particular cases of a general algorithm, and can be described using a small set of concepts. The terminology we employ encompasses the most common terms, which can be found in reference textbooks (for examples, Takeuchi *et al.*, 1984; Jambu, 1991; Legendre and Legendre, 1998; Lebart *et al.*, 2004). Central to the analysis of a dataset of n objects and p descriptors is the question of whether we seek a description of the relationships among the objects or among the descriptors. When analysing genetic markers, the main interest is in finding relationships among objects (genotypes or populations) using p alleles. In this case, data are seen as a cloud of n points embedded inside a p-dimensional space, where each dimension is defined by an allele. Inside this space, *inertia* measures the dispersion of

n points with respect to a given distance: this measurement of variability is used as a criterion that is optimised by the analysis. The directions inside this space reflecting the highest 'variability' (*i.e.*, with maximum inertia) among objects are the *principal axes*, also referred to as the *factors* of the analysis. By extension, a plane formed by two principal axes is often called a *factorial plane*. Each principal axis is defined by p coordinates inside the p-dimensional space, representing the *loadings* of the alleles. The principal axes are orthonormal (*i.e.*, perpendicular and with length one), and can therefore be used as a new basis to represent the n objects. The set of coordinates of the objects in this new basis are the *principal components*, but the terms *scores* (of objects), and *synthetic variables* are also commonly used. Each principal component is associated with an *eigenvalue* that quantifies the amount of inertia contained in the component. Eigenvalues can also be expressed as proportions of the total inertia of the analysis in order to indicate what fraction of the entire genetic variability is represented by the corresponding principal components. The plot of the eigenvalues sorted in decreasing order (the *screeplot*) is the basic tool used to choose which principal components to interpret: it describes how the total inertia is distributed across the principal axes. The basic idea is that a boundary between true structure and random noise would be indicated by a sharp decay between two successive eigenvalues. However, this is a simplistic view, and such a boundary rarely exist in practice: the screeplot merely provides insight about which component likely contains interesting structures, and which does not. Hence, the screeplot and the proportions of inertia associated with the principal components are two complementary tools, respectively indicating the genetic structures to be retained and their magnitude. The last criterion for interpreting principal components is that of the biological meaning, and is sometimes more useful than statistical criteria. In some cases, the first principal components (associated to large inertia) may indicate a trivial structuring, and provide little biological insight. Conversely, principal components associated to smaller eigenvalues might contain biologically relevant information; the interpretation of such components should not be discarded on the basis of a small inertia.

If multivariate analyses are unified by a single algorithm, the core difficulty is in choosing the method that best matches the nature of the data and the questions asked. Because of the variety of questions and data, numerous ordinations in reduced space are used to analyse genetic markers.

Applications to genetic markers

Multivariate analyses are natural tools to extract biological structures from genetic markers, as these data typically contain large numbers of genotypes or populations described by hundreds of alleles (in terms of absolute or relative frequencies). A summary of the application of these methods to genetic markers is provided in Table 1.

Ordinations in reduced space are primarily used to find a few principal components that reflect as much of the genetic variability as possible. PCA was first employed to infer population structuring (Cavalli-Sforza, 1966) and spatial genetic structuring (Menozzi *et al.*, 1978; Bertranpetit and Cavalli-Sforza, 1991; Cavalli-Sforza *et al.*, 1993) in humans. PCA was also used early to infer adaptations from allozyme frequencies, by testing the correlations between principal components of genetic data and principal components of a PCA of environmental variables (Johnson *et al.*, 1969). In disease studies, regression onto the principal components of the PCA has been recently proposed to correct for population stratification (Price *et al.*, 2006). Another method commonly used to infer genetic structuring among genotypes or populations is principal coordinates analysis (PCoA, Gower, 1966; Sanchez-Mazas and Langaney, 1988; Warnes, 2003). While PCA preserves the canonical Euclidean distance among the studied entities, PCoA can be employed to summarize any Euclidean genetic distance between genotypes or populations, but does not provide a representation of the alleles. This offers the advantage of using measures of genetic variability that are directly related to a population genetics model; for instance, PCoA has been used to summarize matrices of pairwise F_{ST} (Zhivotovsky *et al.*, 2003) and of Rogers' distance (Baker and Moeed, 1987). Non-metric dimensional scaling (NMDS, Cox and Cox, 2001) has also been employed to analyse matrices of genetic

distances (Baker and Moeed, 1987; Lessa, 1990). However, NMDS differs from PCoA in that it attempts to preserve the ordering of objects based on their genetic distance rather than their genetic distance *per se*; in this respect, NMDS can be thought of as a non-linear form of PCoA (Lessa, 1990). Note that, unlike other multivariate analyses, the NMDS solution is not analytical: an iterative algorithm aims at finding a good solution, but does not guarantee that this solution is the best. As an alternative to PCA of allele frequencies and PCoA (or NMDS) of genetic distances, correspondence analysis (CA, Greenacre, 1984) can be used to analyse a table of allele counts per populations (CA, She *et al.*, 1987; Li *et al.*, 2002). The last multivariate analysis commonly applied to genetic markers is discriminant analysis (DA, Lachenbruch and Goldstein, 1979). DA is not an fully exploratory approach, in that groups of genotypes must be known in advance. However, it can be used to achieve the best discrimination between groups inside a reduced space, to test for genetic differentiation, and for assignement purposes (Smouse *et al.*, 1982; Beharav and Nevo, 2003).

Other methods have remained somewhat unnoticed, such as constant-row multiple correspondence analysis (CRT-MCA, Guinand, 1996; Guinand *et al.*, 1996), factor analysis (Taylor and Mitton, 1974; Mulley *et al.*, 1979), and distance-based redundancy analysis (db-RDA, Legendre and Anderson, 1999; Geffen *et al.*, 2004). The reason for this may be historical, or could arise from problems associated with using these approaches. For instance, CRT-MCA aims at finding synthetic variables with maximum F_{ST}, but only proposes an approximate solution. Denoting f as a set of frequencies of an allele for q populations, $\bar{f}$ as the mean frequency computed across populations, and var(f) as the variance between populations of f, F_{ST} is defined as var$(f)/\bar{f}(1-\bar{f})$, where $\bar{f}(1-\bar{f})$ is the theoretical variance of f (Weir, 1996, p.166). Unfortunately, the quantity optimised by CRT-MCA is var$(f)/s_f^2$, where s_f^2 is the empirical variance of f ($s_f^2 = \frac{1}{q}\sum_{i=1}^{q}(f_i - \bar{f})^2$). While for arbitrarily large samples s_f^2 converges toward $\bar{f}(1-\bar{f})$, these quantities differ in practice, and the principal components yielded by CRT-MCA do not optimise the F_{ST}. A possible cause for the minimal use of factor analysis is that it was introduced to correlate patterns in allele

frequencies with environmental variables (Taylor and Mitton, 1974), which is not the purpose of this method. In fact, factor analysis estimates a model in which allele frequencies are expressed as a sum of two components: a part common to every allele and a residual part representing allele-specific effects (Seal, 1966, pp.153-180). Lastly, it is not clear why db-RDA has not been applied more often to genetic markers, but this could simply be due to its recent application (Geffen *et al.*, 2004).

While multivariate analyses can be efficiently used to extract information from genetic markers, choosing a method appropriate to the data and the question being asked is sometimes difficult. As a matter of fact, a number of mistakes occur quite frequently in such applications. In the following we point out the major pitfalls, as well as strategies to avoid them.

Misuses, misinterpretations, and specific issues

Ensuring reproducibility

A first concern in data analysis is to ensure reproducibility, or at least to provide all the elements required to evaluate the relevance of the results. Unfortunately, the literature regularly provides examples of studies in which it is almost impossible to know which analyses were actually performed.

The first problem lies in the absence of an accurate description of the method used: reference articles are rarely cited, and abbreviations sometimes do not match the name of the method. For instance, 'PCA' is used to refer to principal coordinates analysis (PCoA) in Pariset *et al.* (2003). Such confusion adds to the ambiguities that already exist between some methods, such as those between PCoA and NMDS. PCoA is also sometimes called 'metric dimensional scaling' (MDS), while NMDS is indifferently abbreviated MDS or NMDS (Legendre and Legendre, 1998). This is all the more confusing since PCoA is routinely used to initialize the algorithm of NMDS (Baker and Moeed, 1987). Papers demonstrating an ambiguity between PCoA and NMDS are not uncommon (for

example, Preziosi and Fairbairn, 1992; Zhivotovsky *et al.*, 2003).

While required, providing a correct reference to a method is usually not sufficient. Some methods exist in different variants, according to the initial transformations of the data. This is particularly true for PCA: while centring (subtracting the mean allele frequency from all observations) is always achieved, scaling of the alleles (dividing each observation by allele-wise values) is optional and can be performed in several ways. Scaling can drastically change the results of a PCA, but is rarely disclosed (for examples, Mitton, 1978; MacHugh *et al.*, 1998; Grivet *et al.*, 2008). In PCoA and NMDS, the genetic distance employed should always be specified, and in the case of NMDS, how the algorithm was initialized should be indicated. An example of such application can be found in Baker and Moeed (1987), who used a NMDS initialized by a PCoA of Rogers' distances of allozyme data to explore the genetic variation among populations of common minas (*Acridotheres tristis*). Lack of accuracy in the description of the method always complicates interpretation of the results, and sometimes brings their validity into question. For instance, some papers show principal components of a PCA that were clearly not centred (their range of variation did not include zero), which indicates an error in the computations of the analysis and invalidates the results (for instance, MacHugh *et al.*, 1997, 1998; Pariset *et al.*, 2003). Moreover, it is difficult to ascertain precisely where the problem came from, as the software used for the computation was not mentioned in these publications.

Making graphics

Another classical problem lies in the graphical display of results. As mentioned previously, the screeplot is the basic tool used to assess which principal components should be interpreted, but it is most often omitted in publications. The amount of inertia associated with each principal component is often indicated, but this information is complementary to the screeplot and cannot be used as a substitute. For instance, in their study of the genetic differentiation among different yak (*Poephagus grunniens*) populations, Xuebin *et al.* (2005)

presented a scatterplot of PCA displaying 80% of the whole variability, but this scatterplot was merely uninformative in terms of genetic differentiation. Conversely, two principal components of PCA containing less than 10% of total inertia provided insights about the phylogeny of different maize subspecies in Matsuoka *et al.* (2002). When used alone, the amount of inertia can therefore be a misleading criterion for choosing the principal components to interprete (see 'Interpreting genetic structures').

Another widespread custom is the use of 3-dimensional scatterplots (for example, van Pijlen *et al.*, 1995; Xuebin *et al.*, 2005). Although these representations add a fancy touch to multivariate analyses, they also have the unfortunate effect of sacrificing the mathematical properties of an analysis, and thus its interpretability. By definition, principal axes and the associated principal components provide the best possible planar representation of the data. If three principal components are retained, their representation requires two factorial planes, with one axis being redundant. Scatterplots in 3-dimensions are ultimately always viewed on a screen or on a sheet of paper, and are thus re-projections of three principal components in two dimensions. The obtained representations are necessarily worse than the true representation of principal components because they no longer have maximum inertia nor orthogonality. Hence, 3-dimensional visualization should be restricted to interactive data analysis (where it can be useful), and is better avoided in publications.

Apart from these pitfalls common to every multivariate analyses, some specific issues also arise when certain methods are applied to genetic markers.

Some specific issues

A first particular issue concerns the use of CA. This method is appropriate for the analysis of a contingency table, that is, a matrix of positive integers (Greenacre, 1984), and is thus appropriate for the analysis of a table of allele counts. A nice example of such an application is provided by She *et al.* (1987), who used a CA of allozyme data to investigate the genetic differentiation between populations of teleost fishes. Interestingly, this study also showed that 'correspondences'

highlighted by CA can reflect linkage disequilibrium existing between alleles. In some cases, CA has been used for allele (relative) frequencies (Li *et al.*, 2002), which has been proven to significantly alter the results of the analysis (Perrière and Thioulouse, 2002). In such a case, it seems much more appropriate to use PCA or PCoA. However, even when CA is correctly used, one should be aware that scarce descriptors are given a stronger weight by the chi-squared distance, which is optimised by the analysis (Legendre and Legendre, 1998, p.285). The typical consequence is that a population possessing a rare allele would appear as an outlier in CA components. Simple simulations show that such an artifactual pattern arises even when studying groups of genotypes randomly chosen from the same population (results not shown). A way to avoid this problem is to remove rare alleles from the data prior to the CA, although this solution requires some investigations regarding which frequency should be considered as 'rare' from the point of view of the CA.

A second specific issue occurs in DA. This method finds principal components maximizing the variance between populations while keeping the variance inside populations constant, assuring optimal discrimination between the populations (Krzanowski and Marriott, 1995, pp.1-56). However, this method involves computation of the Mahalanobis metric (Beharav and Nevo, 2003), which is the inverse of the matrix of covariances between alleles. For this inverse to exist, the covariance matrix must be of full rank, *i.e.*, of rank p if there are p alleles (Harville, 1997, p.80). This is never the case for allele frequencies: each marker spans a space of at most one dimension less than the number of its alleles because any frequency is entirely defined by all the others. That is, if there are k markers, the rank of the covariance matrix is at most $\min(p-k, n)$. Thus, the discriminant analysis can only be performed on a matrix of allele frequencies after removing a given number of alleles, and assuring that there are more objects (genotypes or populations) than alleles. In fact, the number of objects n should be consequently larger than the number of alleles p: Williams and Titus (1988) reported that n should be at least three times larger than p for DA to yield reliable results. Multicollinearity can also exist among alleles (*i.e.*,

when alleles are correlated), especially when linkage disequilibrium occurs. In these cases, the Mahalanobis metric is said to be ill-conditionned, resulting in numerical instability. As a result, principal axes and principal components of DA cannot be computed with accuracy, and small changes in allele frequencies induce large changes in the results (Seber, 1977, pp.319-322). As a consequence, the alleles used in DA should be carefully selected before performing the analysis. An empirical approach consists of retaining only the most frequent allele of each locus (Sagnard *et al.*, 2002), but this does not ensure that the subset of alleles obtained is optimal with regard to discrimination. In fact, it does not even ensure that the multicollinearity problem is solved, since linkage disequilibrium can still exist between the most frequent alleles. One can preferentially use statistical approaches that are especially devoted to the selection of variables in DA (Lachenbruch and Goldstein, 1979), where such approaches proved useful for selecting a subset of best discriminating alleles (Fahima *et al.*, 1999; Beharav and Nevo, 2003). However, investigations should be carried out to assess whether a particular variable selection procedure is preferable to the others in the case of allele frequencies.

Interpreting genetic structures

A major concern in multivariate analysis of genetic markers lies in interpreting the results. This issue can be illustrated by examining one case of misinterpretation, raising the question of which result of a multivariate analysis could be interpreted as genetic structuring.

In the controversy regarding the relevance of definiting human races based on genetic information (Lewontin, 1978; Mitton, 1978; Powell and Taylor, 1978), Mitton (1978) argued that genetic differentiation between 'human races' was important because they clearly appeared as distinct groups on the factorial map of a PCA. This misinterpretation of the results is related to the common mistake of not displaying the screeplot of the analysis along with the values of inertia associated with principal components. Ordinations in reduced space do not summarise the essential part of the genetic variability: they attempt to show as

much genetic variability as possible into a few axes, which is different. Mitton (1978) showed that 'racial' groups were well separated on the factorial plane, and that two principal components were sufficient to assign each population to a given group. However, this did not contradict the well-acknowledged fact that the genetic variability within 'races' is much larger than between 'races' (Edwards, 2003), as suggested by the author. For example, it would be possible to perfectly discriminate two populations using only one allele, but this allele may represent only 1% of the variability of a dataset containing 100 alleles. This point was discussed by Edwards (2003), who emphasised the fundamental difference between being able to assign genotypes to taxonomic groups, and observing larger genetic variability between these taxonomic groups.

We can ask what criterion the principal components of an analysis should meet to be considered as true genetic structuring. The relative amount of inertia cannot be used as a single criterion, because it depends directly on the number of alleles considered. As stressed previously, the screeplot can be used to assess which principal conponents likely contain interesting structures. Recently, Patterson *et al.* (2006) tested the significance of the eigenvalues from a PCA of genetic markers to infer population stratification. Another testing approach to select interpretable principal components of a PCA has been proposed by Dray (2008), and could also be used to identify significant genetic structures. Note that both approaches are reserved to PCA (Patterson *et al.*, 2006; Dray, 2008), and it would be valuable to extend these tests to other multivariate methods. Another way of assessing relevant genetic structures emerging from an ordination method is to quantify the amount of genetic differentiation contained in the principal components. The main difficulty is then identifying clusters of genotypes from the principal components retained. This can be achieved using a given clustering algorithm (Legendre and Legendre, 1998, pp.303-381), such as the unweighted arithmetic average clustering (UPGMA, Rohlf, 1963). It is then possible to measure the amount of genetic differentiation between the obtained clusters of genotypes using classical approaches like the F_{ST}. Note that the obtained statistics can only be used to quantify genetic differentiation, but not

to test it, because the principal components are by definition, optimised with regard to some measurement of genetic differentiation. To conclude this point, the identification of interpretable structures is a major question in multivariate analysis, and is of particular interest when seeking genetic structures from molecular markers.

As we have seen, the application of multivariate analysis to genetic markers can be improved by avoiding a number of pitfalls. However, further improvements can be gained by adapting multivariate methods to several particularities of genetic markers.

Respecting the very nature of data

Scaling in PCA

In many cases, genetic markers are analysed as allele frequencies, which are subjected to a PCoA or a PCA. PCoA is usually well-suited to genetic markers because several genetic distances can be used to summarize the genetic variability. In this case, it is necessary to use an Euclidean distance like Roger's (Weir, 1996, p.197), so that genetic relationships among entities can be fully represented in a plane, and to choose a distance whose underlying model best matches the data (see for instance Weir, 1996, pp.190-198). In the case of PCA, attention must be devoted to the transformations of data: if centring of allele frequencies is almost mandatory, the scaling of allele frequencies can be discussed. The general reason for scaling is to compensate for trivial differences that occur in the variance of the descriptors, for instance, when descriptors are expressed in different units. A reason for not scaling allele frequencies is that doing so is not necessary (scales of variation are inherently the same for every allele), and could mask differences in the genetic variability contained by informative and non-informative markers, ultimately hiding structures in the data. Nonetheless, one good argument for scaling allele frequencies would be to compensate for differences in variance among alleles due to their underlying

binomial nature: the theoretical variance associated with the j^{th} allele frequency, f_j ($j = 1, \cdots, p$ where p is the total number of alleles), is proportional to $f_j(1 - f_j)$. The result is that the variance of an allele frequency is expected to be 'naturally' larger for frequencies close to 0.5, and smaller for frequencies close to 0 or 1. The PCA seeking linear combinations of alleles with maximum variance, alleles with frequencies closer to 0.5 would be favoured by the analysis, while not necessarily reflecting a genetic structure. One way to correct for this is to divide f_j by $\sqrt{f_j(1 - f_j)}$, as has been previously proposed (Cavalli-Sforza *et al.*, 1994, pp.41-42). Mulley *et al.* (1979) used a related standardisation of allele frequencies, which does not amount to unit theoretical variance, but accounts for the number of genotypes used to compute frequencies in each population. Interestingly enough, the variance between populations of the allele frequency standardised by $\sqrt{f_j(1 - f_j)}$ is exactly the classical F_{ST} (Weir, 1996, p.166). Therefore, the between-class PCA (Dolédec and Chessel, 1987), which maximizes the variance between populations, would yield principal components with maximum F_{ST} if performed on allele frequencies centred to a mean of zero and scaled by $\sqrt{f_j(1 - f_j)}$. Even though between-class PCA has only recently been applied to genetic markers by Parisod and Christin (2008) and Jombart (2008, presented as 'inter-class PCA'), this method seems promising for investigating genetic differentiation between groups of genotypes.

Compositional data

The principal particularity of allele frequencies may be that they are sets of compositional data, that is, data with a constant sum for each population and locus. This feature induces non-independence between allele frequencies inside each locus (one frequency can always be deduced from the others), and has several consequences on ordinations in reduced space. Developments in the multivariate analysis of compositional data were led by the work of Aitchison (Aitchison, 1983, 1999, 2003; Aitchison and Greenacre, 2002), but remained mostly ignored in genetics, apart from a few exceptions (Romano *et al.*, 2003; Reyment, 2005). As stressed before, allele frequencies at a given locus are not

independent, since one can be entirely deduced from the others. Populations described by p_j alleles at the j^{th} locus are not embedded inside a p_j-dimensional space, but are instead inside a space whose maximum dimensionality is $(p_j - 1)$, known as a *simplex space* (Aitchison, 2003, pp.24-28). A variety of problems can occur when directly computing an ordination in reduced space in the simplex space (or in a set of simplex spaces in the case of several loci), like the impossibility of identifying structures that are intrinsically non-linear and numerical instability of principal components. The solutions proposed to account for these problems rely on transforming frequencies (mostly using logarithms) and performing a classical analysis like PCA of the obtained data. Reyment (2005) showed that the results of PCA could be strikingly improved by such practices, even when considering a simple log transformation of the data. Henceforth, these approaches should be considered when analysing allele frequencies.

Diversity inside the diversity

A portion of the literature in conservation biology stresses the idea that different genetic markers can provide different information about the genetic diversity of a set of populations (Moazami-Goudarzi and Laloë, 2002). In fact, genetic markers are usually taken as a whole to seek a global, common typology of individuals or populations, without trying to assess if such a common typology exists. There are, however, good reasons for this typology not to occur, the first being that selection can affect different loci in different ways. If this is obvious for selected markers like allozymes, it can also be true for supposedly neutral markers that are physically linked to selected regions of the genome. Interestingly, the first studies linking the genetic variability in allozymes to environmental features analysed each locus separately by PCA (Johnson *et al.*, 1969; Johnson and Schaffer, 1973).

To tackle the question of the typological coherence of genetic markers, the locus must be considered as the unit of analysis. In this perspective, if there are K markers, K analyses should be performed and compared. A class of

multivariate analyses, called the K-table methods (Dray *et al.*, 2007), is devoted to this particular task. Such methods were introduced in genetics by Laloë *et al.* (2007), who used the multiple co-inertia analysis (Chessel and Hanafi, 1996) to compare the typological information provided by different microsatellites. This study showed that microsatellites could provide different pictures of the genetic diversity among populations: while some microsatellites reveal the entire genetic structure, some perceive only particular aspects of the genetic diversity, and others are simply not informative in terms of genetic differentiation patterns. The *typological value* of a marker can be used to quantify the extent to which this marker contributes to displaying a particular genetic structure (Laloë *et al.*, 2007). The application of K-table approaches to genetic markers was further developed by Pavoine and Bailly (2007), who introduced other K-table methods coupled with a multivariate analysis of the biodiversity (Pavoine *et al.*, 2004). Their results confirmed the fact that summing the information coming from different genetic markers, as is usually performed for ordinations in reduced space, does not always provide the most accurate picture of the biodiversity. Note that if K-table methods can suggest that loci experience different selective pressures, they cannot be used as a direct test for these differences. In fact, K-table approaches are first and foremost designed to identify common typologies, and not discrepancies, among a set of markers.

If K-table methods are more complex tools than single-table analyses, their use in genetics should be considered with attention. Note that the linkage of multilocus genetic information to environmental features like in Johnson *et al.* (1969) still raises challenging questions in terms of data analysis: How can we describe the genetic-environment relationships at several loci? What are the different patterns of adaptation among loci?

Linking genetic markers to other data

One of the greatest applications of ordinations in reduced space is in the linkage of genetic markers to other types of data (Johnson *et al.*, 1969; Taylor and Mitton,

1974; Mulley *et al.*, 1979; Barker *et al.*, 1986; Jarraud *et al.*, 2002). This is typically the case in the study of genotype-environment relationships, where multivariate methods can be used to investigate correlations between genetic data and environmental features (Johnson *et al.*, 1969; Mulley *et al.*, 1979). Another application of such approach is to relate genetic information to phenotypic data (for instance, Jarraud *et al.*, 2002). Note that when patterns of selection are being investigated, the genetic diversity should be inferred from non-neutral rather than neutral markers. Various methods are available for coupling two different kinds of information, some of which have been introduced into population genetics. These can be divided into two categories, depending on whether they treat both types of information symmetrically or not. Note that approaches like DA and between-class PCA are also methods for coupling genetic markers with a different information (some partitions of individuals). However, because their aim is very different from the methods presented below (their purpose is to investigate the genetic differentiation between groups of genotypes), DA and between-class PCA are not presented in this section.

Asymmetric methods

The first type of method is formed by *constrained ordinations*, which are devoted to investigating the variability in one dataset that can be explained by another dataset. This is achieved by a multivariate regression of a 'response' dataset onto an 'explanatory' dataset (Ter Braak, 1986). These methods are thus asymmetric, in that the variability in one dataset is explained by another. There are two main techniques in this context: redundancy analysis (RDA, Rao, 1964), which is a constrained version of PCA, and canonical correspondence analysis (CCA, Ter Braak, 1986), which is based on CA. RDA and CCA therefore inherit their properties from PCA and CA: RDA can be used for allele frequencies, while CCA is more appropriate to analysis of tables of allele counts. Both RDA (Kölliker *et al.*, 2008) and CCA (Angers *et al.*, 1999) have proven useful in population genetics, mostly to investigate the portion of the genetic variability that can be explained by a set of environmental variables. For instance, in Angers

et al. (1999), the CCA revealed that the genetic diversity among a set of brook charr populations (*Salvelinus frontalis*) was mainly driven by the structure of the hydrographic network and by a few environmental variables. Another interest of this study is that analyses were led at two different levels, to study the effects of hydrographic and environmental features on the genetic diversity inside, and between populations.

Like discriminant analysis, RDA and CCA involve computation of the Mahalanobis metric which is, in this case, the matrix of covariances between explanatory variables (Legendre and Legendre, 1998). These analyses therefore require that the number of explanatory variables (for instance, environmental variables) be fairly lower than the number of studied objects (genotypes or populations) in order to be computable. Following the previously cited study of Williams and Titus (1988) concerning DA, we can recommend that the number of objects should be at least three times larger than the number of explanatory variables. RDA and CCA also demand that the explanatory variables are reasonably uncorrelated to achieve numerical stability and interpretability of the results. As a rule of thumb, we could suggest to avoid correlations greater than 0.7, so that no more than one half of the variability of any predictor could be explained by another predictor (*i.e.*, $R^2 < 0.5 \Leftrightarrow r < \sqrt{0.5} \simeq 0.7$). Note that genetic markers could also be used as explanatory variables, for example, with an 'explained' dataset of phenotypic traits. In such cases, the dimension of the genetic information should be reduced, either by applying a standard variable selection procedure (for example, forward selection) to the allele frequencies, or by reducing the genetic data to a few principal components using PCA or PCoA.

When the above conditions are respected, constrained ordinations can be efficiently used to explain one kind of variability by another. However, when the purpose of a study is to investigate common patterns of variability in two datasets, or when RDA and CCA cannot be used for technical reasons, an alternative can be found in certain symmetric approaches.

Symmetric methods

Symmetric methods allow one to study the structures common to two datasets by treating the two types of information similarly. They differ from constrained ordinations in the same way that linear regression differs from correlation. Symmetric approaches include canonical correlation analysis (CCorA, Hotelling, 1936; Takeuchi *et al.*, 1984, pp.225-280) and co-inertia analysis (COA, Dolédec and Chessel, 1994; Dray *et al.*, 2003a). CCorA was introduced by Johnson and Schaffer (1973) in order to describe and test the correlations between allele frequencies in allozymes and a set of environmental features. The principle of this technique is to find two sets of orthogonal axes (one for each dataset), such that the obtained pairs of principal components have a maximum squared correlation (Takeuchi *et al.*, 1984, pp.225-229). It is worth noting that Johnson and Schaffer (1973) were following another pioneering work (Johnson *et al.*, 1969) in which the same authors used correlations between principal components of two PCAs (one of allele frequencies, one of environmental variables) to test genetic-environment relationships. A series of subsequent papers provided remarkable illustrations of the insights that CCorA can bring to the study of adaptation (Schaffer and Johnson, 1974; McKechnie *et al.*, 1975; Mulley *et al.*, 1979). A nice example is provided by Mulley *et al.* (1979), which used the CCorA to investigate patterns of adaptation in populations of *Drosophila buzzatii*. The authors have shown that the allelic variation observed at some allozyme loci was significantly correlated to climate descriptors, which strongly suggested the existence of local adaptations in these populations. A recurrent problem in these studies is that gene-flow can act as a confounding effect when assessing genetic-environment correlations. Schaffer and Johnson (1974) addressed this issue by regressing allele frequencies onto spatial coordinates prior to the analysis, and hence removing linear spatial trends from the data. Note that more efficient methods of removing spatial patterns have since been developed, some of which are described in the next section.

A typical problem in CCorA is that, like RDA and CCA, it requires to compute the Mahalanobis metric of both datasets: it cannot be used when

there are more descriptors than studied objects and it requires descriptors to be uncorrelated to yield interpretable results. In some of these cases, a CCorA can still be performed after selecting a small subset of uncorrelated variables (for example, Mulley *et al.*, 1979). A common criticism of CCorA is that pairs of principal components with maximum squared correlation could have a very small variance, and therefore have in general no real biological meaning (Taylor and Mitton, 1974). Taylor and Mitton (1974) suggested that a symmetric analysis should yield pairs of principal components reflecting both a fair amount of variance and be correlated with each other, that is, reflecting common parts of the variability in the two datasets. This is the definition of a method developed later in ecology: the co-inertia analysis (Dolédec and Chessel, 1994; Dray *et al.*, 2003a).

COA has been imported into genetics to relate the genetic variability of several bacterial strains to the expression of toxin genes (Jarraud *et al.*, 2002). It is worth noting that COA is closely related to Procustean analysis (Dray *et al.*, 2003b), which has been proposed for the analysis of genetic markers coupled to other kinds of information (Cavalli-Sforza *et al.*, 1994, p.41), although we were unable to find any applications of this technique to genetic markers. COA finds two sets of principal axes (one for each dataset), such that the pairs of principal components have a maximum squared covariance (*i.e.*, co-inertia). This criterion is particularly interesting since it amounts to maximizing the product of the variances of each principal component and their squared correlations (because $\text{cov}^2(a,b) = \text{var}(a)\text{var}(b)\text{cor}^2(a,b)$). Interestingly, the COA does not require inversion of a covariance matrix; consequently, it does not require the number of descriptors to be lower than the number of objects and it is not hampered by correlations among the descriptors. Moreover, COA relies on a modification of two separate analyses, each of which can be, for instance, a PCA, a PCoA, or a CA. For example, Jarraud *et al.* (2002) employed the co-inertia between a PCoA of a genetic distance matrix derived from AFLP markers and a PCA of distributions of toxin genes in several strains of *Staphylococcus aureus* to assess the evolution of virulence factors with respect to the genetic background of the

strains. The COA appears to be a good alternative to RDA, CCA, and CCorA when these methods cannot be applied for the reasons described above. In other cases, the COA may still be favored whenever the squared covariance criterion is more satisfying than criteria used by other analyses, that is, when one is interested in identifying common patterns of variation between two different sources of information.

Spatial multivariate analysis

Many population genetics studies in which multivariate analyses were used involve georeferenced data. When processes related to gene-flow are being investigated– which may be the most common case –spatial genetic patterns are researched in neutral markers (for example, Menozzi *et al.*, 1978; Cavalli-Sforza *et al.*, 1993). In contrast, when non-neutral markers are used to infer patterns of adaptation, spatial structures induced by gene flow can act as a confounding effect that would have to be removed (Schaffer and Johnson, 1974). As noted by Mulley *et al.* (1979), the drawback of this strategy is that '*if environmental factors with selective effects are stronly correlated to geographic location, adjustement for location may remove a major fraction of the selective effects*'. In such case, it would be worthwhile to compare the selective effects detected with and without removing the effects of spatial patterns. Spatial information can be used in multivariate analysis of genetic markers, to investigate the part of the genetic variability that is or is not spatially structured.

Unfortunately, the methods commonly used to investigate spatial genetic patterns almost never take spatial information into account explicitly, *i.e.*, they do not incorporate spatial information as a component of the criterion optimised by the analysis (Jombart *et al.*, 2008). This contrasts with other methodological frameworks such as analysis of molecular variance (Excoffier *et al.*, 1992) or Bayesian clustering (Pritchard *et al.*, 2000), in which spatially explicit methods are used (respectively, Dupanloup *et al.*, 2002; François *et al.*, 2006). However,

spatial ordinations exist and are widely used in other domains, the closest to genetics being ecology. It is therefore not surprising that spatial ordinations were first proposed to analyse genetic markers in vegetation sciences (Escudero *et al.*, 2003) and landscape genetics (Grivet *et al.*, 2008).

Recently, Grivet *et al.* (2008) used the canonical trend surface analysis (Wartenberg, 1985) to detect spatial patterns using microsatellite markers. This approach relies on performing a CCorA to identify correlations between genetic and spatial data. Grivet *et al.* (2008) used polynomials of spatial coordinates as spatial predictors, while this approach was criticised in ecology (Borcard and Legendre, 2002; Dray *et al.*, 2006), mainly because the obtained variables are generally correlated and can only model broad-scale patterns. Other spatial predictors, Moran's eigenvectors, are now used in ecology (Dray *et al.*, 2006; Griffith and Peres-Neto, 2006). Contrary to polynomials of spatial coordinates, these spatial predictors are uncorrelated, and can model spatial patterns on a wide range of scales. To reveal spatial genetic patterns, Moran's eigenvectors can be used as explanatory variables in a CCA or a RDA of genetic markers. In studies in which spatial structures need to be removed to infer adaptations, Moran's eigenvectors could also be used as covariables in partial RDA or partial CCA (Legendre and Legendre, 1998, pp.769-779).

To our knowledge, the only spatial ordination developed within the genetic framework is spatial principal component analysis (sPCA, Jombart *et al.*, 2008). This method relies on a modification of PCA such that not only the variance of the principal components, but also their spatial autocorrelation, is optimised. Jombart *et al.* (2008) identified various kinds of spatial structuring that can arise in genetic data, and showed that sPCA can be efficiently used to reveal these patterns. In particular, comparison between PCA and sPCA demonstrated that sPCA should be preferred to PCA whenever spatial genetic patterns are researched. Note that a similar approach was developed in the vegetation sciences by Dray *et al.* (2008), who developed a spatial version of CA. While the sPCA is devoted to investigating spatial genetic patterns in allele frequencies, the approach of Dray *et al.* (2008) could be used to study spatial genetic patterns in allele counts.

Perspectives and conclusion

One important observation emerging from this review is that if multivariate analyses proved useful for extracting biological information from genetic markers, their application could sometimes benefit from more rigorous practices. Methods should always be referred to clearly and with a distinction between the method itself and its implementation. An accurate description of an ordination in reduced space would include all data transformations such as centring and scaling in PCA, the chosen distance in PCoA and NMDS, the selection of alleles in DA, or algorithm initialization in NMDS. To facilitate reproducibility, free and script-based software should be favoured over other software. In this context, the R software (R Development Core Team, 2008) is clearly an appealing choice: in addition to allowing exact reproducibility, it provides an interface between a large number of implemented multivariate methods (for example, Chessel *et al.*, 2004; Dray *et al.*, 2007) and genetic marker data (Jombart, 2008), in addition to supporting the usual population genetics tools (Warnes, 2003; Goudet, 2005). From a more theoretical point of view, it seems important to further investigate the relationships between multivariate methods and genetic models. A step in this direction has been made by Patterson *et al.* (2006), who applied recent developments in statistics (summarized in Soshnikov and Fyodorov, 2005) to infer the number of populations in a set of genotypes and define a threshold for genetic structuring to be detectable by PCA.

More generally, several multivariate analyses developed in other disciplines can be adapted to search biological structures within genetic markers. This is clearly the case in spatial genetics, where constrained ordinations based on Moran's eigenvectors (Dray *et al.*, 2006) could be used to investigate or correct for spatial genetic structures. It is also true for K-table methods, which were only recently introduced into population genetics (Laloë *et al.*, 2007; Pavoine and Bailly, 2007), and open appealing perspectives for the study of the genetic diversity. These methods can also be used to investigate common patterns of variation inferred from genetic markers and other sources of information, like biological traits and environmental features. As noted by Patterson *et al.* (2006),

multivariate analysis can analyse larger datasets than other usual approaches such as the Bayesian clustering, and thus represents a relevant approach to extracting information from huge datasets produced by the detailed mapping of genetic variation for a large number of genotypes. This is the case, for instance, with the '1000 Genomes' project (`http://www.1000genomes.org/`), which aims at sequencing one thousand human genotypes in order to provide high-resolution information that is directly valuable for disease studies. Promisingly, a wide range of questions are raised by or through genetic markers, some of which can currently be solved by existing methods. Some of these questions will undoubtedly require specific developments in which multivariate models will have to closely match the genetic concerns, which makes the multivariate analysis of genetic markers a whole area of research in biometry.

Aknowledgement

We are very grateful to Christian Biémont, Sébastien Devillard, F. Stephen Dobson, and Gilles Yoccoz for providing constructive comments on an earlier version of the manuscript.

Aitchison J (1983). Principal component analysis of compositional data. *Biometrics* **70**: 57–65.

Aitchison J (1999). Logratios and natural laws in compositional data analysis. *Mathematical Geology* **31**: 563–589.

Aitchison J (2003). *The statistical analysis of compositional data*. The Blackburn Press.

Aitchison J, Greenacre M (2002). Biplot of compositional data. *Journal of the Royal Statistical Society Series C, Applied statistics* **51**: 375–392.

Angers B, Plante M, Bernatchez L (1999). Canonical correspondence analysis for estimating spatial and environmental effects on microsatellite gene diversity in brook charr (*Salvelinus fontinalis*). *Molecular Ecology* **8**: 1043–1053.

Baker AJ, Moeed A (1987). Rapid genetic differentiation and founder effect in colonizing populations if common mynas (*Acridotheres tristis*). *Evolution* **41**: 525–538.

Barker JSF, East PD, Weir BS (1986). Temporal and microgeographic variation in allozyme frequencies in a natural population of *Drosophila buzzatii*. *Genetics* **112**: 577–611.

Beharav A, Nevo E (2003). Predictive validity of discriminant analysis for genetic data. *Genetica* **119**: 259–267.

Bertranpetit J, Cavalli-Sforza LL (1991). A genetic reconstruction of the history of the population of the Iberian Peninsula. *Annals of Human Genetics* **55**: 51–67.

Borcard D, Legendre P (2002). All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecological Modelling* **153**: 51–68.

Cavalli-Sforza LL (1966). Population structure and human evolution. *Proceedings of the Royal Society of London Series B* **164**: 362–379.

Cavalli-Sforza LL, Menozzi P, Piazza A (1993). Demic expansions and human evolution. *Science* **259**: 639–646.

Cavalli-Sforza LL, Menozzi P, Piazza A (1994). *The history and geography of human genes*. Princeton University Press.

Chessel D, Dufour AB, Thioulouse J (2004). The ade4 package-I- one-table methods. *R News* **4**: 5–10.

Chessel D, Hanafi M (1996). Analyse de la co-inertie de K nuages de points. *Revue de statistique appliquée* **XLIV (2)**: 35–60.

Ciofi C, Wilson GA, Beheregaray LB, Marquez C, Gibbs JP, Tapia W, *et al.* (2006). Phylogeographic history and gene flow among giant galápagos tortoises on southern Isabela Island. *Genetics* **172**: 1727–1744.

Cox RF, Cox MAA (2001). *Multidimensional scaling*. Chapman & Hall/CRC.

Dolédec S, Chessel D (1987). Rythmes saisonniers et composantes stationnelles en milieu aquatique. I. description d'un plan d'observation complet par projection de variables. *Acta Oecologica, Oecologia Generalis* **8**: 403–426.

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

Dray S (2008). On the number of principal components: A test of dimensionality based on measurements of similarity between matrices. *Computational statistics & data analysis* **52**: 2228–2237.

Dray S, Chessel D, Thioulouse J (2003a). Co-inertia analysis and the linking of ecological tables. *Ecology* **84**: 3078–3089.

Dray S, Chessel D, Thioulouse J (2003b). Procrustean co-inertia analysis for the linking of multivariate datasets. *Ecoscience* **10**: 110–119.

Dray S, Dufour AB (2007). The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software* **22(4)**: 1–20.

Dray S, Dufour AB, Chessel D (2007). The ade4 package - II: Two-table and K-table methods. *R News* **7**: 47–54.

Dray S, Legendre P, Peres-Neto P (2006). Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbours matrices (PCNM). *Ecological Modelling* **196**: 483–493.

Dray S, Saïd S, Debias F, Chessel D (2008). Spatial ordination of vegetation data using a generalization of Wartenberg's multivariate spatial correlation. *Journal of Vegetation Science* **19**: 45–56.

Dupanloup I, Schneider S, Excoffier L (2002). A simulated annealing approach to define the genetic structure of populations. *Molecular Ecology* **11**: 2571–2581.

Edwards AWF (2003). Human genetic diversity: Lewontin's fallacy. *BioEssays : news and reviews in molecular, cellular and developmental biology* **25**: 798–801.

Escoufier Y (1987). The duality diagramm : a means of better practical applications. In: Legendre P, Legendre L (eds.), *Development in numerical ecology*, NATO advanced Institute , Serie G .Springer Verlag, Berlin, pp. 139–156.

Escudero A, Iriondo JM, Torres ME (2003). Spatial analysis of genetic diversity as a tool for plant conservation. *Biological Conservation* **113**: 351–365.

Excoffier L, Smouse PE, Quattro JM (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: applications to human mitochondrial DNA restriction data. *Genetics* **131**: 479–491.

Fahima T, Sun GL, Beharav A, Krugman T, Beiles A, Nevo E (1999). RAPD polymorphism of wild emmer wheat populations, *Triticum dicoccoides*, in Israel. *Theoretical and Applied Genetics* **98**: 434–447.

Falush D, Stephens M, Pritchard JK (2003). Inference of population structure using multilocus genotype data : linked loci and correlated allele frequencies. *Genetics* **164**: 1567–1587.

Fisher RA (1952). Statistical methods in genetics. *Heredity* **6**: 1–12.

François O, Ancelet S, Guillot G (2006). Bayesian clustering using hidden markov random fields in spatial population genetics. *Genetics* **174**: 805–816.

Geffen E, Anderson MJ, Wayne RK (2004). Climate and habitat barriers to dispersal in the highly mobile grey wolf. *Molecular Ecology* **13**: 2481–2490.

Goudet J (2005). HIERFSTAT, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Notes* **5**: 184–186.

Gower JC (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* **53**: 325–338.

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

Griffith D, Peres-Neto P (2006). Spatial modeling in ecology: the flexibility of eigenfunction spatial analyses. *Ecology* **87**: 2603–2613.

Grivet D, Sork VL, Westfall RD, Davis FW (2008). Conserving the evolutionary potential of california valley oak (*Quercus lobata* Née): a multivariate genetic approach to conservation planning. *Molecular Ecology* **17**: 139–156.

Guinand B (1996). Use of a multivariate model using allele frequency distributions to analyse patterns of genetic differentiation among populations. *Biological Journal of the Linnean Society* **58**: 173–195.

Guinand B, Bouvet Y, Brohon B (1996). Spatial aspects of genetic differentiation of the European chub in the Rhone River basin. *Journal of Fish Biology* **49**: 714–726.

Hanotte O, Bradley DG, Ochieng JW, Verjee Y, Hill EW, Rege JEO (2002). African pastoralism: genetic imprints of origins and migrations. *Science* **296**: 336–339.

Harville DA (1997). *Matrix algebra from a statistician's perspective*. Springer, New York.

Hotelling H (1936). Relations between two sets of variables. *Biometrika* **28**: 321–327.

Jambu M (1991). *Exploratory and multivariate data analysis*. Academic Press, Inc.

Jarraud S, Mougel C, Thioulouse J, Lina G, Meugnier H, Forey F, *et al.* (2002). Relationships between *Staphylococcus aureus* genetic background, virulence factors, *agr* groups (alleles), and human disease. *Infection and Immunity* **70**: 631–641.

Johnson FM, Schaffer HE (1973). Isozyme variability in species of the genus drosophila. VII. Genotype-environment relationships in populations of *D. melanogaster* from the eastern United States. *Biochemical Genetics* **10**: 149–163.

Johnson FM, Schaffer HE, Gillaspy JE, Rockwood ES (1969). Isozyme genotype-environment relationships in natural populations of the harvester ant, *Pogonomyrmex barbatus*, from Texas. *Biochemical Genetics* **3**: 429–450.

Jombart T (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**: 1403–1405.

Jombart T, Devillard S, Dufour AB, Pontier D (2008). Revealing cryptic spatial patterns in genetic variability by a new multivariate method. *Heredity* **101**: 92–103.

Kölliker R, Bassin S, Schneider D, Widmer F, Fuhrer J (2008). Elevated ozone affects the genetic composition of *Plantago lanceolata* L. populations. *Environmental Pollution* **152**: 380–386.

Krzanowski WJ, Marriott FHC (1995). *Multivariate analysis. Part 2: classification, covariance structures and repeated measurements*. Halsted Press, and John Wiley & Sons.

Lachenbruch PA, Goldstein M (1979). Discriminant analysis. *Biometrics* **35**: 69–85.

Laloë D, Jombart T, Dufour AB, Moazami-Goudarzi K (2007). Consensus genetic structuring and typological value of markers using multiple co-inertia analysis. *Genetics Selection Evolution* **39**: 545–567.

Lebart L, Morineau A, Piron M (2004). *Statistique exploratoire multidimensionnelle*. DUNOD.

Legendre P, Anderson DJ (1999). Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. *Ecological Monographs* **69**: 1–24.

Legendre P, Legendre L (1998). *Numerical ecology*. Elsevier Science B. V., Amsterdam.

Lessa EP (1990). Multidimensional analysis of geographic genetic structure. *Systematic Zoology* **39**: 242–252.

Lewontin RC (1978). Single- and multiple-locus measures of genetic distance between groups. *The American Naturalist* **112**: 1138–1139.

Li MH, Zhao SH, Bian C, Wang HS, Wei H, Liu B, *et al.* (2002). Genetic relationships among twelve chinese indigenous goat populations based on microsatellite analysis. *Genetic Selection Evolution* **34**: 729–744.

MacHugh DE, Loftus RT, Cunningham P, Bradley DG (1998). Genetic structure of seven European cattle breeds assessed using 20 microsatellite markers. *Animal Genetics* **29**: 333–340.

MacHugh DE, Shriver MD, Loftus RT, Cunningham P, Bradley DG (1997). Microsatellite DNA variation and the evolution, domestication and phylogeny of taurine and zebu cattle (*Bos taurus* and *Bos indicus*). *Genetics* **146**: 1071–1086.

Matsuoka Y, Vigouroux Y, Goodman MM, Jesus Sanchez G, Buckler E, Doebley J (2002). A single domestication for maize shown by multilocus microsatellite genotyping. *Proceedings of the National Academy of Sciences of the United States of America* **99**: 6080–6084.

McKechnie SW, Ehrlich PR, White RR (1975). Population genetics of euphydryas butterflies. I. genetic variation and the neutrality hypothesis. *Genetics* **81**: 571–594.

Mcrae BH, Beier P, Huynh LY, Keim P (2005). Habitat barriers limit gene flow and illuminate historical events in a wide-ranging carnivore, the American puma. *Molecular Ecology* **14**: 1965–1977.

Menozzi P, Piazza A, Cavalli-Sforza LL (1978). Synthetic maps of human gene frequencies in Europeans. *Science* **201**: 786–792.

Mitton JB (1978). Measurement of differentiation: reply to Lewontin, Powell and Taylor. *The American Naturalist* **112**: 1142–1144.

Moazami-Goudarzi K, Laloë D (2002). Is a multivariate consensus representation of genetic relationships among populations always meaningful? *Genetics* **162**: 473–484.

Moazami-Goudarzi K, Laloë D, Furet JP, Grosclaude F (1997). Analysis of genetic relationships between 10 cattle breeds with 17 microsatellites. *Animal Genetics* **28**: 338–345.

Mulley JC, James W, Barker JSF (1979). Allozyme genotype-environment relationships in natural populations of *Drosophila buzzatii*. *Biochemical Genetics* **17**: 105–126.

Pariset L, Savarese MC, Cappuccio I, Valentini A (2003). Use of microsatellites for genetic variation and inbreeding analysis in Sarda sheep flocks of central Italy. *Journal of Animal Breeding and Genetics* **120**: 425–432.

Parisod C, Christin PA (2008). Genome-wide association to fine-scale ecological heterogeneity within a continuous population of *Biscutella laevigata* (brassicaceae). *New Phytologist* **178**: 436 – 447.

Patterson N, Price AL, Reich D (2006). Population structure and eigenanalysis. *PLoS genetics* **2**: 2074–2093.

Pavoine S, Bailly X (2007). New analysis for consistency among markers in the study of genetic diversity: development and application to the description of bacterial diversity. *BMC Evolutionary Biology* **7**: 156.

Pavoine S, Dufour AB, Chessel D (2004). From dissimilarities among species to dissimilarities among communities: a double principal coordinate analysis. *Journal of Theoretical Biology* **228**: 523–537.

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

Perrière G, Thioulouse J (2002). Use and misuse of correspondence analysis in codon usage studies. *Nucleic Acids Research* **30**: 4548–4555.

Powell JR, Taylor CE (1978). Are human races "substantially" different genetically? *The American Naturalist* **112**: 1139–1142.

Preziosi RF, Fairbairn DJ (1992). Genetic population structure and levels of gene flow in the stream dwelling waterstrider *Aquarius* (= *Gerris*) *remigis* (Emiptera: Geridae). *Evolution* **46**: 430–444.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**: 904–909.

Pritchard JK, Stephens M, Donnelly P (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. **URL:** *http://www.R-project.org*

Rao CR (1964). The use and interpretation of principal component analysis in applied research. *Sankhya, A* **26**: 329–359.

Reyment RA (2005). The statistical analysis of multivariate serological frequency data. *Bulletin of Mathematical Biology* **67**: 1303–1313.

Rohlf FJ (1963). Classification of *Aedes* by numerical taxonomic methods (diptera:culicidae). *Annals of the Entomological Society of America* **56**: 798–804.

Romano V, Calì F, Ragalmuto A, D'Anna RP, Flugy A, De Leo G, *et al.* (2003). Autosomal microsatellite and mtDNA genetic analysis in Sicily (Italy). *Annals of Human Genetics* **67**: 42–53.

Sagnard F, Barberot C, Fady B (2002). Structure of genetic diversity in *Abies ablab* Mill. from southwestern Alps: multivariate analysis of adaptive and non-adaptative traits for conservation in France. *Forest ecology and management* **157**: 175–189.

Sanchez-Mazas A, Langaney A (1988). Common genetic pools between human populations. *Human Genetics* **78**: 161–166.

Schaffer HE, Johnson FM (1974). Isozyme allelic frequencies related to selection and gene-flow hypotheses. *Genetics* **77**: 163–168.

Seal HL (1966). *Multivariate Statistical Analysis for Biologists*. Methuen and co.

Seber GAF (1977). *Linear regression analysis*. John Wiley & Sons.

She JX, Autem M, Kotulas G, Pasteur N, Bonhomme F (1987). Multivariate analysis of genetic exchanges between *Solea aegyptiaca* and *Solea senegalensis* (Teleosts, Soleidae). *Biological Journal of the Linnean Society* **32**: 357–371.

Smouse PE, Spielman RS, Park MH (1982). Multiple-locus allocation of individuals to groups as a function of the genetic variation within and differences among human populations. *The American Naturalist* **119**: 445–463.

Soshnikov A, Fyodorov YV (2005). On the largest singular values of random matrices with independent Cauchy entries. *Journal of Mathematical Physics* **46**: 033302.

Takeuchi K, Yanai H, Mukherjee BN (1984). *The foundations of multivariate analysis: a unified approach by means of projection onto linear subspaces*. Wiley Eastern Limited.

Taylor CE, Mitton JB (1974). Multivariate analysis of genetic variation. *Genetics* **76**: 575–585.

Ter Braak CJF (1986). Canonical correspondence analysis : a new eigenvector technique for multivariate direct gradient analysis. *Ecology* **67**: 1167–1179.

van Pijlen IA, Amos B, Burke T (1995). Patterns of genetic variability at individual minisatellite loci in mike whale *Balaenoptera acutorostrata* populations from three different oceans. *Molecular Biology and Evolution* **12**: 459–472.

Warnes GR (2003). The genetics package. *R News* **3**: 9–13.

Wartenberg DE (1985). Canonical trend surface analysis: a method for describing geographic pattern. *Systematic Zoology* **34(3)**: 259–279.

Weir BS (1996). *Genetic data analysis II*. Sinauer Associates, Sunderland, Massachussetts.

Williams BK, Titus K (1988). Assesment of sampling stability in ecological applications of discriminant analysis. *Ecology* **69**: 1275–1285.

Xuebin Q, Jianlin H, Chekavora I, Badamdorj D, Rege JEO, Hanotte O (2005). Genetic diversity and differentiation of Mongolian and Russian yak populations. *Journal of Animal Breedings and Genetics* **122**: 117–126.

Zhivotovsky LA, Rosenberg NA, Feldman MW (2003). Features of evolution and expansion of modern humans, inferred from genomwide microsatellite markers. *American Journal of Human Genetics* **72**: 1171–1186.

Titles and legends to tables

Table 1: Multivariate analyses applied to genetic markers. Each method is indicated by its most frequent name and abbreviation. The 'criterion' is the quantity optimised by the principal components of the method. The 'Application' column gives the reference of an early and representative publication using the method to analyse genetic markers. [1] AFLP: amplified fragment length polymorphism. [2] SSR: single sequence repeats.

Tables

Table 1

Method	Criterion	Application	Data
Principal component analysis (PCA)	Variance (same as squared Euclidean distances)	(Cavalli-Sforza, 1966)	Allozymes
Principal coordinates analysis (PCoA)	Any Euclidean distance	(Sanchez-Mazas and Langaney, 1988)	Allozymes
Non-metric dimensional scaling (NMDS)	Ordering of objects	(Lessa, 1990)	Rogers' and Nei's distances
Correspondence analysis (CA)	Chi-squared distance	(She *et al.*, 1987)	Allozymes
Discriminant analysis (DA)	Variance between groups / total variance	(Smouse *et al.*, 1982)	Allozymes
Constant-row total multiple correspondence analysis (CRT-MCA)	Correlation ratio	(Guinand, 1996)	Allozymes
Factor analysis (FA)	'Common effect' in allele frequencies	(Taylor and Mitton, 1974)	Allozymes
Canonical correspondence analysis (CCA)	Chi-squared distances in predicted data	(Angers *et al.*, 1999)	Microsatellites
Redundancy analysis (RDA)	Variance of predicted data	(Kölliker *et al.*, 2008)	AFLP[1] and SSR[2]
Canonical correlation analysis (CCorA)	Squared correlation between pairs of scores	(Johnson and Schaffer, 1973)	Allozymes
Co-inertia analysis (COA)	Squared covariance between pairs of scores	(Jarraud *et al.*, 2002)	AFLP[1]
Multiple co-inertia analysis (MCOA)	Squared covariance between a set of scores	(Laloë *et al.*, 2007)	Microsatellites
Spatial principal component analysis (sPCA)	Product of variance and spatial autocorrelation	(Jombart *et al.*, 2008)	Microsatellites

1.5 Organisation du manuscrit

A présent que les termes et l'état du dialogue interdisciplinaire ont été explicités, nous pouvons aborder la présentation de notre contribution méthodologique à l'analyse des données de marqueurs moléculaires. Le chapitre 2 présente les débuts de cette thèse, qui ont été marqués par une collaboration avec deux chercheurs de l'INRA de Jouy-en-Josas, centrée sur la question de la cohérence typologique des marqueurs génétiques. Ceci nous a mené à introduire les méthodes dites K-tableaux en génétique, en proposant d'utiliser l'analyse de coinertie multiple (Chessel & Hanafi, 1996) pour répondre à la question posée. Le chapitre 3 présente l'*analyse en composantes principales spatiales* (ou *spatial principal component analysis*, sPCA), une méthode développée pour analyser la structuration spatiale de la variabilité génétique observée au niveau individuel ou populationnel. Cette méthode constitue une réponse possible à la question de l'identification de la structure génétique des populations naturelles. Nous présentons au chapitre 4 le package *adegenet* pour le logiciel R (R Development Core Team, 2008), dans lequel nous fournissons un ensemble d'outils pour la gestion, la manipulation et l'analyse (essentiellement multivariée) des marqueurs génétiques. Les chapitres suivants constituent les extensions méthodologiques qui, tout en provenant de l'analyse des données génétiques, sortent à des degrés divers de ce contexte. Le chapitre 5 présente l'*analyse de structure multi-échelle* (ou *multi-scale pattern analysis*, MSPA), une méthode permettant d'étudier les principales échelles de la structuration spatiale d'un ensemble de variables. Pour des raisons culturelles, et parce qu'elle est très générale, cette méthode a été développée dans le contexte de l'analyse de données écologiques. Cependant, nous montrerons que la MSPA est aussi pleinement applicable aux données de fréquences alléliques, et peut être utilisée pour explorer la structuration spatiale d'un ensemble de génotypes. Le chapitre 6 montre que l'on peut étendre la sPCA à l'analyse des structures phylogénétiques dans un ensemble de traits biologiques, en décrivant l'*analyse en composantes principales phylogénétiques* (ou *phylogenetic principal component analysis*, pPCA). Cette généralisation, comme nous le montrons finalement dans le chapitre 7, peut en fait être étendue à d'autres contextes, et nous permet de proposer un cadre général à la recherche de structures autocorrélées dans tout type de données multivariées. En conclusion, nous dressons un bilan de notre contribution à l'analyse des données de marqueurs génétiques, en détaillant des points de critiques ou des perspectives qui nous semblent intéressants.

Chapitre 2

Cohérence typologique des marqueurs génétiques

Sommaire

2.1 Introduction

2.1.1 Contexte général

La question de la cohérence typologique des marqueurs moléculaires a été posée dès lors que l'on a utilisé des analyses multivariées pour extraire de l'information de ces données. Dans leur étude pionnière des relations génotype-environnement, Johnson *et al.* (1969) utilisent une ACP pour résumer l'information de chaque marqueur (isozymes) ; les mêmes auteurs introduisent l'analyse canonique des corrélations (Hotelling, 1936) en génétique en faisant une analyse par marqueur (Johnson & Schaffer, 1973). Cette démarche est alors naturelle : les marqueurs utilisés sont susceptibles d'être sélectionnés de façon différente, donc d'exhiber des structures différentes, et c'est la recherche de ces structures qui motive l'étude. Les résultats obtenus dans cette étude (Johnson *et al.*, 1969) et par la suite (Johnson & Schaffer, 1973; Schaffer & Johnson, 1974; Mulley *et al.*, 1979) montrent qu'en effet les marqueurs n'ont que peu ou pas de cohérence typologique vis-à-vis du lien avec l'environnement. Dans le même temps, Cavalli-Sforza (1966) montre que cinq locus enzymatiques analysés en une seule ACP sont suffisamment cohérents pour fournir une typologie claire d'un ensemble de populations humaines. La question n'est donc pas tranchée.

Avec l'avènement des marqueurs « neutres » (Schlötterer, 2004), la question de la cohérence typologique des marqueurs subsiste, d'abord parce que cette supposée neutralité n'est pas toujours vérifiée. Au-delà de la question de la cohérence des marqueurs, c'est celle de leur utilité qui peut être posée (Rosenberg, 2005). L'aspect méthodologique de ces questions a été initialement abordé dans le cadre d'une collaboration entre Daniel Chessel, Denis Laloë et Katayoun Moazami-Goudarzi de l'INRA de Jouy-en-Josas (Laloë *et al.*, 2002). La problématique s'est avérée particulièrement intéressante : la génétique interroge la biométrie dans un champ qui lui est familier : celui des *méthodes K-tableaux*, c'est-à-dire des méthodes conçues pour analyser simultanément un ensemble de K tableaux appariés par les lignes ou par les colonnes.

Toutefois, les méthodes K-tableaux forment un champ complexe et semble-t-il plutôt méconnu de l'analyse multivariée, et l'introduction de ces méthodes en génétique constitue déjà un problème en soi. Ma contribution a été de réaliser cette introduction, via la proposition de l'analyse de coinertie multiple (ACOM, Chessel & Hanafi, 1996) pour étudier la cohérence typologique de marqueurs génétiques multialléliques. Ces travaux ont été réalisés dans la poursuite de la collaboration avec Denis Laloë et Katayoun Moazami-Goudarzi. Ils ont donné lieu à une première publication dans les actes du colloque du Bureau des Ressources Génétiques (Jombart *et al.*, 2006), présentée dans l'annexe 1, puis à une publication présentée plus bas (Laloë *et al.*, 2007). Si cette contribution met en oeuvre une seule méthode, il ne faut pas oublier que celle-ci intervient également comme représentante d'une classe de méthodes. Avant de présenter l'article proprement dit, il convient donc de situer l'ACOM dans son contexte méthodologique. Nous présentons en particulier deux autres méthodes K-tableaux « concurrentes » de l'ACOM, qui sont plus loin illustrées et comparées aux résultats obtenus dans Laloë *et al.* (2007).

2.1.2 Eléments méthodologiques

Les méthodes K-tableaux s'inscrivent dans un problématique particulière, qu'il convient d'abord d'expliciter. Puisque nous insistons sur le fait que l'ACOM est également proposée comme représentante des méthodes K-tableaux, deux autres méthodes seront ensuite présentées : l'analyse factorielle multiple (AFM, Escofier, 1994) et STATIS (Lavit & Escoufier, 1994).

a. Problématique générale

L'approche K-tableaux semble pertinente dès lors que les marqueurs génétiques utilisés possèdent un nombre conséquent de formes alléliques, ce qui peut par exemple être le cas pour les microsatellites. Il est alors possible d'obtenir une typologie d'entités (génotypes ou populations) pour chaque marqueur, et la recherche d'une typologie consensuelle entre marqueurs est alors justifiée.

On notera $\mathbf{X}$ la matrice $n \times p$ contenant l'intégralité des données centrées par colonne, portant les n entités en lignes et les p allèles en colonnes. On considère que $\mathbf{X}$ est une juxtaposition de K tableaux (un par marqueur) de dimensions $n \times p_k$, p_k étant le nombre de formes alléliques du marqueur k. On est donc en présence d'un ensemble $\mathcal{E}$ de K triplets statistiques :

$$\mathcal{E} = \{\mathfrak{T}_k \ | k = 1, \ldots, K\} \text{ avec } \mathfrak{T}_k = (\mathbf{X}_k, \mathbf{Q}_k, \mathbf{D}) \tag{2.1}$$

où $\mathbf{Q}_k$ est une métrique de $\mathbb{R}^{p_k}$ et $\mathbf{D}$ est une matrice diagonale contenant les poids des lignes de $\mathbf{X}$.

L'analyse de l'ensemble de triplets statistiques $\mathcal{E}$ (Eqn. 2.1) pose plusieurs questions, certaines étant d'un intérêt particulier lorsque l'étude porte sur des marqueurs génétiques. On dégage ici deux objectifs principaux. La première question est celle de l'existence d'une cohérence typologique entre les marqueurs. Celle-ci peut être abordée de différentes façons : en quantifiant la cohérence typologique globalement, en en donnant une mesure par marqueur, ou encore en examinant les changements de position des entités d'un marqueur à l'autre sur des cartes factorielles comparables. Un second objectif consiste à identifier, si elle existe, la nature de la typologie consensuelle, c'est-à-dire à obtenir une représentation « globale » des entités (éventuellement des allèles), visualiser le « message commun » donné par les marqueurs au sujet des entités étudiées. La diversité de points de vue sur les façons d'aborder ces deux objectifs est une composante essentielle des différences entre les méthodes K-tableaux.

b. Analyse factorielle multiple

L'analyse factorielle multiple (Escofier, 1994) constitue une approche pragmatique du problème posé par K-tableaux appariés par les lignes. Elle répond à trois objectifs : i) fournir une typologie globale des entités ii) fournir une typologie globale des variables (dans notre cas, des allèles) et iii) comparer les tableaux (*i.e.*, les marqueurs). Le schéma général de la procédure est présenté à la figure 2.1.

Avant de fournir une typologie globale d'entités, l'AFM égalise les contributions des différents

tableaux en les pondérant par l'inverse de la première valeur propre du triplet statistique correspondant. De cette façon, les nuages de n points-entités dans les K différents espaces ($\mathbb{R}^{p_k}$) sont redimensionnés de façon à ce que leur plus grande dimension soit toujours de norme unitaire. Si l'on note α_k l'inverse de la première valeur propre du triplet $\mathfrak{T}_k$, on définit alors une matrice diagonale $\mathbf{Q}$ d'ordre p contenant les pondérations des colonnes de $\mathbf{X}$:

$$\begin{pmatrix} \alpha_1\mathbf{Q}_1 & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & & & \vdots \\ \vdots & & \alpha_k\mathbf{Q}_k & & \vdots \\ \vdots & & & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & \alpha_K\mathbf{Q}_K \end{pmatrix}$$

L'analyse factorielle multiple de $\mathcal{E}$ (Eqn. 2.1) est l'analyse du triplet $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$. Elle fournit des axes principaux (notés de façon générique $\mathbf{u}$) $\mathbf{Q}$-orthonormés maximisant (Eqn. 1.8) :

$$\|\mathbf{XQu}\|^2_{\mathbf{D}} \tag{2.2}$$

On a donc une représentation des entités de variance maximale, qui est en un sens une typologie consensus. Chaque axe principal est un vecteur de $\mathbb{R}^p$ pouvant être vu comme la concaténation de K vecteurs, chacun étant associé à un tableau ($\mathbf{u}^T = [\mathbf{u}_1 \dots \mathbf{u}_K]$). La ligne i du tableau $\mathbf{X}$ possède alors une coordonnée pour le tableau k donnée par :

$$\alpha_k \mathbf{X}_{k[i]} \mathbf{Q}_k \mathbf{u}_k \tag{2.3}$$

où $\mathbf{X}_{k[i]}$ est la $i^{\text{ème}}$ ligne du tableau $\mathbf{X}_k$ (considérée comme vecteur-ligne). On peut donc avoir une représentation des entités étudiées dans chaque tableau. Notons que la typologie consensus est la somme de ces représentations car :

$$\mathbf{X}_{[i]}\mathbf{Qu} = \sum_{k=1}^{K} \alpha_k \mathbf{X}_{k[i]} \mathbf{Q}_k \mathbf{u}_k \tag{2.4}$$

Pour comparer les différences entre les typologies de chaque marqueur et la typologie consensus, on peut placer chaque observation i à la moyenne de ses coordonnées par tableau en prenant :

$$\frac{1}{K}\alpha_k \mathbf{X}_{k[i]} \mathbf{Qu} = \frac{1}{K} \sum_{k=1}^{K} \alpha_k \mathbf{X}_{k[i]} \mathbf{Q}_k \mathbf{u}_k \tag{2.5}$$

La représentation correspondante donnera une image de la cohérence typologique des marqueurs.

Enfin, on peut quantifier la participation de chaque marqueur à la typologie consensus en exprimant l'inertie projetée par rapport à l'inertie totale :

$$\frac{\|\alpha_k \mathbf{X}_k \mathbf{Q}_k \mathbf{u}\|^2_{\mathbf{D}}}{\|\mathbf{XQu}\|^2_{\mathbf{D}}} \tag{2.6}$$

On retiendra de l'AFM que c'est une approche simple permettant de représenter conjointement, et avec un critère d'optimalité, les lignes et les colonnes de $\mathbf{X}$. Le compromis se réduit aux composantes principales de l'ACP du tableau général ($\mathbf{X}$) avec une pondération particulière des colonnes. Le choix de la surpondération des allèles par l'inverse de la première valeur propre des analyses séparées des marqueurs est un parti pris : les marqueurs ayant une inertie très structurée, c'est-à-dire que l'on peut résumer en peu d'axes, auront un poids faible dans l'analyse. A l'inverse, les marqueurs contenant beaucoup d'allèles dont chacun porte une information particulière, peu redondante avec les autres, auront un poids fort (Pagès, 1996). C'est un choix discutable : les marqueurs fournissant les typologies les plus claires sont avant tout ceux pour lesquels l'inertie est la plus structurée (réduite à peu d'axes), mais ce sont ceux auxquels l'analyse accorde le moins d'importance *a priori*.

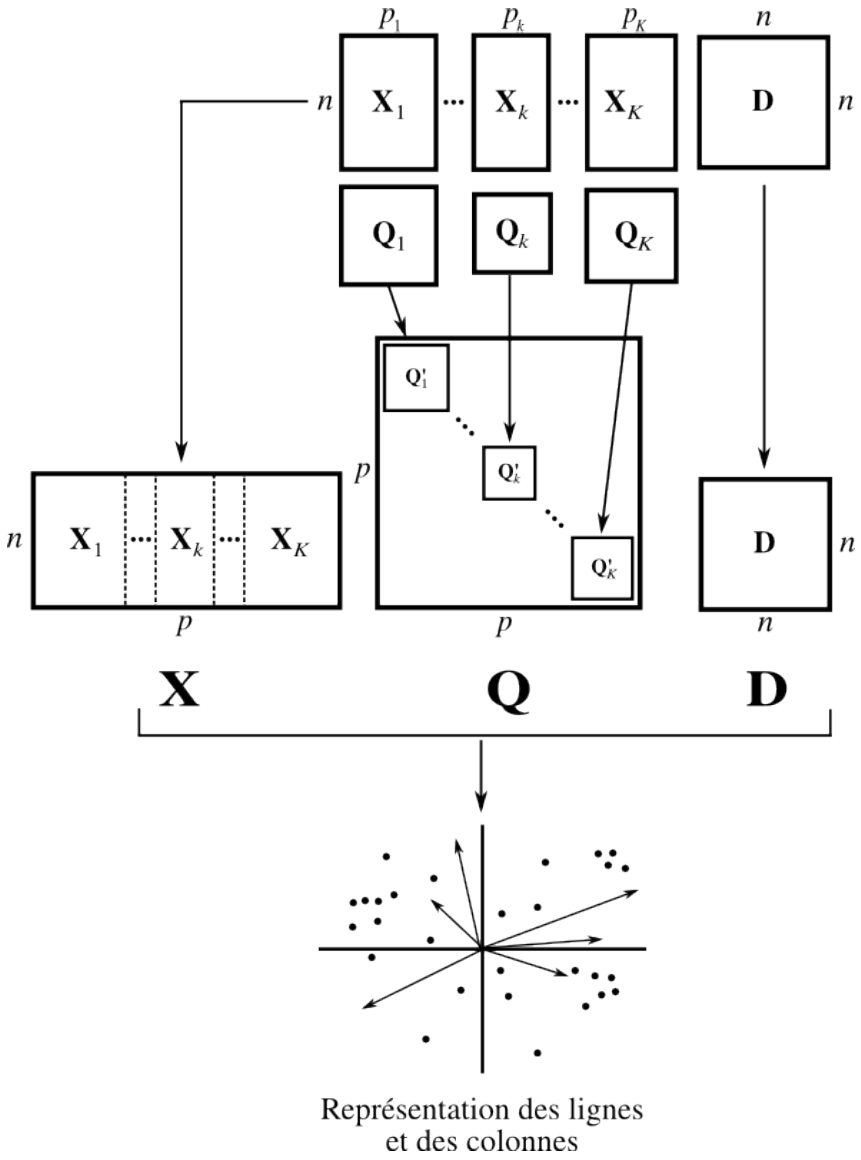


FIG. 2.1: Schéma de l'analyse factorielle multiple. La matrice $\mathbf{X}$ est obtenue par juxtaposition des $\mathbf{X}_k$. Les matrices $\mathbf{Q}_k$, $\mathbf{D}$ et $\mathbf{Q}$ sont diagonales. $\mathbf{Q}'_k = \alpha_k \mathbf{Q}_k$, où α_k est l'inverse de la première valeur propre du triplet $\mathfrak{T}_k = (\mathbf{X}_k, \mathbf{Q}_k, \mathbf{D})$. L'analyse factorielle multiple est l'analyse du triplet $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$. L'ordination obtenue est une représentation des lignes et des colonnes de l'ensemble des données, mais pouvant donner une sous-représentation par tableau.

c. STATIS

L'approche STATIS (Lavit & Escoufier, 1994) repose sur la comparaison de K matrices de données appariées par les lignes et/ou par les colonnes. Dans le cas des marqueurs moléculaires, seul l'appariement par ligne (*i.e.* « entités » : génotypes ou populations) est pertinent. C'est d'ailleurs la présentation originelle de la méthode. Le schéma général de l'analyse est présenté figure 2.2.

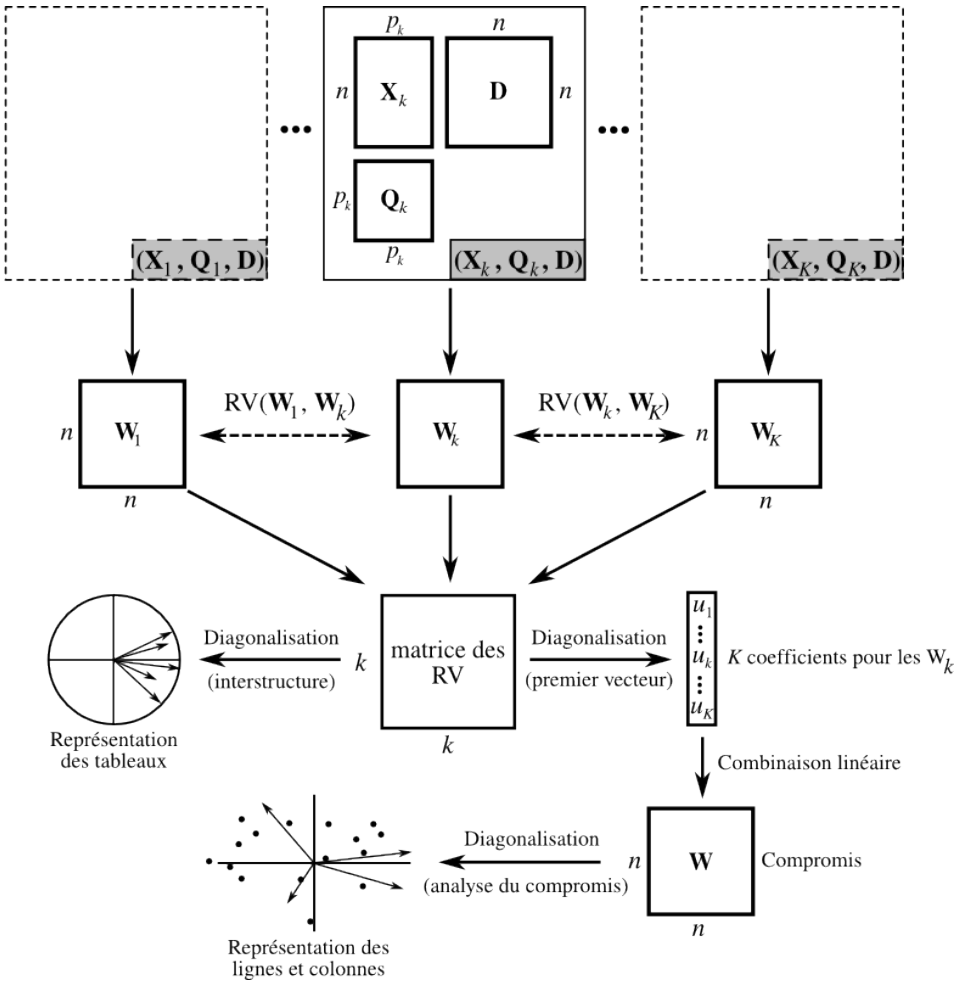


FIG. 2.2: Schéma de la méthode STATIS sur K tableaux appariés par les lignes. Chaque triplet statistique $\mathfrak{T}_k = (\mathbf{X}_k, \mathbf{Q}_k, \mathbf{D})$ donne une matrice symétrique de produit scalaires $\mathbf{W}_k$. La comparaison de ces opérateurs d'inertie donne une matrice de coefficients RV entre les $\mathbf{W}_k$, qui est diagonalisée. D'une part, cette diagonalisation fournit une représentation euclidienne des tableaux (*interstructure*). D'autre part, le premier vecteur propre $\mathbf{u} = [u_1 \dots u_K]^T$ possède des coefficients tous positifs, qui sont utilisés pour former des combinaisons linéaires des $\mathbf{W}_k$, et ainsi obtenir un opérateur d'inertie « moyen », ou *compromis* $\mathbf{W}$. L'analyse du compromis fournit une représentation des lignes et des colonnes de l'ensemble des triplets statistiques $\mathcal{E}$.

Le premier problème est de comparer des objets comparables. Il n'est pas directement possible de définir de distance ou de similarité entre K matrices de dimensions différentes. Par contre, chaque triplet fournit une matrice de produits scalaires entre lignes de $\mathbf{X}_k$ de la forme :

$$\mathbf{W}_k = \mathbf{X}_k \mathbf{Q}_k \mathbf{X}_k^T \mathbf{D} \tag{2.7}$$

Ces produits scalaires contiennent l'information de la configuration du nuage des n objets permettant d'en produire une typologie (Eqn. 1.6). La ressemblance entre ces matrices définies pour chaque marqueur est donc source d'information en terme de cohérence typologique.

De même que pour des vecteurs, on peut définir des produits scalaires, des normes qui y sont associées, et des cosinus entre matrices (Harville, 1997, pp.58-61). Le produit scalaire entre deux matrices $\mathbf{A} = [a_{ij}]$ et $\mathbf{B} = [b_{ij}]$ de mêmes dimensions (aussi appelé *produit de Frobenius*) est défini par la forme bilinéaire symétrique :

$$\begin{array}{rccl} \langle \cdot, \cdot \rangle : & \mathbb{R}^{n\times p} \times \mathbb{R}^{n\times p} & \longrightarrow & \mathbb{R} \\ & (\mathbf{A}, \mathbf{B}) & \longmapsto & \langle \mathbf{A}, \mathbf{B} \rangle = \mathrm{tr}(\mathbf{A}^T\mathbf{B}) = \sum_{i=1}^{n}\sum_{j=1}^{p} a_{ij}b_{ij} \end{array} \tag{2.8}$$

où $\mathrm{tr}(\mathbf{A}^T\mathbf{B})$ désigne la trace de $\mathbf{A}^T\mathbf{B}$.

Ce produit scalaire est associé à une norme, dite *norme de Frobenius* :

$$\begin{array}{rccl} \|\cdot\| : & \mathbb{R}^{n\times p} & \longrightarrow & \mathbb{R}_+ \\ & \mathbf{A} & \longmapsto & \|\mathbf{A}\| = \sqrt{\mathrm{tr}(\mathbf{A}^T\mathbf{A})} \end{array} \tag{2.9}$$

La distance entre deux matrices de mêmes dimensions est alors définie par :

$$\begin{array}{rccl} d : & \mathbb{R}^{n\times p} \times \mathbb{R}^{n\times p} & \longrightarrow & \mathbb{R}_+ \\ & (\mathbf{A}, \mathbf{B}) & \longmapsto & \|\mathbf{A} - \mathbf{B}\| = \sqrt{\mathrm{tr}((\mathbf{A} - \mathbf{B})^T(\mathbf{A} - \mathbf{B}))} \end{array} \tag{2.10}$$

La méthode STATIS utilise ces résultats pour comparer les matrices de produits scalaires entre les lignes de $\mathbf{X}$ (Eqn. 2.7). Soient $\mathbf{W}_k$ et $\mathbf{W}_l$ deux de ces matrices ; on peut mesurer leur similarité par leur produit scalaire :

$$\langle \mathbf{W}_k, \mathbf{W}_l \rangle = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{kij}w_{lij} \tag{2.11}$$

Mais comme noté par Lavit & Escoufier (1994), une forte dissimilarité dans (Eqn. 2.11) peut être due à une différence de configuration des nuages aussi bien qu'à une différence de taille. On utilise donc un produit scalaire entre matrices normées, noté RV (Escoufier, 1973) :

$$\mathrm{RV}(\mathbf{W}_k, \mathbf{W}_l) = \frac{\langle \mathbf{W}_k, \mathbf{W}_l \rangle}{\|\mathbf{W}_k\|\|\mathbf{W}_l\|} \tag{2.12}$$

ce qui peut être développé en :

$$\mathrm{RV}(\mathbf{W}_k, \mathbf{W}_l) = \frac{\mathrm{tr}(\mathbf{D}\mathbf{X}_k\mathbf{Q}_k\mathbf{X}_k^T\mathbf{X}_l\mathbf{Q}_l\mathbf{X}_l^T\mathbf{D})}{\sqrt{\mathrm{tr}(\mathbf{D}\mathbf{X}_k\mathbf{Q}_k\mathbf{X}_k^T\mathbf{X}_k\mathbf{Q}_k\mathbf{X}_k^T\mathbf{D})}\sqrt{\mathrm{tr}(\mathbf{D}\mathbf{X}_l\mathbf{Q}_l\mathbf{X}_l^T\mathbf{X}_l\mathbf{Q}_l\mathbf{X}_l^T\mathbf{D})}} \tag{2.13}$$

L'écriture (Eqn. 2.12) montre qu'il s'agit exactement du cosinus de l'angle entre $\mathbf{W}_k$ et $\mathbf{W}_l$ (Harville, 1997, pp.60). Le développement (Eqn. 2.13) suggère que nous sommes dans un cas

particulier, car :

$$\begin{aligned} \mathrm{tr}(\mathbf{D}\mathbf{X}_k\mathbf{Q}_k\mathbf{X}_k^T\mathbf{X}_l\mathbf{Q}_l\mathbf{X}_l^T\mathbf{D}) &= \mathrm{tr}(\mathbf{X}_k\mathbf{Q}_k\mathbf{X}_k^T\mathbf{D}\mathbf{X}_l\mathbf{Q}_l\mathbf{X}_l^T\mathbf{D}) \\ &= \mathrm{tr}(\mathbf{X}_k^T\mathbf{D}\mathbf{X}_l\mathbf{Q}_l\mathbf{X}_l^T\mathbf{D}\mathbf{X}_k\mathbf{Q}_k) \\ &= \mathrm{coinertie}(\mathfrak{T}_k, \mathfrak{T}_l) \end{aligned} \tag{2.14}$$

Le numérateur du coefficient RV correspond donc à la coinertie entre les deux triplets statistiques $\mathfrak{T}_k$ et $\mathfrak{T}_l$ (Dolédec & Chessel, 1994; Dray *et al.*, 2003). Cette quantité étant positive (c'est une somme de carrés de covariances entre les variables des tableaux $\mathbf{X}_k$ et $\mathbf{X}_l$), le coefficient RV est compris entre 0 et 1, et non entre -1 et 1.

La méthode STATIS diagonalise la matrice symétrique d'ordre K des coefficients RV entre opérateurs d'inertie $(\mathbf{W}_k, k = 1, \ldots, K)$, qui fournit une typologie des différents opérateurs. C'est l'étape dite de l'*interstructure*, qui donne une image de la cohérence typologique des marqueurs. Le théorème de Perron-Frobenius assure que toutes les composantes $(u_1, \ldots, u_K)$ du premier vecteur propre $\mathbf{u}$ de cette matrice sont positives. Ce sont des coefficients de combinaisons linéaires définissant un *compromis* $\mathbf{W}$ entre les opérateurs $\mathbf{W}_k$:

$$\mathbf{W} = \sum_{k=1}^{K} u_k \mathbf{W}_k \tag{2.15}$$

qui est l'opérateur moyen le plus proche de tous les autres au sens du produit scalaire, c'est-à-dire maximisant :

$$\sum_{k=1}^{K} \langle \mathbf{W}, \mathbf{W}_k \rangle^2 \tag{2.16}$$

Lavit & Escoufier (1994) proposent également différentes options pour standardiser les coefficients u_k ou les opérateurs $\mathbf{W}_k$ pour construire le compromis. La version de STATIS implémentée dans *ade4* (Dray *et al.*, 2007) utilise par exemple des combinaisons linéaires d'opérateurs standardisés $\frac{\mathbf{W}_k}{\|\mathbf{W}_k\|}$.

L'analyse du compromis (diagonalisation de la matrice $\mathbf{W}$) fournit un ensemble de vecteurs propres de $\mathbb{R}^n$ donnant une représentation consensuelle des individus (génotypes ou populations). Après $\mathbf{D}$-normalisation, ces vecteurs propres sont analogues aux composantes principales normées d'un schéma classique (Eqn. 1.6, vecteurs $\mathbf{v}_k$, en rouge sur (Fig. 1.2)). Une représentation des allèles est obtenue par projection $\mathbf{D}$-orthogonale des colonnes des $\mathbf{X}_k$ sur ces vecteurs propres.

On retiendra de la méthode STATIS qu'elle repose avant tout sur une mesure élégante de la co-structuration de deux tableaux, le coefficient RV, qui mesure le lien entre deux opérateurs d'inertie, et donc entre la configuration de deux nuages de points. L'analyse du tableau des RV entre les différents triplets statistiques fournit d'abord une représentation euclidienne des marqueurs donnant une image de leur cohérence typologique. Cette analyse fournit également un compromis, qui est l'opérateur d'inertie le plus lié (au sens du carré du produit de Frobenius)

aux opérateurs d'inertie de chaque tableau. C'est une configuration du nuage de points qui représente au mieux (pour le critère défini) l'ensemble des K configurations : c'est une expression du « message commun » apporté par les marqueurs au sujet des entités étudiées. Il en découle une représentation graphique des entités, et éventuellement des allèles, mais celle-ci n'est pas optimale. En ce sens, on peut penser que STATIS est une méthode avant tout axée sur la cohérence typologique, et secondairement sur la typologie elle-même.

d. Analyse de coinertie multiple

L'analyse de coinertie multiple (Chessel & Hanafi, 1996) est comme son nom l'indique une extension multi-tableaux de l'analyse de coinertie (Dolédec & Chessel, 1994; Dray *et al.*, 2003). La méthode est présentée dans le cadre de l'analyse de marqueurs génétiques dans l'article qui suit (Laloë *et al.*, 2007). Sans pour autant entrer dans les détails des principes mathématiques de la méthode qui seront exposés dans l'article, nous pouvons donner quelques éléments justifiant l'utilisation de l'ACOM plutôt que l'AFM ou STATIS.

L'ACOM considère que chaque triplet statistique $\mathfrak{T}_k$ fournit une typologie, c'est-à-dire un ensemble restreint d'axes principaux ayant une inertie projetée forte. Elle diffère en cela fondamentalement de l'AFM, qui privilégie la multidimensionnalité des données de chaque tableau (Pagès, 1996). Dans le cas des marqueurs génétiques, on peut dire que l'ACOM recherche des variables de synthèses communes entre les marqueurs, alors que l'AFM recherche des ensembles d'allèles pour chaque marqueur portant une information consensuelle. L'ACOM recherche un ensemble d'axes $\mathbf{u}_k \in \mathbb{R}^{p_k}$, $\mathbf{Q}_k$-orthonormés pour chaque $\mathfrak{T}_k$ et des composantes principales associées $(\mathbf{X}_k\mathbf{Q}_k\mathbf{u}_k)$, ainsi qu'un score $\mathbf{v} \in \mathbb{R}^n$ dit de *référence*, tels que les composantes soient les plus proches possible de la référence au sens du carré de covariance (ou *coinertie*), c'est-à-dire maximisant :

$$\sum_{k=1}^{K} w_k \mathrm{cov}^2(\mathbf{X}_k\mathbf{Q}_k\mathbf{u}_k, \mathbf{v}) = \mathrm{var}(\mathbf{X}_k\mathbf{Q}_k\mathbf{u}_k)\mathrm{var}(\mathbf{v})\mathrm{cor}^2(\mathbf{X}_k\mathbf{Q}_k\mathbf{u}_k, \mathbf{v}) \tag{2.17}$$

où w_k est une pondération du tableau k, et où variances, covariances et corrélations sont calculées avec la métrique $\mathbf{D}$. Les scores fournis sont donc un compromis entre optimalité typologique (variances maximisées) et ressemblance à la référence (carrés des corrélations maximisés). La valeur typologique d'un marqueur l est calculée comme sa contribution à la coinertie totale :

$$\frac{w_l \mathrm{cov}^2(\mathbf{X}_l\mathbf{Q}_l\mathbf{u}_l, \mathbf{v})}{\sum_{k=1}^{K} w_k \mathrm{cov}^2(\mathbf{X}_k\mathbf{Q}_k\mathbf{u}_k, \mathbf{v})} \tag{2.18}$$

Les allèles peuvent être représentés sur ces typologies, mais sans propriété d'optimalité.

L'ACOM possède donc l'intérêt de fournir des typologies pour chaque triplet statistique, et une typologie de référence qui met en lumière les ressemblances entre typologies. Elle propose en outre une mesure naturelle de la contribution de chaque marqueur à la typologie de référence (Eqn. 2.18). Cette approche permet donc de remplir les principaux objectifs fixés. L'ACOM peut d'une certaine façon être vue comme un intermédiaire entre l'AFM et STATIS, s'intéressant plus

à la cohérence typologique que la première, et plus à la typologie d'ensemble et aux typologies séparées que la seconde. Ces propriétés justifient le choix de l'ACOM pour l'introduction des méthodes K-tableaux en génétique, réalisée dans l'article qui suit.

2.2 Article 2 : Consensus genetic structuring and typological value of markers using multiple co-inertia analysis

Article paru en 2007 dans *Genetics Selection Evolution*, **39** : 545-567.

Genet. Sel. Evol. 39 (2007) 545–567
© INRA, EDP Sciences, 2007
DOI: 10.1051/gse:2007021

Available online at:
www.gse-journal.org

Original article

Consensus genetic structuring and typological value of markers using multiple co-inertia analysis

Denis LALOË[a*], Thibaut JOMBART[b], Anne-Béatrice DUFOUR[b], Katayoun MOAZAMI-GOUDARZI[c]

[a] Station de génétique quantitative et appliquée UR337, INRA, 78352 Jouy-en-Josas, France
[b] Université de Lyon, Université Lyon 1, CNRS, UMR 5558, Laboratoire de biométrie et biologie évolutive, 69622 Villeurbanne Cedex, France
[c] Laboratoire de génétique biochimique et de cytogénétique UR339, INRA, 78352 Jouy-en-Josas, France

(Received 23 October 2006; accepted 20 April 2007)

Abstract – Working with weakly congruent markers means that consensus genetic structuring of populations requires methods explicitly devoted to this purpose. The method, which is presented here, belongs to the multivariate analyses. This method consists of different steps. First, single-marker analyses were performed using a version of principal component analysis, which is designed for allelic frequencies (%PCA). Drawing confidence ellipses around the population positions enhances %PCA plots. Second, a multiple co-inertia analysis (MCOA) was performed, which reveals the common features of single-marker analyses, builds a reference structure and makes it possible to compare single-marker structures with this reference through graphical tools. Finally, a typological value is provided for each marker. The typological value measures the efficiency of a marker to structure populations in the same way as other markers. In this study, we evaluate the interest and the efficiency of this method applied to a European and African bovine microsatellite data set. The typological value differs among markers, indicating that some markers are more efficient in displaying a consensus typology than others. Moreover, efficient markers in one collection of populations do not remain efficient in others. The number of markers used in a study is not a sufficient criterion to judge its reliability. "Quantity is not quality".

congruence / multiple co-inertia analysis / biodiversity / microsatellite / allelic frequencies

1. INTRODUCTION

Today, a large number of studies are aimed at investigating the genetic structuring of populations within species. The goal of such studies is first to provide

* Corresponding author: denis.laloe@jouy.inra.fr

Article published by EDP Sciences and available at http://www.gse-journal.org or http://dx.doi.org/10.1051/gse:2007021

insight into the management and conservation of today's animal and plant genetic resources, the history of populations: demography [7, 39], origin and migration routes for human populations [14] or the history of livestock domestication [9, 11]. Epidemiological considerations can also motivate such studies in human populations [56]. However, the most common justification of these studies is their importance for quantifying biodiversity and thus for establishing priorities in conservation programs [10, 22, 41, 59, 64].

Under the coordination of the FAO, an initiative called the measurement of domestic animal diversity (MoDAD) was started in order to provide technical recommendations for studies in farm animals [24]. Among the many DNA tools available, microsatellites are the most widely used mainly because of their high variability. Within this context, an FAO/ISAG advisory group has been formed to recommend species-specific lists of microsatellite loci (about 30 per species) for the major farm animal species (cattle, buffalo, yak, goat, sheep, pig, horse, donkey, chicken and camelids; http://dad.fao.org/en/refer/library/guidelin/marker.pdf). The adherence to such recommendations permits reasonable comparisons of parallel or overlapping studies of genetic diversity and it is a necessary prerequisite to combine results in meta-analyses [60]. Within this context, Baumung *et al*. [5] published the results from a survey concerning 87 projects of genetic domestic studies in domestic livestock. In their article, they underline that the recommended markers are well known and used in 79% of the projects.

Generally, in these studies on genetic structuring, two methods were performed: phylogenetic reconstruction [46, 57, 67] and/or multivariate procedures [8, 15, 63, 65, 69]. In phylogenetic reconstruction, a consensus tree is typically built to summarize information and measure the reliability of the tree. Several methods have been proposed for inferring consensus trees, among them the maximum agreement subtree, the strict consensus, the majority tree, the Adams consensus and the asymmetric median tree [12, 52].

However, construction of trees using admixed populations, as is the case in livestock species, violates the principles of phylogeny reconstruction [25, 64]. In this situation, multivariate procedures are recommended. The most common method to analyze allelic frequency data is the principal component analysis (PCA) [6, 33, 34, 36, 37, 48]. Using such methods may result in a non consensus representation, due to the incongruence among markers [50]. Weak congruence could also explain some of the low bootstrap values which are typically reported in several studies in the following species: beef cattle [13, 43, 45, 47, 51, 67], goats [35, 42], sheep [63, 70], and natural populations, such as white-tailed deer [20].

The markers involved in such studies are chosen to be neutral. One of the main principles of population genomics states that neutral markers across the genome will be similarly affected by demography and the evolutionary history of populations [44]. Accordingly, these markers should be congruent, *i.e.* should reveal the same typology among populations.

Nevertheless, neutral markers may be influenced by selection on nearby (linked) loci, and, then, reveal different patterns of variation.

Thus, a method explicitly devoted to exhibit a consensus in a multivariate framework is necessary. In this context, the markers of interest should be both highly variable and congruent in order to perform a consensus typology. The multiple co-inertia analysis (MCOA) is dedicated to this purpose. MCOA was first described by Chessel and Hanafi [17], and is used in ecology [4, 30].

In this paper, we address the capacity and efficiency of marker panels to exhibit a genetic structuring and measure the contribution of each specific marker by MCOA. In the genetic framework, this ordination method identifies the structures of populations common to many tables of allelic frequencies. First, single marker analyses were performed. Allelic frequencies are a special case of compositional data [1,3]: they consist of vectors of positive values summing to one. De Crespin de Billy *et al*. [19] introduced a specifically designed principal component analysis (%PCA) for this kind of data. This method can be used together with a biplot representation [27], which permits an interpretation of the location of a population in terms of its allelic frequencies. Adding confidence ellipses [29] around the population points on the resulting plot improves the visual assessment of the separating power of the markers. It also allows accounting for the uncertainty due to the size of the sampled population. Second, MCOA simultaneously finds ordinations from the tables that are most congruent. It does this by finding successive axes from each table of allelic frequencies, which maximize a covariance function. This method permits the extraction of common information from separate analyses, in the setting-up of a reference typology, and the comparison of each separate typology to this reference typology. Finally, to quantify the efficiency of a marker, we introduce the typological value (TV), which is the contribution of the marker to the construction of the reference typology.

Hence, we reply to the following practical questions. Which markers contribute most to the typology of populations? Do efficient markers in one collection of populations remain efficient in others? Does the number of markers ensure the reliability of the typology?

In this article, we provide a short background to MCOA, we describe the typological value and we study the interest and efficiency of this method using a bovine data set.

2. MATERIALS AND METHODS

2.1. Single marker analyses

Each marker yields allelic frequencies that define Euclidian distances between the populations in a multidimensional space. The principal component analysis [33, 34] can be used to find a plane on which the populations are scattered as much as possible, *i.e.* conserving the distances among populations as best as possible. However, this method does not take into account the true nature of the data. Since allelic frequencies are positive and sum to one, they are compositional data [1]. Aitchison addressed some issues specific to the multivariate analysis of such data [1–3] and showed that centered PCA performs better when compositional data are transformed using log ratios or other logarithmic data transformations [55]. An appealing alternative to these approaches is to use a principal component analysis of proportion data (%PCA) [19]. Indeed, the typologies provided by this analysis are directly interpretable in term of allelic frequencies, which is at least discussed in former methods [68].

The %PCA yields the same axes as a classical centered PCA, and the distances between the scores of the populations are exactly the same as in PCA. Thus the typology of the populations is not altered. %PCA differs from PCA in that the cloud of points corresponding to the populations is not constrained to be at the origin. Instead, the populations are placed by averaging with respect to their allelic frequencies. The score s_i of a population i onto an axis $\mathbf{u}$ is computed as the mean of the allele coordinates (denoted u_j, $1 \leq j \leq p$) weighted by the corresponding allelic frequencies (f_{ij}): $s_i = \sum_{j=1}^{p} f_{ij} u_j$.

This method makes it possible to draw meaningful biplots [19], where both populations and alleles are represented, respectively by points and arrows. In such biplots, the closer the populations are to an allele, the higher the corresponding frequencies are.

To improve the typologies of populations obtained by %PCA, we propose confidence ellipses as a visual tool to assess the genetic differences between populations. Indeed, it should be valuable to take the precision of the population frequency estimates into account. Since these frequencies are just estimates of the real ones, they may change from one sample to another. The

consequence for the typology is that the coordinates of any population fluctuate around the true, unknown position. Hence, we can determine a confidence ellipse [29], inside which the true population can be expected to be located, with a given probability. This probability P is linked to a size factor S by:

$$P = 1 - \exp\left(-\frac{S^2}{2}\right)\cdot$$

Using a PCA appropriate for allelic frequencies and confidence ellipses around population positions should help to interpret the different typologies provided by the markers. At this point, the multiple co-inertia makes it possible to carry out a comparison between these typologies.

2.1.1. Multiple co-inertia analysis

Multiple co-inertia analysis is an ordination method, which simultaneously analyzes K tables describing the same objects (in rows) with different sets of variables (in columns). The mathematical principles of the method are fully described by their authors [17], but we provide essential steps in the appendix; examples of its utilization can be found in ecology studies [4, 30].

Within the MCOA framework, K sets of variables produce K typologies of the same objects on the basis of any single-table analysis, such as PCA or correspondence analysis. MCOA relies on the idea that there may be congruent structures among these typologies. The MCOA coordinates the K separate PCA, in order to facilitate their comparison and emphasize their similarities. A reference ordination is then constructed, which best summarizes the congruent information among the sets of variables. It can thus be considered as a "reference structure" (also called "reference").

We apply the MCOA to analyze a set of n populations typed on K markers. The method provides a set of K coordinated %PCA, each corresponding to a given molecular marker. These analyses can be interpreted like previous %PCA since populations are placed by averaging with respect to the alleles. However, these analyses display both scattered and congruent typologies, which can thus be compared. So, the criterion of the scores of maximum variance (used in %PCA) is no longer sufficient, and the correlation of the scores with the reference must be taken into account. To consider these two aspects, the MCOA maximizes the sum of the co-inertias (*i.e.* squared covariances) between the scores of populations of the coordinated analyses, and the reference. Let $\mathbf{l}_k^r$ be the r^{th} scores of populations in the coordinated %PCA of a marker k (with $1 \leq k \leq K$),and $\mathbf{v}^r$ be the r^{th} reference scores. The criterion optimized in

MCOA is then:

$$\sum_{k=1}^{K} w_k \ \mathrm{cov}^2(\mathbf{l}_k^r, \mathbf{v}^r) = \sum_{k=1}^{K} w_k \ \mathrm{var}(\mathbf{l}_k^r) \ \mathrm{var}(\mathbf{v}^r) \ \mathrm{corr}^2(\mathbf{l}_k^r, \mathbf{v}^r) \tag{1}$$

where w_k is a given weight for the marker k. These weights can be chosen according to the nature and disparity of the markers. We choose here uniform weights ($w_k = \frac{1}{K}$) for every marker, but it is possible, for instance, to choose w_k so that markers of different types are on the same level of variation.

The optimized criterion (1) guarantees that the typologies are scattered (maximization of the variance of the scores) and emphasizes their common structure (maximization of the squared correlation). This matches our definition of what a "good marker" is, from a typological point of view: a marker which can separate the populations well, and which separates them like many other markers. Mathematically, this exactly corresponds to the contribution of a marker to the MCOA criterion:

$$w_k \ \mathrm{cov}^2(\mathbf{l}_k^r, \mathbf{v}^r) = w_k \ \mathrm{var}(\mathbf{l}_k^r) \ \mathrm{var}(\mathbf{v}^r) \ \mathrm{corr}^2(\mathbf{l}_k^r, \mathbf{v}^r). \tag{2}$$

2.2. Typological value

If the maximum of (1) is noted λ_r, we can define the typological value (TV) of the marker k as its relative contribution to the previous criterion:

$$TV_r(k) = \frac{w_k \ \mathrm{cov}^2(\mathbf{l}_k^r, \mathbf{v}^r)}{\lambda_r}. \tag{3}$$

Contrary to (2), this expression is a proportion and can be expressed as a percentage. It corresponds to the ability of the marker k to display the r^{th} reference structure. The higher it is, the better it displays the r^{th} structure of the reference. As a consequence, it can be used to compare the typological values of a set of markers on a given structure. Whenever a structure is expressed by more than one axis of the reference, (3) can be extended by summing separately the numerator and denominator. For example, if an interesting structure of populations is expressed by scores i and j, (3) is generalized as:

$$TV_{i,j}(k) = \frac{w_k \ \mathrm{cov}^2(\mathbf{l}_k^i, \mathbf{v}^i) + w_k \ \mathrm{cov}^2(\mathbf{l}_k^j, \mathbf{v}^j)}{\lambda_i + \lambda_j}.$$

A last question to be tackled concerns the number of existing common structures. This is the number of scores to be kept for the reference and for each

coordinated analysis. This number is chosen according to the decrease of λ_r, as is the case in PCA with eigenvalues. However, this choice is made easier than in PCA, since MCOA eigenvalues have the status of squared PCA eigenvalues, the differences between high ones (interesting structures) and low ones would be clearer in MCOA.

These methods are available in the ade4 package [18] of the R software [54].

2.3. Application to data

Blood samples of 755 unrelated animals from 16 cattle breeds were analyzed:

– **11 from France**: Aubrac (Aub, n = 50), Bazadaise (Baz, n = 47), Blonde d'Aquitaine (Blo, n = 61), Bretonne Pie noire (Bre, n = 31), Charolaise (Cha, n = 55), Gasconne (Gas, n = 50), Limousine (Lim, n = 50), Maine-Anjou (Mai, n = 49), Montbeliarde (Mon, n = 31), Normande (Nor, n = 50) and Salers (Sal, n = 50). Samples were collected throughout France;

– **5 from West Africa**: Lagunaire (Lag, n = 51), N'Dama (N'Da, n = 30), Somba (Som, n = 50), Sudanese Fulani Zebu (Zeb, n = 50) and Borgu (Bor, n = 50). The Borgu breed is a crossbred between West African shorthorn cattle and zebu. West African populations were collected in three neighboring countries: Benin, Togo and Burkina Faso. This West African data set has been taken from [49].

All breeds were genotyped for 30 microsatellite loci recommended for genetic diversity studies by the EC-funded European cattle diversity project (Resgen CT 98-118) and the FAO. Details on primers, original references and experimental protocols (conditions of PCR, multiplexing) can be found at http://dad.fao.org/en/refer/library/guidelin/marker.pdf.

These 30 microsatellites were genotyped using an ABI 377 sequencer or by Labogena (www.labogena.fr) using an ABI 3700 sequencer.

To standardize genotypes between our laboratory and Labogena and in order to limit genotyping errors during laboratory experiments, we used three reference animals as controls in each gel run. To limit scoring errors, the results were recorded by two independent scorers [53].

3. RESULTS AND DISCUSSION

We first ran a %PCA on each microsatellite table of allelic frequencies (single-marker analysis). Corresponding plots are drawn on the same scale for six markers on Figure 1. For each marker, the first two axes of the %PCA are

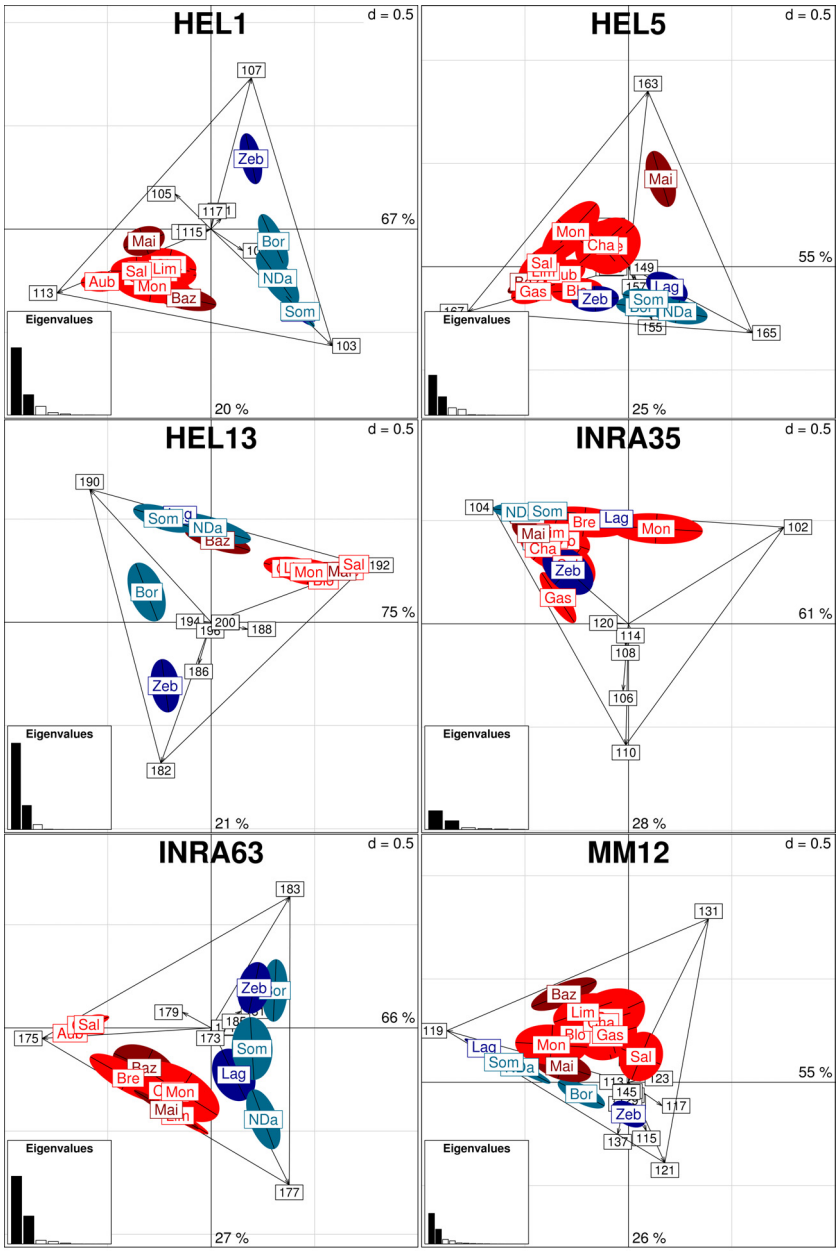


Figure 1. Single marker %PCA (first two axes). The populations are labelled in their confidence ellipse (P = 0.95), within an envelope formed by the alleles (arrows). Figures are on the same scale as indicated by the mesh of the grid (d = 0.5). Eigenvalue percents are indicated for each axis. The colors are based on the most congruent differentiation in the reference scores.

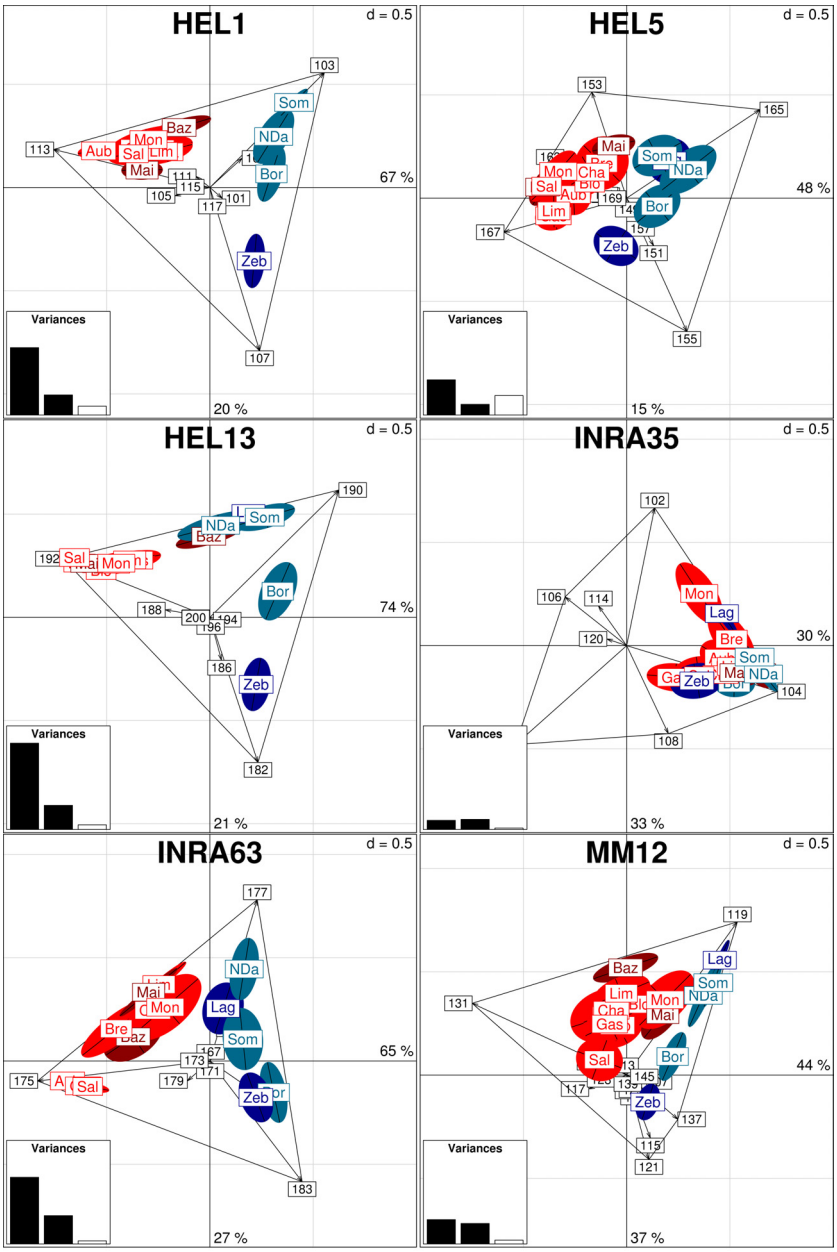


Figure 2. Single marker coordinated %PCA (first two axes). The populations are labelled in their confidence ellipse (P = 0.95), within an envelope formed by the alleles (arrows). Figures are on the same scale as indicated by the mesh of the grid (d = 0.5). Variance percents are indicated for each axis). The colors are based on the most congruent differentiation in the reference scores.

shown. Alleles are represented by arrows, the most discriminating ones being joined by lines. A confidence ellipse (P = 0.95) accounting for the number of sampled animals is drawn around each population point. The barplot of eigenvalues is drawn at the bottom left. It indicates the relative magnitude of each axis with respect to the total variance. The higher the eigenvalue is, the higher the Euclidean distances are among populations. For example, for HEL13, the first axis accounts for 75% of the total variance and the second axis accounts for 21%.

For this marker, the populations are mainly structured by three alleles, alleles 182, 190 and 192, their allelic frequencies varying strongly according to populations (from 0 to 0.59 for 182, from 0.02 to 0.70 for 190 and from 0.05 to 0.94 for 192). The breeds are mainly differentiated by their respective allelic frequencies for these alleles. The Sudanese Fulani Zebu breed and Borgu lie along the line 182–190 and African taurine breeds and French breeds lie along the line 190–192. For example, allele 192 was highly frequent in French breeds (0.94 in Salers), and allele 190 was frequent in African taurine breeds (0.70 for Somba), while allele 182 was very rare in African taurine populations, absent in the French populations and present with a frequency of 0.59 in the Sudanese Fulani Zebu breed. Thus allele 182 could be a zebu diagnostic allele.

Some other alleles are located close to the center of the plot, because they are rare: 178, 184, 194, 196 and 200, with maximal allelic frequencies of 0.01, 0.01, 0.07, 0.02 and 0.01, respectively. The last two alleles (186 and 188) lie in an intermediate position: allele 186 was detected with a frequency of 0.17 in the Sudanese Fulani Zebu breed and it was nearly absent in the remaining breeds. Allele 188 was detected only in French breeds with a maximal allelic frequency of 0.26 for the Blonde d'Aquitaine breed. Drawing a confidence ellipse leads to a graphical assessment of the population structuring. Four clusters can be pointed out: the French breeds (without the Bazadaise breed), the African taurine breeds and Bazadaise breed, the Borgu breed and the Sudanese Fulani Zebu breed.

When all the markers are considered, it is easy to see that the efficiency of each marker differs. Some did not exhibit any clustering (INRA35), others exhibited some clusters but not always the same. For example HEL1 and HEL13 separated three clusters: French taurine, African taurine and African Zebu. Some microsatellites *i.e.* MM12 separated the African taurine breeds from the zebu breed. Within the French cluster, INRA63 separated three breeds and HEL5 isolated the Maine-Anjou breed from the others.

Figure 1 is a graphical tool, which compares the usefulness of markers for separating populations. However, the axes of each %PCA differ from one

marker to another, and cannot be interpreted in the same way. Axis 1 of the HEL1 plot is not the same as Axis 1 of the MM12 plot. Single-marker structures cannot be easily compared by looking at factorial maps of separate uncoordinated analyses. The multiple co-inertia analysis deals with this problem, through coordinated analyses, where axes of each plot tend to display the same structures.

Coordinated %PCA plots are drawn on the same scale for the six markers on Figure 2. Ellipses and proximities between alleles and populations can be interpreted in the same way as in Figure 1. However, the barplot at the bottom left of the plot no longer represents eigenvalues, but the variance of the scores according to the different axes. For instance, populations are more scattered along the first axis for HEL13 than for HEL1, or INRA63.

A comparison of Figure 1 with Figure 2 shows that some markers fit the common structures quite well. For instance, the first two axes of the plots of HEL1, HEL13 and INRA63 are almost identical. Some others remain non efficient *e.g.* INRA35. However, for MM12 and HEL5, the situation is more interesting. For MM12, axis 1 in Figure 1 is more or less axis 2 in Figure 2 of the common structure exhibited by MCOA. Concerning HEL5, in Figure 1 the most obvious feature is the separation of the Maine-Anjou breed from the others. However this marker exhibits the common structure as indicated in Figure 2.

Therefore, the non-coordinated analyses answer the question: does the marker separate the populations while the coordinated analysis answers the question: how does the marker separate the populations regarding the common structure.

The decrease of eigenvalues shows three main structures in the reference typology. The first three axes of the reference typology are shown in Figures 3A (axes 1 and 2) and 3B (axes 1 and 3). The first axis clearly distinguishes French breeds from African breeds. The second axis separates African breeds into three groups: Taurine breeds, Borgu and Zebu. The intermediate position of the Borgu is explained because this breed is an African shorthorn $\times$ Zebu crossbred. The third axis separates French breeds into three clusters. The first cluster is mainly composed of southwestern French breeds and the Montbeliarde breed, the second is composed of Charolaise and Bretonne Pie Noire breeds and the third distinguishes the Maine-Anjou breed. Note that these clusters mainly fit with history and geography except for the Charolaise and Bretonne Pie Noire cluster.

The relationship between a single marker analysis (Fig. 2) and the MCOA (Fig. 3a) is illustrated by a cohesion plot, which is the superimposition of the

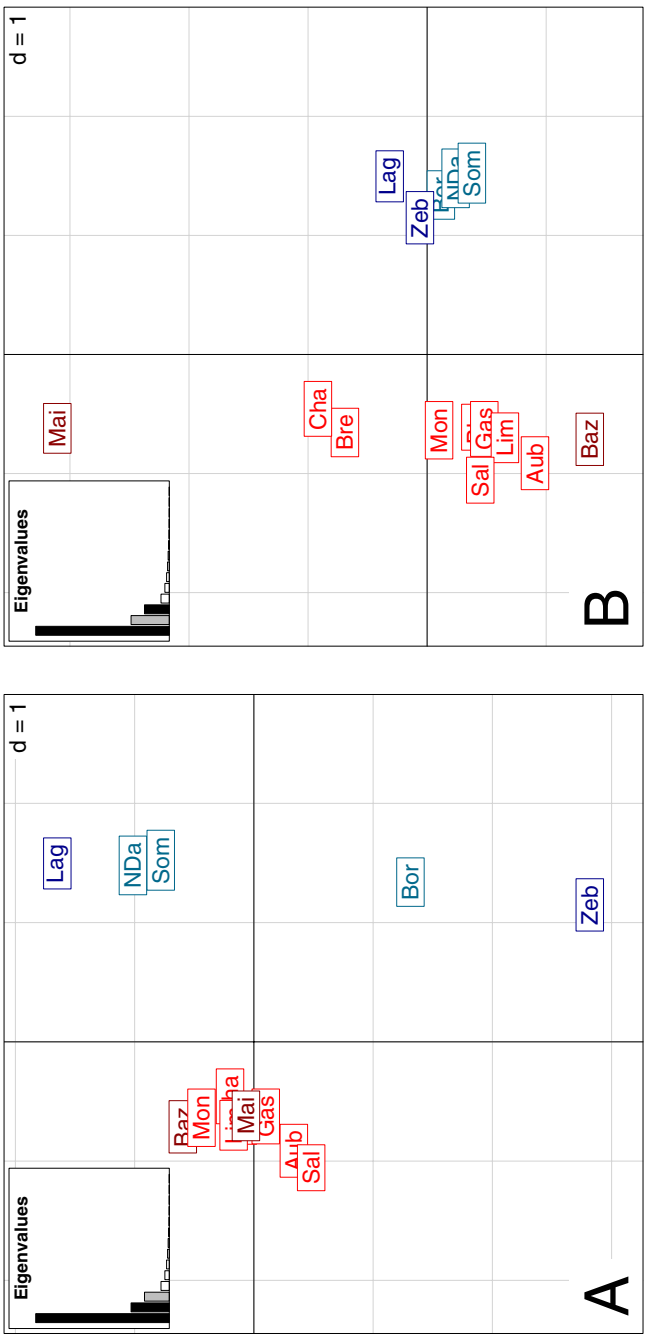


Figure 3. Reference scores of the multiple co-inertia analysis, displaying the most congruent structures among markers, on the planes 1–2 (Fig. 3A) and 1–3 (Fig. 3B). A common scale is used (d = 1) for both plots. The colors indicate African breeds in blue and French breeds in red (for the figure in color see online version).

two corresponding plots (Fig. 4). In this figure, the location of each data point can be indicated using an arrow. The tip of the arrow is used to show a location in the single marker analysis and the start of the arrow is the location of the breed in MCOA analysis. If both typologies strongly agree, the arrows would be short. Equally, a long arrow demonstrates a locally weak relationship among structures.

Of the six microsatellites, INRA35 exhibits the longest arrows and is thus the less congruent marker. With the MM12 marker, the direction of the arrows is mainly horizontal, showing discrepancies along the first axis (separation between France and Africa), while there is a good adequacy for the second axis (separation between African taurine breeds and zebu breeds). However, HEL1 reproduces the reference almost perfectly. HEL13 is also a structuring marker for all the breeds except for the Bazadaise breed, which is clustered with African taurine breeds.

Diagrams of typological values are plotted in Figures 5A (1^{st} axis), 5B (2^{nd} axis) and 5C (3^{rd} axis). The heterogeneity of typological values increases with the number of the axis. In order to obtain a total percentage equal or greater than 50%, nine markers are needed for axis 1, eight markers for axis 2, and only six for axis 3. Minimum value is close to 0 for the three axes (0.11% (INRA35), 0.07% (SPS115) and 0.02% (ILSTS005) for axes 1 to 3, respectively). The maximum percentage (8.3%) for axis 1 is reached by HEL13. This marker is also the most important for axis 2, with a typological value percentage equal to 9.0%. For axis 3, the typological values reach a maximum percentage of 11.5%, for HEL5.

Some markers do not contribute to the population structuring, whatever the axes: INRA35, INRA5 and SPS115. However, the typological values vary according to the structures. For example, HEL13, which is the most important marker for axes 1 and 2, is among the worst markers for axis 3 (typological value percentage of 0.21%). Conversely, HEL5 is the most important marker for axis 3, but not for axes 1 and 2. MM12 contributes mostly to axis 2, but not to the other axes.

Thus, efficient markers for distinguishing African from French breeds are not necessarily the same as for distinguishing within Africa or within France. Correlations between typological values vary from 0.55 (axis 1 – axis 2) to −0.13 (axis 2 – axis 3). However, typological values are robust with respect to the set of populations that are involved in the analysis. Analyzing the subset of French populations leads to typological values that are very well correlated with the whole dataset ($r = 0.89$).

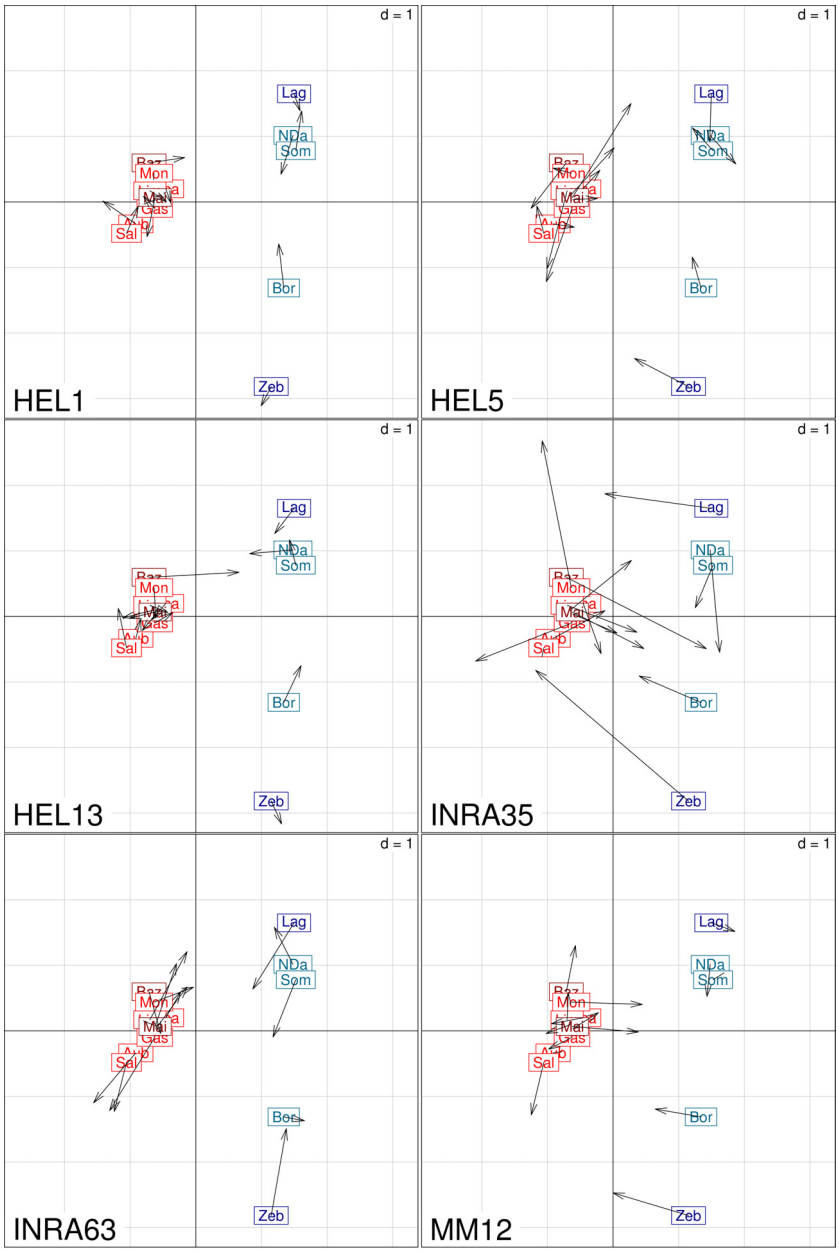


Figure 4. Cohesion plots showing the differences between the reference typology (labels and arrows origin) and the coordinated single-marker analyses (normed scores) on the first two axes. The arrows represent the typological "mistakes" displayed by the markers. The longer an arrow is, the greater the mistake is. A common scale is used (d = 1) for all plots.

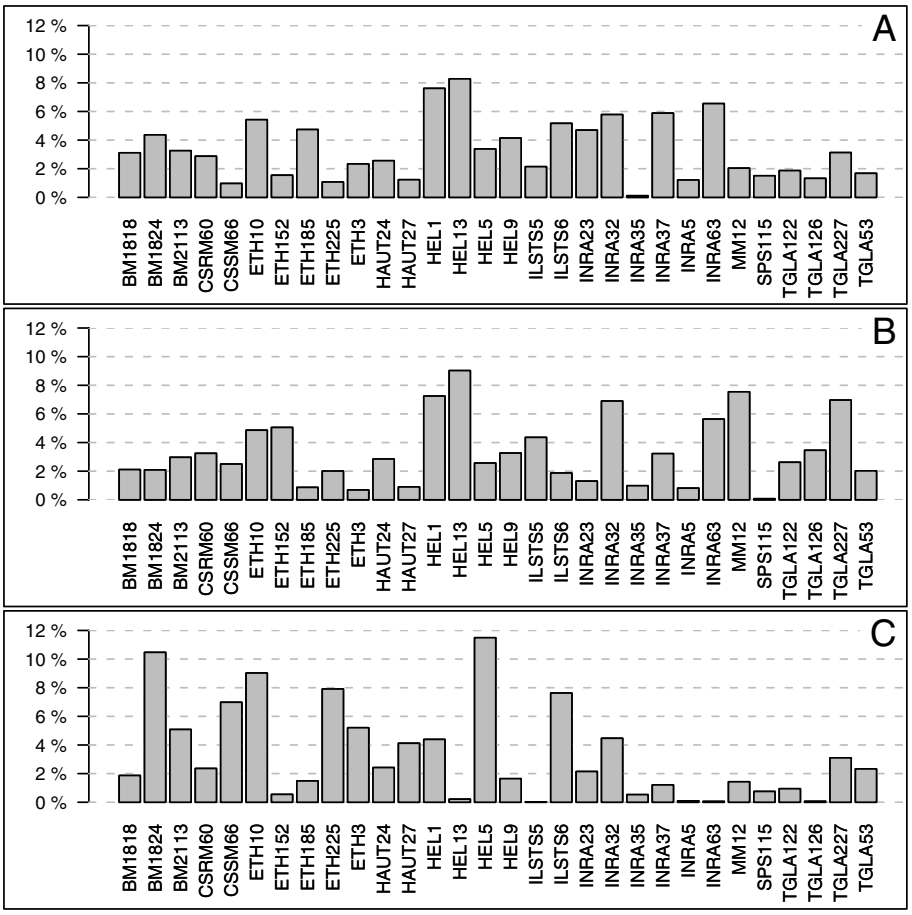


Figure 5. Diagrams of typological values components, in percentages, for the three reference structures, corresponding to (A) Africa-France separation (B) within Africa differentiation and (C) within France differentiation.

4. CONCLUSION

In this paper, we describe the MCOA in the context of a population genetic structuring analysis. This methodology is easy to use and could be of general applicability for livestock species. The efficiency of a set of markers is addressed with graphical tools and quantitative measures. This method is implemented in the ade4 package [18] of the R software [54].

This method is independent of the mutation model of the markers used, and thus can be applied to various types of markers (*e.g.*, proteins, blood groups,

microsatellites, amplified fragment length polymorphism, single nucleotide polymorphisms).

The choice of a weighting scheme should be thought according to the nature of the markers involved in the study. A uniform weighting may be sensible if only one type of markers is used, as in this paper. However, weighting each marker by its total inertia will give the same scale of differentiation for each marker. These two weighting options are available in the ade4 package. Moreover, thanks to the flexibility of the method, the user may supply any weighting scheme of his/her own choice, which could be based, for instance, on the number of alleles of the marker.

Separate coordinated plots show how the markers separate the populations regarding a common structure, while superimposed plots visually address the discrepancies among the common structure and one single-marker structure.

The quantitative measure of typological value includes two aspects: the ability to perform a typology of populations and the degree of congruence with the reference. Population structure is more easily exhibited using markers with high typological values, than using those with low values. We show that efficient markers in one collection of populations do not remain efficient in others. Typological values of markers are structure-dependent. When strongly different populations such as French and African populations are considered, all markers roughly equally reproduce the main features of the typology. However, this is not the case for closely related populations because only a few markers reproduce the reference typology. Thus, caution is needed in evaluating populations based on molecular studies if a small number of efficient loci are used. These results contradict the idea [61, 62] that increasing the number of markers will increase the reliability of the typology analysis: quantity is not quality.

As such, a marker selection method based on the typological value should select an efficient, not to say the most efficient, subset of markers for exhibiting a consensus population structuring. In this respect, a general algorithm, and particularly stopping rules for determining an optimum number of selected markers should be investigated, as in [38,40] or [66] in a classical PCA context.

Towards a quality process, it is important to check ***data*** (sampling strategy, DNA, experimental protocol, tracking of genotyping errors [53], standardization of data), ***tools*** (choice of markers [58]), ***methods*** (suitability of the method to the data and scientific goal [61,71]) and ***the computer programs*** (well established and recommended by experts [21,32]). This process has been initiated in livestock species by FAO guidelines [24], including recommended ISAG/FAO sets of genetic markers for domestic species. In this respect, MCOA should

play a major role in the choice of panels of markers, which is essential for an efficient design of population genetic analyses of species. A large number of genetic diversity studies for livestock species has been carried out, some concern livestock from a single country [23, 41, 67], others have examined diversity and distribution of livestock at the regional level [13, 22, 26] or even at the scale of nearly an entire continent or all over the world [16,28,31,63]. Since such studies are still continuing and have financial constraints, it is important to have a measure that permits the elimination of non efficient markers from studies. If no previous data are available, another application of the MCOA is to study a subset of the populations, and remove the less informative markers when completing the analysis. Luikart *et al.* [44] advocate the importance of identifying "outlier loci" to avoid biased estimates of population parameters. With that respect, MCOA and typological values should also be efficient tools to differentiate neutral markers from markers likely to be selected from the selection of a subset of markers, or for the comparison of the degree of differentiation in neutral marker loci and genes coding quantitative traits [58, 64].

ACKNOWLEDGEMENTS

We acknowledge the assistance of the respective breeders associations in the collection of French cattle samples. We acknowledge the following persons for their help in planning and conducting the sampling missions for African samples: V. Codja (Bénin), N.T. Kouagou (Togo), I. Sidibé (Burkina Faso). We also thank J.A. Lenstra for his coordination of the wider European project. This work was funded by the European Commission, Contract CT98-118, Inco-Dc Erbic 18Ct960031 and by the Bureau des Ressources Génétiques and the French Ministry in charge of Ecology (MEDD), Contract 14-A/2003. We thank Daniel Chessel for his very helpful comments and interesting discussions about compositional data and the multiple co-inertia analysis. We thank two referees for a thorough review and constructive comments.

REFERENCES

[1] Aitchison J., Principal component analysis of compositional data, Biometrika 70 (1983) 57–65.

[2] Aitchison J., Logratios and natural laws in compositional data analysis, Math. Geol. 31 (1999) 563–589.

[3] Aitchison J., Greenacre M., Biplot of compositional data, Appl. Stat. 51 (2002) 375–392.

[4] Bady P., Dolédec S., Dumont B., Fruget J.-F., Multiple co-inertia analysis: a tool for assessing synchrony in the temporal variability of aquatic communities, C. R. Biol. 327 (2004) 29–36.

[5] Baumung R., Simianer H., Hoffmann I., Genetic diversity studies in farm animals – a survey, J. Anim. Breed. Genet. 121 (2004) 361–373.

[6] Baumung R., Cubric-Curik V., Schwend K., Achmann R., Solkner J., Genetic characterisation and breed assignment in Austrian sheep breeds using microsatellite marker information, J. Anim. Breed. Genet. 123 (2006) 265–271.

[7] Beaumont M.A., Recent developments in genetic data analysis: what can they tell us about human demographic history? Heredity 92 (2004) 365–379.

[8] Beja-Pereira A., Alexandrino P., Bessa I., Carretero Y., Dunner S., Ferrand N., Jordana J., Laloë D., Moazami-Goudarzi K., Sanchez A., Canon J., Genetic characterization of southwestern European bovine breeds: a historical and biogeographical reassessment with a set of 16 microsatellites, J. Hered. 94 (2003) 243–250.

[9] Beja-Pereira A., Caramelli D., Lalueza-Fox C., Vernesi C., Ferrand N., Casoli A., Goyache F., Royo L.J., Conti S., Lari M., Martini A., Ouragh L., Magid A., Atash A., Zsolnai A., Boscato P., Triantaphylidis C., Ploumi K., Sineo L., Mallegni F., Taberlet P., Erhardt G., Sampietro L., Bertranpetit J., Barbujani G., Luikart G., Bertorelle G., The origin of European cattle: evidence from modern and ancient DNA, Proc. Natl. Acad. Sci. USA 103 (2006) 8113–8118.

[10] Bennewitz J., Kantanen J., Tapio I., Li M.H., Kalm E., Vilkki J., Ammosov I., Ivanova Z., Kiselyova T., Popov R., Meuwissen T.H., Estimation of breed contributions to present and future genetic diversity of 44 North Eurasian cattle breeds using core set diversity measures, Genet. Sel. Evol. 38 (2006) 201–220.

[11] Bruford M.W., Bradley D.G., Luikart G., DNA markers reveal the complexity of livestock domestication, Nat. Rev. Genet. 4 (2003) 900–910.

[12] Bryant D., A classification of consensus methods for phylogenies, in: Janowitz M., Lapointe F.-J., McMorris F.R., Mirkin B., Roberts F.S. (Eds.), Bioconsensus, DIMACS, AMS, 2003, pp. 163–184.

[13] Canon J., Alexandrino P., Bessa I., Carleos C., Carretero Y., Dunner S., Ferran N., Garcia D., Jordana J., Laloë D., Pereira A., Sanchez A., Moazami-Goudarzi K., Genetic diversity measures of local European beef cattle breeds for conservation purposes, Genet. Sel. Evol. 33 (2001) 311–332.

[14] Cavalli-Sforza L.L., The Human Genome Diversity Project: past, present and future, Nat. Rev. Genet. 6 (2005) 333–340.

[15] Cavalli-Sforza L.L., Menozzi P., Piazza A., The history and geography of human genes, Princeton University Press, 1994, 1088 p.

[16] Chenyambuga S.W., Hanotte O., Hirbo J., Watts P.C., Kemp S.J., Kifaro G.C., Gwakisa P.S., Peterson P.H., Rege J.E.O., Genetic characterization of indigenous goats of Sub-Saharan Africa using microsatellite DNA markers, Asian-Australas. J. Anim. Sci. 17 (2004) 445–452.

[17] Chessel D., Hanafi M., Analyses de la co-inertie de K nuages de points, Rev. Stat. Appl. 44 (1996) 35–60.

[18] Chessel D., Dufour A.B., Thioulouse J., The ade4 package - I: One-table methods, R-News 4 (2004) 5–10.
[19] De Crespin de Billy V., Dolédec S., Chessel D., Biplot presentation of diet composition data: an alternative for fish stomach contents analysis, J. Fish Biol. 56 (2000) 961–973.
[20] DeYoung R.W., Demarais S., Honeycutt R.L., Rooney A.P., Gonzales R.A., Gee K.L., Genetic consequences of white-tailed deer (*Odocoileus virginianus*) restoration in Mississippi, Mol. Ecol. 12 (2003) 3237–3252.
[21] Excoffier L., Heckel G., Computer programs for population genetics data analysis: a survival guide, Nat. Rev. Genet. 7 (2006) 745–758.
[22] Fabuel E., Barragan C., Silio L., Rodriguez M.C., Toro M.A., Analysis of genetic diversity and conservation priorities in Iberian pigs based on microsatellite markers, Heredity 93 (2004) 104–113.
[23] Fang M., Hu X., Jiang T., Braunschweig M., Hu L., Du Z., Feng J., Zhang Q., Wu C., Li N., The phylogeny of Chinese indigenous pig breeds inferred from microsatellite markers, Anim. Genet. 36 (2005) 7–13.
[24] FAO, Secondary Guidelines for Development of National Farm Animal Genetics Resources Management Plans. Measurement of Domestic Animal Diversity (MoDAD), Original Working Group Report, FAO, Rome, 1998.
[25] Felsenstein J., How can we infer geography and history from gene frequencies? J. Theor. Biol. 96 (1982) 9–20.
[26] Freeman A.R., Meghen C.M., Machugh D.E., Loftus R.T., Achukwi M.D., Bado A., Sauveroche B., Bradley D.G., Admixture and diversity in West African cattle populations, Mol. Ecol. 13 (2004) 3477–3487.
[27] Gabriel K.R., The biplot graphic display of matrices with application to principal component analysis, Biometrika 58 (1971) 453–467.
[28] Hanotte O., Bradley D.G., Ochieng J.W., Verjee Y., Hill E.W., Rege J.E., African pastoralism: genetic imprints of origins and migrations, Science 296 (2002) 336–339.
[29] Healy M.J.R., Drawing a probability ellipse, J. R. Stat. Soc. Ser. C-Appl. Stat. 21 (1972) 202–204.
[30] Hedde M., Lavelle P., Joffre R., Jiménez J.J., Decaëns T., Specific functional signature in soil macro-invertebrate biostructures, Funct. Ecol. 19 (2005) 785–793.
[31] Hillel J., Groenen M.A., Tixier-Boichard M., Korol A.B., David L., Kirzhner V.M., Burke T., Barre-Dirie A., Crooijmans R.P., Elo K., Feldman M.W., Freidlin P.J., Maki-Tanila A., Oortwijn M., Thomson P., Vignal A., Wimmers K., Weigend S., Biodiversity of 52 chicken populations assessed by microsatellite typing of DNA pools, Genet. Sel. Evol. 35 (2003) 533–557.
[32] Holmes S., Multivariate Data Analysis: The French Way, to appear in Festchrift for David Freeman, IMS lecture notes (2006) http://www-stat.stanford.edu/~susan/papers/dfc.pdf [consulted: 20 March 2007].
[33] Hotelling H., Analysis of a complex of statistical variables into principal components, J. Educ. Psychol. 24 (1933) 417–441.
[34] Hotelling H., Analysis of a complex of statistical variables into principal components (continued from September issue), J. Educ. Psychol. 24 (1933) 498–520.

[35] Iamartino D., Bruzzone A., Lanza A., Blasi M., Pilla F., Genetic diversity of Southern Italian goat populations assessed by microsatellite markers, Small Ruminant Res. 57 (2005) 249–255.

[36] Ibeagha-Awemu E.M., Jann O.C., Weimann C., Erhardt G., Genetic diversity, introgression and relationships among West/Central African cattle breeds, Genet. Sel. Evol. 36 (2004) 673–690.

[37] Jann O.C., Ibeagha-Awemu E.M., Ozbeyaz C., Zaragoza P., Williams J.L., Ajmone-Marsan P., Lenstra J.A., Moazami-Goudarzi K., Erhardt G., Geographic distribution of haplotype diversity at the bovine casein locus, Genet. Sel. Evol. 36 (2004) 243–257.

[38] Joliffe I.T., Discarding variables in a principal component analysis. I: Artificial data., Appl. Stat. 22 (1972) 373–374.

[39] Knowles L.L., The burgeoning field of statistical phylogeography, J. Evol. Biol. 17 (2004) 1–10.

[40] Krzanowski W.J., A stopping rule for structure-preserving variable selection, Stat. Comput. 6 (1996) 51–56.

[41] Kumar S., Gupta J., Kumar N., Dikshit K., Navani N., Jain P., Nagarajan M., Genetic variation and relationships among eight Indian riverine buffalo breeds, Mol. Ecol. 15 (2006) 593–600.

[42] Li M.H., Zhao S.H., Bian C., Wang H.S., Wei H., Liu B., Yu M., Fan B., Chen S.L., Zhu M.J., Li S.J., Xiong T.A., Li K., Genetic relationships among twelve Chinese indigenous goat populations based on microsatellite analysis, Genet. Sel. Evol. 34 (2002) 729–744.

[43] Liron J.P., Peral-Garcia P., Giovambattista G., Genetic characterization of Argentine and Bolivian Creole cattle breeds assessed through microsatellites, J. Hered. 97 (2006) 331–339.

[44] Luikart G., England P.R., Tallmon D., Jordan S., Taberlet P., The power and promise of population genomics: from genotyping to genome typing, Nat. Rev. Genet. 4 (2003) 981–994.

[45] MacHugh D.E., Loftus R.T., Cunningham P., Bradley D.G., Genetic structure of seven European cattle breeds assessed using 20 microsatellite markers, Anim. Genet. 29 (1998) 333–340.

[46] Marletta D., Tupac-Yupanqui I., Bordonaro S., Garcia D., Guastella A.M., Criscione A., Canon J., Dunner S., Analysis of genetic diversity and the determination of relationships among western Mediterranean horse breeds using microsatellite markers, J. Anim. Breed. Genet. 123 (2006) 315–325.

[47] Martin-Burriel I., Garcia-Muro E., Zaragoza P., Genetic diversity analysis of six Spanish native cattle breeds using microsatellites, Anim. Genet. 30 (1999) 177–182.

[48] Menozzi P., Piazza A., Cavalli-Sforza L.L., Synthetic maps of human gene frequencies in europeans, Science 201 (1978) 786–792.

[49] Moazami-Goudarzi K., Laloë D., Is a multivariate consensus representation of genetic relationships among populations always meaningful? Genetics 162 (2002) 473–484.

[50] Moazami-Goudarzi K., Laloë D., Furet J.P., Grosclaude F., Analysis of genetic relationships between 10 cattle breeds with 17 microsatellites, Anim. Genet. 28 (1997) 338–345.

[51] Moazami-Goudarzi K., Belemsaga D., Ceriotti G., Laloë D., Fagbohoun F., Kouagou N., Sidibé I., Codjia V., Crimella M., Grosclaude F., Touré S., Caractérisation de la race bovine Somba à l'aide de marqueurs moléculaires, Rev. Elev. Med. Vet. Pays Trop. 54 (2001) 1–10.

[52] Phillips C., Warnow T.J., The asymetric median tree - a new model for building consensus trees, Discrete Appl. Math. 71 (1996) 311–335.

[53] Pompanon F., Bonin A., Bellemain E., Taberlet P., Genotyping errors: causes, consequences and solutions, Nat. Rev. Genet. 6 (2005) 847–859.

[54] R Development Core Team, R: a Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria (2006), ISBN 3-900051-07-0, http://www.R-project.org.

[55] Reyment R., The statistical analysis of multivariate serological frequency data, Bull. Math. Biol. 67 (2005) 1303–1313.

[56] Rosenberg N.A., Pritchard J.K., Weber J.L., Cann H.M., Kidd K.K., Zhivotovsky L.A., Feldman M.W., Genetic structure of human populations, Science 298 (2002) 2381–2385.

[57] Royo L.J., Alvarez I., Beja-Pereira A., Molina A., Fernandez I., Jordana J., Gomez E., Gutierrez J.P., Goyache F., The origins of Iberian horses assessed via mitochondrial DNA, J. Hered. 96 (2005) 663–669.

[58] Schlötterer C., The evolution of molecular markers-just a matter of fashion? Nat. Rev. Genet. 5 (2004) 63–69.

[59] Simianer H., Using expected allele number as objective function to design between and within breed conservation of farm animal biodiversity, J. Anim. Breed. Genet. 122 (2005) 177–187.

[60] Simianer H., Meyer J.N., Past and future activities to harmonize farm animal biodiversity studies on global scale, Arch. Zootec. 52 (2003) 193–199.

[61] Takezaki N., Nei M., Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA, Genetics 144 (1996) 389–399.

[62] Talle S.B., Chenyabuga W.S., Fimland E., Syrstad O., Meuwissen T., Klungland H., Use of DNA technologies for the conservation of animal genetic resources: A review, Acta Agric. Scand. Sect. A Anim. Sci. 55 (2005) 1–8.

[63] Tapio M., Miceikiene I., Vilkki J., Kantanen J., Comparison of microsatellite and blood protein diversity in sheep: inconsistencies in fragmented breeds, Mol. Ecol. 12 (2003) 2045–2056.

[64] Toro M.A., Caballero A., Characterization and conservation of genetic diversity in subdivided populations, Philos. Trans. R. Soc. Lond. B Biol. Sci. 360 (2005) 1367–1378.

[65] Uzun M., Gutierrez-Gil B., Arranz J.J., San Primitivo F., Saatci M., Kaya M., Bayon Y., Genetic relationships among Turkish sheep, Genet. Sel. Evol. 38 (2006) 513–524.

[66] Wang A., Gehan E.A., Gene selection for microarray data analysis using principal component analysis., Stat. Med. 24 (2005) 2069–2087.

[67] Wiener P., Burton D., Williams J.L., Breed relationships and definition in British cattle: a genetic analysis, Heredity 93 (2004) 597–602.
[68] Woronow A., Regression and discrimination analysis using raw compositional data – Is it really a problem? in: Proceedings of the 3rd Annual Conference of the International Association for Mathematical Geology, Universitate Politecnica de Catalunya, Barcelona, 1997, pp. 157–162.
[69] Xuebin Q., Jianlin H., Lkhagva B., Chekarova I., Badamdorj D., Rege J.E., Hanotte O., Genetic diversity and differentiation of Mongolian and Russian yak populations, J. Anim. Breed. Genet. 122 (2005) 117–126.
[70] Yue-Hui M., Shao-Qi R., Shen-Jin L., Guan-Yu H., Wei-Jun G., Hong-Bin L., Xia L., Qian-Jun Z., Jun G., Phylogeography and origin of sheep breeds in Northern China, Conserv. Genet. 7 (2006) 117–127.
[71] Zhang D.X., Hewitt G.M., Nuclear DNA analyses in genetic studies of populations: practice, problems and prospects, Mol. Ecol. 12 (2003) 563–584.

APPENDIX: MCOA PRINCIPLES

Notations: We consider K tables $\mathbf{X}_k$ having the same rows, but different columns. Each table defines a cloud of n points in a p_k-dimensional space of real numbers, $\mathbb{R}^{p_k}$. Distances between two points in $\mathbb{R}^{p_k}$ are computed using $\mathbf{Q}_k$ metric ($p_k \times p_k$).

Let $\mathbf{D}$ be a $n \times n$ diagonal matrix containing the weights of the n points and used to compute the distances between the variables in the $\mathbb{R}^n$ space.

Let w_k be the weight of each table. Here we used the uniform weighting $w_k = \frac{1}{K}$ for $k = 1, K$.

The tables $\mathbf{X}_k$ are centered by columns. We note $\mathbf{X}_k^{\mathrm{T}}$ the transposed matrix of $\mathbf{X}_k$.

First step: The aim of the MCOA is to find a set of $\mathbf{Q}_k$-normed vectors in each space $\mathbb{R}^{p_k}$, called the co-inertia axes ($\mathbf{u}_1^1 ... \mathbf{u}_k^1 ... \mathbf{u}_K^1$), and a reference vector $\mathbf{v}^1$ $\mathbf{D}$-normed in $\mathbb{R}^n$ maximizing:

$$\sum_{k=1}^{K} w_k (\mathbf{X}_k \mathbf{Q}_k \mathbf{u}_k^1 | \mathbf{v}^1)_{\mathbf{D}}^2,$$

where $(\mathbf{X}_k \mathbf{Q}_k \mathbf{u}_k^1 | \mathbf{v}^1)_{\mathbf{D}}$ is the scalar product of $\mathbf{X}_k \mathbf{Q}_k \mathbf{u}_k^1$ and $\mathbf{v}^1$ computed with the $\mathbf{D}$ metric. The vectors are centered and then, this scalar product is a covariance. Note that row scores onto co-inertia axes are the scores of the coordinated analyses: $\mathbf{X}_k \mathbf{Q}_k \mathbf{u}_k^1 = \mathbf{l}_k^1$.

Let us consider the matrix $\mathbf{Y}_1$ composed of the juxtaposed weighted tables:

$$\mathbf{Y}_1 = \left[\sqrt{w_1}\, \mathbf{X}_1 \middle| \ldots \middle| \sqrt{w_k}\, \mathbf{X}_k \middle| \ldots \middle| \sqrt{w_K}\, \mathbf{X}_K \right].$$

Chessel and Hanafi [17] showed that $\sum_{k=1}^{K} w_k(\mathbf{X}_k\mathbf{Q}_k\mathbf{u}_k^1|\mathbf{v}^1)^2_{\mathbf{D}}$ is maximum for λ_1, the first eigenvalue of the PCA of $\mathbf{Y}_1$.

So,

– the reference score $\mathbf{v}^1$ is the first principal component of this PCA and,

– the vectors $\mathbf{u}_k^1$are obtained by: $\mathbf{u}_k^1 = \frac{\mathbf{X}_k^{\mathrm{T}}\mathbf{D}\mathbf{v}^1}{\|\mathbf{X}_k^{\mathrm{T}}\mathbf{D}\mathbf{v}^1\|_{\mathbf{Q}_k}}$.

Following steps: We note r the number of chosen structures in the reference, *i.e.* the total number of steps ($i = 1, r$). The aim of the MCOA is to find another set of $\mathbf{Q}_k$-normed co-inertia axes ($\mathbf{u}_1^i ... \mathbf{u}_k^i ... \mathbf{u}_K^i$), and a reference vector $\mathbf{v}^i$ $\mathbf{D}$-normed in $\mathbb{R}^n$ maximizing:

$$\sum_{k=1}^{K} w_k(\mathbf{X}_k\mathbf{Q}_k\mathbf{u}_k^i|\mathbf{v}^i)^2_{\mathbf{D}}$$

under the additional constraints that $\mathbf{u}_k^i$ is orthogonal to $\mathbf{u}_k^{i-1}$ and $\mathbf{v}^i$ is orthogonal to $\mathbf{v}^{i-1}$.

Let us consider the orthonormal co-inertia basis $\mathbf{U}_k = \left\{\mathbf{u}_k^1, ..., \mathbf{u}_k^{i-1}\right\}$ for each table $\mathbf{X}_k$.

Let $\mathbf{P}_k$ be the projector onto $\mathbf{U}_k$.

Let us then consider the matrix $\mathbf{Y}_i$ composed of the juxtaposed weighted tables:

$$\mathbf{Y}_i = \left[\sqrt{w_1}\,\mathbf{X}_1 - \sqrt{w_1}\,\mathbf{X}_1\mathbf{P}_1^{\mathrm{T}} \middle| \ldots \middle| \sqrt{w_k}\,\mathbf{X}_k - \sqrt{w_k}\,\mathbf{X}_k\mathbf{P}_k^{\mathrm{T}} \middle| \ldots \middle| \sqrt{w_K}\,\mathbf{X}_K - \sqrt{w_K}\,\mathbf{X}_K\mathbf{P}_K^{\mathrm{T}} \right].$$

Chessel and Hanafi [17] showed that $\sum_{k=1}^{K} w_k(\mathbf{X}_k\mathbf{Q}_k\mathbf{u}_k^i|\mathbf{v}^i)^2_{\mathbf{D}}$ is maximum for λ_i, the first eigenvalue of the PCA of $\mathbf{Y}_i$.

So,

– the reference score $\mathbf{v}^i$ is the first principal component of this PCA and,

– the vectors $\mathbf{u}_k^i$ are obtained by the following: $\mathbf{u}_k^i = \frac{\mathbf{X}_k^{\mathrm{T}}\mathbf{D}\mathbf{v}^i}{\|\mathbf{X}_k^{\mathrm{T}}\mathbf{D}\mathbf{v}^i\|_{\mathbf{Q}_k}}$.

Finally, MCOA yields r orthonormal row scores $\mathbf{V} = \left\{\mathbf{v}^1, ..., \mathbf{v}^r\right\}$ (the reference scores), r orthonormal co-inertia axes $\mathbf{U}_k = \left\{\mathbf{u}_k^1, ..., \mathbf{u}_k^r\right\}$ in each p_k-dimensional space and the corresponding row scores $\mathbf{L}_k = \left\{\mathbf{l}_k^1, ..., \mathbf{l}_k^r\right\}$ (the scores of the coordinated analyses).

2.3 Discussion

Comme il a été souligné plus haut, l'ACOM n'est pas la seule approche K-tableaux pertinente dans l'analyse des données de marqueurs génétiques. On se propose ci-dessous de comparer les résultats fournis par l'AFM et par STATIS sur les données de Laloë *et al.* (2007). Les analyses présentées sont réalisées avec le logiciel R (R Development Core Team, 2008) et les packages *ade4* et *adegenet*.

2.3.1 Illustration

Les données illustrant l'ACOM dans l'article qui précède constituent le jeu de données `microbov` du package *adegenet*. On commence par charger les données :

```
> library(adegenet)
> data(microbov)
> microbov


   ######################
   ### Genind object ###
   ######################
- genotypes of individuals -

S4 class:  genind
@call: genind(tab = truenames(microbov)$tab, pop = truenames(microbov)$pop)

@tab:  704 x 373 matrix of genotypes

@ind.names: vector of  704 individual names
@loc.names: vector of  30 locus names
@loc.nall: number of alleles per locus
@loc.fac: locus factor for the  373 columns of @tab
@all.names: list of  30 components yielding allele names for each locus
@ploidy:  2

Optionnal contents:
@pop:  factor giving the population of each individual
@pop.names:  factor giving the population of each individual

@other: a list containing: coun  breed  spe
```

L'objet `microbov` contient 704 génotypes pour 30 marqueurs microsatellites totalisant 373 allèles. On sait également à laquelle des races suivantes les génotypes appartiennent :

```
> microbov@pop.names


            P01             P02             P03             P04
      "Borgou"          "Zebu"      "Lagunaire"         "NDama"
            P05             P06             P07             P08
       "Somba"        "Aubrac"       "Bazadais" "BlondeAquitaine"
            P09             P10             P11             P12
 "BretPieNoire"     "Charolais"       "Gascon"      "Limousin"
            P13             P14             P15
  "MaineAnjou"   "Montbeliard"       "Salers"
```

On commence par obtenir les fréquences alléliques par race :

```
> obj <- genind2genpop(microbov)


 Converting data from a genind to a genpop object...

...done.


> bov.freq <- makefreq(obj, missing = "mean")$tab
```

```
 Finding allelic frequencies from a genpop object...

...done.
```

Après centrage des fréquences, `bov.freq` est converti en objet `ktab`, qui est la classe d'objet utilisée dans *ade4* pour gérer les données K-tableaux.

```
> bov.freq <- scalewt(bov.freq, scale = FALSE)
> bov.ktab <- ktab.data.frame(as.data.frame(bov.freq), microbov$loc.nall,
+     colnames = unlist(microbov$all.names), tabnames = microbov$loc.names)
> class(bov.ktab)
```

```
[1] "ktab"
```

```
> names(bov.ktab)
```

```
 [1] "INRA63"  "INRA5"   "ETH225"  "ILSTS5"  "HEL5"    "HEL1"    "INRA35"
 [8] "ETH152"  "INRA23"  "ETH10"   "HEL9"    "CSSM66"  "INRA32"  "ETH3"
[15] "BM2113"  "BM1824"  "HEL13"   "INRA37"  "BM1818"  "ILSTS6"  "MM12"
[22] "CSRM60"  "ETH185"  "HAUT24"  "HAUT27"  "TGLA227" "TGLA126" "TGLA122"
[29] "TGLA53"  "SPS115"  "lw"      "cw"      "blo"     "TL"      "TC"
[36] "T4"      "call"
```

```
> bov.ktab$INRA63[1:5, 1:5]
```

```
                  167          171          173        175         177
Borgou    -0.0006535948 -0.001361111 -0.004947064 -0.3696217 -0.04745415
Zebu      -0.0006535948  0.008638889 -0.004947064 -0.3396217 -0.15745415
Lagunaire  0.0091503268 -0.001361111 -0.004947064 -0.2527590  0.11391840
NDama     -0.0006535948 -0.001361111 -0.004947064 -0.3762884  0.33254585
Somba     -0.0006535948 -0.001361111 -0.004947064 -0.2896217  0.10254585
```

L'objet `bov.ktab` contient maintenant les fréquences alléliques centrées pour les 15 races de bovins, formant un tableau pour chaque marqueur. Par exemple, `bov.ktab$INRA63` contient les fréquences centrées des allèles 167, 171, ... du marqueur microsatellite INRA63.

On peut à présent effectuer l'AFM et l'analyse STATIS de ces données :

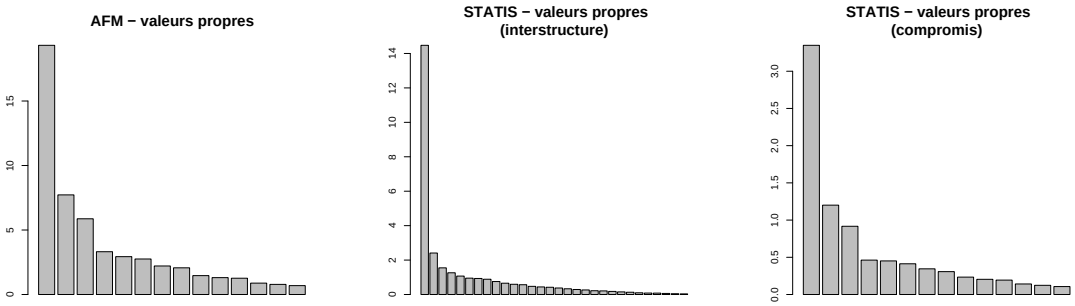
```
> bov.afm <- mfa(bov.ktab, scannf = FALSE, nf = 3)
> barplot(bov.afm$eig, main = "AFM - valeurs propres", cex.main = 1.5)
```

```
> bov.statis <- statis(bov.ktab, scannf = FALSE, nf = 3)
> barplot(bov.statis$RV.eig, main = "STATIS - valeurs propres\n(interstructure)",
+     cex.main = 1.5)
```

```
> barplot(bov.statis$C.eig, main = "STATIS - valeurs propres\n(compromis)",
+     cex.main = 1.5)
```

Ces figures sont assez cohérentes. L'AFM voit trois structures principales (Fig. 2.3a) dans les données, ce qui est confirmé par l'analyse du compromis de STATIS (Fig. 2.3c). En outre, STATIS confirme l'existence d'une structuration commune forte au sein des marqueurs (Fig. 2.3b).

On peut comparer la valeur typologique des marqueurs perçue par les deux méthodes. Ces indicateurs étant de nature différente, cette comparaison est nécessairement limitée.

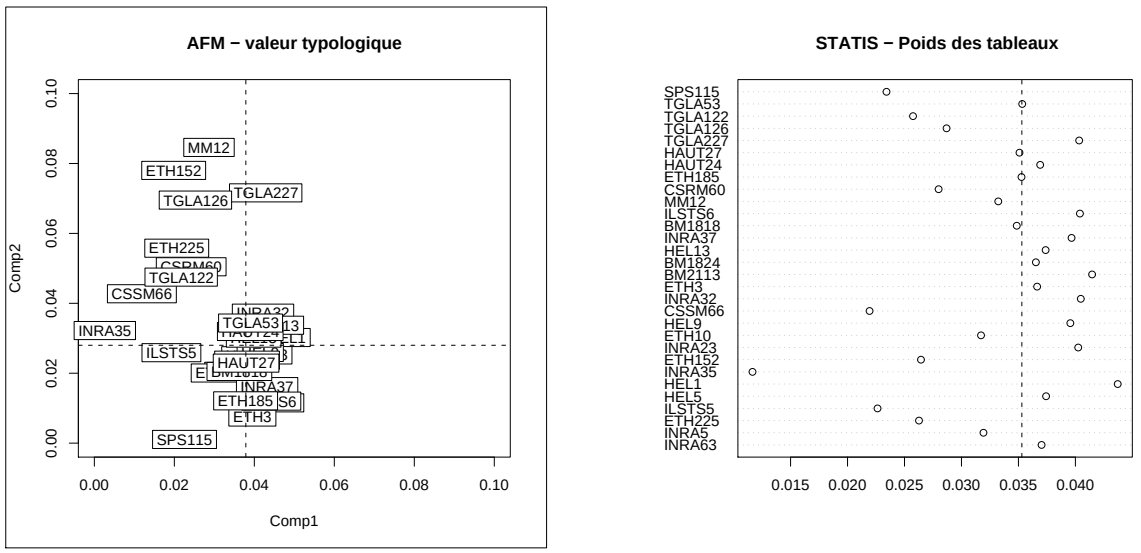


(a) AFM

(b) STATIS - interstructure. Valeurs propres de la matrice des RV (Fig. 2.2)

(c) STATIS - analyse du compromis. Valeurs propres de l'opérateur d'inertie moyen $\mathbf{W}$ (Fig. 2.2)

Fig. 2.3: Valeurs propres de l'analyse factorielle multiple (AFM) et de l'analyse STATIS des données `microbov`.



(a) AFM - Contribution à l'inertie totale des tableaux (Eqn. 2.6).

(b) STATIS - Pondération des tableaux pour la construction du compromis (Eqn. 2.15, u_k).

Fig. 2.4: Valeurs typologiques vues par l'analyse factorielle multiple (AFM) et l'analyse STATIS (données `microbov`). Les droites en pointillés indiquent les valeurs médianes sur chaque axe.

```
> typval.afm <- prop.table(as.matrix(bov.afm$link), 2)
> plot(typval.afm[, 1:2], type = "n", xlim = c(0, 0.1), ylim = c(0,
+     0.1), main = "AFM - valeur typologique")
> s.label(typval.afm[, 1:2], add.p = TRUE)
> abline(v = median(typval.afm[, 1]), lty = 2)
> abline(h = median(typval.afm[, 2]), lty = 2)
```

```
> typval.statis <- bov.statis$RV.tabw/sum(bov.statis$RV.tabw)
> dotchart(typval.statis, label = rownames(bov.statis$RV.coo),
+     main = "STATIS - Poids des tableaux")
> abline(v = median(typval.statis), lty = 2)
```

L'AFM donne une valeur typologique de chaque marqueur pour chaque composante de la typologie. En l'occurence, on constate que si la majorité des marqueurs contribue à parts égales à la première structure, la seconde composante est avant tout le fruit de quatre marqueurs dont le poids est élevé par rapport à l'ensemble (Fig. 2.4a, MM12, ETH152, etc.). La vision de

STATIS est plus globale : elle montre d'abord que l'ensemble de marqueurs est majoritairement cohérent, mais que certains participent peu au message commun (Fig. 2.4b, INRA35, CSSM66, SPS115, etc.). Ces résultats sont cohérents avec ceux obtenus par l'ACOM (Laloë *et al.*, 2007), tout en étant moins riches, l'ACOM donnant une vision plus précise de la contribution des marqueurs pour chaque structure.

On peut enfin s'intéresser à la typologie consensuelle identifiée par ces deux méthodes.

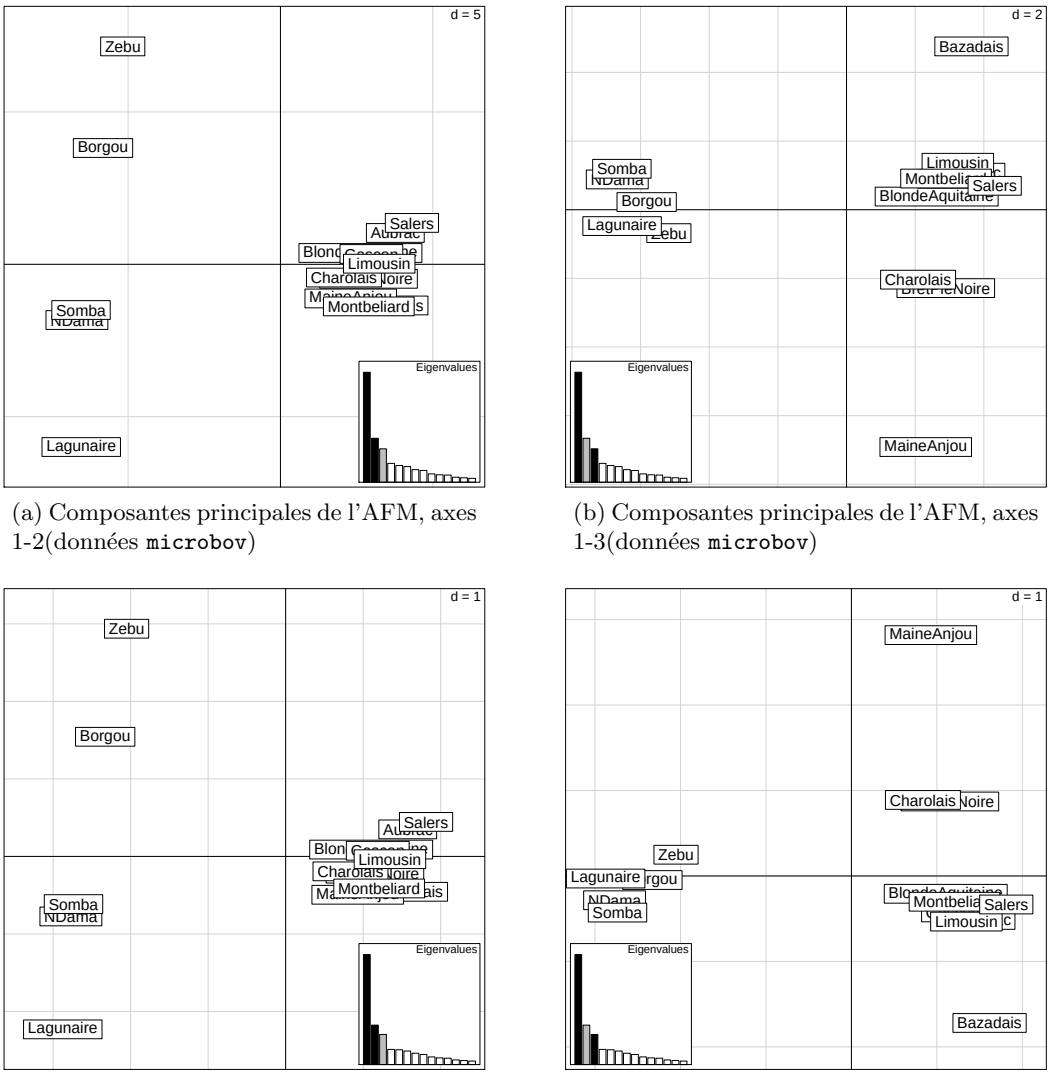
```
> s.label(bov.afm$li, clab = 1.2)
> add.scatter.eig(bov.afm$eig, 3, 1, 2, posi = "bottomright")

> s.label(bov.afm$li, 1, 3, clab = 1.2)
> add.scatter.eig(bov.afm$eig, 3, 1, 3)

> s.label(bov.statis$C.li, clab = 1.2)
> add.scatter.eig(bov.statis$C.eig, 3, 1, 2, posi = "bottomright")

> s.label(bov.statis$C.li, 1, 3, clab = 1.2)
> add.scatter.eig(bov.statis$C.eig, 3, 1, 3)
```

Les images "consensus" de la diversité génétique vue par les deux méthodes diffèrent peu entre elles (Fig. 2.5), et sont aussi semblables à la référence de l'ACOM (Laloë *et al.*, 2007, voir figure 3 de l'article).

2.3.2 Perspectives

Dans l'illustration qui précède, la structuration des données est dominée par deux axes de différenciation génétique au sein des races bovines. Bien que cette structuration fasse consensus entre les méthodes (AFM, STATIS, et ACOM), on voit déjà que les informations et les points de vue des méthodes diffèrent pour ce qui est de définir une valeur typologique des marqueurs. Ces différences transparaissent dans l'étude de Pavoine & Bailly (2007), qui ont poussé plus loin l'application des méthodes K-tableaux aux données moléculaires en proposant des extensions multi-tableaux de la double analyse en coordonnées principales (Pavoine *et al.*, 2004) suivant l'AFM, STATIS et l'ACOM. Un problème se pose donc : à la question posée, trois réponses sont proposées qui sont, si ce n'est différentes, du moins nuancées. Mais il est sans doute plus juste de percevoir ces différences comme une richesse plutôt qu'un inconvénient : elles procèdent simplement d'approches différentes du même problème. D'un point de vue pragmatique, on peut néanmoins se rassurer en observant que les trois méthodes sont unanimes quant à la principale question posée : les marqueurs génétiques peuvent fournir une information partiellement cohérente, et il est légitime de s'interroger, au cas par cas, sur l'existence de cette cohérence avant de réunir l'information en une seule analyse de la diversité génétique. C'est dans ce sens que Pavoine & Bailly (2007) concluent :

> *All three methods proposed can be used for a better description of inter-population genetic diversity measured over more than one locus. They imply a new reflection on the role of means in measures of diversity : can we work on average information over locus, or do we first need to examine the differences among the patterns of diversity given by the locus ? [...] can we build a unique, very synthetic measure of biodiversity, or do we have to make up our mind to define several conflicting measures ?*



(a) Composantes principales de l'AFM, axes 1-2(données `microbov`)

(b) Composantes principales de l'AFM, axes 1-3(données `microbov`)

(c) Compromis de l'analyse STATIS, axes 1-2(données `microbov`)

(d) Compromis de l'analyse STATIS, axes 1-3(données `microbov`)

FIG. 2.5: Typologies consensus de l'analyse factorielle multiple (AFM) et STATIS(données `microbov`).

Une autre question est soulevée par Pavoine & Bailly (2007), qui regrettent que STATIS soit une méthode axée sur les ressemblances entre marqueurs, et ne permette donc pas d'apprécier les différences entre ceux-ci. Cette critique peut être tempérée. Tout d'abord, nous avons vu que STATIS pouvait mettre en lumière des marqueurs déviant nettement du compromis général (FIG. 2.4). Mais surtout, s'il est vrai que STATIS n'est pas optimale du point de vue de la mise en valeur des différences entre les marqueurs, ceci est aussi vrai pour l'AFM et la MCOA. Il s'agit donc d'un point sensible, qui requiert une attention toute particulière : les méthodes K-tableaux sont généralement orientées vers la recherche de ressemblances entre les tableaux ; ici, nous sommes dans un cas où les dissemblances sont également des sources potentielles d'information biologique. L'étude de la diversité typologique des marqueurs génétiques est donc une question particulièrement intéressante, dans la mesure où nous n'en possédons pas, pour le moment, la réponse.

Chapitre 3

A la recherche de structures génétiques spatialisées

Sommaire

3.1 Introduction

L'identification de structures génétiques spatialisées est une question relativement ancienne dont les réponses méthodologiques, ou du moins leur application, sont finalement très récentes. L'*analyse en composantes principales spatiales* (ou *spatial principal component analysis*, sPCA, Jombart *et al.*, 2008) est une de ces réponses formulée dans le cadre de l'analyse multivariée. Le but de cette partie est de situer la méthode dans son contexte, d'abord biologique, puis méthodologique.

3.1.1 De l'intérêt des structures spatiales en génétique

L'identification de structures spatiales dans la variabilité génétique est une problématique ancienne, aussi ancienne sans doute que l'étude de la structuration d'une population d'individus en sous-populations. En réalité, l'information moléculaire est souvent utilisée pour inférer des processus biologiques directement en relation avec la distribution spatiale des individus, qu'il s'agisse d'isolement par la distance (Wright, 1943), de métapopulations (Hanski & Simberloff, 1997) ou de barrières aux flux géniques (Slatkin, 1985). Il n'est donc pas surprenant qu'un grand nombre, si ce n'est une majorité d'études portant sur la structuration génétique d'individus comme de populations soit intrinsèquement spatialisée (Jombart *et al.*, in revision). Il est par exemple frappant que la popularisation de l'ACP en génétique soit due à son utilisation pour révéler des structures génétiques spatiales permettant d'inférer les grandes migrations humaines à l'échelle du globe (Menozzi *et al.*, 1978; Bertranpetit & Cavalli-Sforza, 1991; Cavalli-Sforza *et al.*, 1993, 1994). L'ACP n'est pourtant pas une méthode spatialisée : le critère maximisé par les composantes principales ne prend pas en compte l'information spatiale et n'a aucune propriété d'optimalité vis-à-vis de la structure spatiale. Cette remarque peut être généralisée à toutes les méthodes multivariées courantes. Dans le cas des travaux initiés par Menozzi *et al.* (1978), l'essentiel de la variabilité est spatialisée, et la PCA suffit à détecter les structures spatiales.

Dans des cas moins tranchés, les limites des méthodes non spatialisées apparaissent (Wartenberg, 1985; Jombart *et al.*, 2008). Dans certains cas, il est évident qu'on tente à tout prix de détourner un outil statistique de son usage premier pour étudier des structures spatiales. Par exemple, on peut lire dans Preziosi & Fairbairn (1992) :

> *Multidimensional scaling (MDS) (Wilkinson, 1990) was used as an alternative method of examining the association between genetic and geographic distance. If an association exists between genetic and geographic distance, then the map produced by the multidimensional scaling procedure resembles the geographic map.*

La même idée est poussée à l'extrême dans Pariset *et al.* (2003), qui superposent la carte factorielle d'une analyse en coordonnées principales avec une carte géographique pour inférer des structures spatiales.

Il est donc manifeste que des méthodes dédiées à l'identification de structures génétiques spatiales sont requises. Pour ce faire, il est essentiel qu'une méthode soit *spatialement explicite*, c'est-à-dire qu'elle incorpore l'information spatiale comme une partie intégrante du modèle des données, ou comme une composante du critère optimisé dans le cas de l'analyse multivariée.

3.1.2 Contexte méthodologique

Des méthodes spatialement explicites ont récemment vu le jour dans différents contextes méthodologiques de l'analyse des marqueurs génétiques. Une version spatialisée de l'analyse de variance moléculaire (AMOVA, Excoffier *et al.*, 1992) a été proposée par Dupanloup *et al.* (2002) et s'est montrée utile en phylogéographie (Pramual *et al.*, 2005; Tolley *et al.*, 2006). L'approche de groupement bayésien implémentée dans le logiciel STRUCTURE (Pritchard *et al.*, 2000; Falush *et al.*, 2003) a également été élargie à l'étude des structures spatiales par Guillot *et al.* (2005, 2006) puis améliorée par François *et al.* (2006). Notons cependant que toutes ces méthodes supposent que la structuration génétique revêt la forme de groupes d'individus, proches sur le plan génétique et spatial. Ces approches sont donc discutables dans le cas d'individus ou de populations structurés en cline. Elles constituent néanmoins des étapes importantes pour la prise en compte de l'information spatiale dans l'analyse de données génétiques. La sPCA s'incrit dans ce courant d'adaptation des méthodologies existantes à l'analyse des structures spatiales. Elle émerge néanmoins d'un contexte méthodologique particulier, dont nous identifions à présent les principaux éléments.

a. L'origine de la méthode

La sPCA prend ses origines dans les travaux de Moran (1948, 1950) sur la mesure de l'autocorrélation spatiale d'une variable, popularisés par Cliff & Ord (1973, 1981) et étendus au cas multivarié par Wartenberg (1985). La thèse de Sébastien Ollier (2004) fournit une présentation remarquable de ces approches, ainsi qu'une comparaison intéressante avec une démarche analogue consistant à étendre la mesure de variance locale de Geary (1954) au cas multivarié (Lebart, 1969; Banet & Lebart, 1984). Les versions multivariées basées sur l'**indice de Moran** (I) et sur celui de Geary (c) ont par ailleurs été réunies par Thioulouse *et al.* (1995). Cependant, pour comprendre l'origine de la sPCA, seule la connaissance des démarches basées sur l'indice de Moran est nécessaire.

b. Le I de Moran

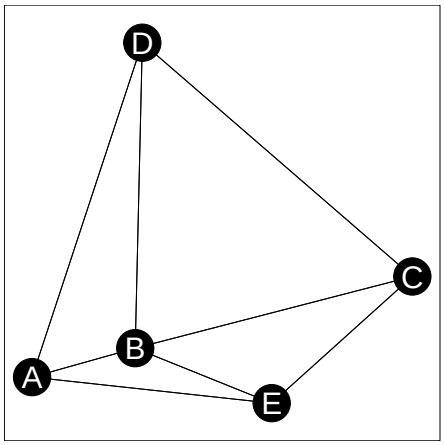
L'indice de Moran est sans doute la mesure la plus courante de l'autocorrélation spatiale (Cliff & Ord, 1973, 1981). La présentation qui suit est volontairement succincte, et ne vise qu'à rappeler des éléments qui seront utiles par la suite. La formulation matricielle la plus simple de l'indice de Moran est définie pour un vecteur $\mathbf{z}$ de $\mathbb{R}^n$ centré et réduit ($\langle \mathbf{z}, \mathbf{1}_n \rangle_{\frac{1}{n}\mathbf{I}_n} = 0$ et $\|\mathbf{z}\|^2_{\frac{1}{n}\mathbf{I}_n} = 1$) par :

$$I(\mathbf{z}) = \mathbf{z}^T \mathbf{M} \mathbf{z} \tag{3.1}$$

où $\mathbf{M}$ est une matrice symétrique dérivée de la matrice d'adjacence $\mathbf{M}_0$ d'un graphe de voisinage entre observations par :

$$\mathbf{M} = \frac{\mathbf{M}_0}{\mathbf{1}_n^T \mathbf{M}_0 \mathbf{1}_n} \tag{3.2}$$

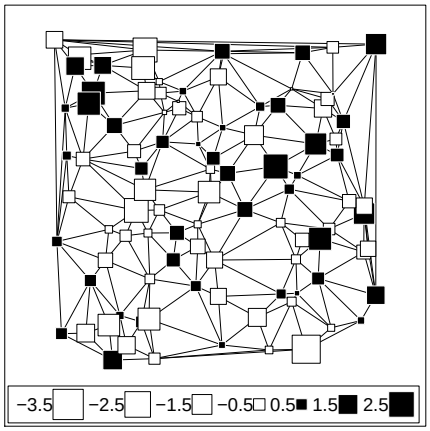
c'est-à-dire après une normalisation globale. Un exemple de matrice d'adjacence d'un graphe est fourni à la figure 3.1.



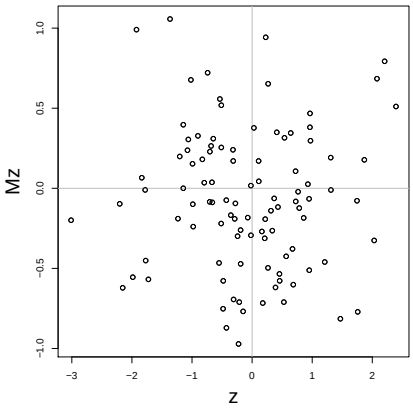
$$\begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \end{pmatrix} \end{array}$$

FIG. 3.1: Exemple de graphe de voisinage (triangulation de Delaunay) et sa matrice d'adjacence. La matrice d'adjacence correspond à la matrice $\mathbf{M}_0$ (EQN. 3.2) ; la matrice $\mathbf{M}$ (EQN. 3.1) est obtenue en divisant $\mathbf{M}_0$ par la somme de tous ses termes, qui vaut ici 16. On note $\mathbf{z} = [z_A \dots z_E]^T$ le vecteur des valeurs associées aux points $\{A \dots E\}$.

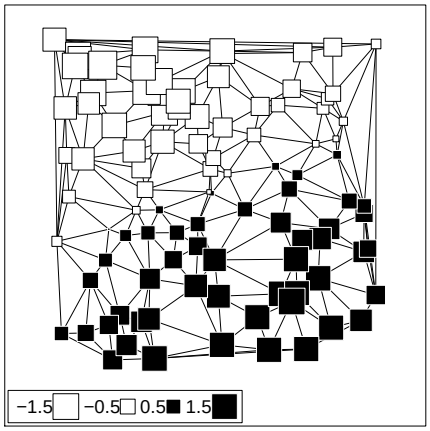
La forme (EQN. 3.1) associe à un vecteur un scalaire, mais n'est pas un produit scalaire puisque la matrice $\mathbf{M}$ n'est pas définie positive. Le vecteur $\mathbf{Mz}$, nommé **vecteur lissé**, a pour $i^{\text{ème}}$ terme une combinaison linéaire des valeurs voisines (c'est-à-dire reliées sur le graphe) du site i (FIG. 3.1). Par exemple, la valeur de $\mathbf{Mz}$ pour le point A de la figure 3.1 vaut $\frac{1}{16}(z_B + z_D + z_E)$. Les termes de $\mathbf{Mz}$ ne sont pas donc pas des moyennes (les sommes marginales de $\mathbf{M}$ ne sont pas unitaires), mais ce sont des mesures globales des valeurs de $\mathbf{z}$ au voisinage de chaque point. On comprend alors que la valeur de $I(\mathbf{z})$ est élevée lorsque des valeurs extrêmes et de même signe sont associées sur le graphe (FIG. 3.2c-d), et inversement fortement négative lorsque des sites voisins portent des valeurs extrêmes et de signe opposé (FIG. 3.2e-f). Lorsque les valeurs sont distribuées aléatoirement sur le graphe (FIG. 3.2a-b), $I(\mathbf{z})$ prend une valeur voisine de zéro, dont l'espérance est $-1/(n-1)$ (Cliff & Ord, 1981). Enfin, il est à noter que $I(\mathbf{z})$ varie sur un intervalle qui est défini par la matrice de pondérations de voisinage $\mathbf{M}$ (de Jong *et al.*, 1984). On reviendra sur cette observation et sur les implications qui en découlent dans la discussion.



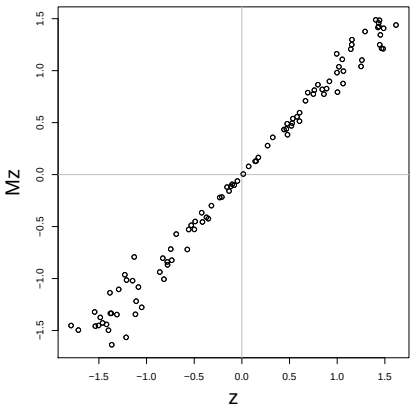
(a) Autocorrélation spatiale nulle (cartographie de **z**).



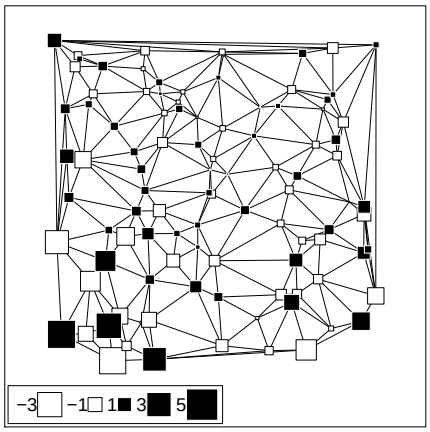
(b) Autocorrélation spatiale nulle (relation entre **z** et **Mz**).



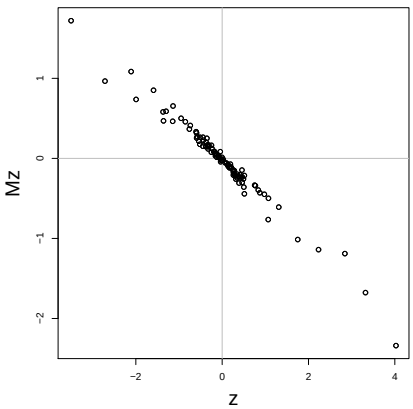
(c) Autocorrélation spatiale positive (cartographie de **z**).



(d) Autocorrélation spatiale positive (relation entre **z** et **Mz**).



(e) Autocorrélation spatiale négative (cartographie de **z**).



(f) Autocorrélation spatiale négative (relation entre **z** et **Mz**).

FIG. 3.2: Différents cas d'autocorrélation spatiale d'une variable **z** centrée et réduite (gauche). Les graphiques de droite montrent la relation entre **z** et son vecteur lissé **Mz**.

c. L'approche de Wartenberg

Wartenberg (1985) fut sans doute le premier à formuler une analyse multivariée basée sur l'indice de Moran, et qui est en réalité très proche de la sPCA. Cette méthode, la *multivariate spatial correlation* (MSC), repose sur la formulation matricielle de l'indice de Moran pour une variable centrée et réduite (Eqn. 3.1). Wartenberg remarque que la matrice $\mathbf{M}$ peut être utilisée pour mesurer la co-structure spatiale de deux variables $\mathbf{z}_1$ et $\mathbf{z}_2$, en utilisant la forme bilinéaire symétrique $\mathbf{z}_1^T\mathbf{M}\mathbf{z}_2 = \mathbf{z}_2^T\mathbf{M}\mathbf{z}_1$. La généralisation au cas multivarié est immédiatement obtenue par le produit $\mathbf{Z}^T\mathbf{M}\mathbf{Z}$.

La matrice $\mathbf{Z}^T\mathbf{M}\mathbf{Z}$ n'est pas définie positive, mais elle est symétrique et admet donc une base de vecteurs propres orthonormés. Les composantes principales $\mathbf{a}_i$ ($i = 1, \ldots, r$ où r est le rang de $\mathbf{Z}^T\mathbf{M}\mathbf{Z}$) associées aux valeurs propres positives de $\mathbf{Z}^T\mathbf{M}\mathbf{Z}$ sont des vecteurs de $\mathbb{R}^n$ maximisant successivement $\mathbf{a}_i^T\mathbf{M}\mathbf{a}_i$ sous la contrainte $\mathbf{a}_i^T\mathbf{M}\mathbf{a}_j = 0$ pour $i \neq j$. Ce sont donc des combinaisons linéaires de variables ayant une autocorrélation spatiale fortement positive, donc des structures spatiales. Les composantes principales liées à des valeurs propres négatives ne sont pas interprétées. Le commentaire de l'auteur à ce sujet laisse entrevoir un malaise certain (Wartenberg, 1985) :

> *These negative eigenvalues are as important as positive eigenvalues but are of qualitatively different type. They represent spatial interaction (covariance) that is more important than spatial pattern (variance). A thorough discussion of this topic is beyond the scope of this paper and will be presented elsewhere. To avoid this situation, data yielding negative eigenvalues are not used in this paper.*

On peut penser que Wartenberg est alors gêné par l'interprétation de ces composantes principales, mais il peut aussi s'agir de réticences culturelles (dogmatiques ?) vis-à-vis des valeurs propres négatives, associée en analyse multivariée « classique » à des distances non-euclidiennes, donc non représentables dans un plan, donc inexploitables. Ces composantes sont pourtant interprétables : il s'agit de combinaisons linéaires des variables ayant une autocorrélation spatiale fortement négative ; dans le contexte génétique, nous verrons qu'il s'agit des *structures locales* définies dans l'article présenté plus bas.

Bien que la MSC s'adresse aux géographes et non aux généticiens, Wartenberg, qui vient à l'époque de travailler avec Sokal sur l'isolement par la distance (Sokal & Wartenberg, 1983), illustre sa méthode en réanalysant les marqueurs génétiques de Menozzi *et al.* (1978). Cependant, la méthode n'a eu qu'un impact relatif en géographie jusqu'à ce qu'elle soit redécouverte par Anselin *et al.* (2002), et aucun en génétique.

d. L'approche MULTISPATI

L'approche MULTISPATI (pour *multivariate spatial analysis based on Moran's I*) apparaît d'abord dans la thèse de Sébastien Ollier (2004) et est présentée dans Dray *et al.* (2008). Il s'agit d'une approche générale et applicable à n'importe quel triplet statistique $\mathfrak{T} = (\mathbf{X}, \mathbf{Q}, \mathbf{D})$. MULTISPATI généralise la MSC de Wartenberg (1985) en trois points : i) $\mathbf{X}$ n'est pas nécessairement centrée et réduite par colonne ii) la matrice de pondérations de voisinage (notée

$\mathbf{W}$) peut être non-symétrique iii) des pondérations de lignes ($\mathbf{D}$) et de colonnes ($\mathbf{Q}$) de $\mathbf{X}$ peuvent être prises en compte.

Au lieu de $\mathbf{Z}^T\mathbf{MZ}$, MULTISPATI diagonalise la matrice $\mathbf{Q}$-symétrique :

$$\mathbf{H} = \frac{1}{2}\mathbf{X}^T(\mathbf{W}^T\mathbf{D} + \mathbf{DW})\mathbf{XQ} \tag{3.3}$$

et fournit des composantes principales décomposant le produit de l'inertie et de l'indice de Moran (Dray *et al.*, 2008). Dans le cas du triplet $\mathfrak{T}' = (\mathbf{X}_0, \mathbf{I}_p, \frac{1}{n}\mathbf{I}_n)$ où $\mathbf{X}_0$ est une matrice centrée et normée par colonne et d'une matrice de pondération de voisinage globalement normalisée ($\mathbf{W} = \mathbf{M}$), MULTISPATI est exactement la MSC de Wartenberg.

Simultanément à la proposition de ce cadre général d'ordination sous contrainte spatiale (Dray *et al.*, 2008), nous avons développé la sPCA pour l'analyse de structures génétiques spatiales. Dans le contexte génétique, il était en effet hasardeux de vouloir présenter une méthode basée sur le schéma de dualité, celui-ci étant inconnu des généticiens. Nous avons donc fait le choix de restreindre notre approche à l'analyse en composantes principales, qui est elle bien mieux connue. Fondamentalement, la sPCA est une analyse MULTISPATI basée sur le triplet $\mathfrak{T}'' = (\mathbf{X}, \mathbf{I}_p, \frac{1}{n}\mathbf{I}_n)$ où $\mathbf{X}$ est centrée par colonne, utilisant une matrice de pondération de voisinage normalisée par ligne.

Si la sPCA n'est donc pas nouvelle sur le plan mathématique, son application aux marqueurs génétiques pour rechercher des structures génétiques spatiales est par contre plus novatrice. Par ailleurs, deux tests multivariés contre l'absence de structure spatiale ont été développés conjointement à la sPCA. Ces éléments ont fait l'objet de la publication suivante. Son implémentation, ainsi qu'une application biologique seront ensuite présentées.

3.2 Article 3 : Revealing cryptic spatial patterns in genetic variability by a new multivariate method

Article paru en 2008 dans *Heredity* **101** : 92-103

Heredity (2008) 101, 92–103
© 2008 Nature Publishing Group All rights reserved 0018-067X/08 $30.00
www.nature.com/hdy

ORIGINAL ARTICLE

Revealing cryptic spatial patterns in genetic variability by a new multivariate method

T Jombart, S Devillard, A-B Dufour and D Pontier
Laboratoire de Biométrie et Biologie Evolutive, UMR-CNRS 5558, Université de Lyon, Université Lyon 1, Villeurbanne Cedex, France

Increasing attention is being devoted to taking landscape information into account in genetic studies. Among landscape variables, space is often considered as one of the most important. To reveal spatial patterns, a statistical method should be spatially explicit, that is, it should directly take spatial information into account as a component of the adjusted model or of the optimized criterion. In this paper we propose a new spatially explicit multivariate method, spatial principal component analysis (sPCA), to investigate the spatial pattern of genetic variability using allelic frequency data of individuals or populations. This analysis does not require data to meet Hardy–Weinberg expectations or linkage equilibrium to exist between loci. The sPCA yields scores summarizing both the genetic variability and the spatial structure among individuals (or populations). Global structures (patches, clines and intermediates) are disentangled from local ones (strong genetic differences between neighbors) and from random noise. Two statistical tests are proposed to detect the existence of both types of patterns. As an illustration, the results of principal component analysis (PCA) and sPCA are compared using simulated datasets and real georeferenced microsatellite data of Scandinavian brown bear individuals (*Ursus arctos*). sPCA performed better than PCA to reveal spatial genetic patterns. The proposed methodology is implemented in the adegenet package of the free software R.
Heredity (2008) **101,** 92–103; doi:10.1038/hdy.2008.34; published online 30 April 2008

Keywords: landscape genetics; Moran's *I*; multivariate analysis; principal component analysis; spatial genetics

Introduction

Recently, growing attention is being devoted to taking landscape information into account in genetic studies (Manel *et al*., 2003). Among landscape features, space is most likely to influence the genetic structuring of a set of individuals or populations (Manel *et al*., 2004; Coulon *et al*., 2006). This structuring can exhibit different patterns, such as isolation by distance (Wright, 1943), clines (Haldane, 1948), metapopulations (Hanski and Simberloff, 1997; Kerth and Petit, 2005) and barriers to gene flow (Slatkin, 1985). Moreover, as technological advances have made obtaining spatial information easier, there is strong interest in including this information in the analysis of genetic data.

Exploiting the geographic dimension of genetic data is not new. Spatial information can be used *a posteriori* for graphical display purposes (for example, Bertranpetit and Cavalli-Sforza, 1991; Manel *et al*., 2004) or to measure spatial autocorrelation (for example, Sokal and Wartenberg, 1983; Sokal *et al*., 1986; Bertorelle and Barbujani, 1995; Smouse and Peakall, 1999). Such methods are useful descriptive tools to visualize, quantify and test spatial structure, but are not properly designed to investigate spatial patterns. For instance, ordinary ordination methods (Bertranpetit and Cavalli-Sforza, 1991) may reveal spatial patterns wherever they are obvious, but they are not constrained to do so. To investigate spatial structures other than the most evident, a method should be spatially explicit, that is, it should directly take spatial information into account as a component of the adjusted model or of the optimized criterion, thereby focusing on the part of the variability, which is spatially structured.

Such methods have been developed using different approaches. Within the analysis of molecular variance (AMOVA) framework (Excoffier *et al*., 1992), the spatial analysis of molecular variance (SAMOVA; Dupanloup *et al*., 2002) has proven useful for phylogeographic studies (Pramual *et al*., 2005; Tolley *et al*., 2006) to assess the spatial structure of a known number of populations. Within the Bayesian clustering framework, GENELAND (Guillot *et al*., 2005) and, more recently, a hierarchical Markov random field (HMRF) model (François *et al*., 2006) were proposed as improvements of STRUCTURE (Pritchard *et al*., 2000; Falush *et al*., 2003) by integrating geographic information to infer the number of populations and detect the genetic discontinuities among these populations (Coulon *et al*., 2006). Combining wombling and Bayesian assignment, Manel *et al*. (2007) proposed a method to detect genetic boundaries among multilocus genotypes. However, these approaches rely on a genetic model and require populations to meet Hardy–Weinberg equilibrium expectations (although the HMRF model allows inbreeding) and for linkage equilibrium to exist between loci (see Kaeuffer *et al*., 2007). This might be a problem as such expectations are unrealistic in many

Correspondence: Dr T Jombart, Laboratoire de Biométrie et Biologie Evolutive, UMR-CNRS 5558, Université Lyon 1—CNRS, 43 bd du 11 novembre 1918, Villeurbanne Cedex 69622, France.
E-mail: jombart@biomserv.univ-lyon1.fr
Received 12 February 2008; revised 19 March 2008; accepted 26 March 2008; published online 30 April 2008

cases and robustness of these methods have not been evaluated yet. Another, maybe more concerning, issue with these methods resides in the clustering approach itself: assigning individuals to groups is a likely inappropriate strategy when individuals are genetically structured as a cline. A last approach would be to use a Mantel correlogram (Legendre and Legendre, 1998, pp 736–738) to assess the variation of spatial autocorrelation in allelic frequencies across scales. However, this method is not wholly satisfying as it would only allow to detect spatial structuring, but would not permit to visualize the corresponding spatial patterns.

An appealing alternative for exploring genetic data is offered by reduced space ordination methods because their utilization is not contingent on a particular genetic model. Hardy–Weinberg equilibrium or linkage equilibrium are thus no longer required. Basically, these methods aim at summarizing strongly multivariate data into a few uncorrelated components, forming the so-called 'reduced space'. For this summary to be meaningful, the components are chosen so as to reflect most of the variability in data, as defined by an optimized criterion (for example, variance among observations). Such methods can be applied on allelic frequency data to obtain a summary of the genetic variability among individuals or populations. A great illustration of such practice was offered by Menozzi *et al*. (1978), who used a principal component analysis (PCA; Pearson, 1901) to investigate the spatial patterns of the genetic variability, obtaining the well-known synthetic maps of human gene frequencies. More recently, PCA proved useful to correct for population stratification (Price *et al*., 2006) and to infer and test the number of subpopulations represented in a set of genotypes (Patterson *et al*., 2006). However, PCA can be criticized when applied to reveal spatial patterns. Indeed, this method finds synthetic variables on which the variance among genotypes is maximized, but does not take spatial information into account. PCA seeks genetic variability, not spatial structures; it is not a likely optimal method for revealing *cryptic spatial patterns*, that is, spatial patterns that are not associated to the highest genetic variation.

In this paper, we propose a new method, the spatial principal component analysis (sPCA), as a tool to investigate cryptic spatial patterns of genetic variability using georeferenced multilocus genotypes. Our method relies on a modification of PCA so that not only the variance between the studied entities (individuals or populations), but also their spatial autocorrelation is taken into account. The main results of the analysis are maps of entities scores allowing a visual assessment of the spatial genetic structures. The obtained scores reveal two types of patterns, which we define as *global* and *local* structures (*sensu* Thioulouse *et al*., 1995). Although both types express a fair amount of genetic variability, global structures display positive spatial autocorrelation whereas local ones display negative spatial autocorrelation. In other words, a global pattern would differentiate between two spatial groups or find a cline (or any intermediate state), whereas local scores would retrieve stronger genetic differences among neighbors than among random pairs of entities. As the studied entities can be genotypes or groups of genotypes (later referred to as 'populations', in a broad sense), global and local structures encompass a wide range of biological situations. For instance, global patterns of genotypes could indicate population patches in an island model, as well as cline wherever isolation by distance occurs. Local structures could arise when individuals from the same genetic pool are selected to avoid each other (repulsion) or to be attracted by individuals from other genetic pools. Similarly at the population level, global and local patterns may result from stratification (Price *et al*., 2006) or from adaptations to environmental variables that are inherently spatially structured ('spatial dependence'; *sensu* Wagner and Fortin, 2005).

First, we explain how the spatial information is modeled explicitly through a connection network (Legendre and Legendre, 1998, pp 752–756) to be used in further computations. Then, we detail the meaning of Moran's index of spatial autocorrelation (*I*; Moran, 1948, 1950) which is incorporated into the sPCA criterion, and show how it can identify global and local patterns in allelic frequency data. We then demonstrate how sPCA yields independent components optimizing the product of the variance and Moran's *I*. As an aid to choose the sPCA scores to be interpreted, we developed two multivariate tests against the absence of global and local patterns. Our approach is illustrated and compared to PCA using simulated and real datasets. We conclude by discussing the prospective applicability of this method for the analysis of genetic data. The developed methodology is implemented in the adegenet package (Jombart, 2008) of the free software R (Ihaka and Gentleman, 1996; R Development Core Team, 2008).

Methods

Modeling spatial information

The first step of a spatially explicit method is to define how spatial information is introduced in the method. In sPCA, the detection of spatial structures uses the well-known Moran's *I* (Moran, 1948, 1950), which relies on the comparison of the value of a quantitative variable (for example, allelic frequency) observed at one site (that is, individual or population) to the values observed at neighboring sites. This approach thus requires 'neighboring sites' to be defined. This is usually achieved by building a connection network (also called neighboring graph) which uses an objective criterion to define which entities are neighbors, and which are not. To simplify the definition of spatial structures provided in this paper, the term 'neighbors' is here restrained to immediate neighbors, that is, two vertices of the same edge of the connection network.

Several algorithms, whose review is beyond the scope of the present paper, can be used to build a connection network (Legendre and Legendre, 1998, pp 752–756). Although other spatially explicit methods, such as SAMOVA (Dupanloup *et al*., 2002) or GENELAND (Guillot *et al*., 2005), impose a specific connection network (the Delaunay triangulation; Upton and Fingleton, 1985), sPCA can use any graph. This plasticity makes sPCA adaptable to various spatial distributions. For instance, Delaunay triangulation or Gabriel graph (Gabriel and Sokal, 1969) would be adapted to evenly distributed entities, whereas distance-based neighborhood would be more appropriate to aggregated distributions. Moreover, connection networks can be refined

manually to include empirical knowledge of the spatial connectivity among entities. Once the connection network is defined, the spatial information is stored in a binary connection matrix **M**, which is symmetrical and its lines and columns correspond to the same biological entities (as in a distance matrix). The values of **M** are 1 if the two considered entities are connected, and 0 otherwise. This matrix is used in the computation of Moran's *I* and therefore in sPCA.

Measuring spatial autocorrelation of an allelic frequency

Let us consider one allelic frequency measured on n individuals or populations. Once the binary connection matrix **M** is obtained, the spatial autocorrelation of this frequency can be quantified using Moran's *I*. The general form of this index can be written using matrix notation (Cliff and Ord, 1981, p 119), where **x** is the vector of n centered allelic frequencies and W is the sum of all the terms of **M**:

$$I(\mathbf{x}) = \frac{\mathbf{x}^{\mathrm{T}}\mathbf{M}\mathbf{x}}{W}\,\frac{n}{\mathbf{x}^{\mathrm{T}}\mathbf{x}} \qquad (1)$$

The meaning of this index depends only on its first component; the effect of the second component $n/\mathbf{x}^{\mathrm{T}}\mathbf{x}$ (which is the inverse of the variance of **x**) is to scale the variable so that *I* only reflects its spatial structure, not its variability. In this paper we use a version of *I* in which **M** is standardized so that the rows sum to one (Cliff and Ord, 1973, p 13). Denoting by **L** the resulting matrix, (1) becomes:

$$I(\mathbf{x}) = \frac{\mathbf{x}^{\mathrm{T}}\mathbf{L}\mathbf{x}}{n}\,\frac{n}{\mathbf{x}^{\mathrm{T}}\mathbf{x}} = \frac{\mathbf{x}^{\mathrm{T}}\mathbf{L}\mathbf{x}}{\mathbf{x}^{\mathrm{T}}\mathbf{x}} \qquad (2)$$

The expected value when the frequency observed at a site is independent of its neighbors (the null value, denoted I_0) equals $-1/(n-1)$ under a nonparametric model of the $n!$ possible permutations of the data (Cliff and Ord, 1973, pp 29 and 32). Note that if *I* is to be interpreted quantitatively, its range of variation, which depends on the connection network, should be taken into account (De Jong *et al*., 1984).

This index has a straightforward interpretation. Let i and j indicate a row and a column of **L** ($i=1, n$; $j=1, n$). The row i contains positive values if i and j are neighbors, and 0 otherwise. As the terms of row i sum to one, these values are weights. Hence, the lag vector **Lx** computes, for a given entity, the mean frequency of its neighbors (Anselin, 1996). It follows that $\mathbf{x}^{\mathrm{T}}\mathbf{L}\mathbf{x}$ is the scalar product of the allelic frequencies and their lag vector ($\mathbf{x}^{\mathrm{T}}\mathbf{L}\mathbf{x} = \langle \mathbf{x} \,|\, \mathbf{L}\mathbf{x} \rangle$): the frequency observed for any entity is multiplied by the mean frequency of its neighbors, and the obtained values are added over all entities.

Two types of spatial structuring can be observed in individuals or populations, whenever allelic frequency observed among neighbors are more similar or more dissimilar than expected in a random spatial distribution. These cases are illustrated using 20 fictitious populations (Figure 1). Patches of similar allelic frequencies (Figure 1a) lead to a highly positive *I* because the allelic frequency observed in a population is positively correlated to the allelic frequency of its neighbors (Figure 1c). Conversely, different neighbor to neighbor allelic frequencies (Figure 1b) lead to a highly negative *I*, as the value taken by any population is negatively correlated to those taken by its neighbors (Figure 1d). These two patterns are global and local structures as defined by Thioulouse *et al*. (1995). In the sPCA context, we define global and local patterns as entities being more genetically similar (respectively dissimilar) to their immediate neighbors than expected in a random spatial distribution.

As we have shown, Moran's *I* can be used to numerically detect such patterns using the frequencies of a single allele. Now we tackle the following question: how to reveal these patterns using a complete set of alleles?

Spatial principal component analysis

Two different objectives arise when analyzing a set of georeferenced allelic frequencies. On the one hand, we would like to summarize the genetic variability among the biological entities (individuals or populations) into a few informative components. On the other hand, we would also like to reveal existing spatial patterns.

A convenient solution to the first problem is to use a centered PCA (Pearson, 1901; Menozzi *et al*., 1978). This method analyzes a table **X** of p centered allelic frequencies (displayed in columns) measured on n biological entities (rows). The allelic frequencies define Euclidian distances between the n entities in $\mathbf{R}^p$, the p-dimensional space of real numbers. Finding the line of closest fit through the n points (Pearson, 1901) is the same as finding an axis in $\mathbf{R}^p$ on which the projections of the n entities are as widely scattered as possible, that is, where the Euclidian distances between the entities are best preserved. To fulfill this property, PCA seeks a scaled vector **u** ($\|\mathbf{u}\|^2 = 1$) containing p loadings (one per allele) so that the entities scores onto this axis ($\boldsymbol{\phi} = \mathbf{X}\mathbf{u}$) have a maximum variance. This can be reformulated as the maximization of:

$$\| \mathbf{X}\mathbf{u} \|^2_{1/n} = \frac{1}{n}(\mathbf{X}\mathbf{u})^{\mathrm{T}}\mathbf{X}\mathbf{u} = \frac{1}{n}\mathbf{u}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{u} \qquad (3)$$

where $\|\mathbf{X}\mathbf{u}\|^2_{1/n} = \mathrm{var}(\boldsymbol{\phi})$. The solution is given by the first eigenvector of $(\frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{X})$, which yields scores whose variance is maximized and equates to the highest eigenvalue. The second objective can be tackled by testing the spatial autocorrelation of the PCA scores, to assess whether they display significant spatial structure (Wartenberg, 1985b). However, these scores are appropriate only to summarize genetic variability and are in no way designed to reveal spatial patterns. Thus, there is a need for a methodology summarizing the genetic diversity and revealing spatial structures at the same time.

sPCA encompasses these two objectives. This new method finds a few independent synthetic variables that no longer optimize the variance of the entities' scores (as in PCA), but the product of their variance and of Moran's *I*. sPCA is closely related to Wartenberg's multivariate spatial correlation (MSC; Wartenberg, 1985a), but MSC constrains all alleles to have the same variance. This has the undesirable effect of masking the variability of the most informative alleles. Our method is also linked to that of Thioulouse *et al*. (1995), which also focuses on global and local structures. However, their method differs from sPCA in two ways. Firstly, it introduces nonuniform row weights giving more importance to the entities with many neighbors, whereas sPCA gives equal weights to all entities. Secondly, Thioulouse *et al*. (1995)

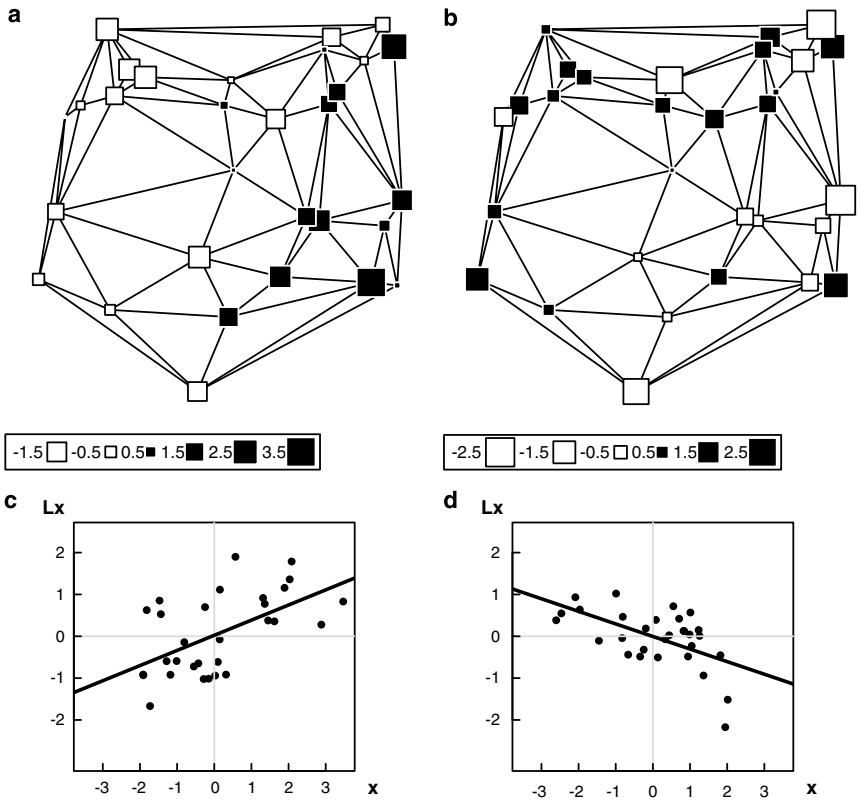


Figure 1 Illustration of global and local patterns of an allelic frequency for 20 fictitious populations overlying their sampling area. Each square represents the frequency of a population. Edges correspond to the connection network (Gabriel's graph). (**a**) Example of global structure, corresponding to $I(\mathbf{x}) > I_0$. (**b**) Example of local structure, corresponding to $I(\mathbf{x}) < I_0$. (**c**) Moran's scatterplot showing that in the global structure (a), the allelic frequency **x** of a population is positively correlated with the mean frequency of its neighbors, **Lx**. The line corresponds to the linear regression of **Lx** on **x**. (**d**) Conversely, the Moran's scatterplot associated with the local structure (**b**) shows that frequency **x** of a population is negatively correlated with the mean value of its neighbors, **Lx**.

used a globally standardized connection matrix instead of the row-standardized matrix **L**, and thus lost the meaning of the lag vector **Lx**.

sPCA seeks scaled axes $\mathbf{v}$ ($\|\mathbf{v}\|^2 = 1$) in $\mathbf{R}^p$ so that entity scores $\boldsymbol{\psi} = \mathbf{Xv}$ are both scattered and spatially autocorrelated. Similarly to the centered PCA (3), this relies on identifying the extreme values of a function (denoted C for 'criterion'):

$$C(\mathbf{v}) = \mathrm{var}(\mathbf{Xv}) I(\mathbf{Xv}) = \frac{1}{n}(\mathbf{Xv})^{\mathrm{T}}\mathbf{LXv} = \frac{1}{n}\mathbf{v}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{LXv} \quad (4)$$

We show that the solution is given by the eigenvectors of the symmetric matrix $\frac{1}{2n}\mathbf{X}^T(\mathbf{L} + \mathbf{L}^{\mathrm{T}})\mathbf{X}$ associated with the highest and lowest eigenvalues (Supplementary Appendix A). As with PCA, other eigenvectors associated with less extreme eigenvalues display weaker structuring under the orthogonality constraint.

Although PCA and sPCA rely on a common approach, two major differences between these analyses must be underlined. Firstly, sPCA does not decompose the total variance into decreasing additive components. Instead, the product of the variance $\mathrm{var}(\boldsymbol{\psi})$ and the spatial autocorrelation $I(\boldsymbol{\psi})$ is separated into positive, null and negative components. Indeed, if the variance is always positive, the spatial autocorrelation can be positive as well as negative. Hence, while PCA focuses on the scores associated to the highest eigenvalues, sPCA encompasses two types of informative scores, both reflecting an aspect of the spatial patterning of the genetic variability. On the one hand, scores with a strong variance and a highly positive spatial autocorrelation (that is, global structures) correspond to highly positive eigenvalues. On the other hand, scores with a strong variance and a highly negative spatial autocorrelation (that is, local structures) correspond to highly negative eigenvalues. Note that these negative eigenvalues are thus useful tools to detect local patterns, and should not be ignored, as it was done in MSC (Wartenberg, 1985a).

Secondly, it makes no sense to compare a sPCA eigenvalue to the sum of all eigenvalues (as done in PCA) because this sum itself has no meaning: it can be low if there is no structure at all, as well as when there are strong global and local structures. Therefore, the percentage of total criterion associated to a given eigenvalue cannot be used as a rule to choose the structures to retain. However, as in other multidimensional methods, an abrupt decrease of the eigenvalues is likely to indicate the boundary between strong and weak structures (Legendre and Legendre, 1998). The interesting patterns are displayed graphically, and their spatial

autocorrelation is measured using Moran's *I*. Note that it is meaningless to test the *I* of the sPCA scores, as is done in PCA (Wartenberg, 1985b) because the sPCA scores are already optimized regarding spatial autocorrelation.

Multivariate tests to detect global and local structuring

Sometimes, the sPCA eigenvalues may not clearly indicate if global and/or local structures should be interpreted. A first aid would be to assess if there are significant global and local patterns in the data. We developed two statistical tests (a global and a local test) to answer to these questions.

These tests rely on the spectral decomposition of the row-standardized connection matrix **L** into Moran's eigenvector maps (MEMs; Griffith, 1996; Dray *et al*., 2006). These vectors are uncorrelated variables modeling different spatial structures; they were initially used in geography for spatial filtering purposes, that is, to remove spatial autocorrelation from the residuals of a statistical model (Griffith, 2000). In ecology, MEMs are used as explanatory variables in linear modeling approaches to model complex spatial patterns (Dray *et al*., 2006; Griffith and Peres-Neto, 2006). Each of these spatial predictors is associated to a Moran's *I* and can therefore be characterized either as a global or a local pattern. We denote $\mathbf{E}^+$ the matrix whose columns are the global MEMs of **L**, and $\mathbf{E}^-$ the matrix storing local MEMs. As there are always $(n-1)$ MEMs for n locations, these vectors can fully decompose a centered allelic frequency **x** into global and local spatial structures using simple linear regression. Note that this decomposition is not subject to multicollinearity troubles because MEMs are orthogonal: each MEM explains a different part of the variance of **x**, which is measured by the corresponding coefficient of determination (R^2). This can be applied to the p centered allelic frequencies of matrix **X**, yielding $p\mathbf{x}(n-1)$ coefficients of determination (one for each allele/MEM combination) which are stored separately for $\mathbf{E}^+$ and $\mathbf{E}^-$ (see detailed computations in Supplementary Appendix B). The R^2 of alleles with vectors of $\mathbf{E}^+$ are used in the global test, whereas R^2 computed with MEMs of $\mathbf{E}^-$ are used in the local test.

The basic idea behind our testing procedures is that if a global (respectively local) pattern exists among individuals (or populations), a large number of alleles is expected to be fairly correlated to at least one vector of $\mathbf{E}^+$ (respectively $\mathbf{E}^-$). To detect this, the mean R^2 across alleles is computed for each MEM. Denoting these means by t_j ($j=1, q$), a vector **t** containing all t_j is then obtained ($\mathbf{t}=[t_1 \ldots t_j \ldots t_q]^T$). To detect an eventual MEM with which all alleles would be significantly correlated, the test statistic used in both procedures is the maximum of **t** values, denoted max(**t**). The null hypothesis ($H0$) is that allelic frequencies of the individuals (or populations) are distributed at random on the connection network. Alternative hypotheses are that allelic frequencies of the studied entities display at least one global (respectively local) spatial structure. The distribution of max(**t**) under H_0 is obtained by a Monte Carlo procedure involving a large (say at least 999) number of permutations. For each permutation, the rows of **X** are randomized and max(**t**) is computed. In both tests, the P-value is defined as the relative frequency of permuted statistics equal to or higher than the initial value of max(**t**). We verified that the type I errors of both tests were correct using simulated datasets (see Supplementary Appendix B).

Illustrations

Simulated data: simple structures

This illustration compares the results of PCA and sPCA on three simulated datasets containing simple spatial structures: two global patterns (patches, cline) and one local structure (repulsion). For discontinuous patterns (patches and repulsion, Figures 2a, b, e and f), three populations of 500 diploid individuals were simulated using EASYPOP version 2.0.1 (Balloux, 2001), using a hierarchical island model to have different levels of genetic differentiation. Migration rate between populations 1 and 2 was set to 0.005 and to 0.002 between population 3 and the other two. Genotypes consisting in 20 microsatellite-like loci were obtained after 1000 generations using a KAM mutation model (that is, loci mutate to any new allelic state with the same probability) with a mutation rate of 0.0001 and 50 possible allelic states. For the continuous pattern (cline, Figures 2c and d), 4 populations of 500 diploid genotypes were simulated under an isolation-by-distance process. Spatial coordinates of the four populations were set in a two-dimensional space to (0, 0), (0, 2), (2, 0) and (2, 2). Dispersal distances were drawn from a negative exponential distribution with a mean of 1. EASYPOP computes the migration rates as $\exp(-r^*d_{ij})$, where d_{ij} is the distance between populations i and j, and r is the inverse of the dispersal distance. The other input parameters of this simulation were the same as in the simulation of discontinuous patterns. All analyzed data were obtained by randomly sampling genotypes from the created populations. Spatial coordinates were defined using the R software to create the various spatial structures.

Three datasets of 80 georeferenced genotypes were created: (1) two patches of 35 individuals from populations 1 and 2 with 10 individuals from population 3 randomly distributed; (2) 40 individuals from populations 1 and 2 forming a cline with 40 individuals from population 3 and 4 randomly distributed; (3) 30 individuals from population 3 distributed in repulsion among a total of 50 individuals of populations 1 and 2.

Data were analyzed using the R software, especially the *ade4* package for multivariate analysis (Chessel *et al*., 2004; Dray *et al*., 2007), *spdep* for spatial methods (Bivand, 2007) and *adegenet* for genetic data handling, sPCA and global/local tests (Jombart, 2008). The same procedure was applied to each dataset: first, data were analyzed by PCA, using Moran's *I* test to detect spatial structuring in the PCA scores; second, data were analyzed by sPCA using global and local tests (with 9999 permutations) as an aid to select the structures to be interpreted. All connection networks were defined using the Delaunay triangulation (Upton and Fingleton, 1985), a common graph that underlies several other methods (Dupanloup *et al*., 2002; Guillot *et al*., 2005; François *et al*., 2006). The patches of the first dataset were retrieved by the first PCA scores (Figure 2a), which were significantly autocorrelated ($I=0.228$, $P=0.0005$). However, these patches appeared more clearly on the first global scores of sPCA

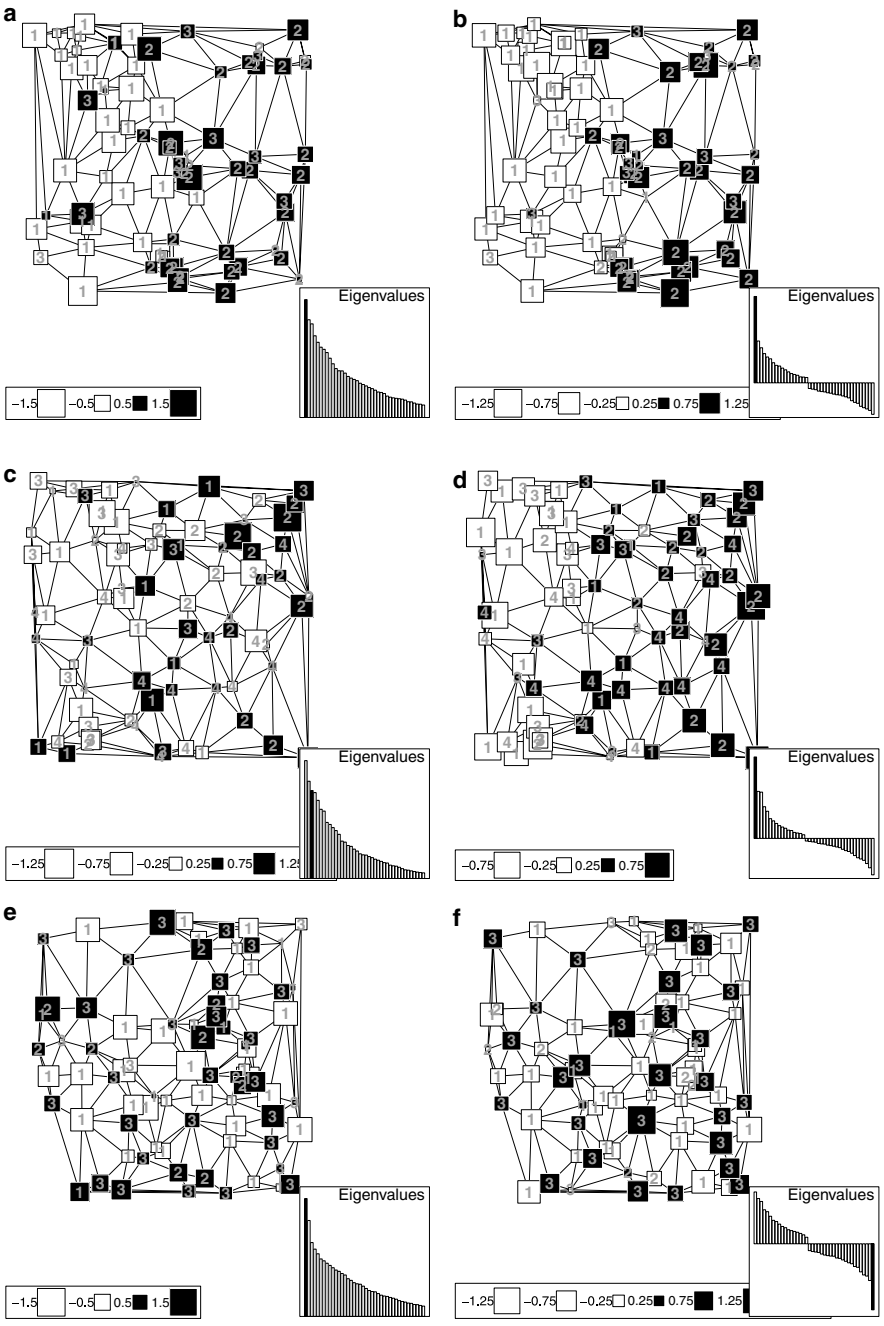


Figure 2 Analyses of simple global and local structures among 80 genotypes from three different populations by principal component analysis (PCA) and spatial PCA (sPCA). Each square represents the score of a genotype and is positioned by its spatial coordinates. The eigenvalues corresponding to the displayed scores are filled in black on the screeplots. Numbers indicate the population to which genotypes belong. (**a**, **b**) Two patches with random noise. (**c**, **d**) A cline with random noise. (**e**, **f**) Repulsion with random noise. (**a**, **c**, **e**) First PCA scores. **(b**, **d**) First global scores of sPCA. (**f**) First local scores of sPCA.

(Figure 2b). The global test confirmed the existence of global pattern ($\max(\mathbf{t}) = 0.0166$, $P = 0.0011$), whereas the local test did not detect any local structure ($\max(\mathbf{t}) = 0.0140$, NS). In the second dataset, PCA overlooked the cline, showing a weak spatial pattern on the third principal component (Figure 2c; $I = 0.128$, $P = 0.022$), whereas sPCA completely retrieved it (Figure 2d). Note that the first global structure is indeed a cline—and not patches—because genotypes situated in the middle of the distribution have less extreme scores

(smaller squares). The global test confirmed the presence of global structure ($\max(\mathbf{t}) = 0.0161$, $P = 0.0038$), whereas the local test detected no local pattern ($\max(\mathbf{t}) = 0.0117$, NS). In the third dataset, the local pattern (repulsion among genotypes from population 3) was not identified by PCA (Figure 2e; $I = -0.054$, NS). On the contrary, the first local score of sPCA revealed this pattern: large black squares (population 3) are rarely found as neighbors and tend to be surrounded by white ones (genotypes from other populations) more often than at random. The global test did not detect any global pattern ($\max(\mathbf{t}) = 0.0132$, NS), whereas the local test was significant ($\max(\mathbf{t}) = 0.0174$, $P = 0.0008$).

Simulated data: complex structures in individuals

This illustration compares the results of PCA and sPCA using a simulated dataset in which different structures are mixed. Four populations of 500 diploid individuals were simulated using EASYPOP, following a hierarchical island model. Migration rate between populations 1 and 2 was set to 0.005, and to 0.002 for other populations. Otherwise, all parameters were those used in the previous illustration. A random sample of 80 genotypes was obtained, with unequal sample sizes (from population 1 to 4, sizes were 30, 30, 10, 10). Spatial coordinates were defined so that: (1) the 60 individuals from populations 1 and 2 were structured as a cline; (2) the 10 individuals from population 3 were distributed randomly; (3) the 10 individuals from population 4 were structured in repulsion.

This dataset was analyzed as previously, first by PCA and then by sPCA. The Delaunay triangulation was employed to model the spatial connectivity among genotypes. Two axes were retained for PCA (Figures 3a and c). No clear spatial pattern was revealed by PCA. The cline between populations 1 and 2 seemed split between the first (Figure 3a) and the second scores (Figure 3c), whereas the local structure induced by individuals from population 4 does not appear clearly on either axis. The Moran's I tests did not detect significant autocorrelation in either scores ($I = 0.081$, NS; $I = -0.014$, NS). On the contrary, sPCA revealed both structures. The first global and local scores were retained (Figures 3b and d). The global scores clearly differentiated populations 1 and 2, even if it was not clear whether this global structure consisted in two patches or in a cline (Figure 3b). The global

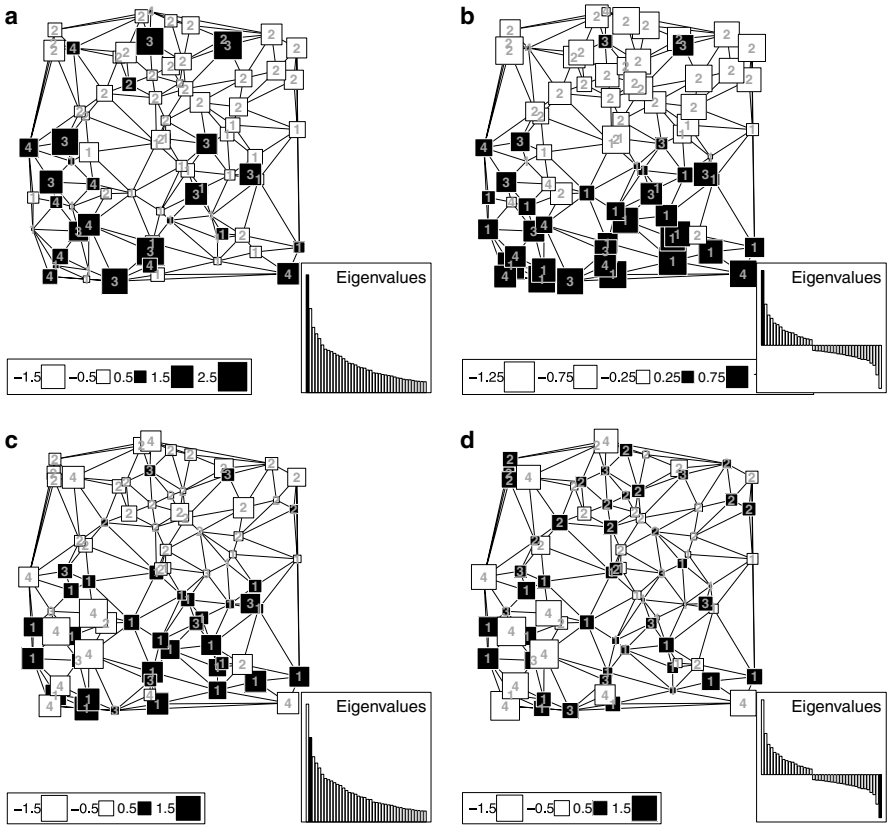


Figure 3 Analyses of complex global and local structures among 80 genotypes from four different populations by principal component analysis (PCA) and spatial PCA (sPCA). Each square represents the score of a genotype and is positioned by its spatial coordinates. The eigenvalues corresponding to the displayed scores are filled in black on the screeplots. Numbers indicate the population to which genotypes belong. (**a**) First PCA scores. (**b**) First global scores of sPCA. (**c**) Second PCA scores. (**d**) First local scores of sPCA.

test confirmed that a global structure existed ($\max(\mathbf{t}) = 0.0200$, $P = 0.0005$). The first local score clearly emphasized to genetic differences of individuals from population 4 (large white squares) from their immediate neighbors (other populations). The local test was consistently significant ($\max(\mathbf{t}) = 0.0270$, $P = 0.0001$).

Simulated data: complex structures in populations

This illustration had the same objective as the previous one, but involved populations rather than individuals. Three populations of 500 diploid individuals were simulated using EASYPOP, following an island model with a migration rate of 0.01. Other parameters of simulations were the same as in previous illustrations. Subpopulations (16) were created by taking random samples of 30 individuals from a given population; 10 subpopulations were thus obtained from population 1 and 2, and 6 were drawn from population 3. Spatial coordinates of subpopulations were defined so that: (1) the 20 subpopulations from populations 1 and 2 were structured in two patches; (2) the 6 subpopulations from population 3 were distributed following a local pattern.

After transforming the data into allelic frequencies for each subpopulation, a PCA and an sPCA were performed. The Delaunay triangulation was used to model the spatial connectivity among subpopulations. The PCA eigenvalues showed that two strongly structured axes were to be retained (Figures 4a and c). This was likely due to the fact that the variability among subpopulations was essentially an interpopulation variability: only two axes are required to differentiate three populations. The first PCA scores displayed a significant spatial structure (Figure 4a; $I = 0.265$, $P = 0.0096$), but it was merely as a by-product: it simply differentiated the population 1 from the two others. Similarly, the second PCA scores differentiated the population 3 from the others (Figure 4c), but these scores were not spatially structured ($I = 0.031$, NS). The sPCA eigenvalues clearly showed that one global and one local axes were to be retained (Figures 4b and d). The first global scores (Figure 4b) found the two patches of subpopulations from populations 1 and 2, giving rather low values to the scores of population 3 (small squares). The global test detected the existence of spatial pattern ($\max(\mathbf{t}) = 0.131$, $P = 0.0065$). The local scores highlighted the differences between subpopulations from population 3 with the neighboring subpopulations (Figure 4d). The local test was also significant ($\max(\mathbf{t}) = 0.133$, $P = 0.0053$).

Scandinavian brown bear data

The Scandinavian brown bear dataset illustrated other methods such as the wombling approach of Manel *et al*.

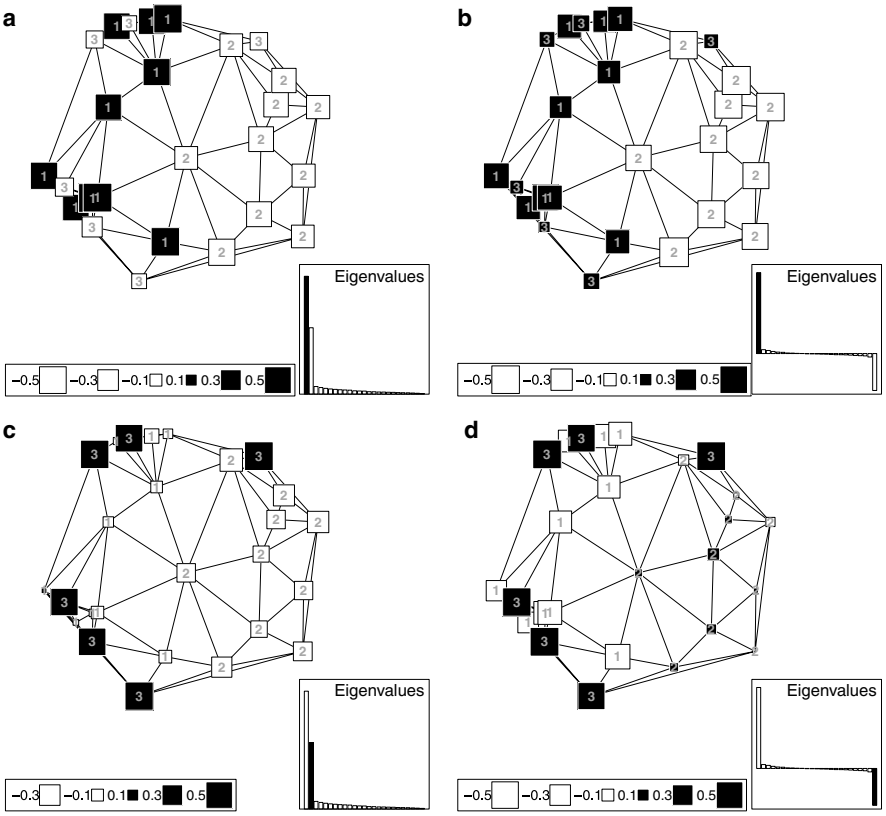


Figure 4 Analyses of complex global and local structures among 16 subpopulations from three populations by principal component analysis (PCA) and spatial PCA (sPCA). Each square represents the score of a subpopulation and is positioned by its spatial coordinates. The eigenvalues corresponding to the displayed scores are filled in black on the screeplots. Numbers indicate the population to which subpopulations belong. (**a**) First PCA scores. (**b**) First global scores of sPCA. (**c**) Second PCA scores. (**d**) First local scores of sPCA.

(2007). These data contain the georeferenced genotypes of 964 brown bears sampled over 200 000 km² and typed for 18 microsatellite markers. A more complete description of the data will be found in Waits *et al*. (2000). Former studies stressed the need for identifying management units (MUs) among Scandinavian brown bear for conservation purposes (Waits *et al*., 2000). Using density indicators, Swenson *et al*. (1998) suggested four different MUs. There seems to be a general agreement that the southernmost group is strongly differentiated from all the others because of different lineages (Manel *et al*., 2004). Nonetheless, the number of MUs to be considered is still discussed: using microsatellites, Waits *et al*. (2000) confirmed the four groups suggested previously (Swenson *et al*., 1998), whereas more recent studies found only three MUs (Manel *et al*., 2007), considering northern individuals as from one MU instead of two. Expressed in terms of sPCA, different MUs would appear as global structures, each global score potentially differentiating between two MUs. Thus, these data seem appropriate to illustrate how sPCA can identify several spatial groups.

First, a centered PCA was performed on the allelic frequencies of the individuals. The first three scores were significantly positively autocorrelated, but only the first

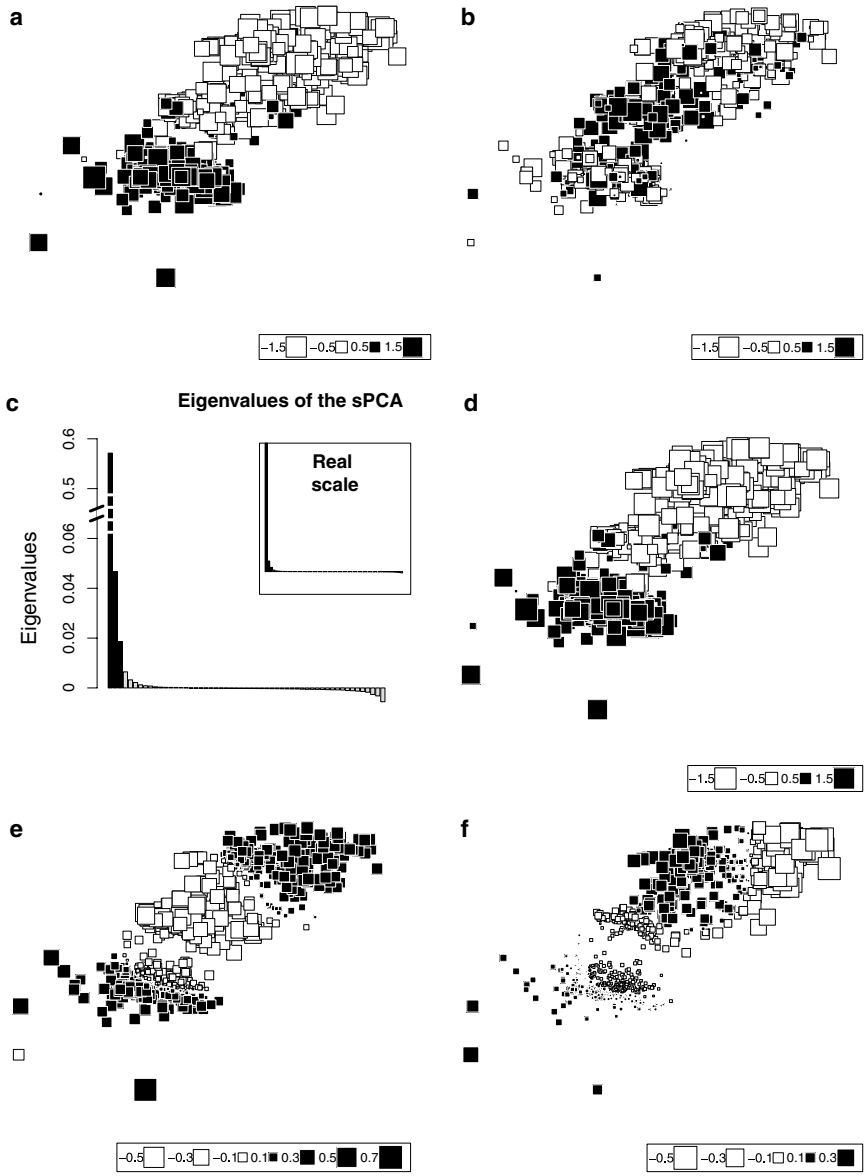


Figure 5 Analyses of Scandinavian brown bears data. (**a**) First and (**b**) second scores of the centered principal component analysis (PCA), displaying significant spatial structure ($I = 0.647$, $P = 0.001$; $I = 0.125$, $P = 0.001$). (**c**) Screeplot of spatial PCA (sPCA). The first eigenvalue (see real scale screeplot, top-right box) was truncated to better appreciate the others. Retained structures are filled in black. (**d**–**f**) First, second and third global scores of sPCA. Autocorrelation statistics were respectively: $I = 0.686$, $I = 0.147$, and $I = 0.122$ ($I_0 \approx 0$).

two had biologically meaningful I values ($I = 0.647$, $P = 0.001$; $I = 0.125$, $P = 0.001$; $I_0 \approx 0$). The first PCA scores differentiated the southern MU from all the others (Figure 5a). The second PCA scores (Figure 5b) were more difficult to interpret, but seemed to correspond in part to the middle MU identified in previous studies (Swenson *et al*., 1998; Manel *et al*., 2004). The third scores displayed a very small spatial autocorrelation, and the associated map was not interpretable (result not shown).

Second, we proceeded to sPCA. We used a distance-based connection network because the spatial distribution was fairly aggregated; such a graph ensures that genotypes inside aggregates have more neighbors than outliers. The threshold distance between any two neighbors was chosen as the minimum distance so that no individual was excluded from the graph. We call the resulting graph a *minimum distance neighboring graph*. The first sPCA eigenvalue was strikingly large compared to the others, but with no doubt the first three eigenvalues and corresponding scores were to be retained (Figure 5c). The first scores revealed the same pattern as in PCA (Figure 5d) and separated individuals from the southern MU from all the others, like in previous studies. This pattern was associated to a strong spatial autocorrelation ($I = 0.686$). The second sPCA scores ($I = 0.147$) clearly differentiated individuals from the 'middle' subpopulation (Waits *et al*., 2000) from the others (Figure 5e). By combining the first two global scores we thus recovered the three subpopulations found in previous studies (for example, Manel *et al*., 2007). But more interestingly, our analysis retrieved an additional weaker structure: undoubtly the third global scores ($I = 0.122$) showed an east–west differentiation among northern individuals (Figure 5f). Contrary to the first pattern (Figure 5d), this structure does not show sharp boundaries between patches, but rather progressive changes from one patch to another, suggesting an isolation-by-distance process or progressive introgression. This may be the reason why a method based on boundary detection (Manel *et al*., 2007) overlooked this structure. The global test confirmed the existence of at least one global pattern ($\max(\mathbf{t}) = 0.0533$, $P = 0.0001$) without detecting local structuring ($\max(\mathbf{t}) = 0.0043$, NS).

Discussion

We propose a spatially explicit multivariate method, sPCA, as a new tool to explore georeferenced multilocus genotypes and, therefore, to try to understand how geographical and environmental features structure genetic information. Although ordinary centered PCA yields scores that summarize the genetic variability among considered entities (individuals or populations), sPCA adds the constraint that the provided scores should be spatially autocorrelated and, thus, focuses on the spatial pattern of genetic variability. Two types of patterns are discriminated: global and local structures, corresponding respectively to large positive and large negative eigenvalues. Maps of sPCA scores are used to visually assess these patterns. As an aid to the interpretation of sPCA results, two Monte Carlo tests are proposed to detect the existence of global and local patterns. Simulated data illustrated that sPCA can retrieve simple structures (patches, clines, repulsion) as well as more complex patterns among genotypes or populations, and performs better in this task than PCA. The global and local tests efficiently detected the existing patterns, with a reliable type I error, and can therefore be used to assess which kind of pattern should be interpreted. sPCA also retrieved already known patterns in Scandinavian brown bear dataset, as well as more cryptic structures, which were overlooked by another method (Manel *et al*., 2007), but were biologically expected (Swenson *et al*., 1998).

Several points relative to the method should be discussed. Firstly, the spatial information is integrated using a connection network. This widely used approach allows taking virtually any type of spatial information into account. Contrary to other spatially explicit methods (Dupanloup *et al*., 2002; Guillot *et al*., 2005), we do not impose a specific connection network; one would have to choose from existing algorithms, and refine it according to what is known about the ecological connectivity in the system. It is important to keep in mind that sPCA is not intended to study the spatial connectivity among the considered entities; it aims at finding spatial structuring given that connectivity.

Secondly, sPCA is proposed mainly as an exploratory tool. For this purpose, our approach seems relevant as it is a reduced space ordination method; no assumptions are made about the data model. It is thus free, for instance, from modeling constraints like Hardy–Weinberg equilibrium assumptions, which are often violated when considering markers involved in selection processes. This is in contrast to, for instance, STRUCTURE (Pritchard *et al*., 2000; Falush *et al*., 2003), which assumes both Hardy–Weinberg equilibrium and linkage equilibrium. Nonetheless, further investigations should be devoted to link sPCA to existing population genetics models. Indeed, the ability of spatial autocorrelation based methods (of which sPCA is one) for inferring genetic processes has been a controversial topic (Sokal and Wartenberg, 1983; Sokal *et al*., 1989; Slatkin and Arter, 1991a, b), but useful studies have shown that Moran's I can be linked to population genetics models (Hardy and Vekemans, 1999). Similarly, a recent study demonstrated that the number of significant eigenvalues of PCA can be directly related to the number of subpopulations in a set of genotypes (Patterson *et al*., 2006). Such development with sPCA would surely enhance the interpretation of the provided results.

Thirdly, the efficiency of sPCA in different population genetics scenario remains to be investigated further, as it was done with spatial autocorrelation. For instance, we did not tackle the relative power of the analysis to reveal patterns due to directional selection (Epperson, 1990) or isolation by distance (Barbujani, 1987; Epperson, 1995). The influence of other parameters, such as the connection network or the level of genetic differentiation, should also be evaluated. These topics as well as comparisons of sPCA to other methods will be investigated using simulations in a next paper.

To conclude, we have shown that sPCA can be used and is useful at the scale of individuals as well as at a population scale. This suggests that our method could be an appropriate tool in different domains. As sPCA can be performed on data from individuals with no *a priori* knowledge of the studied system, our method should become a useful tool in landscape genetics studies (Manel *et al*., 2003), to link the revealed genetic patterns

to landscape features and to explain genetic discontinuities in terms of environmental, behavioral or physiological barriers. Indeed, the sPCA scores can be correlated to other variables or included as dependent or independent variables in models, as long as their spatial autocorrelation is taken into account (Anselin, 2002). Moreover, sPCA can assess the genetic structuring of a set of fragmented populations, which seems especially relevant in conservation biology where this is common. It is particularly important to identify the most isolated populations, when introducing new individuals to maintain genetic diversity or to predict the spatial spread and maintenance of an introduced disease to control pest species. In these cases, sPCA may help to develop appropriate management and surveillance strategies for a disease. Therefore, the proposed method can be seen as a versatile tool for investigating the genetic structuring of set of individuals or populations, within different contexts.

Acknowledgements

This project would not have been possible without the help of Daniel Chessel, who provided insightful discussions about the method and whose comments greatly improved earlier versions of this manuscript. We thank the three anonymous reviewers and the associated editor whose comments also improved a former version of this manuscript. We also thank Jon E Swenson, Stéphanie Manel and Pierre Taberlet for providing the data collected by the Scandinavian Brown Bear Research Project. Finally, we are very grateful to John O'Brien, Nicolas Perrin and Christian Biémont for their helpful comments.

References

Anselin L (1996). Spatial analytical perspectives on GIS. In: Fisher M, Scholten H, Unwin D (eds). *The Moran Scatterplot as an ESDA Tool to Assess Local Instability in Spatial Association*. Taylor and Francis: London. pp 111–125.

Anselin L (2002). Under the hood. Issues in the specification and interpretation of spatial regression models. *Agric Econ* **27**: 247–267.

Balloux F (2001). EASYPOP (version 1.7): a computer program for population genetics simulations. *J Hered* **92**: 301–302.

Barbujani G (1987). Autocorrelation of gene frequencies under isolation by distance. *Genetics* **117**: 777–782.

Bertorelle G, Barbujani G (1995). Analysis of DNA diversity by spatial autocorrelation. *Genetics* **140**: 811–819.

Bertranpetit J, Cavalli-Sforza L (1991). A genetic reconstruction of the history of the population of the Iberian Peninsula. *Ann Hum Genet* **55**: 51–67.

Bivand R (2007). spdep: Spatial dependence: weighting schemes, statistics and models. R package version 0.4-9.

Chessel D, Dufour A-B, Thioulouse J (2004). The ade4 package-I-one-table methods. *R News* **4**: 5–10.

Cliff A, Ord J (1973). *Spatial Autocorrelation*. Pion: London.

Cliff A, Ord J (1981). *Spatial Processes. Model & Applications*. Pion: London.

Coulon A, Guillot G, Cosson J-F, Angibault J, Aulagnier S, Cargnelutti B *et al.* (2006). Genetic structure is influenced by landscape features: empirical evidence from a roe deer population. *Mol Ecol* **15**: 1669–1679.

De Jong P, Sprenger C, van Veen F (1984). On extreme values of Moran's *I* and Geary's *c*. *Geogr Anal* **16**: 17–24.

Dray S, Dufour A-B, Chessel D (2007). The ade4 package—II: Two-table and *K*-table methods. *R News* **7**: 47–54.

Dray S, Legendre P, Peres-Neto P (2006). Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbours matrices (PCNM). *Ecol Modell* **196**: 483–493.

Dupanloup I, Schneider S, Excoffier L (2002). A simulated annealing approach to define the genetic structure of populations. *Mol Ecol* **11**: 2571–2581.

Epperson B (1990). Spatial autocorrelation of genotypes under directional selection. *Genetics* **124**: 757–771.

Epperson B (1995). Spatial distribution of genotypes under isolation by distance. *Genetics* **140**: 1431–1440.

Excoffier L, Smouse P, Quattro J (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: applications to human mitochondrial DNA restriction data. *Genetics* **131**: 479–491.

Falush D, Stephens M, Pritchard J (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**: 1567–1587.

François O, Ancelet S, Guillot G (2006). Bayesian clustering using hidden Markov random fields in spatial population genetics. *Genetics* **174**: 805–816.

Gabriel K, Sokal R (1969). A new statistical approach to geographic variation analysis. *Syst Zool* **18**: 259–278.

Griffith D (1996). Spatial autocorrelation and eigenfunctions of the geographic weights matrix accompanying geo-referenced data. *Can Geogr* **40**: 351–367.

Griffith D (2000). A linear regression solution to the spatial autocorrelation problem. *J Geogr Syst* **2**: 141–156.

Griffith D, Peres-Neto P (2006). Spatial modeling in ecology: the flexibility of eigenfunction spatial analyses. *Ecology* **87**: 2603–2613.

Guillot G, Estoup A, Mortier F, Cosson JF (2005). A spatial statistical model for landscape genetics. *Genetics* **170**: 1261–1280.

Haldane J (1948). The theory of a cline. *J Genet* **48**: 277–284.

Hanski I, Simberloff D (1997). Metapopulation biology: ecology, genetics and evolution. In: Hanski I, Gilpin M (eds). *The Metapopulation Approach, Its History, Conceptual Domain, and Application to Conservation*. Academic Press. pp 5–26.

Hardy O, Vekemans X (1999). Isolation by distance in a continuous population: reconciliation between spatial autocorrelation analysis and population genetics models. *Heredity* **83**: 145–154.

Ihaka R, Gentleman R (1996). R: A language for data analysis and graphics. *J Comput Graph Stat* **5**: 299–314.

Jombart T (2008). *adegenet*: a R package for the multivariate analysis of genetic markers. *Bioinformatics* (doi:10.1093/bioinformatics/btn129; e-pub ahead of print, 8 April 2008).

Kaeuffer R, Réale D, Coltman DW, Pontier D (2007). Detecting population structure using STRUCTURE software: effect of background linkage disequilibrium. *Heredity* **99**: 374–380.

Kerth G, Petit E (2005). Colonization and dispersal in a social species, the Bechstein's bat (*Myotis bechtsteinii*). *Mol Ecol* **14**: 3943–3950.

Legendre P, Legendre L (1998). *Numerical ecology, Developments in Environmental Modelling*. Elsevier Science B.V.: Amsterdam.

Manel S, Bellemain E, Swenson J, François O (2004). Assumed and inferred spatial structure of populations: the Scandinavian brown bears revisited. *Mol Ecol* **13**: 1327–1331.

Manel S, Berthoud F, Bellemain E, Gaudel M, Luikart G, Swenson JE *et al.* (2007). A new individual-based spatial approach for identifying genetic discontinuities in natural populations. *Mol Ecol* **16**: 2031–2043.

Manel S, Schwartz MK, Luikart G, Taberlet P (2003). Landscape genetics: combining landscape ecology and population genetics. *Trends Ecol Evol* **18**: 189–197.

Menozzi P, Piazza A, Cavalli-Sforza L (1978). Synthetic maps of human gene frequencies in Europeans. *Science* **201**: 786–792.

Moran P (1948). The interpretation of statistical maps. *J R Stat Soc Ser B* **10**: 243–251.

Moran P (1950). Notes on continuous stochastic phenomena. *Biometrika* **37**: 17–23.

Patterson N, Price A, Reich D (2006). Population structure and eigenanalysis. *PLoS Genet* **2**: 2074–2093.
Pearson K (1901). On lines and planes of closest fit to systems of points in space. *Philos Mage* **2**: 559–572.
Pramual P, Kuvangkadilok C, Baimai V, Walton C (2005). Phylogeography of the black fly *Simulium tani* (Diptera: Simuliidae) from Thaïland as inferred from mtDNA sequences. *Mol Ecol* **14**: 3989–4001.
Price A, Patterson N, Plenge R, Weinblatt M, Shadick N, Reich D (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**: 904–909.
Pritchard J, Stephens M, Donnelly P (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.
R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. *R Foundation for Statistical Computing*. Vienna: Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.
Slatkin M (1985). Gene flow in natural populations. *Annu Rev Ecol Syst* **16**: 393–430.
Slatkin M, Arter H (1991a). Spatial autocorrelation methods in population genetics. *Am Nat* **138**: 499–517.
Slatkin M, Arter H (1991b). Spatial autocorrelation methods in population genetics. *Am Nat* **138**: 522–523.
Smouse P, Peakall R (1999). Spatial autocorrelation analysis of individual multiallele and multilocus genetic structure. *Heredity* **82**: 561–573.
Sokal R, Jacquez G, Wooten M (1989). Spatial autocorrelation analysis of migration and selection. *Genetics* **121**: 845–855.
Sokal R, Smouse P, Neel J (1986). The genetic structure of a tribal population, the Yanomama Indians. XV. Patterns inferred by autocorrelation analysis. *Genetics* **114**: 259–287.
Sokal R, Wartenberg D (1983). A test of spatial autocorrelation analysis using an isolation-by-distance model. *Genetics* **105**: 219–237.
Swenson J, Sandegren F, Soderberg F (1998). Geographic expansion of an increasing brown bear population: evidence for presaturation dispersal. *J Anim Ecol* **67**: 819–826.
Thioulouse J, Chessel D, Champely S (1995). Multivariate analysis of spatial patterns: a unified approach to local and global structures. *Environ Ecol Stat* **2**: 1–14.
Tolley K, Burger M, Turner A, Matthee C (2006). Biogeographic patterns and phylogeography of dwarf chameleons (*Bradypodion*) in an African biodiversity hotspot. *Mol Ecol* **15**: 781–793.
Upton G, Fingleton B (1985). *Spatial Data Analysis by Sample. Vol. 1: Point Pattern and Quantitative Data*, Volume 1 of Spatial Data Analysis by Example. Wiley: New York.
Wagner H, Fortin M-J (2005). Spatial analysis of landscapes: concepts and statistics. *Ecology* **86**: 1975–1987.
Waits L, Taberlet P, Swenson J, Sandegren F, Franzen R (2000). Nuclear DNA microsatellite analysis of genetic diversity and gene flow in the Scandinavian brown bear (*Ursus arctos*). *Mol Ecol* **9**: 421–431.
Wartenberg D (1985a). Multivariate spatial correlations: a method for exploratory geographical analysis. *Geogr Anal* **17**: 263–283.
Wartenberg D (1985b). Spatial autocorrelation as a criterion for retaining factors in ordinations of geographic data. *Math Geol* **17**: 665–682.
Wright S (1943). Isolation by distance. *Genetics* **28**: 114–138.

Supplementary Information accompanies the paper on Heredity website (http://www.nature.com/hdy)

Appendix A : rational of the spatial Principal Component Analysis

In this appendix, the following notations are used :

$\mathbf{X}$ is the n-by-p table of centred allelic frequencies, where rows are observations (on individuals or populations) and columns are alleles.

$\mathbf{L}$ is a row-standardized connection matrix associated to the connection network among genotypes or populations.

$\mathbf{v}$ refers to any scaled vector of p loadings, so that $\|\mathbf{v}\|^2 = 1$.

$\boldsymbol{\alpha}$ refers to any vector of n row scores obtained by linear combinaison of the columns of $\mathbf{X}$ so that : $\boldsymbol{\alpha} = \mathbf{X}\mathbf{v}$.

$C(\mathbf{v})$ is a quantity serving as a score criterion in the analysis, defined as : $C(\mathbf{v}) = var(\mathbf{X}\mathbf{v})I(\mathbf{X}\mathbf{v}) = \frac{1}{n}(\mathbf{X}\mathbf{v})^T\mathbf{L}\mathbf{X}\mathbf{v} = \frac{1}{n}\mathbf{v}^T\mathbf{X}^T\mathbf{L}\mathbf{X}\mathbf{v}$.

$\mathbf{w}$ refers to any scaled vector of p loadings provided by the sPCA, thus verifying $\|\mathbf{w}\|^2 = 1$.

$\boldsymbol{\psi}$ refers to any vector of n row scores obtained by linear combinaison of the columns of $\mathbf{X}$ so that : $\boldsymbol{\psi} = \mathbf{X}\mathbf{w}$.

The purpose of the spatial Principal Component Analysis is to find the extrema of :

$$C(\mathbf{v}) = \frac{1}{n}\mathbf{v}^T\mathbf{X}^T\mathbf{L}\mathbf{X}\mathbf{v} = \frac{1}{n}\boldsymbol{\alpha}^T\mathbf{L}\boldsymbol{\alpha}$$

which we rewrite, posing $\mathbf{A} = \frac{1}{n}\mathbf{X}^T\mathbf{L}\mathbf{X}$:

$$C(\mathbf{v}) = \mathbf{v}^T\mathbf{A}\mathbf{v}$$

The solution to this problem is well-known when $\mathbf{A}$ is symmetric. This is, however, not the case because $\mathbf{L}$ is not symmetric itself. To solve this problem, we seek a symmetric matrix $\mathbf{B}$ so that :

$$C(\mathbf{v}) = \mathbf{v}^T\mathbf{B}\mathbf{v}$$

The expression $\mathbf{v}^T\mathbf{A}\mathbf{v}$ is a scalar, so $\mathbf{v}^T\mathbf{A}\mathbf{v} = (\mathbf{v}^T\mathbf{A}\mathbf{v})^T = \mathbf{v}^T\mathbf{A}^T\mathbf{v}$. Thus we have :

$$\begin{aligned}
\mathbf{v}^T\mathbf{A}\mathbf{v} &= \frac{1}{2}(\mathbf{v}^T\mathbf{A}\mathbf{v} + \mathbf{v}^T\mathbf{A}^T\mathbf{v}) \\
&= \frac{1}{2}(\mathbf{v}^T(\mathbf{A} + \mathbf{A}^T)\mathbf{v}) \\
&= \mathbf{v}^T(\frac{1}{2}(\mathbf{A} + \mathbf{A}^T))\mathbf{v}
\end{aligned}$$

where $(\mathbf{A} + \mathbf{A}^T)$ is symmetric because $(\mathbf{A} + \mathbf{A}^T)^T = \mathbf{A}^T + \mathbf{A}$. As a consequence, $C(\mathbf{v}) = \mathbf{v}^T\mathbf{B}\mathbf{v}$ with :

$$\mathbf{B} = \frac{1}{2}(\mathbf{A} + \mathbf{A}^T) = \frac{1}{2n}\mathbf{X}^T(\mathbf{L} + \mathbf{L}^T)\mathbf{X}$$

Hence, we can find the extrema of $C(\mathbf{v}) = \mathbf{v}^T\mathbf{B}\mathbf{v}$ using the existing solution (Harville, 1997, p533-534). It is shown that if $\mathbf{w}_1$ and $\mathbf{w}_r$ are the eigenvectors of $\mathbf{B}$ associated to λ_1 and λ_r, respectively the highest and lowest eigenvalues of $\mathbf{B}$, then :

$$\lambda_r = \mathbf{w}_r^T\mathbf{B}\mathbf{w}_r \leq \mathbf{v}^T\mathbf{A}\mathbf{v} \leq \mathbf{w}_1^T\mathbf{B}\mathbf{w}_1 = \lambda_1$$

and so :

$$\lambda_r = var(\boldsymbol{\psi}_r)I(\boldsymbol{\psi}_r) \leq C(\mathbf{v}) \leq var(\boldsymbol{\psi}_1)I(\boldsymbol{\psi}_1) = \lambda_1$$

Appendix B : diagram of the multivariate tests of global and local structuring and associated computations

Computations of the test statistic

The testing procedure (Figure 1) is the same for both multivariate tests (global or local structuring). The matrix of allelic frequencies $\mathbf{X}$ (n individuals or genotypes ; p alleles) is first centred and scaled. The obtained matrix is denoted $\mathbf{Y}$. The matrix $\mathbf{E}$ is obtained like in Griffith (1996) and Dray et al. (2006) by the eigen analysis of the connection matrix associated to the connection network between genotypes (or populations). Its columns $\mathbf{e}_j$ ($j = 1, q$) are the centred and scaled Moran's eigenvector maps (MEMs), which model either global or local structures (Dray et al. 2006). For the purpose of our tests, $\mathbf{E}$ contains 'global MEMs' ($\mathbf{E} = \mathbf{E}+$) for the test of global structuring and 'local MEMs' ($\mathbf{E} = \mathbf{E}-$) for the local test.

The coefficients of determinations (R^2) are computed after linear regression of each allele on each MEM, giving a matrix $\mathbf{S}$ containing p x q values of R^2 computed as :

$$\mathbf{S} = \frac{(\mathbf{Y}^T\mathbf{E}) \bullet (\mathbf{Y}^T\mathbf{E})}{n^2}$$

where $\mathbf{Y}^T$ is the transposed matrix of $\mathbf{Y}$ and where '$\bullet$' denotes the Hadamard product.

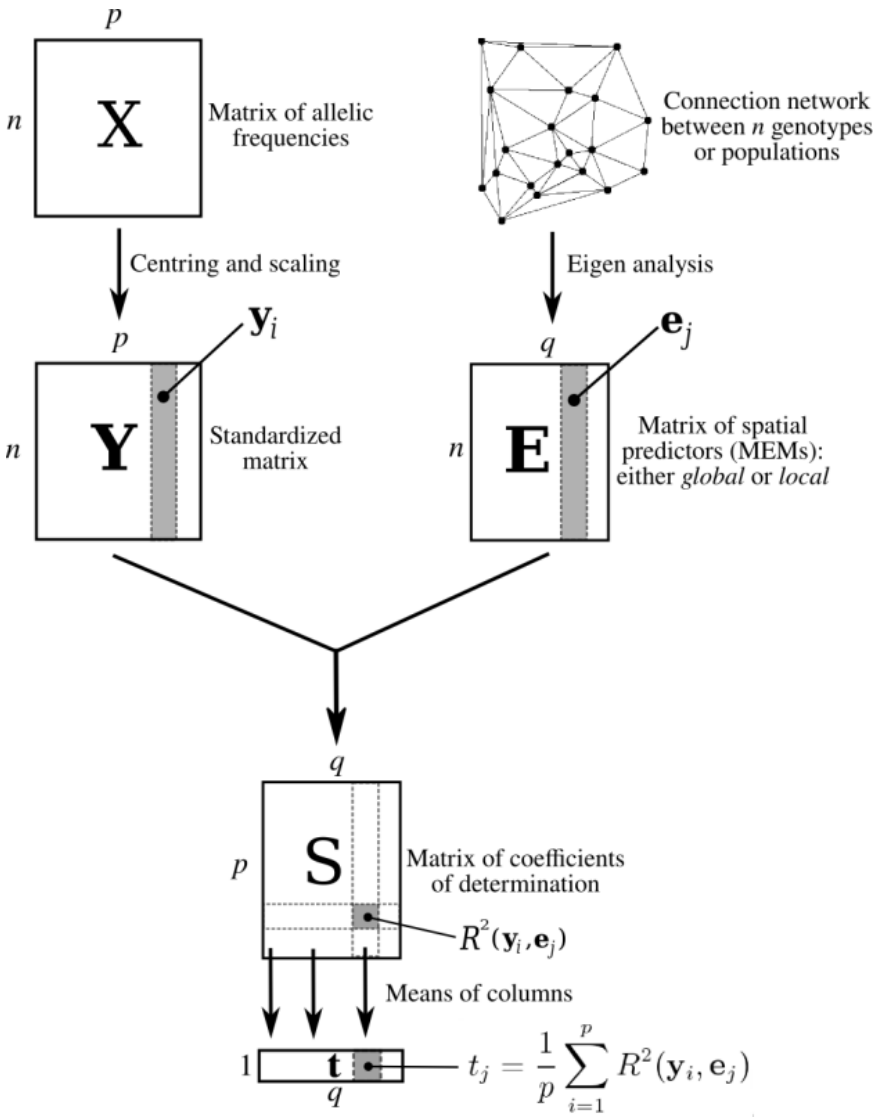


FIG. 1 – Diagram of the testing procedure

As MEMs are orthogonal vectors, the coefficients of determination of alleles obtained from regression onto $\mathbf{E}+$ are independent from of those obtained with $\mathbf{E}-$. Practically, this implies that the test statistics in global and local tests are independent, in the sense that the value of the first is not conditionned by the value of the second (and reciprocally). Note that this is also true for the reference distributions of both statistics.

The squared correlations are then averaged for each MEM, giving a value t_j for the j^{th} allele defined by :

$$t_j = \frac{1}{p} \sum_{i=1}^{p} R^2(\mathbf{y}_i, \mathbf{e}_j)$$

The value of t_j is 'large' (respectively 'small') when alleles are in average 'strongly' (respectively 'weakly') correlated to the j^{th} MEM. The $(n-1)$ values of t_j are stored in the vector $\mathbf{t}$, directly computed as :

$$\mathbf{t} = \frac{1}{p} \mathbf{1}_p^T \mathbf{S}$$

where $\mathbf{1}_p$ is the p-dimensional vector whose components are all 1. The test statistic is defined as the maximum of all components of $\mathbf{t}$, $\max(\mathbf{t})$.

Computation of the reference distribution

The 'reference distribution' is defined as the distribution of the test statistic ($\max(\mathbf{t})$) under the null hypothesis H_0 that allelic frequencies of the genotypes (or populations) are distributed at random on the connection network. The alternative hypothesis H_1 depends on the type of test :

- global test : allelic frequencies of the genotypes (or populations) display at least one global structure
- local test : allelic frequencies of the genotypes (or populations) display at least one local structure

In both tests, the reference distribution is approximated by a Monte Carlo procedure involving a large number of permutations (at least 999) of the rows of $\mathbf{Y}$ (genotypes or populations), the test statistic being computed from each permutation. The p-value is computed as the relative frequency of permuted statistics equal to or higher than the initial value of $\max(\mathbf{t})$.

Assessment of type I error

The actual type I error of both tests was assessed using simulated datasets of allelic frequencies with random spatial coordinates. We used datasets with different number of observations (25, 50, 100, 200) and different number of alleles (50, 100, 150). For each size of dataset, 200 simulations were performed and the results were pooled across simulations for each test, yielding a total of 2400 simulations per test. The actual type I error was measured as the relative frequency of reject of H_0 given different nominal α levels (Table 1). The estimated type I error was always very close to the chosen α level.

Nominal α level	Observed type I error (global test)	Observed type I error (local test)
0.1	0.0925	0.1042
0.05	0.0425	0.0512
0.01	0.0075	0.0062

TAB. 1 – Estimations of actual type I errors of the global and local tests assessed through simulations, for three nominal α levels. 2400 simulations of datasets with different size (see text) were performed for each test.

References

Dray S, Legendre P, Peres-Neto P (2006) Spatial modelling : a comprehensive framework for principal coordinate analysis of neighbours matrices (PCNM). *Ecological Modelling* **196** : 483-493.

Griffith D (1996) Spatial autocorrelation and eigenfunctions of the geographic weights matrix accompanying geo-referenced data. *Canadian Geographer* **40** : 351-367.

Harville D (1997) *Matrix algebra from a statistician's perspective.* Springer, New York.

3.3 Mise en oeuvre de la méthode

3.3.1 L'implémentation

La sPCA est implémentée dans le package *adegenet* (Jombart, 2008) du logiciel R (R Development Core Team, 2008). Au travers de la reproduction de l'analyse d'un jeu de données simulées, on illustre les principaux aspects de cette implémentation. Les données simulées présentées dans Jombart *et al.* (2008) sont disponibles sous le nom de `spcaIllus`. Il s'agit d'une liste d'objets `genind`, dont seul le premier sera utilisé.

```
> data(spcaIllus)
> myObj <- spcaIllus$dat2A
> myObj


   ####################
   ### Genind object ###
   ####################
- genotypes of individuals -

S4 class:  genind
@call: old2new(object = obj)

@tab:  80 x 192 matrix of genotypes

@ind.names: vector of  80 individual names
@loc.names: vector of  20 locus names
@loc.nall: number of alleles per locus
@loc.fac: locus factor for the  192 columns of @tab
@all.names: list of  20 components yielding allele names for each locus
@ploidy:  2

Optionnal contents:
@pop:  factor giving the population of each individual
@pop.names:  factor giving the population of each individual

@other: a list containing: xy
```

L'objet `myObj` contient 80 génotypes provenant de 3 populations simulées sous le logiciel Easypop (Balloux, 2001) en utilisant un modèle en îles. La description précise des données sera trouvée dans Jombart *et al.* (2008). La définition d'une matrice de pondération de voisinage, et donc d'un graphe de voisinage sous-jacent, est un pré-requis à l'analyse. Une première difficulté vient du fait que les différents graphes disponibles sous R sont répartis dans plusieurs packages, sous différents formats (Tab. 3.1).

Afin de simplifier la tâche de l'utilisateur, la fonction `chooseCN` (pour "choose a connection network") fournit une interface, interactive ou non, pour le choix du graphe de voisinage. Outre les graphes mentionnés plus haut (Tab. 3.1), `chooseCN` peut également construire une matrice de pondération de voisinage normalisée par ligne dont les termes sont les inverses de la distance géographique, à un exposant près. On peut illustrer son fonctionnement simplement, en fournissant différents graphes pour la distribution spatiale des génotypes de `myObj`.

```
> par(mfrow = c(3, 2), mar = c(0.5, 0.5, 3, 0.5))
> cn1 <- chooseCN(myObj@other$xy, ask = FALSE, type = 1)
> title("Triangulation de Delaunay", cex.main = 1.5)
> invisible(chooseCN(myObj@other$xy, ask = FALSE, type = 2))
> title("Graphe de Gabriel", cex.main = 1.5)
> invisible(chooseCN(myObj@other$xy, ask = FALSE, type = 3))
> title("Graphe des voisins relatifs", cex.main = 1.5)
> invisible(chooseCN(myObj@other$xy, ask = FALSE, type = 4))
> title("Arbre de longueur minimale", cex.main = 1.5)
> invisible(chooseCN(myObj@other$xy, ask = FALSE, type = 5, d1 = 0,
+     d2 = 3))
> title("Voisinage par la distance (0-3)", cex.main = 1.5)
> invisible(chooseCN(myObj@other$xy, ask = FALSE, type = 6, k = 8))
> title("K plus proches voisins (K=8)", cex.main = 1.5)
```

TAB. 3.1: Principaux graphes de voisinage disponibles dans le logiciel R

Type de graphe	Fonction	Package	Classe	Fonctions de conversion
Delaunay triangulation	`tri2nb`	`tripack`	`nb`	`nb2neig`
Gabriel graph	`gabrielneigh`	`spdep`	`graph`	`graph2nb`, `nb2neig`
Relative neighbours	`relativeneigh`	`spdep`	`graph`	`graph2nb`, `nb2neig`
K nearest neighbours	`knearneigh`	`spdep`	`knn`	`knn2nb`, `nb2neig`
Neighbourhood by distance	`dnearneigh`	`spdep`	`nb`	`nb2neig`
Minimum spanning tree	`mstree`	`ade4`	`neig`	`neig2nb`

Les graphes de voisinage peuvent être édités manuellement pour enlever ou ajouter des connections (argument `edit.nb`).

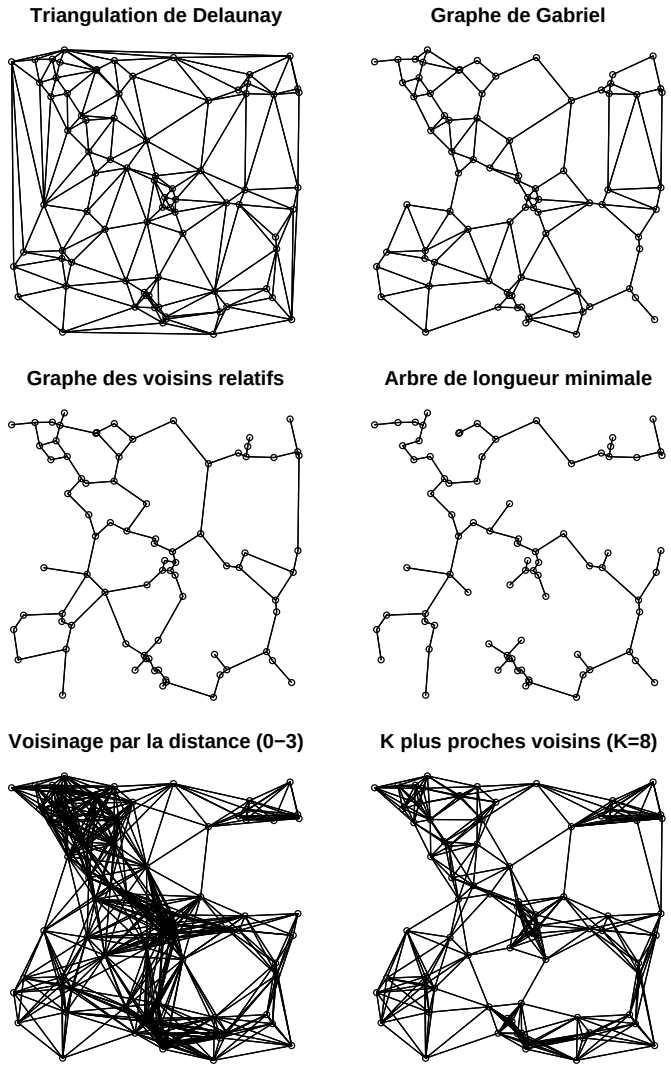


FIG. 3.3: Différents graphes de voisinage disponibles par `chooseCN` (données : `spcaIllus$dat2A`).

La fonction `spca` est la procédure effectuant les calculs. Elle accepte un grand nombre d'arguments, dont peu seront tous utiles à la fois.

```
> args(spca)
```

```
function (obj, xy = NULL, cn = NULL, scale = FALSE, scannf = TRUE,
    nfposi = 1, nfnega = 1, type = NULL, ask = TRUE, plot.nb = TRUE,
    edit.nb = FALSE, truenames = TRUE, d1 = NULL, d2 = NULL,
    k = NULL, a = NULL, dmin = NULL)
NULL
```

Par exemple, les coordonnées spatiales (`xy`) peuvent être omises, auquel cas elles seront recherchées dans `obj$other$xy`. Le graphe de voisinage (`cn`) peut être spécifié, ou construit à partir des données par un appel interne à `chooseCN`. On utilise le graphe de Delaunay précédemment défini pour la sPCA (`cn=cn1`) ; on précise qu'on ne décide pas des structures retenues de manière interactive (`scannf=FALSE`), que l'on conserve trois structures globales (`nfopsi=3`) et aucune structure locale (`nfnega=0`) :

```
> mySpca <- spca(myObj, cn = cn1, scannf = FALSE, nfposi = 3, nfnega = 0)
> class(mySpca)


[1] "spca"


> mySpca


        ########################################
        # Spatial principal component analysis #
        ########################################
class: spca
$call: spca(obj = myObj, cn = cn1, scannf = FALSE, nfposi = 3, nfnega = 0)

$nfposi: 3 axis-components saved
$nfnega: 0 axis-components saved
Positive eigenvalues: 0.2309 0.1118 0.09379 0.07817 0.06911 ...
Negative eigenvalues: -0.08421 -0.07376 -0.06978 -0.06648 -0.06279 ...

  vector length mode    content
1 $eig   79     numeric eigenvalues

  data.frame nrow ncol content
1 $c1        192  3    principal axes: scaled vectors of alleles loadings
2 $li        80   3    principal components: coordinates of entities
3 $ls        80   3    lag vector of principal components
4 $as        2    3    pca axes onto spca axes

$xy: matrix of spatial coordinates
$lw: a list of spatial weights (class 'listw')

other elements: NULL
```

De façon interne, les calculs sont effectués par la fonction `multispati` du package *ade4*, ce qui reflète le fait que la sPCA ne soit qu'un cas particulier de MULTISPATI (Dray *et al.*, 2008).

Un objet de la classe `spca` est à peu de choses près un objet `multispati`. La principale différence tient à l'affichage de l'objet (plus explicite pour la `spca`) et au fait que le graphe de voisinage et les coordonnées spatiales soient intégrés à l'objet `spca`, ce qui n'est pas le cas de `multispati`. Cela permet de ré-effectuer tous les calculs de l'analyse, notamment celui du I de Moran (voir `summary.spca) et screeplot.spca`, et de fournir des représentations graphiques à partir du seul objet `spca` (voir `plot.spca`).

Lorsque l'on effectue une sPCA, une première question consiste à savoir ce que la sPCA apporte par rapport à une ACP classique. Comparer les variances et l'autocorrélation des composantes principales respectives permet de se faire une idée. Cette information est donnée par la fonction `summary.spca` :

```
> summary(mySpca)
```

```
Spatial principal component analysis

Call: spca(obj = myObj, cn = cn1, scannf = FALSE, nfposi = 3, nfnega = 0)

Connection network statistics:
          I0       Imin     Imax
 -0.01265823 -0.5153152 1.015754

Scores from the centred PCA
             var       cum      ratio       moran
Axis 1 0.4710811 0.4710811 0.06566234  0.22779127
Axis 2 0.3906820 0.8617631 0.12011812  0.12057199
Axis 3 0.3743071 1.2360702 0.17229147 -0.02673497

sPCA eigenvalues decomposition:
             eig       var     moran
Axis 1 0.2308786 0.3761419 0.6138072
Axis 2 0.1118472 0.2308986 0.4843998
Axis 3 0.0937875 0.2049956 0.4575099
```

Dans cet exemple, le gain en autocorrélation spatiale (I de Moran) est important (de 0,23 à 0,61) et la perte de variance raisonnable (de 0,47 à 0,38). Les valeurs minimale et maximale (I_{min} et I_{max}) sont déterminées comme dans de Jong *et al.* (1984) ; I_0 est la valeur "nulle" du I (*i.e.*, attendue en absence d'autocorrélation spatiale), qui vaut exactement $-1/(n-1)$.

La question découlant de cette information est celle de l'identification de seuils, en termes de variance et d'autocorrélation, à partir desquels les composantes principales de la sPCA ne reflètent plus de structure biologique. La fonction graphique `screeplot.scpa` décompose chaque valeur propre en fonction de la variance et de l'autocorrélation correspondante (Fig. 3.4).

```
> screeplot(mySpca, main = "Décomposition des valeurs propres\nd'une sPCA")
```

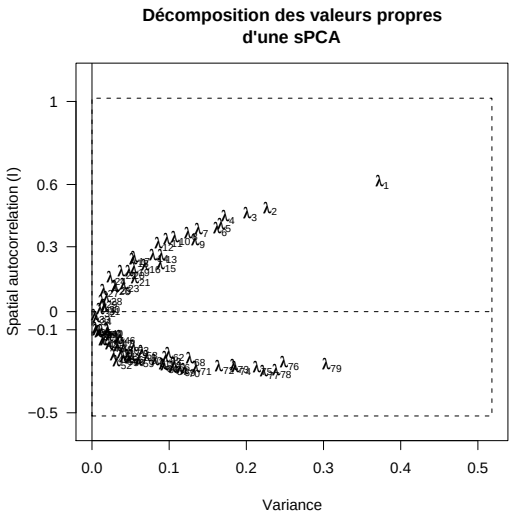


Fig. 3.4: Décomposition des valeurs propres d'une sPCA par la fonction `screeplot` (données : `spcaIllus$dat2A`). Les segments en pointillés indiquent en abscisse la variance maximale d'une composante principale, et en ordonnée les valeurs minimales et maximales du I de Moran (I_{min} et I_{max}) ; le segment central indique I_0.

Dans le cas de cette illustration, la première valeur propre λ_1 se distingue clairement du lot par sa variance et son autocorrélation spatiale : seule la composante principale correspondante sera interprétée. La fonction `plot.spca` fournit plusieurs représentations graphiques d'une composante donnée, ainsi que d'autres informations relatives à l'analyse (Fig. 3.5).

```
> plot(mySpca, axis = 1)
```

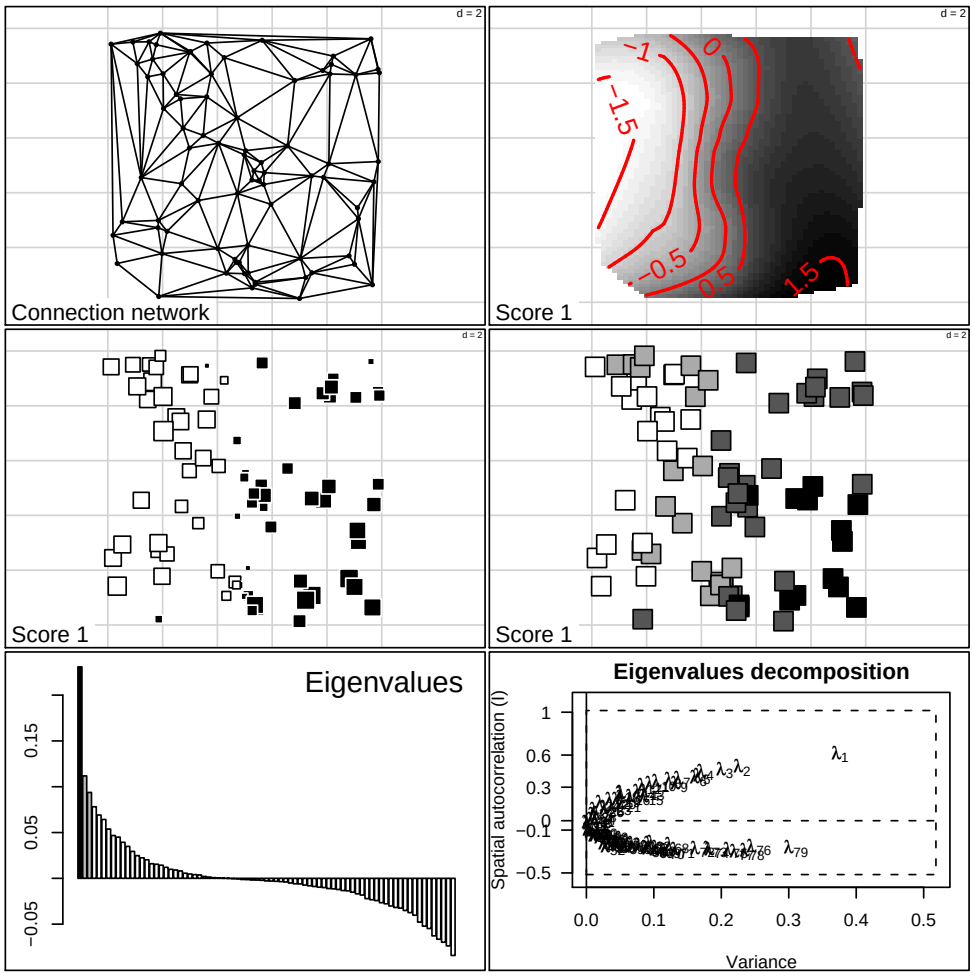


FIG. 3.5: Représentation graphique d'un objet `spca` par la fonction `plot.spca`. En haut : graphe de voisinage et représentation d'une composante principale avec interpolation et courbes de niveau. Au milieu : deux autres représentations d'une composante principale dans l'espace. En bas : graphe des valeurs propres, et décomposition par la fonction `screeplot.spca`.

Enfin, les tests globaux et locaux sont implémentés par les fonctions `global.rtest` et `local.rtest`. Ces fonctions requièrent une matrice de données et une pondération de voisinage, et retourne un objet de la classe `randtest`, qui est la classe standard des tests par permutations du package *ade4*.

```
> global.rtest(myObj$tab, mySpca$lw)


Monte-Carlo test
Call: global.rtest(X = myObj$tab, listw = mySpca$lw)

Observation: 0.01658103

Based on 499 replicates
Simulated p-value: 0.008
Alternative hypothesis: greater

     Std.Obs  Expectation     Variance
4.448953e+00 1.288479e-02 6.902459e-07


> local.rtest(myObj$tab, mySpca$lw)


Monte-Carlo test
Call: local.rtest(X = myObj$tab, listw = mySpca$lw)

Observation: 0.01397349

Based on 499 replicates
Simulated p-value: 0.142
Alternative hypothesis: greater

     Std.Obs  Expectation     Variance
1.061005e+00 1.314863e-02 6.044021e-07
```

Ici seul le test global est significatif, confirmant l'existence de structuration globale et l'absence de structure locale.

3.3.2 Application aux données du chamois des Bauges

Les résultats présentés plus bas sont le fruit d'une collaboration avec Stéphanie Cassar et Anne Loison, et ont donné lieu à la rédaction d'une publication (Cassar *et al.*, in revision). Cette collaboration impliquait également l'analyse d'un autre jeu de données portant sur le chevreuil (*Capreolus capreolus*), qui n'est pas présenté dans cette thèse, mais qui devrait également motiver la rédaction d'un article.

a. Les données

Le jeu de données `rupica` contient les génotypes géoréférencés de 335 chamois (*Rupicapra rupicapra*) du massif des Bauges (France) pour 9 marqueurs microsatellites. On se propose d'illustrer la sPCA en analysant ce jeu de données.

La question posée est celle du statut génétique de ces individus, pour lequel aucune connaissance *a priori* n'est disponible. On suppose que le relief est un élément structurant pour les populations de chamois, les individus se réunissant en groupes autour des sommets et évitant les vallées. Néanmoins, les observations de terrain sur le sujet sont encore en cours d'acquisition. On commence par charger les packages requis ainsi que le jeu de données proprement dit (`ripuca.RData`) :

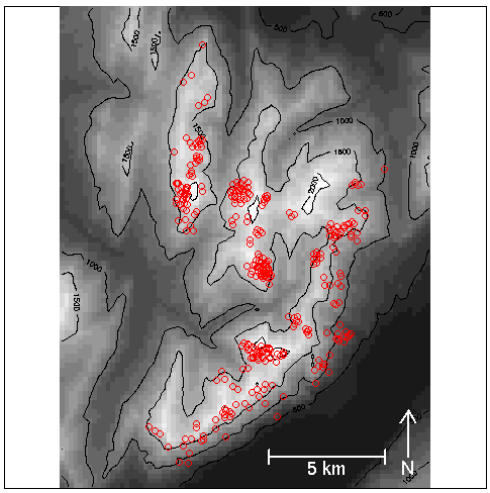


FIG. 3.6: Répartition des génotypes échantillonnés (cercles rouges) dans l'espace géographique

```
> library(adegenet)
> library(adehabitat)
> library(spdep)


> load("rupica.RData")
> rupica


   ######################
   ### Genind object ###
   ######################
- genotypes of individuals -

S4 class:  genind
@call: NULL

@tab:  335 x 55 matrix of genotypes

@ind.names: vector of  335 individual names
@loc.names: vector of  9 locus names
@loc.nall: number of alleles per locus
@loc.fac: locus factor for the  55 columns of @tab
@all.names: list of  9 components yielding allele names for each locus
@ploidy:  2

Optionnal contents:
@pop:  - empty -
@pop.names:  - empty -

@other: a list containing: xy  sex  age  behav  mnt  showBauges
```

L'object `rupica` est un objet `genind` contenant plusieurs éléments additionnels dans l'item `other`; en particulier, la fonction `showBauges` présente une carte topographique sommaire de l'aire d'échantillonnage.

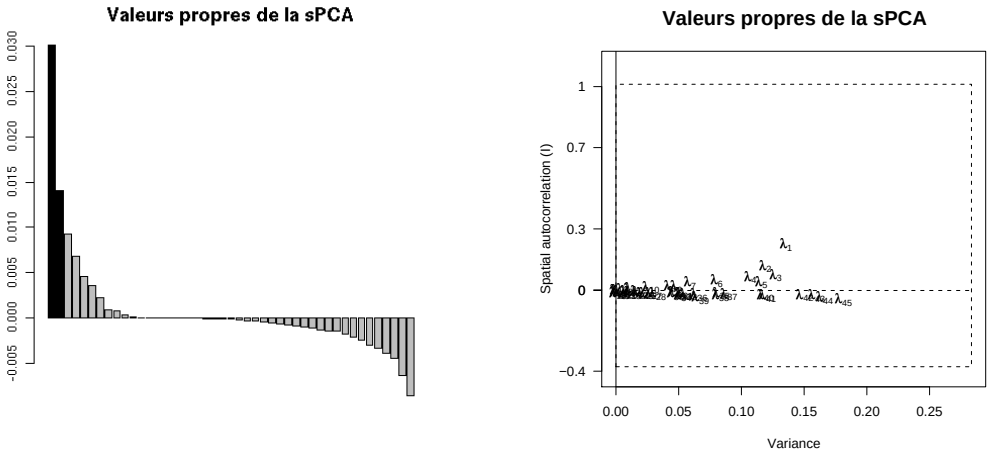
```
> mnt <- rupica$other$mnt
> showBauges <- rupica$other$showBauges
> showBauges()
> points(rupica$other$xy, col = "red")
```

b. Analyse du jeu de données

On ne présente ici que la sPCA. Pour prendre en compte la distribution spatiale des génotypes, il semble logique de considérer qu'un individu situé au centre d'un agrégat possède

plus de voisins qu'un individu périphérique. Cette propriété devrait être reflétée par le graphe de voisinage. Pour ce faire, nous pouvons utiliser une définition du voisinage basée sur la distance : seront considérés comme voisins deux individus séparés par une distance inférieure à une distance seuil. Le choix le plus naturel pour ce seuil est d'utiliser la distance à laquelle les espaces vitaux des individus se recoupent. L'espace vital du chamois dans cette réserve pouvant être modélisé par un disque d'environ 420 ha en moyenne (Darmon, pers. com.), on choisira donc une distance seuil de 2300m.

```
> spca1 <- spca(rupica, type = 5, d1 = 0, d2 = 2300, plot = FALSE,
+     scannf = FALSE, nfposi = 2, nfnega = 0)
> barplot(spca1$eig, col = rep(c("black", "grey"), c(2, 100)),
+     main = "Valeurs propres de la sPCA", cex.main = 1.5)


> par(cex.main = 1.5)
> screeplot(spca1, main = "Valeurs propres de la sPCA")
```



(a) Screeplot (b) Décomposition des valeurs propres

FIG. 3.7: Valeurs propres de la sPCA des données des chamois des Bauges

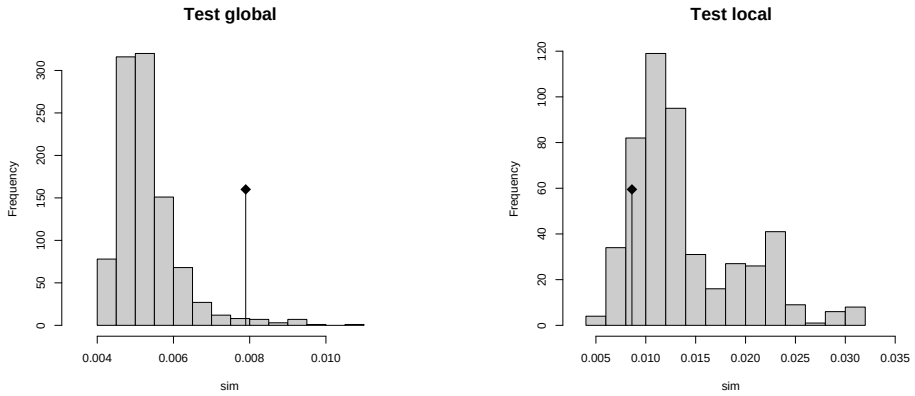
On retient deux structures globales correspondant aux deux premières valeurs propres λ_1 et λ_2 de l'analyse (FIG. 3.7). Notons que si la première valeur propre se détache nettement des autres, elle ne représente que la moitié de la variance pouvant être représentée par une seule dimension (FIG. 3.7b). S'il existe bien une structure globale "significative", celle-ci ne représente pas la majeure partie de la variabilité génétique.

On procède aux tests de structures globales et locales présentés dans Jombart *et al.* (2008) :

```
> Gtest <- global.rtest(rupica@tab, spca1$lw, nperm = 999)
> Ltest <- local.rtest(rupica@tab, spca1$lw, nperm = 999)

> plot(Gtest, main = "Test global", cex.main = 1.5)

> plot(Ltest, main = "Test local", cex.main = 1.5)
```

On constate qu'il y a effectivement une structuration globale significative (FIG. 3.8a), et pas de structure locale (FIG. 3.8b).

On peut visualiser les deux structures globales retenues en cartographiant les scores (*i.e.*, les composantes principales) lissés de la sPCA.



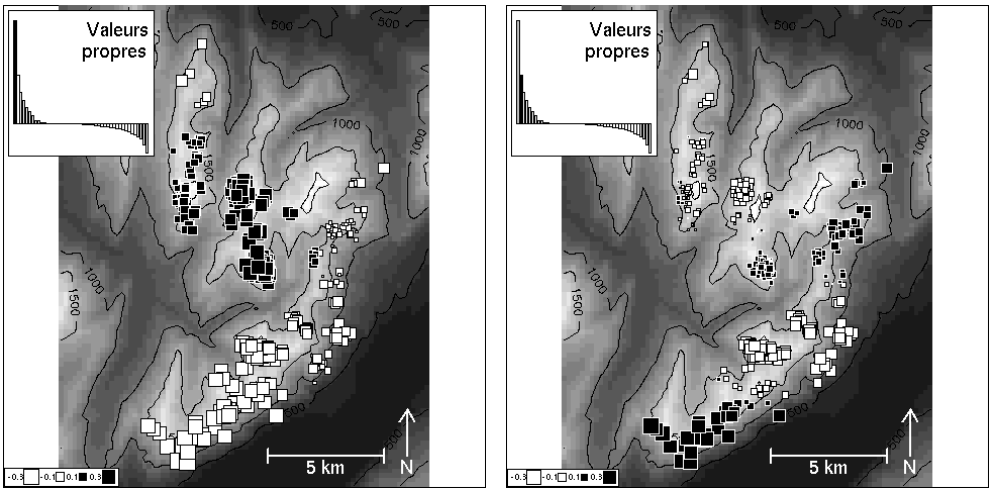
(a) max(**t**) = 0.008, p = 0.020 (999 permutations)

(b) max(**t**) = 0.009, p = 0.88 (999 permutations)

FIG. 3.8: Test global et local appliqués aux données des chamois des Bauges. Les histogrammes correspondent aux valeurs permutées de la statistique de test. La statistique observée est indiquée en noir.

```
> showBauges(, labcex = 1)
> s.value(spca1$xy, spca1$ls[, 1], add.p = TRUE, csize = 0.5)
> add.scatter.eig(spca1$eig, 1, 1, 1, posi = "topleft", csub = 1.5,
+     ratio = 0.3, sub = "Valeurs \npropres")
```

La première composante principale lissée oppose clairement deux sous-populations séparées par la principale vallée (FIG. 3.9a). Cette structuration est cohérente avec l'idée que les vallées constituent des freins à la dispersion du chamois (Loison *et al.*, 1999).



(a) Première composante globale

(b) Seconde composante globale

FIG. 3.9: sPCA des données des chamois des Bauges : composantes principales lissées

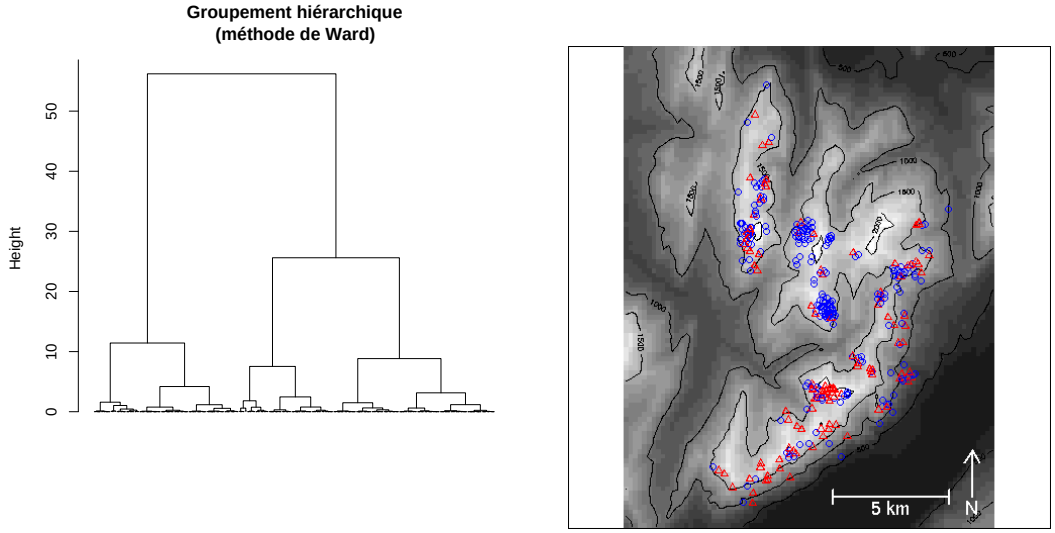
```
> showBauges(, labcex = 1)
> s.value(spca1$xy, spca1$ls[, 2], add.p = TRUE, csize = 0.5)
> add.scatter.eig(spca1$eig, 2, 2, 2, posi = "topleft", csub = 1.5,
+     ratio = 0.3, sub = "Valeurs \npropres")
```

La seconde composante principale lissée marque surtout une opposition est-ouest au sein des individus des massifs du sud (FIG. 3.9b). On note que les génotypes les plus au nord, ayant des scores proches de zéro, sont peu concernés par cette structuration. Étant donnée l'absence de

barrière environnementale entre ces deux patches, il est plus probable que cette structure reflète une structuration sociale.

La première composante principale globale de la sPCA (Fig. 3.9a) implique tous les génotypes, contrairement à la seconde qui ne montre qu'une différenciation génétique entre les génotypes du sud (Fig. 3.9b). On peut essayer de quantifier la différenciation génétique observée sur la figure 3.9a. On commence par définir formellement les deux groupes de génotypes perçus (Fig. 3.9a) ; on utilise le groupement hiérarchique de Ward (Legendre & Legendre, 1998, pp.329-333) pour définir deux groupes bien distincts (Fig. 3.10a).

```
> distScores <- dist(spca1$li[, 1])
> arb <- hclust(distScores, method = "ward")
> plot(arb, main = "Groupement hiérarchique\n(méthode de Ward)",
+     xlab = "", sub = "", lab = FALSE, hang = 0)
```



(a) Dendrogramme

(b) Représentation des groupes dans l'espace géographique

Fig. 3.10: Groupement de Ward basé sur la première composante principale globale de la sPCA.

On peut également représenter ces groupes dans l'espace géographique (Fig. 3.10b). La différenciation géographique apparaît moins clairement que les scores lissés (Fig. 3.9a), mais est plus proche de la réalité biologique.

```
> subpop <- cutree(arb, k = 2)
> showBauges()
> points(rupica$other$xy, pch = subpop, col = c("blue", "red")[subpop])
```

Une fois les deux groupes d'individus définis, la fonction `fstat` permet de calculer le F_{ST} entre ces groupes.

```
> rupica.Fst <- fstat(rupica, subpop, fstonly = TRUE)
> rupica.Fst
```

```
[1] 0.03698181
```

Celui-ci est d'environ 0.04, ce qui correspond à une différenciation génétique modérée, mais qui est exactement du même ordre que les valeurs de F_{ST} observées dans les données simulées illustrant la sPCA (Jombart *et al.*, 2008).

On terminera par une représentation simultanée des deux scores globaux de la sPCA en utilisant le système de couleurs RGB comme suggéré par Menozzi *et al.* (1978). L'idée est d'exprimer chaque composante principale comme une intensité de couleur (Rouge, Verte ou Bleue), et d'utiliser les compositions résultantes pour résumer l'information. Ici, seules les couleurs rouges et vertes entrent en jeu, puisque seules deux composantes ont été retenues. La représentation graphique (Fig. 3.11) est obtenue par la fonction `colorplot`, qui est en cours de développement.

```
> rupica$other$showBauges()
> colorplot(spca1$xy, spca1$li, cex = 1.5, add.plot = TRUE)
```

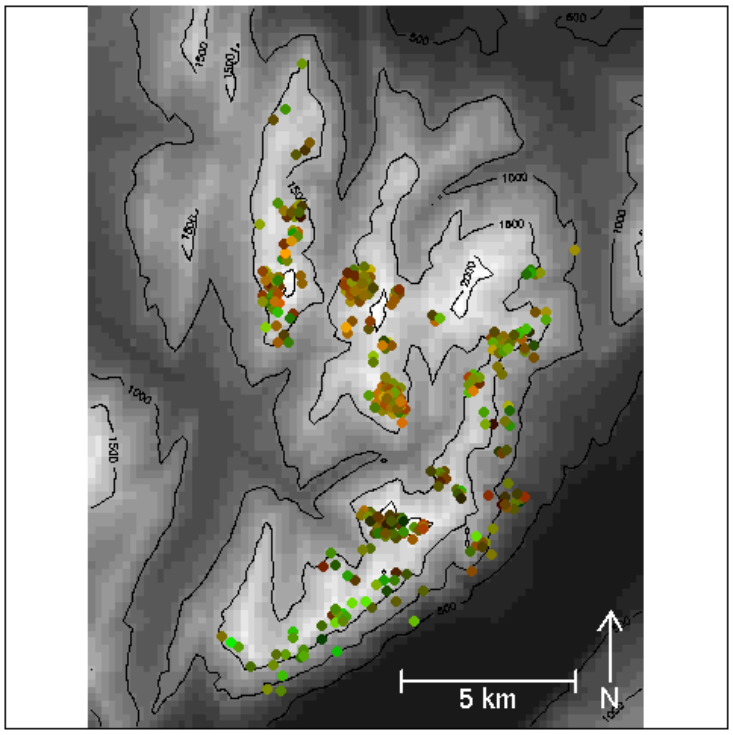


Fig. 3.11: Représentation en couleur (fonction `colorplot`) des composantes globales de sPCA des données des chamois des Bauges. Les deux premières composantes globales sont traduites en niveaux de rouge et de vert, les compositions résultantes résumant les proximités génétiques entre individus.

On observe un noyau de génotypes proches génétiquement au sud-ouest (Fig. 3.11, en vert), ainsi qu'un autre au nord (Fig. 3.11, tons orangés), la zone est semblant être un mélange des deux catégories. Bien que la représentation basée sur le système RGB semble intéressante, on notera que ses propriétés sont mal connues ; on ne peut donc que recommander la prudence quant à son utilisation.

3.4 Discussion

3.4.1 Critique de la méthode

a. Asymétrie du I de Moran

Une première critique pouvant être faite sur le I de Moran concerne la distribution de cet indice, et plus précisément son intervalle de variation. Les valeurs minimales ($I_{\min}$) et maximales ($I_{\max}$) du I dépendent entièrement de la matrice de pondérations de voisinage utilisée. Dans la pratique, on observe souvent un intervalle de variation plus large dans les valeurs positives que dans les valeurs négatives, c'est-à-dire :

$$|I_{\min} - I_0| < |I_{\max} - I_0| \tag{3.4}$$

Il en résulte qu'à variances égales, les valeurs propres des structures locales seraient nécessairement plus faibles que les valeurs propres globales. Les structures locales seraient donc plus difficiles à détecter.

On peut étudier empiriquement ce phénomène en observant la distribution de $I_{\min}$ et $I_{\max}$ pour différents graphes de voisinage. On propose d'étudier ces distributions pour la triangulation de Delaunay, le graphe de Gabriel, les K plus proches voisins et le voisinage par la distance pour des distributions de 100 points selon une loi uniforme, puis selon une loi normale. Par souci de simplicité, on fixe arbitrairement $K = 10$, et le voisinage par la distance utilisera la distance minimale telle que chaque point ait au moins un voisin.

La fonction `simExtrI` effectue ces simulations selon une loi (argument `distr`), un type de graphe (`type`) et un nombre d'itérations (`nsim`).

```
> simExtrI <- function(distr = c("unif", "norm"), type = 1, nsim = 100) {
+     if (!require(ade4))
+         stop("ade4 package required")
+     distr <- match.arg(distr)
+     vecImin <- numeric(nsim)
+     vecImax <- numeric(nsim)
+     if (distr == "unif") {
+         for (i in 1:nsim) {
+             xy <- matrix(runif(200), ncol = 2)
+             cn <- chooseCN(xy, type = type, plot = FALSE, k = 10,
+                 d1 = 0, d2 = "dmin", res = "listw")
+             rangeI <- range(attr(orthobasis.listw(cn), "values"))
+             vecImin[i] <- rangeI[1]
+             vecImax[i] <- rangeI[2]
+         }
+     }
+     if (distr == "norm") {
+         for (i in 1:nsim) {
+             xy <- matrix(rnorm(200), ncol = 2)
+             cn <- chooseCN(xy, type = type, plot = FALSE, k = 10,
+                 d1 = 0, d2 = "dmin", res = "listw")
+             rangeI <- range(attr(orthobasis.listw(cn), "values"))
+             vecImin[i] <- rangeI[1]
+             vecImax[i] <- rangeI[2]
+         }
+     }
+     return(list(Imin = vecImin, Imax = vecImax))
+ }
```

Par exemple pour une simulation :

```
> simExtrI(nsim = 1)
```

```
$Imin
[1] -0.5036986

$Imax
[1] 0.9762682
```

On effectue maintenant l'ensemble des simulations :

```
> simUnifDelau <- simExtrI(nsim = 1000)
> simUnifGab <- simExtrI(type = 2, nsim = 1000)
> simUnifKnn <- simExtrI(type = 6, nsim = 1000)
> simUnifNDist <- simExtrI(type = 5, nsim = 1000)
> simNormDelau <- simExtrI(distr = "norm", nsim = 1000)
> simNormGab <- simExtrI(distr = "norm", type = 2, nsim = 1000)
> simNormKnn <- simExtrI(distr = "norm", type = 6, nsim = 1000)
> simNormNDist <- simExtrI(distr = "norm", type = 5, nsim = 1000)
> save.image()
```

On représente enfin les distributions correspondantes :

```
> par(mfcol = c(4, 2), mar = c(2, 1, 4, 1), yaxt = "n")
> hist(unlist(simUnifDelau), nclass = 150, col = "blue", border = "blue",
+     main = "Uniforme\nDelaunay", xlab = "I", xlim = c(-1, 1))
> abline(v = -1/99, lwd = 3, lty = 3)
> hist(unlist(simUnifGab), nclass = 150, col = "blue", border = "blue",
+     main = "Uniforme\nGabriel", xlab = "I", xlim = c(-1, 1))
> abline(v = -1/99, lwd = 3, lty = 3)
> hist(unlist(simUnifKnn), nclass = 150, col = "blue", border = "blue",
+     main = "Uniforme\nK plus proches voisins", xlab = "I", xlim = c(-1,
+         1))
> abline(v = -1/99, lwd = 3, lty = 3)
> hist(unlist(simUnifNDist), nclass = 150, col = "blue", border = "blue",
+     main = "Uniforme\nDistance minimale", xlab = "I", xlim = c(-1,
+         1))
> abline(v = -1/99, lwd = 3, lty = 3)
> hist(unlist(simNormDelau), nclass = 150, col = "blue", border = "blue",
+     main = "Normale\nDelaunay", xlab = "I", xlim = c(-1, 1))
> abline(v = -1/99, lwd = 3, lty = 3)
> hist(unlist(simNormGab), nclass = 150, col = "blue", border = "blue",
+     main = "Normale\nGabriel", xlab = "I", xlim = c(-1, 1))
> abline(v = -1/99, lwd = 3, lty = 3)
> hist(unlist(simNormKnn), nclass = 150, col = "blue", border = "blue",
+     main = "Normale\nK plus proches voisins", xlab = "I", xlim = c(-1,
+         1))
> abline(v = -1/99, lwd = 3, lty = 3)
> hist(unlist(simNormNDist), nclass = 150, col = "blue", border = "blue",
+     main = "Normale\nDistance minimale", xlab = "I", xlim = c(-1,
+         1))
> abline(v = -1/99, lwd = 3, lty = 3)
```

Les résultats montrent d'abord qu'il y a peu de différences entre loi uniforme et loi normale (Fig. 3.12), sauf peut-être pour le voisinage par la distance minimale (Fig. 3.12, graphiques du bas), pour lequel la variance de $I_{\max}$ semble plus forte pour une loi normale que pour une loi uniforme. Par contre, on constate que le choix du graphe influence clairement les valeurs de $I_{\min}$, alors que $I_{\max}$ est en général stable autour de 1. On peut donc penser que si le graphe de Gabriel est assez "impartial" vis-à-vis du type de structure (globale ou locale), le graphe de Delaunay et celui des K plus proches voisins mettront plus difficilement en évidence des structures locales. Enfin, on constate une grande dispersion des valeurs extrême du I pour le voisinage par la distance minimale utilisé pour une loi normale. Ce résultat peut être expliqué par le fait que la distance à laquelle les points sont considérés comme voisins est directement dépendante des "outliers", qui conditionnent donc les propriétés du graphe.

On retiendra que le choix du graphe peut donc, dans certains cas, influencer l'efficacité de la sPCA pour détecter des structures locales.

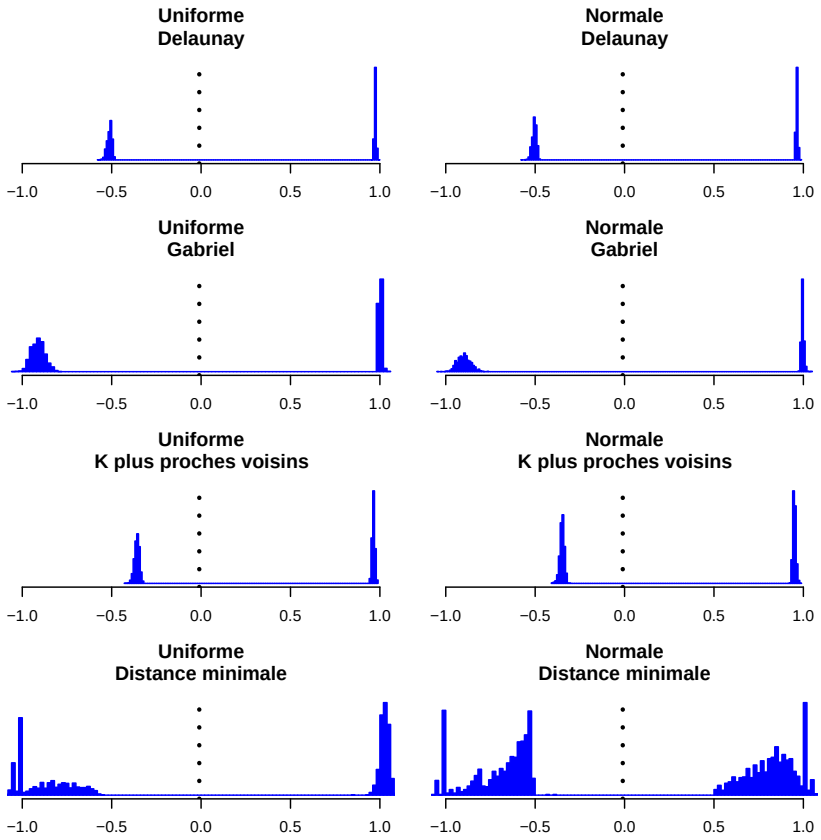


FIG. 3.12: Distributions empiriques (histogrammes bleus) des valeurs minimales $I_{\min}$ et maximales $I_{\max}$ de l'indice de Moran, pour différentes distributions et différents graphes. La ligne en pointillés indique I_0, la valeur du I correspondant à l'absence d'autocorrélation.

b. Structures artefactuelles

Lorsqu'on développe un nouveau test statistique visant à comparer des mesures sur un ensemble d'échantillons, on s'intéresse à sa capacité à voir les différences (sa puissance), mais aussi à son erreur de type I, c'est-à-dire à la fréquence à laquelle le test voit des différences là où elles n'existent pas. A l'émergence d'une nouvelle méthode exploratoire, la première question est celle de son efficacité à trouver des structures dans un jeu de données. Peu souvent, on s'intéresse à sa capacité à trouver des structures qui n'existent pas. Cette question est particulièrement délicate, puisqu'à la différence d'un test statistique (à l'interprétation, on l'espère, plus universelle), les résultats d'une méthode exploratoire sont sujets à interprétation, et donc en partie subjectifs.

On renonce donc à évaluer des équivalents d'erreurs de type I et II dans le cas de la sPCA. Mais si l'on a vu dans l'article présenté que la sPCA est plus efficace que l'ACP pour révéler des structures spatiales, on peut se demander quels sont les résultats des méthodes en absence de structure.

La tâche est assez simple : il suffit d'analyser un jeu de données n'ayant aucune structuration génétique et de comparer les résultats. On utilisera pour ce faire le jeu de données `sim2pop`, qui contient 130 génotypes simulés pour 20 microsatellites répartis en deux populations. On ne garde

que les 100 premiers génotypes qui appartiennent tous à une même population panmictique.

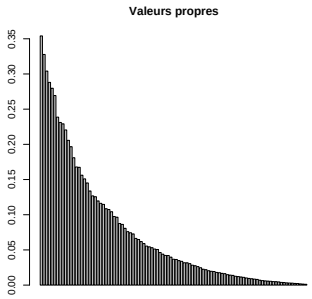
```
> data(sim2pop)
> obj <- sim2pop[1:100]
> obj


   ####################
   ### Genind object ###
   ####################
- genotypes of individuals -

S4 class:  genind
@call: .local(x = x, i = i, j = j, drop = drop)

@tab:  100 x 241 matrix of genotypes

@ind.names: vector of  100 individual names
@loc.names: vector of  20 locus names
@loc.nall: number of alleles per locus
@loc.fac: locus factor for the  241 columns of @tab
@all.names: list of  20 components yielding allele names for each locus
@ploidy:  2

Optionnal contents:
@pop:  factor giving the population of each individual
@pop.names:  factor giving the population of each individual

@other: a list containing: xy
```

On définit aléatoirement des coodonnées spatiales pour chaque génotype :

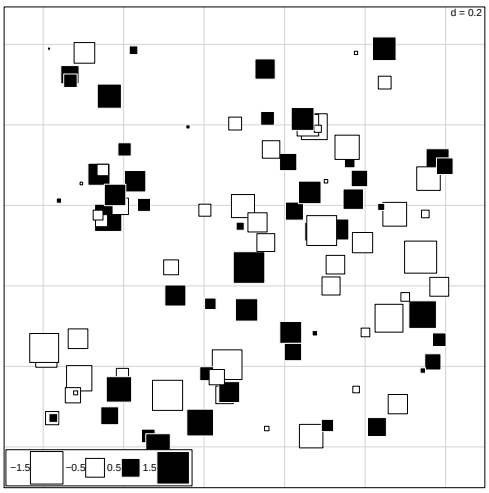
```
> xy <- matrix(runif(200), ncol = 2)
```

On procède à l'ACP du jeu de données, en cartographiant la première composante principale.

```
> myPca <- dudi.pca(obj, scale = FALSE, scannf = FALSE, nf = 1)
> barplot(myPca$eig, main = "Valeurs propres")


> s.value(xy, myPca$li[, 1], include.ori = FALSE, addaxes = FALSE)
```



(a) Graphe des valeurs propres



(b) Cartographie de la première composante principale

FIG. 3.13: Analyse en composantes principales de l'objet `obj`, contenant 100 génotypes d'un groupe panmictique.

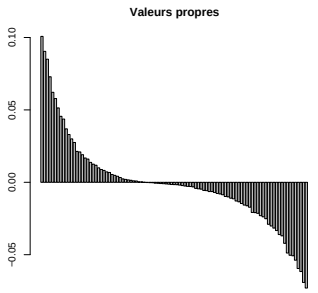
La figure (FIG. 3.13) montre que l'ACP ne révèle aucune structure génétique spatialisée.

A présent, on réitère la même opération en utilisant la sPCA.

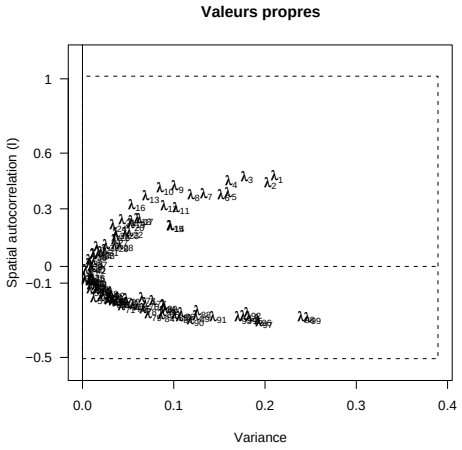
```
> mySpca <- spca(obj, xy = xy, scannf = FALSE, nfposi = 1, type = 1,
+     plot = FALSE)
> barplot(mySpca$eig, main = "Valeurs propres")
```

```
> screeplot(mySpca, main = "Valeurs propres")
```

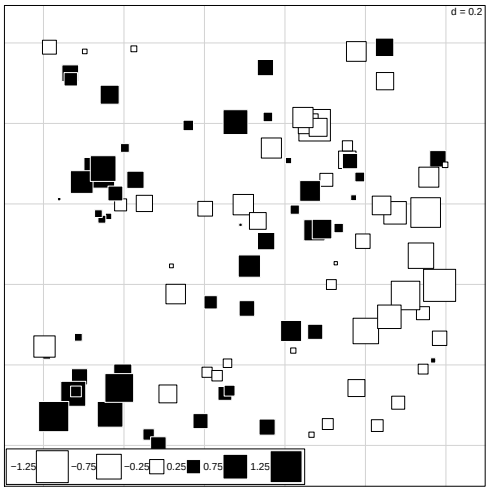
```
> s.value(xy, mySpca$li[, 1], include.ori = FALSE, addaxes = FALSE)
```



(a) Graphe des valeurs propres



(b) Décomposition des valeurs propres



(c) Cartographie de la première composante principale

FIG. 3.14: Analyse en composantes principales spatiales (sPCA) de l'objet `obj`, contenant 100 génotypes d'un groupe panmictique, utilisant la triangulation de Delaunay.

On constate que si les valeurs propres de l'analyse suggèrent l'absence de structure (FIG. 3.14a et b), la cartographie de la première composante principale globale semble par contre révéler deux patches de génotypes (FIG. 3.14c).

Ceci est encore plus flagrant si on examine les scores lissés (FIG. 3.15)

```
> s.value(xy, mySpca$ls[, 1], include.ori = FALSE, addaxes = FALSE)
```

On peut s'interroger sur l'origine de cette structure purement artefactuelle. L'explication la plus vraisemblable tient au critère optimisé : la sPCA recherche des combinaisons linéaires

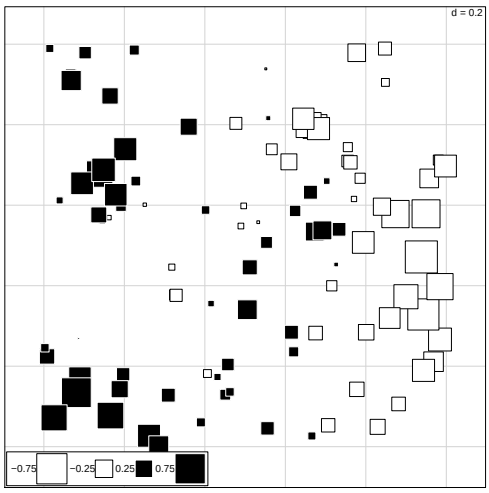


FIG. 3.15: sPCA de l'objet `obj`, contenant 100 génotypes d'un groupe panmictique, utilisant la triangulation de Delaunay. Cartographie de la première composante principale globale lissée

d'allèles ayant des valeurs extrêmes pour le produit de la variance et du I de Moran. Cette optimisation demande donc à la fois de maximiser la variance et de fournir une valeur d'autocorrélation extrême (positive ou négative). En l'absence de structuration génétique, le nuage des génotypes est sphérique : il n'y a plus alors d'axe principal, et toutes les directions se valent en terme d'inertie projetée. La maximisation de la variance n'est donc plus une contrainte. Or, parmi toutes ces directions équivalentes du point de vue de l'inertie, on peut toujours en trouver qui, par chance, optimise l'autocorrélation.

Lorsqu'il n'y a pas de structuration (non spatiale) dans les données, c'est-à-dire lorsque la notion d'axe principal est caduque, on trouvera donc souvent, si ce n'est toujours, la cartographie des scores de la sPCA interprétable. Il est par conséquent essentiel de n'interpréter celle-ci que lorsque le graphe des valeurs propres et leur décomposition en terme de variance et d'autocorrélation indiquent clairement une structuration. Par ailleurs, les tests globaux et locaux apportent également une information sur l'existence respective des deux types de structures. En l'occurence, ces tests confirment l'absence de structures des deux types :

```
> global.rtest(obj$tab, mySpca$lw)


Monte-Carlo test
Call: global.rtest(X = obj$tab, listw = mySpca$lw)

Observation: 0.009162928

Based on 499 replicates
Simulated p-value: 0.722
Alternative hypothesis: greater

      Std.Obs   Expectation      Variance
-6.877335e-01  9.485695e-03  2.202605e-07


> local.rtest(obj$tab, mySpca$lw)


Monte-Carlo test
Call: local.rtest(X = obj$tab, listw = mySpca$lw)

Observation: 0.009711178

Based on 499 replicates
```

```
Simulated p-value: 0.5
Alternative hypothesis: greater

      Std.Obs   Expectation      Variance
-1.978862e-01  9.819502e-03  2.996571e-07
```

3.4.2 Perspectives

L'introduction de la sPCA en génétique devrait offrir de nouvelles perspectives pour l'analyse de la structuration spatiale de la variabilité génétique. La méthode offre en effet une alternative aux méthodes spatialement explicites de groupement bayésien (Guillot *et al.*, 2005, 2006; François *et al.*, 2006) qui sont critiquables à plusieurs niveaux. Outre le fait qu'elle soient exigentes en temps de calcul (de l'ordre de plusieurs jours sur une machine standard pour STRUCTURE, contre quelques secondes pour la sPCA et les tests associés), ces approches font des hypothèses fortes sur les données telles que l'état d'équilibre d'Hardy-Weinberg, ou l'absence de déséquilibre de liaison. Par ailleurs, ces approches imposent une unique vision du voisinage : par exemple, Geneland (Guillot *et al.*, 2005, 2006) ne peut prendre en compte l'information spatiale qu'au travers de la triangulation de Delaunay, bien que ce graphe ne puisse être adapté à toutes les distributions spatiales. En outre, les méthodes de groupements sont par définition des stratégies inadaptées lorsque les génotypes ou les populations sont structurés selon un gradient de différenciation génétique.

La sPCA ouvre également des pistes de recherche prometteuses. Dans le cadre de la génétique du paysage, on cherche à corréler la structuration spatiale de la variabilité génétique et celle de l'environnement (Manel *et al.*, 2003). Dans ce contexte, la sPCA remplit le pré-requis implicite de l'identification des structures génétiques spatiales. La question de la génétique du paysage se pose alors en terme de couplage d'analyses sous contraintes spatiales. Il sera sans doute nécessaire de s'interroger sur les mesures possibles de la co-structuration spatiale, une question déjà abordée en géographie (Lee, 2001) mais qui demeure ouverte.

Par ailleurs, on peut également s'intéresser à la cohérence typologique des marqueurs moléculaires, abordée dans le chapitre précédent, dans le contexte spatial. Si un ensemble de marqueurs génétiques montre une structure spatiale, quelle est la participation de chaque marqueur à cette structure ? Dans le cas de marqueurs neutres, on s'attendrait à ce que tous les marqueurs exhibent la même structuration spatiale avec une magnitude comparable. Cependant, Smouse & Peakall (1999) remarquent que :

> *Contrary to theoretical expectations, however, spatial structure is rarely consistent across loci or sites*

En réalité, certains marqueurs supposés neutres peuvent être associés à des régions sélectionnées du génome, cette sélection pouvant induire différentes structures spatiales. Cependant, l'analyse séparée de chaque locus pour étudier l'existence de structures spatiales est problématique. Par exemple dans Devillard *et al.* (sous presse, jeu de données `nancycats`), nous appliquons des tests du I de Moran sur les premières composantes principales d'ACP faites séparément pour chaque marqueur de l'étude, en négligeant d'utiliser des corrections pour les tests multiples. Sur cette erreur (la mienne, en l'occurence), nous avons conclu à l'existence d'une structuration spatiale

faible, alors qu'il est très vraisemblable qu'il n'en existe aucune. Il semble donc que la question de la cohérence des marqueurs génétiques en terme de structuration spatiale soit aussi intéressante, si ce n'est plus, que celle déjà abordée de leur cohérence typologique.

Chapitre 4

Un cadre technique pour l'analyse des marqueurs génétiques

Sommaire

4.1 Point de départ

Dans le dialogue entre mathématique et biologie auquel le biométricien participe, l'outil informatique joue un rôle prépondérant : la « meilleure méthode » pour analyser les données, c'est souvent celle qui est disponible. Cette affirmation est sans doute appelée à être de plus en plus contredite si l'on considère l'expansion de logiciels scientifiques libres, qui mettent à disposition de l'utilisateur des outils de qualité, sans cesse éprouvés et rectifiés. Dans le domaine de la statistique et de l'analyse de données (au sens large), le logiciel R (R Development Core Team, 2008) est devenu un outil incontournable, qui implémente un nombre de méthodes sans cesse croissant, dont un grand nombre d'analyses multivariées et un ensemble non négligeable d'outils pour l'analyse de données génétiques.

L'analyse multivariée de marqueurs génétiques dans le logiciel R n'est donc pas, en premier lieu, un problème de disponibilité des méthodes. La bibliothèque de fonctions (ou 'package') *ade4* (Chessel *et al.*, 2004; Dray *et al.*, 2007) propose un vaste choix d'ordinations en espace réduit implémentées dans le cadre du schéma de dualité (Dray & Dufour, 2007). Si ces méthodes demandent parfois certaines adaptations à la donnée moléculaire, elles sont globalement utilisables, et utilisées, dans leur forme première (Jombart *et al.*, in revision). La première difficulté posée à l'analyse multivariée de marqueurs génétiques dans R est avant tout d'ordre pratique : avant d'analyser les données, il faut pouvoir les lire, les stocker et les manipuler. Et souvent, les exporter, puisque une analyse standard de ces données fait appel à de nombreux outils autres que l'analyse multivariée. Le package *adegenet* a été développé pour satisfaire ces besoins, et d'autres apparus par la suite. Ce chapitre présente le contexte dans lequel le package est apparu, son état actuel et quelques perspectives de développement.

4.1.1 L'analyse de données génétiques dans R

On peut situer les débuts « officiels » de l'analyse de marqueurs moléculaires dans R à l'apparition du package *genetics* (Warnes, 2003), dont les première versions sont disponibles dès mai 2001. Ces fonctionnalités sont alors nécessairement restreintes : on peut stocker des données génotypiques ou haplotypiques, calculer des fréquences alléliques, et effectuer le test d'Hardy-Weinberg. Les évolutions du package (`http://cran.r-project.org/web/packages/genetics/ChangeLog`) prennent logiquement la direction d'outils « classiques » de génétique des populations, comme des mesures du déséquilibre de liaison. Au moment où commence cette thèse, d'autres packages viennent compléter le choix d'outils pour la génétique des populations, tels que *hierfstat* qui propose des mesures et des tests de la structuration hiérarchique d'un ensemble de génotypes (Goudet, 2005). L'analyse multivariée des marqueurs génétiques n'est alors possible que via *ade4*, mais demeure un calvaire pour l'utilisateur occasionel.

4.1.2 Les données génétiques dans ade4

Comme son nom l'indique, *adegenet* procède du package *ade4*, qui était le premier à mettre des éléments à disposition pour l'analyse multivariée de marqueurs génétiques. En premier lieu, *ade4* définit la classe d'objets `genet`, qui a inspiré en partie les classes définies dans *adegenet*. La classe `genet` est une liste possédant des éléments définis, l'élément central étant un tableau

de fréquences alléliques (`$tab`) portant des génotypes en ligne et des allèles en colonne. On parle bien de fréquence au niveau individuel : les génotypes sont codés de façon à ce que la somme des termes soit unitaire par locus. Pour des génotypes diploïdes, un homozygote sera codé par un 1 à l'allèle concerné, alors qu'un hétérozygote portera 0,5 à deux allèles (les allèles non portés étant codés par 0). Un objet `genet` peut s'obtenir depuis des tableaux de chaînes de caractères (`char2genet`), de dénombrements d'allèles par population (`count2genet`), ou de fréquences alléliques (`freq2genet`). Une contrainte surprenante est que la classe `genet` suppose que les groupes de génotypes (appelés plus loin « populations ») soient connus à l'avance. Lorsque ce n'est pas le cas, la fonction `fuzzygenet` permet de créer des objets ressemblant à la classe `genet`, mais sans classe officielle.

On peut illustrer ces différences en utilisant le jeu de données `casitas`. Ce jeu de données issu du logiciel GENETIX (Belkhir *et al.*, 2001) contient les génotypes de 74 souris pour 15 marqueurs microsatellite.

```
> library(ade4)
> data(casitas)
> head(casitas)[, 1:5]
```

```
     Aat    Amy    Es1    Es2   Es10
1 100100 080080 094094 100100 100100
2 100100 080100 094094 100100 100100
3 100100 080080 094094 100100 100100
4 100100 080080 094094 100100 100100
5 100100 080080 094094 100100 100100
6 100100 080100 094094 100100 100100
```

La documentation du jeu de données nous informe que les individus sont répartis en 4 espèces, comme suit :

```
> species <- factor(rep(1:4, c(24, 11, 9, 30)))
```

L'information sur l'espèce est requise par la classe `genet`, et donc par `char2genet` :

```
> casi.genet <- char2genet(casitas, species)
> class(casi.genet)
```

```
[1] "genet" "list"
```

```
> names(casi.genet)
```

```
[1] "tab"        "center"     "pop.names"  "all.names"  "loc.blocks"
[6] "loc.fac"    "loc.names"  "pop.loc"    "all.pop"
```

Néanmoins il arrivera très souvent que les groupes de génotypes ne soient pas connus à l'avance. En biologie de la conservation, c'est même ce que l'on cherche à déterminer en premier lieu (Moritz, 1994). A cet égard, l'analyse multivariée se montre particulièrement pertinente, de par son côté exploratoire. On peut toutefois conserver cette démarche exploratoire en utilisant `fuzzygenet` :

```
> casi.bis <- fuzzygenet(casitas)
> class(casi.bis)
```

```
[1] "data.frame"
```

```
> names(attributes(casi.bis))
```

```
[1] "names"      "row.names"  "class"      "col.blocks" "all.names"
[6] "loc.names"  "row.w"      "col.freq"   "col.num"
```

On constate que l'objet `casi.bis` diffère en plusieurs points de `casi.genet` : ce n'est pas une classe particulière, mais un `data.frame` ayant quelques attributs supplémentaires, recoupant en partie les éléments de la classe `genet`.

Cette confusion, sans doute rédibitoire pour l'utilisateur non averti, montre les limites de cette implémentation. Par ailleurs, la classe `genet` n'offre aucun moyen de manipuler l'information. Par exemple, si l'on veut isoler un sous-échantillon de génotypes ou de marqueurs, il faudra effectuer la sélection pour tous les éléments de l'objet (un par un), en faire une nouvelle liste et lui assigner la classe `genet`. Cependant, on ne peut reprocher à un package dédié à l'analyse multivariée une gestion limitée de la donnée moléculaire : les besoins à couvrir sont nombreux, et motivent pleinement le développement d'un package propre.

4.1.3 Les besoins

Le premier besoin technique était donc de définir des classes d'objets stables, qui puissent représenter les types de données les plus courants, et qui permettent un accès immédiat à l'analyse multivariée. Il est aussi nécessaire que les classes définies soient cohérentes avec une démarche exploratoire dans laquelle les groupes de génotypes ne sont pas connus à priori. Ce qui n'empêche pas de travailler au niveau populationnel lorsque cela est possible. La représentation des données devrait aussi, idéalement, permettre de prendre en compte des données ayant différents niveaux de ploïdie.

Une fois la représentation de l'information définie, la principale tâche consiste à lire les données. Pour faire de l'analyse de marqueurs génétiques dans R, la première chose à reconnaître est que personne, ou presque, n'en fait. Il est donc essentiel de permettre à l'utilisateur curieux d'importer simplement ses données sous R depuis d'autres logiciels. La seconde déconvenue, qui survient dès lors que l'on s'intéresse aux logiciels de génétique des populations, est le manque d'unité : il existe une multitude[1] de logiciels de génétique des populations utilisant différents formats de données, proposant différentes fonctionnalités, et ayant une interopérabilité limitée (Excoffier & Heckel, 2006). Dans cette nébuleuse effrayante (pour le programmeur, du moins), il faut toutefois reconnaître que certains logiciels, STRUCTURE en tête (Pritchard *et al.*, 2000; Falush *et al.*, 2003), sont plus usités que d'autres. Par ailleurs, il est possible de « naviguer » entre une majorité de ces logiciels en n'utilisant qu'un nombre restreint de formats de données.

Une fois les données disponibles dans R et sous une forme adéquate, il est nécessaire de pouvoir manipuler aisément l'information : séparer les données par marqueur, par population, isoler certains génotypes selon des critères complexes, passer du niveau individuel au niveau populationnel, etc. Par ailleurs, puisque l'analyse multivariée n'est souvent qu'un outil parmi d'autres dans les mains du généticien, il est essentiel de rendre accessibles les fonctionnalités proposées par d'autres packages. Ce n'est que lorsque ces besoins seront satisfaits, une fois

[1] le site `http://linkage.rockefeller.edu/soft/` en recence 240, mais ceci inclut également des logiciels de reconstruction phylogénétique

débarrassé des obstacles pratiques à l'analyse des données, que l'on pourra réfléchir plus avant sur l'utilisation de l'analyse multivariée pour extraire de l'information des marqueurs génétiques.

4.2 Le package adegenet pour le logiciel

4.2.1 Présentation

Sous l'impulsion de Daniel Chessel, qui m'a encouragé à « *prendre ce qu'il avait pour la génétique dans ade4 pour en faire un package à part entière* », *adegenet* a donc été créé. Le package tente de répondre aux besoins précédemment énoncés : il définit des classes d'objets pour lesquels l'analyse multivariée peut être directement appliquée, permet une manipulation avancée de l'information, que ce soit au niveau individuel ou populationnel, et favorise l'interopérabilité en proposant l'import de données depuis les formats les plus courants.

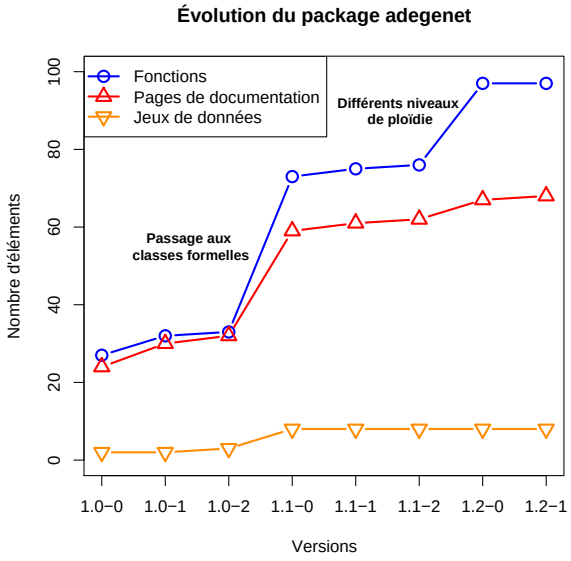


FIG. 4.1: Evolution d'*adegenet* en termes de fonctions, de documentation et de jeux de données

De nouvelles fonctionnalités sont apparues au fur et à mesure du développement (FIG. 4.1), incluant deux innovations principales : l'utilisation de classes formelles (dites « S4 »), plus robustes pour la représentation des données, et la généralisation des fonctionnalités du package à des niveaux de ploïdie quelconques, essentielle pour l'étude de certains végétaux. *adegenet* a également servi de support à l'implémentation de différentes méthodes, dont la *spatial principal component analysis* (sPCA, Jombart *et al.*, 2008), et l'algorithme de Monmonier (Monmonier, 1973), dont la présence dans un package dédié à la génétique peut être justifée par son application en génétique (Manni *et al.*, 2004) bien que sa formulation soit plus générale.

Le package a fait l'objet d'une note d'application dans la revue *Bioinformatics*, présentée dans la prochaine section. La documentation, point essentiel de la diffusion des méthodes, de leur bonne utilisation, et du dialogue avec l'utilisateur, prend la forme d'un manuel détaillant le contenu du package (Annexe 2) et d'un tutoriel (Annexe 3). Pour faciliter les interactions avec les utilisateurs ou des contributeurs potentiels, un site internet dédié à *adegenet* a également

été mis en place (`http://adegenet.r-forge.r-project.org/`), ainsi qu'une liste de diffusion (`adegenet-forum@lists.r-forge.r-project.org`).

4.2.2 Un exemple : le jeu de données `casitas`

A titre de comparaison avec *ade4*, on peut s'intéresser à nouveau au jeu de données `casitas`, mais cette fois-ci dans *adegenet*. La lecture du jeu de données est opérée par `df2genind`, qui reprend l'idée des fonctions `char2genet` et `fuzzygenet` tout en étant beaucoup plus générale. On restaure au passage les noms des populations de *Mus musculus* : *domesticus*, *castaneus*, *musculus* et *casitas*.

```
> levels(species) <- c("dome", "cast", "musc", "casi")
> casi <- df2genind(casitas, pop = species)
> casi


   ######################
   ### Genind object ###
   ######################
- genotypes of individuals -

S4 class:  genind
@call: df2genind(X = casitas, pop = species)

@tab:  74 x 38 matrix of genotypes

@ind.names: vector of  74 individual names
@loc.names: vector of  15 locus names
@loc.nall: number of alleles per locus
@loc.fac: locus factor for the  38 columns of @tab
@all.names: list of  15 components yielding allele names for each locus
@ploidy:  2

Optionnal contents:
@pop:  factor giving the population of each individual
@pop.names:  factor giving the population of each individual

@other: - empty -
```

Avant de procéder à une analyse multivariée de ce jeu de données, on peut faire appel à des outils classiques de génétique des populations implémentés dans d'autres packages (*genetics* et *hierfstat*). On peut tester l'équilibre d'Hardy-Weinberg par population et locus (un test non paramétrique est aussi disponible) :

```
> HWE.test(casi, res.type = "matrix")


         Aat       Amy       Es1         Es2        Es10       Hbb        Gpd1
P1        NA 0.6450944        NA          NA          NA        NA          NA
P2 0.5912833        NA 0.1096466 0.011013755          NA 0.6286169          NA
P3        NA        NA        NA 0.002699796 0.002699796 0.5485062          NA
P4 0.7150007 0.6612572 0.8021296 0.068916418          NA 0.9435582 0.003432669
        Idh1         Mod1      Mod2       Mpi          Np      Pgm1      Pgm2
P1 0.5705442 0.9863810792        NA        NA          NA        NA        NA
P2 0.7639536 0.0009111189        NA 0.6286169 0.03851353 0.3937445        NA
P3        NA 0.8455454553        NA 0.2254423 0.73888268 1.0000000 0.6434837
P4 0.7052516 0.5353423642 0.3764151        NA          NA        NA 0.8709928
   Sod
P1  NA
P2  NA
P3  NA
P4  NA
```

Les NA proviennent de données manquantes ; on peut s'intéresser à un test en particulier :

```
> allHWEtests <- HWE.test(casi)
> allHWEtests$Es2$P3
```

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  tab
X-squared = 9, df = 1, p-value = 0.0027
```

On peut également tester la différenciation génétique entre les 4 populations. On peut mesurer le F_{ST} et tester sa significativité par l'approche de Goudet *et al.* (1996) :

```
> fstat(casi, fstonly = TRUE)

[1] 0.4643312

> Gtest <- gstat.randtest(casi)

> plot(Gtest, main = "Test de la statistique G", cex.main = 1.5)
```

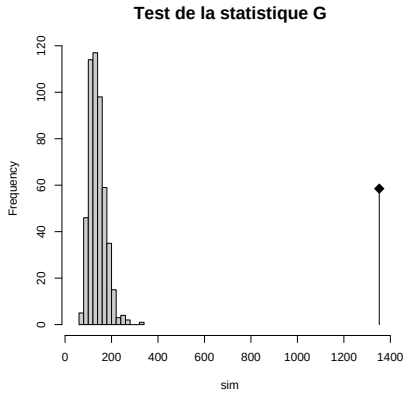


FIG. 4.2: Test de la statistique G (Goudet *et al.*, 1996) des données `casitas`. L'histogramme donne la distribution des valeurs simulées ; la valeur initiale du G est indiquée à droite.

Les populations considérées sont extrêmement différenciées sur le plan génétique (FIG. 4.2). On pourrait procéder à d'autres analyses ou manipulations, mais il est peut-être plus intéressant d'illustrer une analyse multivariée. On constate d'abord que les effectifs par population sont très hétérogènes :

```
> table(casi$pop)

P1 P2 P3 P4
24 11  9 30
```

Or si l'on procède à une analyse au niveau individuel, les populations les plus représentées auront un poids exagéré. On peut définir de nouveaux poids et les ajouter à l'objet `casi` :

```
> row.w <- rep(1/(c(24, 11, 9, 30)), (c(24, 11, 9, 30)))
> row.w <- row.w/sum(row.w)
> casi$other$row.w <- row.w
```

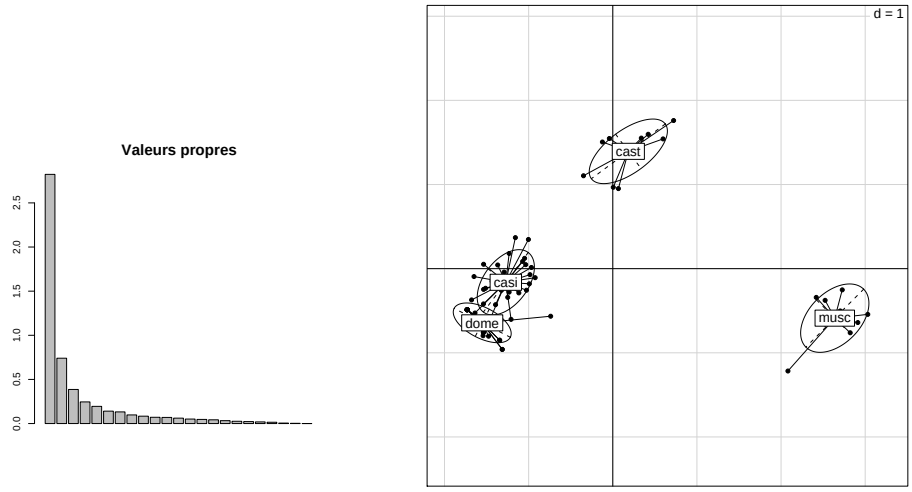
Pour effectuer une analyse multivariée, il faut d'abord se débarasser des données manquantes. Si l'on effectue une analyse en composantes principales (ACP), on peut remplacer les données manquantes par les fréquences alléliques calculées sur l'ensemble des génotypes :

```
> casiNoNA <- na.replace(casi, method = "mean")
```

```
 Replaced 21 missing values


> casiNoNA


   ######################
   ### Genind object ###
   ######################
- genotypes of individuals -

S4 class:  genind
@call: df2genind(X = casitas, pop = species)

@tab:  74 x 38 matrix of genotypes

@ind.names: vector of  74 individual names
@loc.names: vector of  15 locus names
@loc.nall: number of alleles per locus
@loc.fac: locus factor for the  38 columns of @tab
@all.names: list of  15 components yielding allele names for each locus
@ploidy:  2

Optionnal contents:
@pop:  factor giving the population of each individual
@pop.names:  factor giving the population of each individual

@other: a list containing: row.w
```

On peut enfin effectuer une ACP utilisant la nouvelle pondération (Fig. 4.3). Il est à noter que les procédures d'analyse d'*ade4* peuvent directement utiliser un objet `genind` ou `genpop`.

```
> casi.pca <- dudi.pca(casiNoNA, scale = FALSE, scannf = FALSE,
+     nf = 2, row.w = casiNoNA$other$row.w)
> barplot(casi.pca$eig, main = "Valeurs propres", cex.main = 1.5)


> s.class(casi.pca$li, casi$pop, lab = casi$pop.names)
```



(a) Graphe des valeurs propres

(b) Composantes principales 1 et 2 ; les populations sont représentées par des ellipses contenant 95% d'inertie

Fig. 4.3: Analyse en composantes principales pondérée des données `casitas`.

On renverra au tutoriel (Annexe 3) pour de plus amples démonstrations des fonctionnalités du package. A présent que l'histoire du développement d'*adegenet* a été abordée, nous pouvons fournir une description plus globale du package. C'est l'objet de la publication qui suit.

4.3 Article 4 :

adegenet : a R package for the multivariate analysis of genetic markers.

Jombart T.

Article paru en 2008 dans *Bioinformatics* **24** : 1403-1405

BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 11 2008, pages 1403–1405
doi:10.1093/bioinformatics/btn129

Genetics and population analysis

adegenet: a R package for the multivariate analysis of genetic markers

Thibaut Jombart*

Université de Lyon, Université Lyon 1, CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, France

Received on February 12, 2008; revised on April 2, 2008; accepted on April 4, 2008

Advance Access publication April 8, 2008

Associate Editor: Alex Bateman

ABSTRACT

Summary: The package *adegenet* for the R software is dedicated to the multivariate analysis of genetic markers. It extends the *ade4* package of multivariate methods by implementing formal classes and functions to manipulate and analyse genetic markers. Data can be imported from common population genetics software and exported to other software and R packages. *adegenet* also implements standard population genetics tools along with more original approaches for spatial genetics and hybridization.

Availability: Stable version is available from CRAN: http://cran.r-project.org/mirrors.html. Development version is available from *adegenet* website: http://adegenet.r-forge.r-project.org/. Both versions can be installed directly from R. *adegenet* is distributed under the GNU General Public Licence (v.2).

Contact: jombart@biomserv.univ-lyon1.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Genetic markers are now widely used in many fields of population biology, and can be analysed using various approaches. Among these, multivariate methods such as principal component analysis (PCA) are compelled to play an important role because they can summarize the genetic variability without making strong assumptions about an evolution model: they do not rely on Hardy–Weinberg equilibrium, nor do they suppose the absence of linkage disequilibrium. This is especially valuable when no or very little information is known about the system under study, as is frequent in landscape genetics (Manel *et al*., 2003). Recently, multivariate methods have proven useful to assess the consensus genetic structuring among a set of genetic markers (Laloë *et al*., 2007), as well as to investigate the spatial pattern of the genetic variability (Jombart *et al*., in press). However, multivariate methods currently available in population genetics software are very restricted, despite the fairly large number of these programs (Excoffier and Heckel, 2006). An exception to this is the R software (R Development Core Team, 2008) which contains both packages devoted to multivariate methods like *ade4* (Chessel *et al*., 2004; Dray *et al*., 2007), and packages dedicated to the analysis of genetic markers (http://cran.r-project.org/web/views/Genetics.html). Currently there are no bridges between multivariate analysis packages and genetic marker packages, and genetic markers data cannot be readily analysed using multivariate approaches. The purpose of *adegenet* is to build this connection. This package aims at extending the *ade4* package so that genetic markers can be analysed using multivariate methods. This is achieved by defining new classes of objects to represent genetic markers, and providing functions to import, export and manipulate these objects. Moreover, *adegenet* also implements some usual population genetics methods, as well as more original tools for spatial genetics and data simulation. This article presents an overview of these functionalities.

*To whom correspondence should be addressed.

2 CONTENT

2.1 Data representation

Basic genetic markers data are genotypes obtained for a set of markers, each allele being coded by a character string (Warnes, 2003). In order to use statistical methods, such information cannot be used directly, and needs to be recoded numerically into a matrix of allelic frequencies. In *adegenet*, allelic frequencies of genotypes are stored inside objects of the class `genind`, which is the basic class of the package. In addition to allelic frequencies stored in a `@tab` component (the '@' designs a slot), a `genind` object stores other useful information whose description is provided by the R command `class?genind`. This class `genind` was designed to allow flexibility (it can virtually store any relevant information about genotypes) but also to be robust. As `genind` is a formal class (or 'S4 class' in R language), it is naturally robust: the content of an object is checked for validity when it is created and each time it is modified, which considerably limits the risks of having wrong or missing items in it. Moreover, `genind` internally uses generic labels for markers, alleles and genotypes, so that missing or redundant user-defined labels cannot originate an error in further analyses. Whenever the study involves groups of genotypes (say, 'populations') rather than genotypes, `genpop` objects are used. This formal class is very similar to `genind`, except that `@tab` contains counts of alleles per population instead of allelic frequencies of genotypes. Objects of both classes can be analysed by multivariate methods using

© The Author 2008. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

the @tab slot as input. Main available functions to import to, export from, manipulate and analyse genind and genpop objects are listed in Supplementary Material.

2.2 Functionalities

Great attention was devoted to developing input/output functions, because interoperability of data is crucial to facilitate data analysis. Until now, data could only be imported into R from FSTAT (Goudet, 2002) using the *hierfstat* package (Goudet, 2005). Currently, *adegenet* can read files from the software GENETIX (Belkhir *et al*., 1996–2004), STRUCTURE (Pritchard *et al*., 2000), FSTAT (Goudet, 2002), and Genepop (Raymond and Rousset, 1995), which are among the most common data formats in population genetics software (Excoffier and Heckel, 2006). Data can also be read inside R from a data.frame of genotypes coded by character strings (using df2genind), and exported back (genind2df). Outputs are possible from genind to the R packages *genetics* (Warnes, 2003) and *hierfstat* (Goudet, 2005), using genind2genotype and genind2hierfstat, respectively. Note that the data representation in the *genetics* package was intended to be consensual, and is used by many other R packages. Moreover, the output of genind2df is customisable and can be designed to fit usual formats like those of GENETIX or STRUCTURE.

To perform analyses at a population level, a genind object can be translated into a genpop object using genind2genpop. Other data manipulations include splitting information by marker (seploc) or by population (seppop), computing allelic frequencies for populations (makefreq), or subsetting genotypes, populations or alleles according to a given criterion. Basic methods are implemented such as the Hardy–Weinberg equilibrium test for all combinations of populations and markers (HWE.test.genind), a matrix of *P*-values allowing a quick overview of the results. Observed and expected heterozygosity, number of alleles by marker or population, sample sizes and other miscellaneous information are provided by summary functions. Missing data can be replaced in different ways—which is required by most statistical methods—using na.replace. Several genetic distances among populations can be computed using dist.genpop. Goudet's *G* statistic (Goudet *et al*., 1996) can be tested by a Monte–Carlo procedure to assess the hierarchical structuring of a set of genotypes (gstat.randtest).

The last goal of *adegenet* is to implement more original methods, either by extending existing ones, or by proposing new methods. Hybridization between individuals from two genind objects can be simulated using hybridize, which can be useful to evaluate the power of methods based on genetic differentiation. Monmonier's algorithm (Monmonier, 1973), which is used to infer genetic boundaries among geo-referenced genotypes (Manni *et al*., 2004), has been extended to include different degrees of connectivity among genotypes (monmonier) and implemented with an optimization function (optimize.monmonier). Finally, recently developed methods to investigate spatial patterns of the genetic variability (Jombart *et al*., in press) are also part of the package (functions spca, global.rtest and local.rtest).

3 EXAMPLE

This example illustrates how a theoretical hybrid population would appear on a typology provided by a multivariate method. First, we load the required packages, and the dataset microbov containing 30 microsatellite markers for 704 genotypes of 15 cattle breeds, described in Laloë *et al*. (2007).

```
> library(adegenet)
> library(ade4)
> data(microbov)
```

To simulate a hybrid population, two parent breeds (Salers and Zebu) are isolated:

```
> temp <- seppop(microbov)
> salers <- temp$Salers
> zebu <- temp$Zebu
```

The hybrid population ('Zebler') is obtained using the hybridize function (with n = 40 simulated genotypes). All data are pooled in a new object newbov:

```
> zebler <- hybridize(salers, zebu,
+ pop = "Zebler", n = 40)
> newbov <- repool(microbov, zebler)
```

Now we seek a typology displaying the diversity between breeds. For this, the inter-class PCA (Dolédec *et al*., 1987) is appropriate: this modification of PCA maximizes the variance between populations (here, breeds), instead of the total variance. Missing data are replaced (na.replace) before performing a centred PCA (dudi.pca) and an inter-class PCA (between):

```
> newbov <- na.replace(newbov, method =
"mean")
> pca1 <- dudi.pca(newbov$tab, center = TRUE,
+ scale = FALSE, scannf = FALSE)
> bet1 <- between(pca1, fac = newbov$pop,
+ scannf = FALSE, nf = 3)
```

The resulting typology (Fig. 1) is obtained by:

```
> s.class(bet1$ls, fac = newbov$pop,
+ clab = 1.2, lab = newbov$pop.names)
> add.scatter.eig(bet1$eig, nf = 2, xax = 1,
+ yax = 2, pos = "bottomright", csub = 1.2)
```

The first principal axis of the analysis (Fig. 1) differentiates African and French breeds, while the second axis expresses the genetic variability between African breeds. Interestingly enough, the simulated hybrid population (Zebler) appears between its parent populations (Salers and Zebu).

4 CONCLUSION

The first contribution of the R package *adegenet* is to implement classes and functions to facilitate the multivariate analysis of genetic markers. This led to define new formal classes for genotypes (genind) or groups of genotypes (genpop), which can be used as input to multivariate methods proposed in the R software. Several functions are also implemented to manipulate and analyse these objects, including recent development in spatial genetics and data simulation. By assuring a good interoperability of data, *adegenet* contributes to making the R software a unifying platform for the analysis of genetic markers.

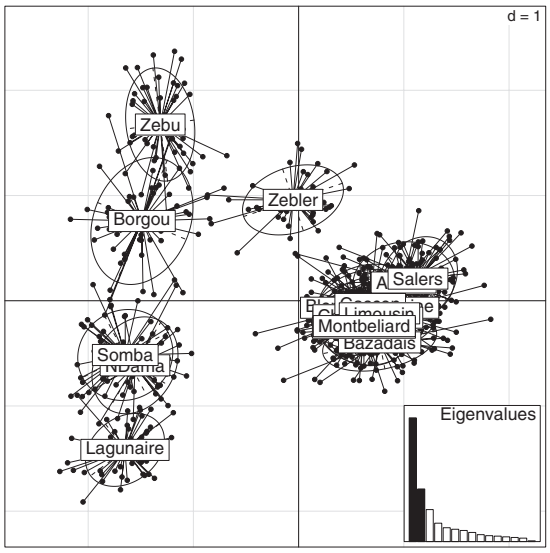


Fig. 1. Typology of cattle breeds (object `newbov`) obtained by inter-class PCA. Eigenvalues corresponding to the represented components are filled in black. Points represent genotypes; breeds are labelled inside their 95% inertia ellipses.

ACKNOWLEDGEMENTS

The author is grateful to R-Forge for hosting *adegenet*, to P. Sólymos for his contribution and to A.-B. Dufour, S. Devillard, D. Laloë and D. Pontier for their constructive comments.

Conflict of Interest: none declared.

REFERENCES

Belkhir,K. *et al.* (1996–2004) GENETIX 4.05, logiciel sous Windows TM pour la génétique des populations. Laboratoire Genome, Populations, Interactions, CNRS UMR 5000, Université de Montpellier II, Montpellier, France. URL: http://www.genetix.univmontp2.fr/genetix/intro.htm.

Chessel,D. *et al.* (2004) The ade4 package-I-one-table methods. *R News*, **4**, 5–10.

Dolédec,S. and Chessel (1987) Rythmes saisonniers et composantes stationnelles en milieu aquatique. *Acta Oecologica, Oecologia Generalis*, **8**, 403–426.

Dray,S. *et al.* (2007) The ade4 package – II: two-table and *K*-table methods. *R News*, **7**, 47–54.

Excoffier,L. and Heckel,G. (2006) Computer programs for population genetics data analysis: a survival guide. *Nat. Rev. Genet.*, **7**, 745–758.

Goudet,J. (2002) Fstat 2.9.3.2. URL: http://www2.unil.ch/popgen/softwares/fstat.htm.

Goudet,J. (2005) Hierfstat, a package for R to compute and test hierarchical F-statistics. *Mol. Ecol. Notes*, **5**, 184–186.

Goudet,J. *et al.* (1996) Testing differentiation in diploid populations. *Genetics*, **144**, 1933–1940.

Jombart, T. *et al.* (in press) Revealing cryptic spatial patterns in genetic variability by a new multivariate method. *Heredity*.

Laloë,D. *et al.* (2007) Consensus genetic structuring and typological value of markers using multiple co-inertia analysis. *Genet. Sel. Evol.*, **39**, 545–567.

Manel,S. *et al.* (2003) Landscape genetics: combining landscape ecology and population genetics. *Trends in Ecol. Evol.*, **18**, 189–197.

Manni,F. *et al.* (2004) Geographic patterns of (genetic, morphologic, linguistic) variation: how barriers can be detected by "Monmonier's algorithm". *Hum. Biol.*, **76**, 173–190.

Monmonier,M. (1973) Maximum-difference barriers: an alternative numerical regionalization method. *Geogr. Anal.*, **3**, 245–261.

Pritchard,J. *et al.* (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

R Development Core Team (2008) *R: A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing, Vienna, Austria.

Raymond,M. and Rousset,F. (1995) Genepop (version 1.2): population genetics software for exact tests and ecumenicism. Journal of *Heredity*, **86**, 248–249.

Warnes, G. (2003) The genetics package. *R News*, **3**, 9–13.

4.4 Perspectives

Le package *adegenet* n'est certainement pas à considérer comme un produit entièrement fini, immuable et complet. S'il comble en bonne partie les besoins qui ont été énoncés plus haut, nombre d'améliorations, de nouvelles fonctionnalités, et espérons-le, peu de correctifs restent à apporter. Il serait sans doute plus juste de considérer ce package comme un objet évoluant au gré des interactions avec les utilisateurs, et sans doute plus tard avec l'apparition de nouvelles données et de nouvelles problématiques.

adegenet s'insère dans une démarche visant à promouvoir l'analyse de données génétiques dans le logiciel R. Cette démarche est motivée par le fait que R mette à disposition un ensemble considérable d'outils statististiques, de données et de documentation, tout en étant un logiciel libre, c'est-à-dire gratuit et dont le code source est pleinement accessible. Ceci, associé au fait que R passe par l'édition de scripts et facilite la diffusion des données, permet une reproductibilité parfaite des analyses, qui fait trop souvent défaut dans la littérature (Jombart *et al.*, in revision). La volonté de faire de R une plateforme pour l'analyse de marqueurs génétiques a été clairement affichée par les créateurs du package *genetics* (Warnes, 2003), qui ont abandonné leur projet initial au profit du projet *R-genetics*, plus vaste et plus fédérateur, qui vise à proposer un grand ensemble d'outils pour l'analyse des marqueurs moléculaires (`http://rgenetics.org`). Cette initiative, assez similaire au projet *phylobase* pour les méthodes comparatives (Bolker *et al.*, 2007), devrait à terme constituer le fer de lance de l'analyse de données génotypiques dans R. On pourrait critiquer le fait qu'*adegenet* soit resté en marge de ce projet. En réalité, le rythme de développement de ces deux projets a été très différent, et *adegenet* est une démarche entièrement personnelle là où *Rgenetics* est un effort collectif. Néanmoins, les prochains développements d'*adegenet* viseront clairement à établir des liens avec *Rgenetics*, dès lors que la représentation des données dans ce projet aura été stabilisée. La prochaine étape consistera en un effort de documentation, pour illustrer comment l'ensemble des outils mis à disposition permettront d'extraire de l'information biologique des marqueurs génétiques. Dans cette optique, la mise à disposition de la liste de discussion R-sig-genetics (`https://stat.ethz.ch/mailman/listinfo/r-sig-genetics`) devrait favoriser l'échange d'informations et d'idées entre utilisateurs et développeurs, et affermir le support du dialogue interdisciplinaire requis par l'analyse de données.

Chapitre 5

Structures spatiales à plusieurs échelles

Sommaire

5.1 Introduction

Lorsque l'on s'intéresse à la structuration spatiale d'un ensemble de variables, une question venant immédiatement est celle de l'échelle à laquelle les processus engendrant cette structuration agissent. Nous avons développé une nouvelle méthode, la *multi-scale pattern analysis* (MSPA, Jombart *et al.*, sous presse), pour aborder cette question. A la différence de la sPCA présentée au chapitre 3, la MSPA n'a pourtant pas été proposée dans le cadre génétique, mais en écologie. La raison en est culturelle : bien qu'intéressante dans le contexte génétique, la question de l'identification des principales échelles de la structuration spatiale est avant tout écologique. En outre, les outils statistiques sur lesquels repose la MSPA sont de mieux en mieux connus en écologie, et parfaitement inconnus en génétique. Nous avons donc choisi une solution de facilité pour le développement de cette méthode : la proposer à un public averti pour l'utiliser par la suite dans un cadre plus large, et plus particulièrement pour étudier les échelles auxquelles la variabilité génétique est spatialement structurée.

Ce chapitre présente la MSPA après avoir défini le cadre biologique et méthodologique dans lequel la méthode a émergé. Après l'article proprement dit, nous insistons sur les liens méthodologiques existants entre la MSPA et les tests globaux et locaux proposés avec la sPCA (Jombart *et al.*, 2008). Enfin, nous illustrons l'intérêt de la méthode en génétique à travers une application à des données réelles.

5.1.1 Une question écologique

L'écologie s'intéresse à une multitude de phénomènes biologiques, des interactions biotiques au sein d'une population aux effets du changement climatique sur la distribution des espèces à l'échelle du globe. C'est donc par essence une science multi-échelle (Turner *et al.*, 1989). Le terme d'échelle est ici utilisé en tant que paradigme recouvrant un ensemble de notions par ailleurs définies dans la littérature (Turner *et al.*, 1989; Dungan *et al.*, 2002). L'identification de structures biologiques et les inférences qui peuvent en découler sont donc intimement liées à l'échelle à laquelle l'observateur se place, ce qui est résumé par Wiens (1989) :

> *Acts in what Hutchinson (1965) has called the 'ecological theatre' are played out on various scales of space and time. To understand the drama, we must view it on the appropriate scale.*

Levin (1992) poursuit en notant :

> [...] *no description of the variability and predictability of the environment makes sense without reference to the particular range of scales that are relevant to the organisms or processes being examined.*

On comprend alors que l'identification de structures, et de structures spatiales en particulier, bénéficierait largement de plans expérimentaux explicitement multi-échelles (Turner *et al.*, 1989; Wiens, 1989).

Mais pour des raisons évidentes, les plans expérimentaux couvrant réellement plusieurs échelles sont sans doute rares, et les études dites "multi-échelle" correspondent souvent à

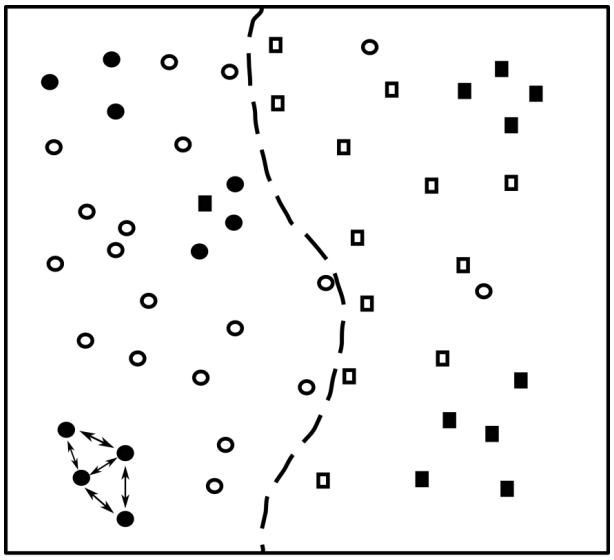


FIG. 5.1: Exemple de structuration à différentes échelles dans une distribution spatiale d'objets. Les carrés sont séparés des ronds en deux patches de grande taille par une frontière partiellement perméable (ligne en pointillés). Cette structure engendrée par le milieu relève de la **dépendance spatiale** *sensu* Wagner & Fortin (2005). Les symboles noirs sont réunis en patches de plus petite taille par un phénomème d'attraction locale entre objets. Il s'agit d'**autocorrélation spatiale** *sensu* Wagner & Fortin (2005). Ici, deux tailles de patches indiquent l'existence de processus sous-jacents ayant lieu à des échelles différentes.

l'identification de différents patrons spatiaux au sein d'un plan d'échantillonnage mené à une seule échelle (par exemple, Brind'Amour *et al.*, 2005). Il est néanmoins possible, au sein d'une distribution d'objets biologiques, de reconnaitre par exemple des patches de tailles différentes, et de supposer que les processus qui les engendrent opèrent à différentes échelles (FIG. 5.1). Au sein des structures spatiales observables dans une distribution d'espèces, Wagner & Fortin (2005) distinguent la **dépendance spatiale**, induite par des structures environnementales, et l'**autocorrélation spatiale** véritable, provenant d'interactions entre organismes (FIG. 5.1). La dépendance spatiale est supposée induire des structures à large échelle, tandis que l'autocorrélation spatiale est supposée créer des structures à des échelles plus fines (Legendre, 1993).

Les méthodes statistiques utilisées pour étudier les échelles auxquelles des structures spatiales sont observées dans une distribution d'objets sont nombreuses, et ont fait l'objet de plusieurs revues (Legendre & Fortin, 1989; Dale *et al.*, 2002; Liebhold & Gurevitch, 2002; Perry *et al.*, 2002; Ollier, 2004). L'approche qui nous concerne ici est celle des vecteurs propres associés à l'indice de Moran (Moran, 1948, 1950; Cliff & Ord, 1981), d'abord utilisés en géographie (Griffith, 1996, 2000) puis importés en écologie (Dray *et al.*, 2006; Griffith & Peres-Neto, 2006), où cette approche contient celle des coordonnées principales de matrice de voisinage (Borcard & Legendre, 2002; Borcard *et al.*, 2004) comme un cas particulier (Dray *et al.*, 2006). L'intérêt de ces vecteurs est qu'ils sont capables de modéliser, étant donnée une mesure des proximités spatiales entre objets, différentes structures observables dont des patches de différentes tailles et des différences locales fortes. Chacun de ces vecteurs étant associé à une valeur d'autocorrélation spatiale décroissante, les notions d'autocorrélation et d'échelle tendent à se confondre. Cet amalgame est à l'origine de

la définition de structures **globales** et **locales** proposée par Thioulouse *et al.* (1995) et reprise dans Jombart *et al.* (2008), correspondant respectivement à une autocorrélation positive et négative. Dans le contexte des vecteurs propres de Moran, l'idée d'échelle est donc à prendre avec précaution : s'il est peut-être abusif de dire que chaque vecteur modélise une échelle différente, il est toutefois légitime de penser que ces vecteurs peuvent modéliser des structures spatiales causées par des processus ayant bien lieu à différentes échelles. Ces précautions étant prises, on peut s'intéresser à la nature mathématique de ces vecteurs, avant d'en montrer l'usage qui en est fait dans la MSPA.

5.1.2 Vecteurs propres de Moran

L'indice de Moran défini au chapitre 3 (Eqn. 3.1) dans le contexte de sa généralisation multivariée par Wartenberg (1985) est une version simplifée, où la variable concernée est centrée et réduite. Néanmoins, une définition plus générale du I de Moran pour un vecteur $\mathbf{x}_0$ $(\frac{1}{n}\mathbf{I}_n)$-centré de $\mathbb{R}^n$ est (Cliff & Ord, 1981, p.119) :

$$I_{\mathbf{W}}(\mathbf{x}_0) = \frac{n}{\mathbf{1}_n^T\mathbf{W}\mathbf{1}_n}\frac{\mathbf{x}_0^T\mathbf{W}\mathbf{x}_0}{\mathbf{x}_0^T\mathbf{x}_0} \tag{5.1}$$

où $\mathbf{W}$ est une matrice de pondération de voisinage symétrique d'ordre n. Notons qu'une matrice de pondération de voisinage non symétrique $\mathbf{V}$ peut être rendue symétrique en prenant (Cliff & Ord, 1981, p.18 ; Tiefelsdorf & Boots, 1996) :

$$\mathbf{W} = \frac{1}{2}(\mathbf{V} + \mathbf{V}^T) \tag{5.2}$$

ce qui n'altère pas la valeur du $I_{\mathbf{W}}(\mathbf{x}_0)$ car pour tout $\mathbf{a} \in \mathbb{R}^n$:

$$\begin{aligned}
\mathbf{a}^T\mathbf{W}\mathbf{a} &= \frac{1}{2}\mathbf{a}^T(\mathbf{V} + \mathbf{V}^T)\mathbf{a} \\
&= \frac{1}{2}\mathbf{a}^T\mathbf{V}\mathbf{a} + \frac{1}{2}\mathbf{a}^T\mathbf{V}^T\mathbf{a} \\
\mathbf{a}^T\mathbf{W}\mathbf{a} &= \mathbf{a}^T\mathbf{V}\mathbf{a}
\end{aligned} \tag{5.3}$$

Il est également possible de relaxer l'hypothèse du centrage de la variable, en remarquant que :

$$\begin{aligned}
\mathbf{x}_0 &= \mathbf{x} - \mathbf{1}_n\bar{x} \\
&= \mathbf{x} - \mathbf{1}_n(\mathbf{1}_n^T(\frac{1}{n}\mathbf{I}_n)\mathbf{x}) \\
&= (\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T)\mathbf{x} \\
\mathbf{x}_0 &= \mathbf{P}\mathbf{x}
\end{aligned} \tag{5.4}$$

où $\bar{x}$ est la moyenne de $\mathbf{x}$ et où $\mathbf{P} = (\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T)$ est le projecteur $(\frac{1}{n}\mathbf{I}_n)$-orthogonal sur $\mathbf{1}_n^{\perp}$. On note que ce projecteur est par définition $(\frac{1}{n}\mathbf{I}_n)$-symétrique et idempotent ($\mathbf{P}\mathbf{P} = \mathbf{P}$).

En utilisant (Eqn. 5.4) dans (Eqn. 5.1), et en posant $\mathbf{W}' = \frac{n}{\mathbf{1}_n^T\mathbf{W}\mathbf{1}_n}\mathbf{W}$ (c'est-à-dire en normalisant globalement $\mathbf{W}$), on obtient la formulation largement répandue (de Jong *et al.*,

1984; Tiefelsdorf & Boots, 1995; Griffith, 1996) :

$$I_{\mathbf{W}}(\mathbf{x}) = \frac{\mathbf{x}^T\mathbf{P}\mathbf{W}'\mathbf{P}\mathbf{x}}{\mathbf{x}^T\mathbf{P}\mathbf{x}} \tag{5.5}$$

La question abordée par de Jong *et al.* (1984) et reprise par Tiefelsdorf & Boots (1995) consiste à trouver, pour une pondération de voisinage donnée $\mathbf{W}$, des variables centrées, normées et non corrélées entre elles, dont l'autocorrélation décroit successivement, de la valeur maximale à la valeur minimale. On peut maintenant reformuler ce problème comme la recherche de vecteurs de $\mathbb{R}^n$, $(\frac{1}{n}\mathbf{I}_n)$-centrés et orthogonaux, trouvant les extrêmes de (Eqn. 5.5). La solution à ce problème est donnée par la diagonalisation de $\mathbf{P}\mathbf{W}'\mathbf{P}$, fournissant une base :

$$\mathbf{U} = [\mathbf{u}_1|\mathbf{u}_2|\cdots|\mathbf{u}_{n-1}] \tag{5.6}$$

de vecteurs propres $\mathbf{P}$-orthonormés ($\mathbf{U}^T\mathbf{P}\mathbf{U} = \mathbf{I}_{n-1}$) (Harville, 1997, pp.533-534 et de Jong *et al.*, 1984; Tiefelsdorf & Boots, 1995) .

On peut montrer que les $\mathbf{u}_j$ $(j = 1, \ldots, n-1)$ sont de moyennes nulles en remarquant que $\mathbf{1}_n$ est un vecteur propre de $\mathbf{P}\mathbf{W}'\mathbf{P}$ associé à la valeur propre nulle :

$$\begin{aligned}
\mathbf{P}\mathbf{W}'\mathbf{P}\mathbf{1}_n &= \mathbf{P}\mathbf{W}'(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T)\mathbf{1}_n \\
&= \mathbf{P}\mathbf{W}'(\mathbf{1}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\mathbf{1}_n) \\
&= \mathbf{P}\mathbf{W}'(\mathbf{1}_n - \mathbf{1}_n) \\
\mathbf{P}\mathbf{W}'\mathbf{P}\mathbf{1}_n &= 0
\end{aligned} \tag{5.7}$$

Donc, les vecteurs $\mathbf{u}_j$ sont orthogonaux à $\mathbf{1}_n$, et donc de moyennes $\frac{1}{n}\mathbf{u}_j^T\mathbf{1}_n$ nulles.

On peut également montrer que les $\mathbf{u}_j$ $(j = 1, \ldots, n-1)$ sont orthonormés pour la métrique canonique :

$$\begin{aligned}
\mathbf{U}^T\mathbf{P}\mathbf{U} &= \mathbf{U}^T(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T)\mathbf{U} \\
&= \mathbf{U}^T(\mathbf{U} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\mathbf{U}) \\
&= \mathbf{U}^T(\mathbf{U} - \mathbf{1}_n \underbrace{(\frac{1}{n}\mathbf{1}_n^T\mathbf{U})}_{\mathbf{0}_{n-1}^T}) \\
\mathbf{U}^T\mathbf{P}\mathbf{U} &= \mathbf{U}^T\mathbf{U} = 0
\end{aligned} \tag{5.8}$$

Donc, les vecteurs $\mathbf{u}_j$ ont bien une variance unitaire et sont non corrélés entre eux. Les $\mathbf{u}_j$ sont généralement nommés **vecteurs propres de Moran**, mais Dray *et al.* (2006) ont montré que (Eqn. 5.5) englobe également les coordonnées principales de matrice de voisinage (CPMV, ou PCNM en anglais, Borcard & Legendre, 2002; Legendre & Borcard, 2003; Borcard *et al.*, 2004), auquel cas la matrice $\mathbf{W}$ est une matrice de proximités géographiques tronquée à une valeur seuil arbitraire.

On retient des vecteurs propres de Moran qu'ils possèdent des caractéristiques particulièrement intéressantes. Ils modélisent différentes structures spatiales observables du point de vue de l'indice de Moran pour une pondération de voisinage donnée. Ils peuvent donc être utilisés pour faire du **filtrage spatial**, c'est-à-dire être introduits comme covariables dans un modèle linéaire pour éliminer l'autocorrélation spatiale des résidus (Griffith, 2000; Tiefelsdorf & Griffith, 2007). Dans cette optique, leur orthogonalité est évidemment une caractéristique précieuse. On peut au contraire les utiliser comme variables explicatives d'intérêt pour quantifier la structuration spatiale d'une variable à différentes échelles, et c'est l'approche que nous adoptons dans la *multi-scale pattern analysis*, qui fait l'objet de la publication suivante.

5.2 Article 5 : Finding essential scales of spatial variation in ecological data : a multivariate approach

Manuscrit accepté le 26 août 2008 dans la revue *Ecography*.

Running head: Finding scales of spatial patterns

Finding essential scales of spatial variation in ecological data: a multivariate approach

Thibaut Jombart[1], Stéphane Dray[1], Anne-Béatrice Dufour[1]

[1] Université de Lyon, F-69000, Lyon ; Université Lyon 1 ; CNRS, UMR5558, Laboratoire de Biométrie et Biologie Evolutive, F-69622, Villeurbanne, France.

Corresponding author's email: jombart@biomserv.univ-lyon1.fr

ABSTRACT

The identification of spatial structures is a key step in understanding the ecological processes structuring the distribution of organisms. Spatial patterns in species distributions result from a combination of several processes occuring at different scales: identifying these scales is thus a crucial issue. Recent studies have proposed a new family of spatial predictors (PCNM: principal coordinates of neighbours matrices; MEMs: Moran's eigenvectors maps) that allow for modelling of spatial variation on different scales. To assess the multi-scale spatial patterns in multivariate data, these variables are often used as predictors in constrained ordination methods. However, the selection of the appropriate spatial predictors is still troublesome, and the identification of the main scales of spatial variation remains an open question. This paper presents a new statistical tool to tackle this issue: the *multi-scale pattern analysis* (MSPA). This ordination method uses MEMs to decompose ecological variability into several spatial scales and then summarizes this decomposition using graphical representations. A canonical form of MSPA can also be used to assess the spatial scales of the species-environment relationships. MSPA is compared to constrained ordination using simulated data, and illustrated using the famous oribatid mites dataset. The method is implemented in the free software R.

Keywords: autocorrelation, MEM, Moran's I, multivariate, ordination, PCNM, principal component analysis, scale, spatial patterns

INTRODUCTION

The study of spatial patterns has been and still is a fecund paradigm in ecology (Legendre 1993). Indeed, the identification of spatial structures is a key step toward an improved understanding of the ecological processes structuring the distribution of organisms (Legendre and Fortin 1989). The distribution of species is influenced by several environmental variables, some of which are inherently spatially structured; the resulting spatial patterns observed in species communities are referred to as induced *spatial dependence* (Legendre 1993, Wagner and Fortin 2005). Such patterns are mainly expected to occur on broad scales (Wiens 1989, Legendre 1993). In contrast, contagious biotic processes like dispersal, mating or competition can give rise to *spatial autocorrelation* which likely results in intermediate to small-scale spatial structures (Wiens 1989, Legendre 1993, Wagner and Fortin 2005). In fact, any empirically observed spatial pattern can be a combination of several processes occuring on different scales, the identification of which is a crucial issue in ecological studies (Wiens 1989, Menge and Olson 1990, Dungan et al. 2002).

The study of spatial processes led to several methodological innovations, for which the modelling of spatial patterns retained particular attention (Legendre and Fortin 1989, Legendre 1993, Legendre and Legendre 1998). Several approaches have been proposed to introduce space into ecological models in order to identify spatial patterns or on the contrary to remove the effects of spatial structures. Trend surface analysis (Legendre and Fortin 1989, Borcard et al. 1992) has used polynomial expressions of spatial coordinates to create spatial predictors. However, this approach suffered from certain flaws, the strongest being the non-independence of the created variables and difficulties in interpreting high-degree terms of the polynomials. Another limitation was that trend surface analysis could not correctly model fine-scale patterns (Borcard and Legendre 2002).

Borcard and Legendre (2002) proposed the use of principal coordinates of neighbours matrices (PCNM) as spatial predictors. PCNM are uncorrelated variables dissecting the spatial variability into different but complementary scales. The fact that PCNM are uncorrelated variables makes them ideal candidates as predictors in linear models because they are not subject to multicollinearity troubles (Borcard and Legendre 2002, Dray et al. 2006, Griffith and Peres-Neto 2006). Recent theoretical works (Dray et al. 2006, Griffith and Peres-Neto 2006) demonstrated that PCNM are in fact particular cases of eigenvectors of a spatial weighting matrix (Griffith 1996; 2000), called Moran's eigenvectors maps (MEMs) by Dray et al. (2006). In geography, these eigenvectors are used for spatial filtering purposes, *i.e.* to remove spatial autocorrelation from residuals of a model so that standard statistical tools can be used (Getis and Griffith 2002). In ecology, PCNM and MEMs are used to dissect the spatial patterns of the ecological variability into separate scales. These variables can be used as spatial predictors in multiple regressions (Brind'Amour et al. 2005). When dealing with multivariate data, PCNM and MEMs can also serve as spatial predictors in constrained

ordinations (Borcard et al. 2004) such as redundancy analysis (RDA, Rao 1964) or canonical correspondence analysis (CCA, Ter Braak 1986). It should be noted that for n sites, PCNM and MEM approaches respectively produce $2n/3$ and $(n-1)$ spatial predictors. Hence, variables must be selected in order to avoid overfitting in statistical models. Due to differences of objectives in geography and ecology, the selection procedures are different. In geography, Tiefelsdorf and Griffith (2007) developed a method to select spatial predictors that minimize the autocorrelation in residuals. However, ecologists favored the classical forward selection that optimizes the fit of the model. Unfortunately, this approach often overestimates the number of retained predictors (Dray et al. 2006): methods based on forward selection thus tend to identify more structuring scales than there actually are. Therefore, the question remains as how to pin down the main scales of spatial variation in ecological data.

In this paper, we propose a new exploratory approach to tackle this issue. This method, called *multi-scale pattern analysis* (MSPA), describes the correlation structure among a set of ecological variables and all possible scales of variation modeled by MEM, and can be applied to both quantitative and qualitative variables. The MSPA yields graphical representations allowing a visual assessment of the essential scales of spatial variation along with the variables exhibiting the strongest spatial patterns. After describing the method, we compare the results of MSPA and RDA using simulated data. An application of MSPA is also provided using the famous oribatid mites dataset (Borcard et al. 1992, Borcard and Legendre 1994) which raises the question about the scales of variation in species-environment relationships. For this purpose, we show how MSPA can be conducted after a multivariate regression to obtain a canonical MSPA. MSPA is implemented in the free software R (R Development Core Team 2008).

STATISTICAL METHOD

Our approach involves two steps: i) the decomposition of a set of ecological variables in terms of spatial scales and ii) the analysis of this decomposition. The whole procedure is summarized in Figure 1, and computation details are provided in Appendix A.

Decomposing variables in terms of spatial scales

As advocated in recent papers (Dray et al. 2006, Griffith and Peres-Neto 2006), MEMs can efficiently decompose the spatial variability of a variable into a set of different scales. The spatial distribution of n sites is first modelled by a connection network (Legendre and Legendre 1998, pp. 752-756). The doubly centred spatial weighting matrix of this network is then diagonalized as described in Dray et al. (2006) to obtain $(n-1)$ centred, scaled, and uncorrelated MEMs ($\mathbf{u}_j$, $j = 1, \ldots, n-1$). The vectors $\mathbf{u}_j$ model spatial patterns at different scales, from the largest scale

($\mathbf{u}_1$) to the finest scale ($\mathbf{u}_{n-1}$). They can be used to decompose the variability of quantitative or qualitative variables in terms of scales.

While a quantitative variable can be used as it is, a qualitative variable needs to be transformed first. Each level of a qualitative variable $\mathbf{x}$ is recoded by a dummy variable, *i.e.* a vector whose components are 1 where the level is observed and 0 otherwise. For instance, a variable $\mathbf{x} = [a, a, b, b, c]$ will be recoded by three vectors: $\mathbf{x}_a = [1, 1, 0, 0, 0]$, $\mathbf{x}_b = [0, 0, 1, 1, 0]$, and $\mathbf{x}_c = [0, 0, 0, 0, 1]$. This transformation allows levels of a factor to be treated like quantitative variables. Henceforward, we shall consider a centred and scaled variable $\mathbf{y}$ which can correspond to a quantitative variable or to the level of a factor. In some cases, one may be more interested in studying the variability of $\mathbf{y}$ which can be predicted by a set of explanatory variables, rather than in $\mathbf{y}$ itself. For instance, one can seek multi-scale spatial patterns in the environmental component of a species variability. In such a case, we can proceed like in other canonical approaches (Rao 1964, Ter Braak 1986): $\mathbf{y}$ can be submitted to a linear regression onto a set of variables to obtain a predicted vector $\hat{\mathbf{y}}$. The *canonical MSPA* is then a usual MSPA in which vectors of predictions $\hat{\mathbf{y}}$ replace original observations $\mathbf{y}$ (see Appendix A). Similarly, one could perform a *partial canonical MSPA* to study multi-scale spatial patterns in a variable after removing the effects of a set of covariates. For instance, we could investigate the main scales of spatial variation in species varibility that is not explained by a set of environmental variables. This can be achieved using the same approach as in canonical MSPA, but using the residuals of linear regression onto covariates ($\mathbf{y} - \hat{\mathbf{y}}$) rather than the predicted values ($\hat{\mathbf{y}}$) in further computations.

The variability of $\mathbf{y}$ is decomposed through linear regressions using the set of vectors $\mathbf{u}_j$ as explanatory variables. It is worth recalling that as all $\mathbf{u}_j$ are uncorrelated, they explain different and complementary components of variation and are not subject to multicollinearity problems. Moreover, as there are $(n-1)$ regressors, this model explains 100% of the variance of $\mathbf{y}$. Determination coefficients (R^2) are used to measure the strength of association between $\mathbf{y}$ and the different MEMs: they represent the proportion of the variance of $\mathbf{y}$ explained by each MEM (all R^2 summing to one). The $(n-1)$ coefficients of determination compose the *scale profile* of $\mathbf{y}$. A strong $R^2(\mathbf{y}, \mathbf{u}_j)$ would indicate that $\mathbf{y}$ exhibits spatial patterns at the j^{th} scale, while $(n-1)$ evenly distributed coefficients would denote an absence of spatial pattern.

This operation can be extended to multivariate data. Let $\mathbf{Y}$ be a n by q matrix of transformed variables, *i.e.* including centred and scaled quantitative variables and dummy vectors. The q by $(n-1)$ matrix of coefficients of determination $\mathbf{S}$ is then obtained by:

$$\mathbf{S} = \frac{1}{n^2}(\mathbf{Y}^T\mathbf{U} * \mathbf{Y}^T\mathbf{U})$$

where '$*$' denotes the Hadamard product (*i.e.*, element-wise product) and where $\mathbf{Y}^T$ is the

transposed matrix of $\mathbf{Y}$. Each row of $\mathbf{S}$ corresponds to a variable (or to a level of a factor) and sums to one, while each column corresponds to a MEM.

Finding essential scales of spatial variation

The matrix $\mathbf{S}$ is analysed using an adapted version of PCA. This adaptation involves three points: i) a particular centring, ii) an adapted row weighting, and iii) a graphical representation exploiting the row-sum constraint of $\mathbf{S}$ (*i.e.*, all R^2 of a single variable sum to one).

First, columns of $\mathbf{S}$ must be centred to define a point of reference corresponding to a non-informative state. Usual centring by subtracting the mean of the columns to each value of $\mathbf{S}$ is irrelevant as it would reduce the information given by the most structuring scales (*i.e.*, MEMs having high R^2 on average). Here, centring should be done by subtracting the value of R^2 that would be observed when a variable is not spatially structured at any scale. In the case of normally-distributed variables, this expected R^2 equals $1/(n-1)$ (Kendall and Stuart 1961, p. 341). In the case of non-normal variables, the expected value of R^2 is determined using a non-parametric approach proposed by Peres-Neto et al. (2006): the rows of $\mathbf{Y}$ are randomly permuted, breaking possible spatial structures in the data, and a new matrix of R^2 is computed. This operation is performed a large number of times (1000 by default), giving a distribution of expected R^2 for each value in $\mathbf{S}$. The means of these distributions are used as centring values. In both parametric and non-parametric cases, the expected R^2 (denoted $E(R^2(\mathbf{y}_i, \mathbf{u}_j))$ in Figure 1) are subtracted from the terms of $\mathbf{S}$, yielding a matrix $\mathbf{Z}$. Each row of $\mathbf{Z}$ measures the difference between the scale profile of a variable and the expected profile of a variable exhibiting no spatial structure at any scale (*i.e.*, the 'null profile').

Second, it must be considered that ordinary PCA would give equal weights ($1/q$) to all rows of $\mathbf{Z}$. This would be unfortunate as these rows can correspond to a quantitative variable or to a level of a qualitative variable. In other words, a qualitative variable with four levels would have four times the weight of a quantitative variable in the analysis. Moreover, centred and scaled dummy vectors coding a qualitative variable are, by construction, linearly dependent. This could induce spurious correlations between the corresponding scale profiles. Hence, if a level exhibits spatial structures at particular scales, the other levels of the same variable also convey, in part, this information. To avoid such redundancies to affect the method, we define row weights for $\mathbf{Z}$ so that all variables (quantitative and qualitative) have the same weight. If there were originally p variables, the weight given to each one should be $1/p$. As in multiple correpondence analysis, dummy vectors are weighted proportionnally to the number of observations of the corresponding level: a modality observed k times would be given the weight $k/(pn)$, so that all modalities of a single variable sum

to $1/p$. The diagonal matrix of row weights is denoted $\mathbf{D}$. MSPA is a PCA of $\mathbf{Z}$ in which rows are weighted by $\mathbf{D}$, without additional centring or scaling of the columns of $\mathbf{Z}$. This PCA is the eigen analysis of $\mathbf{Z}^T\mathbf{DZ}$. Note that because the columns of $\mathbf{Z}$ are neither centred to mean zero nor scaled to unitary variance, the diagonalized matrix $\mathbf{Z}^T\mathbf{DZ}$ is not a covariance nor a correlation matrix. In fact, $\mathbf{Z}^T\mathbf{DZ}$ simply is the symmetric matrix of scalar products between the columns of $\mathbf{Z}$ computed with the metric $\mathbf{D}$. MSPA yields *synthetic scales* summarizing the differences between the variables according to their multi-scale spatial patterns. Principal axes provide a new orthonormal basis onto which variables are represented so that their inertia (*i.e.*, the squared Euclidean distances between the scale profiles) is maximized.

Third, it must be emphasized that the analysed matrix $\mathbf{S}$ contains compositional data, as each scale profile sums to one (Aitchison 2003). de Crespin de Billy et al. (2000) developed a PCA for compositional data (%PCA) which is employed here. The %PCA generates biplots using the principal axes previously defined, but each variable is represented at the centre of the MEM coordinates weighted by its original profile (R^2 coefficients). Technically, this is achieved by projecting the non-centred matrix $\mathbf{S}$ instead of $\mathbf{Z}$ onto the principal axes. Note that the squared distances among scale profiles are still optimized by this projection. What %PCA adds to PCA is that a scale profile can be inferred by the position of a variable with respect to the MEMs: the closer a variable is to a given scale, the closer the corresponding R^2 is to one.

Finally, MSPA provides informative biplots representing the most structured variables inside an envelope formed by the most structuring scales. As usual in reduced space ordination methods, the number of retained axes should be chosen according to the decrease of eigenvalues, which represents the amount of structure explained by each axis. MSPA will be implemented in the next release of the ade4 package (Chessel et al. 2004, Dray et al. 2007) of the free software R (R Development Core Team 2008). Functions in R language are provided in Appendix B.

ILLUSTRATIONS

The R code allowing the reproduction of these analyses is provided in Appendix C.

Simulated data

The construction of this dataset is detailed in Appendix D. It contains measurements of 35 variables (V1-V35) for 100 observations distributed on a 10 by 10 regular grid and linked using the rook connection (*i.e.* neighbours share one edge, Legendre and Legendre 1998, p. 752). The MEMs ranged from $\mathbf{u}_1$ (largest scale) to $\mathbf{u}_{99}$ (finest scale). Seven spatially structured variables (V1-V7) were obtained by linear combinations of MEMs with the addition of random noise (see Appendix D). The other variables (V8-V35) were drawn randomly from a normal variate distribution.

Variables V1-V3 exhibited spatial patterns at the largest scales ($\mathbf{u}_1$, $\mathbf{u}_2$, and $\mathbf{u}_3$), while V4 was structured at an intermediate scale ($\mathbf{u}_{44}$, $\mathbf{u}_{45}$, and $\mathbf{u}_{46}$), and V5-V7 were structured at the finest scales ($\mathbf{u}_{97}$, $\mathbf{u}_{98}$, and $\mathbf{u}_{99}$).

This dataset was analysed using both MSPA and RDA. MSPA clearly showed three axes to be retained (Figures 2A-B). On the MSPA biplots (Figures 2A-B), the most structuring scales correspond to the MEMs that are the closest to the circle of radius one. Note that because MEMs are orthogonal, a variable cannot be perfectly correlated to several MEMs at the same time. Hence, it is unlikely that several MEMs will be given a strong loading on the same axis of MSPA. The variables exhibiting the strongest spatial patterns are the furthest from the origin. The first axis identified variables V5-V7 as being structured at the finest scales, while the second axis retrieved large-scale structured variables V1-V3 (Figure 2A). The third axis (Figure 2B) found the medium-scale pattern in V4. It is worth noting that MSPA successfully identified all structures of these simulated data, without finding any artifactual patterns.

Prior to the RDA, forward selection was applied to choose the relevant MEMs using the R package 'packfor' proposed by S. Dray (`http://biomserv.univ-lyon1.fr/~dray/software.php`). The best model for a theoretical α level of 5% retained 14 MEMs, including the 9 structuring and 5 non-structuring MEMs (see Appendix E for complete results). This reinforces the criticisms made by Dray et al. (2006) about using forward selection to select relevant MEMs. RDA was performed with simulated data as response variables and the 14 selected MEMs as predictors. Two axes were interpreted although the decrease of eigenvalues suggested only one axis (Figures 2C-D). The analysis detected large and fine scales but ommited the intermediate one (Figure 2C). However, the difference between structuring and non-structuring scales was far less obvious than in MSPA (Figures 2A-C). Moreover, all structured variables were not revealed: only V2, V3, V5, and V7 seemed to contain spatial patterns (Figure 2D).

Empirical data: oribatid mites

This illustration involves the famous oribatid mites dataset (Borcard et al. 1992, Borcard and Legendre 1994), which is available in the ade4 package as the dataset 'oribatid'. The data contained a table of 5 environmental variables (quantitative and qualitative data) measured on 70 georeferenced sites and a table giving counts of oribatid mites for 35 species at the same sites. Our purpose was i) to investigate the scales of spatial variation in both tables separately and ii) to study how species multi-scale patterns were linked to those of the environment. We used the Delaunay triangulation (Upton and Fingleton 1985) to model the spatial connectivity among the sites. MEMs ranged for both datasets from $\mathbf{u}_1$ (largest scale) to $\mathbf{u}_{69}$ (finest scale). Both species and environmental

data were regressed onto spatial coordinates, as it was done in previous studies, to yield comparable results (see Borcard et al. 2004).

The eigenvalues of the MSPA of environmental data showed that two axes should be retained (Figure 3A). The principal axes mainly represented two large scales ($\mathbf{u}_2$ and $\mathbf{u}_3$), but other finer scales like $\mathbf{u}_6$ or $\mathbf{u}_{13}$ also contributed fairly to both. The density of shrub cover (qualitative variable 'shrub') displayed the strongest structuring: a high density of shrubs ('shrub.many') exhibited a strong pattern at scale $\mathbf{u}_3$, while the absence of shrubs ('shrub.none') was linked to the scale $\mathbf{u}_2$. These results are consistent with previous results of Borcard et al. (2004).

Species data were Hellinger-transformed prior to MSPA (Legendre and Gallagher 2001). Two principal axes were retained (Figure 3B). The species were mainly structured at large scales ($\mathbf{u}_2$, $\mathbf{u}_3$, and $\mathbf{u}_4$), but intermediate scales also contributed to the axes ($\mathbf{u}_4$, $\mathbf{u}_5$, $\mathbf{u}_7$, and $\mathbf{u}_8$). Some species displayed moderate spatial patterns at the largest scales (R^2 around 0.3), but none were distinguished from the others by a strong spatial pattern.

These results raised the question of how the multi-scale spatial patterns of species are determined by environmental variables. To study species-environment relationships at multiple scales, a canonical MSPA was performed: species data were predicted by multiple regression onto environmental variables, and the obtained predictions were submitted to a MSPA. This method can be employed to investigate the environmental components of the multi-scale spatial patterns of the species. The first two eigenvalues were clearly larger than the others and were therefore retained (Figure 4A). Environmental variables were projected onto the principal axes as supplementary individuals (Figure 4B): the first spatial structure ($\mathbf{u}_2$) was mainly related to the absence of shrubs ('shrub.none'), while the second pattern ($\mathbf{u}_3$) was linked to large quantities of shrubs ('shrub.many'). Strikingly, the multi-scale structuring of species was no longer composed of large and intermediate scales as seen in the previous analysis (Figure 3B), but only consisted of two large scales ($\mathbf{u}_2$ and $\mathbf{u}_3$, Figure 4A). Moreover, some species like TVEL (*Tecticepheus velatus*) or Trimalaco2 (*Trimalaconothrus* species) exhibited stronger structuring than others, with R^2 values around 0.5, showing that half of the variance predicted by environment was of spatial essence (Figure 4A). This analysis reinforced the idea that *spatial dependence* (*i.e.*, spatial patterns of species induced by environment) mainly occurs at relatively large scales, as opposed to *spatial autocorrelation*.

DISCUSSION

This paper presents multi-scale pattern analysis (MSPA) as a new tool to investigate the scales of spatial variation in ecological data, using quantitative and qualitative variables. Classically, such investigation has been performed by constrained ordinations in which MEMs are used as spatial

predictors after a forward selection procedure. However, as underlined by Dray et al. (2006) and shown in our simulated example, this method can provide very large type I error: it tends to find more structuring scales than there actually are. On the contrary, MSPA does not rely on testing procedures; it can be used for a preliminary exploration of data, to assess the existence of multi-scale spatial patterns (using the eigenvalues screeplot) and to identify both structuring scales and structured variables (using biplots). Moreover, MSPA can be extended to canonical and partial canonical MSPA by performing a multivariate regression of data onto a set of explanatory variables, as done in constrained ordinations (Rao 1964, Ter Braak 1986; 1988). Canonical MSPA can be used to extract the environmental component of the multi-scale spatial patterns of species, thereby focusing on *spatial dependence*. Partial canonical MSPA can be employed to study multi-scale spatial patterns in species after removing the environmental effect from species data; provided all relevant environmental variables have been used in this operation, one would get rid of *spatial dependence* and observe true *spatial autocorrelation* (*sensu* Wagner 2004).

Several points concerning the method shall be discussed. First, we used MEMs to model spatial scales, which was not their initial purpose: they were by-products of the decomposition of Moran's index of spatial autocorrelation (de Jong et al. 1984, Tiefelsdorf and Boots 1995). But, as underlined by Griffith (2000), these vectors have excellent properties to be used as multi-scale spatial predictors. They are centred (all MEMs have a mean of zero), scaled (all MEMs have a norm equaling one), orthogonal and uncorrelated (MEMs are not subject to multicollinearity problems), and each MEM models a different scale of variation. It could be argued that modelling the concept of scale (which is continuous and can involve an infinite number of levels) using a discrete approach (the number of MEMs is finite) is somewhat clumsy. This would not be justified, however, because MSPA seeks linear combinations of MEMs; there is an infinite number of such combinations, which likely enables the method to detect very complex spatial patterns.

Other spatial predictors besides MEM may have been considered, such as PCNM (Borcard and Legendre 2002, Borcard et al. 2004). Indeed, many of the mathematical properties of MEM (*e.g.*, orthonormality) are also found in PCNM (Borcard and Legendre 2002, Borcard et al. 2004) because these are particular cases of MEM (Dray et al. 2006). However, the original PCNM approach yields $2n/3$ spatial predictors, which is usually insufficient to decompose the whole variability of n observations. On the contrary, $(n-1)$ MEMs always explain 100% of the variability of n centred observations. Moreover, MEMs are not contingent upon the choice of a particular threshold, contrary to PCNM.

A last concern is about the dimensionality of the matrix of centred R^2 ($\mathbf{Z}$) which is submitted to a particular PCA. This matrix has q variables (or species) in rows and $(n-1)$ MEMs in columns, where q would often be lower than $(n-1)$. In such cases, the results of PCA can be numerically

instable: adding or removing a variable could induce important changes in the principal components (Costello and Osborne 2004; 2005). Numerical instability would be problematic when one wants to draw conclusions beyond the sample that is studied, to make inferences about the population from which observations were drawn (Costello and Osborne 2005). In some cases, ecological descriptors could be used to infer the multi-scale spatial structuring of a larger set of variables (or species), such as an ecosystem. Then, numerical stability would be required, since introducing a new descriptor in MSPA should not change the assessment of the main structuring scales of the ecosystem. To achieve numerical stability, it would be necessary to measure all relevant descriptors of the ecosystem at a large number of sites, which would likely result in an extensive experimental design. However, MSPA will most often be employed to describe the main scales of spatial variation in a set of variables or species, without drawing conclusions about other variables or species. In such cases, the PCA of $\mathbf{Z}$ is simply used as a descriptive tool, to summarize the information contained by the matrix of scale profiles into a few dimensions, and numerical stability is no longer required (Joliffe 2004, Costello and Osborne 2005).

As a conclusion, MSPA is an exploratory tool adapted to the multi-scale nature of spatial patterns. Its function is to peer at the ecological variability through 'scale filters' — the MEMs — which can detect many different scales of spatial variation. Furthemore, MSPA could also be applied in other domains in which MEMs can be used, for instance in time series and in phylogeny (Peres-Neto 2006), which increases significantly the potential of the method.

ACKNOWLEDGEMENTS

We are grateful to Pedro Peres-Neto and two anonymous reviewers for providing constructive comments on an earlier version of the manuscript.

LITERATURE CITED

Aitchison, J. 2003. The statistical analysis of compositional data. The Blackburn Press.

Borcard, D. and Legendre, P. 1994. Environmental control and spatial structure in ecological communities: an example using oribatid mites (Acari, Oribatei). Environ. Ecol. Stat. 1: 37–61.

Borcard, D. and Legendre, P. 2002. All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. Ecol. Model. 153: 51–68.

Borcard, D. et al. 2004. Dissecting the spatial structure of ecological data at multiple scales. Ecology 85: 1826–1832.

Borcard, D. et al. 1992. Partialling out the spatial component of ecological variation. Ecology 73: 1045–1055.

Brind'Amour, A. et al. 2005. Multiscale spatial distribution of a littoral fish community in relation to environmental variables. Limnol. Oceanogr. 50: 465–479.

Chessel, D. et al. 2004. The ade4 package-I- One-table methods. R News 4: 5–10.

Costello, A.B. and Osborne, J.W. 2004. Sample size and subject to item ratio in principal components analysis. Pract. Assess. Res. Eval. 9 (11).

Costello, A.B. and Osborne, J.W. 2005. Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. Pract. Assess. Res. Eval. 10 (7).

de Crespin de Billy, V. et al. 2000. Biplot presentation of diet composition data: an alternative for fish stomach contents analysis. J. Fish Biol. 56: 961–973.

de Jong, P. et al. 1984. On extreme values of Moran's I and Geary's c. Geogr. Anal. 16: 17–24.

Dray, S. et al. 2007. The ade4 package - II: Two-table and K-table methods. R News 7: 47–54.

Dray, S. et al. 2006. Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbours matrices (PCNM). Ecol. Model. 196: 483–493.

Dungan, J., et al. 2002. A balanced view of scale in spatial analysis. Ecography 25: 626–640.

Getis, A. and Griffith, D. 2002. Comparative spatial filtering in regression analysis. Geogr. Anal. 34: 130–140.

Griffith, D. 1996. Spatial autocorrelation and eigenfunctions of the goegraphic weights matrix accompanying geo-referenced data. Can. Geogr. 40: 351–367.

Griffith, D. 2000. A linear regression solution to the spatial autocorrelation problem. J. Geogr. Sys. 2: 141–156.

Griffith, D. and Peres-Neto, P. 2006. Spatial modeling in ecology: the flexibility of eigenfunction spatial analyses. Ecology 87: 2603–2613.

Joliffe, I.T. 2004. Principal component analysis. Springer.

Kendall, M. and Stuart, A. 1961. The advanced theory of statistics. Charles Griffin & Company Limited.

Legendre, P. 1993. Spatial autocorrelation: trouble or new paradigm? Ecology 74: 1659–1673.

Legendre, P. and Fortin, M. 1989. Spatial pattern and ecological analysis. Vegetatio 80: 107–138.

Legendre, P. and Gallagher, E. 2001. Ecologically meaningful transformations for ordination of species data. Oecologia 129: 271–280.

Legendre, P. and Legendre, L. 1998. Numerical ecology. Elsevier Science B.V., Amsterdam.

Menge, B. and Olson, A. 1990. Role of scale and environmental factors in regulation of community structure. Trends Ecol. Evol. 5: 52–57.

Peres-Neto, P. 2006. A unified strategy for estimating and controlling spatial, temporal and phylogenetic autocorrelation in ecological models. Oecol. Brasil. 10: 105–119.

Peres-Neto, P. and Legendre, P. and Dray, S. and Borcard, D. 2006. Variation partitioning of species data matrix: estimation and comparison of fractions. Ecology 87: 2614–2625.

R Development Core Team, 2008. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL: `http://www.R-project.org`.

Rao, C.R. 1964. The use and interpretation of principal component analysis in applied research. Sankhya, A 26: 329–359.

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

Ter Braak, C., 1988. Partial Canonical Analysis. Pages 551–558 *in* H. Bock, editor. Classification an related methods of data analysis. North Holland.

Tiefelsdorf, M. and Boots, B. 1995. The exact distribution of Moran's I. Envir. Plan. A 27: 985–999.

Tiefelsdorf, M. and Griffith, D. 2007. Semiparametric filtering of spatial autocorrelation: the eigenvector approach. Envir. Plan. A 39: 1193–1221.

Upton, G. and Fingleton, B. 1985. Spatial data analysis by sample. Vol. 1: Point pattern and quantitative data. Wiley, New York.

Wagner, H. 2004. Direct multi-scale ordination with canonical correspondence analysis. Ecology 85: 342–351.

Wagner, H. and Fortin, M.-J. 2005. Spatial analysis of landscapes: concepts and statistics. Ecology 86: 1975–1987.

Wiens, J. 1989. Spatial scaling in ecology. Funct. Ecol. 3: 385–397.

FIGURE LEGENDS

Figure 1:
Diagram of the computations of MSPA. Computation details are provided in Appendix A. 'MEMs' stands for 'Moran's eigenvector maps'.

Figure 2:
Analyses of simulated data: (A) MSPA biplot, axes 1-2; screeplot indicates displayed eigenvalues in black and retained ones in grey. MEMs are represented by arrows; those being well represented are close to the circle of unity radius; 'd' indicates the mesh of the grid. (B) MSPA biplot, axes 1-3. (C) Correlation circle of the RDA of simulated data predicted by 14 MEMs selected by forward selection. (D) Variable scores of the RDA of simulated data (14 MEMs retained as predictors).

Figure 3:
Biplots of the MSPA of oribatid mites dataset (same representations as in figures 1A-B). The qualitative variables are positionned at the average of the coordinates of their modalities. (A) MSPA of environmental variables. (B) MSPA of species.

Figure 4:
MSPA of oribatid mite species predicted by environmental data. (A) MSPA biplot. (B) projection of environmental variables onto MSPA axes.

FIGURES

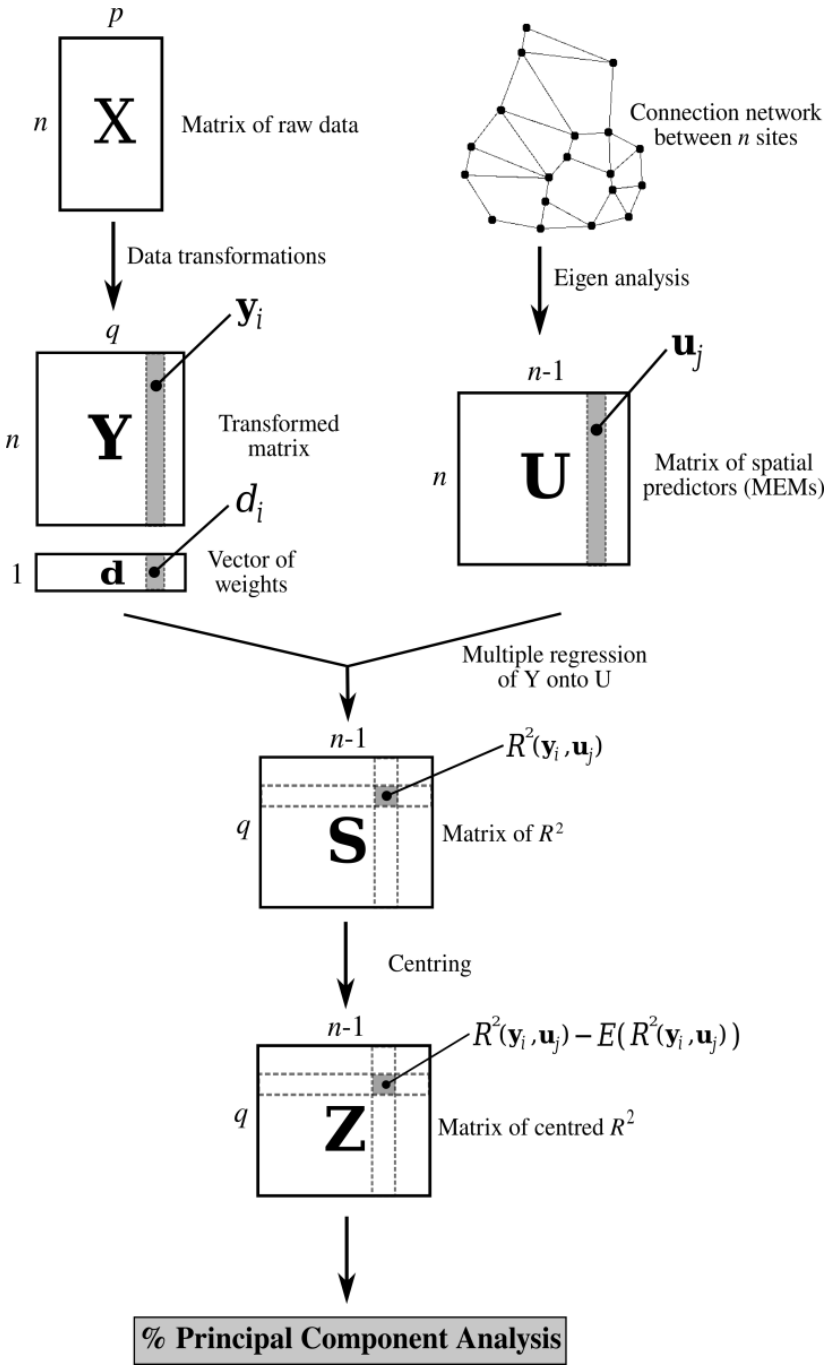


Figure 1

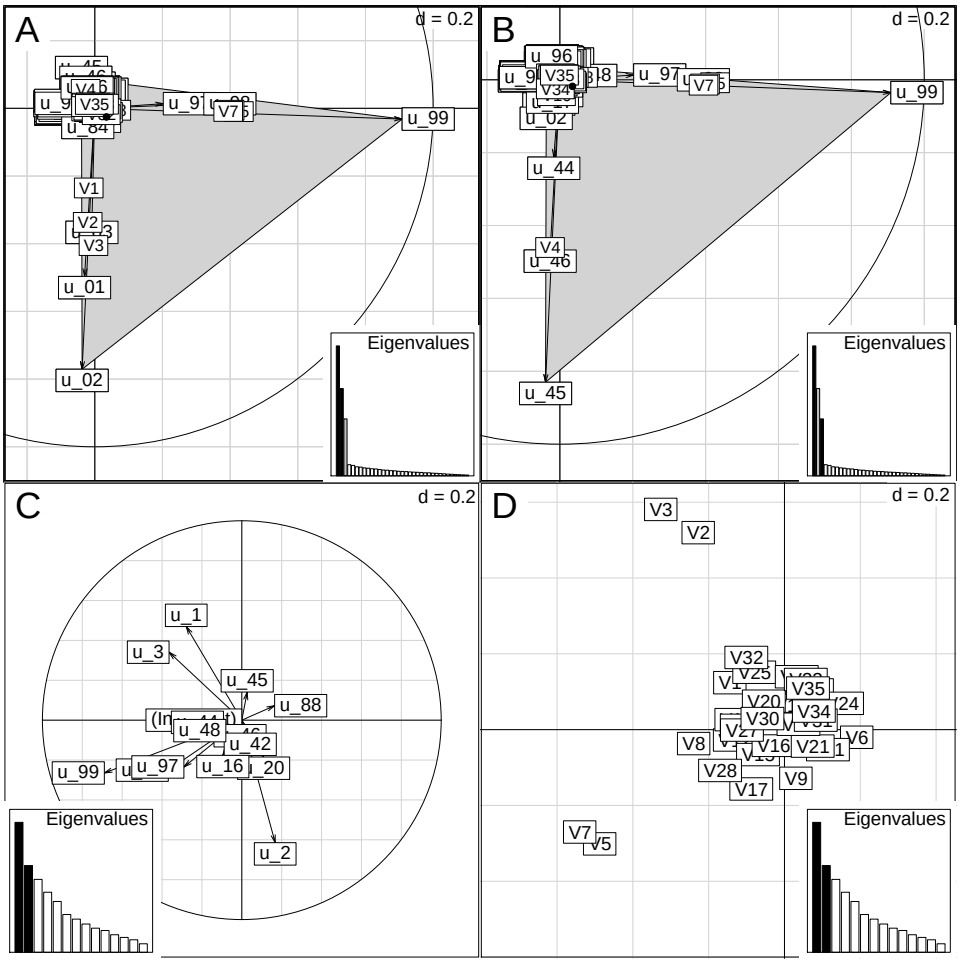

A
d = 0.2
u_99
V1
V2
V3
u_01
u_02
Eigenvalues
B
d = 0.2
u_96
u_97
V35
V7
u_99
u_44
V4
u_46
u_45
Eigenvalues
C
d = 0.2
u_1
u_3
u_45
u_88
u_48
u_42
u_99
u_97
u_16
u_20
u_2
Eigenvalues
D
d = 0.2
V3
V2
V32
V35
V20
V34
V30
V8
V16
V21
V6
V28
V9
V17
V7
V5
Eigenvalues

Figure 2

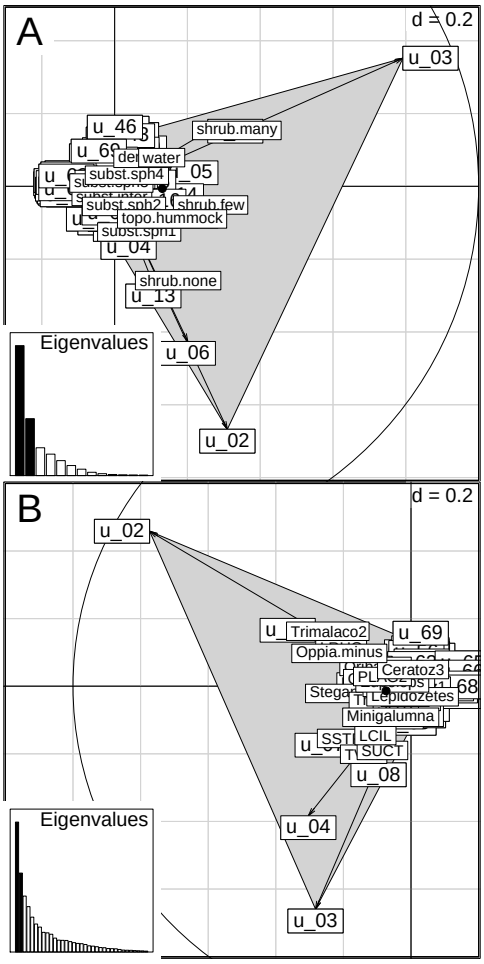

A
d = 0.2
u_03
u_46
shrub.many
water
subst.sph4
05
shrub.few
topo.hummock
u_04
shrub.none
u_13
Eigenvalues
u_06
u_02
B
d = 0.2
u_02
Trimalaco2
u_69
Oppia.minus
Ceratoz3
Minigalumna
LCIL
SUCT
u_08
u_04
u_03
Eigenvalues

Figure 3

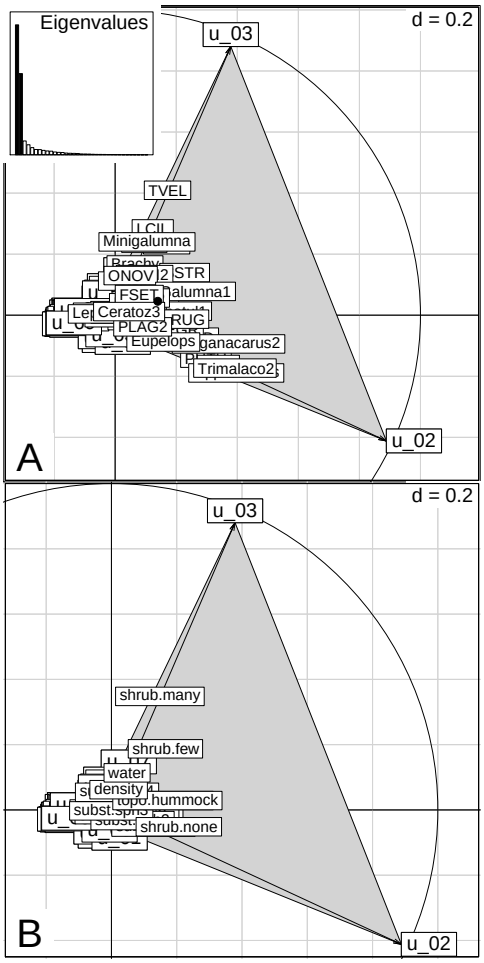

Eigenvalues
d = 0.2
u_03
TVEL
Minigalumna
ONOV
STR
alumna1
Ceratoz3
RUG
PLAG2
Eupelops
ganacarus2
Trimalaco2
u_02
A
d = 0.2
u_03
shrub.many
shrub.few
water
density
topo.hummock
shrub.none
u_02
B

Figure 4

5.3 Dicussion

Bien que développée en écologie, la MSPA peut être utilisée dans d'autres contextes, notamment en génétique. Cette discussion s'attache dans un premier temps à expliciter les liens existants entre la MSPA et les tests globaux et locaux proposés avec la sPCA (Jombart *et al.*, 2008). Dans un second temps, nous proposons d'illustrer l'intérêt de la méthode pour l'analyse de données génétiques spatialisées en ré-analysant les données des ours de Scandinavie illustrant la sPCA.

5.3.1 Liens avec les tests globaux et locaux

Si les tests contre l'absence de structures spatiales globale et locale ont été proposés en complément de la sPCA, ces tests sont en réalité beaucoup plus proches de la MSPA que de la sPCA. Cette proximité apparaît dès lors que l'on compare les procédures.

Si l'on examine la procédure de test décrite à l'annexe B de la sPCA (chapitre 3) et la figure 1 de l'article présentant la MSPA, on remarque immédiatement une partie commune (Fig. 5.2), correspondant à la régression d'une matrice de vecteurs de $\mathbb{R}^n$ ($\mathbf{Y} = [\mathbf{y}_1|\mathbf{y}_2|\cdots|\mathbf{y}_q]$) $\frac{1}{n}\mathbf{I}_n$-centrés et réduits sur un ensemble de vecteurs de Moran ($\mathbf{U} = [\mathbf{u}_1|\mathbf{u}_2|\cdots|\mathbf{u}_r]$). On obtient une matrice $q \times r$ $\mathbf{S}$ de coefficients de détermination ($\mathbf{S} = [s_{ij} = R^2(\mathbf{y}_i, \mathbf{u}_j)]$ par le produit :

$$\mathbf{S} = \frac{1}{n^2}\mathbf{Y}^T\mathbf{U} \bullet \mathbf{Y}^T\mathbf{U} \tag{5.9}$$

où $\bullet$ désigne le produit d'Hadamard. Les quelques différences entre les deux approches portent sur les dimensions des matrices. En MSPA, les transformations de p variables peuvent former q nouveaux vecteurs (Fig. 5.2, gauche) et la totalité des vecteurs propres de Moran est utilisée ($r = n - 1$). Pour les tests globaux et locaux, les transformations de variables correspondent au centrage et à la réduction des allèles ($p = q$), mais seule une partie des vecteurs propres de Moran est utilisée (correspondant aux valeurs propres supérieures ou inférieures à $-1/(n-1)$).

Ces détails mis à part, les deux approches ont en commun qu'elles quantifient la structuration spatiale d'un ensemble de variables en mesurant la part de variance de chaque variable $\mathbf{y}_i$ expliquée linéairement par chaque vecteur de Moran $\mathbf{u}_j$. La MSPA résume la matrice $\mathbf{S}$ par une analyse multivariée alors que les tests globaux et locaux utilisent la plus forte moyenne par colonne de $\mathbf{S}$ comme statistique de test. Notons que cette dernière approche est pertinente dans le contexte génétique, parce qu'on s'attend globalement à ce qu'une majorité d'allèles exhibent la même structure spatiale. Si cette structuration spatiale correspond même grossièrement à un des vecteurs de Moran, alors les allèles auront en moyenne un R^2 élevé pour ce vecteur. Cette hypothèse ne tient plus dans le contexte écologique, dans la mesure où il est parfaitement envisageable, et même probable, que différents descripteurs environnementaux ou différentes espèces possèdent des structures spatiales différentes.

En conséquence, il est peu probable que les tests multivariés contre l'absence de structuration spatiale proposés avec la sPCA soient d'un usage pertinent en écologie. L'échange dans l'autre sens semble beaucoup plus pertinent, et l'on montre maintenant que la MSPA peut s'avérer utile pour l'analyse de données génétiques géoréférencées.

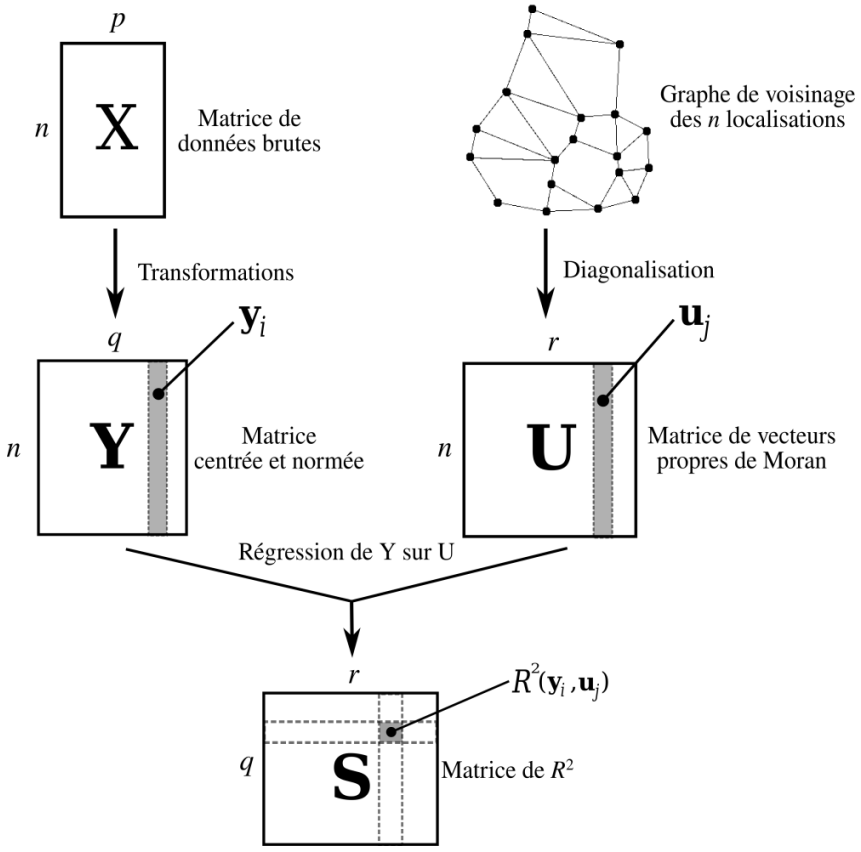


FIG. 5.2: Procédures communes à la MSPA et aux tests globaux et locaux

5.3.2 Une illustration en génétique

Les données de cette illustration sont décrites dans Waits *et al.* (2000), et ont été utilisées dans la présentation de différentes méthodes dont la (Jombart *et al.*, 2008). Elles comportent les génotypes géoréférencés de 964 ours bruns (*Ursus arctos*) pour 18 marqueurs microsatellites. L'identification d'unités de gestion au sein de cette population est considérée comme une question centrale (Taberlet *et al.*, 1995; Swenson *et al.*, 1998; Waits *et al.*, 2000) mais encore non tranchée (Swenson *et al.*, 1998; Waits *et al.*, 2000; Manel *et al.*, 2004, 2007). Un façon nouvelle de l'aborder est d'identifier à quelles échelles les données génétiques sont structurées spatialement, en utilisant la MSPA.

On commence par charger les données contenues dans un objet `genind` :

```
> load("ours.rda")
> ours


   ######################
   ### Genind object ###
   ######################
- genotypes of individuals -

S4 class:  genind
@call: old2new(object = ours)

@tab:  964 x 139 matrix of genotypes

@ind.names: vector of  964 individual names
```

```
@loc.names: vector of  18 locus names
@loc.nall: number of alleles per locus
@loc.fac: locus factor for the  139 columns of @tab
@all.names: list of  18 components yielding allele names for each locus
@ploidy:  2

Optionnal contents:
@pop:  - empty -
@pop.names:  - empty -

@other: a list containing: xy
```

La MSPA est implémentée par les fonctions `mspa` et `scatter.mspa`, qui devraient prochainement intégrer le package `sedaR` (`http://r-forge.r-project.org/projects/sedar/`), un projet visant à réunir les diverses méthodes d'écologie spatiale sous une seule coupe.

```
> source("mspa.R")
> args(mspa)


function (dudi, lw, scannf = TRUE, nf = 2, centring = c("param",
    "sim"), nperm = 1000)
NULL
```

La fonction `mspa` requiert un objet de la classe `dudi`, qui n'est utilisé que comme moyen de stockage des données transformées et des pondérations des descripteurs, et une liste de pondérations de voisinage (`lw`). Elle permet de choisir entre centrage paramétrique ou non paramétrique des coefficients de détermination (`centring = c("param","sim")`), et le cas échéant de choisir le nombre de permutations réalisées (`nperm = 1000`).

On effectue la MSPA des données génétiques en utilisant un graphe de voisinage par la distance minimale (telle que chaque génotype ait au moins un voisin), comme dans Jombart *et al.* (2008).

```
> ours.cn <- chooseCN(ours$other$xy, type = 5, d1 = 0, d2 = "dmin",
+     plot = FALSE, res = "listw")
> ours.pca <- dudi.pca(ours$tab, scannf = FALSE)


> ours.mspa <- mspa(ours.pca, ours.cn, scannf = FALSE, nf = 4)


> ours.mspa


Duality diagramm
class: mspa dudi
$call: mspa(dudi = ours.pca, lw = ours.cn, scannf = FALSE, nf = 4)

$nf: 4 axis-components saved
$rank: 136
eigen values: 0.00854 0.0006662 0.0003967 0.0002123 0.0001587 ...
  vector length mode    content
1 $cw    963    numeric column weights
2 $lw    139    numeric row weights
3 $eig   136    numeric eigen values

  data.frame nrow ncol content
1 $tab       139  963  modified array
2 $li        139  4    row coordinates
3 $l1        139  4    row normed scores
4 $co        963  4    column coordinates
5 $c1        963  4    column normed scores
other elements: ls R2 meanPoint varweights


> barplot(ours.mspa$eig[-1], main = "Valeurs propres\n(première valeur supprimée)")
> add.scatter(barplot(ours.mspa$eig, main = "Graphe non tronqué"),
+     posi = "topright", inset = c(0.1, 0.2), ratio = 0.4)
```

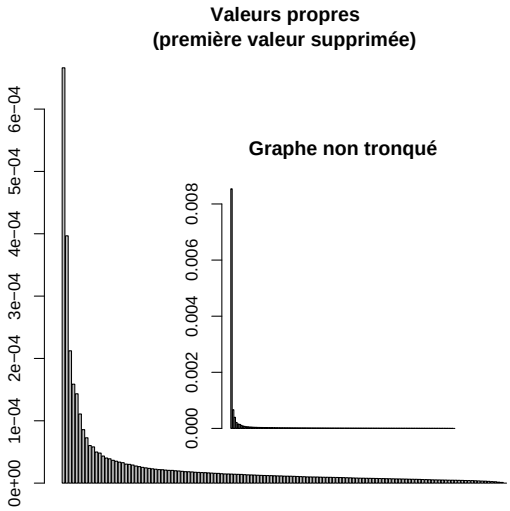
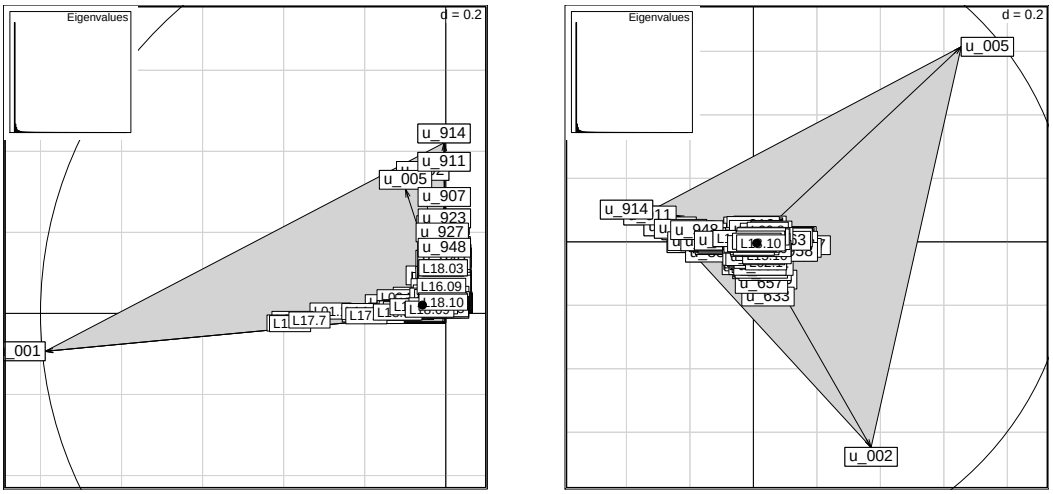


FIG. 5.3: Graphe des valeurs propres de la MSPA (ours bruns de Scandinavie)



(a) Axes 1-2 (b) Axes 3-4

FIG. 5.4: Biplots de la MSPA (ours bruns de Scandinavie)

```
> scatter(ours.mspa, clab.var = 1, clab.sca = 1.2)
```

```
> scatter(ours.mspa, xax = 3, yax = 4, clab.var = 1, clab.sca = 1.2)
```

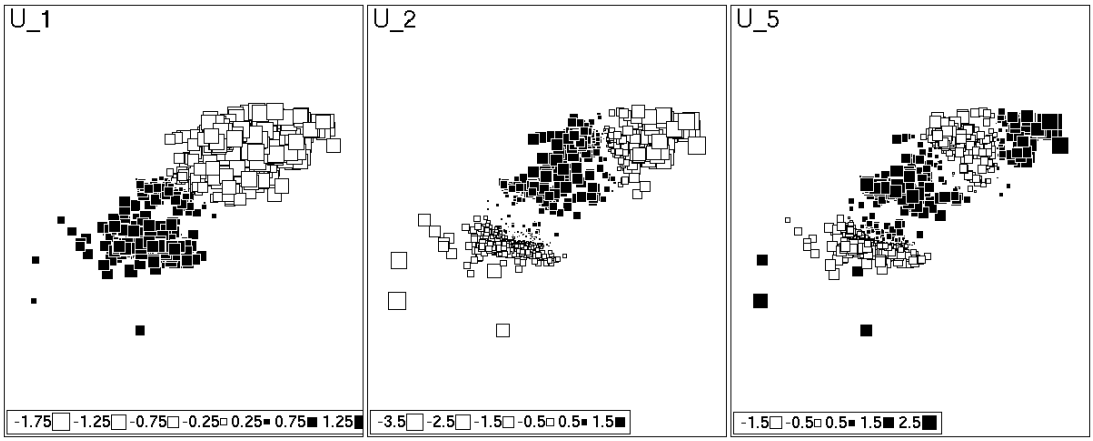
Les résultats de cette analyse sont intéressants à plusieurs titres. En considérant les quatre axes retenus (FIG. 5.3) dans leur ensemble, on trouve trois échelles larges prédominantes (FIG. 5.4, u_1, u_2 et u_5). On peut représenter ces vecteurs propres de Moran dans l'espace géographique :

```
> U <- as.matrix(orthobasis.listw(ours.cn))
```

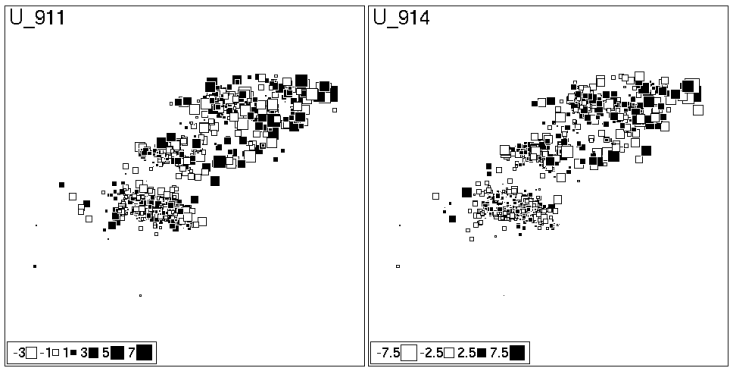
```
> par(mfrow = c(1, 3), cex = 1.5)
> for (i in c(1, 2, 5)) s.value(ours$other$xy, U[, i], include.ori = FALSE,
+     addaxes = FALSE, grid = FALSE, csize = 0.7, sub = paste("U",
+         i, sep = "_"))
```

On constate que ces vecteurs (Fig. 5.5a) correspondent assez bien aux structures mises en évidence par la sPCA dans Jombart *et al.* (2008). Néanmoins, il faut garder à l'esprit que ces structures ne sont pas directement interprétables en tant que structures génétiques : ces variables modélisent simplement des structures spatiales auxquelles les allèles sont globalement corrélés. Dans le cas présent, on pourra conclure que les échelles « larges » (Fig. 5.5a) de la structuration spatiale révélées par la MSPA sont cohérentes avec les structures mises en évidence par la sPCA.

```
> par(mfrow = c(1, 2), cex = 1.5)
> for (i in c(911, 914)) s.value(ours$other$xy, U[, i], include.ori = FALSE,
+     addaxes = FALSE, grid = FALSE, csize = 0.7, sub = paste("U",
+         i, sep = "_"))
```



(a) Structures à large échelle ($\mathbf{u}_1$, $\mathbf{u}_2$ et $\mathbf{u}_5$)



(b) Structures à fine échelle ($\mathbf{u}_{911}$ et $\mathbf{u}_{914}$)

Fig. 5.5: Cartographie des vecteurs propres de Moran (ours bruns de Scandinavie)

La façon dont les données sont structurées semblent différer entre (Fig. 5.4)a et (Fig. 5.4)b.

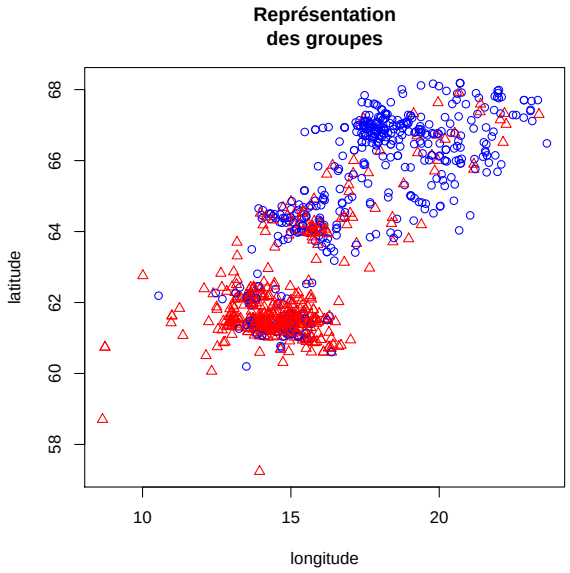
En effet, on voit que $\mathbf{u}_1$ est particulièrement structurante pour quelques allèles (Fig. 5.4a, *e.g.*, `L12.3`, `L17.7`), alors qu'aucun allèle ne semble plus lié que les autres à $\mathbf{u}_2$ et $\mathbf{u}_5$ (Fig. 5.4b). Il est admis que les individus du sud (Fig. 5.5a, gauche) forment une lignée distincte des autres (Taberlet *et al.*, 1995). On peut penser que les quelques allèles qui ressortent à ce niveau sont typiques de cette différenciation (allèles *privés*). On peut isoler les quelques allèles concernés et vérifier cette hypothèse :

```
> selAll <- colnames(ours$tab)[which(abs(ours.mspa$li[, 1]) > 0.3)]
> selAll
```

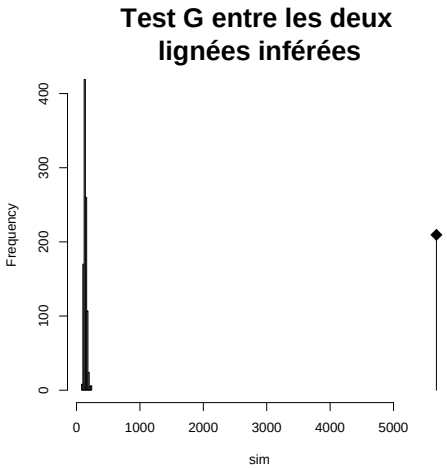
```
[1] "L09.6" "L12.2" "L12.3" "L17.7"
```

On utilise ces quatre allèles pour tenter de définir les deux lignées avec un groupement de Ward :

```
> subX <- ours$tab[, selAll]
> myTree <- hclust(dist(subX), method = "ward")
> lignee <- cutree(myTree, 2)
> plot(ours$other$xy, type = "n", xlab = "longitude", ylab = "latitude",
+     main = "Représentation\ndes groupes")
> points(ours$other$xy, pch = lignee, col = c("blue", "red")[lignee])
```



(a) Représentation dans l'espace des deux groupes obtenus



(b) Test G de la différenciation génétique des deux groupes obtenus

FIG. 5.6: Séparation de deux lignées d'ours bruns obtenue par groupement de Ward sur les 4 allèles liés à $\mathbf{u}_1$ (FIG. 5.4a)

On peut également tester la différenciation génétique entre les deux groupes obtenus par le test de la statistique G (Goudet *et al.*, 1996). Notons que cette démarche n'est pas circulaire puisque les allèles sélectionnés ne sont choisis que sur des critères spatiaux, et non de différenciation génétique.

```
> ours.gtest <- gstat.randtest(ours, pop = newFac, nsim = 999)
```

```
> plot(ours.gtest, main = "Test G entre les deux \nlignées inférées",
+     cex.main = 2)
> fstat(ours, pop = lignee, fstonly = TRUE)
```

```
[1] 0.07798527
```

Les quatre allèles retenus permettent à eux seuls de différencier les deux lignées (FIG. 5.6) de façon très significative (FIG. 5.6b). On a donc pu mettre en évidence grâce à la MSPA l'existence d'allèles privés caractérisant les deux lignées.

Enfin, on note que si les échelles les plus structurantes sont larges, quelques échelles très fines semblent aussi structurer les données (FIG. 5.4b). Ces structures, principalement $\mathbf{u}_{911}$ et $\mathbf{u}_{914}$, correspondent à des différences non aléatoires entre génotypes proches (FIG. 5.5b). Cette structuration locale, non mise en évidence par la sPCA (Jombart *et al.*, 2008), pourrait indiquer un processus d'évitement de la consanguinité.

Cette illustration montre que si la MSPA a été développée dans le contexte écologique, son application sort largement de ce cadre, et peut être en particulier pertinente dans le cadre génétique. En retrouvant des structures déjà connues, et en révélant de nouvelles structures ayant une interprétation biologique claire (l'existence d'allèles caractéristiques des deux lignées) ou ouverte (des différences génétiques locales), la méthode offre de nouvelles perspectives pour l'analyse des données génétiques géoréférencées. On retiendra cependant que si la MSPA semble capable d'identifier les principales échelles de la structuration spatiale d'un ensemble de génotypes, cette méthode ne permet pas, contrairement à la sPCA, de visualiser directement les structures spatiales. Loin d'être redondantes, ces deux approches apportent donc des informations complémentaires.

Chapitre 6

L'étude des structures phylogénétiques

Sommaire

6.1 Introduction

Cette partie constitue la première extension de la sPCA (Jombart *et al.*, 2008) hors du contexte spatial. On montre que la méthode peut être adaptée à la recherche de structures phylogénétiques dans un ensemble de traits biologiques. Par analogie à la sPCA, nous avons nommé cette méthode l'*analyse en composantes principales phylogénétiques* (ou *phylogenetic Principal Component Analysis* (pPCA) en anglais). Par « structures phylogénétiques », on entend la part de la variabilité des traits qui peut être expliquée par les relations phylogénétiques entre taxons. Nombre de méthodes permettant de mesurer, de supprimer ou d'étudier ces structures phylogénétiques sont largement utilisées dans le contexte spatial. Dans un premier temps, nous explicitons certaines de ces relations qui mettent en lumière la proximité, du point de vue biométrique, entre ces deux champs de recherche. Notons que cette proximité porte sur les objets et est indépendante des processus particuliers qui génèrent les structures, ce qui n'est pas pour autant un obstacle à l'utilisation de méthodes communes (Legay, 1997, p.54) :

> *compte tenu des relations partielles entre sujet et modèle* [...] *deux sujets explorés par un même modèle peuvent n'avoir aucun rapport entre eux : relever d'un même outil ne crée pas de liens de parenté sur le fond.*

La pPCA proprement dite fait l'objet d'une publication en préparation, qui est ensuite présentée. La fin de ce chapitre consiste en une illustration biologique de la pPCA, et en une discussion des perspectives offertes par la méthode.

6.1.1 L'autocorrélation phylogénétique : un problème original (?)

La méthode comparative (Harvey & Pagel, 1991) repose sur la comparaison de traits mesurés pour un ensemble de taxons. Ces comparaisons mettent en oeuvre des outils statistiques nécessitant l'indépendance des observations, indépendance qui est violée dans certains cas par l'existence d'une *autocorrélation phylogénétique* : deux taxons tendent à avoir des traits d'autant plus similaires qu'ils ont divergé d'un ancêtre commun depuis peu de temps. On peut considérer ce problème comme propre aux données comparatives, chercher à modéliser cette non indépendance par des modèle évolutifs et développer des outils statistiques dédiés ; c'est le point de vue soutenu par Harvey & Pagel (1991, p. v) :

> *In short, all useful comparative methods are based on explicit models of evolutionary change.*

Le point de vue du biométricien peut néanmoins différer. Sans pour autant nier l'intérêt des modèles d'évolution, on peut voir dans le problème de l'autocorrélation phylogénétique, un « simple » problème d'autocorrélation, pour lequel la statistique spatiale propose nombre de solutions. On peut en particulier remarquer que certaines de ces solutions sont actuellement appliquées en phylogénie, et que d'autres pourraient l'être. Notre approche procède directement de cette démarche, et avant de l'aborder, il convient présenter les autres éléments qui s'inscrivent en son sein.

La première importation d'une méthode de statistique spatiale en phylogénie fut celle des modèles autorégressifs (Cliff & Ord, 1973; Cheverud *et al.*, 1985; Gittleman & Kot, 1990; Cornillon *et al.*, 1999). Cette approche implique la définition du *vecteur lissé* $\mathbf{Wy}$ d'une variable $\mathbf{y}$, où $\mathbf{W}$ est une matrice de connectivité phylogénétique entre taxons normalisée par ligne, strictement équivalente à une matrice de pondération de voisinage utilisée dans le calcul du I de Moran (Eqn. 3.1). Le modèle autoregressif proposé par Cheverud *et al.* (1985) est un modèle de type :

$$\mathbf{y} = \alpha\mathbf{Wy} + \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \tag{6.1}$$

où α est un coefficient, $\mathbf{X}$ une matrice de prédicteurs, $\boldsymbol{\beta}$ un vecteur de coefficients associés, et $\mathbf{e}$ un vecteur de résidus. Notons que d'autres formes de modèles autorégressifs peuvent être utilisés, notamment pour incorporer le vecteur lissé des variables explicatives $\mathbf{X}$ (Anselin, 2002).

Le vecteur lissé étant défini, l'utilisation du I de Moran pour mesurer et tester l'autocorrélation phylogénétique était assez immédiate (Gittleman & Kot, 1990). Notons que Gittleman & Kot (1990) définissent la connectivité entre taxons à différentes échelles de la phylogénie et introduisent donc dans le même temps la notion de corrélogramme (Sokal & Oden, 1978) phylogénétique.

Les vecteurs propres de Moran (de Jong *et al.*, 1984; Tiefelsdorf & Boots, 1995; Griffith, 1996), qui sont utilisés pour le filtrage spatial en géographie (Griffith, 2000; Getis & Griffith, 2002; Tiefelsdorf & Griffith, 2007) et comme prédicteurs spatiaux en écologie (Dray *et al.*, 2006; Griffith & Peres-Neto, 2006), ont été proposés par Peres-Neto (2006) pour supprimer l'autocorrélation phylogénétique des données comparatives. Cette approche consiste à utiliser les vecteurs de Moran (Eqn. 5.6) d'une matrice de proximités phylogénétiques comme covariables d'un modèle. Suivant une démarche analogue, Diniz-Filho *et al.* (1998) utilisent les coordonnées principales d'une matrice de distances phylogénétiques comme covariables. Dans les deux cas, les modèles utilisés sont de la forme :

$$\mathbf{y} = \mathbf{E}\boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \tag{6.2}$$

où $\mathbf{E}$ est une matrice de prédicteurs spatiaux et $\boldsymbol{\alpha}$ un vecteur de coefficients associés.

Enfin, de même qu'en statistique spatiale (Dormann *et al.*, 2007), les moindres carrés généralisés sont utilisés en phylogénie pour incorporer la structure d'autocorrélation des résidus dans un modèle (Grafen, 1989; Martins & Hansen, 1997; Garland & Ives, 2000), qui possède alors la forme :

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{6.3}$$

où les résidus $\boldsymbol{\epsilon}$ suivent une loi normale multivariée $\mathcal{N}(\mathbf{0}, \mathbf{V})$ où $\mathbf{V}$ est une matrice de covariances des résidus.

On constate donc que nombre de méthodes utilisées en analyse comparative pour gérer le problème de l'autocorrélation phylogénétique ont été développées, ou bien sont également utilisées, en statistique spatiale. On peut s'étonner du fait que ces liens n'aient été jusqu'alors que peu reconnus (Ives & Zhu, 2006; Peres-Neto, 2006). Il est vraisemblable que d'autres méthodes spatiales (revues dans Dormann *et al.*, 2007) trouveraient également une application dans l'analyse de données comparatives. C'est en particulier le cas de la sPCA (Jombart *et al.*, 2008),

qui trouve une extension directe dans l'analyse de données comparatives avec la pPCA.

6.1.2 Origine de la méthode : le test d'Abouheif

Le test d'Abouheif est à l'origine un test de l'hypothèse d'indépendance phylogénétique entre obervations d'une variable quantitative ou qualitative sur un ensemble de taxons (Abouheif, 1999). Cette approche repose sur une mesure de proximité phylogénétique correspondant à la fréquence des permutations d'un arbre positionnant deux feuilles immédiatement côte-à-côte. Comme il est en pratique impossible de couvrir l'ensemble des permutations possibles d'un arbre, Abouheif (1999) recourt à une approximation des proximités phylogénétiques basée sur un grand nombre de permutations. Dans sa thèse, Ollier (2004) propose une solution analytique et exacte pour le calcul de ces proximités. La matrice des proximités phylogénétiques entre feuilles ($\mathbf{A} = [a_{ij}]$ avec $i, j = 1, \ldots, n$ où n est le nombre de feuilles) est définie par :

$$a_{ij} = \frac{1}{\prod_{p \in P_{ij}} dd_p} \text{ avec } i \neq j \text{ et } a_{ii} = 1 - \sum_{j=1, i \neq j}^{n} a_{ij} \tag{6.4}$$

où P_{ij} est l'ensemble de noeuds internes reliant les feuilles i et j, et dd_p est le nombre de descendants directs du noeud p. Ollier (2004) montre en outre que le test d'Abouheif est en fait un test du I de Moran utilisant la matrice $\mathbf{A}$ comme matrice de proximité (Eqn. 5.1, $\mathbf{A}$ remplaçant $\mathbf{W}$). Ces résultats ont été approfondis et diffusés par Pavoine *et al.* (2008). Nous inscrivant dans la lignée de ces travaux, il était donc logique de développer une extension de la sPCA pour l'analyse de données comparatives basée sur la proximité d'Abouheif.

Néanmoins, il semble ici important de clarifier un point : la pPCA peut utiliser n'importe quelle mesure de proximité phylogénétique, la matrice $\mathbf{A}$ n'étant qu'un cas particulier. Quelques arguments présentés dans l'article qui suit justifient ce choix. Mais comme on le verra par la suite, cette question reste ouverte.

6.2 Article 6 : Exploring phylogeny as a source of ecological information : a methodological approach

Article en préparation.

Exploring phylogeny as a source of ecological information: a methodological approach

Thibaut Jombart [a,*], Sandrine Pavoine [b],
Anne-Béatrice Dufour [a], Dominique Pontier [a]

[a] *Université de Lyon; université Lyon 1; CNRS; UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, 43 boulevard du 11 novembre 1918, Villeurbanne F-69622, France.*

[b] *Museum d'Histoire Naturelle de Paris; UMR 5173 MNHN-CNRS-P6 'Species conservation, restoration and monitoring of populations'; CRBPO, 55 rue Buffon, 75005, Paris, France*

Abstract

Key words:
phylogeny, phylogenetic principal component analysis, spatial principal component analysis, reduced space ordination, comparative method, simulations

1 Introduction

The importance of phylogeny in the study of biological traits among several taxa has long been recognized. For instance, Gregory (1913) and Osborn (1917) opposed two sources of biological variation, the *habitus* (*i.e.*, adaptation) and the *heritage* (*i.e.*, phylogenetic inertia). In their famous criticism of the adaptationist paradigm, Gould and Lewontin (1979) underlined the importance of the constraints imposed by the phylogeny to the variability observed among organisms. In comparative studies, the effect of phylogeny has merely been perceived as a source of nuisance, since it induces non-independence among the traits observed in taxa, and thus violates one of the basic assumptions required by most statistic tools.

* Corresponding author.
Email address: jombart@biomserv.univ-lyon1.fr (Thibaut Jombart).

Comparative methods were especially designed to solve this problem. The most common method developed in this context surely is the phylogenetic independent contrasts (PIC, Felsenstein, 1985) which, given a quantitative trait measured on n non-independent taxa, provides $(n-1)$ node values that are not phylogenetically autocorrelated under a Brownian motion model. In fact, PIC are now recognized (Grafen, 1989; Rohlf, 2001) as a particular case of generalized least squares (GLS, Rao and Toutenburg, 1999), which purpose is to incorporate the non-independence of observations inside a linear model by specifying the autocovariance matrix of the residuals. But where PIC rely on a Brownian motion model, GLS can incorporate virtually any model of evolution (Hansen and Martins, 1996) as far as a symmetric autocovariance matrix is provided. As stressed by Rohlf (2006), these approaches do not actually remove phylogenetic autocorrelation from data; they take this autocorrelation into account to provide more precise estimators of correlation or regression coefficients. But no matter how these method proceed, their purpose still is, in some ways, to cope with the phylogenetic signal.

We advocate that the phylogenetic signal is more than only a nuisance to the analysis of comparative data, but is also a source of relevant and interesting biological information. A similar statement was made in spatial ecology (Legendre, 1993), in which the study of spatial patterns emerged as a fecund paradigm. Following this idea, we propose to explore this phylogenetic signal, to see how traits are correlated to the phylogeny.

A step toward this direction was accomplished when Gittleman and Kot (1990) proposed to use Moran's I (Moran, 1948, 1950), a statistic originally developed to quantify spatial patterns (Cliff and Ord, 1981), to detect a phylogenetic signal in a given trait. Moran's I takes large positive values when closely related taxa tend to have similar traits, and conversely takes large negative values when close taxa tend to have dissimilar traits. This statistic relies on comparisons of the values of a trait across taxa with respect to their degree of relatedness in a phylogeny. It can thus incorporate proximities derived from various models of evolution. For instance, it has been demonstrated by Pavoine et al. (2008) that Abouheif's test (Abouheif, 1999) is in fact a Moran's I test using a particular evolution model.

In this paper, we present a multivariate approach to investigate the phylogenetic structure in a set of quantitative traits based on the detection of phylogenetic autocorrelation by Moran's I. The proposed methodology is derived from the spatial principal component analysis (sPCA, Jombart et al., 2008), a method developed to analyse spatial patterns of the genetic variability, and here adapted to infer phylogenetic structures in comparative data. Our approach is also related to the spatial ordination proposed by Dray et al. (2008) to study spatial patterns in vegetation data. First, we present Moran's I in a comprehensive framework, linking this statistic to other phylogenetic meth-

ods, and showing that it can incorporate proximities among taxa derived from any model of evolution. Then, we explain how the sPCA can be extended to a *phylogenetic principal component analysis* (pPCA) to investigate phylogenetic variation in comparative data. The proposed methodology is illustrated and evaluated using extensive simulations, and implemented in the free software R (R Development Core Team, 2008).

2 Statistical Method

2.1 A comprehensive framework for Moran's I

Adapting the former definition of Moran's I (Cliff and Ord, 1973, p13) to the phylogenetic context (Gittleman and Kot, 1990), this index can be defined as:

$$I(\mathbf{x}) = \frac{\mathbf{x}^T\mathbf{W}\mathbf{x}}{n}\frac{1}{\text{var}(\mathbf{x})} \tag{1}$$

where $\mathbf{x}$ is the centred vector of a trait observed on n taxa, $\text{var}(\mathbf{x})$ is the usual variance of $\mathbf{x}$, and $\mathbf{W}$ is a matrix of proximities among taxa ($\mathbf{W} = [w_{ij}]$ with $i, j = 1, \ldots, n$), whose diagonal terms equate zero ($w_{ii} = 0$), and all its rows summing to one ($\sum_{j=1}^{n} w_{ij} = 1$). The null value, *i.e.* the expected value when no phylogenetic autocorrelation arises, is $I_0 = -1/(n-1)$. In its initial formulation (Gittleman and Kot, 1990), $\mathbf{W}$ contained binary weights before normalization: w_{ij} was set to 1 if taxon i shared a common ancestor with taxon j at a given taxonomic level, and to 0 otherwise. Hence, taxa were considered as either phylogenetically related, or not, and Moran's I compared the trait of a taxon to the mean trait in related taxa to detect phylogenetic autocorrelation. To achieve better resolution in these comparisons, we propose using as terms of $\mathbf{W}$ any measurement of phylogenetic proximity valued in $\mathbb{R}$ meeting the following requirements:

$$\begin{cases} w_{ij} \geqslant 0 & \forall\ i, j = 1, \ldots, n \\ w_{ii} = 0 & \forall\ i = 1, \ldots, n \\ \sum_{j=1}^{n} w_{ij} = 1 & \forall\ j = 1, \ldots, n \end{cases} \tag{2}$$

Then, Moran's I compares the value of a trait in one taxon (terms of $\mathbf{x}$) to a weighted mean (terms of $\mathbf{W}\mathbf{x}$) in which the closest taxa are given stronger weights. This extension gives the index a great flexibility to measure phylogenetic autocorrelation, because phylogenetic proximities can be derived from

any model of evolution.

This formulation of Moran's I relates the index to other methodological approaches. The test proposed by Abouheif (1999), first appearing to rely on a flaw of the representation of a tree (non-uniqueness of representation), turned out to be a Moran's I test using a Monte-Carlo procedure to generate the reference distribution, and a particular measure of phylogenetic proximity for $\mathbf{W}$ (Pavoine et al., 2008).

Moran's I is also deeply related to autoregressive models. In their simplest form, these models are written as (Cheverud and Dow, 1985; Cheverud et al., 1985):

$$\mathbf{x} = \rho\mathbf{W}\mathbf{x} + \mathbf{e} \tag{3}$$

where ρ is an autocorrelation coefficient, and $\mathbf{e}$ is a vector of residuals. The matrix of phylogenetic relatedness $\mathbf{W}$ (Cheverud and Dow, 1985; Cheverud et al., 1985) is exactly the weight matrix of our definition of Moran's I (equation 1). The only difference between the two approaches is that autoregressive models perform the regression of $\mathbf{x}$ on $\mathbf{W}\mathbf{x}$, while I computes the inner product between both vectors (numerator of equation 1) to measure phylogenetic autocorrelation.

Lastly, the weighting matrix $\mathbf{W}$ is also the core of another approach producing variables modeling phylogenetic structures (Peres-Neto, 2006). Like Moran's I, this approach was initially developed in spatial statistics (Griffith, 1996), and consisted in finding eigenvectors of a doubly centred spatial weighting matrix (Dray et al., 2006). Applied to a matrix of phylogenetic proximity $\mathbf{W}$, this method yields uncorrelated variables modeling different observable phylogenetic patterns, each related to a value of Moran's I. One can perform the regression of a variable $\mathbf{x}$ onto these eigenvectors to 'remove' the phylogenetic autocorrelation from $\mathbf{x}$ (Peres-Neto, 2006). Another application, which was extensively exploited here (see section 3.1), is using these eigenvectors to simulate what we further call 'global' and 'local' phylogenetic structures.

2.2 Global and local phylogenetic structures

From equation 1, it can be seen that the meaning of Moran's I is given by the inner product of a centred trait $\mathbf{x}$ and the vector of mean traits weighted by phylogenetic proximities, $\mathbf{W}\mathbf{x}$. Hence, Moran's I will be significantly higher (respectively lower) than I_0 when closely related taxa tend to have similar (respectively dissimilar) values for the studied trait. We shall refer to the cor-

responding phylogenetic patterns as *global* and *local* structures. Clearly, the definition of phylogenetic proximities in **W** will condition the measurement of global and local structures. As shown by Pavoine et al. (2008), not all phylogenetic proximities are equal in detecting phylogenetic structuring. Especially, the phylogenetic proximities underlying Abouheif's test (matrix $\mathbf{A} = [a_{ij}]$ in Pavoine et al. (2008)) proved superior to several common phylogenetic proximities for testing phylogenetic inertia in traits simulated under Brownian and Ornstein-Uhlenbeck (OU) processes. More generally, the matrix **W** can be derived from any model of evolution which seems appropriate to the data, taking branch lengths into account whenever these are accurately estimated, and relying only on the topology in other cases.

Biologically, global and local patterns have different meanings. Global patterns are likely the most common, and reflect the general idea of phylogenetic inertia: traits observed in a set of taxa are not independent, but tend to be more similar in closely related taxa (*e.g.*, Figure 1A). Most common explanations for this phenomenon are inheritance from a common ancestor, niche conservatism, or similar adaptive strategies (Harvey and Pagel, 1991). Traits whose evolution can be satifyingly modeled by a Brownian or by an OU process generally display global patterns (Abouheif, 1999; Pavoine et al., 2008). Local structures may be more rarely observed, but are biologically meaningful as well. A local structuring would be observed whenever closely related taxa tend to be more different with respect to a given trait than randomly chosen taxa (*e.g.*, Figure 1E). This can occur, for instance, when the trait under study is selected towards different optimal values, *i.e.* when different evolutive strategies are observed in closely related taxa. If both kinds of pattern can easily be assessed when a few traits are under study, there is a need for a method that can efficiently retrieve such patterns in multivariate data.

2.3 The phylogenetic principal component analysis

The spatial principal component analysis (Jombart et al., 2008) summarizes strongly multivariate data into a few components expressing both a fair part of variability and strong spatial patterns. This is achieved by including Moran's I as a part of the criterion optimized by the principal components. The extension of sPCA to phylogenetic principal component analysis (pPCA) simply relies on using Moran's index to measure phylogenetic inertia rather than spatial autocorrelation. The purpose of pPCA is to summarize a set of p quantitative traits measured on n taxa into a few synthetic variables displaying the part of the variability which is phylogenetically structured. We denote $\mathbf{X}$ the $n \times p$ matrix of centred traits (viewed as n points in $\mathbb{R}^p$), and $\mathbf{W}$ the matrix of phylogenetic weights used in the computation of Moran's I (equation 1).

Because of its good performances in detecting phylogenetic structures (Pavoine et al., 2008), we used the phylogenetic proximity matrix **A** underlying the test of Abouheif (1999) to define **W**. The terms of **A** are defined as:

$$a_{ij} = \frac{1}{\prod_{p \in P_{ij}} dd_p} \text{ for } i \neq j \text{ and } a_{ii} = 1 - \sum_{j=1, i \neq j}^{n} a_{ij} \tag{4}$$

where P_{ij} is the set of internal nodes on the shortest path from tips i to j, and dd_p is the number of direct descendants from the internal node p. However, the matrix **W** differs from **A** in that its diagonal terms have to be null (equation 2). Hence, the rows of the proximity matrix have to be re-standardized, leading to:

$$w_{ij} = \frac{a_{ij}}{\sum_{j=1, i \neq j}^{n} a_{ij}} = \frac{a_{ij}}{1 - a_{ii}} \tag{5}$$

Note that all mathematical results presented below remain valid for any other measurement of phylogenetic proximity, as long as conditions (2) are verified. It can be shown that using the matrix **A** instead of **W** in a pPCA, and more generally any proximity matrix with a non-null diagonal, yields far less interpretable results (Appendix A). Mathematically, the purpose of the pPCA is to find the appropriate loadings $\mathbf{u} \in \mathbb{R}^p$ (with $\|\mathbf{u}\|^2 = 1$) giving the extremums of the function:

$$\begin{aligned} f : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^p &\longrightarrow \mathbb{R} \\ (\mathbf{X}, \mathbf{W}, \mathbf{u}) &\longmapsto \text{var}(\mathbf{Xu}) I(\mathbf{Xu}) \end{aligned} \tag{6}$$

The solution to this problem is given by the diagonalization of the matrix $\frac{1}{2n}\mathbf{X}^T(\mathbf{W} + \mathbf{W}^T)\mathbf{X}$ (Jombart et al., 2008). It results in a set of loadings $\{\mathbf{u_1}, \ldots, \mathbf{u_k}, \ldots, \mathbf{u_r}\}$ with $\mathbf{u_k} \in \mathbb{R}^p$ forming linear combinations of traits $\mathbf{Xu}_k$ (the so-called principal components) associated with decreasing eigenvalues λ_k, so that:

$$\text{var}(\mathbf{Xu_k}) I(\mathbf{Xu}_k) = \lambda_k \tag{7}$$

The largest eigenvalues likely correspond to a large variance and a strong positive I, indicating global structures. Conversely, the lowest (*i.e.*, most negative) eigenvalues correspond to a high variance and a large negative I, indicating local structures. As in other reduced space ordinations, the eigenvalues indicate the quantity of structure expressed by each synthetic variable. Hence, a sharp decrease in the screeplot is likely to indicate the boundary between strong and weak structures. The amount of variance and phylogenetic autocorrelation in each principal component $\mathbf{Xu}_k$ can be computed for a better interpretation

of each structure. Moreover, the loadings $\mathbf{u}_k$ can be used to infer which traits exhibit the most a given structuring.

3 Simulations

3.1 Data simulation

Extensive simulations were conducted to evaluate the sensitivity of the pPCA to different parameters. Datasets were simulated with different characteristics concerning the type of tree, the tree size, the type, strength and numbers of phylogenetically structured traits, and the total number of traits (including structured traits). These parameters are summarized in Table 1. Five types of trees were chosen to reflect various topologies: completely symmetric trees (Figure 1A), trees obtained by random clustering of tips (Figure 1B), the Yule model (Figure 1C), the biased model (Kirkpatrick and Slatkin, 1993, Figure 1D), and completely asymmetric trees (Figure 1E). The possible tree sizes were 16, 32, and 128 (powers of two were chosen because completely symmetric trees exist only in these cases). For each tree, tips data were simulated including both structured traits (*i.e.*, global and/or local structures), and random variates drawn from a normal distribution. Structured traits were obtained using the eigenvectors of a phylogenetic proximity matrix (Peres-Neto, 2006), added with random noise. The amount of added random noise was used to define different strength of patterns, with standard deviation equalling 0.5, 0.75 and 1 (eigenvectors being scaled to unitary variance). The measure of proximity among taxa defined previously (matrix $\mathbf{W}$,(5)) was used for this method too. Note that this does not involve any kind of circularity, as the computations of principal components in pPCA are not related to the eigenanalysis of the doubly centred matrix $\mathbf{W}$. When several structures of the same type (global or local) were simulated in a given dataset, these were derived from the same eigenvector, so that we could evaluate the performance of pPCA when a 'consensus' phylogenetic signal exists in a set of traits (*e.g.*, Figure 1B). Moreover, this was consistent with the fact that several phylogenetically structured traits are expected to exhibit the same structuring, either because these structures are caused by the same evolutive process, or because all traits are correlated to another structured trait. Structured traits varied in nature, number, and in the amount of random noise added to the eigenvectors (see Table 1). Lastly, the possible sizes of the datasets (including the structured traits) were of 10, 20, or 50 traits. The various combinations of parameters led to 810 different datasets, that were each simulated 200 times, resulting in a total of 162 000 simulated datasets.

All simulations were performed in the R software (R Development Core Team, 2008). Trees were simulated using the R packages ape (Paradis et al., 2004) and apTreeshape (Bortolussi et al., 2006), and self-defined routines for the symmetric model. Traits were simulated using the ade4 package (Chessel et al., 2004; Dray et al., 2007), and data manipulations and graphics were performed using the phylobase package (Bolker et al., 2007).

Tree model	Tree size[1]	Structures[2]	Random noise[3]	Number of traits[4]
Symmetric	16	1/0	0.5	10
Random clustering	32	0/1	0.75	20
Yule	128	3/0	1	50
Biased		0/3		
Asymmetric		1/1		
		3/3		

Table 1
Parameters of the simulated data. Each dataset was defined by one value in each column. (1) expressed in number of tips. (2) number of structured traits (global/local). (3) standard deviation of normal variates added to structured traits. (4) includes the structured traits.

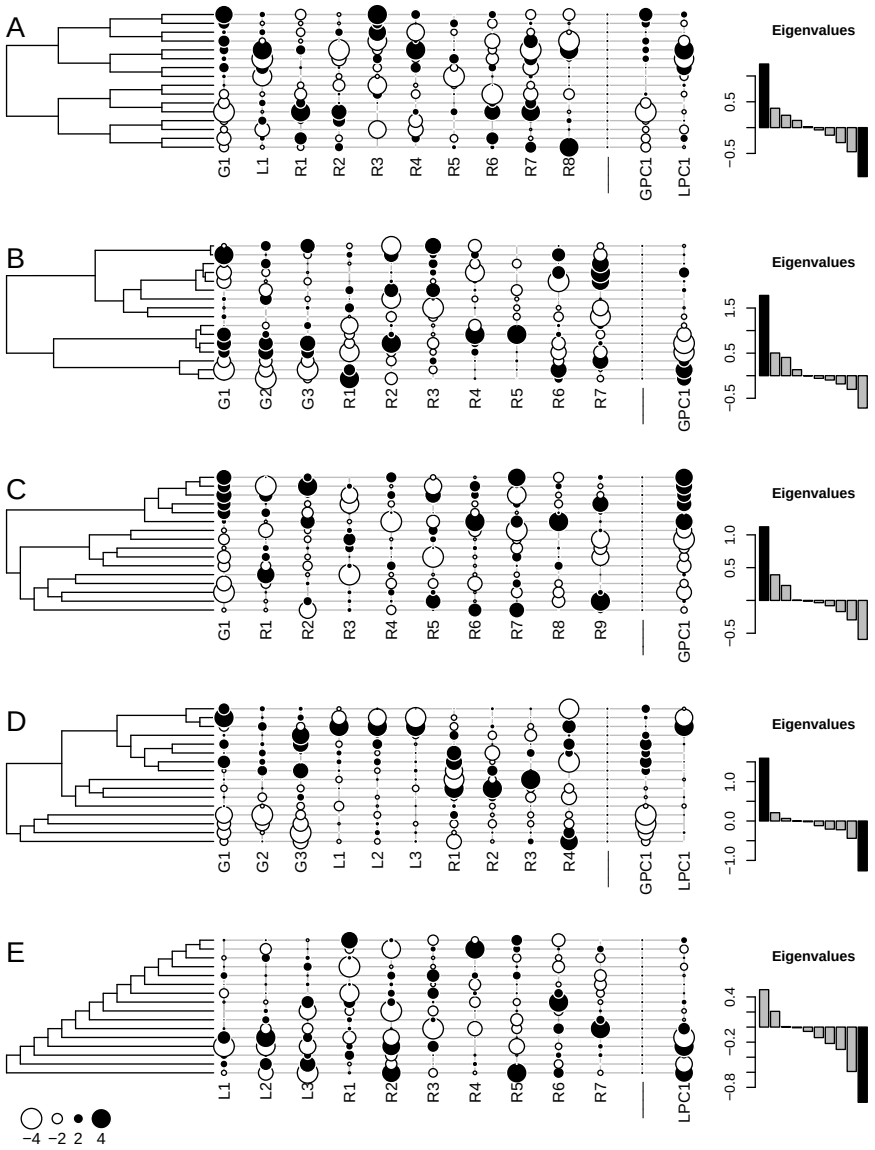


Fig. 1. Examples of simulated datasets and corresponding pPCA. Circles represent the values of simulated traits and of the retained principal components of pPCA on the right (negative values in white, positive ones in black). Gi: i^{th} global structure. Li: i^{th} local structure. Ri: i^{th} random trait. GPC1: first global principal component of pPCA (*i.e.*, associated to the largest positive eigenvalue). LPC1: first global local component of pPCA (*i.e.*, associated to the largest negative eigenvalue). Right screeplots show pPCA eigenvalues, with retained structures in black. (A) Symmetric tree; random noise added structures ('noise') equaled 0.5. (B) Random clustering of tips; noise=1. (C) Yule model; noise=0.5. (D) biased model; noise=0.75. (E) Assymetric tree; noise=1.

3.2 Data analysis

Each dataset was analyzed by a pPCA using the matrix **W** (5) to model phylogenetic proximities. Because of the huge number of simulated datasets, we had to reduce each analysis to a measure of how well the structured traits were retrieved by the analysis. This was achieved by computing the absolute value of Spearman's ranks correlation ($|\rho|$) between the structured traits and the principal components of pPCA. In each case, only the first global and/or local components (*i.e.*, those associated to the most extreme eigenvalues) were retained. When the dataset included several structured traits, the values of $|\rho|$ were averaged per type of structure (global or local). Hence, we obtained one or two $|\rho|$ per simulated dataset, roughly indicating whether the analysis performed well ($|\rho|$ close to one), or not ($|\rho|$ close to zero).

Then, we built a linear model to assess the overall performance of the method, and the influence of different parameters on the performance of the pPCA. Because $|\rho|$ varies from 0 to 1 and not on $\mathbb{R}$, we used $\text{logit}(|\rho|) = \log\frac{|\rho|}{1-|\rho|}$ as the explained variable. When interpreting coefficients of the model, predictions $\hat{\mu}$ were re-transposed onto the $|\rho|$ scale, that is, replacing $\hat{\mu}$ by $\frac{1}{1+e^{-\hat{\mu}}}$. The explanatory variables were: the type of tree (factor 'tree', the biased model being the intercept), the type of structuring (factor 'strutype', with level 'global' at the intercept), the number of tips ('ntips', intercept=16), the total number of traits ('ntraits', intercept=10), the standard deviation of the random noise added to structured variables ('noise', intercept=0.5), and the number of structures (factor 'nstruc', 1 being intercept).

3.3 Results

All explanatory variables had a very significant effect on the response variable (Appendix B, Table B.1), which was trivial given the huge number of observations. All coefficients of the model (Appendix B, Table B.2) can therefore be interpreted quantitatively. Nonetheless, these interpretations should be tempered by the fact that this model explained only 30% of the total variance. The average $|\rho|$ was satisfying ($\text{IC}_{\alpha=0.01} = [0.667; 0.669]$), showing that phylogenetic structures were globally well retrieved by pPCA. The strongest effect was by far that of the type of structure: global patterns were more easily retrieved than local structures, with a difference of 0.29 in predicted $|\rho|$ (later denoted $\Delta_{\tilde{|\rho|}}$). Another important effect was the number of random traits (keeping the number of structured traits constant): patterns were slightly more difficult to detect when a lot of random traits were present ($\Delta_{\tilde{|\rho|}} = 0.16$ between 10 and 50 traits). Among weaker effects, the results of pPCA seemed

better in larger trees: $\Delta_{|\tilde{\rho}|}$ between trees with 16 and 128 tips was 0.11. The 'strength' of the structured traits also had an effect on the results: patterns that were generated using larger amounts of random noise were more difficult to retrieve ($\Delta_{|\tilde{\rho}|} = 0.10$ between noise of 0.5 and 1). Lastly, the tree type and the number of structured traits did not seem to influence much the ability of the analysis to identify phylogenetic patterns.

4 Discussion

In this paper, we presented a multivariate method to investigate the phylogenetic patterns in a set of quantitative traits. This method, the phylogenetic principal component analysis (pPCA), is adapted from a spatial method (Jombart et al., 2008), and relies on the Moran's I, an autocorrelation index that can be used to measure the non-independence in a series of observations of a trait with respect to the phylogenetic proximity among taxa. The pPCA was evaluated through extensive simulations, whose results shall now be discussed.

The first result emerging from this sensitivity study was that pPCA performed well to retrieve phylogenetic structures, even in some cases where only 1 out of 50 traits was phylogenetically structured. Among the different parameters under study, it appeared that the type of structure first influenced the results of pPCA: global patterns were more easily identified than local structures. This can be explained by the fact that the distribution of Moran's I is not symmetric around the null value. Given a phylogenetic proximity matrix, the range of Moran's I is fixed (de Jong et al., 1984), and it appeared from our simulations that the minimum values of I were closer to the null value than maximum values (results not shown). As the discovery of phylogenetic patterns by pPCA relies in part on finding synthetic traits with extreme values of I, global structures (associated to large positive I) would be more easily detected than local structures (associated to large negative I). Interestingly, the method seemed rather insensitive to the type of tree, meaning the pPCA can be used with virtually any kind of tree.

Another point is the method used to simulated phylogenetic structures. In this study, we used a yet uncommon approach, which relies in finding eigenvectors of a phylogenetic proximity matrix. The obtained variables model different global and local structures, that can be combined to form complex phylogenetic patterns. This method had several advantages over more classical approaches like trait simulation under a Brownian model or an OU process. First, the algorithm we used was much faster than current implementations of these processes, which permitted to test the effects of a wide range of parameters. The sensitivity study could not have been conducted, or only on a

very small set of parameters, using trait simulation under Brownian or OU models. Second, while it was clear that global structures can be modeled using a Brownian or an OU process, it seemed more difficult to simulate local patterns using these approaches. Lastly, we wanted to evaluate the pPCA in the most frequent case where branch lengths are unknown (or not known with accuracy): contrary to Brownian and OU processes, the approach we used could simulate phylogenetic structures without defining branch lengths.

A parameter that could have been investigated in this paper was the influence of the phylogenetic proximity used by the pPCA. As a short answer, we could argue that the proximity measure we used in this paper proved more powerful to detect phylogenetic structures than several other common measures (Pavoine et al., 2008). Another answer would be that the phylogenetic structures viewed through different phylogenetic proximities is an open question that could motivate whole studies. Such a study was clearly out of the scope of the present paper. Nevertheless, there is a way of assessing whether the phylogenetic proximities used in a pPCA are adapted to a given dataset: one can perform Moran's I test using this proximity matrix to assess whether significant structures are detected, and then use the relevant proximities in pPCA. In this context, our sensitivity study was performed in favorable conditions for the pPCA, since the proximity matrix was by definition the most adapted to data. However, the results obtained by Pavoine et al. (2008) clearly showed that the proximity matrix underlying the test of Abouheif (1999) was well-suited to identifying various phylogenetic structuring.

To conclude, this paper illustrates the large intersection between spatial and phylogenetic methodological issues. Of course, spatial and phylogenetic patterns are generated by very different processes, but the mathematical tools that can be used to measure and model these patterns are deeply related. This is because both rely on the concept of autocorrelation, which could be formulated out of a particular context, as the non-independence among observations of a variable given a set of underlying proximities. This would explain why several spatial methods developed in Ecology were adapted to Phylogenetics (Cheverud et al., 1985; Gittleman and Kot, 1990; Diniz-Filho et al., 1998; Desdevises et al., 2003; Giannini, 2003). Originally, spatial autocorrelation was perceived by ecologists as problem because it prevented the use of standard statistical tools, by violating the assumption of independence among observations. However, this 'annoying feature' turned out to be a fecund paradigm as ecologists got interested in studying spatial patterns rather than simply trying to get rid of them. Rewording Legendre (1993), we may now also ask the question: *phylogenetic inertia, trouble or new paradigm?*

5 Acknowledgements

We are grateful to the CCIN2P3 for providing access to their computers, and particularly to Simon Penel for his help.

References

Abouheif, E., 1999. A method for testing the assumption of phylogenetic independence in comparative data. Evolutionary Ecology Research 1, 895–909.

Bolker, B., Butler, M., Cowan, P., de Vienne, D., Jombart, T., Kembel, S., Orme, D., Paradis, E., Zwickl, D., 2007. phylobase: Base package for phylogenetic structures and comparative data. R package version 0.3.
URL `http://phylobase.R-forge.R-project.org`

Bortolussi, N., Durand, E., Blum, M., François, O., 2006. apTreeshape: statistical analysis of phylogenetic tree shape. Bioinformatics 22, 363–364.

Chessel, D., Dufour, A.-B., Thioulouse, J., 2004. The ade4 package-I- one-table methods. R News 4, 5–10.

Cheverud, J. M., Dow, M. M., 1985. An autocorrelation analysis of genetic variation due to lineal fission in social groups of Rhesus macaques. American Journal of Phyisical Anthropology 67, 113–121.

Cheverud, J. M., Dow, M. M., Leutenegger, W., 1985. The quantitative assessment of phylogentic constaints in comparative analyses: sexual dimorphism in body weights among primates. Evolution 39, 1335–1351.

Cliff, A. D., Ord, J. K., 1973. Spatial autocorrelation. Pion, London.

Cliff, A. D., Ord, J. K., 1981. Spatial Processes. Model & Applications. Pion, London.

de Jong, P., Sprenger, C., van Veen, F., 1984. On extreme values of Moran's I and Geary's c. Geographical Analysis 16, 17–24.

Desdevises, Y., Legendre, L., Azouzi, L., Morand, S., 2003. Quantifying phylogenetically structured environmental variation. Evolution 57 (11), 2647–2652.

Diniz-Filho, J. A. F., de Sant'Ana, C. E. R., Bini, L. M., 1998. An eigenvector method for estimating phylogenetic inertia. Evolution 52, 1247–1262.

Dray, S., Dufour, A.-B., Chessel, D., 2007. The ade4 package - II: Two-table and K-table methods. R News 7, 47–54.

Dray, S., Legendre, P., Peres-Neto, P., 2006. Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbours matrices (PCNM). Ecological Modelling 196, 483–493.

Dray, S., Saïd, S., Debias, F., Chessel, D., 2008. Spatial ordination of vegetation data using a generalization of Wartenberg's multivariate spatial correlation. Journal of Vegetation Science 19, 45–56.

Felsenstein, J., 1985. Phylogenies and the comparative method. The American Naturalist 125, 1–15.

Giannini, N. P., 2003. Canonical phylogenetic ordination. Systematic Biology 52, 684–695.

Gittleman, J. L., Kot, M., 1990. Adaptation: statistics and a null model for estimating phylogenetic effects. Systematic Zoology 39, 227–241.

Gould, S. J., Lewontin, R. C., 1979. The spandrels of san marco and the panglossian paradigm: a critique of the adaptationist program. Proceedins of the Royal Society of London, Series B 205, 581–598.

Grafen, A., 1989. The phylogenetic regression. Philosophical Transactions of the Royal Society of London Series B - Biology 326, 119–157.

Gregory, W. K., 1913. Convergence and applied phenomena in the mammalia. Report of the British Association for the Advancement of Science IV, 525–526.

Griffith, D. A., 1996. Spatial autocorrelation and eigenfunctions of the geographic weights matrix accompanying geo-referenced data. The Canadian Geographer 40, 351–367.

Hansen, T. F., Martins, E. P., 1996. Translating between microevolutionary process and macroevolutionary patterns: the correlation structure of interspecific data. Evolution 50 (4), 1404–1417.

Harvey, P. H., Pagel, M., 1991. The Comparative Method in Evolutionary Biology. Oxford University Press.

Jombart, T., Devillard, S., Dufour, A.-B., Pontier, D., 2008. Revealing cryptic spatial patterns in genetic variability by a new multivariate method. Heredity 101, 92–103.

Kirkpatrick, M., Slatkin, M., 1993. Searching for evolutionary patterns in the shape of a phylogenetic tree. Evolution 47, 1171–1181.

Legendre, P., 1993. Spatial autocorrelation: trouble or new paradigm? Ecology 74, 1659–1673.

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

Moran, P. A. P., 1950. Notes on continuous stochastic phenomena. Biometrika 37, 17–23.

Osborn, H. F., 1917. Heritage and habitus. Science 45, 660–661.

Paradis, E., Claude, J., Strimmer, K., 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20, 289–290.

Pavoine, S., Ollier, S., Pontier, D., Chessel, D., 2008. Testing for phylogenetic signal in life history variable: Abouheif's test revisited. Theoretical Population Biology 73, 79–91.

Peres-Neto, P., 2006. A unified strategy for estimating and controlling spatial, temporal and phylogenetic autocorrelation in ecological models. Oecologica Brasiliensis 10, 105–119.

R Development Core Team, 2008. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
URL `http://www.R-project.org`

Rao, D. R., Toutenburg, H., 1999. Linear Models: least squares and alternatives. Springer, Ch. The generalized linear regression model, pp. 89–110.

Rohlf, F. J., 2001. Comparative methods for the analysis of continuous variables: geometric interpretations. Evolution 55 (11), 2143–2160.

Rohlf, F. J., 2006. A comment on phylogenetic correction. Evolution 60, 1509–1515.

A Appendix A: pPCA using Abouheif's matrix (A)

In this section, we demonstrate that the quantity optimized by the pPCA when the condition $w_{ii} = 0$ is violated (equation 2) is no longer interpretable as the product of variance and Moran's I (6). As a particular case, the matrix underlying Abouheif's test (**A**, (4)) cannot be directly used in lieu of **W**. As stated in the text, the extremums of (6) are given by the eigenvectors of the symmetric matrix $\frac{1}{2n}\mathbf{X}^T(\mathbf{W}+\mathbf{W}^T)\mathbf{X}$. Denoting $\mathbf{u}$ one of these eigenvectors, and λ the associated eigenvalue, we have:

$$\frac{1}{2n}\mathbf{u}^T\mathbf{X}^T(\mathbf{W}+\mathbf{W}^T)\mathbf{X}\mathbf{u} = \lambda \tag{A.1}$$

$$\frac{1}{n}\mathbf{u}^T\mathbf{X}^T\mathbf{W}\mathbf{X}\mathbf{u} = \lambda \tag{A.2}$$

$$\frac{1}{n}(\mathbf{X}\mathbf{u})^T\mathbf{W}\mathbf{X}\mathbf{u} = \lambda \tag{A.3}$$

Which, using equation 1, amounts to:

$$\mathrm{var}(\mathbf{X}\mathbf{u})I(\mathbf{X}\mathbf{u}) = \lambda \tag{A.4}$$

Hence, each eigenvalue λ equals the product of the variance and of Moran's I of the corresponding synthetic variable $\mathbf{X}\mathbf{u}$.

Now, let us define a matrix of phylogenetic proximity $\mathbf{Y} = [y_{ij}]$ having the same off-diagonal terms as **W**, but with $y_{ii} \neq 0$. We denote $\mathbf{D} = [d_{ij}]$ $(i, j = 1, \ldots, n)$ the diagonal matrix defined as $d_{ii} = y_{ii}$ and $d_{ij} = 0\ \forall i \neq j$. Performing a pPCA using **Y** to model phylogenetic proximities amounts to replace **W** by **Y** in (A.1). We denote $\mathbf{v}$ the obtained eigenvectors and γ the corresponding eigenvalue:

$$\frac{1}{n}(\mathbf{X}\mathbf{v})^T\mathbf{Y}\mathbf{X}\mathbf{v} = \gamma \tag{A.5}$$

$$\frac{1}{n}(\mathbf{X}\mathbf{v})^T(\mathbf{Y}-\mathbf{D})\mathbf{X}\mathbf{v} + \frac{1}{n}(\mathbf{X}\mathbf{v})^T\mathbf{D}\mathbf{X}\mathbf{v} = \gamma \tag{A.6}$$

$$\frac{1}{n}(\mathbf{X}\mathbf{v})^T\mathbf{W}\mathbf{X}\mathbf{v} + \frac{1}{n}\|\mathbf{X}\mathbf{v}\|^2_{\mathbf{D}} = \gamma \tag{A.7}$$

$$\mathrm{var}(\mathbf{X}\mathbf{v})I(\mathbf{X}\mathbf{v}) + \frac{1}{n}\sum_{i=1}^{n} y_{ii}[\mathbf{X}\mathbf{v}]_i^2 = \gamma \tag{A.8}$$

where $[\mathbf{X}\mathbf{v}]_i$ is the i^{th} component of the synthetic variable $\mathbf{X}\mathbf{v}$. Hence, the eigenvalues of a pPCA computed with matrix **Y** no longer finds the extremum

of (6). Indeed, a positive term ($\frac{1}{n}\sum_{i=1}^{n} y_{ii}[\mathbf{Xv}]_i^2$) is added to the function whose extremums are retreived. Note that when $\mathbf{Y} = \mathbf{A}$, this quantity in itself may be interesting independently from pPCA. Because the terms a_{ii} are weights measuring the 'originality' of taxon i with respect to the phylogeny (Pavoine et al., 2008), and because $\mathbf{Xv}$ is centred, the term $\sum_{i=1}^{n} a_{ii}[\mathbf{Xv}]_i^2$ represents the variance of the synthetic variable weighted by the originality of the taxa. It follows that the extreme values of (A.8) are much more difficult to interprete in biological terms than those of (A.4). This is made even more difficult as we do not know the respective contributions of the terms of (A.8).

B Appendix B: tables of the analysis of the model

This appendix presents two tables corresponding to the analysis of the model described in section 3.2.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
fac.tree[1]	4	4499	1129	1637.81	$< 2.2e^{-16}$
fac.strutype[1]	1	80838	80838	36699.06	$< 2.2e^{-16}$
ntips[3]	1	8009	8009	11661.69	$< 2.2e^{-16}$
ntraits[4]	1	21225	21225	30907.10	$< 2.2e^{-16}$
noise[5]	1	5824	5824	8481.20	$< 2.2e^{-16}$
fac.nstruc[6]	1	280	280	653.58	$< 2.2e^{-16}$
Residuals	215990	92414	0.4279		

Table B.1
ANOVA of the model. Factor are preceded by 'fac'. (1) type of tree. (2) type of structure (global or local). (3) number of tips. (4) total number of traits. (5) number of structured traits (1 or 3). (6) standard deviation of the random noise added to the structured traits.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.554	$4.496e^{-03}$	345.727	$< 2e^{-16}$
fac.tree-clust[1]	0.037	$4.451e^{-03}$	8.307	$< 2e^{-16}$
fac.tree-comb[2]	0.249	$4.451e^{-03}$	55.995	$< 2e^{-16}$
fac.tree-sym[3]	0.366	$4.451e^{-03}$	82.222	$< 2e^{-16}$
fac.tree-yule[4]	0.035	$4.451e^{-03}$	7.912	$< 2.55e^{-15}$
fac.strutype-local[5]	−1.224	$2.815e^{-03}$	−434.665	$< 2e^{-16}$
ntips[6]	0.004	$2.846e^{-03}$	136.812	$< 2e^{-16}$
ntraits[7]	−0.018	$8.281e^{-03}$	−222.726	$< 2e^{-16}$
fac.nstruc-3[8]	0.072	$2.815e^{-03}$	25.565	$< 2e^{-16}$
noise[9]	−0.804	$6.895e^{-03}$	−116.673	$< 2e^{-16}$

Table B.2
Coefficients of the model. Factor are preceded by 'fac', followed by the levels. (1) trees obtained by random clustering of tips. (2) comb-like model (completely asymmetric trees). (3) completely symmetric trees. (4) Yule model. (5) local phylogenetic structure. (6) number of tips. (7) total number of traits. (8) number of structured traits (1 or 3). (9) standard deviation of the random noise added to the structured traits.

6.3 Discussion

6.3.1 Illustration

Les données nous permettant d'illustrer la méthodes ont été publiées dans Bauwens & Diaz-Uriarte (1997) et constituent le jeu de données `lizards` du package *ade4* (Chessel *et al.*, 2004; Dray *et al.*, 2007), décrit dans une fiche thématique en français (`http://pbil.univ-lyon1.fr/R/pps/pps063.pdf`). La gestion des données compositionnelles est effectuée par le package *phylobase* (Bolker *et al.*, 2007), un effort de développement coopératif émergent du « R hackathon on comparative methods », organisé par NESCent en décembre 2007 à Durham, NC, USA. Cet événement visait à fédérer les principaux packages du logiciel R (R Development Core Team, 2008) implémentant des méthodes comparatives, en définissant des formats de données canoniques et en facilitant la manipulation de ces objets. Ayant eu la chance de participer à ce hackathon, j'ai contribué depuis lors à la création du package *phylobase*, principalement au niveau de la définition des classes d'objets et de leurs représentations graphiques. Le package *phylobase* offre de multiples avantages pour le traitement des données comparatives, qu'il serait trop long de détailler ici. Il serait par ailleurs mal venu de présenter dans le bilan d'un travail personnel le fruit d'un travail éminemment collectif. On se contentera donc de soutenir, avec quelques bonnes raisons, que les formats d'objets définis dans *phylobase* formeront probablement les fondations sur lesquelles reposeront les prochaines implémentations de méthodes comparatives dans R. Le principal type d'objet revêt la classe `phylo4d`, pour phylogénie (`phylo`) et données (`d`) en classe S4 (`4`), c'est-à-dire en classe formelle du language R.

Dans cette optique, il était indispensable d'implémenter la pPCA directement pour des objets de la classe `phylo4d`. Actuellement, la méthode est sommairement implémentée par trois fonctions non documentées :

- `ppca`, la fonction effectuant les calculs et créant un objet de la classe `ppca`
- `plot.ppca`, la fonction graphique associée aux objets ppca
- `proxphylo4`, une fonction appelée par `ppca` qui calcule différentes matrices de proximités phylogénétiques pour des objets `phylo4` (représentation des arbres dans *phylobase*).

Ces fonctions intégreront le futur package *adephylo*, qui devrait être disponible publiquement à la fin de l'année 2008.

On commence par charger les packages requis, le code source de ces fonctions et les données :

```
> library(phylobase)
> library(ade4)
> source("ppca.R")
> source("prox.R")
> data(lizards)
> names(lizards)
```

```
[1] "traits" "hprA"   "hprB"
```

```
> dim(lizards$traits)
```

```
[1] 18  8
```

```
> names(lizards$traits)
```

```
[1] "mean.L"   "matur.L"  "max.L"    "hatch.L"  "hatch.m"  "clutch.S" "age.mat"
[8] "clutch.F"
```

L'objet `lizards` est une liste contenant deux phylogénies (`hprA` et `hprB`) basées respectivement sur des critères moléculaires et sur des critères morphologiques. L'élément `traits` est un `data.frame` contenant 8 traits biologiques pour chacun des 18 taxons. Ces traits sont :

- `mean.L` : longueur moyenne adulte des femelles (en mm)
- `matur.L` : longueur moyenne des femelles à la maturité (en mm)
- `max.L` : longueur maximale des femelles (en mm)
- `hatch.L` : longueur moyenne des petits à l'éclosion (en mm)
- `hatch.m` : masse des petits à l'éclosion (en g)
- `clutch.S` : taille de portée (en nombre d'oeufs)
- `age.mat` : âge moyen à maturité (en nombre de mois d'activité, *i.e.* hors hibernation)
- `clutch.F` : nombre de portées par année.

Une première observation de la nature de ces données montre qu'on peut s'attendre à ce qu'un effet « taille » masque la part intéressante de la variabilité des données. On peut s'en convaincre par une ACP normée (seul le premier axe est retenu) (Fig. 6.1).

```
> liz.pca <- dudi.pca(lizards$traits, scannf = FALSE, nf = 2)
> s.corcircle(-liz.pca$co, clab = 1.4)
> add.scatter.eig(liz.pca$eig, 1, 1, 1, csub = 1.2)
```

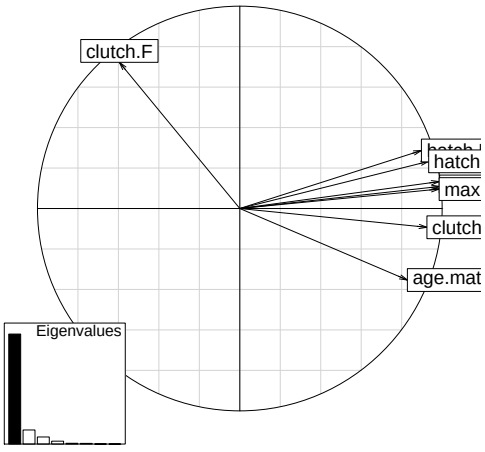


Fig. 6.1: Analyse en composantes principales (ACP) normée des données `lizards`. Cercle des corrélations et valeurs propres.

Et en effet, si l'on représente les données sur la phylogénie, on constate que la seule information visible est qu'un taxon est plus gros que tous les autres (Fig. 6.2). On crée au passage un objet `phylo4d` contenant la phylogénie moléculaire (`liz.tre`), les données (`lizards$traits`) et la première composante principale de l'ACP normée (`liz.pca$li[,1]`) :

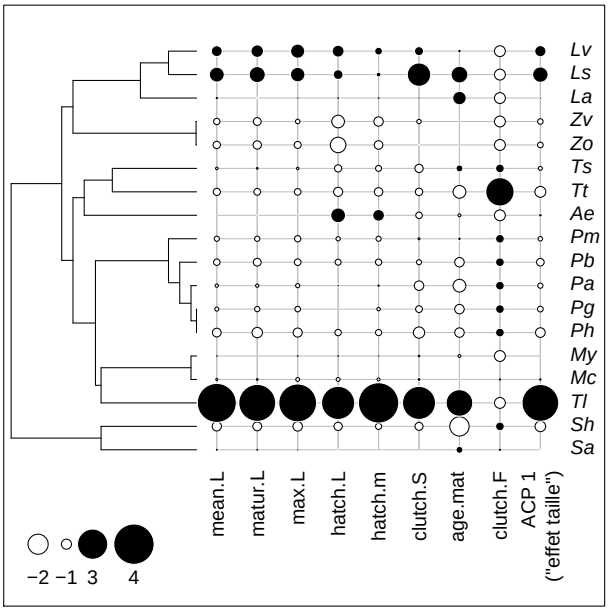


FIG. 6.2: Représentation graphique des données `lizards` et de la première composante principale de leur ACP normée. Les variables sont représentées centrées et normées.

```
> liz.tre <- read.tree(tex = lizards$hprA)
> liz.4d <- phylo4d(liz.tre, cbind(lizards$traits, -liz.pca$li[,
+     1]))
> par(mar = rep(0.1, 4))
> plot(liz.4d, var.lab = c(names(lizards$traits), "ACP 1\n(\"effet taille\")"),
+     show.node = FALSE, cex.lab = 1.2)
```

L'effet « taille » donc est enlevé par projection des données sur le sous-espace orthogonal à la première composante principale de l'ACP (objet `temp`). Une alternative aurait été d'utiliser la taille moyenne des femelles adultes (`mean.L`) en remplacement de la première composante principale. On crée un objet `phylo4d` qui contiendra ces données ainsi que la phylogénie (moléculaire) associée.

```
> liz.res <- pcaivortho(liz.pca, liz.pca$l1[, 1], scannf = FALSE)
> round(cor(X, liz.pca$li[, 1]), 14)
```

```
          [,1]
mean.L       0
matur.L      0
max.L        0
hatch.L      0
hatch.m      0
clutch.S     0
age.mat      0
clutch.F     0
```

```
> liz.newd4 <- phylo4d(liz.tre, liz.res$tab)
> par(mar = rep(0.1, 4))
> plot(liz.newd4, show.node = FALSE, cex.lab = 1.2)
```

Les données dont on a supprimé l'effet « taille » ont maintenant une signification moins triviale (FIG. 6.3). La pPCA permet d'en clarifier le message. La version de l'analyse ici effectuée est celle proposée dans l'article, à ceci près que les données ne sont pas normées : en fait, elles avaient déjà été normées avant d'enlever "l'effet taille". Cette opération modifie profondément les normes des différents vecteurs, et il est normal qu'une variable qui ne contenait presque que l'effet taille voit sa variance fortement diminuée après que cet effet ait été enlevé.

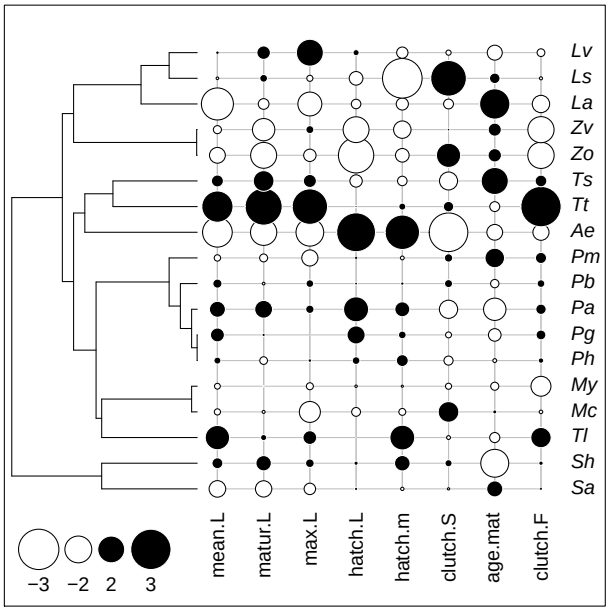


FIG. 6.3: Représentation graphique des données `lizards` après avoir enlevé "l'effet taille" par régression sur la première composante principale de l'ACP normée. Les variables sont représentées centrées et normées.

```
> liz.ppca <- ppca(liz.newd4, scale = FALSE, scannf = FALSE, nfposi = 1,
+     nfnega = 1)
> names(liz.ppca)
```

```
[1] "eig"    "nfposi" "nfnega" "c1"     "li"     "ls"     "as"     "call"
[9] "tre"
```

```
> tempcol <- rep("grey", 7)
> tempcol[c(1, 7)] <- "black"
> barplot(liz.ppca$eig, main = "Valeurs propres de la pPCA", cex.main = 1.8,
+     col = tempcol)
```

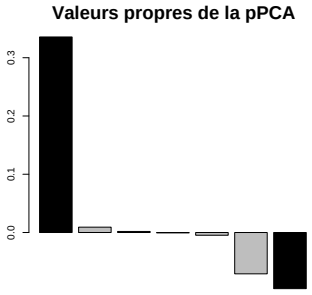


FIG. 6.4: Graphe des valeurs propres de pPCA des données `lizards`. Les structures positivement autocorrélées sont à gauche, celles négativement corrélées sont à droite. Les valeurs propres correspondant aux structures retenues sont indiquées en noir.

La pPCA met en valeur une structure globale forte, et sans doute une structure locale plus faible (FIG. 6.4). On peut représenter ces structures graphiquement :

```
> par(mar = rep(0.1, 4))
> plot(liz.ppca, ratio.tree = 0.7, var.lab = c("Composante \nglobale 1",
+     "Composante \nlocale 1"), cex.lab = 1.2)
```

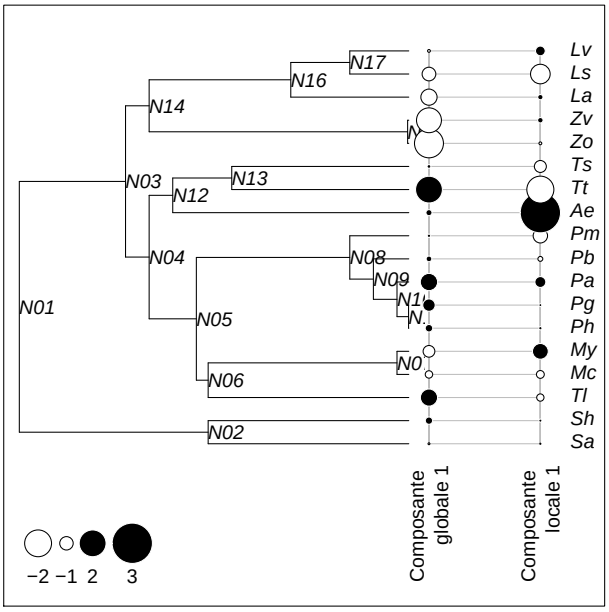
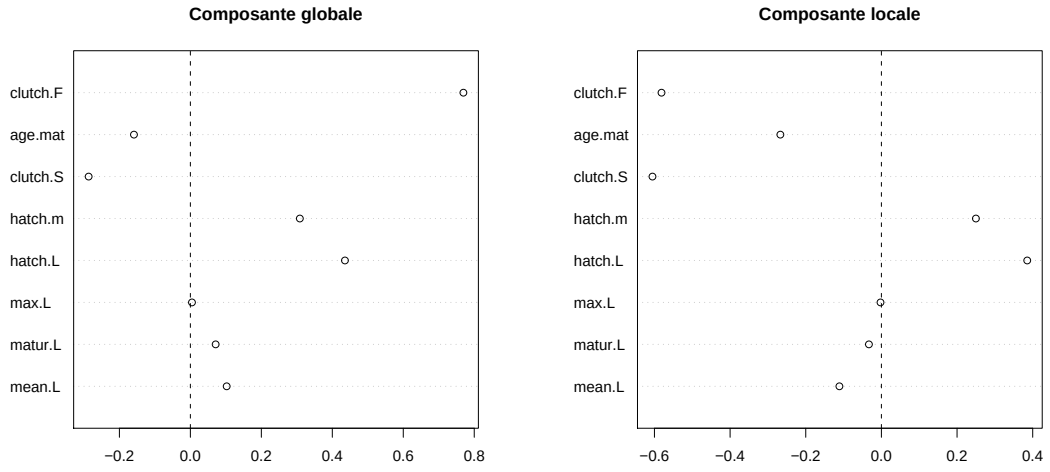


FIG. 6.5: Premières composantes principales globales et locales de la pPCA des données `lizards`. Les variables sont centrées et normées.

Pour interpréter ces structures, il est indispensable d'examiner les contributions des variables :

```
> dotchart(liz.ppca$c1[, 1], lab = rownames(liz.ppca$c1), main = "Composante globale")
> abline(v = 0, lty = 2)

> dotchart(liz.ppca$c1[, 2], lab = rownames(liz.ppca$c1), main = "Composante locale")
> abline(v = 0, lty = 2)
```



(a) Contributions des variables à la première composante globale

(b) Contributions des variables à la première composante locale

FIG. 6.6: Contributions des variables aux composantes principales globales et locales de la pPCA des données `lizards`.

La première composante globale de la pPCA (Fig. 6.5, gauche) oppose les descendants du noeud N14 (scores fortement négatifs) au reste de l'arbre. La contribution des variables à cette structure (Fig. 6.6a) indique que les descendants de N14 se reproduisent peu souvent (`clutch.F`$---$), et produisent des oeufs nombreux (`clutch.S`$+$) et de petite taille (`hatch.m`$--$, `hatch.L`$--$), par rapport aux autres taxons. Une explication possible de cette structure serait que ces espèces sont contraintes par l'environnement à un nombre d'événements reproductifs limités, et produisent un grand nombre d'oeufs dont la survie individuelle est faible. Notons qu'à l'instar de Bauwens & Diaz-Uriarte (1997), on observe un compromis évolutif entre nombre d'oeufs et taille des nouveaux-nés.

La première composante locale de la pPCA (Fig. 6.5, droite) met en exergue deux stratégies différentes entre taxons proches (*Tt* : *Takydromus tachydromoides* ; *Ae* : *Acanthodactylus erythrurus*). Les contributions des variables (Fig. 6.6b) identifient *T. tachydromoides* comme une espèce produisant de nombreux descendants (`clutch.S`$++$, `clutch.F`$++$) de petite taille (`hatch.m`$-$,`hatch.L`$--$) par opposition à *A. erythrurus*, qui produit peu de descendants (`clutch.S`$--$, `clutch.F`$--$) de grande taille (`hatch.m`$+$,`hatch.L`$++$). Encore une fois, le compromis évolutif entre nombre d'oeufs et taille des nouveaux-nés est observé.

Quoique simple, cet exemple illustre l'intérêt majeur de la pPCA, qui réside dans l'étude plutôt que dans l'élimination de l'autocorrélation phylogénétique au sein de données multivariées. Plus que de révéler un compromis évolutif déjà connu entre taille et nombre d'oeufs, cette approche permet ici d'identifier deux grandes stratégies biodémographiques (structure globale, (Fig. 6.5, gauche)), et met en valeur deux stratégies évolutives particulièrement différentes entre deux taxons proches (structure locale, (Fig. 6.5, droite)).

6.3.2 La similarité d'Abouheif

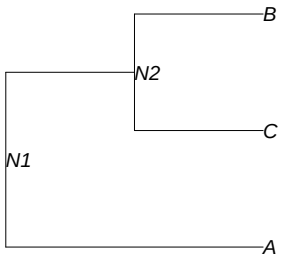
Comme nous l'avons mentionné dans l'article, la mesure de proximité phylogénétique sous-tendue par le test d'Abouheif n'est qu'une mesure parmi d'autres pouvant être utilisée par la pPCA. Notre préférence est à la fois un héritage lié aux conditions d'émergence de la méthode, et un choix motivé par les bons résultats obtenus par Pavoine *et al.* (2008) pour tester l'existence de structures issues d'un mouvement brownien. Néanmoins, il sera sans doute nécessaire de s'interroger plus avant sur la pertinence de ce choix. Dans l'article présenté plus haut, nous avons déjà critiqué la matrice d'Abouheif, dont il est préférable de supprimer la diagonale pour que les résultats de la pPCA soient interprétables en termes de variance et d'autocorrélation. Nous soulevons ici une autre critique pouvant être faite à cette mesure de proximité phylogénétique.

La matrice d'Abouheif sous-tend une distance phylogénétique entre taxons qui peut être définie comme l'inverse des similarités définies dans (Eqn. 6.4) :

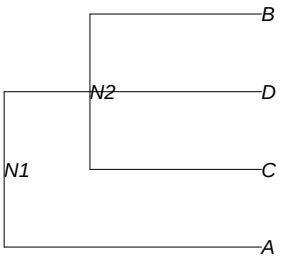
$$d_{AB} = \frac{1}{a_{AB}} = \prod_{p \in P_{AB}} dd_p \tag{6.5}$$

où P_{AB} est l'ensemble de noeuds internes reliant les feuilles A et B, et dd_p est le nombre de descendants directs du noeud p. On peut se faire une idée du comportement de cette distance en calculant d_{AB} pour différents arbres (FIG. 6.7)

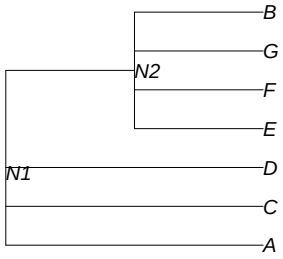
```
> library(ape)
> library(phylobase)
> tre1 <- as(read.tree(text = "(A,(C,B));"), "phylo4")
> plot(tre1, show.node = T, cex = 2)


> tre2 <- as(read.tree(text = "(A,(C,D,B));"), "phylo4")
> plot(tre2, show.node = T, cex = 2)


> tre3 <- as(read.tree(text = "(A,C,D,(E,F,G,B));"), "phylo4")
> plot(tre3, show.node = T, cex = 2)


> tre4 <- as(read.tree(text = "(A,(C,(D,(E,B))));"), "phylo4")
> plot(tre4, show.node = T, cex = 2)
```



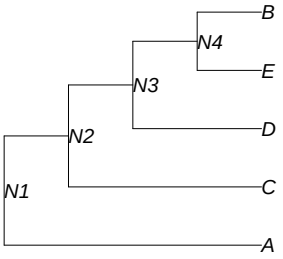
(a) $d_{AB} = 2 \times 2 = 4$



(b) $d_{AB} = 2 \times 3 = 6$



(c) $d_{AB} = 4 \times 4 = 16$



(d) $d_{AB} = 2 \times 2 \times 2 \times 2 = 16$

FIG. 6.7: Quatre phylogénies simples illustrant la distance d'Abouheif

La position des taxons A et B ne change pas dans les trois premiers arbres : seuls changent les nombres de descendants par noeud (FIG. 6.7a-c). Néanmoins, d_{AB} varie du simple au triple alors qu'on attendrait une distance identique dans les trois cas. Le problème inverse se manifeste aussi : les topologies (FIG. 6.7c) et (FIG. 6.7d) sont très différentes, et on attendrait des distances différentes entre A et B (la première étant inférieure à la seconde). Ce n'est pourtant pas le cas : les distances d'Abouheif sont égales, et ne reflètent donc pas ces différences topologiques. Il semble donc que la distance d'Abouheif possède quelques défauts.

Par ailleurs, un véritable problème se pose du point de vue inférentiel : la distance (et donc la similarité) d'Abouheif ne peut être utilisée que pour décrire de façon relative les liens phylogénétiques entre taxons d'un échantillon. En aucun cas il ne s'agit de l'estimation d'une distance absolue entre taxons. Par exemple, on peut considérer que les figures 6.7a-c sont trois représentations différentes d'une même réalité, différant seulement par l'effort d'échantillonnage. La distance d'Abouheif entre les taxons A et B, présents dans tous les cas, est tributaire de cet échantillonnage.

Malgré ces défauts manifestes, la matrice d'Abouheif semble particulièrement puissante pour détecter des structures phylogénétiques par le test de l'indice de Moran (Pavoine *et al.*, 2008). Elle s'est également avérée satisfaisante pour simuler des structures phylogénétiques globales et locales (Jombart *et al.*, en prép-b). Il semble donc important de cerner les qualités et les défauts de la matrice d'Abouheif afin de savoir dans quelle mesure elle peut être utilisée pour modéliser des structures phylogénétiques. D'une façon plus large, ces éléments soulignent le besoin d'orienter une part des prochaines réflexions vers le choix de mesures adéquates de la proximité phylogénétique.

6.3.3 Perspectives

La pPCA étend le principe de la sPCA à l'analyse de données comparatives. Elle s'intégre par là naturellement dans le flux des méthodes de statistique spatiale importées en biologie comparative. Si le fait de reconnaître des applications nouvelles à une méthode hors du champ dans lequel elle a été développée est un facteur de progrès scientifique, il est nécessaire, voire indispensable, de cerner les limites de ces extensions, et d'identifier les particularités de l'interaction entre méthode et données. Dans le cas de la pPCA, il semble en particulier essentiel de s'intéresser à la mesure des proximités phylogénétiques entre taxons. A ce titre, l'utilisation en pPCA de la proximité phylogénétique sous-jacente au test d'Abouheif (1999) semble satisfaisante dans la pratique (Pavoine *et al.*, 2008; Jombart *et al.*, en prép-b) mais pose quelques problèmes, ou du moins comporte quelques limitations sur le plan théorique.

Il reste que la pPCA offre de nouvelles perspectives pour l'analyse de données comparatives. Une des révolutions de l'écologie statistique a eu lieu lorsque les structures spatiales ont cessé d'être perçues uniquement comme une nuisance pour être reconnues comme porteuses d'information biologique intéressante (Legendre, 1993). Il est manifeste qu'un tel changement de point de vue s'impose pour la méthode comparative. La situation est d'autant plus propice à cette nouvelle approche que les outils informatiques qui y sont nécessaires sont d'ores-et-déjà disponibles. Le logiciel R inclut plusieurs packages dédiés à la gestion de données phylogénétiques et aux analyses comparatives, dont le plus brillant représentant est sans conteste *ape* (Paradis *et al.*, 2004). La gestion des données comparatives est par ailleurs assurée par le package *phylobase* (Bolker *et al.*, 2007), qui fournit un socle solide pour l'implémentation de futures méthodes. Les éléments théoriques et techniques sont donc réunis pour ouvrir le champ de l'étude des structures phylogénétiques au sein de la méthode comparative.

Chapitre 7

Généralisation : l'ordination sous contrainte d'autocorrélation

Sommaire

7.1 Introduction

Le chapitre précédent montre que le principe de l'ordination sous contrainte spatiale proposé en génétique (Jombart *et al.*, 2008) et en écologie (Dray *et al.*, 2008) peut être étendu à l'étude des structures phylogénétiques. Ceci est dû à la nature du I de Moran (Eqn. 5.1), qui repose sur la comparaison d'une variable $\mathbf{x} \in \mathbb{R}^n$ avec son vecteur lissé $\tilde{\mathbf{x}} = \mathbf{W}\mathbf{x}$, où $\mathbf{W}$ peut être une matrice de pondérations aussi bien spatiales que phylogénétiques. Dans le premier cas, on compare les valeurs de $\mathbf{x}$ observées en un point à la valeur moyenne des points voisins dans l'espace. Dans le second, on compare la valeur d'un trait $\mathbf{x}$ prise par un taxon à la valeur moyenne des taxons phylogénétiquement proches. En réalité, la matrice $\mathbf{W}$ peut inclure n'importe quelle mesure de proximité entre les objets étudiés, la seule contrainte étant que sa diagonale soit nulle et ses termes positifs et de somme unitaire par ligne. L'indice de Moran peut donc mesurer l'autocorrélation d'une variable par rapport à n'importe quelle proximité entre objets induite par une structure sous-jacente du plan expérimental : qu'il s'agisse de distribution spatiale, d'agencement temporel, ou encore de position sur une phylogénie.

L'approche sPCA (Jombart *et al.*, 2008) et plus largement, l'approche MULTISPATI (Dray *et al.*, 2008) peuvent donc être utilisées hors du contexte spatial ou phylogénétique. C'est à la formalisation et à l'illustration de ce principe que le manuscrit présenté plus loin est dédié. Puisqu'il s'agit de généralisation, son contenu est nécessairement plus dense en mathématique que les autres articles de cette thèse. Bien qu'il requiert encore nombre d'ajouts et d'améliorations avant une éventuelle soumission, ce manuscrit explicite d'ores-et-déjà le principe de l'analyse multivariée sous contrainte d'autocorrélation. Il montre comment l'indice de Moran peut être généralisé pour mesurer l'autocorrélation d'une variable pour toute mesure de proximité entre objets. Cette extension inclut par ailleurs une pondération des objets, ce qui permet notamment de mesurer l'autocorrélation des composantes principales d'analyses utilisant des poids non uniformes, telles que l'analyse des correspondances (Greenacre, 1984). Le manuscrit montre par ailleurs comment l'indice de Moran, mesure univariée d'autocorrélation, peut être étendu aux données multivariées. L'indice résultant peut servir de base à un test général de l'autocorrélation multivariée. On montre que cet indice peut être pris en défaut par un mélange d'autocorrélations positive et négative, auquel cas un test complémentaire est proposé.

7.2 Article 7 : A general framework for constrained ordinations in reduced space using Moran's I

Article en préparation.

A general framework for constrained ordinations in reduced space using Moran's *I*

Thibaut Jombart[1,2], Stéphane Dray[1], Sandrine Pavoine[3], Anne-Béatrice Dufour[1]

[1] Laboratoire de Biométrie et Biologie Evolutive (UMR-CNRS 5558); Université de Lyon, Univ. Lyon 1, 69622, Villeurbanne Cedex, France.

[2] Current Address: Laboratoire de Biométrie et Biologie Evolutive (UMR 5558); CNRS; Univ. Lyon 1, 43 bd 11 nov 1918, 69622, Villeurbanne Cedex, France.

[3] Somewhere in the Museum d'Histoire Naturelle, Paris.

Abstract

Ordination in reduced space are statistical methods summarizing multivariate data into a few synthetic variables. This is achieved by finding an orthonormal basis onto which data are represented so that the new set of coordinates (or 'scores') has maximum successsive inertia (*i.e.*, squared norm, for a given metric). However, it happens that the maximum inertia criterion is not fully satisfying, in particular when some aspects of the variability matter more than the whole variability of the data. In such cases, constrained ordinations are used: these methods provide scores of maximum inertia under a given numerical constraint. For instance, redundancy analysis and canonical correspondance analysis seek scores with maximum inertia under the constraint that these scores are linear combinations of a set of explanatory variables. In some other cases, objects belong to an underlying scheme (*e.g.*, temporal arrangement, spatial distribution, position in a phylogeny) that could induce interesting patterns in the observed variability. Such patterns result in autocorrelation of variables, that is, in the non-independence of observations given their proximity in the underlying scheme. In this paper, we provide a framework to constrain any ordination in reduced space to provide autocorrelated scores. Starting from the fact that ordination methods rely on finding the extremums of a function computing the inertia of the scores, we propose to modify this function so that both inertia and autocorrelation are taken into account. As our approach relies on the duality diagram theory, which encompasses most ordination methods, our results are very general an can be applied to any kind of multivariate dataset, made of quantitative, qualitative, of mixed variables. Examples of constrained ordinations obtained by this approach are provided in the temporal, spatial, and phylogenetic contexts. The proposed methodology is implemented in the ade4 package for the free software R. It offers a new basis for the development of constrained ordinations in reduced space.

Introduction

Ordinations in reduced space are widely used in biology to explore strongly multivariate data (Legendre & Legendre, 1998). Biological objects are typically described by many variables and these methods can efficiently summarise the variability among objects into a few synthetic variables. Basically, ordination methods achieve this goal by finding an orthonormal basis of principal axes onto which the objects scores have maximum inertia. Resulting scatterplots are used to display as much variability among objects as it is possible on a planar representation.

However, it happens that certain aspects of the variability matter more than the whole variability itself, because they correspond to an empirically meaningful information. In such cases, ordination methods can be modified in order to investigate only the relevant aspects of the variability, rather than the entire variability. These approaches, called *constrained ordinations* (Anderson & Willis, 2003), aim at providing objects scores of maximum inertia under a given constraint, like for instance that the scores should be linearly predicted by a set of explanatory variables, as in redundacy analysis (RDA, Rao, 1964), or canonical correspondence analysis (CCA, Ter Braak, 1986, 1987), or between-classes analysis (Dolédec & Chessel, 1987). Another possible constraint is that scores should be autocorrelated. In this framework, an underlying process defines proximities among objects. Examples of such cases are spatial distributions of objects, position of objects in a connection network, or on the leaves of a phylogeny. Autocorrelation arises whenever the scores taken by two objects are not independent from their proximity (Cliff & Ord, 1973, 1981). It is said to be positive (respectively negative) when close objetcs tend to have similar (repectively dissimilar) scores. And if autocorrelation measures were initially developped to detect spatial patterns, they can be extended to measure any kind of autocorrelation.

In this paper we propose a general framework for constrained ordinations using a generalization of Moran's index of autocorrelation. Our approach generalizes methods of Dray *et al.* (2008) and Jombart *et al.* (2008), both developped to inestigate spatial patterns in multivariate data. This framework uses the duality diagram theory (Escoufier, 1987; Holmes, 2006; Dray & Dufour, 2007) to constrain usual ordinations in reduced space such as Principal Component Analysis (PCA, Pearson, 1901), Correspondance Analysis (CA, Greenacre, 1984), Multiple Correspondance Analysis (MCA, Tenenhaus & Young, 1985), or the analysis of Hill & Smith (1976) for mixed variables, to yield autocorrelated scores. It is thus applicable to quantitative and qualitative data, or even to a mixture of both types of data.

To begin, we present a generalization of Moran's autocorrelation index (I, Moran, 1948, 1950; Cliff & Ord, 1981): i) showing that I can handle any measure of similarity among objects in lieu of usual binary spatial weightings ii) using inertia instead of the variance as a scaling factor and iii)

extending formula to the multivariate case. Resulting statistics are used to test autocorrelation in multivariate data through Monte Carlo procedures. Like in the spatial context (Thioulouse *et al.*, 1995), we call *global* and *local* patterns structures respectively issued from positive and negative autocorrelation.

Whenever global and/or local patterns are detected, we show how it is possible to constrain usual ordination methods to yield autocorrelated scores. Resulting analyses provide synthetic scores that reflect both variability and autocorrelation. Positive eigenvalues correspond to scores with large inertia and positive autocorrelation (global patterns), while negative eigenvalues indicate scores with large inertia and negative autocorrelation (local patterns). This methodoly is illustrated within different contexts, using simulated or empirical datasets. Our approach is implemented in the `ade4` package (Chessel *et al.*, 2004; Dray *et al.*, 2007; Dray & Dufour, 2007) of the free software `R` (R Development Core Team, 2008) .

Generalized Moran's index

The generalization of Moran's index of autocorrelation (I) involves three steps: i) computing autocorrelation given any similarity between object instead of binary weightings ii) allowing non-uniform weights for the objects iii) extending the obtained formula to the multivariate case.

Let x be a centred variable measured on n objects and $\mathbf{x}$ its vector of in $\mathbb{R}^n$. Using matrix notations, Moran's index is defined as (Cliff & Ord, 1981, p. 119):

$$I(\mathbf{x}) = \frac{\mathbf{x}^T \mathbf{M} \mathbf{x}}{W \text{var}(x)} \tag{1}$$

where $\mathbf{M}$ is a spatial weighting matrix, W is the sum of all the terms of $\mathbf{M}$ ($W = \mathbf{1}_n^T \mathbf{M} \mathbf{1}_n$), and var$(x)$ is the empirical variance of x. Basically, $\mathbf{M}$ contains binary weightings derived from a neighbouring graph, where two objects i and j are either connected ($w_{ij} = 1$) or not ($w_{ij} = 0$). The lag vector $\tilde{\mathbf{x}} = \mathbf{M}\mathbf{x}$ then contains for each object the sum of the neighbouring values. However, Cliff & Ord (1981, pp.17-19) showed that any measure of proximity between objects could be used instead of binary weightings. Here, we propose to use any similarity index to define the matrix $\mathbf{M}$, which of course includes the particular case of binary spatial weightings. Note that the resulting matrix $\mathbf{M}$ is not a similarity matrix because its diagonal terms equal zero. We add the constraint that each row of the weighting matrix should sum to one, denoting $\mathbf{W}$ the resulting matrix. As a matter of fact, the i^{th} term of the lag vector $\tilde{\mathbf{x}} = \mathbf{W}\mathbf{x}$ becomes the mean of all $\mathbf{x}$ values weighted

according to their similarity to i and (EQN. 1) simplyfies as:

$$I(\mathbf{x}) = \frac{\mathbf{x}^T\mathbf{W}\mathbf{x}}{n\text{var}(x)} = \frac{\mathbf{x}^T\mathbf{W}\mathbf{x}}{\mathbf{x}^T\mathbf{x}} \tag{2}$$

The essential point here is that these similarities are not necessarily derived from spatial coordinates: in fact, any measure of similarity can be used, which allows to measure autocorrelation in various schemes, varying in natures and dimensionality. For instance, temporal autocorrelation can be measured in time series (one-dimensional scheme) using simple binary weightings: the value taken by x at time t (x_t) is neighbour to x_{t-1} and x_{t+1} (Figures 1A-B). Global (respectively local) patterns will occur when similar (respectively dissimilar) values occur among successive observations (Figures 1A-B).

The most common example of two-dimensional scheme defining autocorrelation surely is offered by spatial data (Figures 1C-D). In this case, binary weightings derived from neighbouring graphs are sometimes not satisfying from an empirical point of view, and can lead to endless discussions about how connections should be defined. As an alternative, the spatial proximities among objects ($\mathbf{W}$) can be defined as the inverse of the spatial distances among objects. The advantage of such a practice is that global patterns (Figure 1C) and local ones (Figure 1D) are no longer dependent on a particular neighbouring graph.

It also happens that the underlying scheme contains more than two dimensions, like in phylogeny (Figures 1E-F). Measuring phylogenetical autocorrelation using $I(\mathbf{x})$ can be achieved using different measures of proximities among leaves. Originally, phylogenetic proximity was defined in a binary way, considering as neighbours the leaves that have a common ancestor at a given level of the tree (Gittleman & Kot, 1990). Alternatively, one can derive a similarity index from a given phylogenetic distance and use it to compute a Moran's I (Pavoine *et al.*, 2008). These are some examples showing that Moran's index can be used to quantify autocorrelation in very different contexts by using similarities to depict the relationships among the objects under study.

In some cases, giving the same weight to every object when measuring autocorrelation is not entirely satisfying. This would occur, for instance, when sampling effort differs among objects, giving observations that are more or less reliable. Then, it would be worthwhile to weight each observation according to the confidence we can have in its value. A useful generalization of $I(\mathbf{x})$ is thus to introduce weights for the objects. The previous expression (EQN. 2) can be reformulated as a ratio of two canonical dot products:

$$I(\mathbf{x}) = \frac{\langle \mathbf{x}, \tilde{\mathbf{x}} \rangle}{\langle \mathbf{x}, \mathbf{x} \rangle} \tag{3}$$

Hence, weights can be included by using a metric $\mathbf{D}$:

$$I(\mathbf{x}) = \frac{\langle \mathbf{x}, \tilde{\mathbf{x}} \rangle_{\mathbf{D}}}{\|\mathbf{x}\|^2_{\mathbf{D}}} \tag{4}$$

This generalization expresses Moran's I as the ratio of the dot product of $\mathbf{x}$ and its lag vector, and the inertia (squared $\mathbf{D}$-norm) of $\mathbf{x}$. A consequence of enabling the use of weights is that the resulting statistic has the advantage of being able to measure autocorrelation in any score provided by an ordination in reduced space using the ordination's weights.

Now that $I(\mathbf{x})$ can handle any proximity measure and any object weights, the remaining issue is to extend this index to measure autocorrelation in multivariate data. Instead of a variable x, a data matrix $\mathbf{X} = [x_{ij}]$ with n objects (rows) and p variables (columns) is considered. In addition to the matrix of object weights $\mathbf{D} = [d_{ij}]$, we consider a $p \times p$ matrix $\mathbf{Q} = [q_{ij}]$ giving the weights of the columns of $\mathbf{X}$. These three matrices form a statistical triplet $\mathfrak{T} = (\mathbf{X}, \mathbf{Q}, \mathbf{D})$, whose inertia is the trace of the square matrix $\mathbf{X}^T\mathbf{D}\mathbf{X}\mathbf{Q}$ (tr$(\mathbf{X}^T\mathbf{D}\mathbf{X}\mathbf{Q})$, Escoufier, 1987; Holmes, 2006; Dray & Dufour, 2007). This inertia simply is the sum of the squared $\mathbf{D}$-norms of all variables of $\mathbf{X}$ weighted by $\mathbf{Q}$:

$$\text{tr}(\mathbf{X}^T\mathbf{D}\mathbf{X}\mathbf{Q}) = \sum_{j=1}^{p} q_{jj} \|\mathbf{X}^j\|^2_{\mathbf{D}} \tag{5}$$

Similarly, an operator can be seeked to compute the dot product between each variable $\mathbf{X}^j$ and its lag vector $\tilde{\mathbf{X}}^j$. This is achieved using the Frobenius product (*i.e.*, dot product between matrices):

$$\langle \mathbf{X}\mathbf{Q}^{1/2}, \tilde{\mathbf{X}}\mathbf{Q}^{1/2} \rangle_{\mathbf{D}} = \text{tr}(\mathbf{X}^T\mathbf{D}\tilde{\mathbf{X}}\mathbf{Q}) \tag{6}$$

where $\tilde{\mathbf{X}} = \mathbf{W}\mathbf{X}$ is the matrix of lag vectors. It can be seen that (EQN. 6) is the sum of the numerators of Moran's I of all variables (EQN. 4) weighted by $\mathbf{Q}$:

$$\text{tr}(\mathbf{X}^T\mathbf{D}\tilde{\mathbf{X}}\mathbf{Q}) = \sum_{j=1}^{p} q_{jj} \langle \mathbf{X}^j, \tilde{\mathbf{X}}^j \rangle_{\mathbf{D}} \tag{7}$$

But this quantity also has another interpretation. Indeed, it can be shown that:

$$\langle \mathbf{X}\mathbf{Q}^{1/2}, \tilde{\mathbf{X}}\mathbf{Q}^{1/2} \rangle_{\mathbf{D}} = \langle \mathbf{X}^T\mathbf{D}^{1/2}, \tilde{\mathbf{X}}^T\mathbf{D}^{1/2} \rangle_{\mathbf{Q}} \tag{8}$$

and provided $\mathbf{D}$ is a diagonal matrix:

$$\langle \mathbf{X}^T\mathbf{D}^{1/2}, \tilde{\mathbf{X}}^T\mathbf{D}^{1/2} \rangle_{\mathbf{Q}} = \sum_{i=1}^{n} d_{ii} \langle \mathbf{X}_{[i]}, \tilde{\mathbf{X}}_{[i]} \rangle_{\mathbf{Q}} \tag{9}$$

where $\mathbf{X}_{[i]}$ is the i^{th} row of $\mathbf{X}$. Therefore, (EQN. 6) also amounts to the weighted mean of the dot products between each multivariate observation $\mathbf{X}_{[i]}$ and its lag vector $\tilde{\mathbf{X}}_{[i]}$. The ratio between (EQN. 6) and (EQN. 5) gives a multivariate extension of the generalized Moran's I (EQN. 4):

$$I(\mathbf{X}) = \frac{\text{tr}(\mathbf{X}^T\mathbf{D}\tilde{\mathbf{X}}\mathbf{Q})}{\text{tr}(\mathbf{X}^T\mathbf{D}\mathbf{X}\mathbf{Q})} = \frac{\sum_{j=1}^{p} q_{jj}\langle\mathbf{X}^j, \tilde{\mathbf{X}}^j\rangle_{\mathbf{D}}}{\sum_{j=1}^{p} q_{jj}\|\mathbf{X}^j\|^2_{\mathbf{D}}} \tag{10}$$

The quantity $I(\mathbf{X})$ is highly positive (respectively negative) when observations in $\mathbf{X}$ tend to be positively (respectively negatively) autocorrelated. Hence, $I(\mathbf{X})$ can be used as a test statistic to detect global or local structures in multivariate data. The initial $I(\mathbf{X})$ value is first computed, and then compared to a distribution of the test statistic obtained by randomly permuting observations (Monte Carlo procedure). But as the terms $(\mathbf{X}^j|\tilde{\mathbf{X}}^j)_{\mathbf{D}}$ can be positive (global patterns) as well as negative (local ones), this statistic will fail to detect a mixture of both structures. To cope with this problem, we propose $J(\mathbf{X})$ as a complementary statistic:

$$J(\mathbf{X}) = \frac{\sum_{j=1}^{p} q_{jj}|\langle\mathbf{X}^j, \tilde{\mathbf{X}}^j\rangle_{\mathbf{D}}|}{\sum_{j=1}^{p} q_{jj}\|\mathbf{X}^j\|^2_{\mathbf{D}}} \tag{11}$$

This statistic will not be able to distinguish between global and local patterns, but will be able to detect both patterns occuring at the same time. Whenever $I(\mathbf{X})$ would not reveal significant structuring, an analogous Monte Carlo test based on $J(\mathbf{X})$ would be applied to distinguish between absence of structuring and occurence of both types of patterns with equal strength.

Constrained ordinations using Moran's I

The purpose of this section is to show how ordinations in reduced space can be constrained to provide autocorrelated scores as measured by $I(\mathbf{x})$ (EQN. 4). Our approach is based on the *duality diagram* (Escoufier, 1987; Holmes, 2006; Dray & Dufour, 2007), which encompasses ordination methods inside a unified framework, and therefore makes our results very general and broadly applicable. We insist on the fact that scores with maximum inertia derive from finding the extrema of a particular function. From this result, we propose a new function that takes both inertia and autocorrelation of the scores into account, therefore finding global and local patterns.

As shown by the duality diagram theory, all ordinations in reduced space are particular cases of the analysis of a statistical triplet $\mathfrak{T} = (\mathbf{X}, \mathbf{Q}, \mathbf{D})$, where $\mathbf{X}$ is a $n \times p$ data matrix ($\mathbf{X} \in \mathbb{R}^{n\times p}$), $\mathbf{Q}$ is a metric in $\mathbb{R}^p$ and $\mathbf{D}$ is a metric in $\mathbb{R}^n$. We denote $\mathcal{T}$ the set of all statistical triplets associated to a data matrix belonging to $\mathbb{R}^{n\times p}$, so that $\mathfrak{T} \in \mathcal{T}$. The term *duality* refers to the fact $\mathbf{X}$ defines a

cloud a n points (observations) in $\mathbb{R}^p$ as well as a cloud of p points (variables) in $\mathbb{R}^n$. The analysis of $\mathfrak{T}$ consists in finding an orthonormal basis in one space so that the cloud of points projected onto this basis (*scores*) have successive maximum inertia. Analysis in $\mathbb{R}^p$ provides a typology of observations, while analysis in $\mathbb{R}^n$ provides a typology of variables. In this paper, we only consider the first case, *i.e.* a typology of objects.

The $\mathbf{Q}$-orthonormal basis $\mathbf{U}$ of $\mathbb{R}^p$ is obtained by the eigenanalysis of the $\mathbf{Q}$-symmetric matrix $\mathbf{X}^T\mathbf{DXQ}$:

$$\mathbf{X}^T\mathbf{DXQU} = \mathbf{U\Lambda} \tag{12}$$

where $\mathbf{U}$ contains the eigenvectors of $\mathbf{X}^T\mathbf{DXQ}$ and $\mathbf{\Lambda} = \text{diag}(\lambda_1 \ \ldots \ \lambda_r)$ is a diagonal matrix of r associated eigenvalues sorted in decreasing order. Typically, ordinations in reduced space are about maximizing the inertia of the scores $\mathbf{XQU}$, the maximum being attained for the first eigenvector $\mathbf{u}_1$:

$$\|\mathbf{XQu}_1\|^2_{\mathbf{D}} = \lambda_1 \tag{13}$$

Another way to look at this property is seeking the extrema of the function f_1:

$$\begin{array}{rccl} f_1: & \mathcal{T} \times \mathbb{R}^p & \longrightarrow & \mathbb{R}_+ \\ & (\mathfrak{T}, \mathbf{u}) & \longmapsto & \mathbf{u}^T\mathbf{Q}^T\mathbf{X}^T\mathbf{DXQu} = \mathbf{u}^T\mathbf{Au} \end{array} \tag{14}$$

where $\mathbf{A}$ is a symmetric matrix. Indeed, f_1 computes the inertia of the scores $\mathbf{XQu}$ with metric $\mathbf{D}$. It is shown that the extrema of f_1 are given by $\mathbf{u}_1$ and $\mathbf{u}_r$, *i.e.* the $\mathbf{Q}$-orthonormal eigenvectors of $\mathbf{A}$ associated to the largest and lowest eigenvalues (Harville, 1997, p533-534). Note that matrix $\mathbf{A}$ is not required to be positive definite for this property to hold.

It is now possible to formulate a new function whose extrema are seeked, that would take both inertia and spatial autocorrelation into account, so as to find global as well as local patterns. To do so, we define the function f_2, which computes the product of inertia and Moran's I (eq. 4) of the scores:

$$\begin{array}{rccl} f_2: & \mathcal{T} \times \mathbb{R}^p & \longrightarrow & \mathbb{R} \\ & (\mathfrak{T}, \mathbf{v}) & \longmapsto & \|\mathbf{XQv}\|^2_{\mathbf{D}} I(\mathbf{XQv}) = \mathbf{v}^T\mathbf{Q}^T\mathbf{X}^T\mathbf{DLXQv} = \mathbf{v}^T\mathbf{Bv} \end{array} \tag{15}$$

However, finding the extrema of this expression is not immediate because $\mathbf{B}$ is not symmetric. We therefore seek a symmetric matrix $\mathbf{C}$ so that:

$$f_2(\mathfrak{T}, \mathbf{v}) = \mathbf{v}^T\mathbf{Cv} \tag{16}$$

As $\mathbf{v}^T\mathbf{B}\mathbf{v}$ is a scalar, we have:

$$\begin{aligned}\mathbf{v}^T\mathbf{B}\mathbf{v} &= \frac{1}{2}(\mathbf{v}^T\mathbf{B}\mathbf{v} + \mathbf{v}^T\mathbf{B}^T\mathbf{v}) \\ &= \frac{1}{2}(\mathbf{v}^T(\mathbf{B} + \mathbf{B}^T)\mathbf{v}) \\ &= \mathbf{v}^T(\frac{1}{2}(\mathbf{B} + \mathbf{B}^T))\mathbf{v} \end{aligned} \quad (17)$$

which gives:

$$\mathbf{C} = \frac{1}{2}(\mathbf{B} + \mathbf{B}^T) = \frac{1}{2}\mathbf{Q}^T\mathbf{X}^T(\mathbf{D}\mathbf{W} + \mathbf{W}^T\mathbf{D})\mathbf{X}\mathbf{Q} \quad (18)$$

It follows that the extrema of f_2 are given by the eigenanalysis of the $\mathbf{Q}$-symmetric matrix $\mathbf{E}$ defined as:

$$\mathbf{E} = \frac{1}{2}\mathbf{X}^T(\mathbf{D}\mathbf{W} + \mathbf{W}^T\mathbf{D})\mathbf{X}\mathbf{Q} \quad (19)$$

which yields a set of $\mathbf{Q}$-orthonormal eigenvectors $\mathbf{V}$ and a diagonal matrix $\mathbf{\Gamma} = \text{diag}(\gamma_1 \dots \gamma_r)$ of r associated eigenvalues sorted in decreasing order, so that:

$$\max(f_2) = \|\mathbf{X}\mathbf{Q}\mathbf{v}_1\|^2_{\mathbf{D}} I(\mathbf{X}\mathbf{Q}\mathbf{v}_1) = \gamma_1 \quad (20)$$

and

$$\min(f_2) = \|\mathbf{X}\mathbf{Q}\mathbf{v}_r\|^2_{\mathbf{D}} I(\mathbf{X}\mathbf{Q}\mathbf{v}_r) = \gamma_r \quad (21)$$

The eigenvectors $\mathbf{V}$ therefore provide a $\mathbf{Q}$-orthonormal basis onto which global and local structures among objects are decomposed. As it has been noticed by Dray *et al.* (2008), the eigenanalysis of $\mathbf{E}$ also is a particular case of co-inertia analysis for fully-matched tables (Torre & Chessel, 1995) between $\mathbf{X}$ and $\tilde{\mathbf{X}}$. Global structures would be indicated by large positive eigenvalues, while local structures would result in large negative eigenvalues. As in usual ordinations in reduced space, a sharp decrease in eigenvalues is likely to indicate the number of axes to be retained. The resulting scores can be graphically displayed to assess visually these patterns and can also be used as response variable or as predictors in modelling approaches.

Illustrations

Our methodology is illustrated within three different contexts: i) the investigation of global and local patterns in simulated time series ii) the study of spatial genetic patterns in Galapagos tortoise populations iii) the research of phylogenetic patterns in a set of morphological traits of teleost fishes.

Simulated time series

Global and local patterns were seeked from a simulated time series dataset. The dataset was composed of the global and local variables from Figures 1A-B and 18 other variables randomly drawn from a normal distribution (considered as random noise). Each variable contained 20 observations. The weighting matrix $\mathbf{W}$ was designed so that x_t was connected to x_{t-1} and x_{t+1} (Figures 1A-B).

The existence of temporal autocorrelation was assessed using $I(\mathbf{X})$ (EQN. 10) Monte Carlo test (9999 randomizations). Likely because both global and local patterns of equal strength existed, this test failed to reveal any autocorrelation (Figure 2A, $I(\mathbf{X}) = -0.048$, NS). As expected, the complementary test based on $J(\mathbf{X})$ (EQN. 11) revealed the existence of temporal autocorrelation (Figure 2B, $J(\mathbf{X}) = 0.256$, $p = 0.008$ on 9999 permutations).

Hence, these structures were investigated using a 'temporally constrained' PCA, in the sense that resulting scores would display temporal autocorrelation. The temporal constraint was included using the $\mathbf{W}$ matrix previously mentionned (Figures 1A-B). Eigenvalues suggested that one global and one local structure should be retained (Figures 2C-D) The first global score of the temporally constrained PCA well retrieved the globally structured variable existing in the data (Figure 2C). The same is true for the first local score which clearly revealed the existing locally structured variable (Figure 2D).

Thus, this temporally constrained ordination was able to find and disentangle one global and one local pattern from a fair amount of random noise (2 variables out of 20 were temporally structured).

Galapagos tortoises

This illustration uses data published in Ciofi *et al.* (2002) and available in the *ade4* package as the dataset *ggtortoises*. Seventeen georeferenced Galapagos tortoises (*Geochelone elephantopus*) populations were genotyped for 10 microsatellite markers. Population samplings ranged from 148 to 269 genotypes per population. The resulting dataset is a contingency table counting alleles per population which can be submitted to a Correspondence Analysis in order to summarise the genetic variability among populations into a few components. However, as the location of all populations are known, the question of the spatial organisation of this variability arises.

First, the $I(\mathbf{X})$ (EQN. 10) Monte Carlo test was used to assess the existence of spatial patterns in alleles counts. The matrix of spatial proximities among populations ($\mathbf{W}$) was defined as the inverse of the squared geographic distances. The test revealed significant global patterns ($I(\mathbf{X}) = 0.032$, $p = 1.10^{-4}$ on 9999 permutations, see Figure 3). Therefore, a constrained Correspondence Analysis was used to find these patterns. The eigenvalues indicated that one structure

should be retained (Figure 3). The population scores onto the first axis of the analysis were mapped in order to visualize the spatial pattern (Figure 3). This structure highlighted the genetic and spatial differentiation between populations from southern Isabella island and all the others. This result was consistent with previous findings showing that populations from southern Isabella were genetically rather homogeneous and well differentiated from all the others (Beheregaray *et al.*, 2004; Ciofi *et al.*, 2006).

Teleost fishes

This illustration uses data published in Rochet *et al.* (2000) and is also available in *ade4* as the dataset *mjrochet*. Data consist in 7 biodemographic traits measured on 49 teleost fish species, whose phylogeny is provided: *Age at sexual maturity* (T_m), *Length at sexual maturity* (L_m), *Length at 5% survival* ($L_{.05}$), *Time to 5% survival* ($T_{.05}$), *Slope of fecundity-length relationship* (F_b), *Fecundity at maturity* (F_m) and *Egg volume* (Egg). As usual in comparative studies, the first concern is to assess the existence of phylogenetic signal in data, *i.e.* the non-independence of observations across phylogeny. Whenever such signal is detected, we propose to reveal the corresponding patterns using a phylogenetically constrained ordination.

Although Rochet *et al.* (2000) already proved the non-independence of observations regarding phylogeny, we assessed it using our $I(\mathbf{X})$ Monte Carlo procedure. Interestingly, our general procedure turns out to be the first multivariate test of phylogenetic autocorrelation. Prior to analyses (autocorrelation test and constrained ordination), all variables but F_b were log-transformed as in Rochet *et al.* (2000) to improve their normality. Size effect was not removed from data as size itself is likely to be phylogenetically structured. The phylogenetic proximity matrix $\mathbf{W}$ was derived from the squared inverse phylogenetic distance between species, as it is common in autoregressive models (Gittleman & Kot, 1990). This test confirmed the existence of positive phylogenetic autocorrelation ($I(\mathbf{X}) = 0.022$, $p = 2.10^{-4}$ on 9999 permutations, Figure 4).

The corresponding global patterns were investigated using a scaled PCA (to account for different units among variables) phylogenetically constrained by matrix $\mathbf{W}$. The resulting scores displayed the part of the variability in biodemographic traits which was phylogenetically structured. The eigenvalues clearly showed that two global scores were to be retained. The variable loadings (Table 1) indicated that the first axis was essentially negatively linked to the *egg volume*. The first global pattern (Figure 4) differentiated Perciformes (from *S. cavalla* to *G. niger*) with positive scores (low egg volumes) from other taxa with large negative scores (*i.e.*, large egg volumes), especially taxa *S. alpinus*, *S. malma* and *M. villosus*. The second axis was positively linked to egg volume and negatively related to length measures (L_m, $L_{.05}$, Table 1). The corresponding

score (Figure 4) essentially differentiated Argentiformes and Salmoniformes (from *E. capensis* to *M. villosus*) from Gadiformes (from *G. morhua* to *M. productus*).

Conclusion

We proposed a general framework to constrain ordinations in reduced space to yield autocorrelated scores, using a generalized form of Moran's I. We also developped a multivariate extension of Moran's index which measures autocorrelation in multivariate data, and allows for testing the existence of *global* and *local* patterns, respectively corresponding to positively and negatively autocorrelated variables. When such patterns are detected, a method is needed to investigate them. While usual ordination methods provide observation scores of maximum inertia, our approach yields scores decomposing the product of inertia and Moran's I. Taking both variability and autocorrelation into account, this criterion allows to identify global and local patterns.

A first observation concerning our method is that it yields positive and negative eigenvalues, while usual ordination methods perform the eigenanalysis of a positive-definite matrix and therefore provide only positive eigenvalues. This is due to the fact that the function whose extrema are found in usual methods compute inertia (EQN. 14), which is by definition positive or null. As the function used in our approach maps data onto $\mathbb{R}$ (EQN. 15), there is no reason why eigenvalues finding its extrema should be only positive. Negative eigenvalues are here as meaningful as positive ones, and do not represent a flaw of the method.

A second issue concerns the diagonalized matrix. In usual ordination methods, eigenvalues and eigenvectors can be computed in $\mathbb{R}^p$ or $\mathbb{R}^n$, according to the smallest dimension, which can sometimes save computational time (Dray & Dufour, 2007). In our approach the diagonalization is always performed in $\mathbb{R}^p$, and the possibility of doing computations in $\mathbb{R}^n$ should be investigated.

Another point concerns the choice of the number of retained axes. Like in other ordination methods, we propose that eigenvalues fairly more extreme than the others (*i.e.*, largely positive or negative) should be retained. This criterion remains debatable, however, and testing procedures for the number of retained axes should be further investigated. Recent works based on permutational procedures (Dray, 2008) or on theoretical distribution of eigenvalues (Soshnikov, 2002; Soshnikov & Fyodorov, 2005) should be considered with interest.

To conclude, we hope the proposed framework will be a basis for new constrained ordination methods to be developped. It already encompasses two spatially constrained methods conceived in different contexts (Dray *et al.*, 2008; Jombart *et al.*, 2008). However, the autocorrelation concept goes far beyond the investigation of spatial patterns. Here, we proposed an approach that can be

applied to investigate spatial, temporal, phylogenetic, and other kinds of patterns in multivariate data.

References

ANDERSON, M. & WILLIS, T. (2003). Canonical analysis of principal coordinates: a useful method of constrained ordination for ecology. *Ecology* **84**, 511–525.

BEHEREGARAY, L., GIBBS, J., HAVILL, N., FRITTS, T., POWELL, J. & CACCONE, A. (2004). Giant tortoises are not slow: rapid diversification and biogeographic consensus in the galápagos. *Proceedings of the National Academy of Sciences* **101**, 6514–6519.

CHESSEL, D., DUFOUR, A.-B. & THIOULOUSE, J. (2004). The ade4 package-I- one-table methods. *R News* **4**, 5–10.

CIOFI, C., WILSON, G., BEHEREGARAY, L., MARQUEZ, C., GIBBS, J., TAPIA, W., SNELL, H., CACCONE, A. & POWELL, J. (2006). Phylogeographic history and gene flow among giant galápagos tortoises on southern Isabela Island. *Genetics* **172**, 1727–1744.

CIOFI, M., C. MILINKOVITCH, GIBBS, J., CACCONE, A. & POWELL, J. (2002). Microsatellite analysis of genetic divergence among populations of giant galapagos tortoises. *Molecular ecology* **11**, 2265–2283.

CLIFF, A. & ORD, J. (1973). *Spatial autocorrelation*. London: Pion.

CLIFF, A. & ORD, J. (1981). *Spatial Processes. Model & Applications*. London: Pion.

DOLÉDEC, S. & CHESSEL, D. (1987). Rythmes saisonniers et composantes stationnelles en milieu aquatique. ii. prise en compte et élimination d'effets dans un tableau faunistique. *Acta Oecologica, Oecologia Generalis* **8**, 403–426.

DRAY, S. (2008). On the number of principal components: A test of dimensionality based on measurements of similarity between matrices. *Computational statistics & data analysis* **52**, 2228–2237.

DRAY, S. & DUFOUR, A.-B. (2007). The ade4 package: implementing the duality diagram for ecologists. *Journal of statistical software* **22(4)**, 1–20.

DRAY, S., DUFOUR, A.-B. & CHESSEL, D. (2007). The ade4 package - II: Two-table and K-table methods. *R News* **7**, 47–54.

DRAY, S., SAÏD, S., DEBIAS, F. & CHESSEL, D. (2008). Spatial ordination of vegetation data using a generalization of Wartenberg's multivariate spatial correlation. *Journal of Vegetation Science* **19**, 45–56.

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

GITTLEMAN, J. L. & KOT, M. (1990). Adaptation: statistics and a null model for estimating phylogenetic effects. *Systematic zoology* **39**, 227–241.

GREENACRE, M. (1984). *Theory and applications of correspondance analysis*. Academic Press.

HARVILLE, D. (1997). *Matrix algebra from a statistician's perspective*. Springer, New York.

HILL, M. & SMITH, J. (1976). Principal component analysis of taxonomic data with multi-state discrete characters. *Technical Bulletin of Faculty of Horticulture Chiba University* **25**, 249–255.

HOLMES, S. (2006). *Festschrift for David Freedman*, chap. Multivariate data analysis: the French way. IMS Lecture Notes.

JOMBART, T., DEVILLARD, S., DUFOUR, A.-B. & PONTIER, D. (2008). Revealing cryptic spatial patterns in genetic variability by a new multivariate method. *Heredity* **101**, 92–103.

LEGENDRE, P. & LEGENDRE, L. (1998). *Numerical ecology*. Elsevier Science B.V., Amsterdam.

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

MORAN, P. (1950). Notes on continuous stochastic phenomena. *Biometrika* **37**, 17–23.

PAVOINE, S., OLLIER, S., PONTIER, D. & CHESSEL, D. (2008). Testing for phylogenetic signal in life history variable: Abouheif's test revisited. *Theoretical population biology* **73**, 79–91.

PEARSON, K. (1901). On lines and planes of closest fit to systems of points in space. *Philos Mag* **2**, 559–572.

R DEVELOPMENT CORE TEAM (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

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

ROCHET, M., CORNILLON, P., SABATIER, R. & PONTIER, D. (2000). Comparative analysis of phylogenetic and fishing effects in life history patterns of teleost fishes. *Oikos* **91**, 255–270.

SOSHNIKOV, A. (2002). A note on universality of the distribution of the largest eigenvalues in certain sample covariance matrices. *Journal of Statistical Physics* **108**, 1033–1056.

SOSHNIKOV, A. & FYODOROV, Y. (2005). On the largest singular values of random matrices with independent Cauchy entries. *Journal of Mathematical Physics* **46**, 033302.

TENENHAUS, M. & YOUNG, F. (1985). An analysis and synthesis of multiple correspondence analysis, optimal scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika* **50**, 91–119.

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.

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

TORRE, F. & CHESSEL, D. (1995). Co-structure de deux tableaux totalement appariés. *Revue de Statistique Appliquée* **43**, 109–121.

Table legends

Table 1: Loadings of the 7 biogeographic variables onto the first two axes of the constrained analysis of Teleost fishes data.

Tables

Table 1

	Axis 1	Axis 2
T_m	−0.371	−0.115
L_m	−0.399	−0.486
$L_{.05}$	−0.269	−0.578
$T_{.05}$	−0.166	−0.128
F_b	0.308	0.065
F_m	0.020	−0.182
Egg	−0.713	0.602

Figure legends

Figure 1: Examples of global (A,C,E) and local (B,D,F) patterns in various schemes. All variables are centred. Each pattern is represented with the matrix of proximities among objects $\mathbf{W}$ (largest proximities in black, zero in white) and the $I(\mathbf{x})$ Monte Carlo test (EQN. 4). (A,B) Autocorrelation in a one-dimensional scheme (time), using binary connections among observations (x_t connected to x_{t-1} and x_{t+1}). (C,D) Autocorrelation in a two-dimensional scheme (space), using the inverse of spatial distances among objects as a measure of proximity. (E,F) Autocorrelation in a high-dimensional scheme (phylogeny), using the inverse of squared phylogenetic distances as a measure of proximity among taxa.

Figure 2: Analysis of simulated time series. Dataset consisted of the two variables from Figure 1A ('global variable') and 1B ('local variable') and 18 other variables drawn from a normal distribution. (A) $I(\mathbf{X})$ Monte Carlo test (EQN. 10), non-significant on 9999 permutations. (B) $J(\mathbf{X})$ Monte Carlo test (EQN. 11); $J(\mathbf{X}) = 0.256$, right-tail p-value equals 0.008 on 9999 permutations. (C) First global score of the constrained PCA (squares and plain line) retrieving the structuring of the 'global variable' (triangles and dashed line); the corresponding eigenvalue is filled in black. (D) First local score of the constrained PCA (squares and plain line) retrieving the structuring of the 'local variable' (triangles and dashed line); the corresponding eigenvalue is filled in black.

Figure 3: First global score of the constrained correspondence analysis of Galapagos tortoise data, using the inverse air distance between populations as a proximity measure. Top-left plot represents the $I(\mathbf{X})$ Monte Carlo test (EQN. 10), $I(\mathbf{X}) = 0.032$, right-tail p-value equaled 1.10^{-4} on 9999 permutations. Top-right plot is the screeplot of eigenvalues (retained and represented structure filled in black).

Figure 4: First global score of the constrained principal component analysis of teleost fishes data, using the inverse phylogenetic distance between taxa as a proximity measure. The score is represented using Cleveland's graph. The upper-left plot represents the $I(\mathbf{X})$ Monte Carlo test (EQN. 10), $I(\mathbf{X}) = 0.008$, right-tail p-value equaled 0.0029 on 9999 permutations. The lower-right plot is the screeplot of eigenvalues (retained and represented structure filled in black).

Figures

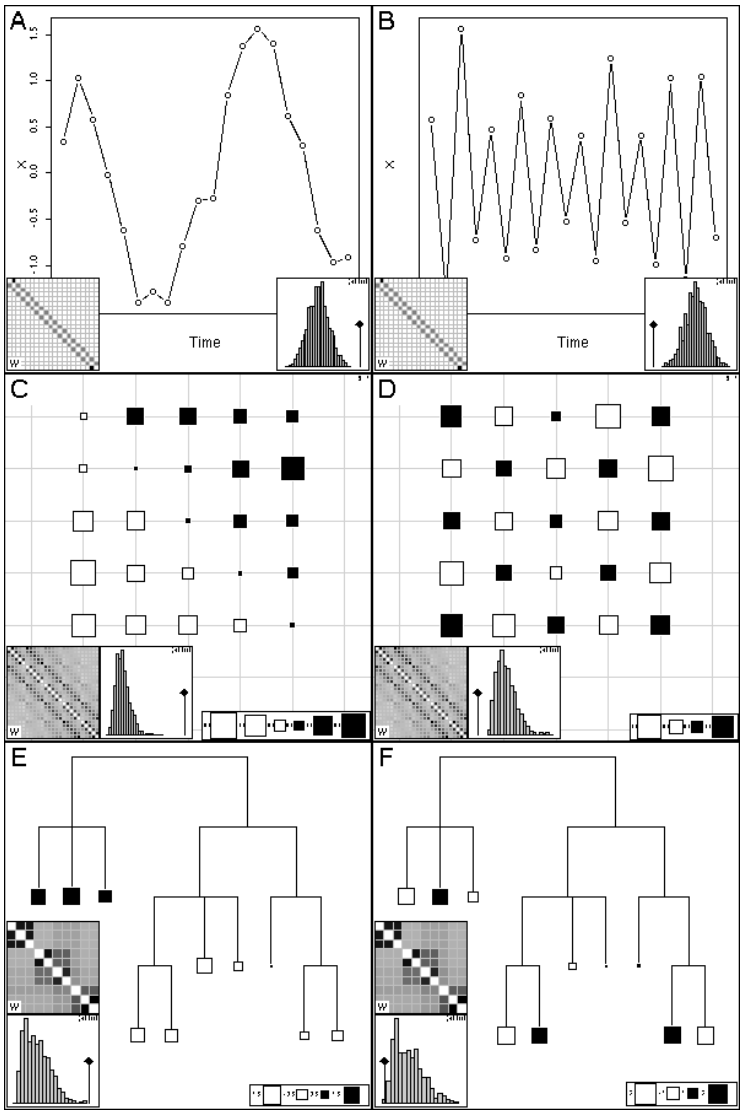


Figure 1

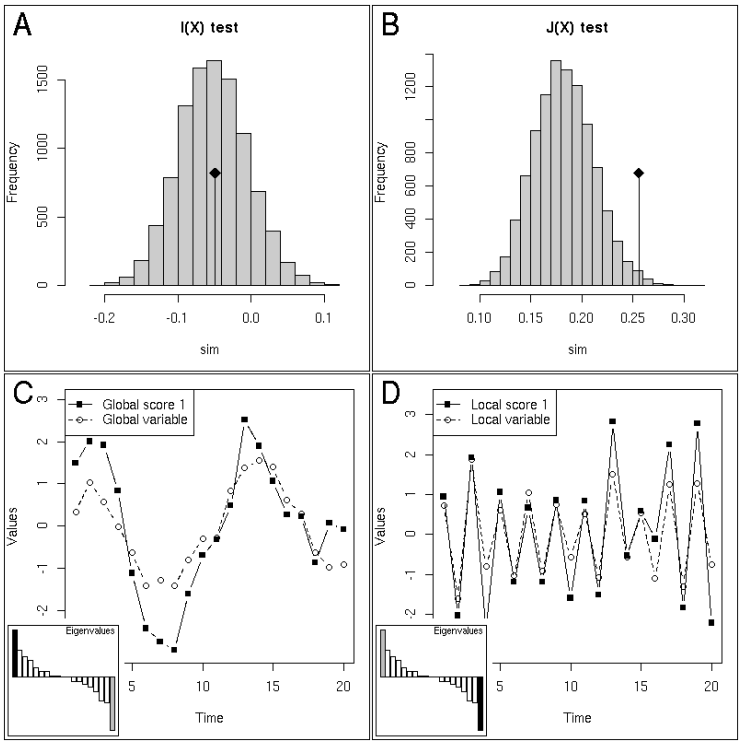

A
I(X) test
Frequency
sim
B
J(X) test
Frequency
sim
C
Global score 1
Global variable
Values
Eigenvalues
Time
D
Local score 1
Local variable
Values
Eigenvalues
Time

Figure 2

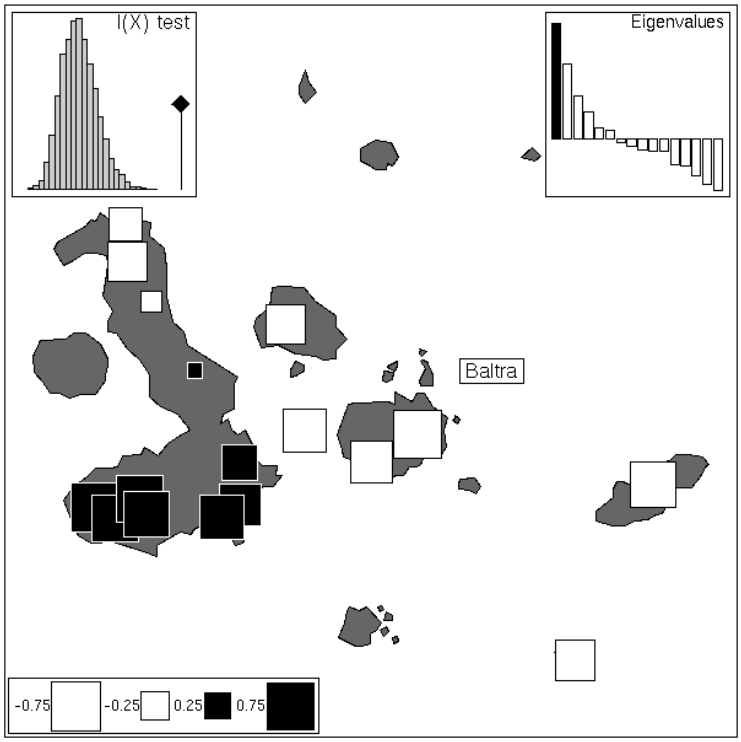


Figure 3

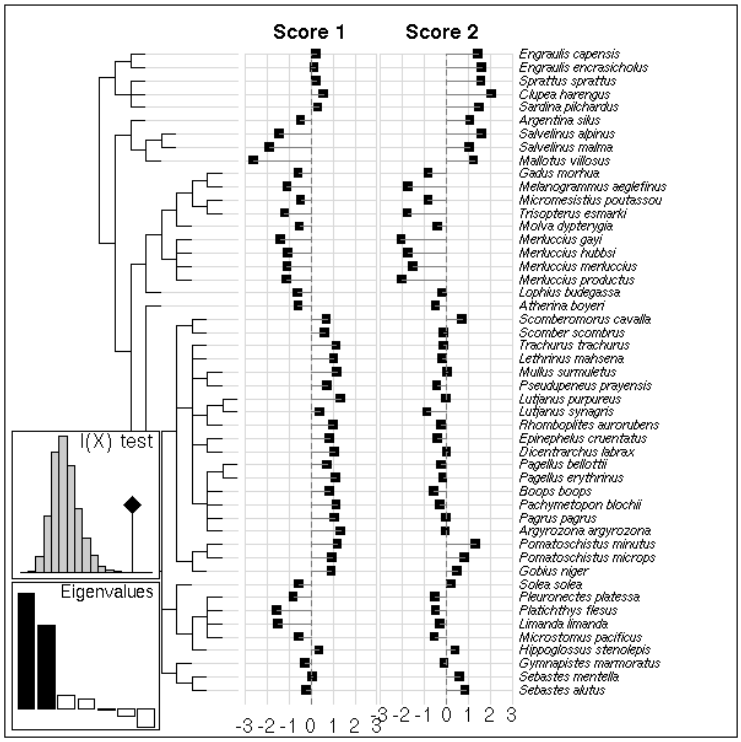

Score 1
Score 2
Engraulis capensis
Engraulis encrasicholus
Sprattus sprattus
Clupea harengus
Sardina pilchardus
Argentina silus
Salvelinus alpinus
Salvelinus malma
Mallotus villosus
Gadus morhua
Melanogrammus aeglefinus
Micromesistius poutassou
Trisopterus esmarki
Molva dypterygia
Merluccius gayi
Merluccius hubbsi
Merluccius merluccius
Merluccius productus
Lophius budegassa
Atherina boyeri
Scomberomorus cavalla
Scomber scombrus
Trachurus trachurus
Lethrinus mahsena
Mullus surmuletus
Pseudupeneus prayensis
Lutjanus purpureus
Lutjanus synagris
Rhomboplites aurorubens
Epinephelus cruentatus
Dicentrarchus labrax
Pagellus bellottii
Pagellus erythrinus
Boops boops
Pachymetopon blochii
Pagrus pagrus
Argyrozona argyrozona
Pomatoschistus minutus
Pomatoschistus microps
Gobius niger
Solea solea
Pleuronectes platessa
Platichthys flesus
Limanda limanda
Microstomus pacificus
Hippoglossus stenolepis
Gymnapistes marmoratus
Sebastes mentella
Sebastes alutus
I(X) test
Eigenvalues
-3 -2 -1 0 1 2 3
-3 -2 -1 0 1 2 3

Figure 4

7.3 Discussion

Le cadre général que nous avons dressé pour l'ordination sous contrainte d'autocorrélation fournit une base à de futurs développements méthodologiques. Si les applications dans le contexte spatial (chapitre 3) et phylogénétique (chapitre 6) ont commencé à montrer l'intérêt de cette démarche, il est certain que ce ne sont que quelques éléments d'un champ d'applications qui reste largement à explorer. Ces travaux sont encore à l'état de perspectives, et il est peu d'éléments que l'on puisse discuter avec pertinence. On peut toutefois mentionner quelques points qui demanderont sans doute une attention particulière.

L'implémentation de la méthode est, pour l'instant, assez mal construite, dans la mesure où elle est clairement orientée vers l'étude de structures spatiales. Il sera donc nécessaire d'ouvrir cette implémentation à un cadre plus large, dont l'aspect spatial ne sera qu'un cas particulier. Certains outils tels que la décomposition graphique des valeurs propres de la sPCA (Fig. 3.4) conserveront un sens indépendamment du contexte de l'étude. D'autres outils, en premier lieu la représentation graphique des structures globales et locales, seront totalement dépendants du contexte. Si les structures spatiales, temporelles ou phylogénétiques peuvent être facilement représentées, il sera par exemple plus délicat de représenter des structures liées à des réseaux d'interactions biologiques (Barabási & Oltvai, 2004), dont la taille peut être un obstacle sérieux à la visualisation de l'information.

Par ailleurs, il est apparu dans le contexte spatial comme dans celui de la phylogénie que la mesure de similarité entre objets introduite par la matrice **W** est un choix déterminant dans la mesure de l'autocorrélation. Il est clair que pour une majorité d'applications, la formulation de la matrice de proximités entre objets sera un point de discussion central. Plutôt qu'un défaut de l'approche, on pourra considérer qu'il s'agit d'une faculté intéressante de la méthode à prendre en compte différents points de vue, ce qui n'empêchera pas de s'interroger sur leur pertinence respective.

En conclusion, il semble que le cadre méthodologique que nous proposons constitue une approche prometteuse pour l'analyse des structures biologiques. Mais pour prometteuse qu'elle soit, cette approche ne pourra se dispenser d'une réflexion biométrique pour chaque nouvelle application.

Conclusion générale

Bilan

Maintenant que les différents éléments qui ont été abordés au cours de cette thèse ont été présentés, nous proposons de résumer l'essentiel de ce travail pour en évaluer plus précisément la portée.

La première tâche qui nous était fixée au commencement de cette thèse était d'examiner dans quelle mesure l'analyse de données pouvait être utilisée pour étudier la cohérence de l'information fournie par différents marqueurs multiallélliques. Cette problématique biologique trouve un ensemble de réponses méthodologiques dans le cadre des méthodes K-tableaux, que nous avons introduites en génétique en présentant une application de l'analyse de co-inertie multiple (Chessel & Hanafi, 1996) à l'étude de la cohérence typologique d'un jeu de marqueurs microsatellites (Jombart *et al.*, 2006; Laloë *et al.*, 2007). Cette approche s'est révélée utile pour mesurer la cohérence de l'information portée par différents locus, et a permis d'identifier une typologie consensuelle ainsi que la participation de chaque locus à cette typologie. Une mesure de la valeur typologique d'un marqueur, basée sur cette participation à la typologie d'ensemble, a également été proposée.

Le second point abordé portait sur l'identification de la structure génétique d'un ensemble de génotypes ou de populations. Partant du constat que ce type de structure est le plus souvent spatialisé, nous avons développé une approche — l'*analyse en composantes principales spatiales* (sPCA) — permettant d'explorer les structures spatiales de la variabilité génétique existant dans un ensemble de génotypes (ou de populations typées) géoréférencées (Jombart *et al.*, 2008). Cette méthode est largement applicable, du fait qu'elle ne repose pas sur un modèle de génétique des populations particulier, et s'est révélée efficace pour identifier des structures génétiques spatiales, y compris dans de faibles niveaux de différenciation génétique.

En parallèle de la réflexion méthodologique s'est posé le problème de la pratique de l'analyse des données de marqueurs moléculaires. Nous avons fait le choix de ne travailler que dans le cadre du logiciel libre, qui offre au travers du logiciel R (R Development Core Team, 2008) un formidable outil pour l'analyse statistique des données. Mais bien que R soit amplement fourni en terme de méthodes multivariées, il manquait clairement un cadre technique qui permette de gérer les données de marqueurs génétiques. Le package *adegenet* (Jombart, 2008) a été développé pour combler cette lacune, et a connu depuis sa création une croissance régulière, intégrant progressivement de nouveaux outils, qu'il s'agisse de méthodes, de procédures pour la manipulation des données ou de simulations.

Un premier pas hors du contexte strictement génétique a été franchi avec le développement d'une méthode d'*analyse des structures spatiales multi-échelles* (MSPA, Jombart *et al.*, sous presse). Cette approche a été développée dans le contexte écologique pour des raisons culturelles, la problématique des structures spatiales à plusieurs échelles étant plus présente dans ce domaine qu'en génétique des populations. Néanmoins, la méthode proposée est très générale, puisqu'elle peut s'appliquer à des données quantitatives, qualitatives ou mixtes. Une première application aux données génétiques a démontré que cette méthodologie était également pertinente pour explorer les principales échelles auxquelles un ensemble de génotypes est structuré dans l'espace.

Enfin, une extension de la sPCA a été développée dans le contexte de la méthode comparative, pour l'exploration des structures phylogénétiques au sein d'un ensemble de traits. La méthode proposée, nommée par analogie à la sPCA l'*analyse en composantes principales phylogénétiques* (pPCA), permet d'identifier la part de la variabilité des traits étudiés qui n'est pas indépendante de la phylogénie. Cette approche suggère, comme cela l'a été plus tôt en écologie (Legendre, 1993), que l'autocorrélation phylogénétique n'est pas qu'une nuisance à l'analyse de données comparatives, mais qu'elle recèle au contraire une information biologique digne d'intérêt. L'extension de la sPCA nous a par ailleurs amené à formuler un cadre plus général pour l'ordination en espace réduit sous contrainte d'autocorrélation. Cette généralisation, dérivée de l'approche MULTISPATI proposée par Dray *et al.* (2008), permet d'adapter de nombreuses analyses multivariées à la recherche de structures autocorrélées par rapport à une mesure de proximité donnée entre observations. On peut donc utiliser cette démarche dans divers contextes, afin de rechercher des structures spatiales, temporelles, phylogénétiques, ou définies par un graphe modélisant des interactions entre objets étudiés.

On espère, dans ce travail de thèse, avoir apporté des arguments à l'idée directrice simple que l'analyse de données est une approche pertinente pour extraire de l'information biologique des données de marqueurs moléculaires. On se propose d'insister sur les points qui semblent les plus pertinents. Tout d'abord, il était nécessaire d'effectuer un bilan de la contribution actuelle de l'ordination en espace réduit à l'analyse des marqueurs génétiques. La revue bibliographique proposée (Jombart *et al.*, in revision) devrait remplir au moins en partie ce besoin ; la première conclusion émergeant de cette revue est que nous avons indentifié un champ de recherche biométrique à part entière. Nous avons observé que certaines pratiques relativement courantes pouvaient être largement améliorées ; ces écueils procèdent vraisemblablement du fait que l'analyse de données est parfois perçue comme une technique, vision dans laquelle la méthode est confondue avec son implémentation, ce qui bannit les espoirs de reproductibilité pourtant essentielle aux résultats scientifiques. Cet obstacle avant tout culturel sera sans doute long à surmonter, mais il n'enlève rien au constat plus réjouissant que l'analyse exploratoire des données génétiques par des méthodes euclidiennes constitue une approche efficace pour aborder de nombreuses problématiques.

Parmi celles-ci, nous pensons avoir contribué à explorer deux thématiques en particulier. D'une part, l'introduction des méthodes K-tableaux a ouvert une discussion sur la mesure de la diversité génétique, discussion dans laquelle différentes méthodes — dont certaines ont été vues au chapitre 2 — offrent plusieurs points de vue originaux. D'autre part, partant du constat que

la variabilité génétique des populations naturelles est très souvent spatialisée, nous avons montré que l'analyse multivariée pouvait proposer des approches efficaces pour explorer ces structures (Jombart *et al.*, 2008, sous presse), en créant au passage une brèche vers les problématiques multi-échelles chères à l'écologie (Jombart *et al.*, sous presse), qui sont également pertinentes en génétique des populations. Ces développements méthodologiques ont été accompagnés d'une mise à disposition des procédures via le développement du package *adegenet* (Jombart, 2008), qui inclut également une variété d'outils pour la gestion et la manipulation des données de marqueurs moléculaires. Les retours obtenus depuis suggèrent que cet outil comble un réel besoin pratique, et démontre qu'une part de la communauté scientifique est ouverte à l'analyse multivariée des données génétiques. Enfin, les extensions méthodologiques développées à partir de la problématique initiale de l'analyse des marqueurs génétiques (Jombart *et al.*, en prép-a, en prép-b) suggèrent que ce champ est particulièrement riche, dans la mesure où il en émerge des modèles féconds au sens de Legay (1997, p.53) :

> *Un modèle est meilleur qu'un autre s'il s'applique à un univers plus large, s'il rend inutile d'autres modèles, s'il permet d'étendre l'utilisation de techniques communes, s'il autorise des comparaisons nouvelles.* [...] *La fécondité d'un modèle est l'ensemble des résultats et des conséquences non prévues que son usage entraîne.*

On peut également faire dans ce bilan un certain nombre de critiques. La première est celle de l'éparpillement : si les extensions méthodologiques qui ont été proposées sont intéressantes sur le plan biométrique, elles nous ont néanmoins entraîné hors de notre problématique initiale, laquelle aurait donc pu recevoir plus d'attention. Une attention particulière aurait pu être accordée à la diversité des marqueurs moléculaires. En effet, une part importante de notre travail a sous-entendu l'usage de marqueurs codominants (allozymes, microsatellites et SNP) ; par exemple, la gestion des données dans *adegenet* ainsi que l'implémentation de la sPCA sont construites pour de tels marqueurs. Bien que les méthodes développées soient également applicables à des données de polymorphisme de tailles de fragments (*e.g.*, Restriction Fragment Length Polymorphism (RFLP), Amplified Fragment Length Polymorphism (AFLP)), nous n'avons pas entamé de réflexion pour ce type de données en particulier. En première approximation, ces données peuvent être traitées comme des présences/absences, mais de récents développements dans la mesure des proximités génétiques entre profils d'AFLP (Gort *et al.*, 2006, 2008) attirent notre attention sur le fait que la situation soit plus complexe.

Par ailleurs, nous n'avons pas mené de réflexion de fond sur l'utilisation des distances génétiques. Nous avons considéré la distance euclidienne entre génotypes ou entre profils de fréquences alléliques entre populations comme une solution satisfaisante, en appliquant éventuellement une standardisation appropriée. Il y a sans doute un intérêt à examiner quelles distances génétiques sont les plus pertinentes dans une ordination en espace réduit. Il est possible qu'une réponse unique n'existe pas, puisque certaines distances reposent sur des modèles de génétique des populations, et donc sur des hypothèses faites quant aux données (Weir, 1996, pp.190-198). Dans le même esprit, nous n'avons fait qu'esquisser un pas vers l'investigation de la nature compositionnelle des données de fréquences alléliques, bien que Reyment (2005) encourage clairement les recherches dans cette direction. Dans le cas des marqueurs multialléliques la situation est en effet originale puisque des fréquences alléliques

définies pour un ensemble de marqueurs forment une juxtaposition de tableaux de données compositionnelles, qui est elle-même de nature compositionnelle. Les implications de cette structure de données sur les résultats d'ordinations en espace réduit n'ont pas été, mais devraient être, évaluées.

On terminera ce bilan en notant que cette thèse a été formatrice du point de vue de la pratique de la biométrie. Cette expérience, bien que courte, nous a permis d'observer certains aspects de la biométrie qui sont venus préciser la vision de l'activité biométrique décrite en introduction. L'expérience la plus riche à cet égard a peut-être été celle du développement informatique. Le dialogue interdisciplinaire caractérisant la recherche en biométrie passe aujourd'hui immanquablement par la mise à disposition des procédures, leur diffusion et leur documentation. L'implémentation, lorsqu'elle n'implique pas les compétences d'un informaticien chevroné, incombe sans doute souvent au biométricien, qui est par définition un touche-à-tout. Cette étape essentielle pour la diffusion des méthodes sert par ailleurs de support au dialogue avec des utilisateurs de différents horizons, qui peut s'avérer particulièrement riche et intéressant.

Il est cependant regrettable que cet aspect de la biométrie, qui est le plus directement utile, soit en même temps parfois le plus méprisé. La critique émane curieusement de biométriciens plus que d'utilisateurs, et consiste à ne reconnaître au programme qu'un statut d'outil, c'est-à-dire l'objet de peu réflexion, tout au plus une démonstration technique vaguement utile à l'analyse de données. Pourtant, Legay (1997, p.23) attire notre attention sur la notion d'outil :

> *Qui veut enfoncer un clou se sert d'un marteau ; le marteau est l'outil, le clou à enfoncer l'objectif ; il n'y a pas d'ambiguïté, il ne pourrait y avoir que maladresse... Si je veux en biologie étudier les ultra-structures cellulaires, le microscope électronique va être l'outil indispensable. Et là les difficultés commencent, car l'outil lui-même est très élaboré ; il a impliqué le travail de nombreux chercheurs et ingénieurs de plusieurs disciplines* [...] *il y a une cascade d'objectifs et d'outils intermédiaires.* [...] *On pourrait faire des remarques analogues en mathématique à propos du calcul matriciel : outil de travail indispensable dans bien des domaines de la science, il est en même temps objet de recherches dans la discipline d'origine.*

Il est donc assez réducteur de concevoir l'implémentation des procédures comme une activité purement technique. Le logiciel R regorge d'exemples de développement de logiciels qui ont requis bien plus que de la technique dans leur élaboration. Le package *ade4* en est une bonne illustration : il s'agit de l'image directe dans le champ des procédures de la théorie du schéma de dualité (Dray & Dufour, 2007), et donc d'un programme qui véhicule une vision très personnelle de la biométrie (Chessel, 1992). Des remarques similaires pourraient être faites pour nombre d'autres logiciels dont l'énumération n'apporterait rien.

Un autre point concerne la diffusion des programmes et des données. Choisir le logiciel libre comme mode de développement s'inscrit pleinement dans la vision de l'activité biométrique en tant que dialogue, dans lequel l'information doit circuler librement. Qui plus est, cette information circule de manière transparente : les programmes ne sont plus des « boîtes noires », mais des objets dont on peut disséquer le contenu, le réutiliser et le modifier à son gré. Dans cette optique, les données devraient être aussi accessibles et transparentes aux biométriciens que

les programmes le sont aux biologistes. On réfère ici encore à Chessel (1992, p.23) :

> *Image d'une méthode pour celui qui l'écrit, le programme change de nature pour celui qui l'emploie, image d'une problématique pour celui qui les acquiert, les données changent de nature quand elles servent d'illustration. La libre circulation des données et des programmes est un facteur décisif de développement : une seule chose est inconcevable, c'est qu'il n'y ait qu'un seul point de vue sur ces objets.*

Et pourtant, force est de constater que si les programmes sont pour beaucoup librement diffusés, les données écologiques (ou d'écologie moléculaire) sont loin d'être massivement distribuées (Parr & Cummings, 2005). Nous avons pu le vérifier au travers de nos collaborations : sur un ensemble de neuf jeux de données analysés au cours de cette thèse et ayant ou devant donner lieu à une publication, trois seulement sont libres. La libre circulation des procédures et des données conçue plus haut est donc encore mythique dans les domaines que nous avons abordés. Notons que d'autres domaines tels que la génomique ou la phylogénie moléculaire semblent à cet égard avoir pris une avance considérable, qui n'est sans doute pas étrangère à leur expansion rapide, et qui ne peut qu'encourager à suivre ce mode de fonctionnement.

Cependant, le développement d'un outil informatique pour l'analyse de données possède ceci de remarquable qu'il met directement en contact le méthodologiste avec un ensemble d'utilisateurs, dont les questions, problématiques, idées, suggestions, ou même contributions sont autant d'échanges scientifiques précieux.

Perspectives

Maintenant que le bilan de ce travail a été dressé, on peut s'intéresser aux perspectives qui en découlent ainsi qu'aux pistes qui ont été ouvertes.

Les méthodes et les outils développés ouvrent des perspectives prometteuses dans différents domaines, qu'il s'agisse d'analyse de données génétiques (méthodes K-tableaux, sPCA, package *adegenet*), de statistique spatiale multivariée (MSPA), ou d'analyse de données comparatives (pPCA). Mais l'impact de ces approches dans leur domaine respectif reste encore à évaluer. De ce point de vue, l'interaction favorisée par *adegenet* avec les utilisateurs laisse entrevoir que la sPCA pourrait occuper un rôle croissant dans l'analyse des données moléculaires géoréférencées. Un effort de documentation — sans doute sous la forme d'un tutoriel dédié — devra cependant être fourni pour accroître l'accès à la méthode. La MSPA devrait également recevoir une certaine attention de la part des écologues intéressés par l'étude des structures spatiales multi-échelles ; la méthode, pour l'instant non officiellement implémentée dans un package R, devrait assez vite intégrer le projet sedaR (`http://r-forge.r-project.org/projects/sedar/`), qui vise à réunir les différents outils d'écologie spatiale implémentés dans R. L'application aux données génétiques demandera un effort supplémentaire, incluant une implémentation dans *adegenet* et surtout une illustration biologique convaincante et librement diffusable[1]. Comme le souligne la conclusion de Jombart *et al.* (sous presse), la MSPA ouvre par ailleurs ses propres perspectives, puisque qu'elle est applicable hors du contexte spatial, par exemple dans le cadre de l'analyse

[1] Les données des ours Bruns (*Ursus arctos*) de Scandinavie constituant l'illustration présentée au chapitre 5, qui montrent l'intérêt de la MSPA dans le contexte génétique, ne sont pas pour l'instant libres de diffusion.

des séries temporelles et des données comparatives. Ce sont là des directions qui restent à explorer. Enfin, la généralisation de l'ordination sous contrainte ouvre un champ d'applications qui reste à découvrir, mais son application dans le cadre de la méthode comparative (pPCA) démontre déjà l'intérêt de cette démarche.

Ces développements méthodologiques appellent pour certains des investigations au-delà de l'application des méthodes, et l'on peut préciser dans quelles directions celles-ci pourraient s'orienter. L'utilisation de méthodes K-tableaux pour analyser la cohérence des marqueurs moléculaires multialléliques ne saurait se dispenser d'une réflexion de fond sur la mesure de la biodiversité. Un marqueur intéressant est-il un marqueur qui exprime un message consensuel ? Ou bien est-ce un marqueur qui porte justement une information originale ? Ce dernier point de vue contient une question nouvelle à laquelle les méthodes K-tableaux proposées ne sauraient répondre efficacement, et interroge donc la biométrie. Par ailleurs, l'article présentant la sPCA (Jombart *et al.*, 2008) souligne l'utilité d'une étude comparant la méthode aux approches Bayésiennes « concurrentes » (Guillot *et al.*, 2005; François *et al.*, 2006) à travers l'analyse de données simulées. Une telle étude demandera d'une part de réfléchir sur les modèles qui généreront les données (qui devront être spatialement explicites) et sans doute de les implémenter, et posera d'autre part la question de critères de comparaison des résultats de méthodes très différentes. Dans un autre contexte, l'extension de la sPCA à l'analyse de structures phylogénétiques pose également une question de fond quant à la mesure des proximités entre taxons pour une phylogénie donnée. Cette mesure doit-elle inclure la longueur des branches, ou au contraire ne prendre en compte que la topologie de l'arbre ? Cette mesure doit-elle simplement être relative à l'échantillon décrit, comme c'est le cas avec la proximité d'Abouheif, ou bien doit-elle revêtir un sens dans l'absolu ? Un dernier point digne d'intérêt consisterait à rechercher les liens existants entre les structures phylogénétiques globales et locales, basées sur l'idée d'autocorrélation, et les structures observées sous des modèles d'évolution connus, tels que le modèle Brownien ou celui d'Ornstein-Uhlenbeck.

Pour conclure, nous noterons que la voie de l'analyse statistique de données génétiques ouverte par Fisher au début du siècle passé, et qui a généré une part considérable de résultats biologiques aussi bien que statistiques, possède encore un bel avenir. Dans cet avenir, les méthodes d'ordinations en espace réduit ont bel et bien une place de choix à occuper. Les progrès technologiques constants mettent à disposition du biologiste des volumes de données sans cesse croissants et de plus en plus complexes, dont l'analyse ne saurait se résumer au calcul de trois statistiques F. Qu'il s'agisse du couplage de données génétiques, morphologiques et environnementales (Germain, 2007), du séquençage massif de génomes humains pour identifier le déterminisme génétique de pathologies majeures (`http://www.1000genomes.org`), de la mesure de la structuration géographique de la biodiversité à l'échelle du globe (Beheregaray, 2008), ou encore de suivi temporel du statut génétique des espèces et des populations à différentes échelles (Schwartz *et al.*, 2006), ces nouvelles données créeront une demande méthodologique forte. Il y a tout lieu de penser que l'analyse de données *sensu* Cailliez & Pages (1976) constituera une approche de plus en plus pertinente de l'analyse des données génétiques, dont on ne fait que

commencer à mesurer la portée.

Références bibliographiques

Abouheif E (1999). A method for testing the assumption of phylogenetic independence in comparative data. *Evolutionary Ecology Research* **1** : 895–909.

Anderson TW (1958). *An introduction to multivariate statistical analysis.* John Wiley & Sons, Inc.

Anselin L (2002). Under the hood. Issues in the specification and interpretation of spatial regression models. *Agricultural Economics* **27** : 247–267.

Anselin L, Syabri I, Smirnov O (2002). Visualizing multivariate spatial correlation with dynamically linked windows. In : Anselin L, Rey S (éds.), *New Tools for Spatial Data Analysis : Proceedings of a Workshop, CSISS, Santa Barbara, CA*.

Balloux F (2001). Easypop (version 1.7) : a computer program for population genetics simulations. *Journal of Heredity* **92** : 301–302.

Bamshad M, Wooding S, Salisbury BA, Stephens JC (2004). Deconstructing the relationship between genetics and race. *Nature Reviews Genetics* **5** : 598–609.

Banet TA, Lebart L (1984). Local and partial principal component analysis (PCA) and correspondence analysis (CA). In : *COMPSTAT 84*, Physica-Verlag, Vienna, pp. 113–123.

Barabási A, Oltvai ZN (2004). Network biology : understanding the cell's functional organization. *Nature Reviews Genetics* **5** : 101–113.

Bauwens D, Diaz-Uriarte R (1997). Covariation of life-history traits in lacertid lizards : a comparative study. *The American Naturalist* **149** : 91–111.

Beheregaray LB (2008). Twenty years of phylogeography : the state of the field and the challenges for the southern hemisphere. *Molecular Ecology* **17** : 3754–3774.

Belkhir K, Borsa P, Chikhi L, Raufaste N, Bonhomme F (2001). GENETIX. Laboratoire Génome, Populations, Interactions, CNRS UMR 5000, Université de Montpellier II, Montpellier, France.

Benzecri JP (1976a). Histoire et préhistoire de l'analyse de données. Partie I - la préhistoire. *Les cahiers de l'analyse de données* **I** : 9–32.

Benzecri JP (1976b). Histoire et préhistoire de l'analyse de données. Partie II - la biométrie. *Les cahiers de l'analyse de données* **I** : 101–120.

Bertranpetit J, Cavalli-Sforza LL (1991). A genetic reconstruction of the history of the population of the Iberian Peninsula. *Annals of Human Genetics* **55** : 51–67.

Bolker B, Butler M, Cowan P, de Vienne D, Jombart T, Kembel S, *et al.* (2007). *phylobase : Base package for phylogenetic structures and comparative data.* R package version 0.3. **URL:** *http ://phylobase.R-forge.R-project.org*

Borcard D, Legendre P (2002). All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecological Modelling* **153** : 51–68.

Borcard D, Legendre P, Avois-Jacquet C, Tuomisto H (2004). Dissecting the spatial structure of ecological data at multiple scales. *Ecology* **85** : 1826–1832.

Brind'Amour A, Boisclair D, Legendre P, Borcard D (2005). Multiscale spatial distribution of a littoral fish community in relation to environmental variables. *Limnology and Oceanography* **50** : 465–479.

Cailliez F, Pages JP (1976). *Introduction à l'analyse de données.* SMASH.

Carroll JD (1968). A generalization of canonical correlation analysis to three or more sets of variables. *Proceeding of the 76th Convention of the American Psychological Association* **3** : 227–228.

Cassar S, Jombart T, Loison A, Pontier D, Dufour AB, Jullien JM, *et al.* (in revision). Spatial genetic structure of alpine chamois (*Rupicapra rupicapra*) : a consequence of landscape features and social factors ? .

Cavalli-Sforza LL (1966). Population structure and human evolution. *Proceedings of the Royal Society of London Series B* **164** : 362–379.

Cavalli-Sforza LL, Menozzi P, Piazza A (1993). Demic expansions and human evolution. *Science* **259** : 639–646.

Cavalli-Sforza LL, Menozzi P, Piazza A (1994). *The history and geography of human genes.* Princeton University Press.

Chessel D (1992). Echanges interdisciplinaires en analyse de données écologiques. Mémoire d'habilitation à diriger des recherches, Université Claude Bernard - Lyon 1.

Chessel D, Dufour AB, Thioulouse J (2004). The ade4 package-I- one-table methods. *R News* **4** : 5–10.

Chessel D, Hanafi M (1996). Analyse de la co-inertie de K nuages de points. *Revue de statistique appliquée* **XLIV (2)** : 35–60.

Cheverud JM, Dow MM, Leutenegger W (1985). The quantitative assessment of phylogentic constaints in comparative analyses : sexual dimorphism in body weights among primates. *Evolution* **39** : 1335–1351.

Cliff AD, Ord JK (1973). *Spatial autocorrelation*. Pion, London.

Cliff AD, Ord JK (1981). *Spatial Processes. Model & Applications*. Pion, London.

Cornillon PA, Pontier D, Rochet MJ (1999). Autoregressive models for estimating phylogenetic and environmental effects : accounting for within-species variations. *Journal of Theoretical Biology* **202** : 247–256.

Dale MRT, Dixon P, Fortin MJ, Legendre P, Myers D, Rosenberg M (2002). Conceptual and mathematical relationships among methods for spatial analysis. *Ecography* **25** : 558–577.

de Jong P, Sprenger C, van Veen F (1984). On extreme values of Moran's I and Geary's c. *Geographical Analysis* **16** : 17–24.

Devillard S, Jombart T, Pontier D (sous presse). Revealing cryptic genetic structuring in an urban population of stray cats (*Felis silvestris catus*). *Mammalian Biology* .

Diniz-Filho JAF, de Sant'Ana CER, Bini LM (1998). An eigenvector method for estimating phylogenetic inertia. *Evolution* **52** : 1247–1262.

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

Dormann CF, McPherson JM, Araújo MB, Bivand R, Bolliger J, Carl G, *et al.* (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data : a review. *Ecography* **30** : 609–628.

Dray S (2003). Éléments d'interface entre analyses multivariées, systèmes d'information géographique et observations écologiques. Thèse de Doctorat, Université Claude Bernard - Lyon 1.

Dray S, Chessel D, Thioulouse J (2003). Co-inertia analysis and the linking of ecological tables. *Ecology* **84** : 3078–3089.

Dray S, Dufour AB (2007). The ade4 package : implementing the duality diagram for ecologists. *Journal of Statistical Software* **22(4)** : 1–20.

Dray S, Dufour AB, Chessel D (2007). The ade4 package - II : Two-table and K-table methods. *R News* **7** : 47–54.

Dray S, Legendre P, Peres-Neto P (2006). Spatial modelling : a comprehensive framework for principal coordinate analysis of neighbours matrices (PCNM). *Ecological Modelling* **196** : 483–493.

Dray S, Saïd S, Debias F, Chessel D (2008). Spatial ordination of vegetation data using a generalization of Wartenberg's multivariate spatial correlation. *Journal of Vegetation Science* **19** : 45–56.

Dungan JL, Perry JN, Dale MRT, Legendre P, Citron-Pousty S, Fortin MJ, *et al.* (2002). A balanced view of scale in spatial analysis. *Ecography* **25** : 626–640.

Dupanloup I, Schneider S, Excoffier L (2002). A simulated annealing approach to define the genetic structure of populations. *Molecular Ecology* **11** : 2571–2581.

Edwards AWF (1990). R. A. Fisher twice professor of genetics : London and Cambridge *or* "a fairly well-known geneticist". *Biometrics* **46** : 897–904.

Edwards AWF (2003). Human genetic diversity : Lewontin's fallacy. *BioEssays : news and reviews in molecular, cellular and developmental biology* **25** : 798–801.

Escofier B (1994). Multiple factor analysis (AFMULT package). *Computational Statistics & Data Analysis* **18** : 121–140.

Escoufier Y (1973). Le traitement des variables vectorielles. *Biometrics* **29** : 751–760.

Escoufier Y (1987). The duality diagramm : a means of better practical applications. In : Legendre P, Legendre L (éds.), *Development in numerical ecology*, NATO advanced Institute , Serie G .Springer Verlag, Berlin, pp. 139–156.

Excoffier L, Heckel G (2006). Computer programs for population genetics data analysis : a survival guide. *Nature Reviews Genetics* **7** : 745–758.

Excoffier L, Smouse PE, Quattro JM (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes : applications to human mitochondrial DNA restriction data. *Genetics* **131** : 479–491.

Falush D, Stephens M, Pritchard JK (2003). Inference of population structure using multilocus genotype data : linked loci and correlated allele frequencies. *Genetics* **164** : 1567–1587.

Fisher RA (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh* **52** : 399–433.

François O, Ancelet S, Guillot G (2006). Bayesian clustering using hidden markov random fields in spatial population genetics. *Genetics* **174** : 805–816.

Fraser DJ, Bernatchez L (2001). Adaptive evolutionary conservation : towards a unified concept for defining conservation units. *Molecular Ecology* **10** : 2741– 2752.

Garland TJ, Ives AR (2000). Using the past to predict the present : confidence intervals for regression equations in phylogenetic comparative methods. *The American Naturalist* **155** : 346–364.

Geary RC (1954). The contiguity ratio and statistical mapping. *The Incorporated Statistician* **5** : 115–145.

Germain E (2007). Approche éco-éthologique de l'hybridation entre le chat forestier d'Europe (*Felis sylvestris sylvestris* Schreber 1777) et le chat domestique (*Felis catus* L.). Thèse de Doctorat, 2C2A-CERFE, Centre de recherche et de formation en éco-éthologie.

Getis A, Griffith D (2002). Comparative spatial filtering in regression analysis. *Geographical Analysis* **34** : 130–140.

Gittleman JL, Kot M (1990). Adaptation : statistics and a null model for estimating phylogenetic effects. *Systematic Zoology* **39** : 227–241.

Gort G, Koopman WJM, Stein A (2006). Fragment length distributions and collision probabilities for AFLP markers. *Biometrics* **62** : 1107–1115.

Gort G, Koopman WJM, Stein A, van Eeuwijk FA (2008). Collision probabilities for AFLP bands, with an application to simple measures of genetic similarity. *Journal of Agricultural, Biological & Environmental Statistics* **13** : 177–198.

Goudet J (2005). HIERFSTAT, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Notes* **5** : 184–186.

Goudet J, Raymond M, Meeüs T, Rousset F (1996). Testing differentiation in diploid populations. *Genetics* **144** : 1933–1940.

Grafen A (1989). The phylogenetic regression. *Philosophical Transactions of the Royal Society of London Series B - Biology* **326** : 119–157.

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

Griffith D (2000). A linear regression solution to the spatial autocorrelation problem. *Journal of Geographical Systems* **2** : 141–156.

Griffith D, Peres-Neto P (2006). Spatial modeling in ecology : the flexibility of eigenfunction spatial analyses. *Ecology* **87** : 2603–2613.

Griffith DA (1996). Spatial autocorrelation and eigenfunctions of the geographic weights matrix accompanying geo-referenced data. *The Canadian Geographer* **40** : 351–367.

Guillot G, Estoup A, Mortier F, Cosson JF (2005). A spatial statistical model for landscape genetics. *Genetics* **170** : 1261–1280.

Guillot G, Mortier F, Estoup A (2006). Geneland : a computer package for landscape genetics. *Molecular Ecology Notes* **5** : 712–715.

Hanski IA, Simberloff D (1997). *Metapopulation biology : Ecology, Genetics and Evolution*, Academic Press, chap. The metapopulation approach, its history, conceptual domain, and application to conservation, pp. 5–26.

Harvey PH, Pagel M (1991). *The Comparative Method in Evolutionary Biology*. Oxford University Press.

Harville DA (1997). *Matrix algebra from a statistician's perspective*. Springer, New York.

Holmes S (2006). *Festschrift for David Freedman*, IMS Lecture Notes, chap. Multivariate data analysis : the French way.

Hotelling H (1933a). Analysis of a complex of statistical variables into principal components. *The Journal of Educational Psychology* **24** : 417–441.

Hotelling H (1933b). Analysis of a complex of statistical variables into principal components (continued from september issue). *The Journal of Educational Psychology* **24** : 498–520.

Hotelling H (1936). Relations between two sets of variables. *Biometrika* **28** : 321–327.

Ives AR, Zhu J (2006). Statistics for correlated data : phylogenies, space and time. *Ecological Applications* **16** : 20–32.

Johnson FM, Schaffer HE (1973). Isozyme variability in species of the genus drosophila. VII. Genotype-environment relationships in populations of *D. melanogaster* from the eastern United States. *Biochemical Genetics* **10** : 149–163.

Johnson FM, Schaffer HE, Gillaspy JE, Rockwood ES (1969). Isozyme genotype-environment relationships in natural populations of the harvester ant, *Pogonomyrmex barbatus*, from Texas. *Biochemical Genetics* **3** : 429–450.

Joliffe IT (2004). *Principal Component Analysis*. Springer Series in Statistics. Springer, 2ème édition.

Jombart T (2008). adegenet : a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24** : 1403–1405.

Jombart T, Devillard S, Dufour AB, Pontier D (2008). Revealing cryptic spatial patterns in genetic variability by a new multivariate method. *Heredity* **101** : 92–103.

Jombart T, Dray S, Dufour AB (sous presse). Finding essential scales of spatial variation in ecological data : a multivariate approach. *Ecography* .

Jombart T, Dray S, Pavoine S, Dufour AB (en prép-a). A general framework to constrained reduced space ordinations using Moran's I .

Jombart T, Moazami-Goudarzi K, Dufour AB, Laloë D (2006). Fréquences alléliques et cohérence entre marqueurs moléculaires : des outils descriptifs. *Les Actes du BRG* **6** : 25–39.

Jombart T, Pavoine S, Dufour AB, Pontier D (en prép-b). Exploring phylogeny as a source of ecological variation : a methodological approach .

Jombart T, Pontier D, Dufour AB (in revision). Genetic markers in the playground of multivariate analysis. *Heredity* .

Joshi A (1997). Sir R A Fisher and the evolution of genetics. *Resonance* **2** : 27–31.

Laloë D, Jombart T, Dufour AB, Moazami-Goudarzi K (2007). Consensus genetic structuring and typological value of markers using multiple co-inertia analysis. *Genetics Selection Evolution* **39** : 545–567.

Laloë D, Moazami-Goudarzi K, Chessel D (2002). Contribution of individual markers to the analysis of the relationships among breeds by correspondence analysis. In : *7th world congress on genetics applied livestock production*. Montpellier, France.

Laval G, Iannuccelli N, Legault C, Milan D, Groenen AM, Guiffra E, *et al.* (2000). Genetic diversity of eleven european pig breeds. *Genetic Selection Evolution* **32** : 187–203.

Lavit C, Escoufier Y (1994). The ACT (STATIS method). *Computational Statistics & Data Analysis* **18** : 97–119.

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

Lebart L, Morineau A, Piron M (2004). *Statistique exploratoire multidimensionnelle.* DUNOD.

Lee SI (2001). Developing a bivariate spatial association measure : an integration of Pearson's r and Moran's I. *Journal of Geographical Systems* **3** : 369–385.

Legay JM (1997). *L'expérience et le modèle - Un discours sur la méthode.* INRA Éditions.

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

Legendre P, Borcard D (2003). Quelles sont les échelles spatiales importantes dans un écosystème ? In : Droesbeke JJ, Lejeune M, Saporta G (éds.), *Analyse statistique de données spatiales*, Paris. Technip. édition.

Legendre P, Fortin MJ (1989). Spatial pattern and ecological analysis. *Vegetatio* **80** : 107–138.

Legendre P, Legendre L (1998). *Numerical ecology.* Elsevier Science B. V., Amsterdam.

Levin SA (1992). The problem of pattern and scale in ecology. *Ecology* **73** : 1943–1967.

Liebhold AM, Gurevitch J (2002). Integrating the statistical analysis of spatial data in ecology. *Ecography* **25** : 553–557.

Loison A, Jullien JM, Menaut P (1999). Subpopulation structure and dispersal in two populations of chamois. *Journal of Mammalogy* **80** : 620–632.

MacHugh DE, Loftus RT, Cunningham P, Bradley DG (1998). Genetic structure of seven European cattle breeds assessed using 20 microsatellite markers. *Animal Genetics* **29** : 333–340.

Manel S, Bellemain E, Swenson JE, François O (2004). Assumed and inferred spatial structure of populations : the Scandinavian brown bears revisited. *Molecular Ecology* **13** : 1327–1331.

Manel S, Berthoud F, Bellemain E, Gaudeul M, Luikart G, Swenson JE, *et al.* (2007). A new individual-based spatial approach for identifying genetic discontinuities in natural populations. *Molecular Ecology* **16** : 2031–2043.

Manel S, Schwartz MK, Luikart G, Taberlet P (2003). Landscape genetics : combining landscape ecology and population genetics. *Trends in Ecology & Evolution* **18** : 189–197.

Manni F, Guérard E, Heyer E (2004). Geographic patterns of (genetic, morphologic, linguistic) variation : how barriers can be detected by "Monmonier's algorithm". *Human Biology* **76** : 173–190.

Martins EP, Hansen TF (1997). Phylogenies and the comparative method : a general approach to incorporating phylogenetic information into the analysis of interspecific data. *The American Naturalist* **149** : 646–667.

McKay JK, Latta RG (2002). Adaptive population divergence : markers, QTL and traits. *Trends in Ecology & Evolution* **17** : 285–291.

Menozzi P, Piazza A, Cavalli-Sforza LL (1978). Synthetic maps of human gene frequencies in Europeans. *Science* **201** : 786–792.

Mitton JB (1978). Measurement of differentiation : reply to Lewontin, Powell and Taylor. *The American Naturalist* **112** : 1142–1144.

Moazami-Goudarzi K, Belemsaga DMA, Ceriotti G, Laloë D, Fagbohoun F, Kouagou NT, *et al.* (2001). Caractérisation de la race bovine Somba à l'aide de marqueurs moléculaires. *Revue d'Elevage et de Médecine Vétérinaire des pays Tropicaux* **54** : 1–10.

Moazami-Goudarzi K, Laloë D (2002). Is a multivariate consensus representation of genetic relationships among populations always meaningful ? *Genetics* **162** : 473–484.

Monmonier M (1973). Maximum-difference barriers : an alternative numerical regionalization method. *Geographical Analysis* **3** : 245–261.

Moran PAP (1948). The interpretation of statistical maps. *Journal of the Royal Statistical Society, B* **10** : 243–251.

Moran PAP (1950). Notes on continuous stochastic phenomena. *Biometrika* **37** : 17–23.

Moritz C (1994). Defining evolutionary significant units for conservation. *Trends in Ecology & Evolution* **9** : 373–375.

Mulley JC, James W, Barker JSF (1979). Allozyme genotype-environment relationships in natural populations of *Drosophila buzzatii*. *Biochemical Genetics* **17** : 105–126.

Ollier S (2004). Des outils pour l'intégration des questions spatiales, temporelles et évolutives en analyse des données écologiques. Thèse de Doctorat, Université Claude Bernard, Lyon 1.

Paetkau D (1999). Using genetics to identify intraspecific conservation units : a critique of current methods. *Conservation Biology* **13** : 1507–1509.

Pagès J (1996). Éléments de comparaison entre l'analyse factorielle multiple et la méthode STATIS. *Revue de Statistique Appliquée* **44** : 81–95.

Palsboll PJ, Bérubé M, Allendorf FW (2006). Identification of management units using population genetic data. *Trends in Ecology & Evolution* **22** : 11–16.

Paradis E, Claude J, Strimmer K (2004). APE : analyses of phylogenetics and evolution in R language. *Bioinformatics* **20** : 289–290.

Pariset L, Savarese MC, Cappuccio I, Valentini A (2003). Use of microsatellites for genetic variation and inbreeding analysis in Sarda sheep flocks of central Italy. *Journal of Animal Breeding and Genetics* **120** : 425–432.

Parr CS, Cummings MP (2005). Data sharing in ecology and evolution. *Trends in Ecology & Evolution* **20** : 362–363.

Pavoine S (2005). Méthodes statistiques pour la mesure de la biodiversité. Thèse de Doctorat, Université Claude Bernard - Lyon 1.

Pavoine S, Bailly X (2007). New analysis for consistency among markers in the study of genetic diversity : development and application to the description of bacterial diversity. *BMC Evolutionary Biology* **7** : 156.

Pavoine S, Dufour AB, 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, Pontier D, Chessel D (2008). Testing for phylogenetic signal in life history variable : Abouheif's test revisited. *Theoretical Population Biology* **73** : 79–91.

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

Peres-Neto P (2006). A unified strategy for estimating and controlling spatial, temporal and phylogenetic autocorrelation in ecological models. *Oecologica Brasiliensis* **10** : 105–119.

Perry JN, Liebhold AM, Rosenberg MS, Dungan J, Miriti M, Jakomulska A, *et al.* (2002). Illustrations and guidelines for selecting statistical methods for quantifying spatial patterns in ecological data. *Ecography* **25** : 578–600.

Piegorsch WW (1990). Fisher's contribution to genetics and heredity, with special emphasis on the Gregor Mendel controversy. *Biometrics* **46** : 915–924.

Pontier J, Dufour AB, Normand M (1990). *Le modèle euclidien en analyse des données.* Ellipses.

Powell JR, Taylor CE (1978). Are human races "substantially" different genetically ? *The American Naturalist* **112** : 1139–1142.

Pramual P, Kuvangkadilok C, Baimai V, Walton C (2005). Phylogeography of the black fly *Simulium tani* (Diptera : Simuliidae) from Thailand as inferred from mtDNA sequences. *Molecular Ecology* **14** : 3989–4001.

Preziosi RF, Fairbairn DJ (1992). Genetic population structure and levels of gene flow in the stream dwelling waterstrider *Aquarius* (= *Gerris*) *remigis* (Emiptera : Geridae). *Evolution* **46** : 430–444.

Pritchard JK, Stephens M, Donnelly P (2000). Inference of population structure using multilocus genotype data. *Genetics* **155** : 945–959.

R Development Core Team (2008). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. **URL:** *http ://www.R-project.org*

Reyment RA (2005). The statistical analysis of multivariate serological frequency data. *Bulletin of Mathematical Biology* **67** : 1303–1313.

Rosenberg NA (2005). Algorithms for selecting informative marker panels for population assignment. *Journal of Computational Biology* **12** : 1183–1201.

Saporta G (1990). *Probabilités, analyse des données et statistique*. Technip.

Schaffer HE, Johnson FM (1974). Isozyme allelic frequencies related to selection and gene-flow hypotheses. *Genetics* **77** : 163–168.

Schlötterer C (2004). The evolution of molecular markers - just a matter of fashion ? *Genetics* **5** : 63–69.

Schwartz MK, Luikart G, Waples RS (2006). Genetic monitoring as a promising tool for conservation and management. *Trends in Ecology & Evolution* **22** : 25–33.

Seal HL (1966). *Multivariate Statistical Analysis for Biologists*. Methuen and co.

Slatkin M (1985). Gene flow in natural populations. *Annual Reviews of Ecology and Systematics* **16** : 393–430.

Smouse PE, Peakall R (1999). Spatial autocorrelation analysis of individual multiallele and multilocus genetic structure. *Heredity* **82** : 561–573.

Sokal RR, Oden NL (1978). Spatial autocorrelation in biology 1. Methodology. *Biological Journal of the Linnean Society* **10** : 199–228.

Sokal RR, Wartenberg DE (1983). A test of spatial autocorrelation analysis using an isolation-by-distance model. *Genetics* **105** : 219–237.

Swenson JE, Sandegren F, Soderberg F (1998). Geographic expansion of an increasing brown bear population : evidence for presaturation dispersal. *Journal of Animal Ecology* **67** : 819–826.

Taberlet P, Swenson JE, Sandegren F, Bjarvall A (1995). Localization of a contact zone between two highly divergent mitochondrial DNA lineages of the brown bear *Urso arctos* in Scandinavia. *Conservation Biology* **9** : 1255–1261.

Takeuchi K, Yanai H, Mukherjee BN (1984). *The foundations of multivariate analysis : a unified approach by means of projection onto linear subspaces*. Wiley Eastern Limited.

Taylor BL, Dizon AE (2002). First policy then science : why a management unit based solely on genetic criteria cannot work. *Molecular Ecology* **8** : S11–S16.

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

Thompson EA (1990). R. A. Fisher's contributions to genetical statistics. *Biometrics* **46** : 905–914.

Tiefelsdorf M, Boots B (1995). The exact distribution of Moran's I. *Environment and Planning A* **27** : 985–999.

Tiefelsdorf M, Boots B (1996). Letters to the editor : the exact distribution of Moran's I. *Environment and Planning A* **28** : 1900.

Tiefelsdorf M, Griffith DA (2007). Semiparametric filtering of spatial autocorrelation : the eigenvector approach. *Environment and Planning A* **39** : 1193–1221.

Tolley KA, Burger M, Turner AA, Matthee CA (2006). Biogeographic patterns and phylogeography of dwarf chameleons (*Bradypodion*) in an African biodiversity hotspot. *Molecular Ecology* **15** : 781–793.

Toro MA (2006). Assessing genetic diversity between breeds for conservation. *Journal of Animal Breeding and Genetics* **123** : 289.

Turner MG, Dale VH, Gardner RH (1989). Predicting across scales : Theory development and testing. *Landscape Ecology* **3** : 245–252.

Wagner HH, Fortin MJ (2005). Spatial analysis of landscapes : concepts and statistics. *Ecology* **86** : 1975–1987.

Waits L, Taberlet P, Swenson JE, Sandegren F, Franzen R (2000). Nuclear DNA microsatellite analysis of genetic diversity and gene flow in the Scandinavian brown bear (*Ursus arctos*). *Molecular Ecology* **9** : 421–431.

Warnes GR (2003). The genetics package. *R News* **3** : 9–13.

Wartenberg DE (1985). Multivariate spatial correlations : a method for exploratory geographical analysis. *Geographical Analysis* **17** : 263–283.

Weir BS (1996). *Genetic data analysis II.* Sinauer Associates, Sunderland, Massachussetts.

Wiens JA (1989). Spatial scaling in ecology. *Functionnal Ecology* **3** : 385–397.

Wright S (1943). Isolation by distance. *Genetics* **28** : 114–138.

Xuebin Q, Jianlin H, Chekavora I, Badamdorj D, Rege JEO, Hanotte O (2005). Genetic diversity and differentiation of Mongolian and Russian yak populations. *Journal of Animal Breedings and Genetics* **122** : 117–126.

Yoccoz G (1988). Le rôle du modèle euclidien d'analyse des données en biologie évolutive. Thèse de Doctorat, Université Claude Bernard - Lyon 1.

Titre en français

Analyses multivariées de marqueurs génétiques : développements méthodologiques, applications et extensions.

Résumé en français

Cette thèse de biométrie s'inscrit dans le cadre de l'analyse multivariée de données de marqueurs moléculaires. Après avoir dressé un bilan de l'état actuel de l'utilisation de l'analyse multivariée pour extraire de l'information biologique des marqueurs génétiques, nous proposons des réponses méthodologiques à des problématiques précises : l'analyse de la cohérence typologique de l'information génétique, la recherche de structures génétiques spatiales, et l'étude des principales échelles de cette structuration. Certaines de ces approches ont donné lieu à des extensions hors du contexte génétique : nous proposons également une méthode d'analyse des structures phylogénétiques, ainsi qu'un cadre plus général à l'analyse multivariée sous contrainte d'autocorrélation. En parallèle du développement méthodologique, un logiciel pour l'analyse des données génétiques et qui implémente les méthodes proposées a été développé.

Mots-clés en français

analyse multivariée ; ordination ; marqueur moléculaire ; génétique ; spatial ; phylogénie ; logiciel

Titre en anglais

Multivariate analyses of genetic markers : methodological developments, applications and extensions.

Résumé en anglais

This PhD thesis in biometry focuses on the multivariate analysis of molecular markers. After providing a review of current applications of multivariate analyses to extract biological information from genetic markers, we present some methods that we have developed to tackle specific issues. First, a methodological proposition is made regarding the study of the typological coherence of genetic markers. Then, we present methods that have been developed to investigate spatial patterns in genetic variability, and to identify these patterns at multiple scales. Some of these approaches have been extended outside the genetic context : we propose a method to investigate phylogenetic patterns in a set of biological traits, and we describe a general framework for multivariate analysis constrained by autocorrelation. The proposed methodology as well as other tools are implemented in a new software dedicated to the analysis of genetic markers.

Mots-clés en anglais

multivariate analysis ; ordination ; molecular marker ; genetics ; spatial ; phylogeny ; software

Discipline : Biostatistique

Intitulé et adresse du laboratoire :

Université de Lyon, F-69000, Lyon ; Université Lyon 1 ; CNRS, UMR5558, Laboratoire de Biométrie et Biologie Evolutive, F-69622, Villeurbanne, France. Laboratoire de Biométrie et Biologie Évolutive - UMR 5558 CNRS

Bâtiment Gregor Mendel - Université Claude Bernard Lyon1